JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA

SLOVENSKEJ AKADÉMIE VIED

1

ROČNÍK 74, 2023

sciendo

SAP
SLOVAK ACADEMIC PRESS

JAZYKOVEDNÝ ČASOPIS
VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS
SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, v. v. i.
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s. r. o.
- elektronicky vo vydavateľstve Sciendo – De Gruyter
  https://content.sciendo.com/view/journals/jazcas/jazcas-overview.xml

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
Kontakt: jazykovedny.casopis@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: https://www.juls.savba.sk/ediela/jc/, https://www.sav.sk/?lang=sk&doc=journal-list&journal_no=26

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2023/74/1) – september 2023

Quartile ranking 2022: Q2
CiteScore 2022: 0,5
SCImago Journal Rank (SJR) 2022: 0,296
Source Normalized Impact per Paper (SNIP) 2022: 0,721

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is covered by the following services: Baidu Scholar; Cabell's Directory; CEJSH (The Central European Journal of Social Sciences and Humanities); CEEOL (Central and Eastern European Online Library); CNKI Scholar (China National Knowledge Infrastructure); CNPIEC – cnpLINKer; Dimensions; DOAJ (Directory of Open Access Journals); EBSCO (relevant databases); EBSCO Discovery Service; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); Genamics JournalSeek; Google Scholar; IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); International Medieval Bibliography; J-Gate; JournalGuide; JournalTOCs; KESLI-NDSL (Korean National Discovery for Science Leaders); Linguistic Bibliography; Linguistics Abstracts Online; Microsoft Academic; MLA International Bibliography; MyScienceWork; Naver Academic; Naviga (Softweco); Primo Central (ExLibris); ProQuest (relevant databases); Publons; QOAM (Quality Open Access Market); ReadCube; SCImago (SJR); SCOPUS; Semantic Scholar; Sherpa/RoMEO; Summon (ProQuest); TDNet; Ulrich's Periodicals Directory/ulrichsweb; WanFang Data; WorldCat (OCLC).

ISSN 0021-5597 (tlačená verzia/print) **ISSN 1338-4287 (verzia online)** **MIČ 49263**

JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA
SLOVENSKEJ AKADÉMIE VIED

1

ROČNÍK 74, 2023

Natural Language Processing and Corpus Linguistics

SLOVKO 2023

Tematické číslo Jazykovedného časopisu venované
počítačovému spracovaniu prirodzeného jazyka a korpusovej lingvistike.

Prizvané editorky:
Katarína Gajdošová
Adriána Žáková

sciendo

SAP
SLOVAK ACADEMIC PRESS

CONTENTS

5 Foreword
6 Predhovor

CORPUS-BASED AND CORPUS-DRIVEN RESEARCH

9 RENATA BRONIKOWSKA: Verbification of Feminine Forms of Adjectives *można* 'possible', *niemożna* 'impossible' and *niepodobna* 'impossible' – Corpus-based Approach
19 EVGENIYA BUDENNAYA, KRISTINA LITVINTSEV AND ANASTASIA YAKOVLEVA: God Knows How It Turns Out: On Three Constructions Including *bog* 'god', *čert* 'devil' and Some Taboo Words in the Russian Language over the Last Three Centuries
32 JAROSLAV DAVID, TEREZA KLEMENSOVÁ AND MICHAL MÍSTECKÝ: Appellativization of Proper Names – In the Perspective of Corpus Analysis
43 MARTIN DIWEG-PUKANEC: The Economy of Czech Exchange in the Slovak Marketplace of Austria after the Fall of Hungary
52 ŁUKASZ GRABOWSKI: Statistician, Programmer, Data Scientist? Who Is, or Should Be, a Corpus Linguist in the 2020s?
60 JAKOB HORSCH: Corroborating Corpus Data with Elicited Introspection Data: A Case Study
70 EDYTA JURKIEWICZ-ROHRBACHER: Dative Ambiguity in Russian: A Corpus Induced Study
81 FILIP KALAŠ: The Competition of German Adjectival Suffixes
92 MARIE KOPŘIVOVÁ AND KATEŘINA ŠICHOVÁ: Proverbs in Contemporary Czech. Corpus Probe into Written Texts
100 MAGDALENA MAJDAK: Keywords in Religious Literature of 17th and 18th Centuries in Light of the Data from the Electronic Corpus of 17th- and 18th-century Polish Texts
108 MARIE MIKULOVÁ: Expressing Measure in Czech (A Corpus-based Study)
119 AKSANA SCHILLOVÁ: Adverbs Derived from Adjectival Present Participles in Polish, Slovak and Czech: A Comparative Corpus-based Study
130 BARBORA ŠTĚPÁNKOVÁ, JANA ŠINDLEROVÁ AND LUCIE POLÁKOVÁ: The Epistemic Marker *určitě* in the Light of Corpus Data
140 Miroslav ZUMRÍK: Comparative Lexical Analysis of Noun Lemmas in Slovak Judicial Decisions

LANGUAGE ACQUISITION, CREATION AND USE OF LANGUAGE RESOURCES

153 CRISTINA FERNÁNDEZ-ALCAINA, EVA FUČÍKOVÁ, JAN HAJIČ AND ZDEŇKA UREŠOVÁ: Spanish Synonyms as Part of a Multilingual Event-type Ontology
163 KATARÍNA GAJDOŠOVÁ, PETRA ŠVANCAROVÁ AND MICHAELA MOŠAŤOVÁ: Errors in the Congruent Attribute among Students Learning Slovak as a Foreign Language (Learner Corpus-based)
173 EDUARD KLYSHINSKY, ANNA BOGDANOVA AND MIKHAIL KOPOTEV: Towards a Corpus-based Dictionary of Verbal Government for the Russian Language
182 VERONIKA KOLÁŘOVÁ, VÁCLAVA KETTNEROVÁ AND JIŘÍ MÍROVSKÝ: Through Derivational Relations to Valency of Non-verbal Predicates in the NOMVALLEX Lexicon

193 MICHAELA NOGOLOVÁ, MICHAELA HANUŠKOVÁ, MIROSLAV KUBÁT AND RADEK ČECH: Linear Dependency Segments in Foreign Language Acquisition: Syntactic Complexity Analysis in Czech Learners' Texts
204 MARTINA WACLAWIČOVÁ: Differences in Spoken Language Processing in General Corpora (ORAL, ORTOFON) and in a Specialized Corpus (DIALEKT) and Their Reflection in the Mapka Application
214 DANIEL ZEMAN, PAVEL KOSEK, MARTIN BŘEZINA AND JIŘÍ PERGLER: Morpho-syntactic Annotation in Universal Dependencies for Old Czech

CORPUS BUILDING

225 ILIA AFANASEV, OLGA LYASHEVSKAYA, STEFAN REBRIKOV, YANA SHISHKINA, IGOR TROFIMOV AND NATALIA VLASOVA: The Effect of (Historical) Language Variation on the East Slavic Lects Lemmatisers Performance
234 VLADIMÍR PETKEVIČ AND HANA SKOUMALOVÁ: Annotation of Analytic Verb Forms in Czech – Complex Cases
244 PETR POŘÍZKA: CapekDraCor: A New Contribution to the European Programmable Drama Corpora
254 ALEXANDR ROSEN: The *InterCorp* Parallel Corpus with a Uniform Annotation for All Languages
266 DMITRI SITCHINAVA: Multiple Interpretation and Fragmented Texts within a Historical Corpus: The Case of Old East Slavic Vernacular Writing
275 LUCIE BENEŠOVÁ, KLÁRA PIVOŇKOVÁ AND MARTIN STLUKA: Lemmatization of the DIA1900 Diachronic Corpus

NATURAL LANGUAGE PROCESSING AND DIGITAL HUMANITIES

287 MARTIN BRAXATORIS AND ANITA BRAXATORISOVÁ: Use of Computer and Corpus Tools in the Research of a 19th Century German-language Manuscript Book of Notes and Extracts
301 NATALIIA ČASNOCHOVÁ ZOZUK: Lexical Diversity and Language Impairment
310 DÁVID DRŽÍK AND KIRSTEN ŠTEFLOVIČ: Text Vectorization Techniques Based on Wordnet
323 DANIEL HLÁDEK, MAROŠ HARAHUS, JÁN STAŠ AND MATÚŠ PLEVA: Slovak Language Models for Basic Preprocessing Tasks in Python
333 RICHARD HOLAJ AND PETR POŘÍZKA: ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech
345 NIKITA LOGIN: Distractor Generation for Lexical Questions Using Learner Corpus Data
357 JAKUB MACHURA, HANA ŽIŽKOVÁ, ADAM FRÉMUND AND JAN ŠVEC: Is it Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction
369 ONDŘEJ PEKÁČEK AND IRENE ELMEROT: When Is a Crisis Really a Crisis? Using NLP and Corpus Linguistic Methods to Reveal Differences in Migration Discourse across Czech Media
381 JÁN STAŠ, DANIEL HLÁDEK AND TOMÁŠ KOCTÚR: Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0
391 MARKÉTA ZIKOVÁ, MARTIN BŘEZINA, RADEK ČECH AND PAVEL KOSEK: Syllabic Consonants in Historical Czech and How to Identify Them

FOREWORD

In Corpus Linguistics – such as in other disciplines – there is an ongoing discussion on the topic of research methods, in the case of Corpus Linguistics regarding e. g. effective corpus building. It is stressed that when analyzing corpora, one should not ignore the effect of genre and register on the quantitative and qualitative features of texts, as recently repeated in the book *Registers in Czech* by Václav Cvrček and colleagues (2020). Similarly, Radek Čech, Pavel Kosek, Ján Mačutek a Olga Navrátilová explain why it preferable to *"sometimes avoid mixing texts"* (2020). The idea behind building corpora has traditionally been to pursue a linear, even exponential growth of corpus size, in other words, "the more data, the better". Nevertheless, with reference to the metaphor of "tide" in the article on *"function and form in language theory and research"* by Robert de Beaugrande (1996), it is possible to say that even this tide is turning. On the one hand, the interdisciplinary approach to complex linguistic issues still stands strong. On the other hand, voices can be heard according to which applying multiple analytic methods and data sources simultaneously is just as important as combining individual domains of knowledge. A case for combining corpus data and elicited introspective data is also made by Jakob Horsch in one of the contributions at the 12th international conference SLOVKO – which we here present to the readers in the special issue of the Journal of Linguistics (*Jazykovedný časopis*). The presented papers can serve as evidence for the fact that in order to provide a satisfactory analysis of multilayered language phenomena, it is not enough to simply enlarge the amount of data or to seek the ultimate, single analytical method. In such methodological polyglossia, the linguists also need a unifying discussion platform, so that the individual approaches can be mutually combined in a fruitful manner. To provide linguists with such a platform belongs to the main aims of our conference. We organized the articles according to the following idea: when traveling, both digression to interesting places, as well as returns to the main research road are important. Therefore, we divided the contributions into four broad thematic sections: 1. Corpus-based and corpus-driven research, 2. Language acquisition, creation and use of language resources, 3. Corpus building, 4. Natural Language Processing and Digital Humanities. We hope you will enjoy the reading.

Miroslav Zumrík

PREDHOVOR

Vo všetkých disciplínach prebieha neustála diskusia o výskumných metódach, v prípade korpusovej lingvistiky napríklad o efektívnom budovaní korpusov. Akcentuje sa, že pri korpusových analýzach netreba prehliadať vplyv žánru a registra na kvantitatívne a kvalitatívne charakteristiky textov, ako sa to pripomína v práci *Registre v češtine* od Václava Cvrčeka a kolektívu (2020). Radek Čech, Pavel Kosek, Ján Mačutek a Olga Navrátilová (2020) podobne vysvetľujú, prečo treba *„texty (niekedy) nemiešať"* (2020). V otázke budovania korpusov sa tradične vychádzalo z úsilia o lineárny až exponenciálny rast veľkosti korpusov v zmysle „čím viac dát, tým lepšie". S odkazom na metaforu vlny v príspevku o *„funkcii a forme v jazykovej teórii a výskume"* od Roberta de Beaugranda (1996) však možno konštatovať, že aj táto vlna sa obracia. V korpusovej lingvistike sa na jednej strane naďalej zdôrazňuje interdisciplinárny prístup ku komplexným lingvistickým otázkam, na druhej strane sa ozývajú hlasy, že okrem prepájania oblastí poznania je potrebné aj aplikovanie viacerých analytických metód a zdrojov dát súčasne. Za kombináciu korpusových a elicitovaných introspektívnych dát sa vyslovuje aj Jakob Horsch v jednom z príspevkov z dvanástej medzinárodnej konferencie SLOVKO, ktoré čitateľom predstavujeme v tomto špeciálnom čísle Jazykovedného časopisu. Prezentované príspevky môžu slúžiť ako dôkaz, že na uspokojivú analýzu mnohovrstvových jazykových javov nestačí iba jednoduché zväčšovanie objemu výskumného materiálu či hľadanie optimálnej analytickej metódy. V takomto metodologickom mnohohlase je zároveň potrebná jednotiaca platforma, aby sa jednotlivé prístupy navzájom dokázali plodne kombinovať. Poskytnúť lingvistom takýto spoločný priestor na diskusiu patrí k hlavným cieľom našej konferencie. Pri členení textov v časopise sme vychádzali z predstavy, že ako pri putovaní krajinou aj v korpusovej lingvistike majú svoje miesto odbočky k zaujímavým miestam, ako aj návraty k hlavnej línii výskumu. Štúdie sme preto rozdelili do štyroch širokých tematických celkov: *1. korpusovo podporovaný a riadený výskum, 2. jazyková akvizícia, tvorba a využívanie jazykových zdrojov, 3. budovanie korpusov, 4. počítačové spracovanie prirodzeného jazyka a digitálne humanitné a spoločenské vedy*. Prajeme vám príjemné čítanie.

Miroslav Zumrík

CORPUS-BASED AND CORPUS-DRIVEN RESEARCH

sciendo
DOI 10.2478/jazcas-2023-0019

VERBIFICATION OF FEMININE FORMS OF ADJECTIVES *MOŻNA* ‘POSSIBLE’, *NIEMOŻNA* ‘IMPOSSIBLE’ AND *NIEPODOBNA* ‘IMPOSSIBLE’ – CORPUS-BASED APPROACH

RENATA BRONIKOWSKA
Department of the History of the 17th and 18th Century Polish,
Institute of Polish Language, Polish Academy of Sciences, Warsaw, Poland

BRONIKOWSKA, Renata: Verbification of Feminine Forms of Adjectives *można* ‘possible’, *niemożna* ‘impossible’ and *niepodobna* ‘impossible’ – Corpus-based Approach. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 9 – 18.

Abstract: The article is devoted to the process taking place in the Middle Polish period, which led to the transformation of nominative, singular, feminine forms of three adjectives (*można* ‘possible’, *niemożna* ‘impossible’ and *niepodobna* ‘impossible’) into verbal lexemes (the so-called predicatives). In this respect, the predicative uses of these forms in the texts collected in the Electronic Corpus of 17th- and 18th-century Polish Texts (up to 1772) were investigated. The progressive verbification of adjectival forms was considered to be indicated by three changes in the constructions where these forms played the role of a predicate: supersession of connections with feminine verb forms by connections with neuter forms, limiting the connections with verbs to the auxiliary verb *być* ‘to be’, and disappearance of connections with the personal form of the verb *być* in the present tense. The research results show that both forms acquired the two most important features characteristic for predicatives during the 17th century. The third of the analysed properties characterizes the form *można/niemożna* from the second half of the 19th century, and the process of its acquisition by the form *niepodobna* has not ended yet.

Keywords: historical syntax, corpus research, categorical change, adjectives, predicatives

1 INTRODUCTION

In modern grammatical descriptions, a class of the so-called predicatives is distinguished among Polish verbs (cf. Bańko 2002, pp. 101–103). These are verbs that do not have the inflectional categories of person, number and gender, and the values of the tense and mood categories are expressed with the auxiliary verb *być* ‘to be’. The lexemes belonging to this class are of a heterogeneous origin – they arose from fixed infinitive forms of verbal lexemes (e.g. *widać* ‘it seems’), noun forms (e.g. *brak* ‘it is lacking’), adverbial forms (e.g. *wolno* ‘it is allowed’) and adjective forms (e.g. *można* ‘it is possible’).

The predicatives formed from adjectives, which are the subject of this article, comprise two examples: the above-mentioned *można*, the negated form of which is

created by adding the particle *nie* 'not' (*nie można* 'it is impossible') and *niepodobna* 'it is impossible' occurring only in the negated form. Both predicatives differ in the stylistic register: *można* is stylistically neutral, while *niepodobna* is used in a bookish style.

The adjectival form in which both lexemes became fixed (which is now the form of the present tense) is the nominative singular feminine form. It is a relic of expressions composed of a feminine noun *rzecz* 'thing' and feminine forms of adjectives *możny* 'possible', *niemożny* 'impossible', *niepodobny* 'impossible'. In the Middle Polish period (16th–18th century), such expressions as a whole were used in the predicative function, however, the semantically empty noun *rzecz* was often omitted, and the feminine form of the adjective took over the predicative function (examples 1, 2).

1. *Ojcze jeśli* ***można rzecz*** *niech odejdzie precz odemnie ten kielich…* 'Father, if it is **a possible thing**, let this cup go away from me...'[1]
2. *Ojcze mój/ Jeśli* ***można****/ niech mię ten kielich minie…* 'My Father/ If it is **possible**/ let this cup pass me by...'

In the Middle Polish period, constructions such as above-mentioned could include other adjectives, e.g. *słuszny* 'right' (example 3).[2]

3. ***Słuszna*** *by Stwórcy, Stworzenie słuchało.* 'It is **right** that the Creation should obey its Creator.'

However, towards the end of the Middle Polish period, predicative constructions with feminine forms of adjectives gradually fell out of use. Their past existence is still reflected in the form of the lexicalized expressions, such as: *dobra nasza* 'good for us', and the predicatives *można* and *niepodobna* described in this article.

The aim of this paper is to present the process of verbification of both forms consisting in losing adjective features and gaining verb features.

2 MATERIAL

The research presented in this article was carried out on corpus material. Unfortunately, the corpus of Polish texts from the 16th century, i.e., from the period when constructions containing the forms *można, niemożna* and *niepodobna* appeared in the Polish language, has not yet been created. The research on the development of

[1] This and the subsequent Middle Polish examples of usage come from the Electronic Corpus of 17th- and 18th-century Polish Texts (up to 1772), characterized in the next section. Citations are given in the modernized notation used in the corpus.

[2] More about such constructions in Bronikowska 2021.

these forms in the Middle Polish period had to be limited to the 17th and 18th centuries. The material covering this period was taken from the Electronic Corpus of 17th- and 18th-century Polish Texts (up to 1772) (Gruszczyński et al. 2022; https://korba.edu.pl/).[3] This corpus, briefly referred to as *Korpus Barokowy* 'the Baroque Corpus' (hence the acronym KorBa), contains over 13 million tokens.[4]

The material for the analysis presented in this article includes:

- 803 predicative usages of the forms *można* (503 quotes) and *niemożna* or *nie można* (300 quotes),
- 622 predicative usages of the forms *niepodobna* or *nie podobna*.

The way of marking the forms of both later predicatives in the KorBa corpus was intended to reflect their uncertain status in the Middle Polish texts. In the 17th and 18th centuries, the forms *można, niemożna* and *niepodobna* in predicative constructions had the character of mixed categories – it is difficult to classify them unequivocally both as adjectives and as predicatives. Therefore, the authors of the corpus decided to mark them as adjective forms of lexemes *możny, niemożny* and *niepodobny* (examples 4, 5 and 6).

4. *...**można*** [ADJ.NOM.SG.F] *być pobożnym i oraz wesołym.* 'it is possible to be both pious and cheerful'
5. *Bez wiary **niemożna*** [ADJ.NOM.SG.F] *się Bogu podobać.* 'without faith it is impossible to please God'
6. ***Niepodobna*** [ADJ.NOM.SG.F] *staremu długo żyć...* 'it is impossible for an old man to live long'

As there were no standards in terms of joint and separable spelling, the texts from the 17th and 18th centuries also contain notations such as *nie można* and *nie podobna*. In the research presented in this article, such expressions were treated as orthographic variants of the forms *niemożna, niepodobna*; both variants were analyzed together.

The data on the negated form *niemożna/nie można* have been also combined with the data on the non-negated form *można*. In the further part of the article, the notation *można/niemożna* is used to denote both the non-negated and negated form of the later predicative *można*.

---

[3] The corpus was created as a result of the project financed under the National Programme for the Development of Humanities for 2013–2018 (NPRH no. 0036/NPRH2/H11/81/2012).

[4] Currently, there is an ongoing work aimed at expanding it – after the current project (no. 0413/NPRH7/H11/86/2018) has been finished, the corpus will cover the entire period of the 17th and 18th centuries, and its size will increase to 25 million tokens.

The search for all the above-mentioned forms was carried out in two stages: the forms of nominative feminine singular of the adjectives *możny, niemożny, niepodobny* and *podobny* (in conjunction with the particle *nie* 'not') were searched automatically in the KorBa corpus. Subsequently, the contexts in which these forms occur in predicative usages were manually selected. The set of quotations, saved in the XLS format, was subjected to an additional annotation, which was intended to mark the syntactic context of the studied constructions.

Some changes in the constructions with *można/niemożna* and *niepodobna* took place after the Middle Polish period, mostly in the 19th century. Studying these changes is difficult as there is not a large, balanced corpus that would include this period. Data from two corpora were used as an aid in the research presented in this article: the small, 1.3-million-token Corpus of Polish texts from 1830–1918 (Bilińska et al. 2016; http://korpus19.nlp.ipipan.waw.pl/) and the relatively large, over 12-million-token, but unbalanced (the so-called opportunistic) Corpus of 19th century texts, covering the period of 1800–1933 (Łaziński et al. 2023; http://www.diaspol.uw.edu.pl/XIX/#!/). The last data source is the National Corpus of Polish (Przepiórkowski et al. 2012; https://nkjp.nlp.ipipan.waw.pl/; hereinafter: NKJP) in which the contemporary Polish texts are gathered.

3 GAINING PREDICATIVE FEATURES BY THE FORMS *MOŻNA/NIEMOŻNA* AND *NIEPODOBNA*

All lexemes currently included in the class of predicatives arose as a result of categorical changes that took place in verbal, substantival, adjectival and adverbial forms, the basic forms of all predicatives are therefore (or at least used to be in the past) homonymous with forms belonging to another part of speech (e.g. *brak* 'lack' – noun vs 'it is lacking' – predicative). This makes it particularly important to define the criteria for including particular word forms in this part of speech. This issue was the subject of a long-term debate, as a result of which various features were pointed out that could be considered as determinants of "predicativeness" (cf. Wróbel 1988; Wiśniewski 1989; Szupryczyńska 1995; Przepiórkowski 2019).

The research in this area, conducted on the material of the contemporary Polish language, presented a synchronic approach. Efforts were made to unambiguously determine whether a given form is or is not a predicative. However, the criteria developed in the course of this discussion can also be used in diachronic research. In this case, it would be about indicating the time when, in the uses of a given form, the features characterizing predicatives begin to prevail over the features of the class to which it originally belonged.

The study presented below discusses three criteria used today to distinguish a class of predicatives. They were treated as a reference point in the analysis of changes that took place over the period under study in syntactic constructions containing the forms *można/niemożna* and *niepodobna*.

3.1 Neuter form of the auxiliary verb *być* 'to be' in the past tense and the subjunctive

The basic criterion used to distinguish predicatives from homonymous forms of other parts of speech, especially nouns, is the grammatical gender taken by the auxiliary verb *być* 'to be'. In this case the forms of the past tense and the subjunctive mood constitute the diagnostic context, because only in them is the generic opposition between the forms of the verb revealed. For example, the verb *być* is combined with the masculine noun *brak* 'lack' by agreement (example 7), and when accompanying the predicative *brak* 'it is lacking' it takes the neuter form (example 8) (cf. Saloni 1974, p. 95).

7. ***Brak*** *jedności* ***był*** [M] *przyczyną klęski...* 'lack of unity was the cause of defeat...' (NKJP)
8. *...**brak było*** [N] *środków na ten cel.* '...there were no funds for this purpose.' (NKJP)

This criterion can be extended to predicatives of non-noun origin, in particular to lexemes derived from adjectives discussed here. Let us compare two examples from the KorBa corpus:

9. *...stamtąd wyniść zgoła **niepodobna była*** [F]. '...it was impossible to get out of there'
10. *...nazad przejść **niepodobna było*** [N]. '...it was impossible to go back'

The feminine form of the verb *być* in example 9 is the result of an agreement with an elided feminine noun *rzecz* 'thing', while neuter form in combination with the form *niepodobna* in example 10 can be interpreted as the analytical form of the predicative.[5]

The corpus data show that in combinations with the forms *można/niemożna* and *niepodobna*, the supersession of the feminine form of the verb *być*[6] by the neuter form took place throughout the 17th century. In the first half of this century, the feminine form of the verb is dominant, while in the second, uses with the neuter form greatly predominate. In the 18th century, not a single occurrence of the feminine form of the verb was recorded in the texts collected in the corpus (Tab. 1 and 2).

[5] This is not the only possible interpretation, because the verb *być* takes the neuter form when combined not only with predicatives, but also with predicative adverbs. The criterion for distinguishing these two grammatical classes will be presented later in the article.

[6] With regard to the form *niepodobna*, the change of the form of two other verbs connected with it was also taken into account: *zdać się* 'to seem' and *widzieć się* 'to seem' (more on such combinations in the next section).

MOŻNA/NIEMOŻNA	1601–1650	1651–1700	1701–1750	1751–1772	Total
feminine gender	2	1			3
neuter gender		11	25	71	107

Tab. 1. Uses of *można/niemożna* with feminine and neuter forms of the verb *być* in the KorBa corpus

NIEPODOBNA	1601–1650	1651–1700	1701–1750	1751–1772	Total
feminine gender	11	4			15
neuter gender	4	22	9	4	39

Tab. 2. Uses of *niepodobna* with feminine and neuter forms of the verbs *być* 'to be', *zdać się* 'to seem' and *widzieć się* 'to seem' in the KorBa corpus

3.2 No connections with verbs other than *być* 'to be'

As noted, the expansion of the neuter form of the verb *być* in conjunction with the forms *można/niemożna* and *niepodobna* is not a proof of their predicative status, because the same process would accompany the transition of adjectival forms to the adverb class. Meanwhile, in the opinion of the majority of contemporary researchers, adverbs in predicative constructions (e.g. *Było duszno*. 'it was stuffy'), the so-called predicative adverbs, do not belong to the class of predicatives.

The main argument in favour of the separateness of these two classes is that predicative adverbs can combine with auxiliary verbs other than *być*, such as *stawać się* 'to get' (e.g. *Stawało się duszno*. 'it was getting stuffy'). If we postulated the existence of the predicative *duszno* 'it is stuffy', we would have to interpret both expressions (*było duszno* and *stawało się duszno*) as variant inflectional forms of its past tense. Due to the difference in meaning of these two expressions such an interpretation is impossible (Wiśniewski 1989, pp. 185–188).

In modern Polish, the lexemes *można* and *niepodobna* do not connect with verbs other than *być*, which, in the light of the argument presented above, is one of the proofs that they belong to the class of predicatives. In the Middle Polish language, there is a difference between the two forms: in the KorBa corpus, the form *można/niemożna* never appears in conjunction with a verb other than *być*, while the form *niepodobna* is used with the above-mentioned verbs *zdać się* 'to seem' and *widzieć się* 'to seem' (example 11).

11. ***Zda się*** *im* ***niepodobna*** *wydołać tej sile…* 'it seems impossible for them to endure this power...'

The corpus data also show that such uses disappear in the second half of the 17th century (Tab. 3). It is worth noting that this process occurred simultaneously with the expansion of the neuter form of verbs in the past tense and the subjunctive.

NIEPODOBNA	1601–1650	1651–1700	1701–1750	1751–1772	Total
zdać się	9	1			10
widzieć się		1			1

Tab. 3. Uses of *niepodobna* with the forms of the verbs *zdać się* and *widzieć się* in the KorBa corpus

3.3 Present tense form without the verb *być* 'to be'

The criterion for distinguishing the forms of predicatives from the forms of other lexemes is also the lack of connections with the form *jest* 'is' (the personal form of the verb *być*) in the present tense. Some researchers point out that the present tense form of the lexemes, which undoubtedly belong to the class of predicatives (e.g. *trzeba* 'must'), does not include the form of the auxiliary verb (e.g. *Trzeba iść.* 'one must go', not: **Trzeba jest iść*.). This is one more feature that distinguishes predicatives from predicative adverbs, where the present tense can be expressed with the form *jest* (e.g. *Trudno iść. = Trudno jest iść*. 'it is hard to go') (cf. Szupryczyńska 1995, p. 175).

This criterion is sometimes considered too restrictive. Przepiórkowski points out that some lexemes (such as *wolno* 'it is allowed') commonly regarded as predicatives, sometimes appear in texts with the form *jest* in the present tense (Przepiórkowski 2019, p. 88). Regarding this feature, there is a difference between lexemes *można* and *niepodobna* in contemporary Polish – the former is never accompanied with the form *jest* in the present tense, while for the latter, sentences as in example 12 are acceptable.

12. *...**niepodobna jest** nauczyć się sztuki komunikowania w przeciągu pół godziny.* '...it is impossible to learn the art of communication in half an hour.' (NKJP).

For the purposes of this article, let us assume that the lack of the form *jest* in the present tense is an additional, auxiliary indicator of the progressive verbification of adjectival forms. As shown in Tab. 5 and 6, in the 17th and 18th centuries, both forms, *można/niemożna* and *niepodobna*, admitted the presence of the form *jest* in the present tense, although in the case of the former, such uses appear less frequently.

MOŻNA/NIEMOŻNA	1601–1650	1651–1700	1701–1750	1751–1772	Total
present tense with the form *jest*	-	1	-	2	3
present tense without the form *jest*	34	119	135	356	644

Tab. 5. Uses of *można/niemożna* in present tense containing and non-containing the form *jest* in the KorBa corpus

NIEPODOBNA	1601–1650	1651–1700	1701–1750	1751–1772	Total
present tense with the form *jest*	6	3	4	5	18
present tense without the form *jest*	91	287	116	44	538

Tab. 6. Uses of *niepodobna* in present tense containing and non-containing the form *jest* in the KorBa corpus

These data do not indicate the complete disappearance of the form *jest* in the present tense with the discussed forms, so the question arises what happened to this structure after the period that is covered by the KorBa corpus. In both available corpora containing texts from the 19th century, the only confirmation of a connection *można jest* comes from 1842, so it can be assumed that such connections disappeared in the first half of the 19th century. The combinations *niepodobna jest* are present in texts throughout the 19th century and at the beginning of the 20th century, and, as evidenced by the data from the NKJP, they can also be used in contemporary Polish.

4 SUMMARY AND CONCLUSIONS

The analysis presented in this article was aimed at examining the verbification of the adjective forms *można/niemożna* and *niepodobna*. The observation of three processes which are the determinants of the ongoing categorical change was carried out on the data collected from several corpora covering the period from the 17th century.

Before summarizing the research results, one reservation should be voiced. Although the work used a large, over 13-million-token corpus, and the examined forms had several hundred occurrences, it turned out that there were very few diagnostic uses in which the features characteristic for the class of predicatives would be revealed. For this reason, the time at which the categorical change described here took place can only be defined within a very broad timeframe. To get a more accurate picture, it will be worth repeating the research presented in this article on more data.

The research results are as follows:

1. Combinations *można/niemożna* and *niepodobna* with the feminine forms of the verbs *być*, *zdać się* and *widzieć się* disappear during the 17th century. From the 18th century, only neuter forms are noted.
2. Throughout the entire period under study, the form *można/niemożna* was not accompanied by verbs other than *być*. The combinations of the form *niepodobna* with the verbs *zdać się* and *widzieć się* vanished in the second half of the 17th century.

3. The *można/niemożna jest* combinations fell out of use probably in the first half of the 19th century. Constructions *niepodobna jest* appeared both in the Middle Polish era and in texts from the 19th and 20th centuries; they also appear in texts of the modern Polish language.

The presented results show that the acquisition of predicative features by the adjectival forms *można/niemożna* and *niepodobna* took place mainly over the course of the 17th century. On the basis of the first two, most important criteria, it can be concluded that there have been predicatives *można* and *niepodobna* in the Polish language since the beginning of the 18th century. The adoption of an additional, third criterion means that the form *można* attained full predicative status in the first half of the 19th century, and the process of verbification of the form *niepodobna* has not ended yet.

The findings made here should be reflected in the tagging of the described forms in the KorBa corpus. The annotation of these forms should be differentiated depending on whether the texts in which they occur were created before or after the completion of the verbification process. In order to do that, it is necessary to develop a general method of describing forms which were subject to a categorical change in the period covered by the corpus. Work on this issue is currently underway.

ACKNOWLEDGEMENTS

The research has been supported by the National Programme for the Development of Humanities (NPRH) under the project "The extending of the Electronic Corpus of 17th- and 18th-century Polish Texts and its integration with the Electronic Dictionary of the 17th- and 18th-century Polish" (no. 0413/NPRH7/H11/86/2018) funded by the Ministry of Science and Higher Education.

References

Bańko, M. (2002). Wykłady z polskiej fleksji. Warszawa: Wydawnictwo Naukowe PWN, 240 p.

Bilińska, J., Derwojedowa, M., Kieraś, W., and Kwiecień, M. (2016). Mikrokorpus polszczyzny 1830–1918. Komunikacja specjalistyczna, 11, pages 149–161. Accessible at: http://www.ks.uw.edu.pl/NUMER_KS11.pdf.

Bronikowska, R. (2021). Unfinished "verbization" process – the development of predicative constructions with an adjective of the feminine gender in the 17th and 18th centuries in the light of corpus data. Polonica, 41, pages 97–110. Accessible at: https://doi.org/10.17651/POLON.41.7.

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., and Woliński, M. (2022). The Electronic Corpus of 17th- and 18th-century Polish Texts. Language Resources and Evaluation, 56, pages 309–332. Accessible at: https://link.springer.com/article/10.1007/s10579-021-09549-1.

Łaziński, M., Górski, R. L., and Woźniak, M. (2023). Korpus XIX w. Uniwersytetu Warszawskiego i IJP PAN. LingVaria, 18, 35(1), pages 125–134. Accessible at: https://doi: 10.12797/LV.18.2023.35.09.

Przepiórkowski, A. (2019). Status gramatyczny predykatywnych *szkoda, wstyd, żal* raz jeszcze. Polonica, 39, pages 85–110. Accessible at: https://doi.org/10.17651/POLON.39.5.

Saloni, Z. (1974). Klasyfikacja gramatyczna leksemów polskich. Język Polski, 54(1), pages 3–13 and 54(2), pages 93–101.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (eds.) (2012). Narodowy Korpus Języka Polskiego. Warszawa: Wydawnictwo Naukowe PWN, 323 p. Accessible at: http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

Szupryczyńska, M. (1995). Jeszcze o tzw. „predykatywach przysłówkowych". Polonica, 17, pages 173–187.

Wiśniewski, M. (1989). Status gramatyczny tzw. przysłówków odprzymiotnikowych typu *duszno, wolno, nieprzyjemnie*. Polonica, 14, pages 183–191.

Wróbel, H. (1988). Przysłówek w strukturze formalnej polskich zdań. Folia Philologica Jugoslavo-Polonica, 1, pages 70–85.

sciendo
DOI 10.2478/jazcas-2023-0020

GOD KNOWS HOW IT TURNS OUT: ON THREE CONSTRUCTIONS INCLUDING BOG 'GOD', ČERT 'DEVIL' AND SOME TABOO WORDS IN THE RUSSIAN LANGUAGE OVER THE LAST THREE CENTURIES

EVGENIYA BUDENNAYA – KRISTINA LITVINTSEVA
– ANASTASIA YAKOVLEVA
Independent researchers

BUDENNAYA, Evgeniya – LITVINTSEVA, Kristina – YAKOVLEVA, Anastasia: God Knows How It Turns Out: On Three Constructions Including *bog* 'god', *čert* 'devil' and Some Taboo Words in the Russian Language over the Last Three Centuries. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 19 – 31.

Abstract: The constructions with the anchor [Noun-Nom Verb (meaning 'to know')] are very productive in Russian. In this article we show that variables such as Bog 'God', čert 'devil' and xer/xren 'X/horseradish' have some common patterns, as well as some shifts with exclusive patterns in semantics and constructionalization.

Keywords: Russian language, construction, semantics, discourse formulae, diachrony

1 INTRODUCTION

The Russian language has three synonymous constructions based on the pattern *X znaet* ('X knows'), which express negative evaluation of an unknown or unspecified entity (in Endresen – Janda 2020) they are tagged as encoding Assessment in relation to knowledge) and can also function as discourse formula: *bog znaet* 'God knows', *čert znaet* 'devil knows' and *xer/xren znaet* 'X/horseradish knows'. They tend to be interchangeable and the pattern itself is described as highly productive (Zhukova 2021, p. 150) and typologically common (Kehayov 2009). All constructions formed accoording to this pattern can be used with interrogatives:

(1) *on dumal o nej* ***bog znaet čto*** 'He thought about her God knows what'
(2) *ona xodit* ***čert znaet gde*** 'She walks devil knows where'
(3) ***Xren znaet*** *počemu ona ušla* 'X knows why she went away'

It should be noted that *xren znaet,* 'X knows', appears in the Russian National Corpus (hereinafter RNC[1]) much later than *bog znaet,* 'God knows', and *čert znaet,* 'devil knows', thus being much more recent, as well as more colloquial and, due to its euphemistic nature, more derogatory. However, over the past three centuries each of these constructions, including *xren znaet*, has undergone a series of micro-diachronic changes in syntax and semantics, which as a result gave rise to fixed variants *(ne) Bog vest'* '(not) God knows' + *interrogative*, *čert-te* 'devil knows', lit. 'devil-*te*'+*interrogative* and *xz* (abbreviation of *xren znaet*), which are far from the original semantics. Thus, *ne bog vest'* expresses mediocrity and not the lack of knowledge, see (4). In these idiomatic variants not only is the interchangeability of the actor impossible (we cannot say **čert vest'* instead of *Bog vest'* or **bog-te* instead of *čert-te* with interrogatives), but the syntax can also be different. Thus, *xz*, originally abbreviated from a clause-like expression *xren znaet* 'X knows', is now often used as a predicate (5). Finally, the construction *čert-te X* (6) has a rather mysterious form, with a not entirely obvious etymology of *-te*, which results in alternative orthographic versions like *čerte X* and *čerti X*.

(4) *Služet* ***ne Bog vest' kakoj*** 'The plot is mediocre', lit. 'The plot is not God knows how' (RNC, 1975–1977)
(5) ***Xz****, možet ty i prav* 'X… knows, maybe you are right' (GICR, 2014)
(6) ***čert-te čto*** *tvoritsja* 'Devil-*te* knows what is happening' (RNC, 2011)

The micro-diachronic development of these constructions has not been previously studied on the corpus material. In related works on Russian they were mostly presented as synonymous (Zhukova 2021; Iomdin 2018). Only recently have the linguistic changes over the last three centuries begun to be treated as a separate subject of study and have been investigated in greater detail (see Rakhilina 2010; Pekelis 2020 inter alia). Our work continues a series of corpus linguistics studies on diachronic changes and focuses on the evolution of these three idioms, which deviated so strongly from the once synonymous variants of the same construction. The data are taken from the main corpus of RNC[2] and from the General Internet-Corpus of Russian (hereinafter GICR).[3] The paper is structured as follows. Section 2 discusses *Bog vest'* and its evolution. Section 3 is focused on *čert-te*, its semantics and the etymology of *-te.* Section 4 presents the results on the syntax and semantics of *xz*. Finally, Section 5 summarizes the results of the study.

[1] https://ruscorpora.ru/en/

[2] https://ruscorpora.ru/en/search?search=CgQyAggBMAE%3D, about 375 million words, contains written texts from the middle of the 18th century to the present day

[3] http://www.webcorpora.ru/en/, more than 20 billion tokens, contains social media materials: networks VKontakte, LiveJournal blogs and the texts of *Žurnal'nyj Zal* (https://magazines.gorky.media/) dated 21st century

2 BOG ZNAET/VEST'

This part of the paper is devoted to the semantics of the earliest attested variant of the construction *X znaet* (X knows): *Bog znaet/vest'/vedaet* (God knows) + *interrogative* (*kto* 'who'/*čto* 'what'/*kak* 'how' etc.). This construction has indefinite, evaluative and intensifying meaning; all of them are attested already in the end of the 18th – beginning of the 19th century. Moreover, the construction has been used in all these meanings till nowadays, and the compositional interpretation is also possible. We will describe the scope of meanings based on corpus data, reconstruct a possible constructionalization path and put forward an explanation of the semantic shift towards indefiniteness and then to an evaluative and intensifying meaning.

Inkova (2006) demonstrates that *Bog znaet* + *interrogative* is less grammaticalized than *Bog vest'* + *interrogative* and can be interpreted compositionally, as 'God knows what/where/why etc.'. Indeed, *Bog znaet* + *interrogative* can be used with the adjective *odin* 'alone', with the adverb *tol'ko* 'only', the predicate is attested not only in present form (Inkova 2006, p. 450). Our data also confirm this thesis: see (7) for the illustration:

(7) *Tol'ko* ***odin bog znal****, čto tvorilos' v jeto vremja v ee duše.* 'Only God alone knew what was happening in her soul at that time.' (RNC, 2003)

The variant *Bog vest'* + *interrogative*, however, seems to be more idiomatic. It can be explained by the fact that the Old Slavonic form *vest'* is not semantically and grammatically transparent for the speakers of modern Russian. Although compositional interpretations (usage with *odin* 'alone' or *tol'ko* 'only', free word order) are attested in the corpus, such contexts are extremely infrequent (only 3 examples from the 19th century were found in RNC).

At the next step the construction *Bog vest'/znaet* (BZ) + *interrogative* evolves the indefinite semantics. So, the observed semantic shift is: *(only) God knows* → *nobody knows* (see (8)):

(8) *Ivan Petrovič idet v poxod i* ***bog vest' kogda*** *vorotitsja v Derbent.* 'Ivan Petrovič goes on a campaign and God knows when he will return to Derbent.' (RNC, 1830–1837)

The indefinite semantics and the place of the BZ-construction in the system of Russian indefinite markers is described in detail in (Inkova 2006); in this study we will concentrate on the development of evaluative and intensifying meanings.

Hence, the categories of time, location and quantity in some contexts allow the interpretation of the construction as the extreme point of the correspondent scale. This way, *BZ kogda* 'God knows when' means 'long time ago', *BZ gde/kuda* 'God

knows where/where to' means 'far', *BZ skol'ko* 'God knows how many/much' means 'a lot' (see (9), where *BZ skol'ko vremeni* 'God knows how much time' refers to a concrete period of two years).

(9) *Emu v samom dele kazalos', čto dva goda* ***— bog vest' skol'ko vremeni****, čto v dva goda on uspeet opjat' nažit' stol'ko, čtoby potjagat'sja s Moskvičom i peretjagat' ego.* 'It really seemed to him that two years were God knows how long, that in two years he would have time to make enough money again to compete with the Muscovite and win him over.' (RNC, 1829)

As for the interrogatives *kto* 'who', *čto* 'what', *kak* 'how', *kakoj* 'what/which', *gde* 'where', they evolve pejorative semantics within the *BZ*-construction. So, *BZ kto* 'God knows who' can be interpreted as a bad person, in (10) *BZ kak* 'God knows how' means 'badly, poorly'.

(10) *Žil on skupo: nedoedal, nedopival, odevalsja* ***bog znaet kak****.* 'He lived sparingly: malnourished, underdrinking, dressed God knows how.' (RNC, 1898)

A possible way of semantic shift is *'nobody knows who/what/how'→ 'anybody/ anything/anyhow'→ 'no matter who/what/how' → negative assessment*.

This process can be classified as *semantic rebranding*: a semantic shift, where the derived meaning is an implicature of the source meaning (see more in Rakhilina et al. 2010, pp. 428–453). It is interesting that in pejorative contexts the BZ-construction can be replaced by the construction *interrogative + popalo* (lit. 'how/ what/who etc. fell'), which also has the semantics of randomness and of a negative assessment.

Furthermore, in some contexts BZ-expressions can also function as intensifiers. In (11) *BZ kak,* 'God knows how', is used as an intensifier and can be interpreted as 'very, a lot'. Cross-linguistically, the semantics of negative assessment is often combined with intensifying meaning (cf. Eng. '*terribly* important'; see more in Lorenz 2002).

(11) *Nakonec počuvstvoval on sebja lučše i obradovalsja* ***bog znaet kak****, kogda uvidel vozmožnost' vyjti na svežij vozdux.* 'At last he felt better and was glad, God knows how, when he saw the opportunity to go out into the fresh air.' (RNC, 1842)

Finally, the *BZ + interrogative* construction develops "de-intensifying" semantics by adding a negative particle *ne*. The expression *ne BZ kakoj/kak etc.* 'Not God knows what/how etc.' functions as a marker of mediocrity (see (12)).

(12) *Okončila* ***ne Bog vest' kak****, s troečkami.* '(She) graduated not even God knows how, with C grades.' (RNC, 2000)

It is remarkable that only *Bog* 'God' is attested with negation and de-intensifying semantics; the construction variants with devil and other creatures/objects are not normally used in such contexts.

The variants *Bog znaet* and more archaic one *Bog vest'* tend to convey different meanings (see Fig. 1). In combination with the conjunction *kak* 'how', *Bog znaet* is more frequently attested than *Bog vest'* almost in all functions. The only semantic domain where *vest'*-variant prevails is de-intensification. A comparative study of the two variants with different conjunctions is beyond the scope of the present paper, but the proposed semantic classification seems to be a good foundation for a detailed synchronic and micro diachronic corpus-based description of the BZ-construction.

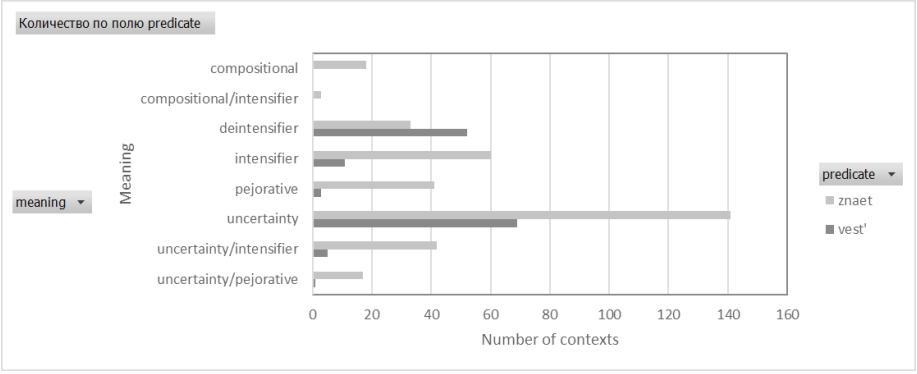

Fig. 1. Semantic distribution of predicates *znajet* and *vest'* 'knows' in the construction *Bog znajet/vest' kak* 'God knows how' in RNC, 18th century – present

3 ČERT-TE + INTERROGATIVE

Unlike the construction with *Bog* 'God', the construction with *čert* 'devil', since the earliest entries, (end of the 18th century) is not used with predicates of knowing other than *znat'*. The constriction *čert-te* seems to be interchangeable with *čert znaet* with interrogatives: *on pošël čert znaet/čert-te kuda* 'he went devil knows where' (= to an unknown place or direction, assessed negatively by the speaker). Semantically, *čert znaet* 'devil knows' + *interrogative* is similar to the construction with *Bog* and can also be used in both evaluative (13) and intensifying (14) contexts, although its usage tends to be more derogatory due to the semantics of *čert* 'devil':

(13) *Malo li detej, kotorye… zanimajutsja* ***čert znaet čem*** 'You never know the children who ... do the devil knows what' (RNC, 2004)

(14) *Ona byla* ***čert znaet, kak*** *xoroša*. 'She was devil knows how beautiful.' (RNC, 2007)

In both (13) and (14) *čert-te* can be used instead of *čert znaet.* However, the origin of *-te* is vague. Historically, *-te* may be treated as a grammaticalized pronoun *te,* a short form of *tebe* (2SG.ACC) which later obtained an *-a* ending (modern *tebja* '2SG.ACC'). In this way, the evolution is likely to be tracked from *čert tebja znaet X* to *čert-te znaet X* and then to *čert-te X.* Yet the RNC demonstrates it the other way round: constructions of type *čert-te* appear much earlier than constructions of type *čert-te znaet* with interrogatives, thus making the origin of *-te* less clear. A more detailed study of the RNC shows the following stages of *čert-te X* evolution:

1. Mid 18th century–1820s: no entries of *čert-te X*, *čert znaet* is attested only as a separate clause expressing dissatisfaction and negative assessment:

(15) ***Čert znaet*****!** *Takoj golos, čto rastajat' možno* 'The devil knows! Such a voice that you can melt' (RNC, 1763–1774)

2. 1830–1880s: *čert znaet* 'devil knows' begins to be used with an object pronoun *tebja* 'you.ACC' as a separate clause expressing doubt and/or uncertainty. However, at this stage no idiomatic constructions with *čert tebja znaet* in an attributive position like *čert tebja znaet čto/gde/kuda* 'Devil knows you what/where/when' are detected:

(16) *Ty, brat,* ***čert tebja znaet***, *poteeš, čto li*. 'You, brother, the devil knows you, sweat or something.' (RNC, 1842)

3. Mid 19th century–1870s: the first entries of *čert-te* in informal imperative constructions expressing dissatisfaction. In (17), *(-)te*-form is a grammaticalized referential pronoun *te* contracted from *tebja*.

(17) *Ax*, ***čert-te voz'mi*** *sovsem! — skazal soldat.* 'Oh, (let) the devil take you! said the soldier.' (RNC, 1863)

4. The 1880s: *čert tebja znaet* gradually develops an evaluative semantics and starts to be used with interrogatives (18). At the same time, the construction *čert-te + interrogative* is also first attested (19). Apparently, *-te* results from the grammaticalization of *tebja* by analogy with the already common form *čert-te* in imperative constructions. However, unlike with imperatives, *-te* in *čert-te* is no longer referential with interrogatives:

(18) *no ved' ty*, ***čert tebja znaet, kakoj lentjaj****!* 'But you are, the devil knows what a slacker!' (RNC, 1886)
(19) *On ej slovo – ona dvadcat'*, *on to po-našinski*, *ona* ***čert-te po kakovskomu*** *norovit…* 'He said a word – she answered twenty, he spoke in our way, she, the devil (knows) in what way tries…' (RNC, 1883)

5. 1917: *čert-te znaet* begins to replace *čert-te* in an attributive position with interrogatives (20). Although the new construction did not become frequent (the RNC contains only 18 entries of *čert-te znaet* + *interrogative* over the last hundred years, see Fig. 2), it is found in texts of different genres (story, novel, diary, serial and article) written by twelve different authors and thus cannot be claimed to be accidental. Hence, we may argue the convergent origin of *čert-te znaet* + *interrogative* from *čert tebja znaet* and *čert-te* + *interrogative*:

(20) *Šokolad stal* ***čert-te znaet kakoj*** *drjan'ju*. 'Chocolate ... became the devil knows what rubbish.' (RNC, 1917)

Fig. 2. Distribution of *čert-te znaet* + *interrogative* by year in RNC, 18th century – present

6. 1975: the first entry of a merged spelling *čerte* instead of the initial separate *čert te* and later hyphenated spelling *čert-te*, the pronominal being completely grammaticalized. In the early 21st century, it also becomes spelled as *čerti* (22), the historical pronominal morphology thus being completely lost:

(21) ***Čerte čto****… Ja javno na pereput'je.* 'What the hell (lit. 'devil what')… I'm clearly at a crossroads.' (RNC, 1975)
(22) *Ja načinaju dumat'* ***čerti čto***. 'I'm starting to think what the hell (lit. 'devil what').' (RNC, 2003)

The loss of referentiality is typical for personal pronouns when performing as components of a discourse formula. Zhukova (2021, pp. 153–154) discusses the desemantization of *ego* '3SG.ACC.M' in constructions like *Bog/čert/xren ego znaet* (lit. 'God knows him') where *ego* no longer designates a concrete person or a situation as a whole. A similar process is documented in the history of an obscene expression *tvoju mat'* ('your mother-ACC', originally a complement of a taboo verb form *eb'* (vulgar 'copulate-IMP') (Uspenskij 2018, p. 256). In modern Russian, *tvoju mat'* no longer applies to the interlocutor and is used as a non-referential curse interjection. Since all expressions with *čert* are derogatory, constructions of type *čert tebja znaet* might have followed a similar path of pronominal desemantization and referentiality loss, giving result to the grammaticalization of *tebja* into *-te* and obtaining a more general meaning of negative assessment. Remarkably enough, according to the RNC, *-te* was initially grammaticalized in imperative constructions like *čert tebja (-> te) deri* 'Let devil tear you', where it functioned as a direct object (a pattern syntactically similar to *eb' tvoju mat'*) and only later expanded to attributive constructions with interrogatives.

4 XER ZNAET/XREN ZNAET/XZ

Fägersten (2012) claims that swearing usually belongs to an emotional situation, while Uspenkij (2018) correlates obscene phrases to ancient mythological, taboo formula; Zaliznjak (2014) demonstrates that elements *x**/xer/xren* (as well as *čert*) have the same semantic, syntactic, morphological and accent properties. All of these assertions should explain the interchangeability of the variable *Bog* 'God'/*čert* 'devil' with the *x**/xren/xer* 'obscene or euphemistic word'.

Due to Zaliznjak (2014, p. 274), the first substitute step was *xer znaet* 'X knows', where *xer* was a name for a Church-Slavonic letter *X* which was initially used as a euphemism to a corresponding obscene word *x*** (23).

(23) *Seli v kakoj-to avtobus i uexali* ***xer znaet kuda***. 'Got on some bus and left to xer knows where.' (GICR, 2014)

The word *xren* is used in the same function in modern Russian (24).

(24) *čast' smotalas' po domam, čast' rvanula v Fili i ešče* ***xren znaet kuda*** *tam*. 'some hit the road home, some rushed to Fili and xren knows where else.' (GICR, 2006)

The anchors *xer znaet/xren znaet* (as well as *x***) have the same constructionalization patterns as anchors *Bog znaet/čert znaet*. They can be used as an intensifier (25) as well as to express negative assessment related to uncertainty (26).

(25) *No eto že* ***xer znaet čto****! Tolja s otčajaniem ogljadel oboix.* 'But this is xer knows what! Tolya looked at them both desperately.' (GICR, 2009)
(26) *Vol'f Messing rabotajuščij v otdele Ananerbe ili* ***xren znaet gde to*** *tam.* 'Wolf Messing working in the Ahnenerbe department or xren knows where.' (GICR, 2010)

Example (27) shows us that *xer znaet/xren znaet* (as well as *x***) can appear not only as an anchor of constructions but also as independent discourse formula (compare to (15)). Thus we can claim a high level of grammaticalization of the construction.

(27) *V a l e r a (požimaet plečami).* ***Xer znaet****, družba, navernoe* 'VALERA (shrugs his shoulders). Xer knows, friendship, probably. (GICR, 2012)'

At the same time a few examples such as (28) and (29) are found. It means that sometimes Russian native speakers can extract the 'inner form' from these idioms, so we can see two multidirectional processes in language. It should be noted that words *vest'* and *vedaet*, which are archaic synonyms to the word *znaet'* 'knows', can be used with *Bog* (8)–(9). Thereby *x*** and its euphemisms retain semantic memory of structural elements of the construction.

(28) *den'* ***xren vedaet*** *kakoj..::)* 'the day is xren knows what.*::)*' (GICR, 2014)
(29) *u nego očerednaja forma bor'by so skinami, buržujami ili ešče* ***x** vest'*** *s kem...* 'he has another form of struggle with skins, bourgeoisie, or even x** (f*ck) knows with whom…' (GICR, 2007)

Another parallel and multidirectional process is a situation in which people use the construction *xz* (from *x znaet* 'x knows' (30)). They sustain the main obscene word and their euphemisms (such as *xer* and *xren*) using the initial of every one of these words to express the same grammar and semantics (31). We have already seen the exact same process when people started using *xer znaet* as the euphemism with the initial letter naming (23). However, now they can sustain the euphemism of the euphemism.

(30) *Vit'ka* ***x znaet*** *gde.* 'Vitka is x knows where.' (GICR, 2006)
(31) *Uexal sjuda, i privet, i* ***xz*** *gde on tut.* 'He left here, and hello, and xz where is he here.' (GICR, 2010)

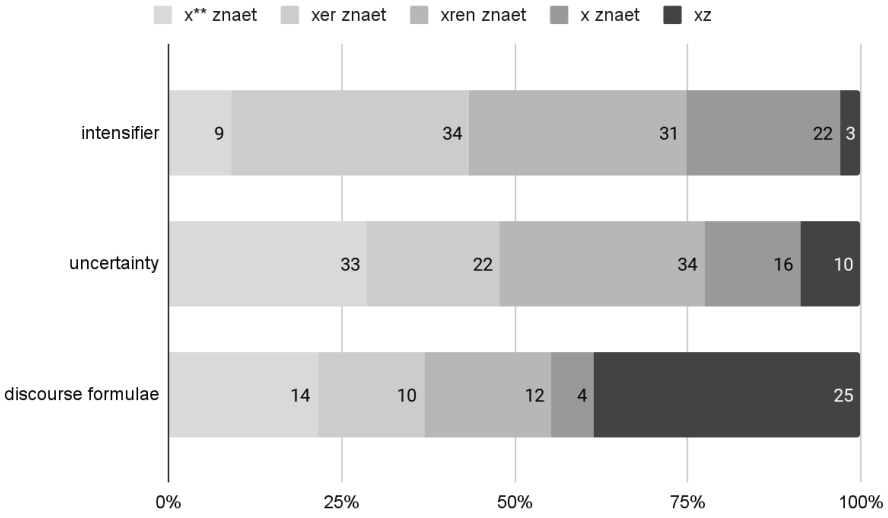

Fig. 3. Semantic distribution of x**/xer/xren/x znaet/xz in GICR, 2004–2015

Fig. 3 shows that *x*** tends to uncertainty and functions as a discours formula, and *xer* and *xren* tend to intensifiers; whereas *xz* is used more often as a discours formula due to its semantic erasure. Interestingly, *xer* is structurally closer to *xren* and *x*, which is explained by the fact that *x*** is simply obscene – it often expresses only emotion, meanwhile *xer/x* is rather an intensifier. It is clear to the speaker that *xer/x* is almost obscene, but he does not use it obscenely, choosing a euphemism as an intensifier, while *xren* is already more erased.

Moreover, today the construction *xz* is not only used as a discourse formula but also as a predicate (32), unlike *Bog vest'* and *čert znaet*. Fig. 4 shows the quantitative distribution (per cent) of usages of the construction *xz* comparing to "entire" forms such as *ja xren znaet* (33) in predicate position.

(32) *Eto gruppa «Ranetki» :)* ***Ja xz****, kto jeto. Serial takoj est', vrode by.* 'This is the band «Ranetki» :) I xz who it is. There is such a series, it seems.' (GICR, 2009)

(33) ***Ja xren znaet****, skol'ko tam ljudej nado i daže ne pomnju sut' vsex.* 'I xren knows how many people are needed there and I don't even remember the essence of it all.' (GICR, 2010)

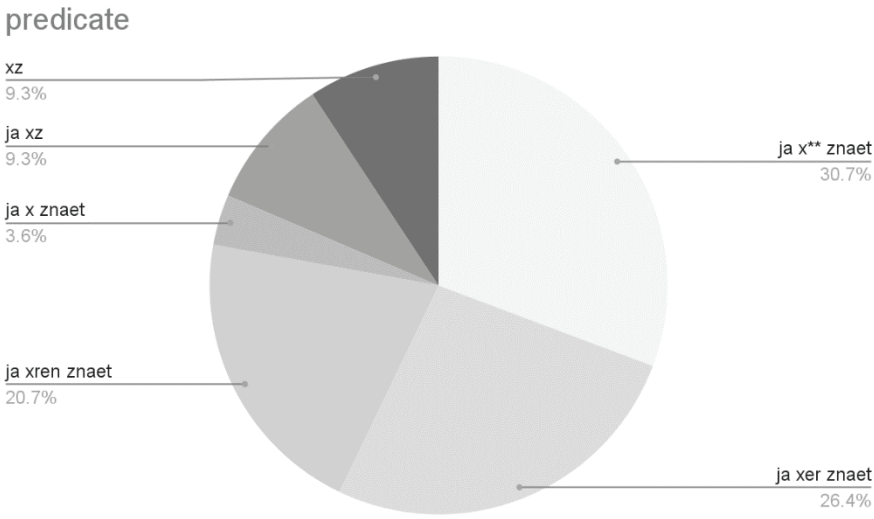

Fig. 4. Quantitative distribution of x**/xer/xren/x znaet/xz using as a predicate in GICR, 2004–2015

At the same time (again as with discussed above examples with *xz* instead of recognizable as obscene *xer znaet* (34)) the situation of rethinking changes the construction (both: the anchor and the meaning). Compare the situation with *ne Bog vest'* (12), although it should be noted that such usages are occasional.

(34) *znakomit'sja mne prixodilos' pust'* ***ne xz kak*** *original'no, no uspešno.* 'I had to get acquainted, even if not xz as originally, but successfully.' (GICR, 2008)

When someone uses *xz* it means they don't know something because it "knows" someone euphemized, however when someone uses *ne xz* ('not xz') the meaning of *x* is erased more, while only the meaning of *z* still exists (however, it also happens the other way around) (see (35)). In other words there is a double constructionalization occurring in the reason of dropping semantics twice (exactly as with *xz*).

(35) *Blin v moej komnate 3 krovati,* ***ja ne xz*** *na kakoj spat'.)))* 'Damn, there are 3 beds in my room, I not xz what to sleep on.)))' (GICR, 2015)

Actually, in the current situation with literally a little anchor *xz* of the construction, we have seen at least three parallel and multidirectional processes.

First, it has started functioning as construction (31), second, it has a double constructionalization twice (32), (35). More than that, we can notice a resemantization in the entire form (29)–(30).

5 CONCLUSION

Thus, we have considered three paths of development of the construction *X znaet* 'X knows' based on the corpus data and revealed the following peculiarities of usage:

1. The construction comprises a broad range of meanings: indefinite, evaluative and intensifying, it can be used as a discourse formula and compositionally. God, devil, and obscene x-words are interchangeable; only God-construction developed the meaning of de-intensification and mediocrity. However, some occasional examples with obscene x-words are also found.
2. We put forward a possible path of constructionalization and the mechanism of the semantic shifts. This phenomenon is mainly described for adjectives and adverbs, but still remains understudied for constructions. Our preliminary results demonstrate that this field of research is rather promising.
3. The devil-construction developed a peculiar variant with -te, which may be interpreted as a grammaticalized pronoun form. We demonstrated that the construction might have undergone the process of pronominal desemantization and referentiality loss: in the data we observe the grammaticalization of *tebja* (you.ACC) into *-te* and the shift towards the semantics of negative assessment.
4. The x-words are remarkable with the development of peculiar syntactic properties: the whole construction becomes abbreviated and starts being used as a predicate ('I x-knows'). So, we observe bleaching of the original semantics and demonstrate the process of double constructionalization and re-semantization of this variant of the construction.

ACKNOWLEDGEMENTS

This work was supported by the Project "Diachronicon": the database on historical changes in Russian constructions.

References

Endresen, A., and Janda, L. A. (2020). Taking Construction Grammar One Step Further: Families, Clusters, and Networks of Evaluative Constructions in Russian. In M. Putnam – M. Carlson – A. Fábregas – E. Wittenberg (eds.): Defining construction: Insights into the emergence and generation of linguistic representations, pages 1–22.

Fägersten, K. B. (2012). Who's Swearing Now? The Social Aspects of Conversational Swearing. Cambridge: Cambridge Scholars Publishing, 337 p.

Inkova, O. (2006). Les indéfinis russes de la série Bog znaet/Bog vest' ('Dieu sait'). In C. Schnedecker – G. Kleiber (éds.): La quantification et ses domaines. Actes du colloque de Strasbourg, 19-21 octobre 2006. Paris: Honoré Champion, pages 449–461.

Iomdin, L. L. (2018). Ešče raz o mikrokonstrukcijax, sformirovannyx služebnymi slovami: To i delo. Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodnoj konferencii "Dialog–2018", 17(24), pages 267–283.

Kehayov, P. (2009). Intensifiers as polarity items: evidence from Estonian. STUF – Language Typology and Universals, 62(1–2), pages 140–164.

Lorenz, G. (2002). Really worthwhile or not really significant. New reflections on grammaticalization, pages 143–161.

Pekelis, O. E. (2020) Zero anaphora in Russian: a microdiachronic analysis. RJANO 39(1), pages 36–59.

Rakhilina, E. V. (ed.). (2010). Lingvistika konstrukcij. Moscow: Azbukovnik, 584 p.

Rakhilina, E. V., Reznikova, T. I., and Karpova, O. S. (2010). Semantičeskie perexody v atributivnyx konstrukcijax: metafory, metonimija i rebrending. In E. V. Rakhilina (ed.). Lingvistika konstrukcij. Moscow: Azbukovnik, pages 398–455.

Uspenskij, B. A. (2018). Mifologičeskij aspekt russkoj ekspressivnoj frazeologii. In Issledovanija po russkoj literature, fol'kloru i mifologii. Moscow: Common place, pages 195–287.

Zaliznjak, A. A. (2014). K voprosu ob akcentnoj evoljucii enklinomenov v russkom jazyke. In Jazyk. Konstanty. Peremennye. Pamjati A. E. Kibrika. St. Petersburg: Aletejja, pages 269–276.

Zhukova, S. Ju. (2021). Russkie diskursivnye formuly v diahroničeskom aspekte. PhD Dissertation, HSE University. Moscow, 283 p.

sciendo
DOI 10.2478/jazcas-2023-0021

APPELLATIVIZATION OF PROPER NAMES – IN THE PERSPECTIVE OF CORPUS ANALYSIS

JAROSLAV DAVID – TEREZA KLEMENSOVÁ
– MICHAL MÍSTECKÝ
Department of Czech Language, Faculty of Arts, University of Ostrava,
Ostrava, Czech Republic

DAVID, Jaroslav – KLEMENSOVÁ, Tereza – MÍSTECKÝ, Michal: Appellativization of Proper Names – In the Perspective of Corpus Analysis. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 32 – 42.

Abstract: The study deals with appellativization of proper names, using as its base selected personal names (surnames). Looking at opinion journalism texts in the Czech National Corpus, corpus SYN, version 11, we investigate aspects of word-formation within appellativization of personal names *Masaryk*, *Beneš*, *Hitler*, *Stalin* – including frequencies of parts of speech and word-formation types (derivation, composition) with respect to their productivity and word-formation potential.

Keywords: appellativization, proper names, personal names, corpus analysis, word formation

1 APPELLATIVIZATION OF PROPRIA

We conduct a corpus analysis of appellativization in Czech, using selected anthroponyms (surnames) of historical figures of modern history as an example. We focus on the diversity found within the appellativization strategies, variations found within parts of speech, and word-formation characteristics connected to the process.

Appellativization is the process by which a proprium (proper name) becomes an appellative lexical unit (a common name), e.g. *Scrooge* becomes a naming of ‘a person who is very unwilling to spend money’, based on the same-name character in Dickens’ story *A Christmas Carol* (1843). A thus created lexical unit acquires the properties of an appellative: it denotes the entire set of objects and it can form the basis of a derivational series. In Czech, in addition, the formal signal of appellativization is the replacement of the initial capital letter by a lower case one (cf. Šrámek 1999, p. 55). Appellativized anthroponyms (personal names) are formed to characterize the external and internal qualities of persons – in general, sources for appellativization are names of mythological and biblical characters, names of historical and contemporary personalities (often politicians; cf.

Děngeová 2010; Jandová 2013), and names of literary characters. To a lesser degree, toponyms (geographical names) serve as bases for appellativization; appellativized toponyms are used mainly to show relations between the place name and a class of objects (e.g. village *Cheddar* > *cheddar cheese*; Pokorná 1978, p. 118).

Formally, appellativization is either direct or indirect. Direct appellativization uses metaphor and metonymy (e.g. *Romeo* > *romeo* 'a man who has a lot of sexual relationships'; *Watt* > *watt* 'a unit of measurement'). Indirect appellativization, while still making use of metaphors and metonymies, utilizes derivation and less often composition (a claim we partly dispute in this paper), e.g. *Marx* > *marxism* 'the teachings of Karl Marx'; *Švejk* > *švejkovat* 'to behave like Švejk'; *švejkomilec* 'a fan of Jaroslav Hašek's novel *The Good Soldier Švejk*'; Hladká 2017; Pokorná 1978; on the forms and changes of word-formation strategies of appellativization in Czech in a synchronic view, see Martincová 2011, pp. 35–36; cf. also Skujiņa 1989; Superanskaya 2012, pp. 113–122). In onomastics, the terms *deonymization* or *deproprialization* are also used, reflecting that the naming has lost the proprial part – the result of the process is referred to as a *deonymic appellative* (Pokorná 1978, p. 118; Šrámek 1999, p. 55). Not only a number of linguistic studies that appeared at the turn of the 2010s (e.g. David 2009; Děngeová 2010; Martincová 2011; Harvalík 2012; Michalec 2012; Jandová 2013), but also an assessment of the situation in languages genetically related to Czech show that we can talk about a tendency to use proper names, especially anthroponyms (family names and especially surnames), to form appellative neologisms. Nowadays, this trend has gained momentum, and it can be considered generally Slavic (see Martincová 2011, p. 22).

2 DATA

The research on appellativization unanimously states that most appellativized lexemes are most often based on anthroponyms followed by toponyms, while chrematonyms (proper names of social events, institutions, organizations, etc.) are used less frequently. These analyses are based on excerption databases, especially from journalism, and/or on texts published on the Internet. These sources, however, may limit the results quantitatively. Our analysis, on the other hand, is based on the set of journalistic texts of the Czech National Corpus, SYN, version 11. Recently, onomastics has adopted corpus-based and generally quantitative analyses to be an integral part of its research (cf. most recently Motschenbacher 2020; David – Klemensová – Místecký 2022; David – Místecký 2023). Despite the fact that in most cases proper names – as complex and often ambiguously defined units – are not tagged in corpora, the quantitative approaches lead to intersubjective and empirically based analyses, especially concerning the grammar of propria.

In formulating the topic of our research, we assumed that a very rich appellative material would be tied to the names of historical figures, in addition to the surnames of politicians. The first step was to select the surnames for our study of appellativization. In order not to choose at random and to avoid potential bias while picking the personal names for our study, we relied on the corpus. In the Czech National Corpus, SYN, version 11, we first restricted the search in the texts via the *KonText* tool, using the attributes doc.txtype_group: NFC (academic literature) and doc.genre: HIS (history, biography). Using the query [tag="N.[FM].*" & lemma="A.*|Á.*|B.*|C.*|Č.*|D.*|Ď.*|E.*|É.*|F.*|G.*|H.*|Ch.*|I.*|Í.*|J.*|K.*|L.*|M.*|N.*|Ň.*|O.*|Ó.*|P.*|Q.*|R.*|Ř.*|S.*|Š.*|U.*|Ú.*|V.*|W.*|X.*|Y.*|Z.*|Ž.*"], we got a set of animate masculine and feminine nouns with a capital initial letter. We consider this to be a sufficient grammatical filter for selecting proper names of persons in the given type of texts. Subsequently, we constructed a frequency-based ranking of lemmas (with a frequency of more than 1,000 occurrences) and from that list, we excluded first names and ambiguous or irrelevant cases (e.g. *Jan*, *Karel*, *Marie*, *Evropa*, *Francie*). On the basis of the criteria above, we obtained a set of four anthroponyms (surnames), which represent four important personalities shaping the modern history of Czechoslovakia and Europe:

- *Masaryk* [the first Czechoslovak president Tomáš Garrigue Masaryk, 1850–1937], absolute frequency 2,600, relative frequency 0.43;
- *Hitler* [German dictator Adolf Hitler, 1889–1945], absolute frequency 2,213, relative frequency 0.37;
- *Stalin* [Soviet dictator Joseph Vissarionovich Stalin, 1878–1953], absolute frequency 1,568, relative frequency 0.26;
- *Beneš* [the second Czechoslovak president Edvard Beneš, 1884–1948], absolute frequency 1,218, relative frequency 0.20.

For these anthroponyms, we analyze the word-formation strategies used to form their derivatives. Thus, we are primarily interested in indirect appellativization.

The next step was to search for lemmas containing the bases of the four surnames in the SYN corpus, version 11; in order to find those, we used the queries [lemma=".*[mM]asaryk.+"], [lemma=".*[hH]itler.+"], [lemma=".*[sS]talin.+"], and [lemma=".*[bB]eneš.+"]. Subsequently, the results were sorted according to the lemma frequency, and the data out of scope of our research were excluded manually from the set. The following forms were excluded: possessive adjectives (e.g. *Masarykův* ‘Masaryk’s’), expressions that were part of foreign language texts (including Latin turns of phrase such as *ad hitlerum*), cases of transonymization (either pure or extended, e.g. chrematonym *Hitlerjugend*, toponym *Stalingrad*; see Šrámek 2004), and other lemmas that were not the result of appellativization (e.g. the proprium *Benešátko* ‘a nickname Masaryk used for Beneš’). The cleaned data were then further analyzed. We always worked with types; we did not consider the

frequencies of their representation, as we were interested in mapping the appellativization strategies in general.

As for the surname *Beneš*, it was necessary to reduce the data set (using negative filters) and then manually sort it very thoroughly. The reason for this is massive homonymy (the surname *Beneš* is still relatively common – as of 2016, there were 9,431 male citizens with this surname in the Czech Republic; Malačka 2011) and the occurrence of derivations unrelated to Edvard Beneš, but included in the query results. These cases included, for example, forms of the female surname *Benešová*, the toponyms *Benešov*, *Benešovice*, *Benešák*, or the group anthroponyms *Benešovici* (aristocratic family) and the inhabitant names *Benešovan*/*Benešák*.

Looking at the word-formation potential, there are significant differences among the anthroponyms studied (see Fig. 1), which can be attributed to various factors. The high number of appellativizations of the surname *Stalin* – in fact, twice as many as the other names – is related to the longevity of Stalin's ideology, which underwent various transformations in the 20th century (*stalinizace* 'Stalinization', *destalinizace* 'de-Stalinization', *restalinizace* 're-Stalinization', *neostalinismus* 'neo-Stalinism') and which needed to be set in a broader context (*prestalinismus* 'pre-Stalinism', *poststalinský* 'post-Stalinist'). A certain role should also be attributed to the richness of synonym series (*stalinský*, *stalinovský*, *stalinistický* 'Stalinist'). There are not any similar contextual detours needed for Hitler's Nazism ideology. Furthermore, the appellativization potential of the surname *Hitler* is reduced by the fact that many German composites (e.g. *Hitlerjunge*) – also found in the corpus – are not native to Czech, and cannot thus be considered the result of the word-forming potential of Czech. Therefore, we did not work with them in the analysis.

Similarly to Hitler's, the life of political philosophy represented by the surname *Masaryk* was also limited in time, which, then, was reflected in the small number of word-formation types. The low number of appellatives derived from the surname *Beneš* suggests that this statesman is not perceived as a creator of a distinctive school of thought one needs to agree/disagree with (with the possible exceptions of the Munich Agreement and the Beneš Decrees). The periods of the end of the First Republic (1935–1938) and the Third Republic (1945–1948) are seen as transitional stages the dynamics of which were largely determined by the development of foreign policy and the personalities of Hitler and Stalin.

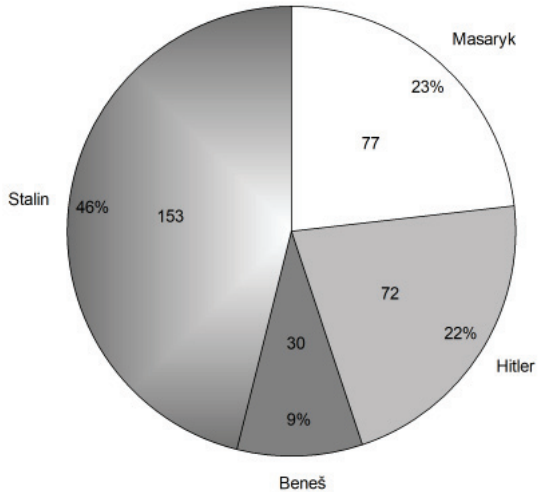

Fig. 1. Frequency of deonymic appellatives derived from the analyzed surnames

3 CORPUS ANALYSIS

As already mentioned above, the quantitative analysis focuses on the word-formation aspects of the appellativization of personal names – the frequency of parts of speech and specific word-formation types with respect to their productivity. We are primarily interested in systemic productivity (potential, langue), but also in real productivity (parole), or latent productivity, realized only rarely (cf. Štícha 2018). Thus, we are paying attention to types here as well.

3.1 Appellativization from the perspective of parts of speech

The overall results, both in percentages and values, are summarized in Fig. 2. The dominance of adjectives is closely related to their ability to form binary and ternary composites "typical of X and Y (or even Z)". "X" in this case represents the analyzed anthroponym. These adjectives – mostly ignored in onomastic studies (see above) – aim to build up a journalistic compression, to indicate the contemporary context (*masarykovsko-benešovský* 'Masaryk-Benešian', *hitlerovsko-henleinovský* 'Hitler-Henleinian', *leninsko-stalinský* 'Lenin-Stalinian'), but also to link different personalities in an innovative way, sometimes with an ironic touch (*masarykovsko-dušínovský* 'Masaryk-Dušínian', *kafkovsko-hitlerovský* 'Kafka-Hitlerian', *verneovsko-stalinský* 'Verne-Stalinian'). Although these compound adjectives appear more as occasionalisms and the results – both orthographically and content-wise – are varied, the potential of this word-formation process is considerable: these types of compound adjectives account for almost 70% of all deonymic adjectives in our data set.

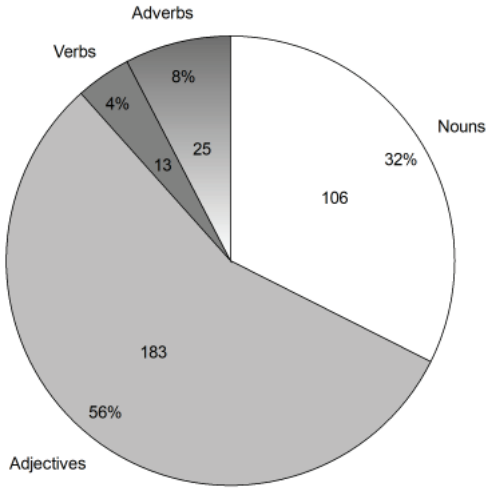

Fig. 2. Deonymic appellatives seen from the parts of speech perspective

The parts-of-speech-based appellativization preferences are shown in detail in Fig. 3. Here, too, there are important differences among individual names, which are related to historical and political circumstances. The predominance of adjectival formation in the case of the anthroponym *Masaryk* is due to his conception of Czech history, which produces a chain of bearers/promoters of his humanitarian ideals (*havlíčkovskomasarykovský* 'Havlíček-Masarykian', *husovsko-palacko-masarykovský* 'Hus-Palacký-Masarykian'), the dispute that resulted from this conception (*masarykovsko-perkařovský* 'Masaryk-Pekařian'), and his anchoring in the democratic current among the Czech intellectuals (*čapkovsko-masarykovský* 'Čapek-Masarykian', *havlovsko-masarykovský* 'Havel-Masarykian'). However, there are also ironic adjectives bringing Masaryk's legacy into a new and unconventional light (*masarykovsko-gandalfovský* 'Masaryk-Gandalfian', *masarykovskogottwaldovský* 'Masaryk-Gottwaldian'). Adjectives, on the other hand, are neglected in the case of Adolf Hitler: his appellativization products are dominated by variously formed and deliberately ironic/derogatory nouns for his supporters (*hitlerovec*, *hitlerčík*, *hitleráček*, *hitlerek*, *hitlerka* 'Hitler-follower'), or for things and phenomena associated with the dictator (*hitlerologie*, *hitlerománie*).

The anthroponyms *Beneš* and *Stalin* are appellativized in a way similar to the surname *Masaryk*. As for *Beneš,* given the relatively low frequency of the appellativized lexical units, what turns out to be prominent are verbs as well (*odbenešit*, *odbenešovávat* 'de-Benešize') and adverbs (*benešovsky*, *probenešovsky* 'Beneš-like', *protibenešovsky* 'anti-Beneš-like'), which are connected to the development of the situation in the Second Republic.

Adjectives related to the name *Stalin* combine the two factors mentioned above – they contextualize the phenomenon within Stalin's ideology (*poststalinský* 'post-Stalinist', *neostalinský* 'neo-Stalinist') and connect the Soviet politician to other personalities or concepts, sometimes in a very complex way (*gestapácko-stalinský* 'Gestapo-Stalinist', *marxisticko-stalinistický* 'Marxist-Stalinist', *poststalinsko-estébácký* 'post-Stalinist-Communist-secret-police', the police being known as StB, read: /ɛstɛːbɛ/).

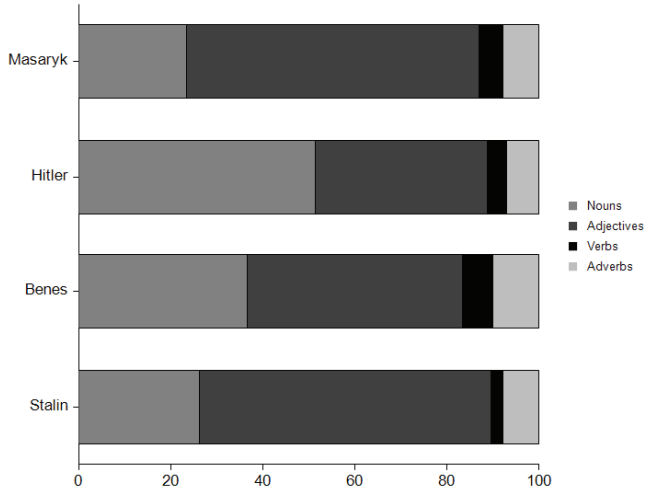

Fig. 3. Deonymic appellatives seen from the parts of speech perspective – detailed analysis

3.2 Appellativization and word-formation strategies

From the point of view of word formation, the products of appellativization were divided into six groups; the main criterion was the mode of formation, with the proprium considered to be the base. We consider this approach transparent and in line with the general idea of the deonymic appellatives originating from the respective proper names. The word formation strategies are thus the following:

a) suffixation (*masarykovský* 'Masaryk.ADJ', *hitlerovec* 'Hitler-follower.N', *benešolog* 'expert on Beneš.N', *staliniáda* 'Stalinias.N');
b) prefixation[1] (*skoro-hitler* 'almost-Hitler', *anti-stalin* 'anti-Stalin');
c) prefixation + suffixation (*anti-masarykovský* 'anti-Masaryk.ADJ', *pohitlerovský* 'post-Hitlerian', *odbenešit* 'deBenešize.V', *exstalinista* 'ex-Stalinist');
d) composition and juxtaposition (two-member composites: *Hitlerjóga* 'yoga in the style of Hitler');

[1] It is to be noted that the prefix/prefixoid and suffix/suffixoid distinctions were not taken into account in our paper and that some words in this category may be considered compounds, such as *skoro-hitler*. In our viewpoint, we considered *skoro-* a prefix, as it is quite productive and can be theoretically added to all sorts of words.

e) composition + suffixation (two-member composites: *masarykovsko-havlovský* 'Masaryk.ADJ-Havel.ADJ', *hitlerovsko-darwinovský* 'Hitler.ADJ-Darwin.ADJ', *benešovsko-nejedlovský* 'Beneš.ADJ-Nejedlý.ADJ', *gottwaldovsko-stalinský* 'Gottwald.ADJ-Stalin.ADJ');
f) composition + suffixation (three-member composites: *jiráskovsko-masarykovsko-komunistický* 'Jirásek.ADJ-Masaryk.ADJ-Communist.ADJ', *benešovsko-gottwaldovsko-stalinský* 'Beneš.ADJ-Gottwald.ADJ-Stalin.ADJ', *stalinisticko-nacionalisticko-mafiánský* 'Stalin.ADJ-nationalistic-Mafia ADJ');
g) blending (Stalin + *noviny* 'newspaper' > *Haló stalinoviny* – an occasionalism for the extremist left-wing periodical *Haló noviny*).

Looking on the overall results depicted in Fig. 4, none of the word-formation strategies can be described as dominant. The most frequent ones are composition and suffixation with two-member composites (type d), and prefixation and suffixation (type c). As far as composition is concerned, coordinating composites (*masarykovsko-čapkovský* 'Masaryk.ADJ-Čapek.ADJ', *stalinsko-zemanovský* 'Stalinist-Zeman.ADJ') predominate, which are, however, relatively rare in the Czech system otherwise (cf. Bozděchová 2018, p. 935). This type became popular in Czech as late as the 19th century; it first appeared for naming of colours, and it gained a certain productivity in the second half of the 20th century only (Šlosar 1999, pp. 66–67). We have already tried to explain the background of these occurrences above (see Section 3.1).

In terms of word-formation strategies, composition is the most prominent in adjectives; it is rarely (only a few lexical units) represented in nouns. It is then mostly used expressively, cf. *masarykobijec* 'Masaryk-beater' is attested already before 1918, and we also found expressions clearly influenced by German: *hitlerjóga* 'yoga in the style of Hitler', *kulturhitlerismus* 'cultural Hitlerism'. Composition is represented minimally for verbs (*heilhitlerovat* 'hail to Hitler') and adverbs (*masarykovsko-havlovsky* 'Masaryk-Havel-like') as well.

In contrast to composition, prefixation and suffixation are standard and continuously productive in the Czech word formation – in the case of the deproprial lexical units analyzed here, they show an attitude (*protistalinský* 'anti-Stalinist', *promasarykovsky* 'pro-Masaryk') and contextualize the political ideologies, as discussed above (*protohitlerovský* 'proto-Hitlerian', *restalinizace* 're-Stalinization'). Some of the prefixes (especially the attitudinal ones *pro-*, *anti-*, *anti-*) were used with all the four analyzed propria; however, there were also differences. Once we speak about the effort to get rid of the given political figure's influence, for the two foreign politicians – Stalin and Hitler – the foreign prefix *de-* is used (*destalinizátor* 'de-Stalinizer'), while in the case of the Czech politicians Masaryk and Beneš, the Czech prefix *od-* is predominant (*odmasarykovštění* 'de-Masarykization'). The richest repertoire of prefixes is exhibited by the appellatives formed from the surname *Stalin*; some of these are found exclusively with this base (*arci-*, *ex-*, *super-*, *ultra-*).

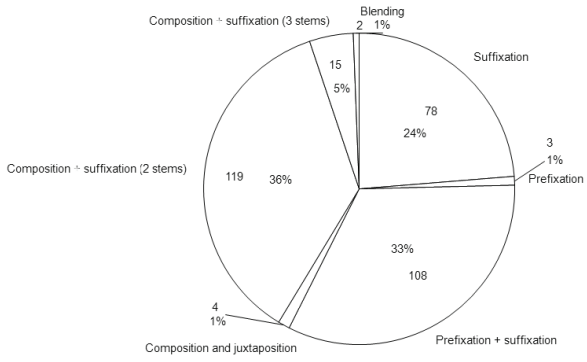

Fig. 4. Word-formation strategies as used with the deonymic appellatives

Fig. 5 provides a more detailed view of the word-formation situation. Unlike the analysis of the parts of speech, where the appellatives formed from the names *Masaryk* and *Stalin* showed some similarities (see Fig. 2), in terms of word formation, each proprium behaves specifically. Terms derived from the name *Hitler* use suffixes very often; this is related to the derivational richness of the names of his followers (see Section 3.1), while in the case of the anthroponym *Stalin*, the prefixal-suffixal and compositional strategies dominate, and we also find here the otherwise rather sparse compositions of three-word bases. The latter type appears both in expected contexts and also in defamiliarizations (*stalinsko-gottwaldovsko-reicinovský* 'Stalin.ADJ-Gottwald.ADJ-Reicin.ADJ', *stalinsko-hitlerovsko-náserovský*, 'Stalin.ADJ-Hitler.ADJ-Nasserian.ADJ') and is intended to highlight the pervasiveness of Stalinist ideology and to reveal (or construct) its surprising connections. The composition strategy is the most prominent in the name *Masaryk*: it is related to his conception of Czech history and its sources and resonances (see Section 3.1).

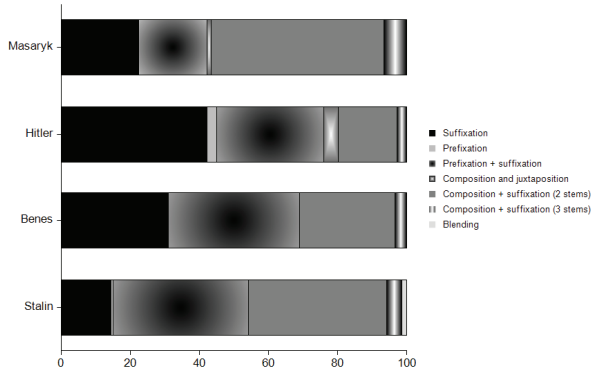

Fig. 5. Word-formation strategies as used with the deonymic appellatives – detailed analysis

4 CONCLUSION

The corpus-based analysis of appellativized names formed from the surnames of the politicians *Masaryk*, *Beneš*, *Hitler*, and *Stalin* has shown several specific features of the appellativization of anthroponyms. In terms of parts of speech, these names predominantly produce descriptive adjectives. The most frequent word-formation strategies are composition and suffixation (two-member composites), and prefixation and suffixation. An important factor influencing the word-forming potential is the prominence/controversiality of the politician in question and the continuous updating of his legacy, as shown by the units formed from the surname *Stalin*.

We would like to compare the conclusions of this study with the process of appellativization in toponyms (cf. e.g. product names such as *Manchester – manchester*/*manšestr* 'corduroy'). In their case, we expect both a lower word-forming potential and a significantly limited range of word-forming strategies; we can even expect the absence of some strategies used exclusively for anthroponyms (e.g. pluralization).

ACKNOWLEDGEMENTS

The research was supported by the grant project GAČR 22-09310 *Kvantitativní onomastika: východiska, koncepty, aplikace* [Quantitative Onomastics: Starting Points, Concepts, Applications], provided by the Czech Science Foundation (GAČR), Czech Republic.

References

Bozděchová, I. (2018). Tvoření adjektiv IV. Kompozita. In F. Štícha et al.: Velká akademická gramatika spisovné češtiny I, part 2. Praha: Academia, pages 915–942.

Český národní korpus. SYN version 11. Praha: Ústav Českého národního korpusu FF UK 2022. Accessible at: https://www.korpus.cz/kontext/query?corpname=syn_v11. [accessed January 25, 2023].

David, J. (2009). Specifické antroponymické a toponymické formanty v nové češtině a jejich vnímání. In V. Čermák – M. Příhoda (eds.): Slovanský areál a Evropa. Praha: Pavel Mervart – FF UK 2009, pages 229–236.

David, J., and Místecký, M. (2023). Prolegomena ke kvantitativní onomastice. Acta onomastica, 64(2), pages 301–320.

David, J., Klemensová, T., Místecký, M. et al. (2022). Od etymologie ke krajině. Onomastika pro 21. století. Brno: Host.

Děngeová, Z. (2010). Nové lexikální jednotky motivované jmény českých politiků. Bohemistyka, 10(3), pages 167–186.

Harvalík, M. (2012). Appellativisation and Proprialisation: The Gateways between the Appellative and Proprial Spheres of Language. In O. Felecan (ed.): Name and Naming.

Synchronic and Diachronic Perspectives. Newcastle upon Tyne: Cambridge Scholars Publishing, pages 10–17.

Hladká, Z. (2017). Apelativizace. In P. Karlík – M. Nekula – J. Pleskalová (eds.): CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/APELATIVIZACE. [accessed January 25, 2023].

Jandová, E. (2013). Neologismy tvořené z příjmení politiků (od čunkodomků přes bobotrik k polévce havlovačce). In D. Hradilová et al.: Proměny slova. Olomouc: Univerzita Palackého, pages 60–70.

Malačka, O. (2011). KdeJsme.cz. Accessible at: https://www.kdejsme.cz. [accessed February 2, 2023].

Martincová, O. (2011). Depropriální neologismy v současné češtině. In I. Valentová (ed.): Jazykovedné štúdie, 29. Život medzi apelatívami a propriami. Bratislava: Vydavateľstvo SAV, pages 20–37.

Michalec, V. (2013). Depropriální neologismy tvořené z příjmení. In M. Drinov (ed.): Problemi na neologijata v slavjanskite ezici. Sofija: Akad. Izd., pages 153–162.

Motschenbacher, H. (2020). Corpus Linguistic Onomastics. A Plea for a Corpus Based Investigation of Names. Names, 68(2), pages 88–103.

Pokorná, E. (1978). Apelativizovaná jména v české slovní zásobě. Slovo a slovesnost, 39(2), pages 116–124.

Skujiņa, V. (1989). Apelativization and separation as method of word-formation. Baltistica, 3(2), pages 408–413.

Superanskaya, A. V. (2012). Obschaya teoriya imeni sobstvennogo. Moskva: Khnizhnyj dom LIBROKOM.

Šlosar, D. (1999). Česká kompozita diachronně. Brno: Masarykova univerzita.

Štícha, F. (2018). Slovotvorná produktivita. In F. Štícha et al.: Velká akademická gramatika spisovné češtiny I, part 1. Praha: Academia, pages 168–172.

Šrámek, R. (1999). Úvod do obecné onomastiky. Brno: Masarykova univerzita.

Šrámek, R. (2003–2004). Transonymizace v propriální nominaci. Folia onomastica Croatica, 12–13, pages 499–508.

sciendo
DOI 10.2478/jazcas-2023-0022

THE ECONOMY OF CZECH EXCHANGE IN THE SLOVAK MARKETPLACE OF AUSTRIA AFTER THE FALL OF HUNGARY

MARTIN DIWEG-PUKANEC
Department of Slavic Philology, Faculty of Arts,
Constantine the Philosopher University in Nitra, Nitra, Slovakia

DIWEG-PUKANEC, Martin: The Economy of Czech Exchange in the Slovak Marketplace of Austria after the Fall of Hungary. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 43 – 51.

Abstract: Lexical quanta in writing and speech are important indicators of an individual's social class. This paper analyses the lexical features of the utterances produced by writers/speakers from different social classes in terms of word length based on a representative sample of letters from the Kremnica archive. The study found that writers/speakers from different social classes showed different average word length in the 2nd half of the 16th century. According to the analysis of the letters, writers/speakers from the upper classes produced utterances with longer words than those from the lower classes. These differences are explained by factors related to the individual's social class backgrounds and his "right to speech" or "right to be read". The linguistic economy principle, objective of which is to save more time and energy by conveying more information with less effort is thus far from exhaustive and by no means reflects the whole sociolinguistic reality.

Keywords: applied linguistics, economy principle, language management, linguistic marketplace, social class, sociolinguistics, word length

1 THE ECONOMICS OF LINGUISTIC EXCHANGES

Pierre Bourdieu in his paper *L'économie des échanges linguistiques* (1977a) introduces the notion of a linguistic marketplace or linguistic market (*marché linguistique*), which he develops in a significant way in his English article with the same title (1977b). The main idea of the studies is that in the symbolic marketplace, where linguistic exchanges happen, linguistic (or cultural) capital is exchanged, and different varieties in the marketplace have different symbolic market values. The *marché linguistique* in this paper will be, in terms of space, the north-eastern part of the monarchy, what would become Austria, which was called Slovak land (*Windenland* in German), specifically its Czech-speaking population from among the nobility and bourgeoisie in communication primarily with the town of Kremnica. In terms of time, it is the period not long after the defeat of Hungary at Mohács in 1526 and practically the end of its existence as an independent state. The period can be defined as the 2nd half of the 16th century. Language varieties are divided according to social class.

After the fall of Hungary, where Latin was dominantly used in writing, Czech began to dominate the linguistic market in the Slovak land. This was partly a result of political revolution after the defeat, but at the same time it happened gradually as a result of a slow transformation of material and symbolic power relations (like the steady devaluation of French on the world market, relative to English), which was a consequence of an increasingly close relationship of the Slovak land to the Holy Roman Empire including the Czech lands. Those who sought to defend a threatened capital of Latin in the so-called Royal Hungary were forced to conduct a struggle, because they could not save the competence without saving the market, i.e. all the social conditions of the production and reproduction of producers and consumers (cf. Bourdieu 1977b, p. 651). The Latin language began to be clearly dominant again in the territory of the former Hungary only after its liberation. In the period relevant for our research, however, the kingdom was dominated by a population with one mother tongue. This was Slovak, and spoken Slovak, of course, penetrated into the splendid Czech written culture radiating from the Czech lands, since Czech and Slovak have always been mutually intelligible. S. Czambel (1890, p. 29) therefore also calls the Czech of the Royal Hungary of the 2nd half of the 16th century "Czech-Slovak", and considers its most important texts to be A) 15 letters of the Révays, B) 18 letters of the town of Mošovce, C) 9 letters of the Benickýs, D) 10 letters of the Majthénys, and E) 3 official documents. The texts thus determined provide a good basis for a corpus intended for sociolinguistic comparison of lexical quanta in writing and speech. As Grotjahn and Altmann (1993, p. 143) point out, letters as a specific text type have been considered to be "prototypical" texts, optimally representing language due to the interweaving of oral and written components.

However, the comparison in the article is made on the basis of social classes, and Czambel's corpus of "Czech-Slovak" texts needs to be modified somewhat for this purpose. The Révays and the Majthénys had castle estates and can be considered middle aristocracy. The Benickýs, since they did not have castle estates, are considered to be landed gentry. Branches of both the Révays and the Majthénys were included among the titled aristocracy relatively early after the fall of Hungary, where only a very limited circle of aristocracy had held titles (41 persons in 1498, cf. Pukanec 2016, p. 52), while the Benickýs never did, which also partly confirms our division. The small but important town of Mošovce was in constant conflicts with the Révays, thus showing its equality with them, so that the mayor and members of the town council can be considered as middle bourgeoisie, which is roughly at the level of the middle aristocracy. However, since the town *de jure* belonged to the Révays, they should have stood socially below them. Each of the three official documents mentioned by Czambel is written by someone else, and they are also few in number overall, which means that they are not representative enough, so we do not include them in the corpus. We will, however, expand the corpus with at least a few archived Czech documents from this period from the upper aristocracy, namely

the prince of Cieszyn (8 letters), the landed gentry, namely George Jesenský (6 letters), and the petty bourgeoisie, namely the Holeš family (6 letters). This gives us a slightly broader sociolinguistic picture. The letters are published in *Slovenský letopis I-VI* (Sasinek 1876, 66ff., 159ff.; 1877, 76ff.; 1879, 73ff., 241ff.;1880, 165ff., 337ff.; 1881, 66ff.; 1882, 80ff.) and are available at https://www.hathitrust.org/, but some manual editing was needed.

All 72 letters analysed are dated from roughly the same period, which is important because the average word length increases and decreases over the centuries (Bochkarev et al. 2012). The writings are found in the archives of Kremnica, a very rich town at the time, and in the vast majority of cases they are also addressed directly to Kremnica, its mayor, high officials and the town council (upper bourgeoisie). This means that the addressee is virtually always the same, which highlights the differences on the production side (the letter writers) and makes the research more precise. As Bourdieu (1977b, p. 648) states: "The structure of the linguistic production relation depends on the symbolic power relation between the two speakers, i.e., on the size of their respective capitals of authority... Language is not only an instrument of communication or even of knowledge, but also an instrument of power. A person speaks not only to be understood but also to be believed, obeyed, respected, distinguished. Hence the full definition of competence as the right to speech... An adequate science of discourse must establish the laws which determine who (*de facto* and *de jure*) may speak, to whom, and how" – including *how long*, we might add, since finding out this is the purpose of the paper. The production is governed by the structure of the market, and when one side of the communication is constant, we can accurately measure the differences of the other side.

This is the basis of our hypothesis in the paper that the higher a social class, the greater the "right to speech", which is quantitatively reflected in the length of words. The hypothesis is also based on the knowledge of a letter addressed to our great-grandfather John Majthény from his subjects in the 16th century, which is in our family library, where the word length is extremely low, and thus the economy of the text is greater. Contemporary assertion that "linguistic economy principle is one of the generally recognized mechanisms, the objective of which is to save more time and energy by conveying more information with less effort" (Zhou 2012, p. 100) is thus far from exhaustive and by no means reflects the whole sociolinguistic reality. As Bourdieu (1977b, p. 646) puts it: "Language is a *praxis*: it is *made for saying*, i.e., for use in strategies which are invested with all possible functions and not only communication functions. It is made *to be spoken appropriately*." This is the problem of *καιρός* [kairos], of doing the right thing at the right time, which was central to the Sophists. The linguistic economy principle formulated in this way ignores the variations in the structure of the linguistic production relations between a speaker and a receiver, which depend on the interlocutors' positions in the symbolic power

relations. The specific characteristics of the work of linguistic production depend on the linguistic production relation inasmuch as the latter is the actualisation of the objective power relations (e.g. class relations) between two speakers (or the groups to which they belong) (Bourdieu 1977b, p. 651).

2 WORD LENGTH IN THE TEXTS

Measuring the length of words in older texts is not new. Something similar has been attempted, for example, by researchers F. Lian and Y. Li (2019). Their research is one of the first to have at its disposal a large corpus of texts and provides an analysis of word length distribution in German based on 360 texts originated between the 17th and 19th centuries. Our corpus of Czech letters is therefore comparatively smaller, but it has the advantage of uniformity of time period and addressee. The question is naturally how to measure word length in text corpora with software tools so that the measurement results are relevant. As Grotjahn and Altmann (1993, pp. 142–143) state with regard to word length, there are three basic types of measurement: graphic (e.g. letters), phonetic (sounds, phonemes, syllables, etc.), and semantic (morphemes). And, as a consequence, it is obvious "that the choice of the unit of measurement strongly effects the model of word length to be constructed". Of course, we must not forget that there are other lexical quanta than word length, such as lexical diversity or lexical density (cf. Strömquist et al. 2002), but these will not be of interest to us in this paper. Of the above-mentioned methods of measuring word length, if one looks for a method in all languages, the general view prevails that "[f]or evaluating word length one cannot use a different measurement unit but the number of syllables" (Popescu et al. 2013, p. 225). It should be noted here, however, that the letters used for the purposes of this study are written more or less monolingually, i.e., in Middle Czech (or in Slovakized Czech), only in some of them the introductory and concluding formulae are in Latin. In this paper, we have therefore chosen to apply grapheme measurement because the "right to speak" in written communication is also interpreted as the "right to be read" and it is the graphemes that are read.

However, in addition to the graphic measurement, we add at least a phonetic one based on phonemes. Slavic languages with their simple orthographies, where the rule of 1 grapheme = 1 phoneme almost always applies – with the exception of the older digraph orthography, which was also often used in the Middle Czech –, do not need to be measured on syllables, especially when one considers that some syllables are much longer than others (cf. Czech *žbluňk* 'splash' and *a* 'and'). In phonetic transcription, we apply the simplification of digraph orthography as follows: *cz* > *c*/*č*, *dd*, *dt* > *d* (in *Radd[a]* 'council', etc.), *dč*/*dts*/*dž* > *č*, *dz* > *ʒ*, *ff* > *f*, *ch* > *X*, *ij* > *i*/*í*, *ll* > *l*, *pp* > *p*, *rz*/*rž* > *ř* (with the exception of *skrze*, *[]drz[at]*/*[]drz[et]*), *ſſ* > *Š*, *ss*/*sz* > *s*, *th*/*tt* > *t*, *zz* > *z*/*ž*. This simplification relatively does not change the average word

length, because it is more or less constant in all letters, but the average word length in the texts is increased quite considerably by writing out the Czech abbreviation *V.M. > Vaš[e] Milost[]* 'Your Grace'. We chose to write it out in order to measure the speech against the writing to a greater extent. This affected the relative word length in the phonetic transcription versus the graphic transcription significantly, especially in the case of the Benickýs, which requires an explanation. Other abbreviations also occur in the letters. For our research, however, only the Czech abbreviation overused by the town of Mošovce is important: *V.Opp. > Vaš[e] Opatrnost[]* 'Your Prudence'. Moreover, the text is to some extent skewed by proper names, of which there are quite a large number and which are objectively given, but we have decided to count them, because sometimes the boundary between proper and common names is disputable (cf. *Most Sučansky* 'bridge of Súča' (= village in Slovakia)). We take the letters as separate units, not as a sub-corpus for a social class.

2.1 Upper aristocracy

We start our measurements from the highest social class, which in our corpus is represented only by an individual – the prince of Cieszyn. His texts are characterized by the use of the majestic plural (royal *we*), which confirms his inclusion in the upper aristocracy. The prince addressed 7 of his 8 letters directly to the mayor and the town council of Kremnica. The second of the 8 letters (in italics in the Tab. 1) has no direct addressee. It should be noted then that word length is highest here. Except for a few abbreviations, there is no Latin in the letters.

	1.	*2.*	3.	4.	5.	6.	7.	8.	AVG
graphic	4.77	*5.35*	5.15	5.09	5.16	5.18	5.03	5.16	**5.11**
phonetic	4.49	*5.13*	4.72	4.71	4.93	4.88	4.81	4.92	**4.82**

Tab. 1. Upper aristocracy

2.2 Middle aristocracy

In the second social class of our survey, and the second in order within our social stratification, we analyse the letters of two families: the Révays and the Majthénys. Unlike the letters of the prince of Cieszyn, there are several writers, so the letters are of a greater informative value in terms of our research. The letters of the Révay family are mostly addressed to the mayor and the council of Kremnica, but not all of them. Among the 15 letters of the family (Tab. 2 and 3), the letters 1–3 were written by two barons who used the majestic plural. However, their proximity to the upper aristocracy is not confirmed by the word length measurements, and the results are average. In the table, to make this more visible, they are in italics. The other letters (4–15) were written by two persons of the Révay family of a lower social rank. Tab. 4 shows the results concerning the Majthény family. In terms of state administration, the Majthény family as a whole was less important than the

Révay family. The Majthény family addressed 9 out of 10 letters of the corpus to the mayor and the council of Kremnica. The authors of the letters were several members of the family, but all but the last one seem to have originated from the office of *vicecomes* Christopher. Latin formulae are much more frequent in the texts of the Révay family, which may distort the results compared to Majthény's lexical quanta (Latin formulae generally have longer words than Czech ones). Indeed, as some models of word length measurement show, the measurements should have their own specificities for each language (cf. Grzybek 2007, p. 66).

	1.	2.	3.	4.	5.	6.	7.	8.
graphic	*4.56*	*4.83*	*4.85*	4.91	5.16	5.36	4.54	4.40
phonetic	*4.34*	*4.80*	*4.70*	4.55	4.99	5.00	4.44	4.69

Tab. 2. Middle aristocracy – Révay I

	9.	10.	11.	12.	13.	14.	15.	AVG
graphic	4.80	4.63	4.63	5.18	4.53	4.78	4.94	**4.81**
phonetic	4.64	4.54	4.61	5.18	4.85	4.82	4.87	**4.73**

Tab. 3. Middle aristocracy – Révay II

	1.	*2.*	3.	4.	5.	6.	7.	8.	9.	10.	AVG
graphic	4.97	4.51	4.66	4.44	4.34	4.25	4.31	5.19	5.35	5.03	**4.71**
phonetic	4.59	4.34	4.50	4.33	4.25	4.17	4.17	4.99	4.98	4.78	**4.51**

Tab. 4. Middle aristocracy – Majthény

2.3 Middle bourgeoisie

The letters of this social class were almost always written by the mayor and the council of the town of Mošovce, or at least by the mayor of the town, the addressee was almost always the mayor and the council of the town of Kremnica, or at least the mayor of the town. Latin formulae occur in relatively small amounts, comparable to those of the Révay family, so at least with respect to them they do not substantially bias the results. The formulae may, however, increase the length of the words compared to the Majthénys. In some letters the Czech abbreviations *V.M.*, and *V.Opp.* are excessively used, which leads to the fact that the phonetic measurement here reaches values slightly above the level of the middle aristocracy, although the graphic measurement does not.

	1.	2.	3.	4.	5.	6.	7.	8.	9.
graphic	4.61	4.81	5.04	4.85	4.95	4.61	4.88	4.71	4.57
phonetic	4.63	4.57	4.72	4.60	4.68	4.43	4.81	4.59	4.52

Tab. 5. Middle bourgeoisie I

	10.	11.	12.	13.	14.	15.	16.	17.	18.	AVG
graphic	4.52	4.60	5.04	4.99	4.97	4.51	4.89	4.35	4.69	**4.76**
phonetic	4.49	4.73	5.22	5.09	4.79	4.36	4.74	4.40	4.77	**4.67**

Tab. 6. Middle bourgeoisie II

2.4 Landed gentry

The letters of this social class were written by three members of the Benický family and by the nobleman George Jesenský. The letters of Jesenský were all written by the same person to the mayor and the town council of Kremnica except for the last one that was addressed to a high official of the town. The largest number of the letters of the Benický family was written by *vicecomes* Jan. We can only speculate that, while the speech of the Benický family was at the level of the *vicecomes* Majthény, they did not have sufficient *certitudo sui* (self-certainty) in their written speech and the feeling that they had a "right to be read". In any case, the fact that they lacked the *certitudo sui* that is typical of dominant class texts (cf. Bourdieu 1977b, p. 659) proves the excessive use of the abbreviation *V.M.* even in completely unjustified cases ("*V.M. take prosym ze byste V.M.…*" 'I also ask Your Grace that Your Grace...', etc.). There is relatively very little Latin in letters of this class.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	AVG
graphic	3.97	3.95	4.30	4.86	4.50	4.03	4.58	4.74	4.50	**4.38**
phonetic	4.18	4.33	4.55	5.03	4.37	4.35	4.57	4.66	4.61	**4.52**

Tab. 7. Landed gentry – Benický

	1.	*2.*	3.	4.	5.	6.	AVG
graphic	5.19	4.40	4.32	4.56	4.26	4.45	**4.53**
phonetic	4.93	4.26	4.23	4.51	4.23	4.36	**4.42**

Tab. 8. Landed gentry – Jesenský

2.5 Petty bourgeoisie

All the letters attributed to this social class were addressed to the mayor of Kremnica. They show the inferior education of the writers, who are probably father and son of the Holeš family. This is proved by the absence and erroneousness of the Latin: there is only one abbreviation (*m.p.*) and only one phrase (*Post Cripta*, which should have been *Post Scripta*) in Latin. The formulations are cumbersome (frequent beginnings of sentences *Zo se dotize...* 'As regards...') and expressive (e.g. the only rhetorical question in our corpus is found in the letters of this social class). The texts show that the writers are members of the lower class (*Yako Vaša milost roskazete.* 'As Your Grace will command.'). Thematically, too, these letters are considerably

"lower" (*Zo se dotize prasiez...* 'As regards pigs...', *Zo se dotize teho hnoga...* 'As regards the manure...'). The abbreviation *V.M.* is overused in the letters.

	1.	*2.*	3.	4.	5.	6.	AVG
graphic	4.31	4.38	4.27	4.25	4.11	4.09	**4.24**
phonetic	4.46	4.33	4.23	4.23	4.06	4.07	**4.23**

Tab. 9. Petty bourgeiosie

3 CONCLUSIONS

The results of the graphic and phonetic measurements of word length from the highest social class to the lowest are approximately as follows: 5.11 – 4.77 – 4.76 – 4.44 – 4.24, and 4.82 – 4.64 – 4.67 – 4.48 – 4.23, respectively. Since the methods of measurement are not accurate, the corpus of texts is limited, we do not know the symbolic power relations between the speakers and the receiver(s), on which linguistic production substantially depends, nor do we know anything about the personalities of the writers (e.g. their education) apart from their social classes, between which there are not even exact boundaries (the bourgeois used to be nobles, a part of the aristocratic family belonged to the upper, part to the middle or lower aristocracy, etc.), we cannot use these results to their full extent. However, by establishing texts with almost the same addressee in the same period, we have brought a new methodological element to word-length research that makes the measurements more precise. Based on the results of our measurements, we can establish a clear tendency and trend: that is, the less socially significant the class writing and speaking Czech in the Slovak linguistic market in Austria after the fall of Hungary was, the more economical the words in their utterances were. This was due to the fact that they had a *different* (in terms of the above-mentioned linguistic economy principle neither better nor worse) *language management*. However, this study is more quantitative than qualitative in nature and the findings are far from conclusive.

References

Bochkarev, V. et al. (2012). Average word length dynamics as indicator of cultural changes in society. Social Evolution and History, 14(2), pages 153–175.

Bourdieu, P. (1977a). L'économie des échanges linguistiques. In P. Encrevé (ed.): Langue française n°34. Linguistique et sociolinguistique, pages 17–34.

Bourdieu, P. (1977b). The economics of linguistic exchanges. Social Science Information, 16(6), pages 645–668.

Czambel, S. (1890). Slovenský pravopis. Budapest: Viktor Hornyanszky, 272 p.

Grzybek, P. (2007). History and methodology of word length studies. In P. Grzybek (ed.): Contributions to the Science of Text and Language, pages 15–90.

Grotjahn, R. – Altmann, G. (1993). Modelling the distribution of word length: some methodological problems. In R. Köhler – B. Rieger (eds.): Contributions to Quantitative Linguistics. Dordrecht: Kluwer, pages 141–153.

Lian, F., and Li, Y. (2019). Word length distribution in German texts during the 17th-19th century. Journal of Quantitative Linguistics, 28(2), pages 1–21.

Popescu, I.-I. et al. (2012). Word length: aspects and languages. In R. Köhler – G. Altmann (eds.): Issues in Quantitative Linguistics, pages 224–281.

Pukanec, M. (2016). Histoire du royal Pays slovaque. Saint-Denis: Edilivre, 290 p.

Sasinek, F. V. (1876). Slovenský letopis I. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/012238871.

Sasinek, F. V. (1877). Slovenský letopis II. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/012238871.

Sasinek, F. V. (1879). Slovenský letopis III. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/012238871.

Sasinek, F. V. (1880). Slovenský letopis IV. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/100405334.

Sasinek, F. V. (1881). Slovenský letopis V. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/100405334.

Sasinek, F. V. (1882). Slovenský letopis VI. Skalica: Jozef Škarnicel, 348 p. Accessible at: https://catalog.hathitrust.org/Record/100405334.

Strömquist, S. et al. (2002). Towards a cross-linguistic comparison of lexical quanta in speech and writing. Written Language & Literacy, 5(1), pages 45–67.

Zhou, G. (2012). On the embodiment of economy principle in the English language. English Language and Literature Studies, 2(2), pages 100–104.

sciendo
DOI 10.2478/jazcas-2023-0023

STATISTICIAN, PROGRAMMER, DATA SCIENTIST? WHO IS, OR SHOULD BE, A CORPUS LINGUIST IN THE 2020S?

ŁUKASZ GRABOWSKI
Institute of Linguistics, The Faculty of Philology, University of Opole,
Opole, Poland

GRABOWSKI, Łukasz: Statistician, Programmer, Data Scientist? Who Is, or Should Be, a Corpus Linguist in the 2020s?. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 52 – 59.

Abstract: In this short essay, I aim to ruminate on the nature of a corpus linguist's work in the 2020s, a time marked by unprecedented advancements in the field of computer technologies and artificial intelligence. This seems to be particularly relevant considering the theme of the 12th International Conference Slovko 2023, which is "Natural Language Processing and Corpus Linguistics". In the last two decades or so, corpus linguistics has drawn extensively from the fields such as statistics, computer science and data science. In many respects corpus linguistics has served as a significant source of inspiration for progress in the field of natural language processing (NLP), leading to the development of large language models (LLMs) as well as recent introduction of conversational artificial intelligence, among others. Thus, in this paper I will make an attempt at identifying the skills that may help rank-and-file or aspiring corpus linguists to survive and, hopefully, flourish in the research field in the 2020s.

Keywords: corpus linguistics, statistics, computer programming, natural language processing, artificial intelligence

1 INTRODUCTION

Linguistic analyses of extensive text volumes were performed well before the rise of computer technologies (McEnery – Wilson 1996, p. 9), yet the character of those studies was significantly different from modern practices. Back then, limited computer tools and data processing capabilities made such analyses, often conducted manually, laborious, time-consuming, and sometimes unfeasible. However, since the late 1980s and early 1990s, the emergence of "computer corpus linguistics" (Ooi 1998, p. 34) has revolutionized the field[1]: it has become the research area that involves employing computer software and tools as well as various statistical techniques to study language through vast collections of naturally occurring text, enabling investigations into various linguistic phenomena, notably into the interplay

[1] Computer corpus linguistics "involves the study of language on the basis of textual or acoustic corpora, almost always involving the computer in some phase of storage, processing, and analysis of data" (Ooi 1998, p. 34).

of grammatical and lexical units. Today, however, simple descriptive statistics or concordance analysis is often not enough to explain more fine-grained patterns of language use in massive collections of linguistic data. In the last two decades, the range of statistical methods of corpus analysis (e.g. multivariate statistics) as well as data visualization techniques has expanded considerably. Today, in the 2020s, it is no surprise anymore that in many language studies researchers also make ample use of machine learning techniques popular in the field of natural language processing (NLP), also known as computational linguistics (Hirschberg – Manning 2015, p. 261).

In short, NLP researchers focus on the development of algorithms and models for processing, producing, learning and understanding human language (Hirschberg – Manning 2015; Dunne 2022; Jurafsky – Martin 2023). It is often the case that NLP researchers use text corpora compiled by corpus linguists to develop computational models aimed at performing various language-related tasks, such as text classification, text summarization, information retrieval, automatic speech recognition, machine translation, sentiment analysis, authorship analysis, semantic role labelling, coreference resolution, to name but a few (Dunne 2022; Jurafsky – Martin 2023). NLP methods facilitate automation of various routine research tasks performed by corpus linguists, namely corpus queries and corpus analyses, some of them earlier conducted manually, to identify and explore patterns of language use, that is, frequencies and distributions of linguistic forms and their classes in texts. Consequently, Dunne (2022) argues that "corpus analysis can be expanded and scaled up by incorporating computational methods from natural language processing".

There is an important caveat, though. Dunne (2022, p. 1) observes that natural language processing and computational linguistics develop so fast that "these advances have become increasingly disconnected from corpus linguistics and linguistic theory". This observation becomes even more prominent in the light of recent developments in the field of artificial intelligence. Crosthwaite and Baisa (2023, p. 1) argue that "our field [corpus linguistics – ŁG] risks being overshadowed by GenAI [generative artificial intelligence – ŁG] researchers who are essentially just doing what we as corpus linguists already do, but in a way that has finally captured the imagination of the public." It seems that GenAI-assisted linguistic studies will appear sooner rather than later, and they will most probably contribute to the already exponentially growing number of researchers, journals and scholarly papers.[2] Hence, it can be expected that language researchers will face two scenarios.

[2] Hyland (2023, p. 2) cites UNESCO data according to which the number of researchers in the years 2014–2018 "grew three times faster than the world population", and that in the year 2022 alone the publishing giant Elsevier received 2.7 million paper submissions, compared to 2.5 million in 2021; the number of published papers, which was over 600,000 in 2021, represented an increase of 89% compared to the previous decade.

Either the disconnect between corpus linguistic, linguistic theory, on the one hand, and NLP and GenAI, on the other, will grow even further, or the combination of theories, methodologies, tools and solutions from all these fields will produce a desired synergy for linguistic research.[3]

Hence, I believe that it is essential to ruminate over the skills that a (corpus) linguist A.D. 2023, recently plunged into the reality of AI chatbots, such as ChatGPT or Bard, to name but a few, should possess to facilitate realization of the second scenario (i.e., a useful synergy between corpus linguistics and NLP). To put this matter into perspective, I would like to start with a brief and very subjective overview of the changes in the field of corpus linguistics over the last two decades up to now.

2 CORPUS LINGUISTICS: GROWTH AND CHALLENGES

In general, linguists treat language corpora as a source of research material that they use to empirically verify previously formulated linguistic theories or develop new ones. In fact, every linguist is a bit of a corpus linguist, when using language corpora or elements of corpus methodology in his or her work. Above all, however, this group includes scientists who use previously developed language corpora in their research, compile and study new corpora of texts, and oftentimes – using programming skills – develop new tools for the analysis of the corpora they have collected by themselves. Given that the corpus linguistic methodology (corpus-informed, corpus-based or corpus-driven one) is also used in literary, pedagogical, forensic or sociological research, among others, new challenges arise that pertain to the interdisciplinary character of undertaken studies.

To put it very mildly, for corpus linguists the frequency of occurrence of various language units and their classes comes to the fore, which means that they treat the text as a certain probabilistic structure that should be examined using research methods typical of those fields of science dealing with the analysis of large amounts of empirical data, e.g. statistics, sociology, demographics. Here, it is also worth mentioning data science, which has become, in a sense, a separate discipline devoted to data analysis. It seems, therefore, that in order to understand how other researchers analyzed and interpreted linguistic, sociological and psycholinguistic data, it is necessary to have rudimentary statistical knowledge. Currently, this is no longer an isolated opinion, because for a long time we have been witnessing a shift in linguistics, and perhaps to some extent a change in the dominant research paradigm, towards empirical and quantitative methods. Gries (2013, p. 4) emphasized this very clearly, pointing to the three main goals of using quantitative methods: describing

[3] For example, in a preliminary study Lew (2023, June 12, preprint) shows that the quality of CO-BUILD-style AI-generated (by ChatGPT Plus) lexicographic definitions is comparable to the ones written by human lexicographers.

data, explaining the relationships found in the data,[4] and predicting the future state of affairs (also using other data for this purpose, but based on a model developed on the basis of previously obtained data).

The above observation related to the necessity of being familiar with statistics becomes even more important in the current situation, where more and more published studies provide us with the results of various statistical analyses, enriched with histograms, box plots or heat maps etc. At the same time, somewhat as an answer to this trend, many publications and textbooks devoted to statistics in linguistic research, also with elements of programming in R language (which until recently was the international standard in the field of statistical analysis and data visualization, now increasingly supplanted by Python), came out in print in recent years, e.g. Gries 2013; Cantos Gomez 2013; Levshina 2015; Navarro 2015; Desagulier 2017; Brezina 2018; Winter 2019. The increasingly advanced statistics used in corpus linguistic research has been accompanied by the growing use of machine learning methods popular in the field of NLP, which brings the corpus linguist even closer to the role of a data scientist, as also mentioned by Dunne (2022).

It can be argued that despite the complexity of some quantitative, statistical methods, including those used in the area of natural language processing (e.g. neural networks, word or sentence embeddings, transformers, cf. Jurafsky – Marin 2023), their use in linguistically-oriented should be based on strong humanistic premises. With the aid of these methods, coupled with theoretical linguistic foundations, we are now better equipped to detect and comprehend intricate textual, lexical or grammatical patterns present in huge amounts of texts. Such patterns might have been overlooked when relying solely on intuition or individual, subjective impressions, or even conducting rudimentary corpus analysis like reading concordances. It would have been also impossible to precisely classify the huge amounts of linguistic data (e.g. using support vector machines (SVMs), classification decision trees, random forests or other machine learning methods, both supervised and unsupervised). Without the use of different regression methods (depending on the type of dependent variable), it would have been impossible to verify the hypotheses explaining the occurrence of some linguistic features with a reliable degree of precision. Having gained insights into the data using methods popular in NLP, we can then go back to the text, take a closer look at the obtained linguistic patterns and conduct their in-depth interpretation in the full textual, cultural and social context, which is, after all, the essence of research in the humanities.

[4] For example, multivariate methods allow us to examine the relationships between multiple independent variables (e.g. social ones, such as age, gender, time period, or textual ones, such as text type or genre) and their impact on the frequency and distribution of linguistic features (dependent variable), and to verify the obtained results based on a more advanced statistical apparatus (Cantos Gomez 2013, pp. 89–133).

The latest publications also provide us with a growing body of evidence that currently those linguists who employ quantitative methods when conducting empirical research (both descriptive and experimental one) have become increasingly aware of the limitations of custom-designed tools for text analysis (e.g. range of options and functionalities, user-friendliness of the user interface, required degree of technical expertise).[5] Some linguists, literary scholars or specialists in culture studies – identifying themselves with the trend currently known as digital humanities – increasingly learn to write text-analysis software by themselves (e.g. in the form of developing short scripts in R or Python, enabling the extraction of the desired linguistic material from their collections of text or to conducting appropriate analyses). Alternatively, they closely cooperate with specialists in the field of computational linguistics or NLP. In the near future, conversational artificial intelligence (or generative artificial intelligence, GenAI), which – if properly asked (*prompt engineering*) – may lend them a helping hand when pursuing their research, by facilitating comprehension of the intricacies of programming code or a statistical method. For example, in the paper devoted to data-driven language learning (DDL), Crosthwaite and Baisa (2023, pp. 2–3) list a number of advantages that corpora and custom-designed corpus tools hold over GenAI applications. At the same time, however, they also reflect upon the ways GenAI chatbots may overcome some of the limitations of language corpora and corpus software, notably in aspects such as technical complexity, lack of flexibility, data size, etc.

So, should a corpus linguist of the 2020s, specializing, for example, in phraseology, language pedagogy or sociolinguistics, also be an NLP and machine learning expert? It seems that the answer to this question depends largely on the specific needs (research questions) of the researcher as well as the culture of collaboration between researchers from different disciplines. It seems that when working with authentic empirical data (linguistic, sociological, etc.) it is worth understanding (at least conceptually rather than in technical detail) the methods typical of both computational linguistics and NLP simply in order to assess their usefulness. In other words: with a leading research question in mind, it is worth knowing in the first place what we get thanks to using one method rather than another. For example, it is good to know what the use of a selected statistical measure in relation to word co-occurrence brings us (e.g. a particular measure of lexical attraction or repulsion, statistical significance test, measure of effect size), rather than just know the mathematical formula each and every metric. In practice, the answer to the question: "Would it better if I learned how to program?" may also depend on the possibility of finding computer programmers who will develop (and

[5] Popular corpus software among corpus linguists include, among others, e.g. WordSmith Tools (Scott 2022), AntConc (Anthony 2022) or SketchEngine (Kilgarriff et al. 2014).

maybe even run) an appropriate programming script for us, taking into account the specificity of our linguistic dataset, research questions and hypotheses.

At this point, it is extremely important to encourage a culture of collaboration between researchers in the humanities (including linguists) and specialists from other fields of science, especially those outside the humanities (computer scientists, chemists,[6] biologists). Thanks to the research and development projects carried out in recent years, the situation has changed for much better in this respect. For example, thanks to the European project CLARIN,[7] researchers in the humanities and social sciences from all over Europe and beyond gained access to tailor-made computational tools and solutions designed to handle large textual datasets (including their processing, analysis and visualization). It is also possible to contact the helpdesk and developers of such tools, affiliated in national CLARIN consortia, for specialist advice on the use of statistical and NLP methods, and even to ask for assistance during the data analysis itself. After all, the essence of creating a pan-European research infrastructure is the very usefulness of such infrastructure for researchers.

3 CONCLUSIONS: A CORPUS LINGUIST IN THE 2020S

It is important to highlight that my observations pertain to corpus linguistics, a set of methods that involve analyzing extensive empirical textual data. However, this is not the sole approach to linguistic research. Hence, the arguments presented here may not be equally applicable to rationalistic "armchair" linguistics, where researchers themselves construct language examples to test linguistic theories: this practice continues to endure and is also likely to evolve in the future. Also, this brief reflection-overview paper, focusing on the desired skills of contemporary corpus linguists, reveals that certain questions remain without a definitive answer, and the arguments presented here are intentionally selective.

It is evident that the development of both NLP and language engineering is extremely fast, and today corpus linguistics has become increasingly embedded in statistics, data science and machine learning. Further advances in research on artificial intelligence, especially in the area of application of large language models (LLMs), will soon bring further challenges for researchers. In fact, the volume of data that surround us will grow fast: according to the forecast provided by Statista[8] (2023), it will grow from around 125 ZB (zetabytes) in 2023 to more than 180 ZB in 2025, where 1 ZB equals a billion TB (terabytes). Most probably, a significant proportion of this will be electronic textual data produced naturally or artificially generated. In addition, the exponentially growing number of digital publications

[6] An interesting example of such collaboration can be found in Woźniak et al. (2017).

[7] https://www.clarin.eu/

[8] https://www.statista.com/statistics/871513/worldwide-data-created/ (NCES. (n.d.))

means that better and better (more accurate, mathematically-sophisticated etc.) research methods may tip the scales in the race for getting published and becoming visible, even for a brief moment, in the reality of attention economy that has permeated academic publishing, to use the phrase taken from Hyland (2023).

Thus, it can be concluded that a corpus linguist working with linguistic data in the 2020s, be it a stylometrist, sociolinguist or otherwise, should have sufficient skills to understand, at least conceptually, descriptive and inferential statistics as well as the methods used in computational linguistics and NLP. This will allow him or her to better understand the specificity of linguistic data (e.g. recognize the type of variables used in the study, understand what vector text representation is etc.), compare performance of different methods (e.g. in relation to topic modelling or text classification), interpret different visualizations of linguistic data, and maybe even estimate the quality and complexity of a statistical model (e.g. using metrics such as accuracy, precision, recall, F-score, mean error rate, AIC or BIC etc.). At the same time, the knowledge of language theories will still play a particularly important role, as it enables one to interpret the complex patterns and regularities that exist in naturally occurring language. Ultimately, the goal of any corpus analysis is to obtain new knowledge of language, be it a language-system or language in use. For these reasons, a corpus linguist should remain a linguist in the first place. On the technical front, it can be expected that the work of corpus linguists will to a greater degree overlap with the tasks performed by a computer programmer, statistician or data scientist.[9] Hence, at least basic programming skills may come in useful, especially in terms of running programming scripts (e.g. written in Python) either locally or using collaborative programming platforms, such as Jupyter Notebook or Google Collab. It might be also useful to know how to extract linguistic data from various resources or infrastructures (lexical databases, language corpora, etc.) that offer access to researchers via API (Application Programming Interface). The informed use of low-code programming platforms or conversational artificial intelligence (and other AI-assisted solutions) is likely to offer assistance as well. As mentioned earlier, collaboration skills will play a vital role in advancing interdisciplinary research at the intersection of (corpus) linguistics, statistics, and NLP. Now that the period of AI-assisted research has already started, it is essential for corpus linguists to navigate through it with an open-mind and caution.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan. Waseda University. Accessible at: https://www.laurenceanthony.net/software.

[9] At this point, I do not consider whether artificial intelligence will replace corpus linguists altogether, as this is a completely separate issue to me. For the same reason, I do not consider limitations in development of NLP AI-research (consumption of energy, funding etc.).

Brezina, V. (2018). Statistics for Corpus Linguistics. Cambridge: Cambridge University Press, 314 p.

Cantos Gomez, P. (2013). Statistical Methods in Language and Linguistic Research. London: Equinox, 256 p.

Crosthwaite, P., and Baisa, V. (2023). Generative AI and the end of corpus-assisted data--driven learning? Not so fast! Applied Corpus Linguistics, 3(3). Accessible at: https://doi.org/10.1016/j.acorp.2023.100066.

Desagulier, G. (2017). Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics. Berlin: Springer, 366 p.

Dunne, J. (2022). Natural Language Processing for Corpus Linguistics (Elements in Corpus Linguistics). Cambridge: Cambridge University Press, 96 p.

Gries, S. (2013). Statistics for Linguistics with R. Berlin: De Gruyter, 374 p.

Hirschberg, J., and Manning, Ch. (2015). Advances in natural language processing. Science, 349(6245), pages 261–266.

Hyland, K. (2023). Academic publishing and the attention economy. Journal of English for Academic Purposes, 64. Accessible at: https://doi.org/10.1016/j.jeap.2023.101253.

Jurafsky, D., and Martin, J. (2023). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. (Third edition e-book: draft of January 7, 2023). Accessible at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. (accessed on 19 July 2023).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubícek, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), pages 7–36.

Levshina, N. (2015). How to do Linguistics with R: Data exploration and statistical analysis. Amsterdam: John Benjamins, 454 p.

Lew, R. (2023, June 12). ChatGPT as a COBUILD lexicographer. Accessible at: https://doi.org/10.31219/osf.io/t9mbu.

McEnery, T., and Wilson, A. (1996). Corpus Linguistics. Edinburgh: University Press, 256 p.

Navarro, D. (2015). Learning Statistics with R: A tutorial for psychology students and other beginners. (Version 0.6), 599 p. Sydney. University of New South Wales. Accessible at: http://compcogscisydney.org/learning-statistics-with-r/.

NCES. (n.d.). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes). In Statista - The Statistics Portal. Accessible at: https://www.statista.com/statistics/871513/worldwide-data-created/.

Ooi, V. (1998). Computer Corpus Lexicography. Edinburgh: University Press, 224 p.

Scott, M. (2022). WordSmith Tools version 8 (64 bit version) Stroud: Lexical Analysis Software.

Winter, B. (2019). Statistics for Linguists: An Introduction Using R. London: Routledge, 310 p.

Woźniak, M., Wołos, A., Modrzyk, U., Górski, R. L., Winkowski, J., Bajczyk, M., Szymkuć, S., Grzybowski, B., and Eder, M. (2018). Linguistic measures of chemical diversity and the ‘keywords’ of molecular collections. Scientific Reports, 8(1), page 7598.

sciendo
DOI 10.2478/jazcas-2023-0024

CORROBORATING CORPUS DATA WITH ELICITED INTROSPECTION DATA: A CASE STUDY

JAKOB HORSCH
Department of English Language and Linguistics, Faculty of Languages and Literatures, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany

HORSCH, Jakob: Corroborating Corpus Data with Elicited Introspection Data: A Case Study. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 60 – 69.

Abstract: The last decades have seen an exponential growth of corpus sizes. This development has been driven by a desire to investigate rare syntactic phenomena, but issues remain: Corpora are by definition finite samples, but language is by definition infinite, leading to the negative data problem ('absence of evidence is not evidence of absence'). One solution is corroborating corpus data with elicited introspection data that is obtained in a reliable, valid, and objective way. I present a case study to show how this can be done using the Magnitude Estimation Test (MET) method (Hoffmann 2013). Analyzing elicited data from 37 L1 English speakers, I show that introspective data can complement corpus data and lead to interesting new findings.

Keywords: negative data problem, MET, introspective data, comparative correlative, *that*-complementizer

1 INTRODUCTION

The rapid increase of computing power and storage capacity has precipitated an exponential growth of electronic corpora, which have become an indispensable data source in linguistics. This trend is visualized in Fig. 1 (adapted from Anthony 2013, p. 145), which plots the sizes of ten English corpora over the last decades. It is a development that has been motivated by interest in infrequent linguistic phenomena. Examples include the English *way*-construction, for which Brunner and Hoffmann (2020) determined a frequency per million words (pmw) of under 10, or the Comparative Correlative construction, for which Hoffmann et al. indicate a pmw frequency of 30–40 (Hoffmann et al. 2019, p. 32).

Nevertheless, even the largest corpora have an inherent flaw: As language *samples* they are by definition *finite*. However, the object of investigation is by definition *infinite*: As noted by Chomsky, "an essential property of language is that it provides the means for *expressing indefinitely many thoughts* and for reacting appropriately in an *indefinite range of new situations*" (Chomsky 1965, p. 6) (emphasis JH).

This flaw is referred to as the 'negative data problem': "just because a phenomenon cannot be found in a corpus, it cannot be concluded that it is

ungrammatical" (Hoffmann 2011, p. 1). In short, absence of evidence is not evidence of absence. Moreover, there is the 'positive data problem': "just because a construction appears in a corpus it does not automatically follow that it is grammatical" (Hoffmann 2011, p. 1). In short, a corpus can never be fully representative of a language.

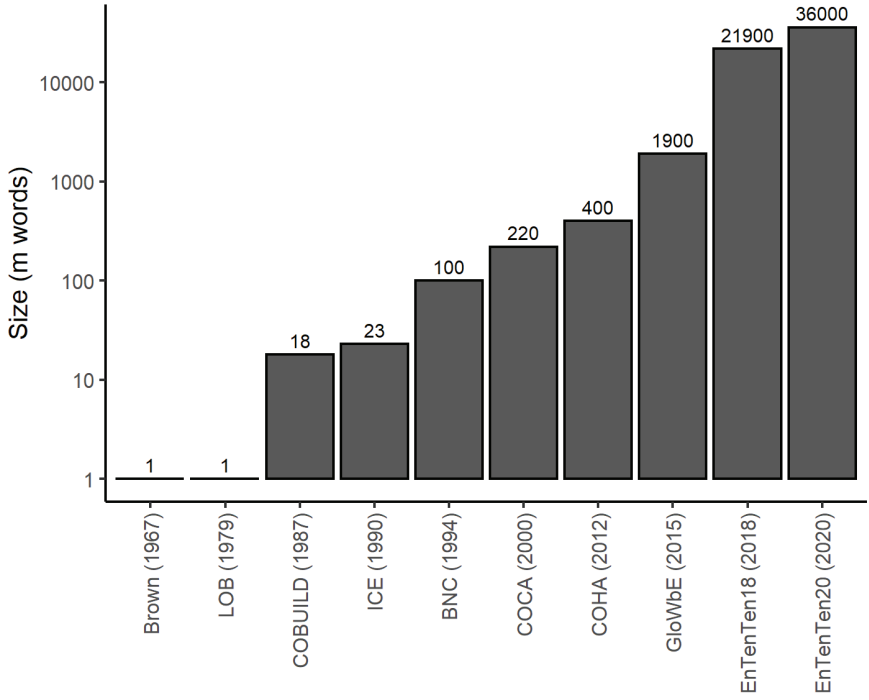

Fig. 1. Growth in size of English corpora over the last decades (logarithmic scale)

Obviously, this is an issue that cannot be resolved by using ever larger corpora, since they will always be finite. One solution is corroborating corpus studies with introspective judgments. Both data types, corpora and introspection, have been deemed "valid, reliable and objective" (Hoffmann 2011, p. 13). However, while corpus studies abound, there is a striking lack of studies based on elicited data.

Of course, elicited data also has disadvantages: Bresnan points out that "constructed sentences […] are often highly artificial [and] isolated from connected discourse" (2007, p. 91); Hoffmann mentions the "unnatural setting and […] unnatural types of stimuli and responses" (2019, p. 16). Nevertheless, it can be useful: Hoffmann notes that "several studies have […] shown that the results from

experimental and corpus-based studies often converge" (2019, p. 18). Apart from Hoffmann's study on preposition placement (2011), a good example that proves this is Bresnan and Ford's study, in which participants "responded reliably to corpus model probabilities" (2010, p. 168).

In the following, I will show how corpus study results can be corroborated using the so-called Magnitude Estimation Test (MET) method (Bard et al. 1996; Cowart 1997, pp. 73–84; Hoffmann 2011, 2013). I used METs to test optional *that*-complementizers in the English Comparative Correlative (CC) construction. The CC consists of two subclauses, C1 and C2; *that*-complementizers are optional in C1 (1) and have been claimed to be possible in C2 "in present-day colloquial registers" (den Dikken 2005, p. 402; cf. also Culicover – Jackendoff 1999, p. 546; Hoffmann 2019, p. 47 (2):

(1) [*The more* **[*that*]**$_{\text{optional THAT-complementizer}}$ *he says*,]$_{C1}$ [*the less I wanna say*.]$_{C2}$
(2) [*the larger the settlement becomes*]$_{C1}$
[*the less* **[*that*]**$_{\text{optional THAT-complementizer (?)}}$ *the reduced number of sites you will have available.*]$_{C2}$

That-complementizers play a role in debates about the syntactic relationship between C1 and C2. Proponents of a hypotactic analysis (e.g. den Dikken 2003, 2005; Borsley 2004) use *that*-complementizers as proof of the subordinate status of C1. Those advocating a paratactic analysis have pointed out that *that*-complementizers are optional and very infrequent in corpus data: In his 1,409-token data set from COCA, Hoffmann (2019) found just 24 *that*-complementizers, and in their 2,041-token BNC data set, Hoffmann et al. (2020) found just 29. In C2, *that*-complementizers are even less frequent: In his BROWN corpus study, Hoffmann could not find any *that*-complementizers in C2 and in his COCA data, he found just six, out of a total of 1,409 CC tokens (2019). Hoffmann et al.'s BNC corpus study yielded just two cases out of 2,041 CC tokens (2020). Accordingly, Hoffmann et al. concluded that *that*-complementizers are "no longer central properties" of the Present-day English (PdE) CC construction, but rather "historical remnants" (2020, p. 200).

However, because *that*-complementizers are so infrequent, they are subject to the negative data and positive data problems. Therefore, Hoffmann has explicitly called for corroborating his corpus study findings with elicited data (Hoffmann 2019, p. 10). Accordingly, I decided to employ the MET method to collect ratings from L1 English speakers. In Section 2 I will describe the MET method and how the data was collected and analyzed. In Section 3 I present the results of the study. In Section 4 I offer a conclusion, arguing that corroborating corpus data with elicited introspective data is not only viable but extremely important in a discipline that has come to rely so heavily on the former.

2 DATA AND METHODOLOGY

2.1 The MET method

In the MET method, participants "judge stimuli relatively to a reference item" (Hoffmann 2013, p. 99), taking advantage of the fact that humans are better at making relative judgments than absolute judgments (Hoffmann 2013, p. 99). The ratings thus obtained translate to "proportional relation among the numbers assigned to different stimuli", which "should reflect the proportions of the stimuli themselves; thus a sentence that is judged twice as good as another should get a number twice as high as that assigned to the other sentence" (Cowart 1997, pp. 73–74).

MET questionnaires present a set of sentences (test items) to participants, who rate them with a number. This number is "proportional to a constant reference sentence" (Hoffmann 2013, p. 99) that has previously been assigned a number by the test subjects. In other words, "subjects do not have to rate stimuli on a scale […] which might artificially limit their choices" but rather "decide on their own scale and make as many fine-grained choices as they deem necessary" (Hoffmann 2013, p. 103). METs also feature grammatical and ungrammatical fillers, which provide baselines against which test items are compared. This makes it possible to assess test items not just in relation to each other: Participants will give low-frequent grammatical phenomena worse ratings than high-frequent grammatical phenomena, but low-frequent grammatical phenomena will still be rated better than ungrammatical fillers; ungrammatical phenomena are expected to receive ratings closer to the ungrammatical fillers than the grammatical ones. Based on corpus studies (see above), test items with no overt *that*-complementizers (Ø–Ø) should be rated best, followed by those with a *that*-complementizer in C1 (*that*–Ø), followed by those with a *that*-complementizer in C2 (*that–that* and Ø–*that*).

The aim of the METs was to present all four of these factor combinations (conditions): NO THAT_{C1}-NO THAT_{C2}, THAT_{C1}-NO THAT_{C2}, THAT_{C1}-THAT_{C2}, NO THAT_{C1}-THAT_{C2}. However, different lexical material (lexicalizations) was used each time, cf. Cowart (1997, pp. 49–50). Thus, a participant "is never able to confront a sentence in quite the same way twice" (Cowart 1997, p. 50). Because there were only four conditions, it was decided to create two lexicalizations for each one. These eight lexicalizations were then used to create all four conditions twice, resulting in 32 tokens, which were split into four material sets of eight tokens each following the Latin squares method (Hoffmann 2011, p. 29; Keller – Alexopoulou 2005, p. 1121). 16 fillers were created, following Cowart's suggestion that the ratio of fillers:test items be at least 2:1 (1997, p. 92).

The fillers were conditional sentences based on four different patterns. For each of these patterns, two grammatical and two ungrammatical lexicalizations were created, the latter featuring erroneous subject-verb agreement (3rd ps. sg. -*s* omission). The resulting set of 24 tokens (eight test items plus 16 fillers) was then randomized. Thus four material sets, each consisting of 24 tokens, were created. The questionnaires also included two training sessions to familiarize participants with the

MET method. To obtain valid, objective, reliable and comparable data, "carefully constructed experimental settings" (Hoffmann 2013, p. 99) had to be ensured: The METs were conducted on-site by the author and his PhD supervisor at the University of Texas at Austin. The 37 participants (17 f, 20 m) were university students that were recruited with the help of university staff and by word of mouth. Ten participants filled out material set 1, and material sets 2, 3, and 4 were each filled out by nine participants, for a total of 37.

2.2 Data analysis

As outlined above, the MET method involves participants generating their own scales. To make the results comparable, *z*-scores were calculated by "subtracting the mean of a variable from an individual value and dividing the result by the variable's standard deviation" (Hoffmann 2013, p. 103). The data were then analyzed using mixed-effects models, which allow the "simultaneous modeling different sources of variability" (Gries 2009, p. 333). As Gries notes, this is much more precise than "simply fitting one regression line over many subjects" (2009, p. 333).

Specifically, I used a variant of stepwise regression that is known as backward selection. This procedure, which involves starting with a full model from which non-significant effects are removed, was carried out by means of the `step()` function from the *lmerTest* package for *R* (Kuznetsova et al. 2020; Kuznetsova et al. 2015). Tab. 1 summarizes the fixed and random effects that were tested:

Fixed effects	**Random effects**
THAT_C1	SUBJECT
THAT_C2	MATERIAL SET
	GENDER
	LEXICALIZATION

Tab. 1. Fixed effects and random effects

3 RESULTS

In total, 37 questionnaires were filled out. Subsequently a full mixed-effects model[1] was subjected to stepwise regression. No random effects were significant, which indicates that the experiment design and implementation were successful in minimizing the influence of sources of variability other than the investigated syntactic variables. However, the fixed effect THAT_C2 (p<0.001***) as well as the interaction THAT_C1:THAT_C2 (p=0.0113*) turned out to be significant. While THAT_C1 (p = 0.284) was not significant, it was retained in the model as part of the significant interaction THAT_C1:THAT_C2.

[1] A model "where the fixed and random parts contain all explanatory variables and as many interactions as possible" (Kuznetsova et al. 2015, p. 33).

In summary, C1 *that*-complementizers did not have a significant influence on the ratings (0.284 n.s.), whereas C2 *that*-complementizers did (p<0.001***). The same is true of the interaction THAT_C1:THAT_C2, although less so (p=0.011*). In the following, I will discuss these results using graphs that plot each condition's *z*-score means as dots with error bars that indicate standard errors (SE). The graphs also show the grammatical[2] (*z*-score mean: 0.428) and ungrammatical (*z*-score mean: −0.930) filler means as horizontal lines, with standard errors (0.050 and 0.033, respectively) indicated by dotted lines. Standard deviations (SD) are not indicated in the graphs. As I will show, the results do not only confirm findings from previous corpus studies, but also uncover interesting new aspects. First, consider Fig. 2, which plots the results for THAT_C1 (left) and THAT_C2 (right). The left plot shows that the *z*-score mean of test items without *that*-complementizers in C1 (Ø) (0.598; SD: 0.770, SE: 0.063) was only slightly higher than that of test items with *that*-complementizers in C1 (*that*) (0.513; SD: 0.682, SE: 0.056). This confirms the results of stepwise regression (see above), which showed THAT_C1 as not significant. Note that both values are well within each other's standard error range, and that neither condition was rated worse than the grammatical fillers.

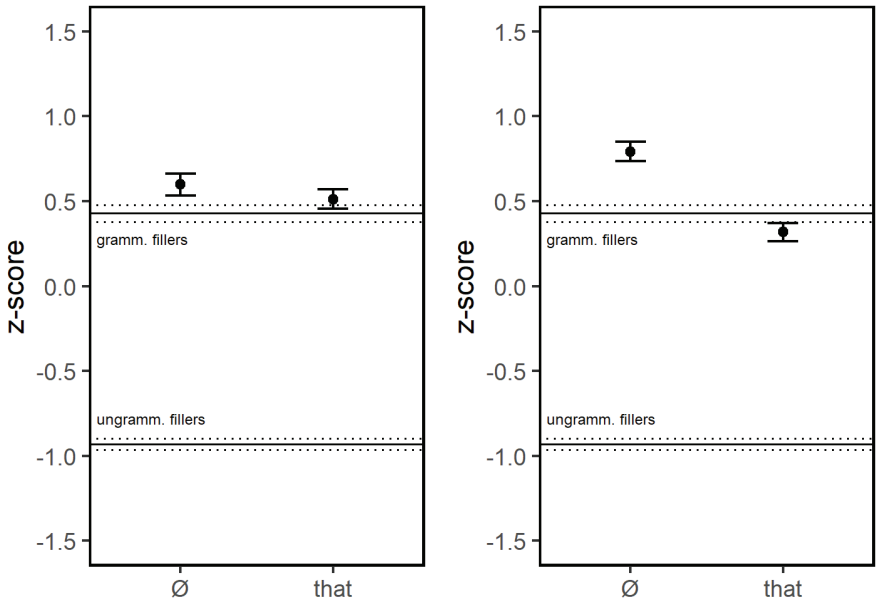

Fig. 2. *Z*-score means of THAT and NO THAT (Ø) in C1 (left) and C2 (right) (n=37)

[2] One grammatical filler was removed from the dataset because it was rated considerably worse than the other grammatical fillers. It appears that this was due to this particular filler being noticeably shorter than the others.

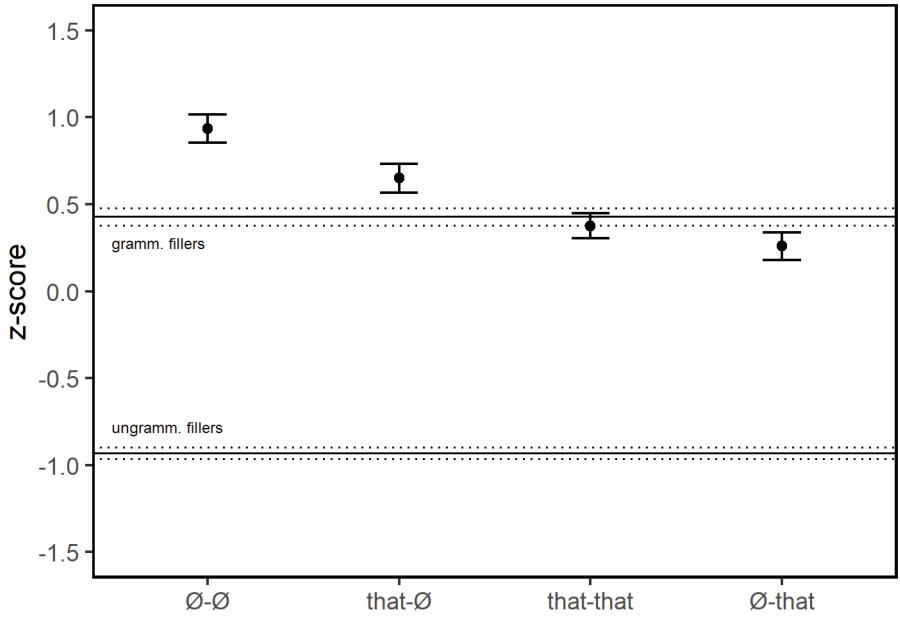

Fig. 3. *Z*-score means of the interaction THAT_C1:THAT_C2 (n=37)

This indicates that the presence or absence of a *that*-complementizer in C1 did not significantly influence the participants' ratings, confirming its status of an optional yet grammatical feature. The right plot in Fig. 2 paints a different picture, again reflecting the findings of stepwise regression: Recall that THAT_C2 was highly significant (p<0.001***). Thus, there is a pronounced difference between *that*-complementizers in C2 (*that*) (0.317; SD: 0.659, SE: 0.054) and no *that*-complementizers in C2 (Ø) (0.793; SD: 0.659, SE: 0.054). Also, *that*-complementizers in C2 were rated worse than the grammatical fillers. This again confirms claims from corpus studies: *That*-complementizers in C2 are indeed perceived as considerably less acceptable than in C1, confirming its lower frequency in corpora and claims about its restriction to colloquial use. Nevertheless, it is noteworthy that C2 *that*-complementizers were still rated considerably better than the ungrammatical fillers. This has serious implications for claims about the status of C2 as main clause: It is difficult to maintain that *that*-complementizers, which are purportedly markers of subordination, are ungrammatical in C2. This in turn challenges a hypotactic analysis of the overall construction.

Finally, Fig. 3 plots the interaction THAT_C1:THAT_C2 that stepwise regression also showed to be significant (although less so, with a p-value of 0.011*). It turns out that these results also correspond to the expectations based on corpus studies: Items

without *that*-complementizers (Ø–Ø) were clearly rated best (*z*-score: 0.937; SD: 0.697; SE: 0.081). *That*-complementizers in C1 (*that*–Ø) were rated second-best and better than the grammatical fillers (*z*-score: 0.650; SD: 0.711; SE: 0.083). This confirms corpus studies that found *that*-complementizers to be extremely rare, concluding that they are "no longer central properties" of the PdE CC construction (Hoffmann et al. 2020, p. 200) but nevertheless not ungrammatical.

There are also some interesting observations to be made about C2 *that*-complementizers. Due to their extremely low frequency in corpora, previous studies were not able to provide persuasive evidence for their acceptability in combination with *that*-complementizers in C1. The MET results show that both conditions, *that*–*that* (*z*-score: 0.647; SD: 0.716; SE: 0.082) and Ø–*that* (*z*-score: 0.647; SD: 0.716; SE: 0.082) were rated considerably worse than Ø–Ø and *that*–Ø. Again, this confirms previous corpus studies that found fewer C2 *that*-complementizers than their C1 counterparts. However, note that while both conditions were rated slightly worse than the grammatical fillers, their *z*-score means are still considerably higher than the ungrammatical fillers. This implies that *that*-complementizers are not ungrammatical and constitutes further evidence against a hypotactic analysis of the English CC construction (see discussion above).

4 CONCLUSION

As I have shown, introspective data that is elicited using the MET method is useful for corroborating corpus studies and thus addressing the 'negative data problem'. In fact, the results presented here do not only confirm findings from previous corpus studies. They also uncovered hitherto unknown aspects of the CC construction: Despite multiple corpus investigations, some of which were based on datasets of several thousand CC tokens (Hoffmann 2019; Hoffmann et al. 2020), it has been impossible to make conclusive claims about the grammaticality of *that*-complementizers in C2, simply for lack of evidence. Using the MET method, it was possible to prove that *that*-complementizers in C2, while rated worse than their counterparts in C1, must nevertheless be considered grammatical, since they were rated much closer to the grammatical filler mean than the ungrammatical filler mean. This, in turn, has implications for hypotactic analyses of the English CC construction: If *that*-complementizers were indeed markers of subordination, they should have been clearly rated as ungrammatical in C2 (i.e., closer to the ungrammatical filler mean), which under a hypotactic analysis functions as main clause.

Thus, the case study demonstrates that empirically obtained introspective data can contribute to clarifying unresolved disputes in the literature, where introspective judgments are usually provided only by individuals. This underscores the importance of the empirical approach in general; as Bresnan notes, "[m]ismatches between grammaticality judgments reported by linguists and the actual language use of

speakers and writers are surprisingly common, particularly in areas of theoretical syntax […] where subtle contrasts are invoked" (2007, p. 84). It is therefore crucial to not only corroborate corpus studies with elicited introspective data, but also test claims that are based solely on the introspection of individuals. This is true despite certain disadvantages that eliciting grammaticality acceptability judgments come with, including the amount of resources required and the fact that what is judged is constructed/artificial and disconnected from discourse (Bresnan 2007, p. 91). All things considered, however, I believe that the advantages outweigh the disadvantages, especially when complemented with corpus data.

Note: The data used in this article are available for download from an OSF repository (https://osf.io/ca9k6/).

ACKNOWLEDGEMENTS

I am grateful to the German Research Foundation for its generous financial support in the context of the project "Comparing English Comparative Correlatives: The more data, the better" (DFG HO 3904/5-1), including funding of a part-time position and covering expenses associated with the METs.

References

Anthony, L. (2013). A critical look at software tools in corpus linguistics. Linguistic Research, 30(2), pages 141–161.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. Language, 72(1), pages 32–68.

Bates, D. (2005). Fitting Linear Mixed Models in R Using the lme4 package. R News, 5(1), pages 27–30.

Borsley, R. D. (2004). An Approach to English Comparative Correlatives. In S. Müller (ed.): Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar, Center for Computational Linguistics, Katholieke Universiteit Leuven, pages 70–92. Stanford, CA: CSLI Publications.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston – W. Sternefeld (eds.): Roots: linguistics in search of its evidential base (Studies in Generative Grammar 96), pages 75–96. Berlin: De Gruyter.

Bresnan, J., and Ford, M. (2010). Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. Language, 86(1), pages 168–213.

Brunner, T., and Hoffmann, T. (2020). The way-construction in World Englishes. English World-Wide, 41(1), pages 1–32.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Cowart, W. (1997). Experimental Syntax: Applying Objective Methods to Sentence Judgements. Thousand Oaks, CA: Sage.

Culicover, P. W., and Jackendoff, R. (1999). The View from the Periphery: The English Comparative Correlative. Linguistic Inquiry, 30(4), pages 543–571.

Dikken, M. den (2003). Comparative correlatives and verb second. Manuscript. CUNY Graduate Centre, ms.

Dikken, M. den (2005). Comparative Correlatives Comparatively. Linguistic Inquiry, 36(4), pages 497–532.

Gries, S. Th. (2009). Statistics for Linguistics with R: A Practical Introduction (Trends in Linguistics: Studies and Monographs: 208). 1st ed. Berlin: De Gruyter Mouton.

Hoffmann, T. (2011). Preposition Placement in English: A Usage-based Approach (Studies in English Language). Cambridge: Cambridge UP.

Hoffmann, T. (2013). Obtaining introspective acceptability judgements. In M. Krug – J. Schlüter (eds.): Research Methods in Language Variation and Change, pages 99–118. Cambridge: Cambridge UP.

Hoffmann, T. (2019). English Comparative Correlatives: Diachronic and Synchronic Variation at the Lexicon-Syntax Interface (Studies in English Language). Cambridge: Cambridge UP.

Hoffmann, T., Brunner, T., and Horsch, J. (2020). English Comparative Correlative Constructions: A Usage-based account. Open Linguistics, 6(1), pages 196–215.

Hoffmann, T., Horsch, J., and Brunner, T. (2019). The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. Cognitive, Linguistics, 30(1), pages 1–36.

Keller, F., and Alexopoulou, T. (2005). A crosslinguistic, experimental study of resumptive pronouns and that-trace effects. In B. G. Bara – L. Barsalou – M. Bucciarelli (eds.): Proceedings of the 27th Annual Conference of the Cognitive Science Society, pages 1120–1125.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13), pages 1–26.

Kuznetsova, A, Brockhoff, P. B. and Christensen, R. H. B. (2020). lmerTest: Tests for random and fixed effects for linear mixed effects models (lmer objects of lme4 package). R package version 3.1-3. Accessible at: https://CRAN.R-project.org/package=lmerTest.

Kuznetsova, A., Christensen, R. H. B., Bavay, C., and Brockhoff P. B. (2015). Automated mixed ANOVA modeling of sensory and consumer data. Food Quality and Preference 40, pages 31–38.

Szmrecsányi, B., Grafmiller, J., Heller, B., and Röthlisberger, M. (2016). Around the world in three alternations: Modelling syntactic variation in varieties of English. English World-Wide, 37, pages 109–137.

sciendo
DOI 10.2478/jazcas-2023-0025

DATIVE AMBIGUITY IN RUSSIAN: A CORPUS INDUCED STUDY

EDYTA JURKIEWICZ-ROHRBACHER[1,2]
[1] Institute of Slavic Studies, University of Regensburg, Regensburg, Germany
[2] Institute of Slavic Studies, University of Hamburg, Hamburg, Germany

JURKIEWICZ-ROHRBACHER, Edyta: Dative Ambiguity in Russian: A Corpus Induced Study. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 70 – 80.

Abstract: When describing the Russian dative case, an observation often made in passing is that its assignment to certain types of arguments is ambiguous, particularly in constructions with a predicative infinitive. Thus far, no studies have put this problem into focus nor described the range of structures to which it applies. I approach this problem with corpus-driven methods. The present study shows that the predicate order and the referential prominence hierarchy can be used as explanatory variables in the modelling of the semantic-syntactic role assignment.

Keywords: dative, Russian, syntactic-semantic role assignment, ambiguity avoidance strategies

1 INTRODUCTION

1.1 Dative attachment ambiguity in Russian

In this article, I discuss the problem of ambiguity for arguments encoded in Russian with the dative case. As observed by Padučeva (2017), for instance, the syntactic and semantic role of dative argument is sometimes ambivalent, as shown in (1).

(1)	*A*	*na*	*drugoj*	*den'*	*vezti.INF*	*produkty*	*bylo*	*uže*
	and	on	second	day	transport	products	AUX	already
	nekomu...							
	nobody.DAT							

Reading 1 'And the next day, there was nobody who could transport the food.'
Reading 2 'And the next day, there was nobody to whom the food could be delivered.'
[A. Marinina. Angely na l'du ne vyživajut 2014; (after Padučeva 2017)]

The lack of congruence between the subject and the predicate causes difficulties in the identification of the potential subject. Additionally, the verb *vezti* 'to transport' encodes an indirect object with a semantic role receiver, overtly marked with the dative. Depending on whether the negative dative pronoun *nekomu* is assigned the

first or second argument role, the sentence receives Reading 1 or Reading 2, respectively. The status of the dative argument in such structures has been often discussed in the literature, but no general consent has been reached so far (see Hansen 2020 for a recent review of the topic). Similarly to Grillborzer (2019), I distinguish only between the first dative argument (D1), which is higher in the syntactic structure, and the second dative argument (D2), which is lower in the syntactic structure in relation to D1. This choice has two motivations. First, the work is theory-neutral in character, and second, it also involves object control structures where D1 is clearly not the subject of the matrix (complement taking) predicate.
As stated by Padučeva (2017), the correct interpretation of dative arguments is often driven by pragmatic factors, that is, it can be obtained from the context, as shown in (2).

(2)	*Mne*	*zvonit*	*nekomu*	*- ja*	*i*	*ne*	*slušaju.*
	me. DAT	call. INF	nobody. DAT	I	FOC	NEG	listen.1SG

Reading 1 'Nobody should call me – I'm not listening.'
Reading 2* 'I have nobody to call – I'm not listening.'
[I. Grekova. Letom v gorode 1962; after Padučeva 2017]

Example (2) is more complex than (1), as it involves two dative arguments, and the role assignment is solved through the context. Maurice (1996) and Weiss (1992) admit that although such sentences are ambiguous with respect to dative argument references, the linear first dative should be assigned to the argument that is higher in syntax, i.e., the first argument, while the other dative argument to the lower one. Bonč-Osmolovskaja (2003, p. 25) suggests that the reversed interpretations in the spoken discourse would be marked prosodically.

1.2 Research motivation

In contrast to ambiguity caused by the morphological poverty in subject/direct object marking in Slavic – best visible in Bulgarian and Macedonian – the theoretically possible ambiguity related to the co-occurrence of two dative arguments in Russian (and possibly other Eastern Slavic languages) has never been addressed. Most works mention it in passing only. Moreover, Maurice (1996) and Weiss (1992) suggest that such ambiguous structures should generally be avoided in language use. Contrary to these normativistic statements, Arnold et al. (2004) provide evidence in a series of experiments on dative constituent ordering in English, whereby language users do not systematically apply constituent ordering as a strategy to avoid ambiguity.

With the increasing role of artificial intelligence and natural language processing, the problem of ambiguity in language use seems relevant, as it poses a challenge, e.g. with syntactic taggers or in neural machine translation (NMT, c.f.

Bawden 2018). Against the widespread assumption that state-of-the-art NMT systems translate texts contextually, I assume that commercial NMT systems have difficulties in solving the dative assignment task appropriately, as it is exemplified in (1T) and (2T), the English translations of (1) and (2).

(1T) DeepL:[1] The next day, there was no one else to carry the food. (28.02.2023 13:53)
Yandex Translate:[2] And the next day there was no one to carry the groceries. (28.02.2023 13:53)
Google Translate:[3] And the next day there was no one to carry food. (28.02.2023 13:53)
ChatGPT: [4] The next day there was no one left to deliver the groceries. (01.03.2023 16:45)

(2T) DeepL: I don't have anyone to call – I don't listen.[5] (28.02.2023 13:57)
Yandex Translate: I have no one to call — I'm not listening. (28.02.2023 13:58)
Google Translate: I have no one to call – I'm not listening. (28.02.2023 13:59)
ChatGPT: I have no one to call me – I am not listening. (01.03.2023 16:43)

In (1T), all the NMT systems interpret the dative argument as an agent (and a syntactically higher argument) and do not recognize the alternative interpretation, although DeepL offers alternating translations in cases of ambiguity. In (2T) the contextual information that the referent marked with 1.SG is not listening, hence, cannot hear if somebody would be calling is given. This, according to Padučeva (2017), is sufficient for the addressee of the message to assign the highest argument position in syntax to the linearly second argument *nekomu*.

Assuming that the task is not performed contextually in (2T), the following principles could explain the dative phrase attachment in an NMT task.

PRINCIPLE 1: The linear first dative should be assigned to the argument which is higher in syntax (Weiss 1993, Maurice 1996).
PRINCIPLE 2: Following (Haspelmath 2021, p. 127), the assignment is performed according to the *referential prominence* of arguments, in this case according to the *person scale* where the assignment of the higher argument/role is preferred to the *locuphoric* pronouns (first/second person) over the *aliophoric* (third person) pronouns.
PRINCIPLE 3: The semantic-syntactic role attachment depends on the linear order of the predicates. That is, the linear organization of dative arguments reflects the linear organization of their governors.

1.3 Research aims

Due to the general lack of data on dative phrase attachment ambiguity in Russian, this pilot study aims to examine the possible range of structures where two datives do not conform to Principles 1–3. The extent to which the topic could

[1] https://www.deepl.com/translator

[2] https://translate.yandex.com/

[3] https://translate.google.com/

[4] https://chat.openai.com/chat with a prompt 'Translate into English:'.

[5] As alternatives: 'I have no one to call, so I don't listen.' and 'I have no one to call – I don't listen.'

be relevant for further studies from the NMT perspective is also assessed. Further, the study analyses which of the principles plays a statistically significant role in the data.

Obtaining and examining an appropriate sample of sentences is methodologically challenging. First, Russian has a free word order so the arguments in question do not have a fixed position in the sentence. Second, some prepositional phrases govern the dative in Russian. Third, adjacent nouns in the dative may form together a constituent, e.g. *Dedu Morozu* 'to/for Father Frost'. Finally, the context should enable the disambiguation. Therefore, I decided to limit the search to personal pronouns in adjacent positions and manually process the obtained sample.

Personal pronouns are an especially interesting object of study with regard to NMT, as they are an ideal setting for studies on the role of context in MT (Müller et al. 2018). Moreover, I expect that when two dative pronouns are placed adjacent the governing relations might be unclear, leading to more ambiguity, which cannot be solved without context.

The remaining part of this paper is structured such that Section 2 presents the research data, as well as the source and method of obtaining them from corpora. In Section 3 I summarize the results. Section 4 is devoted to the discussion of the results and further prospects for research.

2 RESEARCH DESIGN AND DATA

2.1 Research design and data sources

The data comprises sentences from the Russian Timestamped JSI web corpus 2014–2021 (Trampuš – Novak 2012) of 5,788,590,952 tokens in size. I used the Sketch Engine[6] corpus manager using CQL for two adjacent pronominal lower case word form attributes for data retrieval.[7] The 960 well-formed sentences containing two dative forms were locally stored and a native speaker annotated them for the additional linguistic features to allow evaluation of the impact of Principles 1–3. Therefore, I distinguish between the pronoun order (Principle 1), type of relationship between the pronouns with regard to prominence (Principle 2), and the predicate order[8] (Principle 3, see Tab. 1 for a detailed pattern).

[6] https://www.sketchengine.eu/

[7] The exact query form was: [lc="мне|тебе|ему|ей|нам|вам|им"][lc="мне|тебе|ему|ей|нам|вам|им"] and the idiosyncratic forms of the instrumental and locative were erased in post processing.

[8] In the query, I did not control for the syntactic type of predicate (i.e. subject control such as *udatʹsja* 'to manage' and object control, such as *pozvolit*' 'to allow' or similar), so I do not go into more detail as to the word order, because no default neutral word order can be assumed for the obtained heterogenous group of predicates.

Feature	Levels
Dative pronoun order	D1D2 D2D1
Referential prominence hierarchy	N(o) – hierarchy is violated Y(es) – hierarchy is retained E(qual)1 – two locuphoric pronouns E(qual)2 – two aliophoric pronouns
Predicate order	MC – matrix complement order CM – complement matrix order C – only the infinitive predicate

Tab. 1. Features of annotation

3 RESULTS

3.1 Variable distribution in the data set

The data set indicates that none of the principles formulated in Section 1.2 hold in 100%. It contains 893 D1D2 orders of datives and 67 occurrences of D2D1, which clearly violates Principle 1. The distribution across word orders shows interesting properties (see Tab. 2). Most observations (n=658) originate from clauses with single overt predicate, the infinitive. In this group only two observations (0.3%) do not conform to Principle 1, and the reverse order of datives is chosen, as in (4).[9]

(4) – *Delo vse v tom, čto vysokij rezul'tat vsegda podrazumevaet vysokuju stepen' samootdači* – ***ne vam***$_{D2}$ ***mne***$_{D1}$ *ob ėtom* ***rasskazyvat'***$_{C}$.
'The thing is, achieving a high result always implies a high degree of dedication – **I don't need to tell you** that.' [JSI][10]

The relations are not straightforward with structures comprising two predicates. The D1D2 order is generally preferred for both CM and MC orders, which means that reflecting the deep-structure order of arguments seems more important than reflecting the attachment to the governors with regard to their linear order. Additionally, while the proportions in the case of the MC order are very polarized (90.6% of D1D2), the difference between D1D2 and D2D1 order decreases to 14.55 percentage points in CM.

	D1D2	D2D1	∑
C	656	2	658
CM	63	47	110
MC	174	18	192
∑	893	67	960

Tab. 2. Distribution of dative pronoun orders in relation to predicate orders

[9] The sentences in (4–6) are very ambigous, so I recommend consulting the original source given in footnotes for the full contexts.

[10] https://www.sport-express.ru/figure-skating/reviews/988698/

I will now discuss the structure of data in relation to the referential prominence hierarchy (Principle 2). Since Principle 1 seems to have a strong influence on the argument order for structures with one overt infinitive predicate, I will focus on structures where both predicates are overtly marked in the sentence, as they show more variation. The detailed data distribution is presented in Tab. 3.

172 observations clearly conform to the referential prominence hierarchy (56.95%), while 36 observations (11.92%) clearly do not conform to this hierarchy. The remaining 29.80% consists mostly of the cases where two locuphoric pronouns are used. Two aliophoric pronouns appear in only 12 cases.

	D1D2 x MC	D2D1 x MC	D1D2 x CM	D2D1 x CM	∑
E1	57	2	20	3	82
E2	5	1	1	5	12
N	19	0	5	12	36
Y	93	15	37	27	172
∑	174	18	63	47	302

Tab. 3. Relation between dative pronouns to their referential prominence across different predicate orders and dative argument orders

According to Tab. 3, it seems that the exceptions from the D1D2 pattern in the MC order are mainly possible when the hierarchy is preserved. Otherwise, they are extremely rare, as in (5).

(5) *Ja, kstati, nikogda ran'še ob ėtom ne govoril, prosto vy, vidimo, takaja psichoterapevtičnaja dama, čto* ***vam***$_{D2}$ ***mne***$_{D1}$ ***zachotelos'***$_{M}$ ***rasskazat'***$_{C}$.
'By the way, I've never talked about this before, you just seem to be such a psychotherapeutic lady that **I felt like telling you**.'[11]

The picture for CM looks more complex. The D1D2 order is clearly preferred in the case of two locuphoric pronouns, and only slightly preferred when the referential prominence hierarchy is retained. In the case of two aliophoric pronouns and sentences with violated referential prominence hierarchy the pattern corresponding with the order of the governors is preferred over the linear order, replicating the syntactic hierarchy of the arguments, as in (6).

(6) *Moj brat rabotaet v Rossii, no u nego tože sem'ja i ja polagaju, čto* ***pomoč'***$_{C}$ ***nam***$_{D2}$ ***emu***$_{D1}$ ***budet složno***$_{M}$.
'My brother works in Russia, but he also has a family and I think **it will be difficult for him to help us**.'[12]

[11] https://www.svoboda.org/a/31168034.html

[12] https://www.bbc.com/russian/international/2015/07/150720_tajikistan_pamir_badakhshan_mud_slides

In other words, when the syntactically higher argument (D1) is aliophoric, the order of pronouns is parallel to the order of the governors: D1D2 for MC, and D2D1 for CM. In sentences with two locuphoric pronouns, D1D2 order seems to be preferred irrespective of the predicate order. The least clear tendencies concern the cases which conform to Principle 2: for MC 13.88% of the sentences contain D2D1 order, and for CM 42.18%.

3.2 Statistical evaluation of the distribution

I tested the observed tendencies statistically with a generalized linear mixed model fit by maximum likelihood 'glmer' (Bates et al. 2015). Since the number of sentences where the hierarchy is violated, or two aliophoric pronouns are used, is relatively low, I restrict the estimation to a model with variables as shown in Tab. 4. Additionally, I added the matrix verb lexeme as a random effect, since the data set contains multiple observations per lexeme.

Variable	Levels	Type of variable
Dative pronoun order	D1D2 D2D1	Dependent
Predicate order	MC – matrix complement CM – complement matrix	Independent
Referential prominence hierarchy	N(o) – an aliophoric pronoun is the syntactically higher argument (D1) Y(es) – hierarchy is preserved E1 – two locuphoric pronouns	Independent

Tab. 4. Variable description in the regression model

The quantitative results (see Tab. 5 for estimates and significance levels of fixed effects, and Fig. 1 for the estimated probabilities) confirm the intuition from Section 3.1.

Fixed effects:				
	Estimate	Str. Error	z value	Pr(>|z|)
(Intercept)	-0.1725	0.3997	-0.432	0.6660
E1	-1.4772	0.7136	-2.070	**0.0384***
No	1.5255	0.6043	2.524	**0.0116***
MC	-1.9929	0.5019	-3.971	**<0.0001***
E1:MC	0.3906	1.1018	0.354	0.7230
No:MC	-3.0578	1.2756	-2.397	**0.0165***

Tab. 5. Estimates and significance levels in the regression model

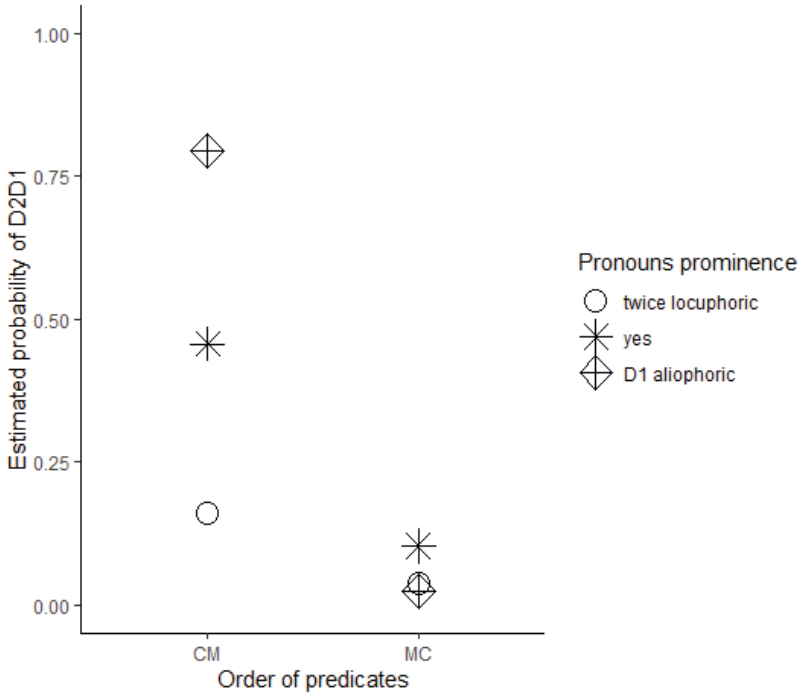

Fig. 3. Estimated probability of D2D1 order with the formula: $P=e^{logO}/(1+e^{logO})$, where P – probability, O – odds.

Both predicate order and referential prominence hierarchy are statistically significant. The reported results are compared to the following baseline: the D2D1 order occurs when pronouns retain the prominence hierarchy in the CM predicate order. When the relationship between pronouns changes and D1 is aliophoric the effect is positive and significant at the p=0.01 level. The probability that D2D1 occurs increases from 0.45, for the baseline, to 0.79. When both pronouns are locuphoric the effect is negative and significant at the p=0.04. The probability that D2D1 occurs decreases to 0.16.

The change in the predicate order has a generally negative effect, significant at the p<0.0001 level. For the pronouns with retained hierarchy the probability of D2D1 decreases to 0.10 and for two locuphoric pronouns to 0.02. This change has an even stronger negative effect (at the p=0.02 significance level) for the aliophoric D1 pronouns. The probability of D2D1 in the MC order is similarly low as for two locuphoric pronouns, that is, 0.02.

4 DISCUSSION

4.1 Central findings of the study

The results described in Section 3 show that ambiguous structures with two dative adjacent pronoun arguments do occur in language use. D2D1 order seems to occur relatively infrequently in the case of mono-predicative infinitive structures. With regard to the studied construction, the problem concerns mainly overtly bi-predicative structures. The results indicate that the order of dative arguments in bi-predicative structures does not always correspond to the structural order of arguments (contrary to Principle 1), nor does it reflect the linear order of predicates on the surface (contrary to Principle 3).

In the case of MC order, the probability of D2D1 order is relatively low. A considerable variation can be observed in situations where the verbal complement appears before the matrix verb. The probability of the reverse order D2D1, reflecting the linear predicate order CM, is in that case generally higher than in MC.

The type of relation between pronouns as to their prominence hierarchy is a statistically significant factor (Principle 2). The most variation is likely to occur when the hierarchy is retained. In other words, language users do not show any ambiguity avoidance strategies when the dative attachment should be made in line with the usual prominence of pronouns. The results are in favour of Haspelmath's claim (2021, p. 161) that "special grammatical markers are preferentially used when the grammatical meaning is least predictable." In this case, I conclude that when the hierarchy is not retained in the CM predicative order:

1. The most likely order is D1D2 if both pronouns are locuphoric, similar to the MC order.
2. The probability of D2D1 is very high if the structurally higher dative (D1) is aliophoric and the lower one (D2) is locuphoric.

Nonetheless, information about the relationship between pronoun arguments is available only contextually. Therefore, I assume that sentences with the reversed order of predicates (CM), where two dative pronoun arguments represent different levels of the prominence hierarchy, can pose interpretational problems for NMT systems and other tools for NLP.

Additionally, the findings are in contradiction with the results of Arnold et al. (2004) concerning PP dative attachment ambiguity. This may suggest that ambiguity avoidance strategies function differently across languages or across phenomena. For example, avoiding ambiguity between the core arguments might be more relevant for speakers than between a core argument and an adjunct. In the future, psycholinguistic tasks could be developed in order to find out whether the results of the present study are replicable in an experimental setting.

4.2 Limitations

In the present study, I could not account for several factors, as it was a pilot study, and the data set was relatively small. As already mentioned, I did not control for the syntactic type of matrix. I handled this problem by introducing a random effect – the matrix lexeme. Nonetheless, it is possible that some predicate orders are more typical for certain syntactic types than for others, and the studied sample could serve to build a new data set stratified across different syntactic predicate classes. Second, only adjacent positions of pronouns were analyzed. The generally low number of observations with an aliophoric D1 could occur because speakers prefer to keep the pronouns separate in such cases, for example, as an ambiguity avoidance strategy. Third, I did not analyze the separate positions of the matrix and complement in relation to the pronouns' positions. These factors could be implemented in future investigations where the query would include a list of matrix predicates obtained in the present study, and other word orders.

ACKNOWLEDGEMENTS

The research has been supported by the Representative for Equal Opportunities of the Faculty of Languages, Literatures and Cultures of the University of Regensburg. I thank Roman Fisun and Konstanzia Lüke for the help with data annotation.

References

Arnold, J. E., Wasow, Th., Asudeh, A., and Alrenga, P. (2004). Avoiding attachment ambiguities: The role of constituent ordering. Journal of Memory and Language, 51, pages 55–70.

Bawden, R. (2018). Going beyond the sentence: Contextual Machine Translation of Dialogue. Ph.D. thesis, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015.) Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), pages 1–48.

Bonč-Osmolovskaja, A. (2003). Konstrukcii s dativnym subjektom v russkom jazyke. PhD thesis, Moscow: MGU.

Grillborzer, C. (2019). Sintaksis konstrukcij s pervym dativnym aktantom: Sinxronnyj i diaxronnyj analiz. Frankfurt: Peter Lang.

Hansen, B. (2020). Subject. in: Encyclopedia of Slavic Languages and Linguistics Online, Editor-in-Chief M. L. Greenberg. Accessible at: http://dx.doi.org/10.1163/2589-6229_ESLO_COM_032471.

Haspelmath, M. (2021). Role-reference associations and the explanation of argument coding splits. Linguistics, 59(1), pages 123–174.

Lopes, A. V., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level Neural MT: A Systematic Comparison. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Maurice, F. (1996). Der modale Infinitiv in der modernen russischen Standardsprache. Peter Lang, Munich.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Proceedings of the Third Conference on Machine Translation.

Padučeva, E. V. (2017). Otricatel'nye mestoimenija predikativy (na ne-). In Russkaja korpusnaja grammatika. Accessible at: http://rusgram.ru/Отрицательныйе_местоимения--предикативы.

Trampuš, M., and Novak, B. (2012). The internals of an aggregated web news feed. In Proceedings of 15th Multiconference on Information Society 2012 (IS-2012). Accessible at: https://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.

Weiss, D. (1992). Infinitif et datif en polonais con temporain: un couple malheureux? In S. Karolak (ed.): Complétude et incomplétude dans les langues romanes et slaves. Actes du VI Colloque international de linguistique romane et slave, Cracovie 29 sept.–3 oct. 1991, pages 443-487. Wydawnictwo Naukowe WSP: Cracow.

sciendo
DOI 10.2478/jazcas-2023-0026

THE COMPETITION OF GERMAN ADJECTIVAL SUFFIXES

FILIP KALAŠ
Department of Linguistics and Translation, Faculty of Applied Languages,
University of Economics, Bratislava, Slovakia

KALAŠ, Filip: The Competition of German Adjectival Suffixes. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 81 – 91.

Abstract: The paper presents a corpus linguistic perspective on two adjectival suffixes *-al* and *-ell* in the German language. Its attention is focused on the distributional frequency of the derived adjectives, the semantic motivation and contextual occurrence through the lens of retried adjective + noun collocations. On top of that, the paper attempts to determine the superiority of such derived adjectives in the specialised vocabulary.
Keywords: suffix, derived adjective, productivity, general corpus, specialised context

1 INTRODUCTORY REMARKS

This study aims to discuss an interesting but rarely analysed, admittedly neglected phenomenon of the derivational morphology. Frankly speaking, there is almost no scientific article that would provide more detailed description of the German derivational suffixes *-al* and *-ell*. Needless to say, almost every book dedicated to German presents the typical methods of word formation and the derivational affixes used within this process supported by few examples, but on the other hand, they do not tend to deal with them in-depth.

Not only the scientific articles and monographies are lacking in giving more information regarding these adjectives, but also the lexicographic works. The information about the meaning and/or possible collocations are in case of many entries insufficient.

These facts were the impetus for the further systematic investigation of the mentioned competing suffixal pair in German as well as for shedding more light on their both quantitative and qualitative analysis.

1.1 Research aims

This paper has set following three aims that are to be accomplished concerning the competing suffixes *-al* and *-ell* in German by means of corpus evidence. Firstly, the distribution frequency of adjectival lemmas will reveal more recurrent suffix. Secondly, the analysis will focus on the semantic distribution of the particular extracted adjectives that detect the tendency to create fixed adjective + noun

collocations. Thirdly, the paper attempts to show the significance of adjectives ending with the abovementioned suffixes within the specialised language (Roelcke 2010). The last one will be achieved by learning the terminological character of the retrieved noun collocates and their proportion among collocates from everyday language.

1.2 Corpus description

To conduct the analysis, we will perform a detailed search in a common German corpus that is accessible for all users after registration. The comparable corpus Araneum Germanicum III Maximum (Benko 2014) which is a balanced corpus covering a wide range of text categories, predominantly fictional and journalistic texts. At the time of writing this paper it contains 8,912,400,350 tokens and 7,518,455,524 words. This corpus is powered by NoSketchEngine. The advantages of this corpus are an easy access and querying, balance and data quantity.

1.3 Methods

In order to obtain an accurate and representative picture as to which are the most common adjectives with the suffixes *-al/-ell* along with their variants, the large comparable corpus was chosen. Regarding corpus queries, the simple search of specific ending was opted, e.g. [lemma=".*iell"] or [lemma=".*ial"] followed by other query modifications. The maximum number of retrieved adjectives was set to ten for each group of suffixal variants. However, this number could not be achieved with suffixes *-iell, -ial, -al* and *-uell*. Simply, there was not enough appropriate candidates.

In the process of manual retrieval many seemingly appropriate candidates had to be filtered out since they failed to meet the requirements for adjectives with competing suffixal pair. The reason for this lies in the fact that we were strictly stuck to the following precondition – the stem of the most frequent retrieved token must be able to create derived adjectives with suffix *-al* (and its variants) as well as with *-ell* (and its variants). Thereby the data objectivity can be guaranteed. Based on this many highly recurrent adjectives had to be excluded from this survey such as international, *sozial* 'social', *digital*, *brutal*, *minimal*, *pauschal* 'general', *royal*, *fundamental*, *horizontal*, *katastrophal* 'catastrophic', *material* etc., because their pendants ending with *-ell* exist neither in modern nor in old German. The adjectives *bakteriell* 'bacterial', *speziell* 'special', *offiziell* 'oficial', *notariell* 'notarial', *textuell* 'textual' do not have their *-al* pendants either.

Furthermore, compound words derived from the same stem were excluded, e.g. *nachindustriell* 'postindustrial', *frühindustriell* 'early industrial', *schwerindustriell* 'heavy industrial', *hochaktuell* 'up-to-the-minute' or *tagesaktuell* 'updated daily' despite high frequency.

Two retrieved lemmas derived by the suffix *-al* (Exponential-, Material-) do not stand as an adjective alone, but exist only as a part of a compound word, e.g.

Exponentialkurve ‘exponential curve’ or *Exponentialfunktion* ‘exponential function’. Needless to say, this concerns explicitly other lemmas with the suffix *-al*, as well, however they are not frequent enough to get into the charts.

The adjectival pairs *partial/partiell, ministerial/ministeriell, personal/personell, real/reell, Individual/individuell, Sexual/sexuell* were randomly subjected to detailed analysis of the potential adjective + noun collocations. The adjective serves as a collocation basis, whereas the noun as the collocate. When retrieving data following conditions had to be met; frequency limit set to three, the 10 collocates sorted by logDice (Brezina 2018) were relevant. It is the statistical measurement which is highly recommended for lexicographic works and whose value should range between 14-5 (the lower the value, the less fixed the collocation is) (Káňa 2014, p. 18).

As Palková claims “the terminological motivation concerns technical terms, i.e., lexical units which are (besides other qualities) defined, fixed and systemised” (Palková 2015, p. 348). We consider this definition as determinative in process of assigning the terminological character to the analysed collocations. We will focus on the lexicographic entry of the particular nominal collocate either in DUDEN or in DWDS, based on which its terminological character can be confirmed (naturally often by subjective assessment).

2 THEORETICAL BACKGROUND

Word-formation is such a confused study at the moment that it would not be possible to write an uncontroversial introduction to the subject (Bauer 1983, p. 13). This quote has still its relevance also after four decades because there are numerous research topics that have stayed unresolved and unanswered. One of them are the word-formation processes. Word-formation is important for any language since it expands the vocabulary and adapts to language users who constantly have new objects, events or processes to name (Kalaš 2020, p. 11). Derivation (Lieber 2016) as one of them along with conversion, compounding, abbreviation, back formation etc. is the process by which new words are produced by modification of existing words. This modification is carried out by means of derivational morphemes (affixes). These can be either bound (-eur, -ismus) or separable (**vor**kommen, **bei**tragen).

Both prefixation and suffixation are highly productive (Trost 2022, p. 228). When it comes to the frequency proportion between suffixes and prefixes, the cross-linguistic distribution of suffixes amounts to 70% of all affixes, the rest is represented by the prefixes (Hall 2000, p. 539). Many languages have a great number of various suffixes leading to a significant dichotomy. A suffix is a specific bound morpheme that is attaches to a root or stem (Römer 2022, p. 710). One group of them is so-called endogenous (native) morphemes that originate from the repertory of the

particular language such as *-er, -el, -ung-, heit* in case of German. The other group is represented by exogenous (foreign) elements, e.g. *-ik, -ie, -ität, -ismus* etc., which have been penetrating the language for ages. These so-called derivational suffixes are word-class specific and they determine the word class of the word to which they attach (cf. Römer 2022).

The object of this paper is the synonymous pair of exogenous (foreign/loan) adjectival suffixes *-al/-ell*. Römer (2022) claims that the productive exogenous suffixes tend to often have grammatical constraints, thus connections with a native base in German are almost impossible (cf. p. 299). The suffix *-al* along with its variants *-ial* and *-ual* (cf. p. 712) belong to the most widespread exogenous suffixes. Its origin is in Latin, but it got into German through French (starting with 16th century, but mainly in 17th and 18th centuries (Schmidt 2000, p. 137)) and English (after the WWII). By means of this suffix an adjective is derived from a noun, e.g. *Zentrum* 'centre'→ *zentral* 'central', whereas the final syllables are often erased as in the mentioned example. Hentschel highlights the fact that large parts of the words formed in this way belong to the specialised vocabulary of various scientific fields (Hentschel 2020, p. 161).

The second suffix *-ell* along with its variants *-iell* and *-uell* originates from French (Donalies 2017). Likewise the previous suffix, this one is applied to build adjectives from foreign and loan words, e.g. *Konzeption* 'concept'→ *konzeptionell* 'conceptional' (Hentschel 2020, pp. 162–163). These synonymous suffixes appear to be used interchangeably thanks to their motivational potency – both derive adjectives from Roman languages; thus, they are generally perceived as equal partners in terms of semantics. Nevertheless, there is a further differentiation of meaning possible (Buscha – Friedrich 1996):

a) referring to a thing named in the first element (reference adjective), e.g. *regional* 'regional', *industriell* 'industrial';
b) expressing the property named in the first element (comparative adjective), e.g. *katastrophal* 'catastrophic', *sensationell* 'sensational';
c) caused by the thing named in the first element, e.g. *hormonal* 'hormonal' compared to *hormonell* 'hormonal' (Buscha – Friedrich 1996, pp. 166–167).

Hentschel (2020) advocates the concept of semantic deviation induced by suffixal change stating that occasionally the suffix *-ell* competes with *-al* with more or less clear differences in meaning in each case. The semantic competition becomes particularly visible where the two suffixes with the same base go together (Donalies 2017), e.g. *nominaler Wert* 'nominal value' vs. *nominelles Mitglied* 'nominal member'. Based on this hypothesis we formulate and work with the term "competing (suffixal) pair". Donalies (2017) concludes that there is still no clear difference and the word-formation trend is based on the preference of one or another suffix. Sometimes the *-al* adjectives like *funktional* 'functional', sometimes the *-ell* adjectives like *experimentell* 'experimental' are more common.

3 RESULTS

3.1 Frequency distribution of suffixes

Fig. 1 depicts the most common adjectives ending with the derivational suffix *-al* and its variant *-ell.* The instances have been excerpted from Araneum Germanicum III Maximum (Benko 2014). A larger proportion is represented by *-ell* adjectives with 10 examples in comparison to 7 examples of *-al* adjectives. Nevertheless, the latter ones outnumber heavily far superior counter-group with the most frequent adjectives *ideal* and *real* which reach from 300,000 up to 684,000 instances. Then the chart indicates a sharp decrease in the frequency.

When it comes to the adjectives ending with suffix *-ell*, the first place is taken by the lemma *generell* 'general' reaching 470,000 instances. Subsequently, the frequency drops to 139,000 instances of the adjective *konventionell* 'conventional' followed by a gradual decrease of the remaining adjectives. There are 3 competing pairs which end up in the chart thanks to their high frequency. The pairs are as follows: *originell* 'original'/original, *personell* 'personal'/personal, *reell* 'real'/real.

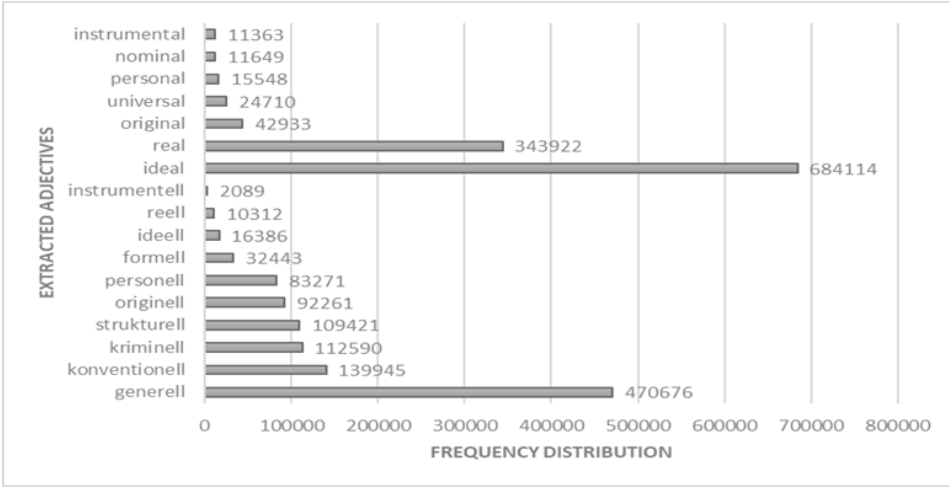

Fig. 1. Frequency of the adjectives derived by suffixes *-al/-ell*

Fig. 2 provides information about the most recurrent adjectives derived by the suffixes *-ial/-iell*. There is a dramatical gap between both groups in terms of distribution making the adjectives with *-iell* endings significantly more frequent than the variants with *-ial.* For example, the adjective *speziell* 'special' achieves more than one million instances followed by other adjectives reaching more than one hundred up to three hundred instances. The least frequent adjective is *ministeriell* 'ministerial' with 1,998 instances.

On the other hand, the first place among adjectives ending with *-ial* takes the lemma *spezial* 'special' reaching 23,551 instances, remarkably leapfrogging the remaining lemmas with the same suffix. Representation in this counter-group is vastly low which can draw the conclusion that the suffix *-iell* undoubtably gains the upper hand over the suffix *-ial* not only in the frequency but also in the number of retrieved data.

Interestingly, there are 3 competing adjectival pairs, e.g. *speziell* 'special'/ *spezial* 'special', *partiell* 'partial'/*partial* 'partial', *ministeriell* 'ministerial'/ *ministerial*.

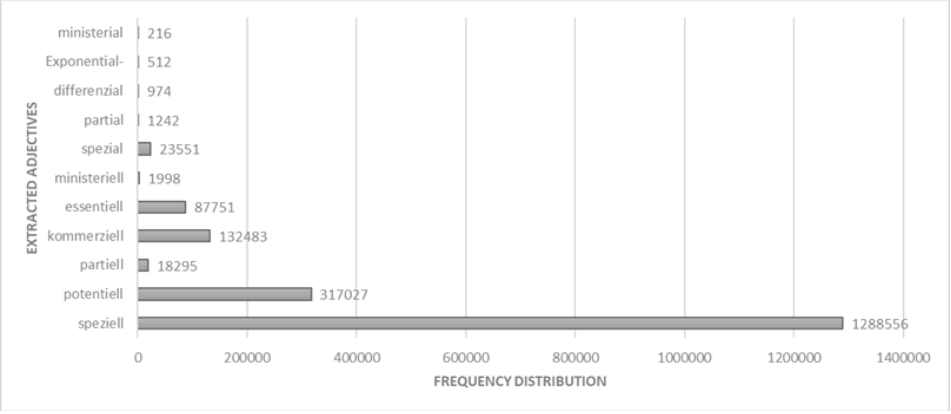

Fig. 2. Frequency of the adjectives derived by suffixes *-ial/-iell*

Fig. 3 shows again the distribution of the adjectives which are derived by the suffixes *-ual* and *-uell*. Upon initial inspection there is a numerical superiority of *-ell* adjectives over *-al* ones. The adjective *aktuell* 'actual' reaches the utmost frequency with more than 2,6 million instances, whereas its counterpart *aktual* 'actual' amounts to rare 388 examples. The second place is taken by the adjective *individuell* 'individual' with more than 1,2 million instances. Less by half has the adjective *eventuell* 'potential' reaching the third place. Afterwards, the chart illustrates slight decrease in the frequency distribution.

Two most numerous representatives of the counter-group are the adjectival stems *Sexual-* with 184 thousand hits and *Individual* exceeding 155 thousand instances followed by two adjectives *prozentual* 'percental' and *virtual*, both overlapping 22 thousand instances. After that, occurrence markedly plunges up to 62 instances of the adjective *asexual*. There are 4 following competing pairs *aktuell/ aktual, individuell/individual, eventuell* 'potential'/*eventual* 'eventual', *manuell* 'manual'/*Manual-*.

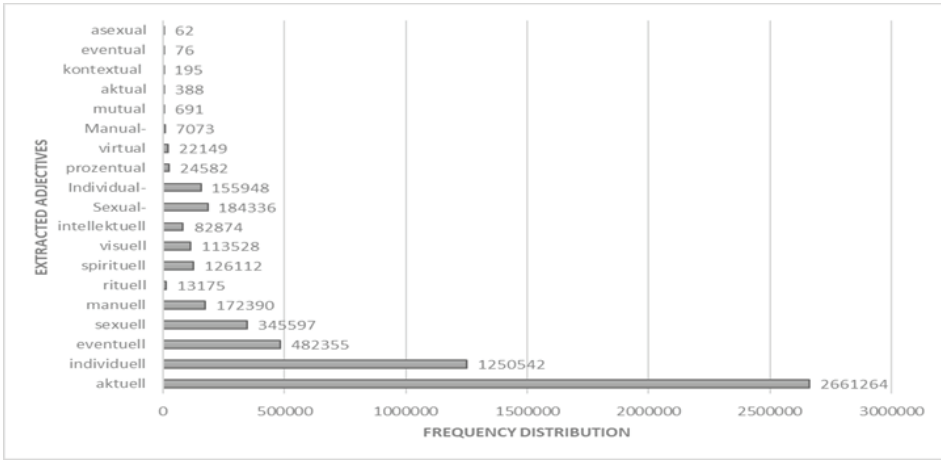

Fig. 3. Frequency of the adjectives derived by suffixes *-ual/-uell*

3.2 ADJ + N collocation

After the conduction of the collocational analysis, it can be stated that none of the analysed adjectival pairs shares mutual collocate(s) retrieved amongst the set range, which is why every collocation is unique. Furthermore, the collocability of the collocation bases varies significantly. Some is small (*partial, ministerial*), some is large (*partiell, personell, sexuell*). It is noteworthy that the adjectives with suffix *-ell* have prevailingly large(r) collocability and/or lexical compatibility than their counter-suffixes.

Tab. 1 depicts the strongest nominal collocates of the adjectival bases *personal* and *personnell*. Remarkable is not only the fact that no collocate occurs with both bases at the same time, but also that the equivalency of the bases contrasts mutually. A considerable (or preferred) equivalency is marked by adding the translation into parentheses. In this way the collocation *personelle Engpässe* would be 'staff shortages' instead of 'personnel shortages', what is nowadays considered to be less frequent. Another fact worth mentioning is that the collocations with *personell* have higher recurrence as well as stronger collocability.

	personal			***personell***	
collocate	frequency	logDice	*collocate*	frequency	logDice
Erzählperspektive '(personal) narrative perspective'	90	7.41	*Engpass* '(staff) shortage'	925	8.61
Erzähler '(personal) narrator'	177	7.13	*Ressource* '(manpower/human) resource'	2474	8.47

Kompetenz ‘(personnel) competence’	674	6.48	*Verflechtung* ‘interweavement’	647	8.23
Erzählsituation ‘(figural) narrative situation’	34	6.14	*Kontinuität* ‘continuity’	695	8.03
Identität ‘identity’	323	6.04	*Ausstattung* ‘(human) resources’	1852	7.88
Seelsorgebereiche ‘pastoral services’	27	5.82	*Besetzung* ‘staff’	890	7.85
Erzählweise ‘narrative style’	29	5.42	*Einzelmaßnahme* ‘(staff-related) measure’	401	7.70
Erzählhaltung ‘narrative attitude’	19	5.28	*Verstärkung* ‘(personnel) reinforcement’	575	7.71
Würde ‘dignity’	92	5.12	*Konsequenz* ‘(personnel) consequence’	1368	7.45
Erzählstil ‘narrative style’	19	5.02	*Kapazität* ‘(staff) capacity’	712	7.32

Tab. 1. Collocability of the adjectival bases *personal/personell*

Another fact is that the particular choice of the suffix is affected also by the context and vice versa. The abovementioned six adjectival pairs (12 adjectives) refer to virtually different context and thus, it is also a significant factor in terms of their usage. Additionally, to understand better the meaning, we present the following 12 adjectives with their primary semantics: (i) *partial* = in our data there is just one co-occurrence with the substantive *Bedingtheit* ‘partial contingency’, (ii) *partiell* – partly, (iii) *ministerial* – in this case there is solely one co-occurrence again – *auf ministerialer Ebene* ‘at ministerial level’, (iv) *ministeriell* – used in context of ministerial affairs/matters, (v) *personal* – it refers to an individual and individual-related matters, in our case it occurs predominantly in literary context (Tab. 1), (vi) *personell* – refers to staff and staff-related matters (Tab. 1), (vii) *real* – it covers concrete and reality-related matters, (viii) *reell* – it expresses honest, sincere things in economics and true and actual things in mathematics, (ix) *Individual* – refers mainly to the needs of individual person, (x) *individuell* – this adjective expresses something distinct and customized, (xi) *Sexual-* in terms of venereal, carnal but also sexual matters, (xii) *sexuell* – it expresses primarily something sexual.

3.3 Terminological character

The graph (Tab. 4) synoptically illustrates the proportion of the excerpted collocates of the particular adjectival bases to the specialised vocabulary. From the

data in the graph, it is apparent that the least proportion (number of specialised words out of 10) is associated with the adjectives ending with *-al* (real, ministerial, partial), whereas the highest number is to assign to the adjectives with suffix *-ell* (partiell, ministeriell, sexuell). Both *individuell* and *Individual-* reach the identical number of specialised collocates. Furthermore, as can be seen from these results, there are two adjectival pairs, by which the adjective with suffix *-al* dominates over its suffixal counter-part.

In the process of determining the terminologisation of the particular collocates, another remarkable aspect was revealed – specialised context. It is evident from the Tab. 1 that the context significantly varies. Whereas the adjective *personal* implies the literary field, *personell,* on the other hand, is used mainly in human resource context. The adjective *real* is linked with no specific context, whereas its counter-part *reell* with maths, e.g. *reeller Vektorraum* 'real vector space'*, reeller Zahlenwert* 'real numerical value'*, reelle Analysis* 'real calculus' or with physics *reelles Zwischenbild* 'real intermediate image'. Another example is the adjective *partial*, which is used in no specialised context, while the adjective *partiell* is applied in many specialised fields as astronomy, e.g. *partielle Sonnenfinsternis* 'penumbral solar eclipse', *partielle Mondfinsternis* 'partial lunar eclipse', maths, e.g. *partielle Differentialgleichung* 'partial differential equation', auto industry, e.g. *partielle UV-Lackierung* 'partial UV varnish', physics, e.g. *partielle Kernschmelze* 'partial core meltdown' or law, e.g. *partielle Gesamtrechtsnachfolge* 'partial universal succession'.

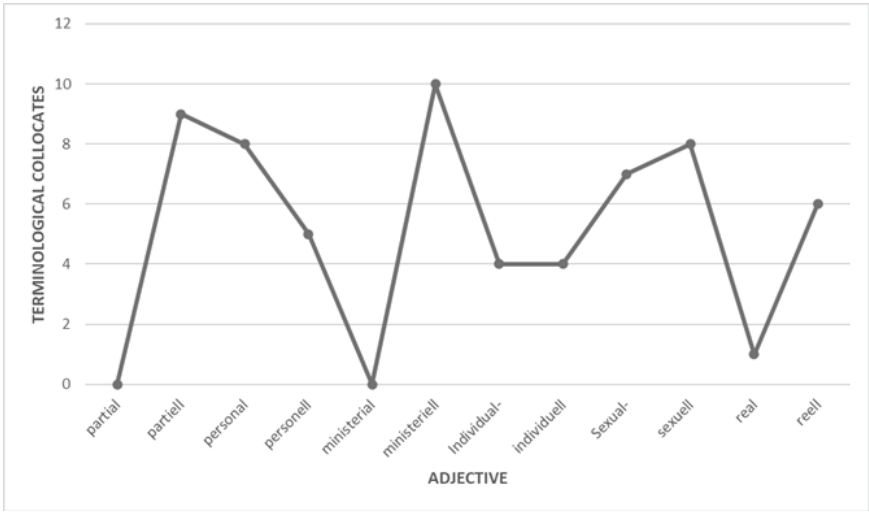

Fig. 4. Proportion of the terminological collocates

4 CONCLUSION

This paper presented the analysis of twelve retrieved adjectives with the suffixes *-al* and *-ell* based on their high recurrence, and attempted to corroborate or to disprove three hypotheses. To summarise, the suffix *-al* and all its derived adjectives are significantly less frequent, they are distinguished by weak lexical compatibility and they occur less in specialised context. Interestingly, some derivates tend to create composite words rather than stand in attributive position (*Sexualleben* 'sexual life'). On the contrary, in case of the suffix *-ell it* is quite the opposite thanks to high occurrence, strong lexical compatibility and together with other collocates the derived adjectives are very often part of different specialised vocabularies. Therefore, these suffixal pairs can be considered to be synonymous from the orthographical point of view. However, it was shown to satisfaction that they are different in terms of usage, semantics, context and lexical compatibility.

ACKNOWLEDGEMENTS

The research has been elaborated within the project of young teachers, scholars and postgraduate students I-23-101-00 *The global finance crisis from economic and linguistic points of view*.

References

Bauer, L. (1983). English word-formation. Cambridge: CUP, 311 p.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka – A. Horák et al. (eds.): Text, Speech and Dialogue. 17th International Conference. Cham: Springer, pages 247–256.

Brezina, V. (2018). Statistics in corpus linguistics. Cambridge: University printing house, 316 p.

Buscha, A., and Friedrich, K. (1996). Deutsches Übungsbuch. Berlin-Schöneberg: Langenscheidt, 237 p.

Digitales Wörterbuch der deutschen Sprache. Accessible at: https://www.dwds.de.

Donalies, E. (2017). Instrumental oder instrumentell? – Bildung von Adjektiven. In Leibniz-Institut für deutsche Sprache: Grammatik in Fragen und Antworten. Accessible at: https://grammis.ids-mannheim.de/fragen/3057.

Duden Online dictionary. Accessible at: https://www.duden.de.

Hall, Ch. J. (2000). Prefixation, suffixation and circumfixation. In Morphology. An international handbook on inflection and word-formation (1), pages 535–545.

Hentschel, E. (2020). Basiswissen deutsche Wortbildung. Tübingen: Narr Francke Attempto Verlag, 241 p.

Kalaš, F. (2020). Analyse der usuellen Wortverbindungen in der Wirtschaftssprache anhand des deutsch-slowakischen Korpus. In S. Adamcová (ed.): Usuelle Wortverbindungen in

der deutschen Wirtschaftssprache und ihre Widerspiegelungen in mehreren Sprachen. Hamburg: Verlag Dr. Kovač, pages 9–64.

Káňa, T. (2014). Sprachkorpora in Unterricht und Forschung DaF/DaZ. Brno: Masarykova univerzita, 212 p.

Lieber, R. (2016). Introducing morphology. Cambridge: University printing house, 243 p.

Palková, L. (2015). Univerbizácia – syntaktická motivácia. In M. Ološtiak (ed): Viacslovné pomenovania v slovenčine. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove, pages 281–356.

Roelcke, T. (2010). Fachsprachen. Berlin: Erich Schmidt Verlag, 269 p.

Römer, Ch. (2022). Ausgewählte Termini im alphabetischen Wörterverzeichnis. In S. J. Schierholz – P. Uzonyi (eds.): Band 1, Formenlehre. Berlin, Boston: De Gruyter.

Schmidt, W. (2000). Geschichte der deutschen Sprache. Stuttgart: S. Hirzel Verlag, 407 p.

Trost, I. (2022). Ausgewählte Termini im alphabetischen Wörterverzeichnis. In S. J. Schierholz – P. Uzonyi (eds.): Band 1, Formenlehre. Berlin, Boston: De Gruyter.

sciendo
DOI 10.2478/jazcas-2023-0027

PROVERBS IN CONTEMPORARY CZECH. CORPUS PROBE INTO WRITTEN TEXTS

MARIE KOPŘIVOVÁ[1] – KATEŘINA ŠICHOVÁ[2]
[1] Institute of Czech and Deaf Studies, Faculty of Arts, Charles University, Prague, Czech Republic
[2] Bohemicum – Center for Czech Studies, University of Regensburg, Regensburg, Germany

KOPŘIVOVÁ, Marie – ŠICHOVÁ, Kateřina: Proverbs in Contemporary Czech. Corpus Probe into Written Texts. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 92 – 99.

Abstract: The paper deals with the possibility of creating a paremiological optimum for students of Czech as a foreign language. The selection of proverbs should reflect the frequency, familiarity with and use of proverbs. The study focuses on the most frequent proverbs in written Czech, using contemporary idiomatically annotated corpora. On this basis, our own minimum was created. The paper compares the results with previous studies on the paremiological minima of Czech (Schindler 1993 and Čermák 2003) and shows the intersection of all three minima.

Keywords: paremiological minimum/optimum, idiom annotation, Czech proverbs, corpus

1 INTRODUCTION

Paremiology has enjoyed scholarly interest for several centuries, and we already have the Book of Proverbs in the Bible, although it contains a somewhat different kind of text than we are accustomed to understand by proverbs today; its content is identical to today's proverbs in that it gives advice on how to live well. The Adagia of Erasmus of Rotterdam from 1500 is considered one of the oldest collections of European phraseology, but there are also extensive collections in Czech, e.g. by Comenius (first edition 1849), Čelakovský (1852) and Flajšhans (1911 and 1913, new edition 2013). All these authors tried to collect as many proverbs as possible, searching them mostly in literature and sometimes translating them from other languages. From today's point of view, they could be called an attempt at phraseological maximum.

Proverbs have been studied from a professional point of view not only in linguistics, which deals with them as special linguistic units and classifies them as idioms of sentence nature (cf. Čermák et al. 2009, hereafter SČFI), but also, for example, in folkloristics, which views them as a condensed experience (e.g.

Bittnerová – Schindler 2003), and phraseodidactics, which focuses on practicing them (cf. Šemelík – Šichová 2016, 2017; Jesenšek 2013). The continued interest in this topic is also evidenced by a number of other activities focused on proverb research, such as the Proverbium journal,[1] the international conference Interdisciplinary Colloquium on Proverbs,[2] and a web database capturing proverbs from different languages.[3]

There are also many popularising book and web collections or teaching materials.[4] Proverbs have a special place in teaching Czech as a mother tongue and as a foreign language and knowledge of their form and meaning is tested in examinations.

The problem is that proverbs are linguistic units of several words and as such are not very frequent in texts. In addition, some proverbs contain archaic realities and may therefore be considered by students to be outdated and no longer in use today. For this purpose, we no longer need the most extensive collection of proverbs, where we do not find information about their use, but an idea of which proverbs are the ones known and used.

As far as the definition of proverb is concerned, we conceive of it rather broadly and lean towards the definition of the paremiologist W. Mieder: "Proverb is commonly thought of as a phrase, saying, sentence, statement, or expression of the folk which contains above all wisdom, truth, morals, experience, lessons, and advice concerning life and which has been handed down from generation to generation." (Mieder 1993, p. 24).

Another important characteristic of proverbs is their stability in sentence form and the fact that they can be understood without adaptation to context (Schindler 1996, p. 265). They share this property with quotations. Therefore, they appear in texts as headings, summaries or comments on certain events, and they tend to be introduced by special introducers or formulas. (cf. Čermák 2007, pp. 549–568).

2 PREVIOUS RESEARCH

Mainly with regard to lexicography (problematic definition of paremiological units and traditionalisation of obsolete or unusual units) and phraseodidactics (the need to select appropriate units for the purpose of teaching a given language as a foreign/second language), the "maximalist demand" and attempts to document all proverbs of a given language, including variants, are abandoned (Ďurčo 2015b, p. 41). There are efforts to produce more quantitatively limited sets, e.g. to filter a certain core of the paremiological vocabulary of a given language.

[1] https://naklada.ffos.hr/casopisi/

[2] https://www.folklore.ee/rl/fo/koostoo/tavira/proceedings.htm

[3] http://www.sprichwort-plattform.org/

[4] e.g. https://www.presto.cz/cz/ceska-prislovi, https://www.knihovnahk.cz/files/tinymce/Prislovi.pdf

The work of the Russian paremiologist Grigoriy Permyakov was groundbreaking in this respect. Permyakov's aim (1971, cited in Mokienko 2012) was to compile the "paremiological fund" of Russian, i.e., to identify the most frequent proverbs. From these, he then tried to identify the minimum set of proverbs known to speakers, the so-called "paremiological minimum" (for Russian, he put it at 300 proverbs). Since Permyakov's idea, attempts have been made to select proverbs in a scientific way for other languages (e.g. for English Mieder 1994; Haas 2008; for German Baur – Chlosta 1996a, 1996b; Ďurčo 2015a; Grzybek 2012; for Slovak Ďurčo 2015b: 196ff.; for more cf. Mokienko 2012, pp. 79–83 and Ďurčo 2015a), while from the point of view of language teaching it has over time proved more appropriate to speak of the paremiological optimum of the respective language rather than the minimum.

The approaches to obtaining material and the criteria for selecting paremiological units vary (Jelínek et al. 2018), partly conditioned by lexicographic and/or linguistic goals (Hrisztova-Gotthard – Varga 2014): analysis of scientific publications (e.g. Mieder 1994), corpus search (e.g. Steyer 2012), analysis of dictionaries (e.g. Hessky – Ettinger 1997).

2.1 Paremiological minima in Czech

For Czech F. Schindler (1993), following Permajakov's model, compiled a paremiological minimum: in a questionnaire survey, he had individual speakers complete the second part of the proverb and then evaluated the degree of knowledge. His great contribution is the very selection of the proverbs to be tested; he used the above-mentioned collections, especially Čelakovský's, and the glossary of the forthcoming 4th volume of the SČFI (Čermák et al. 2009). The selection of proverbs is made more difficult by the lack of a standard definition. However, Schindler, based on his research, established 99 proverbs as the paremiological minimum (Schindler 1996); other proverbs for which respondents showed lower knowledge are included in the book Czech Proverbs (Bittnerová – Schindler 2003). This research was used by F. Čermák as a starting point to verify the frequency of individual proverbs in the corpus (Čermák 2003); for his research he chose the 100-million representative corpus SYN2005 and determined 100 most frequent proverbs in this corpus.

2.2 Paremiological optimum and the prerequisites for its creation

Nowadays, some linguists argue for a critical view of the minima, e.g. Mokienko states: "It is didactic goals and lexicographical description that determine in each individual case the particular paremiological minima suitable for particular addressees or groups of addressees. A fixed paremiological minimum, i.e., a minimum for all language users, is a purely scientific abstraction if not an outright fiction." However, if the knowledge of proverbs is to be taught among students of Czech as a foreign language, it seems essential to create a paremiological optimum,

i.e., a set of proverbs that are frequent (used in Czech texts) and/or known (actively, passively by speakers of Czech). It is important to distinguish between knowledge of form and knowledge of meaning. However, the latter is difficult to verify, and it is similarly problematic to find out directly from speakers which proverbs they actively use in everyday communication (cf. Grzybek 2012).

The aim of our research is to determine the phraseological optimum for learners of Czech with German as their mother tongue. The starting point would be a proverb pool from which proverbs meeting the relevant criteria (frequency, familiarity, usage) would be selected. These basic criteria can be extended by additional criteria specific to certain user groups, thus obtaining paremiological optima suitable for specific purposes. In our case, we could add as an additional criterion, for example, the existence of an equivalent proverb in German.

Some proverbs in European languages have similar origins (cf. e.g. Tölgyesi 2022), but the approach to their use may be different (cf. Schindler 1996, p. 280), which is reflected in teaching and testing. Passive comprehension is also important for learners of Czech as a foreign language, and thus the frequency with which they encounter proverbs plays a large role here. The frequency of occurrence of proverbs can be quite reliably determined in written Czech using corpora. Therefore, in the first step, we only start from the frequency of proverbs and from a comparison with previous paremiological minima (Schindler 1996; Čermák 2007). We want to determine the consistency between the different approaches.

3 DATA AND METHODOLOGY

Compared to Čermák's study, we used more extensive data, in which the annotation of idioms is additionally present (cf. Hnátková 2002). However, proverbs do not form a separate group within the annotation, but are part of a set of sentence idioms, that must be searched manually to decide which ones are proverbs. In this section, we have verified the existence of proverbs in the Dictionary of Proverbs (Čermák 2013). Currently, a more detailed processing of proverbs is underway using the LEMUR database, which will also serve to better annotate proverb fragments. Unfortunately, this database is not yet available to a satisfactory extent (Hnátková et al. 2017).

It is possible to use more extensive data from a larger time period, which includes the syn_v9 corpus (with texts from 1990–2019). This corpus is heavily dominated by journalism, so it is not suitable for our probe; the frequency of proverbs would reflect their use in journalism. To limit the influence of the journalistic genre, a balanced corpus consisting of 4 representative corpora of written Czech was chosen: SYN2000, SYN2005, SYN2010, SYN2015.

This corpus contains 32% fiction, 40% journalistic and 28% professional literature. The highest number of proverbs is found in journalistic, then in fiction,

and the lowest in professional literature, which includes popular literature. To reduce the influence of journalistic literature, three lists of proverbs were created according to each type of text, and the set of those proverbs that appeared in all three lists is hereafter referred to as paremiological minimum 2023 (PM23).

4 COMPARISON OF PAREMIOLOGICAL MINIMA RESULTS

The 100 most frequent proverbs from PM23 were compared with Schindler's (99 proverbs) and Čermák's minimum (100 proverbs). The higher agreement with Čermák's minimum was confirmed, with 57 identical proverbs, and only 31 with Schindler's; agreement with Čermák or Schindler accounted for 62 proverbs. The intersection of all three was 28 proverbs. It is not clear whether the differences are more due to a temporal shift or to the composition of the data. In any case, we believe that these proverbs could form the core of the paremiological optimum.

They are the following proverbs:

1. *Šaty dělají člověka.* 'Clothes make the man.' 'Fine feathers make fine birds.'
2. *Dvakrát měř, (a) jednou řež.* 'Measure twice, cut once.' 'Look twice before you leap.'
3. *Vlk se nažral a koza zůstala celá.* 'The wolf has eaten and the goat has remained entire.'[5]
4. *Stará láska nerezaví.* 'Old love is never forgotten.'
5. *Bližší košile než kabát.* 'Near my coat, but nearer my skin (shirt).'
6. *Bez práce nejsou koláče.* 'No work, no cake.' 'No cross, no crown.'
7. *Kdo nic nedělá, nic nezkazí.* 'He who does nothing, spoils nothing.'
8. *Sejde z očí, sejde z mysli.* 'Out of sight, out of mind.'
9. *Co na srdci, to na jazyku.* 'What's on the heart is on the tongue.'
10. *Lež má krátké nohy.* 'Lies have short wings (legs).'
11. *Kdo hledá, najde.* 'Search and you shall find.'
12. *Lepší vrabec v hrsti než(li) holub na střeše.* 'Better a sparrow in the hand than a pigeon on the roof.' 'A bird in hand is worth two in the bush.'
13. *Všude dobře, doma nejlíp.* 'There is no place like home.'
14. *I mistr tesař se utne.* 'Even a master carpenter gets it wrong.'
15. *Jablko nepadá daleko od stromu.* 'The apple doesn't fall far from the tree.' 'Like father, like son'. 'He is a chip of the old block.'
16. *Láska hory přenáší.* 'Love moves mountains.'
17. *Vrána k vráně sedá.* 'The crow sits with the crow.'
18. *Tonoucí se stébla chytá.* 'A drowning man clutches at a straw.'

[5] For the expression *the wolf has eaten and the goat has remained entire* we can doubt its classification as a proverb if we follow the formal criterion for proverbs, which is its independence and syntactic uninvolvement in the sentence. This expression is usually part of a sentence with a purposeful meaning: to do something so that the wolf eats and the goat stays entire.

19. *Čert nikdy nespí.* ‘Devil never sleeps.’

20. *Kuj železo, dokud je žhavé.* ‘Strike while the iron is hot.’

21. *Co Čech, to muzikant.* ‘Every Czech is a musician.’

22. *Každý svého štěstí strůjcem.* ‘Each man is the author of his own happiness.’ ‘Life is what you make it.’

23. *Kdo dřív přijde, ten dřív mele.* ‘First come, first served.’

24. *Na hrubý pytel hrubá záplata.* ‘For a coarse sack a coarse patch.’ ‘Meet rudeness with rudeness.’

25. *Ráno moudřejší večera.* ‘Morning is wiser than evening.’

26. *Boží mlýny melou pomalu, ale jistě.* ‘The mills of God grind slowly but surely.’

27. *Na každém šprochu pravdy trochu.* ‘There’s a little truth in every speck.’

28. *Mluviti stříbro, mlčeti zlato.* ‘Silver to speak, gold to keep silent.’

The list is dominated by short proverbs, which generally have less variation and their form is more easily reproduced, this ensures a higher frequency. In Schindler’s research, this feature does not play a role, but it is reflected to some extent in the familiarity with proverbs: speakers remember short proverbs better, while longer ones show higher variability. Thus, the intersection of all three minima results in shorter proverbs.

5 FUTURE RESEARCH

We plan to repeat the selection of proverbs using the more extensive and recent data by adding a representative corpus of syn2020. Further work is needed to check the annotation with respect to its scope: different variants of proverbs are not always included in the annotation, while it is the variability that can be high in longer units and should be included in the overall frequency of proverbs. There may also be deliberate modifications, which may indicate that the author considers the proverb to be so well known that he can actualize it. Next, we focus on the distribution of individual proverbs within the data, as there may be a cumulation of proverb occurrences due to the proverb being used as the title of a book, exhibition, or to indicate some current information; for this case, an average reduced frequency (ARF) would be appropriate.

In further research, we also want to focus on the selection of proverbs in lexicographic processing, in the production of textbooks, on comparison with the German paremiological optimum, and on research on proverbs among native speakers in terms of proverb familiarity and usage.

References

Baur, R. S., and Chlosta, Ch. (1996a). Welche Übung macht den Meister? In Fremdsprache Deutsch, 15, pages 17–24.

Baur, R. S., and Chlosta, Ch. (1996b). Sprichwörter: ein Problem für Fremdsprachenlehrer wie lerner?! In Deutsch als Fremdsprache, 33, pages 91–102.

Bittnerová, D., and Schindler, F. (2003). Česká přísloví. Soudobý stav konce 20. století. Praha: Karolinum.

Čermák, F. (2013). Základní slovník českých přísloví: Výklad a užití. Praha: Nakladatelství Lidové noviny.

Čermák, F. (2007). Frazeologie a idiomatika česká a obecná. Czech and General Phraseology. Praha: Karolinum.

Čermák, F. (2003). Paremiological Minimum of Czech: The Corpus Evidence. In H. Burger – A. Häcki Buhofer – G. Gréciano (eds.): Flut von Texten – Vielfalt der Kulturen. Hohengehren: Schneider Verlag, pages 15–31.

Čermák, F., Holub, J., Blatná, R., and Kopřivová, M. (2009). Slovník české frazeologie a idiomatiky (SČFI), vol. 4. Praha: Leda.

Ďurčo, P. (2015a). Diasystematische Differenzen von Sprichwörtern aus der Sicht der kontrastiven Parömiografie. In P. Ďurčo – K. Steyer – K. Hein: Sprichwörter im Gebrauch. Trnava, pages 121–142.

Ďurčo, P. (2015b). Empirical Research and Paremiological Minimum. In H. Hrisztova-Gotthardt – M. A. Varga (eds.): Introduction to Paremiology. A Comprehensive Guide to Proverb Studies. De Gruyter Open, pages 183–205.

Grzybek, P. (2012). Facetten des parömiologischen Rubik-Würfels Kenntnis ≡ Bekanntheit [⇔ Verwendung ≈ Frequenz] ?!? In K. Steyer (ed.): Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie. Narr Verlag, pages 99–138.

Hnátková, M. (2002). Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. Slovo a slovesnost, 63(2), pages 117–126.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., and Vondřička, P. (2017). Eye of a Needle in a Haystack. Multiword Expressions in Czech: Typology and Lexicon. In R. Mitkov (ed.): Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017. London: UK, November, pages 13–14.

Hrisztova-Gotthard, H., and Varga, M. A. (2014). Introduction to Paremiology. A Comprehensive Guide to Proverb Studies. De Gruyter Open.

Hessky, R., and Ettinger, S. (1997). Deutsche Redewendungen. Ein Wörter- und Übungsbuch für Fortgeschrittene. Tübingen.

Jesenšek, V. (2013). Sprichwortgebrauch heute. Linguistische und sprachdidaktische Überlegungen. In Muttersprache 2, pages 81–98.

Jelínek, T., Kopřivová, M., Petkevič, V., and Skoumalová, H. (2018). Variabilita českých frazémů v úzu. Časopis pro moderní filologii, 100(2), pages 151–175.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2021). Korpus SYN, verze 9 z 5. 12. 2021. Ústav Českého národního korpusu FF UK, Praha. Accessible at: https://www.korpus.cz.

Mieder, W. (1994). Paremiological Minimum and Cultural Literacy. In W. Mieder (ed.): Wise Words. Essays on the Proverb. Garland Publishing, pages 297–316.

Mieder, W. (1993): Proverbs are Never Out of Season: Popular Wisdom in the Modern Age. Oxford University Press.

Mokienko, V. M. (2012). Russisches parömiologisches Minimum: Theorie oder Praxis? In K. Steyer (ed.): Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömilogie. Tübingen: Narr Verlag, pages 79–98.

Schindler, F. (1996). Sociolingvistické, paremiologické a paremiografické výsledky empirického výzkumu znalosti přísloví. Slovo a slovesnost, 57(4), pages 264–282.

Schindler, F. (1993). Das Sprichwort im heutigen Tschechischen. Empirische Untersuchung und semantische Beschreibung. Otto Sagner, München.

Steyer, K. (2012). Sprichwortstatus, Frequenz, Musterbildung. Parömiologische Fragen im Lichte korpusmethodischer Empirie. In K. Steyer (ed.): Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömilogie. Narr Verlag, pages 287–314.

Šichová, K., and Šemelík, M. (2016). Was nicht ist, kann noch werden. Zur Parömiodidaktik im Tschechisch-als-Fremdsprache-Unterricht. Teil 1: Eine exemplarische Lehrwerk- und Wörterbuchanalyse. In Beiträge zur Fremdsprachenvermittlung, 57, pages 61–104.

Šichová, K., and Šemelík, M. (2017). Was nicht ist, kann noch werden. Zur Parömiodidaktik im Tschechisch-als-Fremdsprache-Unterricht. Teil 2: Aufgaben und Übungen. In Beiträge zur Fremdsprachenvermittlung, 59, pages 37–76.

Tölgyesi, T. (2022). Deutsch im interlingualen und interkulturellen Vergleich. In S. Szatzker – A. Szilágyi-Kósa (eds.): Hamburg: Dr. Kovač, pages 157–182.

sciendo
DOI 10.2478/jazcas-2023-0028

KEYWORDS IN RELIGIOUS LITERATURE OF 17TH AND 18TH CENTURIES IN LIGHT OF THE DATA FROM THE ELECTRONIC CORPUS OF 17TH- AND 18TH-CENTURY POLISH TEXTS

MAGDALENA MAJDAK
Institute of Polish Language, Polish Academy of Sciences, Warsaw, Poland

MAJDAK, Magdalena: Keywords in Religious Literature of 17th and 18th Centuries in Light of the Data from the Electronic Corpus of 17th- and 18th-century Polish Texts. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 100 – 107.

Abstract: This paper discusses the application of standard keyword extraction methods from corpus linguistics for the study of old Polish language. The unfolding analysis is based on writings included in the Electronic Corpus of 17th- and 18th-century Polish Texts. The aim of this analysis is to select keywords from over two million tokens derived from texts tagged as religion in the corpus and compare them with the reference corpus containing over nine million tokens, while verifying the applicability of the log-likelihood method for the analysis of old Polish language and developing a part of the research model.

Keywords: keywords, keywords in diachronic linguistics, diachronic corpora, Polish language of the 17th and 18th centuries, religious vocabulary

1 INTRODUCTION

This article continues a discussion started several years ago regarding the use of the keyword method for research on the Middle Polish language (Majdak 2016; Majdak 2017). The latter has been largely enabled by the constant evolution of the Electronic Corpus of 17th- and 18th-century Polish Texts (until 1772) (hereinafter: 'corpus' or 'KorBa' from Polish: *Korpus Barokowy* 'Baroque Corpus'), developed at the Institute of Polish Language, Polish Academy of Sciences, with the participation of the Linguistic Engineering Group of the Institute of Computer Science, Polish Academy of Sciences, under the supervision of Włodzimierz Gruszczyński. From 2013 to 2018, a collection of 13.5 million tokens was extracted from texts written and published between 1601 and 1772. The second part of the project, planned for 2019–2023 and currently coming to an end, will result in the corpus doubling its size (up to 25 million tokens) and including records until the end of the 18th century, thus covering two centuries in the history of the Polish language. Our study used the currently available version of the corpus.

The material contained in the corpus seeks to pursue the premise of balance (Adamiec 2015). However, compliance with this requirement is challenging when

selecting contemporary texts and can prove impossible with old material. In the latter case, a two-fold selection process is applied: (1) natural – dictated by historical conditions (only extant sources can be considered); and (2) scientific – based on decisions of contemporary lexicographers. Consequently, the material for the corpus foundation undergoes an (at least) double selection procedure and cannot ensure a complete 1:1 representation of the entire literature from a given period. This results in its double gradation: (1) – actual material; (2) – extant material (that can be found and accessed); and (3) – material selected by a specialist (rather than a collection of texts *in extenso*, the corpus is a selection of samples extracted from them). While KorBa can be considered balanced (or balance-seeking), the question of its representativeness will forever remain open. One could ask what it means for it to be representative. Should it reflect the canons of high or low style, written or colloquial language, dialects or poetry? Should it focus on regional representation, place of origin (publication), 'publishing market' share or a collection of works evoking a real reception, raising awareness and having an actual impact on the intellectual culture of people of the Baroque period?

Access to the collection of 13.5 million tokens provides opportunities for research on the 17th- and 18th-century Polish language material on an unprecedented scale. Nevertheless, it can take us only this much closer to the language and culture of that time. At this stage, conclusions based on the available material, despite the relative extensiveness of the corpus (the largest Polish and Slavic historical corpus thus far), cannot be recognised as reflecting a complete picture of the 17th- and 18th-century Polish literature.

2 RELATED WORKS

Depending on the approach, research based on keywords may have diverse objectives and can focus on searching for words relevant in the output of specific authors (e.g. Zembaty-Michalakowa 1982), works (e.g. Sambor 1969; Rudnicka-Fira 1986), period (e.g. Stachurski 1989), a specific culture and its understanding (Wierzbicka 1997), banner words regarded as the most important terms and determinants of the system of values in a given community (Pisarek 2002), as well as keywords as thematic foci of a given period or community (Williams 1976, et seq.), presented in the form of a dictionary. Studies are also being conducted based on corpora developed for own research purposes, e.g. a corpus of prime ministers' speeches (Kieraś – Zawadzka-Paluektau 2023).[1] The principal aim is essentially to extract words whose frequency of occurrence with regards to a reference set is significantly high (linguistic analyses) or based on the relevance and specificity of a word denoting a crucial cultural category (cultural studies). The first attempts to

[1] For example, corpora developed with the use of Korpusomat, a tool designed by the CLARIN-PL scientific consortium, https://clarin-pl.eu/index.php/en/home/.

use the keyword extraction methods for the study of old Polish language were made several years ago with regards to the political vocabulary from the 17th and 18th centuries (Majdak 2017). This paper is the continuation of that research.

3 OBJECTIVES

Our study had two objectives: (1) to extract keywords from the material consisting of works covered by the corpus and tagged as *religion* (including in the corpus subsets); (2) to verify the applicability of the log-likelihood method for the analysis of old Polish language and to develop a part of the research model.

4 MATERIAL AND SELECTION CRITERIA

The category *religion* was chosen as one of the most represented in the corpus and featuring in all subperiods (halves of centuries).[2] For the purposes of our study, the material was divided into two main parts:

(1) Reference corpus – all texts in the corpus apart from those tagged as *religion* (but including excerpts from the Gdańsk Bible). It covered 639 texts and contained a total of 9,166,861 tokens.[3]

(2) Corpus of religious texts – it covered 91 texts tagged as *religion* and contained a total of 2,064,623 tokens. The 12 excerpts of the Gdańsk Bible were transferred to the reference corpus to ensure that research be conducted on texts embedded in the reality of the 17th and 18th centuries (even if they are translations of older works).

The two-level text search option was used to extract smaller sub-corpora for the material and method testing: (1) thematic, where the category of *religion* was selected; and (2) genre-related (genologic), where search was conducted by category (*factual literature*, *persuasive texts*, *scientific-didactic texts or information and handbooks*), type (*prose*, *poetry*, *drama*), genre (*sermons*, *religious writings*, *nativity plays*, *prayer books*, *catechisms*, *letters*) and author. As a result, we were able to obtain groups of works diversified within the respective classification and select a research sample. The review of the resulting material indicated the need for further differentiation of the general thematic category, perhaps with the use of double determiners, e.g. *political-religious*, *religious-moral*.

Historical background was also considered as an external linguistic criterion – specifically Counter-Reformation as a trend dominating in Polish Baroque literature. This aspect was used for the selection of texts for the third sample where keywords were compared on the example of persuasive texts by Polish military preacher Fabian Birkowski.

[2] For example, the category of *music* is represented by one text in the corpus.

[3] The entire corpus includes precisely 13,445,074 tokens.

5 METHOD

The log-likelihood method was used to extract the keywords. The cut-off point was established at approximately p = 0.0000000001 corresponding to a very high statistical significance (i.e., high 'keyness') to ensure that the list of keywords was neither too long nor cut off arbitrarily. Words that occurred in less than five different texts in entire KorBa were not included in the analysis as containing probable errors, for example, in lemmatisation. The keywords were listed based on entry forms (transcribed lemmata) rather than their forms in texts. Punctuation, foreign words and numerical values were removed from the frequency lists. The log-likelihood method allowed us to assess differences in the frequency of occurrence in both corpora and assign the keyness value to them while considering both 'positive' (significantly more occurrences than in the reference corpus) and 'negative' (significantly fewer occurrences than in the reference corpus) keyness.

6 USAGE EXAMPLES

The list of keywords confirms the effectiveness of the method and its calibration – a set of vocabulary related to religion was obtained. Our analysis of autosemantic words in the corpus of religious texts rendered the following results.

The top 30 nouns included: *God*, *Jesus*, *sin*, *Christ*, *soul*, *(Orthodox) church*, *Satan*, *heaven*, *spirit*, *man*, *faith*, *sacrament*, *word*, *cross*, *witch*, *life*, *angel*, *father*, *world*, *saviour*, *salvation*, *glory*, *body*, *love*, *child*, *scripture*, *the Passion*, *apostle*, *prayer*, *creation*.

The top 30 adjectives included: *saint/holy*, *God's*, *divine*, *heavenly*, *eternal*, *Christlike*, *human*, *of Jesus*, *blessed*, *satanic*, *monastic*, *Christian*, *everlasting*, *beloved*, *Christ's*, *sinful*, *salutary*, *catholic*, *carnal*, *apostolic*, *angelic*, *clerical*, *true*, *pious*, *Lord's*, *eastern*, *worldly*, *just*, *spiritual*, *infernal*.

The top 30 verbs included: *speak*, *create*, *love*, *suffer*, *believe*, *confess*, *say*, *eat*, *save*, *convert*, *desire*, *curse*, *punish*, *receive*, *do*, *pray*, *say*, *be born*, *teach*, *have mercy*, *resurrect*, *imitate*, *lie carnally*, *understand*, *sin*, *condemn*, *celebrate*, *pardon*, *praise*, *crucify*.

The analysis of the first thousand records of occurrences allowed us to identify the following groups: names of divine persons (*God*, *Jesus*, *Christ*, *Saviour*, *Messiah*, *Spirit*, *Trinity*); designations of the Mother of God (*mother*, *Mother of God*, *Virgin Mary*); names of celestial beings and places (*angel*, *Satan*, *devil*, *hell*, *heaven*, *Heavens*); humans and faith (*man*, *creation*, *creature*, *being*, *sinner*, *sin*, *error*, *act*, *deed*, *conscience*, *life*, *death*, *heart*, *gift*, *grace*, *consolation*, *prayer*, *fasting*, *alms*, *exercise*, *punishment*, *suffering*, *cross*, *the Passion*, *sacrament*, *baptism*, *confession*, *penance*, *repentance*, *communion*, *faith*, *will*, *blessing*, *conversion*, *salvation*); names of positive actions, qualities and values

(*love*, *truth*, *mercy*, *goodness*, *perfection*, *reason*, *humility*, *earnestness*, *holiness*, *purity*, *piety*, *lovingkindness*, *wisdom*, *joy*, *gladness*); names of negative actions, qualities and values (*lie*, *ugliness*, *wickedness*, *blasphemy*, *heresy*, *anger*, *obscenity*, *contempt*, *foolishness*, *imperfection*, *vanity*, *wrath*, *fear*); characteristic religious imagery *(shepherd*, *sheep*, *reed*, *bridegroom*, *bride*, *light*, *glory*, *reverence*, *cross*, *thorn*, *supper*, *star*, *neighbour*, *vanity*, *majesty*, *eternity*, *revelation*, *vision*, *child*, *infant*, *baby*, *swaddling cloths*, *son*, *miracle*); names describing believers (*Christian*, *rabbi*, *Muslim*, *Arian*, *Catholic*, *heretic*); church functions (*pope*, *priest*, *Jesuit*, *clergyman*, *parish priest*, *monk*, *minister*, *preacher*); *places of prayer (church*, *Orthodox church*[4]); biblical figures (*apostle*, *prophet*, *patriarch*, *Job*, *Peter*, *Thomas*, *Cain*).

The resulting material could also be organised in a group system, e.g. *saviour*, *saver*, *saved*, *saveable*, whereby the numbers of elements in groups could be considerable.

The material grouping was inspired by different divisions of religious vocabulary to verify which categories and to what extent would be represented, which would dominate or emerge as empty. For various reasons, none of the divisions proved to be entirely adequate. The choice of the classification for the vocabulary organisation deserves a separate study. Referring to the already applied division could contribute to a future comparison study of lexical and semantic fields in the respective historical subperiods in the development of the Polish language. Such a classification could also be used in the *Electronic Dictionary of the 17th- and 18th-century Polish Language*, for which KorBa provides the basis.

The log-likelihood method also allowed us to extract negative keywords that can be arranged in the following thematic categories: political and state-related vocabulary (*army*, *sejm, sejmik*, *election*, *law*, *constitution*); military vocabulary (*camp*, *fortress*, *cannon*); functions (*hetman*, *marshal*, *deputy*, *king*, *prince*, *voievode*, *starosta*, *chancellor*, *castellan*); time division – seasons and months (*winter*, *March*, *September*); food products and spices (*vodka*, *liquor*, *wine*, *beer*, *alcoholic beverage*, *juice*, *salt*, *pepper*, *spices*, *sugar*, *olive oil*, *vinegar*, *peas*); animals (*horse*, *cow*, *hare*); cereals (*grain*, *barley*, *oats*); rivers (*Vistula*); vocabulary related to houses and farmyards (*cottage*, *room*, *barn*, *hallway*, *chamber*, *chimney*, *door*, *porch*, *castle*); measurement units – mass and distance (*mile*, *lot*, *pound*, *ell*); monetary units (*money*, *zloty*,[5] *grosz*, *sterling*, *florin*, *thaler*); specialised vocabulary in mathematics (*line*,[6] *square*, *half*, *triangle*); cities, countries and continents (*Warsaw*, *Poland*,

[4] Among keywords ranked in lower positions there were also *synagogue* and *Afric*.

[5] The Polish term *złoty* can denote a coin (‘zloty‘) and ‘gold’ or ‘golden’. Given that assigning meanings without a context analysis is always a challenging (if not impossible) task, at this stage of our research words were included in categories based on arbitrary decisions.

[6] The noun ‘line’ has many meanings. However, out of 500 example search results in KorBa as many as 454 were derived from the second volume of a geometry textbook by Stanisław Solski (1622–1701).

Republic of Poland, *Kamenets*, *Lviv*, *Vienna*, *Moscow*, *Lithuania*, *Prussia*, *Italy*, *Europe*, *Asia*); nationalities (*Pole*, *Tatar*, *Cossack*, *German*, *Swede*, *French*); administrative division (*country*, *voivodeship*, *city*, *poviat*, *village*); vocabulary related to terrain (*sea*, *land*, *river*, *island*, *mountain*, *field*); kinship, affinity and other relations (*brother*, *wife*, *bachelor*); materials, metals and minerals (*iron*, *coal*, *sulphur*); values (*honour*); weather conditions (*heat*, *humidity*); body parts (*stomach*, *liver*); and surnames (*Potocki*, *Lubomirski*, *Sapieha*, *Chmielnicki*, *Czartoryski*).

The negative keyness can serve as *à rebours* testimony to the importance of words in fields other than religion. It highlights important categories of everyday life, indicating relevant aspects and thematic foci. For example, it reveals a much higher prevalence of nouns denoting values, states, emotions and other abstract concepts in the corpus of religious texts compared to their virtual non-existence in the reference corpus in favour of words describing what lies within sight or related to the organisation of the state and military. These results provide a substantial material for future interpretations.

For the purposes of the material and method testing, two smaller sub-corpora were also analysed: one focused on texts of one genre (*nativity plays*[7]) and the other focused on texts by the same author.

The first set covered 15 texts and featured a total of 41,684 tokens. The use of the log-likelihood method allowed us to extract a typically traditional vocabulary related to Christmas and embedded in the local reality. The following groups of positive key words emerged in the analysis: principal figures (*Jesus*, *Mary*, *Joseph*, *mother*, *Blessed Virgin*, *boy*, *child*); shepherds' names (*Kuba*, *Matys*, *Bartos*, *Wojtek*, *Bartek*, *Janek*, *Wawrzek*); local flavour elements (*flock*, *cattle*, *lamb*, *wolf*, *shepherd*, *shed*, *bag*, *punnet*, *sausage*, *hut*, *innkeeper*, *farmhand*, *master of the house*, *inn*); circumstances of the birth (*manger*, *hay*, *swaddling cloths*, *misery*, *night*, *birth*, *spend the night*); witnesses' reactions (*bow*, *knee*, *gift*, *gospel*, *singing*, *playing*); emotions (*joy*, *gladness*); elements of the supernatural order (*heaven*, *glory*, *angel*); and biblical realities (*Bethlehem*, *Herod*). Illustrating the dynamics of events, the list of verbs started with the activities of shepherds (*sleep*, *talk*, *quarrel*, *chatter*, *trouble*, *play* [*cards and instruments*]), followed by a sequence of actions related to the birth of Jesus from the perspective of shepherds (*be born*, *give birth*, *lie*, *sleep*), their reactions (*hear*, *rise up*, *go*, *follow*, *come*, *greet*, *welcome*, *watch*, *play*, *chant*, *sing*, *long*, *feel*) and references to the spiritual plan (*remit*, *deliver*, *please*). If a group of texts referred to a commonly known story, keywords could be grouped, for example, by elements of action. Key adjectives included the typical folklore references to the birth of Jesus (*beloved*, *poor*, *heavenly*, *little* [*baby*]). In addition to the listed parts of

This finding implies the necessity to verify the corpus for a possible overrepresentation of mathematical texts. If this proves to be the case, the problem will be addressed at later stages of the project.

[7] Also tagged in the corpus as *(Christmas) carols* and *pastorales* or *dialogues*.

speech, another interesting aspect were personal pronouns, indicating the dialogical character of the text (*you* [Singular], *I*, *we*, *you* [Plural]), and pronouns and adverbs (of place and time), highlighting the present character of events (*here*, *already*, *today*, *the present day*). This vocabulary can also be organised into groups, e.g. *sing and singing*; *play and playing*; *give birth*, *birth* and *be born*.

Identified with regards to the reference corpus composed of other religious texts, negative keywords (*Christ*, *faith*, *cross*, *word*, *sin*, *soul*, *saint*, *church*) point to the lack of theological reflection in nativity plays. What emerges from the extracted keywords is the description of the 'stop-motion' nativity scene (*angel*, *kneeling shepherds*, *crib*), typical of the nativity play genre.

The second set, a corpus of works by one author, consisted of four texts by Fabian Birkowski (47,311 tokens), two of which were tagged as religion in the corpus. While not sermons per se, these writings are examples of engaged persuasive literature and Birkowski's style. Compared to the list of keywords for the reference (general) corpus, the differences in the distribution of positive keywords were evident. Word groups of negative emotional language referring to fighting, primarily against heretics, came to the fore, including nouns (*exorbitance*, *tyrant*, *dissenter*, *work*, *crime*, *Luther*, *heretic*, *cross*, *hell*, *cry*, *wound*, *war*, *tombstone*, *Lucifer*, [*wrap-over*] *vest*, *prophet*, *justice*, *janissary*, *sin*, *paganism*, *weight*, *Mehmed*, *freedom*, *apostate*, *condemnation*, *death*, *rope*), verbs (*curse*, *suffer*, *be harmed*, *bellow*, *rebuke*, *reign*, *murder*, *perish*, *liberate*, *disinherit*, *create*, *rout*) and adjectives (*disgraceful*, *cruel*, *heavenly*, *impertinent*, *infernal*, *ungodly*, *devilish*, *double-edged*, *pagan*, *executioner's*). Particularly interesting were negative key nouns identified based on the reference corpus of other religious texts (*Jesus*, *Satan*, [*Orthodox*] *church*, *person*). No other significant autosemantic parts of speech were found.

7 SUMMARY

This study fits into the category of research on both keyword extraction methods and religious vocabulary analysis. It presents a research procedure model using one of the keyword extraction methods for the analysis of material contained in a corpus ofthe old language. Our focus was on the extraction of the dominant lexis in the material composed of works included in the Electronic Corpus of the 17th- and 18th-century Polish Texts (until 1772), tagged as religion also in the corpus subsets. The applicability of the log-likelihood method seems to have been verified, with the study yielding promising results in the form of interesting and extensive material, thus far unstudied in this manner. The method applicability will increase along with the volume of the corpus. Further stages of lexical analyses will provide material for future publications. Certain categories of vocabulary are already emerging with the keyword method, reflecting not only the topics covered in the texts but also the living conditions of people in Poland in the 17th and 18th centuries.

ACKNOWLEDGEMENTS

The research has been supported by the National Programme for the Development of Humanities (NPRH) under the project 'The extending of the Electronic Corpus of 17th- and 18th-century Polish Texts and its integration with the Electronic Dictionary of the 17th- and 18th-century Polish' (no. 0413/NPRH7/H11/86/2018) funded by the Ministry of Science and Higher Education.

I would like to express gratitude to Małgorzate Sobczak for translation.

References

Adamiec, D. (2015). Kryteria doboru tekstów do 'Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)'. Prace Filologiczne, 67, pages 11–20.

Electronic Corpus of 17th- and 18th-century Polish Texts (until 1772). Accessible at: https://korba.edu.pl/korba1.

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., and Woliński, M. (2022). The Electronic Corpus of 17th - and 18th-century Polish Texts. Language Resources and Evaluation, 56, pages 309–332.

Kieraś, W., and Zawadzka-Paluektau, N. (2023). Słowa klucze polskiego dyskursu politycznego na przestrzeni ostatnich stu lat: analiza korpusu exposé premierów (Manuscript submitted for publication).

Majdak, M. (2016). Słowa klucze w materiale historycznym – wyzwania i ograniczenia. Przegląd Humanistyczny, 3, pages 45–56.

Majdak, M. (2017). Słowa ważniejsze niż inne – metoda słów kluczy w badaniu polszczyzny dawnej. Tekst i dyskurs – text und diskurs, 10, pages 229–243.

Pisarek, W. (2002). Polskie słowa sztandarowe i ich publiczność. Kraków: Universitas, 193 p.

Rayson, P., and Potts, A. (2020). Analysing Keyword Lists. In M. Paquot – S. T. Gries (eds.): A Practical Handbook of Corpus Linguistics. Springer, Cham. https://doi.org/10.1007/978-3-030-46216-1_6.

Rudnicka-Fira, E. (1986). Słownictwo 'Dziadów' Adama Mickiewicza w świetle analizy statystycznej (wybór problematyki). Katowice: Uniwersytet Śląski, 142 p.

Sambor, J. (1969). Badania statystyczne nad słownictwem (na materiale 'Pana Tadeusza'). Wrocław: Ossolineum, 163 p.

Stachurski, E. (1998). Słowa-klucze polskiej epiki. Kraków: Wydawnictwo Naukowe WSP, 343 p.

Wierzbicka, A. (1997). Understanding Cultures through Their Key Words. New York: Oxford University Press, 317 p.

Williams, R. (1976). Keywords: A Vocabulary of Culture and Society. London: Croom Helm, 278 p.

Zembaty-Michalakowa, M. (1982). Poezja Juliana Przybosia w świetle badań statystyczno-językowych na tle porównawczym. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, 311 p.

sciendo
DOI 10.2478/jazcas-2023-0029

EXPRESSING MEASURE IN CZECH (A CORPUS-BASED STUDY)

MARIE MIKULOVÁ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

MIKULOVÁ, Marie: Expressing Measure in Czech (A Corpus-based Study). Journal of Linguistics, 2023, Vol. 74, No 1, pp. 108 – 118.

Abstract: In the contribution, we provide a theory-based and corpus-verified description of expressions for measure in Czech. We demonstrate that the measure expressions may modify quantity of entities (*approximately ten boys*), internal characteristics of events (*he works a lot*), properties (*very big*) and relations (*completely without sound*). We distinguish between the measure expressions that are an answer to the question *To what extent?* (Extent-modifiers) and expressions that modify an answer to the question *How many?* (Quantity-modifiers). The Extent-modifiers are formally, structurally and semantically more diverse than the Quantity-modifiers. For the Quantity-modifiers a list of forms and functions is provided. Theoretical knowledge stemming from the analysis will subsequently be used to improve the annotation in the Prague Dependency Treebanks. It can be also useful for other semantically-oriented descriptions of language.

Keywords: measure modifier, syntax, semantic annotation

1 INTRODUCTION

Semantico-syntactic annotation in the Prague Dependency Treebanks (PDT; Hajič et al. 2020a, 2020b) offers broad coverage of semantic phenomena such as semantic role labeling, modalities, discourse, coreference, information structure. However, there is still need for a fine-grained treatment of many categories. The adverbial category is of a special interest, as it covers a wide range of different meanings (temporal, spatial, measure, etc.). These meanings are captured by the *functor* label. However, it is coarse-grained category and a detailed classification is needed for a more precise description of the meaning.

In our previous work we have presented a detailed description of spatial and temporal meanings (Mikulová – Panevová 2021). At present, we focus on refining the classification of the meanings generally related to measure. In the contribution, we propose a formally, syntactically and semantically based analysis of measure expressions verified on a large number of examples from the PDT corpora. Theoretical knowledge stemming from the analysis will subsequently be used to specify and improve the annotation of measure meanings at the PDT tectogrammatical

layer. This effort is in line with the current trend in the field of computational linguistics – to develop language resources annotated semantically.

2 RELATED WORK

Recently, computational linguistics has become increasingly interested in annotation that aims at an adequate description of the meaning. *Measure* is an interesting phenomenon in that effort and has been studied by logicians, theoretical and computational linguists. An approach that combines ideas from formal semantic theories (Barwise – Cooper 1981; Kamp et al. 2011) has been applied in the development of the annotation scheme **QuantML** (Bunt et al. 2022). Under a similar approach, the **Parallel Meaning Bank** (Abzianidze et al. 2017; Bos 2021) is built. It is a project of semantically annotated multilingual corpus, built on the Combinatory Categorial Grammar (Steedman 2001). In the notation of the corpus, for modified numeral expressions (*about 20*) a set of comparison operators to indicate approximate, lower, or higher values is used, e.g. $X \approx Y$ is for approximately equal to, $X < Y$ for less than.

In the framework of the **Uniform Meaning Representation** (Van Gysel et al. 2021), the annotation of the measure phenomenon is adapted from the **Abstract Meaning Representation** project (Baranescu et al. 2013). A measure expression is captured as a separate concept (node in a graph), e.g. a measure modifier of adverbial character (*less powerful*) is captured as a *degree* concept node. UMR also allows to treat *degree* as an attribute of the property concept. It has three values: *equal*, *intensifier* (*very*, *extremely,* superlative form), and *downtoner* (*somewhat*, *relatively,* comparative form). Except for this case, no generalization is made for expressions of the same meaning.

3 MEASURE MEANINGS IN PDT

At the tectogrammatical layer in PDT, the tree-like dependency representation is conceived of as a linguistically structured meaning of the sentence. It contains only content words as its nodes. The types of the semantico-syntactic relations are represented by the *functor* attribute attached to all nodes. Functors represent the semantico-syntactic relations in a generalized way, which does not sufficiently capture more detailed meaningful distinctions. Their fine-grained classification into so-called *subfunctors* is assumed from the early stages of the Functional Generative Description theory within which the project is developed (Panevová 1980). In the current stage of PDT representation, different measure modifiers such as *approximately*, *very*, *completely* have the same value *EXT* (extent) and no fine-grained subcategorization is provided. Illustrative examples of subfunctors were given in Panevová (1980) and Mikulová et al. (2006). However, a comprehensive list of subtle categories has not yet been developed.

4 MEASURE MODIFIERS IN CZECH

The term **measure modifiers** covers expressions primarily of an adverbial nature[1] by which measure meanings are expressed. Based on the large volume of examples from PDT corpora easily available through the ForFun database (Mikulová – Bejček 2018), we analyze the measure modifiers from the formal, syntactic, semantic point of view:

(i) What formal means are used to express measure meanings? (formal aspect)

(ii) What function do the measure modifiers have in a dependency structure? (syntactic aspect)

(iii) What meanings do the measure modifiers express? (semantic aspect)

In the following analysis, we will discuss these aspects in detail.

Within the diverse group of measure modifiers, we distinguish between the measure modifiers that are an answer to the question *To what extent?* and modifiers that modify an answer to the question *How many?* and we divide the modifiers into two groups:[2]

Extent-modifiers modify the internal characteristics of events (*hodně pracuje* 'he works a lot'), properties (*hodně velký* 'very big') and relations (*úplně bez zvuku* 'completely without sound').

Quantity-modifiers relate to counting, they modify an (exact) expression of a number/quantity (*přibližně deset chlapců* 'approximately ten boys').

4.1 Formal point of view

The measure modifiers are primarily expressed by adverbs. However, while the group of adverbs for expressing Quantity-modifiers (*přesně* 'exactly', *téměř* 'almost'; cf. (1) and Tab. 1) is limited, the group of adverbs that modify property, event or relation (Extent-modifiers) is much broader. In addition to the core of set of measure adverbs, which are the same for both groups (cf. *téměř* 'almost' in (1) and (2)), a number of other adverbs perform the function of Extent-modifiers (*trochu* 'a little', *značně* 'greatly' (3), *převážně* 'mostly', *napůl* 'half'). The group of measure adverbs include adverbs with a transparent measure meaning as well as adverbs that are homonymous with intensifying particles (*zcela* 'completely', *vůbec* 'at all'; cf. Šindlerová – Štěpánková 2021). Expressively, adverbs of manner (*hrozně* 'terribly'

[1] Note that expressions of quantity after verbs with the meaning "measurement of some physical property" (*it weighs five kilos*) are obligatory arguments of the respective predicates. In our approach, an expression which is an answer to the question *How many?* is not considered as a measure modifier. As a measure modifier, we only understand a modification of the expressed quantity (*approximately five kilos*).

[2] A similar division can be found in other studies on measure and quantification. Gil (1993) or Keenan (2012) divide measure modifiers according to their character as adverbial or nominal-like. Similarly in Czech grammars, quantitative attribute and adverbial of measure are distinguished. Unlike these divisions based on syntactic behavior our distinction also reflects semantics.

(4), *silně* ‘strongly’) are also used to express a large measure; diminutives (*malinko* ‘very little’ (4), *drobet* ‘very little’) are used to express a small measure. Extent-modifiers are further expressed by prepositional phrases with the nouns *míra* ‘measure’ (5), *výše* ‘amount’, *část* ‘part’, etc. Exact measure can be also expressed in percentages or as a part of the whole (6). Some idiomatic phrases also carry measure meanings (*do puntíku* ‘on the dot’ (7), *celým srdcem* ‘wholeheartedly’). Their usage is limited to certain semantic contexts.[3]

(1) *Náklady činily **téměř** miliardu korun.* ‘Payload was **almost** a milliard of crowns.’
(2) *Situace je **téměř** neřešitelná.* ‘The situation is **almost** unsolvable.’
(3) *Firma byla **značně** oslabena.* ‘The firm was **considerably** weakened.’
(4) *Maminka uměla **hrozně malinko** česky.* ‘Mother spoke **terribly little** Czech.’
(5) *Podnikání má zlaté dno jen **do určité míry**.* ‘Making business is a gold-mine only **to a certain extent**.’
(6) *Akcie s hodnotou 0.5 je nestabilní **z poloviny**, s hodnotou 1.5 **ze 30 %**.* ‘The share with the value of 0.5 is not stable **by half**, with 1.5 **by 30%**.’
(7) *Splnili to **do puntíku**.* ‘They fulfilled it **on the dot**.’

One of the differences between Extent-modifiers and Quantity-modifiers is that the Quantity-modifiers are not only expressed lexically (by a content word), but also using preposition (cf. (8) vs. (9)). E.g. an approximate number is expressed using the preposition *kolem*+2 ‘about’ (9), the quantity “less than the given value including the given value” by the preposition *do*+2 ‘up to’ (10). The meaning “less than the given value” and “more than the given value” is also expressed using the phrase with *než* ‘than’ (11).

(8) *Doba čekání je **zhruba** čtyři roky.* ‘The waiting time is **roughly** four years.’
(9) *Doba čekání je **kolem** čtyř let.* ‘The waiting time is **about** four years.’
(10) *cestující **do** 99 let* ‘passengers **up to** 99 years of age’
(11) v*ýsledek **více než** desetileté práce* ‘the result of **more than** ten years of work’

4.2 Syntactic point of view

Measure modifiers are very interesting especially from a syntactic point of view. The Extent-modifiers roughly correspond to the traditional description of adverbials, i.e., adjuncts that may modify (depend on) a verb (3), adjective (2) or adverb (4). However, from our data analysis, it follows that also the entity itself (expressed by a noun) can be modified to some extent, cf. (12)–(13).

(12) *Je to opravdu **téměř** cestopis.* ‘It is really **almost** a book of travels.’
(13) *Jsem jí **víc než** sestra.* ‘I am **more than** a sister to her.’

[3] A measure modification of nominalized events is expressed using adjectives. Cf.:
(a) *Most nebyl **úplně** uzavřen.* ’The bridge has not been **completely** closed.’
(b) *Nedošlo k **úplnému** uzavření mostu.* ’There was no **complete** closure of the bridge.’

Moreover, the various circumstantial relations expressed by prepositions may be modified themselves to some extent. In (14) the temporal relation "since when" is modified, in (15) the spatial relation "outside" and in (16)–(18) other special types of circumstances are modified.[4]

(14) *Vyrůstal jsem zde **skoro** od narození.* 'I grew up here **almost** from birth.'
(15) *To je **úplně** mimo ves.* 'It is **completely** outside the village.'
(16) *Povedlo se to **částečně** kvůli spojenectví.* 'It succeeded partly because of the alliance.'
(17) *Ceny skončily **téměř** beze změn.* 'The price ended **almost** without a change.'
(18) *Pracuji **převážně** podle plánu.* 'I work **mostly** according to the plan.'

The Quantity-modifiers primarily modify the (exact) expression of the number of entities, i.e., the so-called quantitative attribute (Kopečný 1953) expressed by the cardinal numeral in a noun phrases (1). However, the Quantity-modifiers generally modify expressions of various parts of speech that include a number. This means that they modify all types of numerals: ordinal numerals which are of an adjective nature (19), multiplicative numerals of an adverb nature (20). They further modify compound words, one part of which is a number, typically compound adjectives such as *šestiměsíční* 'six-month' (21), and verbs with meaning "to make X-fold" (*zdvojnásobit* 'double' (22)).

(19) *Dostal se tam **zhruba** třetí den.* 'He got there **approximately** the third day.'
(20) *Zvýšil se **více než** desetinásobně.* 'It went up **more than** by ten times.'
(21) ***maximálně** šestiměsíční vězení* '**maximally** a six-month imprisonment'
(22) *Podpora se **téměř** zdvojnásobila.* 'The support has **almost** doubled.'

Consequently, it is difficult to design an adequate syntactic representation of measure modifiers within the traditional dependency syntax. How to capture that measure expressions in (14)–(18) modify the meaning of the preposition (relation) and not the noun (its internal characteristics) within the same prepositional phrase? Similarly, it is difficult to capture the difference between cases in which a measure expression modifies the entire situation expressed by a verb (*vzorky převážně ztmavly* 'samples mostly darkened') and cases in which it modifies only an internal characteristics of the verb (*nepatrně ztmavly* 'samples darkened slightly'). And how to capture a modification of only part of a word as in (20)–(22)?

4.3 Semantic point of view

Our analysis of measure meanings is based on the assumption that there is no one-to-one relation between meaning (represented by the functor-subfunctor

[4] In the domain of temporal and spatial relations, the measure modification of a relation can be expressed very explicitly (*20 days before the loan*). Here, we leave this issue aside. We also do not consider the cases where measure modification is expressed as a consequence (*a profit so big that no one expected it*), by difference (*profit greater by two percent*), or by comparison (*profit greater than last year*).

combination in PDT) and the way it is expressed in the text. One meaning can be expressed by several different means whereas one form can be used to express different functions. Analyzing fine-grained meanings, we apply the principle of substitutability of forms. As for adverbs (a typical means of measure modifiers), it is not very common for them to have different meanings. Only some adverbs express more measure meanings: *docela* carries the meaning "completely" and is substitutable with *zcela*, *naprosto* 'completely' in (23) and it has the measure-certain meaning "to a certain extent" (cf. Vondráček 1999) and is substitutable with *poměrně* 'relatively' in (24).

(23) *Má to **docela** jinou chuť.* 'It has a **completely** different taste.'

(24) *Anglicky umím **docela** dobře.* 'I know English **relatively** well.'

Meaning	**Forms**	**Examples**
exact exactly the given value	*právě akorát* *přesně*	Trvalo mi to ***akorát*** pět minut. 'It took me **just** five minutes.' Do konce mu chyběla ***přesně*** půlhodina. 'He was **exactly** half an hour away from the end.'
approx approximately the given value	okolo+2 kolem+2	***okolo*** 1000 litrů vína '**around** 1,000 litres of wine' Doba čekání je ***kolem*** čtyř let. 'The waiting time is **around** four years.'
	přibližně zhruba asi tak řádově *plus minus nějaký* *cirka cca víceméně*	Jezdím ***tak*** jednou za měsíc. 'I go **about** once a month.' To jsem já ***plus mínus*** před pěti lety. 'This is me **plus minus** five years ago.' Víno dělám ***cirka*** dvacet let. 'I have been making wine for **approximately** twenty years.' Bylo mi ***nějakých*** šest let. 'I was **about** six years old.'
almost the given value	*téměř skoro takřka* *bezmála málem*	Odešli po ***takřka*** pěti desítkách let. 'They left after **almost** five decades.' ***bezmála*** pět metrů dlouhý vůz '**almost** five meter long car'
less than the given value	pod+4	Inflace klesla ***pod*** dvě procenta. 'Inflation fell **below** two percent.'
	méně než	To je ***méně než*** 4.6 %. 'This is **less than** 4.6%.'
more than the given value	nad+4 přes+4	vedra ***nad*** 40 stupňů Celsia 'heat **above** 40 degrees Celsius' Zajel rekordní kolo ***přes*** 228 km/h. 'He drove a record lap of **over** 228 km/h.'
	více než *pryč*	***více než*** dvojnásobný čistý zisk '**more than** double net profit' Bylo mi osmnáct ***pryč***. 'I was eighteen **gone**.'

Meaning	Forms	Examples
lessincl less than the value incl. the value	(až) do+2 (až) k+3	cestující ***do*** 99 let 'passengers **up to** 99 years of age' Obchody stoupají ***až k*** 145 jenům. 'Deals go **up to** ¥145.'
	maximálně *nejvýše až*	***maximálně*** šestiměsíční vězení '**a maximum of** six months' imprisonment' Je možné převážet ***nejvýše*** dva tisíce korun. 'It is possible to transport **a maximum of** 2,000 crowns.'
moreincl more than the value incl. the value	od+2	Auta jsou k dostání ***od*** 15000 dolarů. 'Cars are available **starting at** $15,000.'
	minimálně *nejméně aspoň* *přinejmenším*	Byl ***minimálně*** dvakrát obětí loupeže. 'He has been robbed **at least** twice.' Je sledován ***přinejmenším*** třemi analytiky. 'He is checked by **at least** three analysts.' Všichni vlastní ***alespoň*** jedno kolo. 'Everyone owns **at least** one bike.'
total	*celkem celkově*	***Celkem*** pětkrát změnili bydliště. 'They changed their place of residence five times **in total**.'
average	*průměrně v průměru*	Stojí to ***průměrně*** 200 korun. 'It costs **an average** of 200 crowns.'
order	*řádově*	průměr jádra je ***řádově*** 10^{-12} cm 'the diameter of the nucleus is **of the order** of 10^{-12} cm' (Cf. Hladiš 1965)

Tab. 1. Measure meanings of Quantity-modifiers

Different formal means and various measure meanings can be also combined and "layered" one on another; cf. (25)–(26).

(25) ***tak asi okolo*** *1000 litrů vína* '**approximately about** 1,000 litres of wine'

(26) *Sehrají* ***minimálně asi*** *dvacet her.* 'They play **minimally about** twenty matches.'

The means that differ stylistically or belonging to a different variety of contemporary Czech (*od – vod* 'from', *velmi* 'very'– *děsně* 'terribly') are considered here as expressions of the same meaning.

Based on the analysis, it is clear that the Extent-modifiers are semantically more diverse than the Quantity-modifiers. Within the Quantity-modifiers, there is a relatively limited repertory of formal means for a limited number of meanings. Particular modifiers place the expressed number on the number line with varying degrees of accuracy. Each partial meaning can be expressed by a set of adverbs, or in the case of a quantified noun phrase, also by means of prepositions. See Tab. 1 for an overview of the measure meanings of Quantity-modifiers.

Meaning	Forms	Examples
large measure	*velmi velice hodně dost moc značně výrazně vysoce hojně ohromně nesmírně mohutně silně mocně notně hrozně příšerně pěkně ve velkém ve velké/značné/hojné míře*	Uveřejňování článků je ***nesmírně*** důležité. 'Article posting is **very** important.' tento ***vysoce*** zajímavý jev 'this **highly** interesting phenomenon' Na staveništi se ***mohutně*** finišovalo. 'They finished **strongly** on the construction site.' Na škole se ***mocně*** krade. 'There is **a lot of** stealing at school.' Je ***pěkně*** vystresovaná. 'She's **pretty** stressed out.' ***V hojné míře*** kouří cigarety. 'He **heavily** smokes cigarettes.'
extreme measure	*krajně extrémně enormně radikálně diametrálně do krajnosti*	Taková věc je ***krajně*** nepříjemná. 'This is **extremely** annoying.' enormně předražené byty '**enormously** overpriced apartments' ***do krajnosti*** vybičovaná inscenace 'a production whipped **to the extreme**'
atmost the largest possible measure	*maximálně nanejvýš úplně zcela naprosto navýsost docela veskrze zhola absolutně totálně nadobro nadevše stoprocentně v plné míře na sto procent*	Situace je ***maximálně*** obtížná. 'The situation is **the most** difficult.' Jezdí se ***nanejvýš*** ohleduplně. 'It is driven with **utmost** consideration.' Žvýkání je činností ***veskrze*** lidskou. 'Chewing is a **completely** human activity.' ***Totálně*** se odvrátila od principů. 'She has **completely** turned away from principles.' ***stoprocentně*** spolehlivé opatření '**100%** reliable measure' Zákrok byl ***v plné míře*** adekvátní. 'The intervention was **fully** adequate.'
oversized measure	*příliš nadměrně nadmíru přes míru přehnaně v nadměrné míře*	Nezkušení kluci ***nadměrně*** riskují. 'Inexperienced guys take **excessive** risks.' Pijí ***přes míru***. 'They drink **excessively**.' Nevystavujte se ***v nadměrné míře*** slunci! 'Do not **over**expose yourself to the sun!'

Tab. 2. Selected meanings of Extent-modifiers

A comprehensive description of the meanings within the group of Extent-modifiers is difficult. Even leaving aside the modification of circumstantial relations ((14)–(18) above), we are still dealing with a wide variety of formal means and diverse meanings; cf. (27)–(29).

(27) *Jsou vypouštěny **stoupající měrou**.* 'They are released **in an increasing extent**.'

(28) *Pracovali **na 83.7 % kapacity**.* 'They worked **at 83,7% of capacity**.'

(29) *Pijí **přes míru**.* 'They drink **beyond measure**.'

In our analysis, we have focused on the modifiers expressed by an adverb. A detailed list with forms and examples for the selected meanings is given in Tab. 2. An overall overview of the basic set of Extent-modifiers meanings is summarized in Tab. 3.

5 CONCLUSION

In the contribution, we propose a formally, syntactically, and semantically based fine-grained classification of measure expressions verified on the analysis of a large number of examples from the Prague Dependency Treebanks. The analysis of the material reveals the diversity of structures and their relative difficulty in classification. We demonstrate that the measure modifiers involve modifying the quantity of entities (*about ten boys*), extent of property expressed by adjectives (*very big*), adverbs (*very slowly*), verbs (*work a lot*), nouns (*more than sister*), and extent of relation expressed by preposition (*almost without sound*). Within this diverse group, we distinguish between the measure modifiers that are an answer to the question *To what extent* (Extent-modifiers) and modifiers that modify an answer to the question *How many* (Quantity-modifiers). The Extent-modifiers are formally, structurally and semantically more diverse than the Quantity-modifiers. For the Quantity-modifiers the comprehensive description of their meanings and forms is provided.

Meaning		**Examples**
exact	exact measure	***akorát*** úspěšný '**just** successful'
approx	approximate measure	***víceméně*** úspěšný '**more or less** successful'
large	large measure	***velmi*** úspěšný '**very** successful'
extreme	extreme measure	***extrémně*** úspěšný '**extremely** successful'
small	small measure	***trochu*** úspěšný '**slightly** successful'
atmost	the largest possible measure	***maximálně*** úspěšný '**maximally** successful'
atleast	the smallest possible measure	***minimálně*** úspěšný '**at least** successful'
over	oversized measure	***nadmíru*** úspěšný '**overly** successful'
approp	appropriate measure	***přiměřeně*** úspěšný '**adequately** successful'
nonapp	non-appropriate measure	***nepřiměřeně*** úspěšný '**disproportionately** successful'
almost	almost appropriate measure	***skoro*** úspěšný '**almost** successful'
prev	prevailing measure	***vesměs*** úspěšný **'mostly successful'**
part	partial measure	***částečně*** úspěšný '**partially** successful'
enough	sufficient measure	***dostatečně*** úspěšný 'successful **enough**'

Meaning		Examples
notenough	insufficient measure	***nedostatečně*** úspěšný '**insufficiently** successful'
relat	relative measure	***relativně*** úspěšný '**relatively** successful'
half	half measure	***napůl*** úspěšný '**half** successful'
medium	medium measure	***středně*** úspěšný '**moderately** successful'
total	total measure	***celkově*** úspěšný '**overall** successful'
average	average measure	***průměrně*** úspěšný '**averagely** successful'
X-multiple	X-multiple measure	***mnohonásobně*** úspěšný '**many times** successful'

Tab. 3. Measure meanings of Extent-modifiers (a preliminary proposal)

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA23-05238S. This research also has used language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

References

Abzianidze, L. et al. (2017). The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, pages 242–247.

Baranescu, L. et al. (2013). Abstract Meaning Representation for Sembanking. In Proceedings of the 7th Linguistic Annotation Workshop, Sophia, pages 178–186.

Barwise, J., and Cooper, R. (1981). Generalized Quantifiers and Natural Language. Linguistics and Philosophy, 4, pages 159–219.

Bos, J. (2021): Quantification Annotation in Discourse Representation. In The 17th Joint ACL – ISO Workshop on Interoperable Semantic Annotation, Groningen.

Bunt, H. et al. (2022). Quantification Annotation in ISO 24617-12. In Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, pages 3407–3416.

Gil, D. (1993). Nominal and verbal quantification. Sprachtypologie und Universalienforschung, 46.4, Berlin: Walter de Gruyter, pages 275–317.

Gysel van, J. E. L. et al. (2021). Designing a Uniform Meaning Representation for Natural Language Processing. Künstl Intell 35, Springer-Verlag, pages 343–360.

Hajič, J. et al. (2020a). Prague Dependency Treebank – Consolidated 1.0. In Proceedings of the 12th International Conference on Language Resources and Evaluation, Marseille, pages 5208–5218.

Hajič, J. et al. (2020b). Prague Dependency Treebank – Consolidated 1.0. LINDAT-CLARIAH, Prague: Charles University. Accessible at: http://hdl.handle.net/11234/1-3185.

Hladiš, F. (1956). K významu a užití příslovce řádově. Naše řeč, 48(4), pages 254–256.

Kamp, H. et al. (2011). Discourse Representation Theory. In Handbook of Philosophical Logic, 15, Elsevier, MIT, pages 125–394.

Keenan, E. L. (2012). The Quantifier Questionnaire. In Handbook of Quantifiers in Natural Language. Studies in Linguistics and Philosophy, 90, Dordrecht: Springer.

Kopečný, F. (1953). Kvantitativní přívlastek a určení míry. Slovo a slovesnost, 14(3), pages 115–121.

Mikulová, M. et al. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical report, 2006/30, Prague: Charles University.

Mikulová, M., and Bejček E. (2017). Prague Database of Forms and Functions 1.0. LINDAT/CLARIAH-CZ, Prague: Charles University. Accessible at: http://hdl.handle.net/11234/1-2542.

Mikulová, M., and Panevová, J. (2021). Formy a funkce okolnostních určení v češtině. Určení prostorová a časová. Prague: Charles University.

Panevová, J. (1980). Formy a funkce ve stavbě české věty. Prague: Academia.

Steedman, M. (2001). The Syntactic Process. Cambridge: The MIT Press.

Šindlerová, J., and Štěpánková, B. (2021). Between Adverbs and Particles: A Corpus Study of Selected Intensifiers. Jazykovedný časopis, 72(2), pages 444–453.

Vondráček, M. (1999). Příslovce a částice – hranice slovního druhu. Naše řeč, 82(2), pages 72–78.

sciendo
DOI 10.2478/jazcas-2023-0030

ADVERBS DERIVED FROM ADJECTIVAL PRESENT PARTICIPLES IN POLISH, SLOVAK AND CZECH: A COMPARATIVE CORPUS-BASED STUDY

AKSANA SCHILLOVÁ
Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic

SCHILLOVÁ, Aksana: Adverbs Derived from Adjectival Present Participles in Polish, Slovak and Czech: A Comparative Corpus-based Study. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 119 – 129.

Abstract: The paper investigates and compares the inventory of adverbs derived from adjectival present participles in the Polish, Slovak and Czech languages. The particularity of these adverbs is that they do not occur in every language, even if it concerns closely related languages. While in Polish and Slovak this type of adverbs is represented by hundreds of lemmas, in Czech it is almost not represented. The comparative analysis is carried out on the data retrieved from the comparable web corpora Aranea. The sets of adverbs extracted from the comparable corpora of the languages examined are analysed by the following criteria: the total number of the adverb lemmas in the corpus, their relative frequency (ipm), morphemic structure features, collocability preferences. The similarities and differences between the adverb sets are established. According to the corpus data, the adverbs derived from adjectival present participles are more widely used in Polish than in Slovak, and in Czech they are a rare phenomenon represented by a limited number of lemmas with a negligible frequency.

Keywords: adverb, adjectival present participle, word-formation, comparable corpora, Aranea

1 INTRODUCTION

The paper focuses on one word-formation type of adverbs which is found in Polish and Slovak (in some other languages as well),[1] but is almost not found in Czech, namely, on adverbs derived from adjectival present participles. From the formal point of view, adverbs with verbal origin ending in *-ąco* in Polish and *-úco*/*-aco* in Slovak are investigated, mainly in the context of restrictions in the formation of similar adverbs in Czech, even though these languages are closely related to each other.

The study is inspired by new publications based on parallel corpora and devoted to the topics of lexical gaps (Perišić 2020), language-specific words (Shmelev 2020), or non-equivalent constructions (Ďurčo – Braxatorisová 2022), viewed from an interlingual perspective.

[1] For example, in Russian (Shmeleva 2016) or English (Killie 2022).

The present comparative study is based on comparable corpora and have the following purposes: a) to extract from the corpora, analyse and compare the sets of Polish, Slovak and Czech adverbs of the type under discussion, b) to establish similarities and differences between these adverb sets, c) based on the data provided by the corpora, to evaluate the usage of this type of adverbs in the languages compared.

2 THEORETICAL BACKGROUND

2.1 Adverbs derived from adjectival present participles in Polish and Slovak

According to Kisiel (2018, p. 193), the adverbs derived from adjectival present participles have not received much attention by Polish language researchers. In (Kisiel 2018), the adverbs ending in *-ąco* are distinguished and discussed due to the function of some of them as the so-called "metapredicative operators" (see also Danielewiczowa 2012; Grochowski 2018), or alternatively due to the undergoing process of the metatextualization of these adverbs, which may be presented as follows: verb → participle → adjective → adverb → metapredicative operator. Moreover, the ability to derive an adverb is considered as a criterion for the adjectivization of a corresponding participle (ibidem).

As regards the Slovak language, the adverbs of the type examined in this paper are also not sufficiently investigated. In the new Slovak word-formation dictionary (Ološtiak et al. 2021), the adverbs ending in *-úco*/*-aco* are not distinguished within the transpositional type of deadjectival qualitative adverbs ending in *-o* and are exampled by one lemma (ibidem, p. 406). In the authoritative Morphology of the Slovak Language (Dvonč et al. 1966), the above adverbs are presented as qualitative adverbs derived from adjectival present participles and are exampled by 28 lemmas (ibidem, p. 580–581). In other grammars, these adverbs are only mentioned and represented by single examples.

In the present paper, it is assumed that the adverbs derived from adjectival present participles are worth special attention, already for the reason that they are not forming in every language, even if it concerns closely related languages.

2.2 Adverbs derived from adjectival present participles as impossible words in Czech

In the Czech linguistic literature, the impossibility of forming adverbs from adjectival present participles[2] ending in *-cí* is either briefly mentioned as a fact (e.g. Štícha et al. 2018, p. 1080), or additionally commented on with the remark that in

[2] In the Czech linguistic tradition, the lexical items ending in *-cí*, that are considered here as (adjectival) present participles, are classified as deverbal adjectives, see the review of different national linguistic interpretations of these words by Kocková (2022, p. 29f.).

Czech there are very few adverbs which can be considered as formed from the adjectival present participles, namely by expanding a base stem by *-n-,* such as *vroucí* 'boiling' – *vroucně* 'fervidly' (Komárek 2006, p. 41; Knappová 1973, p. 16; Stich 1969, p. 63).

In older papers, the forming adverbs according to this model was sharply criticized as incorrect, see, e.g. on the adverb *obdivujícně* 'admiringly' by Zubatý (1920, p. 274). Cf. Grepl and Karlík (1995, p. 219) regard adverbs such as *vyhovujícně* 'suitably' as non-literary.

Nevertheless, these adverbs are sporadically formed by Czech native speakers. So, the database of excerpt material named Neomat,[3] that is storing new Czech lexical items from the 1990s to the present, currently lists more than ten lemmas of adverbs of the above type.

3 DATA AND METHODOLOGY

In the present study, the comparable web corpora Aranea (Benko 2014)[4] are used to collect and compare data about the use of adverbs examined in Polish, Slovak and Czech. These are the following corpora:

Araneum Polonicum Minus (Polish, 15.02) 119 M,

Araneum Slovacum VI Minus Beta (Slovak, 22.01) 125 M,

Araneum Bohemicum IV Minus (Czech, 20.03) 125 M.

The data from the corpora are collected by using the following steps (exemplified by the Polish language):

1) with the CQL query [atag="Av"&word=".*ąco"] a concordance where the adverbs ending in *-ąco* appear is retrieved from the Polish corpus;

2) using the option Frequency (Make frequency list) a frequency list of the adverbs is made;

3) the set of the adverbs presented in the frequency list is thereafter analysed by the following criteria: a total number of the adverb lemmas in the corpus, their relative frequency (ipm), morphemic structure features, collocability preferences.

The Slovak and Czech language data are collected based on the same methodology.

The analysis of the data is supported by online lexicographic resources of the respective languages, such as the Polish Academy of Sciences Great Dictionary of Polish (https://wsjp.pl/), the Dictionary portal of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences (https://slovniky.juls.savba.sk/), the DEBDict General Dictionary Browser that offers direct access to various Czech dictionaries (https://deb.fi.muni.cz/).

[3] https://neologismy.cz

[4] http://unesco.uniba.sk/.

4 RESULTS

4.1 The Polish and Slovak adverbs derived from adjectival present participles, according to the data from the 100-million-word Aranea corpora

The adverb frequency lists extracted from the minus versions of the Araneum Polonicum and Araneum Slovacum corpora (of about 100M words) contain the following number of items: the Polish adverbs ending in *-ąco*, 478; the Slovak adverbs ending in *-úco*, 304.

The Polish and Slovak adverbs ending in *-ąco* and *-úco*, respectively, do not have homonyms in other word classes, therefore, these lists really consist only of the adverbs of this type that occur in the corpora. In this case, the probability of incorrect disambiguation is minimal. Nevertheless, the output number of lemmas in the frequency lists is negligibly distorted because it includes some duplicate lemmas with typos or spelling errors.

In the subsequent quantitative analysis, the presence of misspellings in the lists is ignored. All percentages presented below are rounded to the nearest integer. As illustrated, direct equivalents from the adverbs lists extracted from the corpora are provided.

4.1.1 Similarities

The Polish and Slovak sets of the adverbs investigated have the following common features:

1) Both sets consist of the adverbs with either native or foreign roots, moreover the ratio of these roots in both sets is the same, see Tab. 1.

Root origin	Polish adverbs			Slovak adverbs		
	Number of lemmas	Share in the set, %	Example	Number of lemmas	Share in the set, %	Example
native	379	79	*odstraszająco* 'frighteningly'	240	79	*odstrašujúco* 'frighteningly'
foreign	99	21	*inspirująco* 'inspiringly'	64	21	*inšpirujúco* 'inspiringly'

Tab. 1. The ratio between native and foriegn roots in the adverb sets

2) Both sets include adverbs with negation prefixes, which are moreover directly equivalent to each other, see Tab. 2.

Negation prefixes	Polish adverbs			Slovak adverbs		
	Number of lemmas	Share in the set, %	Example	Number of lemmas	Share in the set, %	Example
Pol. *nie-*, Slov. *ne-* 'non'	27	6	*niewystarczająco* 'insufficiently'	14	7	*nedostačujúco* 'insufficiently'
Pol. *znie-*, Slov. *zne-* 'dis'	3		*zniechęcająco* 'discouragingly'	7		*znechucujúco* 'discouragingly'

Tab. 2. Advers with negation prefixes in the adverb sets

3) Both sets include adverbs with compound stems, see Tab. 3.

Stem type	Polish adverbs			Slovak adverbs		
	Number of lemmas	Share in the set, %	Example	Number of lemmas	Share in the set, %	Example
Compound stem	9	2	*lekceważąco* 'disparagingly'	16	5	*srdcervúco* 'heartbreakingly'

Tab. 3. Adverbs with compound stems in the adverb sets

4) The adverbs in both sets have a similar collocability as well as collocates (verbal, adjectival, or adverbial) that are mostly equivalent, see Tab. 4, Tab. 5. The most frequent verbal collocate of these adverbs is the verb with the meaning 'to act', namely Polish *działać*, Slovak *pôsobiť*. This corresponds to their initial function as adverbs of manner. Further, approximately one-third of both Polish and Slovak adverbs examined is used to modify adjectives.

Top verbal collocates		Number of adverb lemmas collocated	Share in the adverb set, %	Examples
Polish	*działać* 'to act'	181	39	*(działać) kojąco, pobudzająco, uspokajająco* '(to act) soothingly, encouragingly, calmingly'
	wyglądać 'to look '	66	14	*(wyglądać) imponująco, olśniewająco, interesująco* '(to look) impressively, dazzlingly, attractively'
	brzmieć 'to sound'	44	9	*(brzmieć) zachęcająco, interesująco, kusząco* '(to sound) friendly, attractively, temptingly'
Slovak	*pôsobiť* 'to act'	127	42	*(pôsobiť) upokojujúco, osviežujúco, povzbudzujúco* '(to act) calmingly, refreshingly, encouragingly'
	vyzerať 'to look'	25	8	*(vyzerať) vynikajúco, očarujúco, ohromujúco* '(to look) great, charmingly, stunningly '
	znieť 'to sound'	22	7	*(znieť) prekvapujúco, odstrašujúco, šokujúco* '(to sound) surprisingly, frighteningly, shockingly'

Tab. 4. Top verbal collocates (position: -5 to +5) of the adverbs

Top adjectival and adverbial collocates		Number of adverb lemmas collocated	Share in the adverb set, %	Examples
Polish	Adjectives such as *niski* 'low', *wysoki* 'high', *dobry* 'good', *duży* 'big', *piękny* 'nice'	122	26	*Potrzebną wiedzę możesz zdobyć po* ***zaskakująco niskich*** *cenach.* 'You can get the knowledge you need at surprisingly low prices.'
	Adverbs such as *dobrze* 'well', *szybko* 'quickly'	64	14	*Satysfakcjonująca praca sprawia, że* ***zadziwiająco dobrze*** *się trzymają.* 'A satisfying job makes them hold up surprisingly well.'

Top adjectival and adverbial collocates		Number of adverb lemmas collocated	Share in the adverb set, %	Examples
Slovak	Adjectives such as *vysoký* 'high', *nízky* 'low', *dobrý* 'good', *veľký* 'big', *krásny* 'nice'	87	29	*Kvalitné laminátové podlahy sú* ***vynikajúco odolné*** *tak proti vode ako aj opotrebeniu.* 'High-quality laminate floors are excellently resistant to both water and wear.'
	Adverbs such as *dobre* 'well', *rýchlo* 'quickly'	17	6	*Medvieďatá* ***prekvapujúco dobre*** *šplhajú.* 'Bear cubs climb surprisingly well.'

Tab. 5. Top adjectival and adverbial collocates (position: +1) of the adverbs

4.1.2 Differences

Although the sets of the Polish and Slovak adverbs show many similarities, there are also important differences between them.

1) The Polish adverb set includes several lemmas with a high frequency, compared to the frequency of the other lemmas in the set. There are no such high-frequency lemmas in the Slovak set, moreover, in general, the Slovak adverbs of the type discussed have a lower frequency than the respective Polish adverbs, see Tab. 6.

Rank	Polish adverbs		Slovak adverbs	
	Item	ipm	Item	ipm
1	*gorąco*[5] 'hot' or 'fervidly'	3413[6]	*vynikajúco* 'excellently'	6.94
2	*bieżąco* 'constantly'	31.10[7]	*prekvapujúco* 'surprisingly'	4.70
3	*znacząco* 'significantly'	19.56	*horúco*[8] 'hot' or 'fervidly'	2.86
4	*wystarczająco* 'enough'	18.60	*upokojujúco* 'soothingly'	1.15
5	*następująco* 'as follows'	8.62	*šokujúco* 'shokingly'	0.97
…	all the other adverbs in the set	from 5 to 0.01	all the other adverbs in the set	from 0.40 to 0.01

Tab. 6. The most frequent Polish and Slovak adverbs derived from adjectival present participles, according to the comparable corpora

[5] The Polish lemma *gorąco* is used as an adverb of manner as well as predicative adverb and noun, see https://wsjp.pl/.

[6] Including occurrences as a part of the compound adverb *na gorąco* '(serve food) hot' or '(decide) immediately', ipm 2.52.

[7] Including occurrences as a part of the compound adverb *na bieżąco* 'constantly', ipm 30.37.

[8] The Slovak lemma *horúco* is used as adverb of manner as well as predicative adverb, see https://slovniky.juls.savba.sk/.

2) The Polish adverb set includes several lemmas which are mainly used as a part of the compound adverb of the type *na* + adverb ending in *-ąco*,[9] see Tab. 7. Similar adverbs do not occur in the Slovak set.

Item	ipm	Share in the total number of the adverb occurrences, %
na bieżąco 'constantly'	30.37	98
na stojąco 'in a standing position'	1.91	97
na siedząco 'in a sitting position'	0.59	92
na leżąco 'in a lying position'	0.50	97
na klęcząco 'in a kneeling position'	0.08	91

Tab. 7. Polish compound adverbs of the type *na* + adverb ending in *-ąco*

3) The Polish set includes more pairs/groups of adverbs with the same roots, as contrasted with the Slovak set, see Tab. 8.

	The Polish set	The Slovak set
Number of pairs/groups of cognate adverbs	55	20
Share in the adverb set, %	12	7
Examples	*lśniąco* 'glisteningly', *olśniewająco* 'dazzlingly'; *oczyszczająco, przeczyszczająco, czyszcząco* 'in a cleansing manner'; *ochładzająco, schładzająco, wychładzająco, chłodząco* 'coolingly'	*očisťujúco, prečisťujúco* 'in a cleansing manner'; *vysvetľujúco* 'explainingly', *samovysvetľujúco* 'self-explainingly', *zosvetľujúco* 'in a brightening manner'

Tab. 8. Cognate adverbs in the Polish and Slovak adverb sets

4.1.3 Slovak adverbs ending in *-aco*

As for as the Slovak adverbs ending in *-aco*, only 8 relevant but low-frequency lemmas are found in the Slovak corpus. These are the following:

neveriaco 'unbelievingly', 0.24 per million;

tesniaco 'sealingly', *tlmiaco* 'in a dampening manner', *brzdiaco* 'as a brake', 0.02 per million;

páliaco 'stingingly', *ničnehovoriaco* 'unintelligibly', *lahodiaco* 'pleasingly', *hodnotiaco* 'appraisingly', 0.01 per million.

All the above adverbs have their origin in verbs of the pattern *robiť* 'to do' – *robím* 'I do' (Dvonč et al. 1966, p. 455). Due to the fact that in a 100-million-word corpus these

[9] The semantics of these adverbs was examined in the case study by Kościerzyńska (2011).

adverbs are presented by only a few lemmas with an insignificant frequency, while in the same corpus the mentioned verb pattern is presented by more than 30,000 lemmas[10] and the respective present participles are presented by about 1,500 lemmas,[11] it can be concluded that the Slovak adverbs ending in *-aco* are on the periphery of the adverb word-formation type under examination.

4.2 The Czech adverbs with verbal origin ending in *-cně*, according to the data from the 100-million-word Araneum Bohemicum corpus

In the Czech corpus, a total of 41 relevant adverbs are found. These are low frequency lemmas, most of which (34 lemmas, i.e., 83% of the set) have from 4 occurrences to 1 occurrence in the 100-million-word corpus (ipm=0.03–0.01). Nevertheless, several of these adverbs, including 7 lemmas with ipm > 0.03, are found in the Czech dictionaries (see the DEBDict portal mentioned above), such as presented in Tab. 9.

Item	ipm
nevěřícně ‘unbelievingly’	3.23
vroucně ‘fervidly’	0.56
nebojácně ‘fearlessly’	0.44
kajícně ‘contritely’	0.25
bojácně ‘fearfully’	0.19
horoucně ‘fervidly’	0.10
vědoucně ‘knowingly’	0.06
budoucně ‘in the future’, *přejícně* ‘kindly’, *srdcervoucně* ‘heartbreakingly’, *nežádoucně* ‘undesirably’, *vševědoucně* ‘omnisciently’	0.02
žádoucně ‘desirably’, *živoucně* ‘livingly’, *nepřejícně* ‘unkindly’, *nemohoucně* ‘powerlessly’	0.01

Tab. 9. The Czech adverbs ending in *-cně* extracted from the corpus which are represented in dictionaries

Some of the adverbs extracted from the corpus are already registered in the Neomat database, such as:

vynikajícně ‘excellently’, 0.03 per million;

matoucně ‘confusingly’, *ucházejícně* ‘acceptably’, *zarážejícně* ‘surprisingly’, *zneklidňujícně* ‘worryingly’, *dechberoucně* ‘breathtakingly’, 0.01 per million.

[10] See Araneum Slovacum VI Minus Beta (Slovak, 22.01) 125 M, CQL query: [atag="Vb"&lemma=".*it"], lemma frequency list.

[11] See Araneum Slovacum VI Minus Beta (Slovak, 22.01) 125 M, CQL query: [lemma=".*iaci"], lemma frequency list.

All the other adverbs (19 lemmas, i.e., 46% of the set) that appear only in the corpus are formed according to the same pattern, such as *zavádějící* 'misleading' – *zavádějícně* 'misleadingly', 0.03 per million.

If these adverbs are considered as a whole, that is, as a set of adverbs representing a certain word-formation type, then it can be noticed that there are many similarities between the Czech set and the sets of the Polish and Slovak adverbs discussed above, see Tab. 10 (compare with Tab. 1, 2, 3, and 8). However, the Czech set has a small representation of lemmas most of which are felt by native speakers as deviant.

Attribute		Presence of the attribute	Example
Root origin	native	+	*ponižujícně* 'humiliatingly'
	foreign	+	*šokujícně* 'shockingly'
Negation prefixes	*ne-* 'non'	+	*neutichajícně* 'unremittingly'
	zne- 'dis'	+	*znechucujícně* 'discouragingly'
Compound stem		+	*všeříkajícně* 'tellingly'
Cognate adverbs in the set		+	*vědoucně* 'knowingly', *vševědoucně* 'omnisciently'

Tab. 10. Attributes that the Czech adverb set shares with the Polish and Slovak adverb sets

5 CONCLUSION

Thus, the data from the comparable corpora show that the adverbs derived from adjectival present participles have a wider usage in the Polish language, as compared to the Slovak language, especially because a) they are represented by a larger number of lemmas that in general have a higher frequency in the comparable corpus than the respective Slovak adverbs; b) several of the Polish adverbs have undergone the processes of lexicalization (high-frequency lemmas), grammaticalization (multi-word adverbs *na* + adverb ending in *–ąco*) as well as metatextualization which has already been given special attention in the Polish linguistic literature (see Section 2.1). The Czech adverbs that can be considered as formed from adjectival present participles occur as a rare phenomenon in the corpus, although they as a whole share main formal similarity with the Polish and Slovak adverbs.

ACKNOWLEDGEMENTS

The preparation of this article was financed within the statutory activity of the Czech Language Institute of the Czech Academy of Sciences (RVO No. 68378092).

References

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka et al. (eds.): Text, Speech and Dialogue 2014, LNAI 8655. Springer International Publishing, pages 257–264.

Danielewiczowa, M. (2012). W głąb specjalizacji znaczeń: przysłówkowe metapredykaty atestacyjne. Warszawa.

Databáze excerpčního materiálu Neomat (2015). [online]. Verze 3.0. Praha: Oddělení současné lexikologie a lexikografie Ústavu pro jazyk český AV ČR. Accessible at: https://neologismy.cz.

Ďurčo, P., and Braxatorisová, A. (2022). Übersetzungsmöglichkeiten der deutschen äquivalenzlosen Konstruktion Präp + SubAbstr+ sein im Slowakischen am Beispiel von *von Bedeutung / von Belang / von Relevanz sein*. Korpus – gramatika – axiologie, 26, pages 3–17.

Dvonč, L. et al. (1966). Morfológia slovenského jazyka. Bratislava: Vydavateľstvo Slovenskej akadémie vied.

Grepl, M., and Karlík, P. (1995). Příruční mluvnice češtiny. Praha: Nakladatelství Lidové noviny.

Grochowski, M. (2018). Operatory metapredykatywne otwierające pozycję dla komparatywu. Prace Filologiczne, 72, pages 59–70.

Horak, A., Pala, K., and Rambousek, A. (2008). The Global WordNet Grid Software Design. In Proceedings of the Fourth Global WordNet Conference. Szeged: University of Szeged, pages 194–199.

Killie, K. (2022). Extravagance, productivity and the development of -ingly adverbs. In M. Eitelmann – D. Haumann (eds.): Extravagant Morphology: Studies in Rule-Bending, Pattern-extending and Theory-challenging Morphology, John Benjamins Publishing Company, pages 51–72.

Kisiel, A. (2018). O metapredykatywnej funkcji niektórych przysłówków odimiesłowowych zakończonych na -ąco w strukturze tekstu. In A. Dobaczewski et al. (eds.): Sens i konwencje w języku. Toruń: Wydawnictwo Naukowe UMK, pages 193–207. Accessible at: https://www.academia.edu/37101805/.

Knappová, M. (1973). K tvoření příslovcí z přídavných jmen. Naše řeč 56(1), pages 11–18.

Kocková, J. (2022). Neurčité tvary slovesné v češtině, ruštině a němčině a jejich vzájemná ekvivalence. Praha: Academia.

Komárek, M. (2006). Příspěvky k české morfologii. Olomouc: Periplum.

Kościerzyńska, J. (2011). Charakterystyka semantyczna przysłówków typu *na stojąco*, *na leżąco*, *na siedząco*. Linguistica Copernicana, 51(15), pages 153–177.

Ološtiak, M. et al. (2021). Slovník slovotvorných prostriedkov v slovenčine. Prešov: Prešovská univerzita v Prešove.

Perišić, O. (2020). Translating lexical gaps: A contrastive corpus-based analysis. In M. Matešić – A. Memišević (eds.): Language and mind: proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society. Berlin: Peter Lang, pages 93–108.

Shmelev, A. (2020). Russian language specific words in the light of parallel corpora. In Z. Ye – H. Bromhead (eds.): Meaning, life and culture. Acton: Australian National University Press, pages 403–419.

Shmeleva, T. (2016). Otprichastnye narechija v jazyke media. In T. Ignatovich – J. Biktimirova (eds.): Jazyk v razlichnykh sferakh kommunikacii. Chita: Zabajjkalskijj gosudarstvennyjj universitet, pages 147–149.

Slovenské slovníky [Slovak dictionaries] (the dictionary portal of the Ľ. Štúr Institute of Linguistics). Accessible at: https://slovniky.juls.savba.sk.

Stich, A. (1969). Stupňování přídavného jména vzrušující. Naše řeč, 52(1), pages 62–64.

Štícha, F. et al. (2018). Velká akademická gramatika spisovné češtiny. Vol. I (2). Praha: Academia.

Żmigrodzki, P. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. In J. Čibej et al. (eds.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, pages 209–219.

Zubatý, J. (1920). Obdivujícně. Naše řeč, 94(49), pages 274–275.

sciendo
DOI 10.2478/jazcas-2023-0031

THE EPISTEMIC MARKER *URČITĚ* IN THE LIGHT OF CORPUS DATA

BARBORA ŠTĚPÁNKOVÁ[1] – JANA ŠINDLEROVÁ[2]
– LUCIE POLÁKOVÁ[1]
[1] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[2] Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

ŠTĚPÁNKOVÁ, Barbora – ŠINDLEROVÁ, Jana – POLÁKOVÁ, Lucie: The Epistemic Marker *určitě* in the Light of Corpus Data. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 130 – 139.

Abstract: The paper presents a pilot study for a research project on epistemic modality and/or evidentiality markers in Czech. The study focuses on the expression *určitě.* Although, this marker is typically considered to signal high certainty, the dictionary of standard Czech (Slovník spisovné češtiny, SSČ) also offers an alternative meaning of *probability*, indicating a lower degree of certainty. We use parallel data from the InterCorp v15 corpus to determine whether the probability meaning can be identified unequivocally in real language data and whether it correlates with specific translation equivalents, linguistic features, or lexical context. Based on our findings, we propose an alternative method for distinguishing between different shades of meaning based on the communicative functions of the utterances, and we draw conclusions regarding the relevance of individual grammatical and lexical clues in context for future annotations.

Keywords: epistemic markers, Czech, annotation, communicative functions, parallel corpus

1 INTRODUCTION

This paper presents a pilot study for a research project aimed at mapping the semantic space of Czech epistemic and evidential markers. One of the central issues we aim to address is how to approach the description of the meaning of epistemic/evidential markers.

Typically, the meaning[1] of markers expressing epistemic modality[2] is described in terms of a scale (Grepl – Nekula 2017) that covers different degrees of certainty, ranging from the highest certainty (paraphrased as *it is certain that x*) to the lowest

[1] For epistemic markers, we use the term *meaning* in the sense of the term *relational meaning* (cf. Filipec – Čermák 1985, p. 39).

[2] I.e., the ways of expressing the degree of certainty about a propositional meaning (cf. Grepl – Šimík 2017).

degree of certainty (e.g. *hardly*), with medium degrees covering a broad area from probability to possibility (cf. Kořenský et al. 1986, p. 233). While few epistemic particles are semantically unambiguous, most are described as "epistemically underspecified", which means they can be used to signal different degrees of probability, making the interpretation of a text difficult and highly subjective (Ivanová 2017, pp. 242–243).

According to Grepl and Nekula (2017), the Czech word *určitě*[3] is ranked third on the scale of certainty, i.e., even higher than other expressions usually considered representative of high certainty, such as *jistě* 'for sure', or *nepochybně* 'undoubtedly'. However, the most recent complete and representative dictionary of standard Czech, SSČ, splits the meaning of *určitě* into two separate meanings that express different degrees of certainty. The first meaning listed expresses a high degree of certainty (synonymous with *jistě*), while the second meaning, marked as *hovor.* (colloquial), expresses probability, i.e., a medium degree of certainty (synonymous with *patrně* 'apparently'):

SSČ[4]

určitě část. modál. 'modal particle'

1. *vyj. jistotu, jistě:* zítra určitě přijdu; je to určitě on '*expresses certainty, definitely*: I will definitely come tomorrow; it's definitely him'

2. hovor. *vyj. pravděpodobnost, patrně:* určitě na to zapomenu, připomeň mi to 'colloq. *expresses probability, likely*: I'm likely to forget, do remind me'

The tendency to define epistemic modal markers solely based on the degree of certainty they express is also evident in the most recent attempt for a comprehensive dictionary of standard Czech (ASSČ), cf. *bezpochyby*.

Although, the latest lexicon definition of *určitě* distinguishes two separate meanings that claim a difference in the degree of certainty, the example sentence provided for the probability meaning does not offer enough clues to allow for its identification in real data. Our assumption is that the need to distinguish two (or more) meanings of *určitě* does not reflect only the different degree of certainty perceived, but also other linguistic features of the utterance and its context. These features then might be important for the constitution of individual meanings of the epistemic marker.

We pose the following research questions:

1. Can the medium-certainty meaning be distinguished from the full-certainty meaning in the corpus data:

[3] The most common dictionary equivalents of this expression are *surely*, *certainly*, *definitely*, but other translations often appear in specific contexts, see Tab. 1.

[4] Apart from the word class categorization as a modal particle, SSČ also offers *určitě* as an adverb, nested to the adjectival entry *určitý* 'certain'.

a) by comparing the translation equivalents of *určitě* in parallel data? Do translators in any context use medium certainty expressions (e.g. *probably*) as an equivalent to the original data?
b) by the mere native speaker annotator judgment provided for the Czech side of the data?

2. What aspects of an utterance support an unequivocal interpretation of the meaning of *určitě* and can thus be considered key points in defining the meaning of *určitě*?
3. What aspects of an utterance can explain the differences in the meaning of *určitě* and are thus important to be captured in future annotations?

To answer these questions, we have designed a small-scale annotation experiment in which three expert annotators judged the status of *určitě* (or, more precisely, the status of the whole utterance modified by *určitě*) and searched for selected features in its context.

The design of the experiment is described in Section 2. Section 3 discusses the findings with 3.5 presenting an alternative approach to describing the meaning of epistemic markers, namely their communicative functions. Section 4 concludes and offers suggestions for our future work.

2 DATA AND METHODOLOGY

Our corpus study follows the assumption of Aijmer et al. (2006) that comparing data from parallel corpora may be helpful in determining the fine-grained meaning of pragmatic markers, especially because contextual factors become apparent. The interpretation of pragmatic markers is often difficult even for native speakers, due to the natural underspecification of the core meaning in the process of pragmaticalization. The translation process involves looking for functional correspondences within the given context, rather than pure lexical equivalents, providing thus an additional clue for the interpretation.

For this study, we use the Czech (Rosen et al. 2022) and English (Klégr et al. 2022) data of the parallel InterCorp v15 corpus. We use the core part of InterCorp, which consists mainly of fiction. These fictional texts, in contrast to the other parts of InterCorp (subtitles, parliamentary proceedings), represent a suitable basis for our research, as they are presumed to be close to spoken language, e.g. in terms of high frequencies of epistemic markers, a certain degree of subjectivity and elaborate situational context. It also can be assumed that literary translations of fictional texts would be of reliable quality.

For a sample of 300 concordances with Czech as the source language we provided an annotation of the type of meaning of *určitě*. First, we assessed whether the meaning was epistemic or something else (see Section 3). Second, for the epistemic markers, we annotated the degree of certainty perceived (see Section 3.2)

and, additionally, other linguistic features of the utterance that could potentially affect the interpretation of epistemic markers. The repertoire of the annotated features was compiled based on similar previous research (cf. e.g. Pietrandrea 2018, Wiemer – Stathi 2010). The following attributes were captured:

1. lexical environment – predicate lemma, another modal expression in near context (intensifiers, modal markers, modal verbs),
2. grammatical categories of the predicate verb – mood and tense,
3. word order and topic-focus articulation – position within a sentence, scope,
4. annotators' comments – reflections on/arguments for the decision made plus potential other observations.

Separately,[5] we manually identified the English translation equivalents and compared whether the translators tend to use epistemic markers with the degree of certainty corresponding to the annotator intuition about the original (see Section 3.1).

3 ANALYSIS

Out of the 300 instances of *určitě* in this annotation task, 238 were agreed on by all three annotators as signalling epistemic meaning. The remaining 62 cases fall into one of the three non-epistemic categories, adverbs, affirmative use, and other (e.g. irony). As we are mainly interested in the epistemic types of use, these are not further analyzed in this study.

3.1 Translation equivalents

For the 238 epistemic instances of *určitě* described earlier, we manually identified English lexical translation equivalents, if present. The distribution of those with frequency 4 and higher is given in Tab. 1.

The equivalents fall mainly among expressions typically indicating a high degree of certainty in English,[6] translations indicating probability are rare; in 30 cases, there was no equivalent at all. The analysis of the translation equivalents therefore does **not** suggest that there is a discrete meaning with medium certainty in the data.

English equivalent of *určitě*	Frequency
be bound to	14
be sure	48

[5] In order not to be influenced by the translation equivalents.

[6] This claim is supported by an additional annotation of 750 cs > en and 750 en > cs parallel utterances containing *určitě* by a single annotator. Apart from using *probably* as the English translation equivalent (13x) and translating original *probably* to *určitě* (8x) in the opposite direction, there were just 2 instances of other medium-certainty markers (*perhaps*) in the original English data.

English equivalent of *určitě*	Frequency
bet	6
certainly (almost certainly)	39 (5)
definitely	31
for sure	7
must	13
no doubt	6
probably	**4**
surely	11
no equivalent	30

Tab. 1. The most frequent English translation equivalents of *určitě*

3.2 Annotation of the degrees of certainty

For the subsequent annotation tasks, the subset of 238 concordances detected as epistemic readings of *určitě* served as a basis. In this annotation task, the degree of certainty conveyed by the word *určitě* was assessed. We annotated the data according to the meaning split in SSČ (see Section 1), as either expressing high certainty or medium certainty, with a third possibility being "unknown". Given the assumption regarding the degree of certainty of *určitě* articulated above, one might expect frequent disagreements in this respect. The resulting figures (Tab. 2) show that the annotators preferred the high certainty interpretation; the 103 cases agreed on by all three annotators stand for 43% of the data. In contrast, there was almost no agreement (3 cases) on the medium certainty interpretation. The annotations therefore suggest that, in line with expectations, *určitě* is largely perceived as a marker of the high degree of certainty, especially if also the high agreement of just two annotators is considered.

Type of agreement	Count
high certainty (3 agreed)	103
high certainty (2 of 3 agreed)	94
medium certainty (3 agreed)	3
medium certainty (2 of 3 agreed)	3
no agreement	35
total	238

Tab. 2. Agreement of three annotators on the degrees of certainty in the epistemic subset of *určitě*

This claim is further corroborated when we compare the annotators' decision and the interpretation of the translators. As Tab. 3 demonstrates, out of the 103 cases

agreed upon by all three annotators as high-certainty markers, the chosen translation equivalents are in the vast majority (90 cases) also expressions typically perceived as high-certainty markers.[7]

Label	3 annotators agreed	3 annotators and translations agreed	Disagreed with translations	No translation equivalent
high certainty	103	90	3	11
medium certainty	3	2	0	1

Tab. 3. Agreement of 3 annotators and the translators' choice of expression on the degrees of certainty in the epistemic subset of *určitě* in cs > en parallel data

On the other hand, the overall 66% annotator disagreement, with 35 cases even exhibiting complete disagreement (among all three annotators), indicates that there must be other perspectives to consider for the semantic interpretation of these markers (see Section 3.5).

3.3 Relevant interpretation features

The high level of disagreement on the degree of certainty for epistemic uses of *určitě* indicates a general high inconsistency in the perception of features relevant to a given annotation decision. In individual cases, however, specific context patterns or collocations can be discerned, which facilitate the interpretation of a given use.

For this analysis, we used the annotation of the attributes described in Section 2 in order to identify distinctive patterns and clues. The following linguistic features appear to be the most conclusive in our data (instances agreed on by all three annotators):

a) Typically, *určitě* can be considered a high-certainty marker in a **contrastive** relation to another epistemic marker with a different (lower) degree of certainty (1).

(1) *Vyplňuje celou čtvrtou komoru,* ***určitě*** *utlačuje mozeček a* ***možná*** *i kmen.*

'It's filled the whole fourth ventricle, **certainly** it's pressing on the cerebellum and **probably** on the stem too.'

b) High-certainty interpretation is strengthened by the use of **another epistemic signal** with a high degree of certainty, typically the verb *muset* 'must'. In (2), the original Czech utterance combines two high-certainty markers, which is mostly a redundancy. The English translation drops one of them.[8]

[7] See footnote 6 and Tab. 1 for details on English high-certainty markers.

[8] These cases are not to be confused with a co-occurence of *určitě* with an objective meaning of *must* (necessity, duty): *Důstojně odchází,* ***určitě*** *se musí převléci a umýti.* 'He strides off in a dignified manner, but he **must be** in need of a wash and a change of clothes.'

(2) ***Určitě*** *se* ***musela*** *podivit, že je celkem čistá.*

'She **must**'ve been amazed how clean she was.'

In contrast, a co-occurring epistemic signal of a medium certainty weakens the degree of certainty (e.g. *asi* 'might'). In the English translation of (3), by analogy, a medium-certainty marker ('might') is used.

(3) *To právě myslela ty moje potíže v hlavě a ty já teda* ***asi určitě*** *mám.*

'She was thinking about the troubles in my head and they **might** well be something that I do have.'[9]

c) The presence of an **amplifier**, e.g. *zcela, docela* (*fully, completely, quite*) usually confirms the high-certainty interpretation (4).

(4) *Musím si prokousnout ret, protože tentokrát bych se už* ***docela určitě*** *rozbrečela.*

'I had to bite my lip because this time I would **definitely** have burst into tears.' Lit: '...I would **quite surely** have burst...'

On the contrary, the presence of a **downtoner** weakens the degree of certainty (5).

(5) *Hovořil s holohlavým suchým mužíkem,* ***skoro určitě*** *to byl její otec, ani toho jsem živého nikdy neviděl…*

'He was talking to a bald-headed, wizened old man, **almost certainly** her father, whom I hadn't met either.'

Apart from the above mentioned contextual clues, grammatical features (such as sentence-initial position or future tense) also seem to show some interesting tendencies, nevertheless, larger data would be needed to make reliable claims about their impact.

3.4 Analysis of annotators' comments

Based on the annotators' notes, a number of findings have emerged that significantly affect the interpretation of texts but have not yet been systematically addressed here. These findings relate primarily to the areas of semantics and pragmatics.

a) There were several utterances in the data in which the combination of a predicate verb's meaning, past tense, and situational context elicited the reader's expectation of the absence of epistemic marking or the presence of an epistemic marker indicating lower certainty. For example, in (6) the use of a high certainty marker sounds illogical and raises doubts about the interpretation of the utterance.

(6) ***Určitě*** *jsem se do ní zamiloval.* '**I'm sure** I fell in love with her.'

b) The annotators experienced difficulties when the perspectives of the author and the character in the text were blended. For the judgment, it was crucial to identify unequivocally who was evaluating the certainty of the statement, see (7). Note that the English translation interprets the perspective clearly.

[9] One of the three examples in total uniformly annotated as medium certainty.

(7) *...byla ráda, že ten den nemusí vůbec nic důležitého nebo potřebného dělat, protože ten den by **určitě** zkazila cokoli, co by vyžadovalo byť i jen kapku soustředění.*

'...she was glad she had nothing important she needed to do that day, since **she was sure** she would have made a mess of anything that required even a drop of concentration.'

3.5 Communicative functions

The annotation of the degree of certainty did not appear conclusive considering the task of differentiating between distinct uses of the epistemic marker *určitě*. Nevertheless, the annotators were able to intuitively group similar uses on a different basis. It appears that the distinctive feature for describing the differences in use is the intention with which the speaker produces the utterance, i.e., its communicative function. This is not quite surprising since communicative functions are frequently mentioned in connection to epistemic modality (cf. Pietradrea 2018).

For *určitě*, the following types of communicative functions seem relevant[10]:

– assertive functions which can be specified as **assumptions**, regarding past, present or future events.

(8) *Ano, určitě tu je daleko víc lidí mladších než já.*

'Yes, definitely there's a lot more people younger than me around.'

Here the epistemic marker signalizes that the utterance is not presented as an objective fact, but as a (strong) personal belief of the speaker.

(self-)encouragement

(9) *Určitě se vše vyjasní.* 'I'm sure it'll all get cleared up.'

By using *určitě*, the speaker aims at convincing him/herself or the audience about a positive outcome of the situation.

worry

(10) *Sdělila otci, že Věrušku ušpiněná ručka stále svědí, stále si ji umývá a určitě dostane nějaký ekzém.*

'She told Father that her darling little Vera's soiled hand wouldn't stop itching, she kept washing it and was bound to get eczema.'

Here the author expresses negative predictions about the future with the secondary intention to make the audience prevent them.

accusations

(11) *Mládí v hajzlu, revoluce v prdeli, rektor furt v Americe a ty mi určitě s někým zahejbáš!*

'Youth down the tubes, revolution up shit's creek, the dean still in America, and you are definitely cheating on me with someone else!'

– commissive functions which can be specified as **promises.**

(12) *Až to začne jít, dám ti určitě vědět.*

'As soon as it starts to go anywhere, I'll be sure to let you know.'

[10] Here we follow the terminology and classification of Grepl et al. 1995, p. 585ff.

The speaker expresses his (strong) intentions to carry on some action in the future. *Určitě* serves for emphasis on the promise here.

- directive functions which can be specified as **persuasion.**

(13) *Jen se neostýchejte, určitě si chcete sochu podrobně prohlédnout.*

'There's no need to be shy. I am sure you would like to look at the statue more closely.'

In the persuasive contexts, *určitě* is a means of expressing polite urge on someone to act according to the intentions of the speaker.

As can be seen from this list, the degree of certainty is a dominant element of meaning in case of assumptions, while in other cases, it is overshadowed by other functions, e.g. with adding emphasis (fears, promises) or politeness (persuasion) to the statement. We believe that it is exactly the presence of another dominant aspect in the meaning that led the SSČ authors to add the medium certainty meaning. Their example *Určitě na to zapomenu, připomeň mi to* is a manifestation of worry with an explicit request to the partner to prevent the negative outcome.

4 CONCLUSION

Based on the annotations and subsequent analyses, the addressed research questions can be answered as follows:

Ad 1) We were able to document the postulated medium degree of certainty of *určitě* only very rarely in the studied dataset. Similarly, only two translation equivalents with medium certainty in the parallel English translations have been detected (*probably* and *perhaps*) both with low frequencies. The absolute majority of translation equivalents of the epistemic marker *určitě* express a high degree of certainty. In line with this, three annotators agreed on high-certainty meaning of *určitě* in 103 corpus instances, but only on 3 instances on medium-certainty meaning.

Ad 2) As for the aspects of an utterance facilitating the interpretation of the meaning of *určitě,* a distinctive role has been observed in the co-occurrence of *určitě* and other epistemic markers or intensifiers within an utterance – these expressions, in compliance with their own semantics, affect the degree of certainty in the given utterance. Another impactful factor are contrastive or comparative syntactic patterns with epistemic markers where different degrees of certainty are juxtaposed. Further relevant clues tend to appear in combinations, e.g. sentence-initial position and future tense. An analysis of such combinations is subject to more elaborate research, for which we intend to apply statistical methods and tools.

Ad 3) As suggested by the findings in 1 above, the core of the lexical meaning of *určitě* is high certainty. There is not enough evidence in the studied data to support the hypothesis on the existence of an individual meaning of medium certainty.

The analysis of the degree of certainty of the epistemic marker *určitě* has proven to be a less appropriate tool for a deeper description of the meanings and

various fine-grained functions of this expression in a text. It turns out that a more informative method of description, mapping the actual breadth of its use, is an analysis of the communicative functions that this expression may have in different contexts. However, the reliability of such a pragmatic approach also remains to be verified by empirical corpus experiments (i.e. annotation and annotator agreement).

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA23-05240S.

References

Aijmer, K. et al. (2006). Pragmatic markers in translation: a methodological proposal. In K. Fischer (ed.): Approaches to discourse particles. Oxford: Elsevier, pages 101–114.

Akademický slovník současné češtiny (2017–2023). [online]. Praha: ÚJČ AV ČR (ASSČ).

Filipec, J., and Čermák, F. (1985). Česká lexikologie. Praha: Academia, 284 p.

Filipec, J. et al. (2003). Slovník spisovné češtiny pro školu a veřejnost. Praha: Academia. (SSČ).

Grepl, M. et al. (1995). Příruční mluvnice češtiny. Praha: Nakladatelství Lidové noviny.

Grepl, M., and Nekula, M. (2017). Epistémická částice. In CzechEncy – NESČ.

Grepl, M., and Šimík, R. (2017). Epistémická modalita. In CzechEncy – NESČ.

Ivanová, M. (2017). Modálnosť a modálne verbá v slovenčine. Prešov: Filozofická fakulta Prešovskej univerzity, 310 p.

Klégr, A. et al. (2022). Korpus InterCorp – angličtina, verze 14 z 31. 1. 2022. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Kořenský, J. et al. (1986). Mluvnice češtiny 2. Praha: Academia, 536 p.

Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken Italian dialogues. Journal of Pragmatics, 129, pages 171–191.

Rosen, A. et al. (2022). Korpus InterCorp – čeština, verze 14 ze 31. 1. 2022. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Wiemer, B., and Stathi, K. (2010). The database of evidential markers in European languages. A bird's eye view of the conception of the database. Language Typology and Universals, 63, pages 275–289.

sciendo
DOI 10.2478/jazcas-2023-0032

COMPARATIVE LEXICAL ANALYSIS OF NOUN LEMMAS IN SLOVAK JUDICIAL DECISIONS

MIROSLAV ZUMRÍK
Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

ZUMRÍK, Miroslav: Comparative Lexical Analysis of Noun Lemmas in Slovak Judicial Decisions. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 140 – 149.

Abstract: The paper presents a comparative lexical analysis of the most frequent noun lemmas in Slovak judicial decisions. The data was taken from the large corpus of decisions provided by the Ministry of Justice of the Slovak Republic, and compared with lemmas from four other corpora. The style of administrative, legal and other highly formalized texts has been receiving attention in recent years, although research into the style of Slovak judicial decisions remains rather sparse, which could partially stem from their idiosyncratic nature. The paper's aim is thus to focus on the quantitative and qualitative characteristics of noun lemmas found specifically in the style of judicial decisions.

Keywords: judicial decisions, nouns, stylistics, legal linguistics

1 INTRODUCTION

Judicial decisions are a peculiar genre of administrative/legal texts. According to František Štícha (1985, pp. 69–71), the peculiarity of decisions in criminal cases in particular is due to the fact they regularly contain depictions of awkward or extreme life situations that are otherwise not subject to discussion in formal, official settings. Because of that, the language of judicial decisions tends to suffer from stylistic flaws. These shortcomings, as Štícha suggests, result among others from the specific circumstances in which the documents are produced (such as time pressure) and from the very nature of administrative texts. The style of judicial decisions is then characterized by a discrepancy between the elevated legal language and the often mundane content described using this language, overly complicated and prolonged sentences, the use of archaic expressions, as well as the overuse of explicitness and noun phrases. Jozef Mistrík, similarly, describes the style of administrative texts (including judicial decisions) as rather stereotypical, lexically and syntactically poor, simple in composition, yet simultaneously rich in genres, and these texts represent the least researched style in standard Slovak (1997, pp. 458–459).

General claims like the statements above are too vague to be testable, so they need to be transformed into more specific ones (Chlumská 2017, p. 53). In this paper, I will therefore examine the semantic and distributional features of the most frequent

noun lemmas in Slovak judicial decisions, as compared to lemmas in Slovak acts, specialized studies, short stories and texts from a balanced corpus. I will ask the following questions:

1. how numerous and how diverse are the most frequent noun lemmas in the decisions?
2. how do they differ from lemmas in other genres?
3. how could the different numbers and differences be explained?

2 THEORETICAL FRAMEWORK

Accounts on stylistic features of Slovak administrative texts, including noun lexis, have been given by Jozef Mistrík (1975; 1997) or Ján Findra (2004) and more recently by Daniela Slančová et al. (2022). Linguistic interest in legal language, termed "legal linguistics"(as the term is used in Vogel 2019 or Cvrček 2016) follows from the natural connection between language and law: "(l) laws are coded in language, and the processes of the law are mediated through language" (Gibbons 1999, p. 156). The style of legal texts has also been studied in Slovakia (Abrahámová 2012; Slančová et al. 2022, pp. 669–712), focusing mostly on the language of legal acts (Čulenová 2010; Imrichová – Turočeková 2015) or legal terminology, as discussed, for example, with respect to individual Slovak legal terms in shorter journal contributions by Rudolf Kuchár throughout the second half of the 20th century.

The main rationale for choosing the texts of judicial decisions for analysis is the fact that they are produced on a massive scale, judging by their the numbers issued by Slovak courts alone, and thus "all around us", so to say. The digital archive at the Ministry of Justice of the Slovak Republic[1] contains 4,1 M anonymized judicial decisions, with tens of new decisions added every day.

The motivation for focusing on nouns is that this part of speech, together with verbs, represents a "major part of speech" (Chlumská 2017, p. 58), capable of denoting every thinkable aspect of reality, both static, dynamic, physical or abstract (Ružička et al. 1966, p. 62). In this paper, I work with a customized set of semantic categories, which resulted from the most frequent noun lemmas found in the analysed corpora: abstract entity, action, institution, norm, object, person, physical entity, place, proper name, text-related entity, time. At the same time, I am aware that the boundary between abstract and physical objects can be blurred (Cvrček et al. 2020, p. 51). It is also to be noted that an analysis of a single stylistic feature and a singular, one-word expression can prove challenging, since stylistic layers are systematically interconnected (Mistrík 1997, p. 34), representing mutually communicating vessels.

[1] https://obcan.justice.sk/infosud/-/infosud/zoznam/rozhodnutie

3 METHOD

The data from the Slovak corpora was sized down to genre-specific sub-corpora. From these, random samples of 1 M tokens were extracted. The samples were described as to their composition and analysed for the relative frequency and number of lemmas. The minimal frequency threshold for a lemma was set at 5 occurrences. The frequency distribution of lemmas for each genre was visualized; the samples were tested for significant differences in variance. The lemmas were then described quantitatively and semantically. Differences in the relative frequency of lemmas found in several corpora were tested for statistical significance.

4 DATA

The data was taken from the following corpora, compiled at the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences:

1. Corpus of Slovak judicial decisions *OD Justice*, version 2 (ODJ, 10.6 B tokens)
2. Corpus of Slovak legislative documents, version 1.9 (LEG, 44.9 M tokens)
3. Corpus of Slovak professional texts *prim-10.0-public-prf* (PRF, 189 M tokens)
4. Corpus of Slovak fiction *prim-10.0-public-img-sk* (IMG, 96.5 M tokens)
5. Balanced corpus of Slovak texts *prim-10.0-public-vyv* (BAL, 571.5 M tokens)

As the original corpora were different in size and document count, and because they each covered different genres, the following sub-corpora were created:

1. Sub-corpus of first instance decisions in criminal cases (CRC, 80.8 M tokens)
2. Sub-corpus of acts (ACT, 19.6 M tokens)
3. Sub-corpus of studies on marketing communication (STD, 3 M tokens)
4. Sub-corpus of Slovak short stories (COL, 27.6 M tokens)
5. Sub-corpus of the balanced corpus (BAL, 201.4 M tokens)

From these sub-corpora, random samples were created, labelled CRCS, ACTS, STDS, COLS, BALS, each containing 1 M random tokens:

years	**documents**	**words**	**nouns**	**n. lemmas**	**(ipm)**
1. CRCS	2020	9,338	53.8 K – 145	251.4 K	2,027
2. ACTS	2002 – 2022	1,367	72 K – 198	274.5 K	2,388
3. STDS	2011 – 2020	42	123.5 K – 18.6 K	285.3 K	4,383
4. COLS	1960 – 2020	847	263.2 K – 96	182.3 K	5,231
5. BALS	2010 – 2019	91,906	302.3 K – 10	259.7 K	6,300

Tab. 1. Samples used in analysis

The aim behind narrowing down the corpus data was mainly to work with similarly large portions of data and comparably long texts. The texts of first instance decisions in criminal cases are usually one to several pages long, and therefore more similar in length to studies and short stories than to monographs and novels.

The majority of the text samples cover mostly the period of the last 20 years. The relatively low count of documents for studies and short stories results from the fact that these texts are published collectively in proceedings and collections. The low bottom range of word counts could be due to the nature of the texts in question (brief acts and stories). Interestingly, while the relative frequency of noun lemmas in the samples remains approximately the same (with the exception of short stories), the number of noun lemmas in respective samples rises, as if the diversity of the lemmas was increasing. This might also indicate that the respective genres in the given order seem to be less and less formalized: from legally regulated communication during the criminal proceedings, through various walks of life regulated by legal norms, further through the socio-economic reality of marketing communication, to artistic depictions in short stories.

5 ANALYSIS

5.1 Visualization

The relative frequencies of the 50 most frequent lemmas for each sample were first visualized in the graph below. The x-axis represents the rank of lemmas, the y-axis their relative frequencies. The lemmas are only visualized as columns:

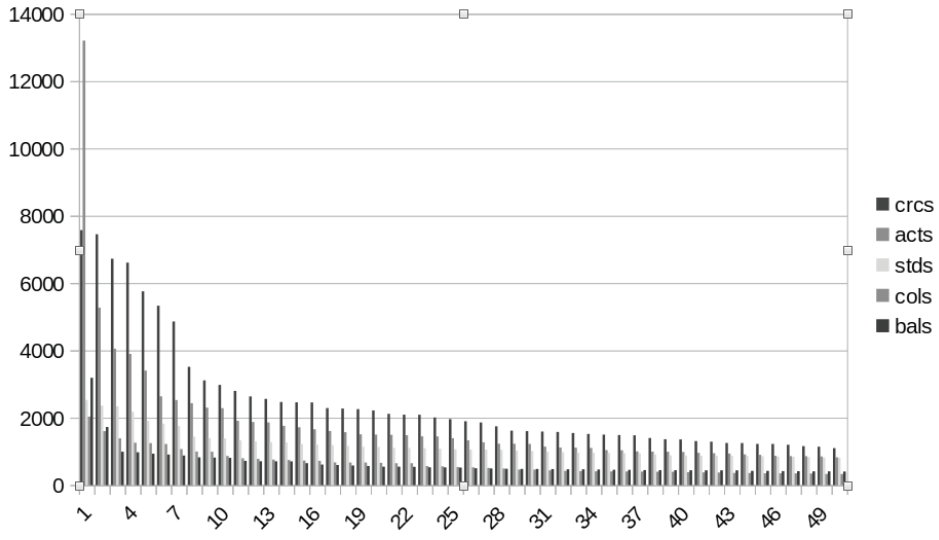

Fig. 1. Relative frequencies of the 50 most frequent lemmas

The red outlier column on the very left, standing for the lemma *zákon* 'act' in the sample of acts, exceeds the relative frequencies not only of the other samples, but also of the other lemmas in the act sample. This is understandable, given the nature of legislation that resembles a continuously modified corpus of legislative documents. These have a strong intertextual dimension, as new amendments refer to other acts and documents. The lemma 'act', however, also has the second most prominent frequency in the sample of decisions, so that the intertextual dimension is present here as well. This could be due to the fact that every decision has to refer to relevant articles of the Criminal Code, Criminal Procedure Code or other laws. Compared with the acts, nevertheless, the lemmas in the decisions sample have higher relative frequencies, although together shaped as a shallower slope, with several lemmas that are similarly frequent. After the bulk of these 7 lemmas, the slope drops down. This could mean that there is no such absolutely prominent lemma in the decisions, compared to acts. The relative frequencies of decisions, acts and studies are quite high, however, but the slopes later converge. The slopes for lemmas found in the short stories and the balanced corpus are lower, almost resembling a flat line. This might indicate a variety of frequent lemmas that are distributed more evenly.

5.2 Statistical testing

The samples, represented by 20 randomly generated lemmas from each sample, were then tested for significant differences in variance, using the online statistic calculator STATISTICS ONLINE.

Since at least the normality of distribution, one of three assumptions for the more robust, parametric ANOVA test, was debatable (the Kolmogorov-Smirnov test indicated normality, Lilliefors test did not), the Bartlett test of equal variance was dropped and the Kruskal-Wallis test, the non-parametric version of ANOVA, was performed:

The Kruskal-Wallis H test indicated that there is a significant difference in the *dependent variable* between the different *groups*, $\chi 2(4) = 55.22$, $p < .001$, with a mean rank score of 79.65 for CRCS, 66.58 for ACTS, 54.13 for STDS, 23.93 for COLS,28.23 for BALS.

The Post-Hoc Dunn's test using a Bonferroni corrected alpha of 0.005 indicated that the mean ranks of the following pairs are significantly different:

1. CRCS-COLS, CRCS-BALS
2. ACTS-COLS, ACTS-BALS
3. STDS-COLS, STDS-BALS.

Both decisions, acts and studies are significantly different from stories and the balanced corpus, while the significance of mutual differences between decisions, acts and studies has not been confirmed.

5.3 Semantic analysis

5.3.1 The lemmas were subsequently ranked according to *ipm* (in decreasing order) and described with the use of semantic categories (ordered alphabetically). Structures erroneously annotated as noun lemmas were excluded from the list. Number of lemmas for each semantic category and sample is given below:

CATEGORIES	CRCS	ACTS	STDS	COLS	BALS
ABSTRACT	10	16	30	7	21
ACTION	15	9	2	1	2
INSTITUTION	1	7	2	0	2
NORM	2	2	0	0	2
OBJECT	2	2	6	4	2
PERSON	7	2	4	12	5
PHYSICAL	0	0	0	10	3
PLACE	2	2	2	9	7
PROPER NAME	0	0	1	0	2
TEXT-RELATED	4	7	0	1	1
TIME	7	3	3	7	4

Tab. 2. Distribution of noun lemmas according to semantic categories

5.3.2 The 10 **abstract** nouns identified in the decisions sample denote basic terms of criminal law (e.g. *trest* ‘sentence’, *škoda* ‘damage’, *náhrada* ‘compensation’), compared to 16 lemmas in the acts denoting components of legislation (*znenie* ‘wording’, *zmena* ‘change’, *podmienka* ‘condition’). A portion of the 30 lemmas in the studies can be classified as general academic vocabulary (Kováříková 2017, pp. 26–27) (*médium* ‘medium’, *informácia* ‘information’, *cieľ* ‘goal’). The 7 lemmas in short stories represent universal existential concepts (*pravda* ‘truth’, *srdce* ‘heart’, *láska* ‘love’) and the 21 lemmas in the balanced corpus is a mixture of categories (*práca* ‘work’, *spoločnosť* ‘society’, *prípad* ‘case’).

The **action** nouns are mostly present in the decision sample (15, such as *odpor* ‘protest’, *rozkaz* ‘order’, *výrok* ‘verdict’), 9 were found in acts (*činnosť* ‘activity’, *konanie* ‘proceedings’, *rozhodnutie* ‘decision’), and only 1 – 2 in studies (*komunikácia* ‘communication’, *výskum* ‘research’), stories (*cesta* ‘road’, ‘journey’) and in the balanced corpus (*zápas* ‘struggle’).

The **institutions** are represented just by one, crucial institution issuing the decisions (*súd* ‘court’), as opposed to 7 institutions in acts (*úrad* ‘office’, *orgán* ‘authority’, ‘body’, *rada* ‘council’, ‘parliament’), only 2 in studies (*firma* ‘firm’, *podnik* ‘company’) and the balanced corpus (*škola* ‘school’, *štát* ‘state’), none in stories.

The names of **norm**-related lemmas are represented with 2 lemmas in the decisions (*zákon* 'act', *poriadok* 'code'), acts (*zákon* 'act', *predpis* 'regulation') and the balanced corpus (*zákon* 'act', *právo* 'law').

The **object**-related lemmas are represented by 2 lemmas in the decisions (*euro* 'euro', *vozidlo* 'vehicle'), acts (*fond* 'fund', *liek* 'medicine') and the balanced corpus (*euro* 'euro', *kniha* 'book'), and by 4 and 6 in stories (*dvere* 'door', *stôl* 'table') and studies (*produkt* 'product', *značka* 'brand'), respectively.

There were 7 **person**-related lemmas in the decisions (*osoba* 'person', *obžalovaný* 'defendant', *prokurátor* 'prosecutor'), which indicates more individual roles and actors when compared to the 2 person-related lemmas in the acts (*osoba* 'person', *zamestnanec* 'employee'), or 4 lemmas in the studies (*človek* 'human', *zákazník* 'customer', *respondent* 'respondee') and 5 in the balanced corpus (*človek* 'human', *dieťa* 'child'). The most person-related lemmas were found in the stories (*človek* 'human', *žena* 'woman', *dieťa* 'child') as these thematize diverse human interactions, mostly within the intimate surroundings of close family members. The decisions deal with human interaction as well, but due to the anonymization process, the parties are only referred to by random letters, not being automatically annotated as individuals and proper names.

The stories sample contained the most lemmas denoting **places** (*dom* 'house', *izba* 'room'), similarly as in the balanced corpus (*mesto* 'city', *svet* 'world', *miesto* 'place'), with only 2 lemmas each in the three remaining samples (*byt* 'apartment', *republika* 'republic', *územie* 'area', *prostredie* 'environment', *svet* 'world').

The names of **physical** entities, such as body parts, were almost exclusively found in stories (*ruka* 'hand', *oko* 'eye', *hlava* 'head') but partially in the balanced corpus as well (*ruka* 'hand', *oko* 'eye', *hlava* 'head'). This might indicate that the decisions, acts and studies, despite regulating social interaction in the physical world, deal mostly with abstract procedures and concepts.

Proper names were almost absent in the samples, with just 1 in the studies (*Slovensko* 'Slovakia') and 2 in the balanced corpus (*Slovensko* 'Slovakia', *Bratislava)*, which can be probably related both to the anonymization of decisions, the normative universality of acts and the high representation of general concepts in the studies.

Text-related noun were found mostly in acts (*odsek* 'article', *slovo* 'word', *písmeno* 'letter') and decisions (*odsek* 'article', *číslo* 'number', *písmeno* 'letter'), which is understandable, given the fact that both types of documents regularly refer to other acts and decisions. In the stories and the balanced corpus, *slovo* 'word' was the most prominent.

Lastly, **time**-related lemmas were found more in decisions (*deň* 'day', *mesiac* 'month', *lehota* 'period') than in the acts (*deň* 'day', *rok* 'year', *lehota* 'period'), which is probably due to the temporal context of any given lawsuit, as compared to the normative, universal character of legislation. The presence of time lemmas is also

prominent in the balanced corpus (*rok* ‘year’, *život* ‘life’, *deň* ‘day’) and in stories, where they also can function as literary motifs (*deň* ‘day’, *rok* ‘year’, *čas* ‘time’).

5.3.3 Finally, the overlapping lemmas, that is, those found in the decisions and in one or more other samples, were identified and their relative frequencies were tested for the significance of differences. This was done by using the Czech corpus calculator CALC. Lemmas significantly different from those in the decisions are marked in bold, those not significantly different in italics.

OVERLAPPING LEMMAS	CRCS	ACTS	STDS	COLS	BALS
vec ‘matter’, ‘issue’, ‘case’	2,257				**442**
konanie ‘proceedings’, ‘procedure’	2,468	**1,760**			
výkon ‘fulfilment’, ‘enforcement’	1,858	**1,447**			
rozhodnutie ‘decision’	1,477	*1,607*			
súd ‘court’	7,574	**1,031**			
značka ‘signature’, ‘brand’	1,289		**872**		
zákon ‘act’	7,451	**13,203**			**468**
euro ‘euro’	1,602				**540**
osoba ‘person’	4,359	*3,898*			
republika ‘republic’	1,155	**2,527**			
odsek ‘article’	2,974	**5,270**			
písmeno ‘letter’	1,221	*1,330*			
deň ‘day’	6,608	**2,431**		**1,245**	**972**
lehota ‘period’	2,089	**936**			
čas ‘time’	1,547		**956**		**932**
rok ‘year’	1,355	**1,858**	**2,359**	**1,068**	**3,186**

Tab. 3. Dispersion of overlapping lemmas throughout the corpora

A significant difference was indicated in a total of 20 pairs. The confidence intervals for the lemmas *rozhodnutie* ‘decision’, *osoba* ‘person’, *písmeno* ‘letter’ were too close to each other, despite a significant difference at the alpha level of 0.05. The majority of overlapping lemmas stem from the legal domain, which correlates with the idea of intertextuality between decisions and acts. The lemmas most dispersed throughout the corpora were the time-related lemmas, with ‘year’ found in all and ‘day’ in almost all samples.

6 DISCUSSION

As shown in Tab. 2, the decisions seem to contain fewer types of lemmas, when compared to other genres or samples, but the relative frequencies of lemmas in the

decisions are higher. In other words, their noun vocabulary is more limited, but used more intensively, which could corroborate J. Mistrík's claim concerning the lexical poverty of administrative texts. The statistic section indicated, however, that significant differences only exist between specialized texts (decisions, acts, studies) and short stories/mixed texts. The semantic analysis of decisions showed a notable presence of nouns denoting "action", mostly in the legal sense, as well as presence of domain-specific legal terms (*samosudca* 'single judge', *(trestný) rozkaz* '(criminal) order' or *odpor* 'protest' (a remedial measure against an order). Since the lexical differences often result from different thematic parameters (Slančová et al. 2022, pp. 445–446), a more robust analysis should also apply more measures, be it lexical density, lexical diversity or some differential measure for the prominence of one word in several contexts, such as the difference index DIN (Chlumská 2017, pp. 55–56).

7 CONCLUSION

From the legal point of view, judicial decisions represent the final stage in the application of law (Prusák 2001, pp. 289–293; Ivor et al. 2017, pp. 165–188). The procedural nature of the stage of decisions within a legal action is then reflected in the texts themselves. Here it materializes as the prominence of action-related nouns, denoting procedures, institutes and measures defined by law. Action-related nouns are to a lesser degree also present in the acts, because these result from a dynamic process – the legislation – as well. The "abstract" lemmas, denoting domain-specific legal terms, are found in the decisions, although they are even more present in the acts and mostly in studies on mass and marketing communication, which focus on a large spectrum of socio-economical phenomena. As complex and legally regulated speech acts, judicial decisions have a prescribed structure which is comparable, but not identical to that of legislative norms (Weinberger 2017, pp. 195–199). This is why some of the person-related lemmas, mostly denoting parts in criminal proceedings, are recurrent. On the other hand, there are several institutional actors involved in the process of legislation, which could be the reason why they appear in the acts more often. Hopefully, this paper has showed that despite its lexical idiosyncrasies – or because of them – the style of judicial decisions deserves further linguistic attention.

ACKNOWLEDGEMENTS

The paper has been written within the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Science, Research and Sport of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

References

Abrahámová, E. (2012). Základy právnej komunikácie. Bratislava: Heuréka, 128 p.

Corpora of the Slovak National Corpus: prim-10.0-public-prf, prim-10.0-public-img-sk, prim-10.0-public-vyv. Accessible at: https://korpus.sk/.

Corpus calculator CALC. Ústav Českého národního korpusu. Accessible at: https://www.korpus.cz/calc.

Corpus of Slovak judicial decisions OD Justice, version 2.0. Accessible at: https://www.juls.savba.sk/justicecorp.html.

Corpus of Slovak legal documents, version 1.9. Accessible at: https://www.juls.savba.sk/legalcorp.html.

Cvrček, F. (2016). Právní informatika a lingvistika. Jurisprudence, 25(6), pages 49–53.

Cvrček, V. et al. (2020). Registry v češtině. Praha: Nakladatelství Lidové noviny, 234 p.

Čulenová, E. (2010). Jazyk a štýl v slovenských zákonoch. Banská Bystrica: Univerzita Mateja Bela, 140 p.

Findra, J. (2004). Štylistika slovenčiny. Bratislava: Osveta, 232 p.

Gibbons, J. (1999). Language and the law. Annual Review of Applied Linguistics, 19, pages 156–173.

Chlumská, L. (2017). Překladová čeština a její charakteristiky. Praha: Nakladatelství Lidové noviny, 150 p.

Imrichová, M., and Turočeková, M. (2015). Lingvistická analýza právnych textov. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove, 107 p.

Ivor, J., Polák, P., and Záhora, J. Trestné právo procesné II. Bratislava: Wolters Kluwer SK, 492 p.

Kováříková, D. (2017). Kvantitativní charakteristiky termínů. Praha: Nakladatelství Lidové noviny, 136 p.

Mistrík, J. (1975). Žánre vecnej literatúry. Bratislava: Slovenské pedagogické nakladateľstvo, 212 p.

Mistrík, J. (1997). Štylistika. 3. vydanie. Bratislava: Slovenské pedagogické nakladateľstvo, 598 p.

Prusák, J. (2001). Teória práva. Bratislava: Právnická fakulta Univerzity Komenského v Bratislave, 340 p.

Ružička, J. et al. (1966). Morfológia slovenského jazyka. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 895 p.

Slančová, D. et al. (2022). Úvod do štúdia interaktívnej štylistiky I, II. Prešov: Prešovská univerzita v Prešove, 481 + 449 p. Accessible at: https://www.pulib.sk/web/kniznica/elpub/dokument/Slancova6.

Statistics calculator STATISTICS ONLINE. Accessible at: https://www.statskingdom.com.

Štícha, F. (1985). O jazyce soudních rozhodnutí. Naše řeč, 68(2), pages 68–77.

Vogel, F. (ed.) (2019). Legal linguistics beyond borders. Language and law in a world of media, globalisation and social conflicts. Berlin: Duncker and Humblot, 384 p.

Weinberger, O. (2017). Norma a instituce. Úvod do teorie práva. Plzeň: Aleš Čeněk, 248 p.

LANGUAGE ACQUISITION, CREATION AND USE OF LANGUAGE RESOURCES

sciendo
DOI 10.2478/jazcas-2023-0033

SPANISH SYNONYMS AS PART OF A MULTILINGUAL EVENT-TYPE ONTOLOGY

CRISTINA FERNÁNDEZ-ALCAINA – EVA FUČÍKOVÁ
– JAN HAJIČ – ZDEŇKA UREŠOVÁ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic

FERNÁNDEZ-ALCAINA, Cristina – FUČÍKOVÁ, Eva – HAJIČ, Jan – UREŠOVÁ, Zdeňka: Spanish Synonyms as Part of a Multilingual Event-Type Ontology. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 153 – 162.

Abstract: This paper presents an ongoing work on the multilingual event-type ontology SynSemClass, where multilingual verbal synonymy is formalized in terms of syntactic and semantic properties. In the ontology, verbs are grouped into synonym classes, both monolingually and cross-lingually. Specifically, verbs are considered to belong to the same class if they both express the same meaning in a specific context, and their valency frame can be mapped to the set of roles defined for a particular class. SynSemClass is built following a bottom-up approach where translational equivalents are automatically extracted from parallel corpora and annotated by human annotators. The task of the annotators consists in mapping the valency frame of a particular verb with the set of roles defined for the class where the verb is included as a potential class member, establishing links to external resources, and selecting relevant examples. The Spanish part of the ontology currently contains 257 classes enriched with Spanish synonyms. The resulting resource provides fine-grained syntactic and semantic information on multilingual verbal synonyms and links to other existing monolingual and multilingual resources.

Keywords: multilingual, ontology, semantics, valency, verbs

1 INTRODUCTION

This paper presents an ongoing work on the construction of a multilingual ontology for events. SynSemClass (SSC) is a multilingual lexicon of verbs organized into classes based on their semantic and syntactic properties. For our purposes, synonymy is defined in terms of contextual synonymy, i.e., a verb is considered a member of a class if it conveys the same or similar meaning expressed by other verbs within the same class, both monolingually and across languages. The construction of SSC involves fine-grained multilingual syntactic-semantic annotation as well as linking to several external resources in different languages (Czech, English, German, and Spanish so far). The information gathered in this lexicon facilitates cross-linguistic comparison, making it a valuable resource for linguistic research. Additionally, SSC also provides curated data useful for Natural Language

Processing tasks, such as cross-lingual synonym discovery, and is used for annotation of UMR (Unified Meaning Representation) (Bonn et al. 2023; Xue et al. 2023).

This paper is structured as follows: Section 2 briefly presents the ontology. Section 3 describes the method used for the extraction and filtering of Spanish candidates. The annotation process and an assessment of its quality are presented in Section 4, and the results obtained so far are described in Section 5. The conclusion and some plans for future work are summarized in Section 6.

2 THE ONTOLOGY

The organization of the ontology into classes revolves around the definition of synonymy as 'contextual synonymy' (Palmer 1981). That is, two or more verbs are considered synonyms (and thus members of the same class) if they express the same or similar meaning in the same context. Some aspects may need clarification: by 'verbs', we refer to verb senses, as we are dealing with cases of partial synonymy; by 'synonyms', we refer to both monolingual and cross-lingual synonyms, since the ontology is multilingual; and by 'context', we refer to the set of semantic roles expressed by the arguments and adjuncts of a verb, either explicitly or implicitly and with possible restrictions.

Therefore, a verb sense in any language is included in a specific class provided that each of the roles defined for the given class can be mapped to the verb valency slots captured in the valency frame. While total mapping between roles and arguments is a requirement, roles can be expressed by different morphosyntactic realizations and additional restrictions may apply, for example, regarding register or domain (Urešová et al. 2018b). Furthermore, each class member (CM) is linked to related entries in a set of preselected language-specific lexical resources, such as VerbNet (Schuler – Palmer 2005) for English, VALLEX (Lopatková et al. 2017; Lopatková et al. 2020) for Czech, and FrameNet des Deutschen (FdD) for German, among others.

The latest version of the ontology, SynSemClass4.0 (June 2022), contains 883 classes (approx. 6,000 CMs) in English and Czech, 61 of which are enriched with German synonyms.[1]

3 DATA PREPARATION

3.1 Main resources

Previous work on the extension of the ontology described a minimal set of resources required for the addition of new languages (Urešová et al. 2022). In particular, the necessary resources required are two: a parallel corpus and (at least) one lexical resource with information on verbal valency.

[1] The ontology is available for browsing and download at: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4746.

Regarding the corpus, the data for Spanish have been extracted from the X-SRL dataset (Daza – Frank 2020), a sentence-aligned parallel corpus containing approx. three million words for the English-Spanish part. The corpus is composed of texts from the Wall Street Journal section of the Penn Treebank and their Spanish translations. Although automatically translated, the quality of the translations was evaluated by human annotators with positive results (Daza – Frank 2020, p. 3909). Texts are tokenized, lemmatized, and POS-tagged.

Regarding the lexical resource, valency information was retrieved from AnCora-ES (Taulé et al. 2008). This verbal lexicon contains 2,820 lemmas (3,938 senses) and was built based on a corpus containing texts from a Spanish newspaper. One of the advantages of using AnCora is that, although monolingual, each sense is linked to several English resources that are also used in the English part of the ontology (specifically, VerbNet, PropBank, FrameNet, WordNet 3.0, and OntoNotes). This feature facilitates annotation process in two ways: first, it makes possible the automatic selection of senses to be imported to the tool used for annotation (see Section 4), thus restricting the number of candidates; second, it simplifies the process of determining verb class membership for human annotators as it offers comparable information between English and Spanish.

3.2 Candidate extraction

Candidate extraction is done in two phases: i) automatic extraction of English-Spanish pairs from the corpus, and ii) data filtering. An illustration of the workflow for data extraction is presented in Fig. 1:

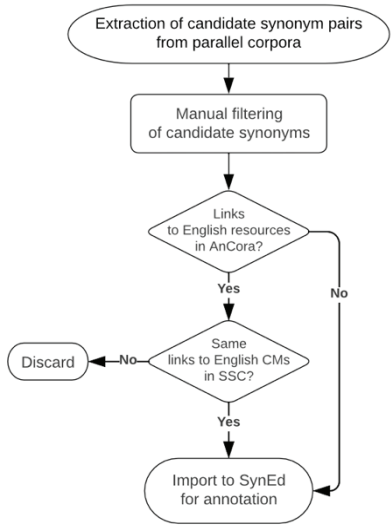

Fig. 1. Candidate extraction workflow

In the first phase, candidate pairs of synonyms are automatically extracted from the parallel corpus. The final dataset amounted to 40,408 verbs divided into 1,715 verbal types. Each Spanish verb in the dataset was paired with its possible English translation in the given context (e.g. *abrir-emerge*, *abrir-leave*, *abrir-come out*, etc.). To prevent an excessive number of irrelevant cases (i.e., incorrect pairings) from being imported into the tool for final manual complex annotation, data was prefiltered in two steps: by lemma and by sense. In the first step, Spanish-English pairs extracted from the corpus were filtered manually by annotators. For each class, annotators discarded the lemmas that did not belong to the class where they were automatically included if they were i) the result of errors during the automatic extraction processes; and ii) verbs that, despite being possible translations of the English verbs, do not express the meaning reflected by the class where they have been automatically included. Based on a sample of 59 classes, amounting to 3,016 verbs, the list of potential candidates was reduced by 68%, i.e., only 990 lemmas were retained for annotation. Although this step drastically reduced the size of the list, a second step in the prefiltering phase was added to restrict the number of candidates available for final annotation. To this aim and based on the information provided by AnCora, the candidates imported for final annotation were of two types: i) AnCora senses linked to the same links to PropBank and/or VerbNet to which English class members in SynSemClass are linked, and ii) AnCora senses with no links to other resources (as these usually represent multi-word expressions with no direct equivalent in English).

4 ANNOTATION

4.1 Annotation setup

The annotation was carried out by three native Spanish speakers who are proficient in English and trained in linguistics. Annotators were provided with annotation guidelines (Fernández-Alcaina et al. 2022) and trained on a preliminary set of classes. To ensure the quality of the annotations, each set of classes was processed by two annotators and the annotations were monitored by one of the authors of this paper. Any discrepancies that arose during the annotation process were discussed as needed. For the task of annotation, which involves mapping roles to arguments, identifying external links, and selecting examples, we used the SynEd editor (Urešová et al. 2018a; Fučíková et al. 2023), a tool specifically designed for this purpose by the SynSemClass maintainers and refactored to adapt it to any number of languages.

4.2 Class membership

The first task of the annotators involves determining whether a given verb belongs to a specific class. A verb is considered a member of a class if i) the meaning conveyed by the Spanish verb in a given context is the same or similar to that

conveyed by its English equivalent, and ii) it is possible to map each role defined in the Roleset for that class to the arguments defined for the verb in the valency frame. In addition to verbal lemmas, SSC includes multi-word expressions, such as idioms and light verb constructions (LVCs). Tab. 1 provides a simplified example of the role-argument mapping of a set of multilingual verbal synonyms in class vec00012 ('An Authority allows an Affected entity to engage in a Permitted entity').

	Authority	Permitted	Affected
allow (EN)	ACT	EFF	PAT
dovolit (CS)	ACT	EFF	ADDR
erlauben (DE)	VA0	VA1	VA2
permitir (ES)	arg0	arg1	arg2

Tab. 1. Role-argument mapping for class members in English, Czech, German, and Spanish in class vec00012 (simplified)

An evaluation of the annotations was performed to assess the quality of the annotated data. The sample used for this aim contains the last set of 42 classes annotated by two pairs of annotators: A1 vs A2 (21 classes, 1,099 verbs) and A1 vs A3 (21 classes, 998 verbs).

In the step of defining class membership, annotators could choose from five labels: 'yes', 'rather_yes', 'no', 'rather_no', and 'deleted'. To facilitate the interpretation of the results obtained, the five labels have been reorganized into two categories: 'yes' (including 'yes' and 'rather_yes') and 'no' (including 'no', 'rather_no', and 'deleted'). Tab. 2 presents the results for the agreement rate and Cohen's κ value (Cohen 1960).

	A1 vs A2	A1 vs A3
Agreement	91%	94%
Cohen's κ	0.62	0.75

Tab. 2. IAA results for Spanish verbal synonyms class membership

The results indicate a high level of agreement among the annotators in both cases (91% and 94%), but these may indicate a biased representation since i) the percentage of agreement between two annotators is expected to be high, and ii) the data distribution is highly biased towards the label 'no'. On the other hand, Cohen's κ is a more informative measure for this purpose as it can correct such bias. For the first pair of annotators (A1 vs A2), κ= 0.62, and the second pair of annotators (A1 vs A3), κ=0.75, thus indicating substantial agreement in both cases.

When compared to the initial set of classes, the results show a notable improvement across subsequent batches as the annotators become more familiar

with the task at hand. Tab. 3 compares the results obtained with the values for the first set of classes annotated (14 classes, 939 verbs). Specifically, for the pair of annotators A1 and A2, the agreement rate remains the same (91%), and κ increases from 0.49 to 0.62. Similarly, for the second pair of annotators, A1 and A3, there is a slight increase in the agreement rate (from 92% to 94%), as well as for κ values (from 0.54 to 0.75).

	A1 vs A2		A1 vs A3	
	First set	Last set	First set	Last set
Agreement	91%	91%	92%	94%
Cohen's κ	0.49	0.62	0.54	0.75

Tab. 3. IAA results for Spanish verbal synonyms class membership for the first set (14 classes) and the last set (21 classes) of classes

One possible explanation for the observed discrepancies in the results obtained for the two pairs of annotators A1 vs. A3 and A1 vs. A2 is that the latter seems to follow a more inclusive approach when determining class membership. This is not unexpected given the semantic complexity of the task.

4.3 Additional information

In the ontology, class members are linked to related entries in other external resources available for Spanish. Specifically, the resources used are two monolingual lexicons (ADESSE and Spanish SenSem), the Spanish version of FrameNet, and the Spanish WordNet 3.0 integrated into the Multilingual Central Repository.

1. ADESSE (García-Miguel et al. 2005) contains 3,400 lemmas extracted from the corpus ARTHUS (1.5 million words). The lexicon provides information regarding argument structure and semantic roles.
2. Spanish SenSem (Alonso et al. 2007) contains the most frequent 250 verbs from the SenSem corpus (Fernández-Montraveta – Vázquez 2014) built on texts from newspapers and literary sources.
3. Spanish WordNet 3.0 is integrated within the Multilingual Central Repository (Gonzalez-Agirre et al. 2012) together with six languages, including English. The MCR is also enriched with semantically tagged glosses and contains ontology information from WordNet Domains, Top Ontology, and AdimenSUMO.
4. Spanish FrameNet (Subirats 2009) contains 1,000 lexical units based on frame semantics and supported by corpus data. It provides syntactic and semantic information for each sense automatically annotated and validated by human annotators.

The resources selected complement the information provided by AnCora in different ways. Specifically, both ADESSE and SenSem provide definitions and

valency frames for each sense that make easier the task of the annotators. Spanish FrameNet uses the same frames used in the original English version, thus facilitating the task of the annotators and giving consistency to the annotation. Similarly, the Spanish WordNet 3.0 also provides consistency and facilitates annotation by integrating Spanish synonyms in multilingual synsets where English equivalent verbs are included.

The last step of the annotation consists in selecting relevant examples to illustrate the meaning of the class. Whenever possible, the argument structure defined for that verb sense must be explicitly realized in the examples.

5 RESULTS

The Spanish part of the SynSemClass currently contains 257 classes enriched with 1,400 Spanish verbal synonyms. It will be available in the next release, SynSemClass 5.0 (planned for 2023). Although classes with Spanish members still represent a small part of the total number of classes (29%), the results obtained so far are relevant for the development of the ontology in several aspects.

In terms of organization, Spanish has offered the opportunity to 'simulate' a scenario where an 'external' team works on the addition of a new language only with central support from the original maintainers.

From a methodological perspective, while the procedure followed for Spanish partly relies on previous work in German, it has been necessary to make some changes in the annotation tool and annotation process in other to adapt to some specifications of Spanish language. The results obtained so far may serve as the basis for future work in Spanish but also for other languages as the tool is expected to continue evolving to adapt to new languages.

In terms of multilinguality, adding a new language (and the first from a different family) contributes to enriching and refining synonymy classes in the ontology by providing linguistic evidence (including special cases, such as LVCs). With the addition of Spanish, it is already clear that as more data from more languages are added, classes will need to be hierarchized in the future (modified by splitting, merging, etc.).

Regarding Spanish resources, to the best of our knowledge, SSC has become the first multilingual richly annotated resource of a general ontology type that includes Spanish. It is also the first to link various existing Spanish lexical resources, in line with other initiatives such as the Unified Verb Index (UVI)[2] for English.

As for the limitations, the development of the ontology still heavily relies on manually processed data with respect to both data filtering and annotation.

[2] https://uvi.colorado.edu/

6 CONCLUSIONS AND FUTURE WORK

This paper has described the progress made so far in the addition of a new language to the multilingual event-type ontology SynSemClass. Although part of the method employed for the inclusion of Spanish is based on previous work in German, some steps needed to be added or modified in the data extraction and preparation phase to accommodate the specific features of the resources used. Similarly, the tool employed for annotation has also undergone some refactorization that allows to include a new language. In terms of the annotation process, some aspects of Spanish required a specific treatment, such as pronominal verbs or differences across geographical varieties. Furthermore, and specifically for Spanish, the result of adding Spanish is a twofold contribution in that it does not only enrich the ontology but also provides a resource that links data from several resources developed for Spanish that were independent up to now and complements them by including senses that are not captured in the resources available. The resulting resource has significant implications for both multilinguality and contrastive language research.

The addition of Spanish to the ontology is one step more towards the creation of a collaborative multilingual event-type ontology. From a more global perspective, plans in the near future include continuing work on the multilingual character of the ontology by adding more languages, which would necessarily imply that the lexicon and the tools continue evolving and adapting to the specification of new languages.

In the long term, the project is a part of a larger project for multilingual knowledge representation, where the SynSemClass classes will serve as a grounding for all events and states by relating all other entities in the resulting representation, which will also be grounded using other means. Although some verb annotation experiments have been done for the previous versions of the ontology, the full specification is still to be developed.

ACKNOWLEDGEMENTS

The work described herein has been supported by the Grant Agency of the Czech Republic under the EXPRO program as project “LUSyD” (project No. GX20-16819X) and uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (project No. LM2018101, supported by the Ministry of Education of the Czech Republic).

References

Alonso, L., Capilla, J. A., Castellón, I., Fernández-Montraveta, A., and Vázquez, G. (2007). The SenSem project: Syntactico-semantic annotation of sentences in Spanish. Recent Advances in Natural Language Processing IV, pages 89–98.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – Volume 1, ACL'98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics

Bonn, J., Myers, S., Gysel van, J. E. L., Denk, L., Vigus, M., Zhao, J., Cowell, A., Croft, W., Hajič, J., Martin, J. H., Palmer, A., Palmer, M., Pustejovsky, J., Urešová, Z., Vallejos, R., and Xue, N. (2023). Mapping AMR to UMR: Resources for Adapting Existing Corpora for Cross-Lingual Compatibility. In Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 74–95, Washington, D.C. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), pages 37–46.

Daza, A., and Frank, A. (2020). X-SRL: A parallel cross-lingual semantic role labeling dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3904–3914. Online. Association for Computational Linguistics. Accesible at: https://aclanthology.org/2020.emnlp-main.321.

Fernández-Alcaina, C., Fučíková, E., and Urešová, Z. (2022). Annotation guidelines for Spanish verbal synonyms in the SynSemClass lexicon. Technical Report 72, ÚFAL MFF UK, 52 p.

Fučíková, E., Hajič, J., and Urešová, Z. (2023). Corpus-Based Multilingual Event-type Ontology: Annotation Tools and Principles. In Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 1–10, Washington, D.C. Association for Computational Linguistics.

García-Miguel, J. M., Costas, L., and Martínez, S. (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. In G. Wotjak – J. Cuartero Otal (eds.): Entre semántica léxica, teoría del léxico y sintaxis. Berlin: Peter Lang Verlag, pages 373–384.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual Central Repository version 3.0. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2525–2529, Istanbul, Turkey. European Language Resources Association (ELRA).

Lopatková, M., Kettnerová, V., Bejcek, E., Vernerová, A., and Žabokrtský, Z. (2017). Valenční slovník českých sloves VALLEX. Nakladatelství Karolinum, Praha.

Lopatková, M., Kettnerová, V., Vernerová, A., Bejcek, E., and Žabokrtský, Z. (2020). VALLEX 4.0 (2021-02-12). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-3524.

Palmer, F. R. (1981). Semantics. 2nd ed. Cambridge: Cambridge University Press.

Taulé, M., Martí, A., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pages 96–101, Marrakech, Morocco. European Language Resources Association (ELRA).

Schuler, K. K., and Palmer, M. S. (2005). Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis. University of Pennsylvania, USA.

Subirats, C. (2009). Spanish FrameNet: A frame semantic analysis of the Spanish lexicon. In H. C Boas (ed.): Multilingual FrameNets in Computational Lexicography. Berlin/New York: De Gruyter Mouton, pages 135–162.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018a). Tools for building an interlinked synonym lexicon network. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018b). Creating a verb synonym lexicon based on a parallel corpus. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 1432–1437, Paris, France. European Language Resources Association.

Urešová, Z., Zaczynska, K., Bourgonje, P., Fučíková, E., Rehm, G., and Hajič, J. (2022). Making a semantic event-type ontology multilingual. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1332–1343, Marseille, France. European Language Resources Association.

Xue, N., Bonn, J., Cowell, A., Hajič, J., Palmer, A., Palmer, M., Pustejovsky, J., Sun, H., Urešová, Z., Wein, S., and Zhao, J. (2023). UMR Annotation of Multiword Expressions. In The 4th International Workshop on Designing Meaning Representation (DMR 2023), June 20, 2023, Nancy, France.

sciendo
DOI 10.2478/jazcas-2023-0034

ERRORS IN THE CONGRUENT ATTRIBUTE AMONG STUDENTS LEARNING SLOVAK AS A FOREIGN LANGUAGE (LEARNER CORPUS-BASED)

KATARÍNA GAJDOŠOVÁ[1] – MICHAELA MOŠAŤOVÁ[2] – PETRA ŠVANCAROVÁ[2]
[1] Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Studia Academica Slovaca – The Centre for Slovak as a Foreign Language, Faculty of Arts, Comenius University, Bratislava, Slovakia

GAJDOŠOVÁ, Katarína – MOŠAŤOVÁ, Michaela – ŠVANCAROVÁ, Petra : Errors in the Congruent Attribute among Students Learning Slovak as a Foreign Language (Learner Corpus-based). Journal of Linguistics, 2023, Vol. 74, No 1, pp. 163 – 172.

Abstract: This paper analyses error rates in the congruent attribute in texts written by students learning Slovak as a foreign language. The material, which was qualitatively analysed, comes from the pilot version of the learner corpus errkorp-pilot. The paper defines the most common types of errors in the congruent attribute and interprets the causes of their origin. The most common errors include the wrong congruence with the grammatical gender and with the case of the defining noun. Errors are usually caused by the following factors: transfer from the L1 language, excessive generalization of the rule in the target language, or insufficient knowledge of the grammatical rule.

Keywords: attribute, congruence, language error, learner corpus, Slovak as a foreign language

1 INTRODUCTION

The aim of the paper is to demonstrate the most relevant types of errors in the congruent attribute by analysing language material from the learner corpus Slovak as a foreign language. Congruence is an important morphosyntactic phenomenon that non-native speakers face when learning Slovak as a foreign language. In Slovak, which is an inflectional language, congruence represents one of the basic formal relationships between the governing and dependent members of a syntagm (Kačala 2014; Ivanová 2020). The main reason for errors is a divergence from at least one of the grammatical categories (gender, number, or case) of the governing member in agreement with the dependent member. An analysis of corpus-based material makes it possible to identify the relevant types of errors in the congruence, describe them, and subsequently interpret them in relation to individual L1 languages and the acquired language level of learners.

2 DATA

We drew the research material from the pilot version of learner corpus including texts written by students learning Slovak as a foreign language. The corpus errkorp-pilot is available through the Slovak National Corpus. The selected corpus contains 137,393 tokens, it is fully lemmatized and morphologically annotated at the error and correct token levels. Errors have been manually annotated with tags that capture the respective types of errors. More information about the annotation method can be found on the project website[1]; the principles of creating of the corpus have been described by M. Mošaťová and K. Gajdošová (2019).

The research sample comprised 172 texts from students with twenty-four different L1 languages[2] with the following allocations: be (2), bg (39), de (24), en (3), es (10), fa (4), fi (1), fr (21), he (1), hr (18), hu (5), it (12), ko (13), lt (3), nl (2), pl (16), pt (2), ro (29), ru (29), sl (8), sr (15), tr (22), uk (29), and zh (11). We searched the corpus using a tag "congr", which identifies different types of errors[3] in grammatical congruence: e.g. *ten centrum (to centrum)* 'the centre', *zaujímavý esej (zaujímavá esej)* 'interesting essay', *veľa ľudí boli (veľa ľudí bolo)* 'a lot of people were', *rámy je vodotesný (rámy sú vodotesné)* 'frames are waterproof', and *ja nerozumieť (ja nerozumiem)* 'I don't understand'. This paper deals exclusively with errors in congruent attributes that are related to an adjective (or other words having the form of an adjective) and the following substantive. When interpreting and discussing individual errors, we took the language of L1 students and their language level of Slovak into account. We obtained this information from the available metadata for the individual texts in the corpus.

3 ANALYSIS

3.1 Errors in gender

While a natural gender appears in all languages, a grammatical gender – as a classification category of nouns – does not exist in many of them (e.g. in Finno-Ugric languages). In some European languages, the original three-value system was reduced to masculine vs feminine (e.g. in Romance languages). The grammatical gender of nouns is primarily an arbitrary category in languages. Confusion about the grammatical gender of nouns thus results in the incorrect formal implementation of congruence in the case of a congruent attribute that is a part of foreigners' relatively frequent errors.

[1] https://korpus.sk/korpusy-a-databazy/korpusy-snk/errkorp/

[2] A list of abbreviations for L1 languages and their meaning can be found at https://korpus.juls.savba.sk/attachments/errkorp_outannot/zoznam_jazyky.txt.

[3] In total, we recorded 319 errors related to the congruent attribute in errkorp-pilot. Their distribution according to analysed types was as follows: 88 errors in gender, 56 errors in the case, 31 errors in the masculine person, 25 spelling errors -i/-y and -í/-ý.

The research sample confirmed phenomena reported by other authors (Kotková 2017; Spáčilová 2019; Imrichová 2023). At the A1 to B1 levels, some cases demonstrate the inconsistency of agreement in gender between a noun and its modifiers (adjective, possessive pronoun). Incorrect gender allocation occurs in the case of nouns ending with weak gender suffixes because "there is no formal criterion for the unambiguous identification of the gender" (Pekarovičová 2022, p. 68): *mám *jeden/jednu posteľ* 'I have one bed' (fr, A1.1); **môj/moje hobby* 'my hobby' (ro, A2.1); **vysoký/vysoká úroveň verejných výdavkov* 'a high level of public finances' (it, B1); *urobil som *dobrý/dobrú vec* 'I have done a good thing' (tr, B1), as well as those ending with strong gender suffixes *-0* (M), *-a* (F), and *-o* (N): e.g. *máme *malý/malú záhrada* 'we have a small garden' (fr, A1.1); *niemame*[4] **dobru/dobré pivo* 'we do not have a good beer' (fr, A1.1); **starý/staré mesto* 'old town' (de, A1.1); **toto/táto žena je Jane* 'this woman is Jane' (bg, A1.1); **moje/moja *ďalšie/ďalšia babička pracuje* 'my other grandmother works' (bg, A1.2); *Tesco je *veľké/veľký supermarket* 'Tesco is a huge supermarket' (es, A1.2). As a transfer from Romanian, the use of the feminine gender in the attribute modifying the lexeme *leto* 'summer' can be interpreted as **moja/moje *ideálna/ideálne leto, *moja/moje *obľúbena/obľúbené leto* 'my ideal summer, my favourite summer' (ro, A2) because the lexeme *vară* (*"leto/summer"*) belongs to the feminine gender in Romanian.

At the B2 to C2 levels, there were interlingual errors in the nouns that indicate a different gender in the equivalent lexeme in the L1 language: e.g. *od žien *taký/také kritérium sa nevyžaduje* 'women are not required the criterion' (uk, B2), cf. *kriterij* (M) in Ukrainian; in English *criterion*. On one hand, these may be intralingual errors caused by an overgeneralization of the grammatical rule of gender identification. For example, if nouns end in a consonant, they incline towards the masculine: e.g. **lacný/lacná čínsky/čínska obuv* 'cheap Chinese shoes' (pl, B2); **rímsky/rímska pop-music* 'Roman pop-music' (hr, B2). Frequent examples of errors in gender appearing in linguodidactic practice include the words *múzeum* 'museum' and *štúdium* 'study', which students usually associate with the masculine gender. We even found evidence in our research sample of students with Russian and Serbian as L1 languages who were at C2 level at the time of producing the text: *skončila som *bakalársky/bakalárske štúdium a univerzita prihlásila mňa na *magisterský/magisterské štúdium* 'I finished bachelor degree and the university enrolled me for the master degree' (ru, C2); **najznámejší/najznámejšie múzeum v Belehrade je *Národný/národné múzeum* 'the most famous museum in Belgrade is the National Museum' (sr, C2).

The interpretation of errors in the following cases was ambiguous: **jeden/jedna forma autocenzúry* 'a form of self-censorship' (sr, B2); **aké/aký problém máte?*

[4] Given examples contain not only errors related to our research, also other types of errors are included. In the text, we present only the correction of error in congruence and we also translate it.

‘what problem do you have?’ (fr, A2). In Serbian, the lexeme *forma* ‘form’ belongs to the feminine as it does in Slovak. In French, the lexeme *problème* belongs to the masculine. We assume this is more of a performance error.

In the case of the lexeme *dievča* ‘girl’, there is often a substitution of the grammatical gender in linguodidactics instead of the natural one: **tá/to dievča je *múdrejšia/múdrejšie ako on* ‘the girl is cleverer than he’ (sr, B2). In this case, the neuter is challenging to accept, especially for students with a Romance L1 (it, es, pt, fr). A specific implementation of the congruence with the neuter noun, expressed in the use of the adjective suffix *-o*, was recorded in students with Bulgarian and Slovenian L1: e.g. *obľúbeno/obľúbené mesto* ‘favourite city’ (bg, A1.1); *Pleven je *veľko/veľké mesto* ‘Pleven is a big city’ (bg, A1.2); **kultúrno/kultúrne podujatie* ‘cultural event’ (sl, B1). There are cases when students used different gender suffixes in the multiple premodifying adjectives: e.g. **moje/môj obľúbený nápoj* ‘my favourite drink’ (ko, A1.2); **môj/moje obľúbené miesto* ‘my favourite place’ (ro, A2); **moje/moja obľúbená zábava* ‘my favourite fun’ (ru, C2); *zázračné *čierno/čierne zrkadielko* ‘a magic black mirror’ (ru, C2). They also used them within the same sentence structure: **moja/moje vlasy su kratky* ‘my hair is short’ (de, A1.1); **moje/moja rodina je veľka* ‘my family is big’ (pt, A1.2); **moj/moje leto minuli rok bol velmi nudný* ‘my last summer was very boring’ (tr, A1.2); **moja/moje *obľúbená/obľúbené leto bola toto leto* ‘my favourite summer was this summer’ (ro, A2); **jeden/jedna veľmi dôležitá inštitúcia* ‘a very important institution’ (tr, B1). Such cases of grammatical structures variability reflect an approximate and transitional state in the acquisition of the grammatical system of the target language.

3.2 Errors in the case

Incorrect congruence in substituting the case suffixes of the congruent attribute occurred to a small extent at the A1 and A2 levels. A higher incidence was observed at the B1 and B2 levels. At the A1 and A2 levels, this was mainly seen in the substitution of indirect case forms by nominative forms (singular and plural): e.g. *bývam na *Veliko/Velikom Tarnove* ‘I live in Veliko Tărnovo’ (bg, A1.1); *nájdete tu tiež *24-hodinová/24-hodinovú lekáreň* ‘you can also find here a 24-hour pharmacy’ (es, A1.2); *potrebuješ *optická/optickú myš* ‘you need an optical mouse’ (es, A1.2). Inconsistency in the acquisition of the congruence category was shown by relatively frequent collocations, where within one sentence structure there were correct and incorrect forms next to each other: e.g. *to je *mojú/moja obľúbená farbu* ‘it is my favourite colour’ (fr, A1.1); *môžete nájsť *zaujímave/zaujímavé veci, počítače, filmy a *ďalších/ďalšie *elektronickych/elektronické veci* ‘you can find interesting things, computers, movies and other electronic things’ *(es,* A1.2); *byvam s *mojím/mojimi rodičmi moja sestrinčka a *starym/starými rodičmi* ‘I live with my sister and grandparents’ (ro, A2.2).

Occasionally, there were some cases with higher language levels where the attribute was used in the basic (nominative) form: *dobré vzdelanie nie je len dôležité pre hospodorstvo ale pre *celá/celú spoločnosť* 'good education is not important for the economy, but for the whole society' (de, B1); *bývam vo *Velko/Velikom Tarnove* 'I live in Veliko Tărnovo' (bg, B1); *ako vidím *moja/moju budúcnosť v práce* 'how I can see my future at work' (hu, B2). At the B1 level, there were also phenomena where one of two attributes was wrong: *sa exportujú do Ruska, Nemecka, USA, Izraela, Českej republiky, Cypru, Spojených *arabským/arabských emirátov* 'are exported to Russia, Germany, the US, Israel, the Czech Republic, Cyprus, the United Arab Emirates' (be, B1); *mám priemernú a *štíhla/štíhlu postavu* 'I have an average and slim figure' (ko, B1).

From the A1 to B2 levels, we found cases of form analogy of the adjectival and substantival suffix: *pod *starom/starým mostom* 'under an old bridge' (de, A1.1); *som *Talianskom/talianskym študentom* 'I am an Italian student' (it, A2.1); *spoznala som sa s ludiam z *celom/celého svetom* 'I met people from all over the world' (de, B1); *z *tom/tohto dôvodom* 'from the reason' (de, B1); *vzťahov medzi *nejakom/nejakým občanom a úradnikom* 'relationships between a citizen and an officer' (it, B1); *najdôležitejší je pochod ludí *zamaskovaní/zamaskovaných za postavu nazvanú Gilles* 'the most important is a march of people masked as Gilles' (fr, B1); *v rodine tiež musí byť *dobrom/dobrým otcom alebo dobrou matkou* 'in the family he also must be a good father or a good mother' (bg, B2). At the B2 and C1 levels, we noticed errors in apposition: e.g. *práve sa pripravoval otvoriť list, napísany na hrubom pergamente, *taký/takom *istý/istom ako obálku* 'he was just about to open a letter written on a thick parchment, the same as the envelope' (bg, B2). In the plural forms of attributive-substantive phrases, there were no significant trends in favour of any case.

At the beginner's level, there was a negative transfer from Russian in the case of the suffix *-e* within the soft adjectival declension in Russian: e.g. *hodila som v les spolu zo *starej/starou mamou* 'I used to walk in the woods with my grandmother' (ru, A2.2), *so *svojej/svojou rodinou* 'with my family' (ru, A2.2).

In the case of students with Polish as an L1 language, we can point out transfer from this language: e.g. *keby som sa narodila v úplne *iným/inom mieste* 'if I was born in a completely different place' (pl, B1); *na *týmto/tomto zápase som nebol* 'I did not attend the match' (pl, B1); *v *každým/každom dome sa rozprava o politike* 'in every house people talk about the politics' (pl, B1); *keby som sa narodila v inom štáte, v inej rodine alebo aj vo *vedlajším/vedľajšom meste* 'if I was born in a different state, different family or in a neighbouring city' (pl, B1). Unlike in Slovak, the LOC[5] sg. and INS sg. forms of adjectives referring to masculine and neuter have the same

[5] In the paper, we use an international abbreviation terminology for grammatical cases: NOM – nominative, GEN – genitive, DAT – dative, ACC – accusative, LOC – locative, INS – instrumental.

grammatical suffix *-ym/-im* in Polish (cf. in Polish *w każdym domu* – in Slovak *v každom dome* – in English *in every house*; in Polish *w nowym samochodzie* – in Slovak *v novom aute* – in English *in a new car*). Students with Bulgarian as an L1 language tend not to decline numerals in an attributive position: e.g. *ja a ona sme spolú (kamaratky) od 6 (*šesť/šiestich) roky* 'she and I have been friends since 6' (bg, A1.2); *tento deň sa oslavuje aj ako deň Svätí *Štyridsať/štyridsiatich mučeníkov* 'this day is the Forty Martyrs celebration' (bg, B1).

Some of the examined phenomena can be seen as performance errors overlapping with spelling errors (omitting a grapheme at the end of the instrumental case): e.g. *hovoril o svojom priateľstve s rôznymi politikmi a *dôležitým/dôležitými ľudmi v našej krajine* 'he spoke about his friendship with different politicians and important people in our country' (hr, C1); *pohľad *Judášovým/Judášovými očami* 'through the eyes of Judas' (hr, B2).

3.3 Errors in the masculine person

Quantitatively fewer represented errors than in cases and the wrong grammatical gender were recorded in the category of the masculine person: "In the plurals in the West Slavic macroarea, a category of masculine person has emerged. Currently, the masculine person in Slovak, Polish, and Upper Sorbian is realized with special endings for a personal masculine noun in the plural nominative and in case homonymy between the genitive and accusative plural (in Upper Sorbian also the dual form)" (Kamenárová 2015, p. 201). The errors occur at all language levels, regardless of whether there is a masculine person category in L1 or not.

In the group of indefinite and limitative pronouns, even at the advanced levels, there were many errors in agreement with the subordinate noun *ľudia* 'people': *ani rodičia, ani *iné/iní *známe/známi ľudia* 'neither the parents, nor other famous people' (uk, B1); **íne/iní ľudia maju radšej letne športy* 'other people like different sports' (de, B2); *naozaj sú aj *také/takí ľudia* 'there indeed are such people' (uk, B2); *sú *také/takí ľudia, ktoré* 'there are people who' (bg, B2); *rešpektovali *iné/iných ľudi* 'they respected other people' (sl, C1); *je najoptimálnejšie pre *niektoré/niektorých ľudí* 'is the most optimal for some people' (ru, C1).

The possessive pronouns in the studied sample do not form a complete paradigm. They are represented only by the forms *môj* 'my', *náš* 'our', and *svoj* 'my'. The incorrect congruence was found mainly in the phrase with the noun *deti* 'children': *aby *naši/naše deti rástli* 'for our children to grow' (bg, B2). The syntactic phrase of **moje/moji kamart rads kino* 'my friend/s like/s going to cinema' (es, A1.2) also contains spelling errors. It can be concluded that the author either wanted to write the form of the singular *môj kamarát* 'my friend' or the plural *moji kamaráti* 'my friends'; although there was an error in both forms, neither of them contains the correct form of NOM pl. There is another occurrence in the following examples: *finančne zabezpečovať *svojich/svoje deti a svoju ženu* 'financially secure my

children and my wife' (uk, B2); *pozerať a počúvať za *svojich/svoje deti* 'to watch and listen to my children' (uk, B2). Here, instead of the correct form of ACC pl., *svoje deti* 'my children' is in GEN pl., which can be interpreted in two ways. The student categorized the lexeme *deti* 'children' as an animate noun and declined it as a masculine person noun or used this form because of the fixed phrases with a verb (*zabezpečovať* 'secure', *počúvať* 'listen to').

There are still very few examples of error occurrences in demonstrative pronouns in the learner corpus (the size of the pilot version of the corpus), but this does not reflect reality, where it turns out that demonstratives (especially composite ones) are just as demanding for students with a Slavic L1 as they are for non-Slavs. Possible causes of the cognitive difficulty in the acquisition of pronouns include internal flexion and the formal similarity of the lexeme (*toto* 'this' – *tieto* 'these' – *tamtie* 'those'*, táto* 'this' – *tieto* 'these' – *tamtie* 'those'*, tento* 'this' – *títo* 'these' – *tamtí* 'those', etc.). Extending the corpus material to other texts and increasing examples of this phenomenon will enable us to confirm or refute this hypothesis in the future. The form of *títo* 'these' in **títo/táto vec dajú vplyv pre divákov* 'these things have an influence on spectators' (tr, B1) may indicate an insufficient knowledge of the grammatical gender of the noun *vec* 'thing'.

Similarly to pronouns, there was a frequently occurring error in qualitative adjectives with the noun *ľudia* 'people': **nové/noví ludia sa prestahaiu do mesta* 'new people will move to the city' (de, B1); **staršе/starší ľudia pozerajú na to všetko inak* 'older people see it all differently' (uk, B2); *nikto viac... ani rodičia, ani *iné/iní *známe/známi ľudia* 'noone else, neither parents, nor other famous people' (uk, B1); *oni su moje *oblubene/obľúbení ľudia* 'they are my favourite people' (ro, A2.2). Other examples with incorrect attributes include *vidíme *nových/nové vystávi maliarov* 'we can see new painters' exhibitions' (fr, B2); **vlastné/vlastní rodičia nemôžu povedať aspoň nejaké dobre slovo* 'own parents cannot say at least a good word' (uk, B2); and *v minulosti boli to *najstaršie/najstarší členovia kmeňa alebo silní vojvodcovia* 'in the past they used to be the oldest members of the tribe or strong dukes' (uk, B2). With the zero grammatical suffix of the noun (*osobnosť-0* 'personality', *slávnosť-0* 'celebration'), students' formal approach to gender probably motivated them to use the masculine. This can be seen in the following examples, where we can also observe the seemingly correct form of the adjective as students assumed that the noun was used in the masculine: *najväčšiu pozornosť vždy získavajú *kontroverzní/kontroverzné osobnosti* 'controversial personalities always get the most attention' (ru, C1); *konajú sa známe *operní/operné slávnosti Richarda Wagnera* 'the famous Richard Wagner Opera Festival is held' (de, B1). This shows that it is "particularly difficult for students of Slovak as a foreign language [...] to acquire the grammatical gender of atypically terminated feminines, to which the speakers, based on the zero suffix, usually mechanically assign the masculine gender" (Spáčilová 2019, p. 208).

This category also includes adjective-substantive collocations, where students correctly identified the soft consonant in the adjective, which then requires frontal unlabialized vowels, but then incorrectly chose the grapheme *i*, which in Slovak expresses a consistency with the personal masculine in the plural: *v Číne nájdeme *najvyšší/najvyššie hory* 'in China there are the highest mountains' (zh, A1.2); *zaujimal som sa aj o *cudzí/cudzie jazyky* 'I was also interested in foreign languages' (de, B2).

The incorrect grammatical morpheme in phrases with the noun *muž* 'man' – *priala by som si *čistý/čistého muž a *vysoký/vysokého* 'I wish for a loyal and tall husband' (ro, A2.2) – shows the form NOM sg. instead of ACC sg. In this example, the category of masculine person is not realized. Again, this error is common for learners where expressions with the masculine person are required; however, we are not able to document it sufficiently in the corpus.

3.4 Spelling errors *-i/-y* and *-í/-ý*

In this category, we present forms that formally reflect only the exchange of the *i/y* and *í/ý* graphemes. Unlike the masculine person category, this type of error is typical for beginners, reflecting the lack of acquisition or grasp of the orthographic principles of Slovak. Any erroneous forms included in this category imply the confusion of the NOM sg. and/or NOM pl. of the attribute modifying the defining masculine noun. We did not classify these occurrences into the masculine person category, because students "phonetically" expressed their agreement with the masculine but with the incorrect spelling, which is a common mistake even among native speakers. Most often, there is a mistake in the agreement with the defining noun *ľudia* 'people' – e.g. *existuje veľa šikanovania a nevhodnych veci čo *mlady/mladí ľudia by nie maly vidiet* 'there is a lot of bullying and inappropriate things that young people should not see' (en, B2); **mlady/mladí ľudia by nie mali mať sociálne siete* 'young people should not join the social media' (en, B2); *milovať a normálne žiť ako *ostatný/ostatní ľudia* 'to love and live like other people do' (pl, B2); the noun *rodičia* 'parents' (*moji *starý/starí rodičia oslavovať 50 rokov od zvadba minylý rok* 'my grandparents celebrated the 50th wedding anniversary last year' (ro, A2.2); *moja najlepšia kamarátka, a moji *starý/starí rodičia* 'my best friend and my grandparents' (sr, B1); and the noun *vojak* 'soldier' (*v roku 1939 tiež boli zavraždeni *poľský/poľskí vojáci* 'in 1939 Polish soldiers were also assassinated' (pl, B1); **nemecký/nemeckí vojaci Hermana Thielea zabili Jana a Štefana* 'German soldiers of Hermann Thiele murdered Jan and Stephan' (sr, B2).

The orthographic errors include ordinal numerals: *žijem so svojim starým otcom a starou mamou a stali sa mi ako *druhy/druhí rodičia* 'I live with my grandfather and my grandmother and they became my second parents' (hr, B2).

The research sample also presented some cases where an erroneous grammatical morpheme was present only in one of the multiple premodifying adjectives: *niektorí *katolícky/katolícki kňazi, chcú ovplyvňovať bežný život* 'some Catholic priests want

to influence the ordinary life' (pl, B2); *mnohí *literárny/literárni kritici a čitatelia spajujú Marakéš s mestečkom Papín* 'many literary critics and readers connect Marrakesh with the city Papín' (pl, C2).

4 CONCLUSIONS

This paper demonstrated several types of error rates in the congruent attribute by analysing language material from the learner corpus errkorp-pilot. Given the limited textual scope in this corpus, qualitative research methods were applied instead of quantitative ones. The most represented group of errors was mistaking the grammatical gender. Errors in gender occurred at all language levels from A1 to C2. While at the beginner levels, most of the errors were interlingual (the students "transferred" the gender from their L1 language) or systematic (the students had not yet acquired the knowledge of which nouns ending in the basic form of a zero grammatical suffix were feminine and which ones were masculine). The errors at the higher language levels seem to have been caused by distraction (performance errors). The material confirmed the empirical experience that the words *dievča* 'girl', *múzeum* 'museum', and *štúdium* 'study' (or other nouns ending in *-um*) were syntagmatically problematic as errors in congruence also occurred in these words at the B2 to C2 levels.

Errors in the attribute's incorrect congruence and the defining noun's case had the largest occurrence at the B1 and B2 levels, which again reflects empirical pedagogical experience. Although systematic errors dominated in this group (in particular, the cases of a form analogy of the adjective and substantive suffix in the case of non-Slavic students), followed by performance errors, there were also cases of transfer of case suffixes from Russian and Polish in the case of students for whom Russian and Polish are L1 languages.

There were errors in the masculine person category where the masculine grammatical morphemes were not realized, while the defining noun was, for example, the nouns *ľudia* 'people', *rodičia* 'parents', and *študenti* 'students'. In most languages that learners have identified as their L1 language, the category of the masculine person does not exist at all. However, in the L1 languages where the masculine person category is expressed, there was an absence of relevant grammatical morphemes. This may be due to several factors (e.g. not knowing the grammatical rule, insufficient proficiency in the rule, or an overgeneralization of the rule of the non-masculine person category).

Spelling errors included the disagreement of adjectives, pronouns, and numerals with the subordinate noun. The errors were caused by the confusion of the NOM sg. and/or NOM pl. masculine nouns. In the examined material, the most frequently error congruence was with the subordinate noun *ľudia* 'people'.

The analysis showed that erroneous attributes, in general, were most often associated with the nouns *ľudia* 'people', *študenti* 'students', and *rodičia* 'parents'.

This is natural since these are frequent nouns in the language as well as in the communication of learners of a given age group, so even attributes such as syntactic additions to these nouns may often contain erroneous suffixes.

The paper is a part of the broader research of linguistic errors in Slovak as a foreign language based on learner corpus ERRKORP, focused on all linguistic levels. The complex results will be published in a monograph in 2024.

ACKNOWLEDGEMENTS

This work was supported by the Slovak Research and Development Agency within contract No. APVV-19-0155 *Language Errors in Slovak as a Foreign Language Based on Learner Corpus*.

References

Imrichová, M. (2023). Slovenčina ako cudzí jazyk na zahraničnej slovakistike (výsledky parciálnych výskumov vybraných jazykových javov). In M. Mošaťová – P. Kollárová (eds.): Slovenčina (nielen) ako cudzí jazyk III. Zborník príspevkov venovaných výskumu a výučbe. 1. zväzok. Bratislava: Univerzita Komenského v Bratislave, pages 386–401.

Ivanová, M. (2020). Syntax slovenského jazyka. 2. dopl. vyd. Prešov: Vydavateľstvo Prešovskej univerzity, 283 p.

Kačala, J. (2014). Jazykové kategórie v slovenčine. Bratislava: Vydavateľstvo Univerzity Komenského, 179 p.

Kamenárová, R. (2015). Gramatické kategórie maskulín a ich vplyv na vývin západoslovanských jazykov. In K. Balleková – Ľ. Králik – G. Múcsková (eds.): Jazykovedné štúdie XXXII. Prirodzený vývin jazyka a jazykové kontakty. Bratislava: VEDA, pages 195–202.

Kotková, R. (2017). Čeština nerodilých mluvčích s mateřským jazykem neslovanským. Praha: Karolinum, 154 p.

Luță, M. F. (2019). Inter-language and trans-language interferences in Romanian students of Slovak studies. In É. Császári – M. Imrichová (eds.): Király Péter 100. Tanulmánykötet. Király Péter Tiszteletére I. Budapest: ELTE BTK, Szláv Filológiai Tanszék, pages 263–273.

Misadová, K. (2011). Kapitoly z morfológie maďarského jazyka. Kontrastívny opis niektorých morfologických javov maďarského jazyka. Vysokoškolské skriptá pre poslucháčov maďarčiny ako cudzieho jazyka. Bratislava: Univerzita Komenského v Bratislave, 131 p.

Mošaťová, M., and Gajdošová, K. (2019). Annotations in the corpus of texts of students learning Slovak as a foreign language (ERRKORP). Jazykovedný časopis, 70(2), pages 345–357.

Pekarovičová, J. (2020). Slovenčina ako cudzí jazyk – predmet aplikovanej lingvistiky. Bratislava: Stimul, 200 p.

Spáčilová, S. (2019). Slovenčina v Maďarsku. Problémy používania numeratívneho genitívu a gramatickej kategórie menného rodu. In É. Császári – M. Imrichová (eds.): Király Péter 100. Tanulmánykötet. Király Péter Tiszteletére I. Budapest: ELTE BTK, Szláv Filológiai Tanszék, pages 274–282.

sciendo
DOI 10.2478/jazcas-2023-0035

TOWARDS A CORPUS-BASED DICTIONARY OF VERBAL GOVERNMENT FOR THE RUSSIAN LANGUAGE

EDUARD KLYSHINSKY[1] – ANNA BOGDANOVA[1]
– MIKHAIL KOPOTEV[2,3]
[1] Independent researchers
[2] Department of Languages, University of Helsinki, Helsinki, Finland
[3] Department of Slavic and Baltic Studies, Stockholm University, Stockholm, Sweden

KLYSHINSKY, Eduard – BOGDANOVA, Anna – KOPOTEV, Mikhail: Towards a Corpus-based Dictionary of Verbal Government for the Russian Language. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 173 – 181.

Abstract: This paper introduces a technique for automatic verbal governance extraction in the Russian language, which encapsulates information on the grammatical features of verb-noun co-occurrences, encompassing both prepositional and non-prepositional dependencies. The construction of the dictionary, a corpus of approximately 3.5 billion words was used. The proposed method involves syntactic parsing of the texts, filtering of resultant outputs, and creating a dictionary of prepositional government. After error filtering, the dictionary contains ca. 18,000 verbs along with NP/PPs governed by these verbs.

Keywords: verbal government, automatic extraction, Russian language

1 INTRODUCTION

Collocational dictionaries hold a significant position in Russian lexicography. However, their utility is often constrained by their relatively modest size and sub-optimal accuracy. Prokopovich et al. (1981) offer combinations for 1219 Russian words, alongside theoretical discourse on nominal and verbal governance. Denisov et al. (1983) furnish a comprehensive description of the 2506 most frequent Russian words, including 727 verbs. Meľčuk and Zholkovsky (1984/2016), among others, supply exhaustive, standardized details concerning word government structure. Regrettably, the dictionary encompasses only 283 entries.

Presently, extensive efforts are being directed towards the development of the Russian Active Dictionary (Apresjan et al. 2014–2017, which encapsulates not only word meanings but also data pertinent to speech production, such as combinatorial characteristics and pragmatic conditions of word usage. Among a variety of features, the dictionary incorporates information on verbal governance, including the most commonly used cases and prepositions marking prepositional and non-prepositional

dependencies. This data is invaluable for researchers and learners of Russian but it may be insufficient for the creation of Natural Language Processing (NLP) systems, which aim to present a comprehensive and complete list. Consequently, existing dictionaries, while they contain a limited number of entries, are infrequently presented in a machine-readable format.

From a Computational Linguistics perspective, the compilation of robust electronic lexicographic resources for the Russian language is a labor-intensive process, and it has seen the development of numerous automated techniques. Most of these methods leverage data from the Russian National Corpus (RNC), a compendium of Russian texts with sophisticated linguistic annotation. An example of these is represented in Biryuk et al. (2008), encompassing 10,015 verb combinations with abstract nouns filling the patterns 'Noun + Verb', 'Verb + Noun', and 'Verb + Adjective + Noun'. This dictionary signifies an attempt to establish a digital resource, grounded in Mel'čuk's Meaning-Text theory, yet employing large corpus data. All combinations within the dictionary are classified according to lexical functions—semantic labels corresponding to case roles, as proposed by Ch. Fillmore. Another online resource for word co-occurrences is the database of Russian lexical constructions known as FrameBank (Lyashevskaya et al. 2011). It comprises a corpus of the 2,500 most frequent Russian verbs and verbal constructions, accompanied by a description of their governance patterns (syntax annotation and semantic role labeling).

Two projects were initiated by the authors of this article. The project 'CoCoCo: Collocations, Colligations, and Constructions' (Kopotev et al. 2015; Kormacheva et al. 2014), utilizes three corpora, RNC, I-RU, and Taiga, to showcase not only collocations but—as indicated by the title—also colligations (grammatical patterns) and constructions (grammatical patterns supplemented by lexical variables). Another resource, called CoSyCo, is introduced in Klyshinsky et al. (2018). It comprises syntactic patterns extracted from a corpus of approximately 17 billion tokens. The resource capitalized on the grammatical and semantic relations between tokens and is primarily devised for NLP and CL tasks.

Consequently, among the current resources targeted at word combinations, those specifically designed for verbal patterns are underrepresented. Both print and online dictionaries are limited in size, online resources are somewhat unspecific in this aspect since they focus on any word collocations for all parts of speech (POS), and they do not incorporate information specific to verbal governance.

With this in mind, the verbal collocation dictionary would serve many users. Firstly, it could be utilized by foreigners learning Russian, as it provides information not only about combinations but also the frequency along with a comprehensive list of examples. Secondly, it may captivate researchers in the fields of theoretical and descriptive linguistics, as it provides a distribution of verb combinations among different genres and semantic classes. Finally, the outcomes of this work may be harnessed by researchers in the field of Natural Language Processing (NLP) as

a benchmark for the evaluation of automated systems. The primary objective of this paper is to describe a corpus, which fulfils specific requirements. Firstly, the corpus must encompass as many diverse verbs as possible to be applicable to NLP tasks; secondly, the corpus must contain less than 5% of incorrect combinations of verbal patterns.

The rest of the paper is structured as follows: Section 2 elucidates the proposed method for the automatic composition of the dictionary of verbal governance. Section 3 furnishes details about the experimental data and results used. The remainder of the paper encompasses a brief discussion of the experimental results and the limitations of the method.

2 PROPOSED METHOD

To fulfil our research objectives, we constructed a syntactically annotated corpus of Russian texts. This enabled the computation of frequencies for syntactic patterns featuring verbal heads and nominal/prepositional dependencies. Nevertheless, the annotated corpus manifested a fairly high degree of errors, necessitating painstaking preprocessing. Consider, for instance, the following sentence:

(1) *Раз за разом в библиотеки приходили новости, которые давали всё меньше подробностей.*
'There came news in the libraries, again and again, which provided fewer and fewer details.'

Theoretically, parsing this sentence could engender multiple errors, thereby compromising statistical data. Initially, the expression *раз за разом* (lit. 'time-NOM for time.INS'; 'again and again' is linked to the verb *приходили* 'come-PAST.PL', which does not govern the nominative case in this position; thus, it would be erroneous if linked to the verb. Secondly, the fixed expression *всё меньше* 'less and less' necessitates the genitive case for the noun *подробностей* 'detail-GEN.PL'; however, the parser could incorrectly associate the noun with the verb *давали* 'give-PAST.PL'. Lastly, the phrase *в библиотеки* 'in the library' can be parsed as в 'to/in' + library-ACC.PL (the most probable analysis) or as 'to/in' library-GEN.SG (incorrect, but still available in a parser). Consequently, the raw frequencies failed to provide relevant data for dictionary construction. This realization necessitated the development of a novel method for text preprocessing.

In the first phase, the lexicon of prepositional governance was established by assembling a comprehensive list of all Russian prepositions and the cases governed by them. The statistics were calculated over the syntactically annotated corpus and normalized solely for prepositions. All instances that exhibited a relative frequency of

less than 1% for a given preposition were considered marginal and, thus, were omitted. These data were inspected and corrected by an expert. The result is a list of prepositions and the cases they govern, henceforth referred to as the Lexicon of Prepositional Governance (hereafter LPG). The list incorporates 132 prepositions, including some compounds such as *за счёт* 'by means of' or *в течение* 'during'. These prepositions may govern, in different combinations, five cases: genitive, dative, accusative, instrumental, and locative. The two marginal cases in Russian, the second genitive (partitive) and second locative, are considered the genitive and locative cases, respectively.

In the second phase, we compiled statistics on the co-occurrence of all prepositional phrases for all verbs in the corpus. Any instances of the cases, which are not attested in the LPG, were eliminated. It is important to note that filtering is not the best decision because many low-frequency co-occurrences are known to be rare, mainly idiomatic, word combinations, necessitating further investigation. Conversely, the data frequently contain a substantial number of errors induced during both text production and parsing. For our project, a classic trade-off between precision and recall is that we elected to decline instances according to the frequency-based filtering.

In the third phase, we examined the non-prepositional case government, which presents challenges due to the specific features of Russian syntax. For instance, a direct object is marked, albeit non-automatically, with the genitive case if a verb is used under negation. This variation isn't exclusive to a specific verb but rather applies to any negated verb. After a thorough preliminary investigation, we opted to apply the following filters to avoid standard variations in Russian syntax:

- All verbs including auxiliary ones should not be negated: *предвещать беду* 'to portend trouble-ACC' becomes *не предвещать беды* 'not to portend trouble-GEN';
- A noun should not form part of a numeral group and, therefore, should not be governed by or agreed with numerals or quantitative adverbs: *трое грустных мужчин* 'three sad men-GEN' VS *три счастливые женщины* 'three happy women- NOM'.

Returning to example (1) above, the genitive case for *библиотеки* 'library-PL. GEN' can be filtered out because it is below the threshold for the preposition в 'to/in'. The phrase *давали всё меньше подробностей* 'gave less and less detail-PL. GEN' is also out of the scope here, since it contains a quantitative adverb.

Despite these filters proving effective in reducing noise, they have not entirely eradicated it. Consequently, we decided to filter out all combinations with a relative frequency below a certain threshold: 5% for genitive, 0.2% for dative, and 1% for accusative and instrumental cases. The locative case has no threshold as it cannot be used without a preposition. Similarly, the nominative case was completely excluded due to its standard syntactic linkage with virtually any verb, save a few exceptions. To encapsulate, our approach comprises several stages:

- assembling the syntactically tagged corpus;
- computing statistics of co-occurrence for verbs and dependent NP/PPs;
- formulating a lexicon of prepositional governance using calculated statistics and implementing expert filtering;
- creating a list of dependent prepositional phrases for all verbs and filtering it following the lexicon of prepositional governance;
- and constructing and filtering a list of nominal dependencies for all verbs.

3 EXPERIMENTAL SETUP AND THE RESULTS OF EXPERIMENTS

3.1 Used texts and tools

The textual corpora employed in this study are detailed in Tab. 1. These collections, sourced from the Internet, were tagged using the DeepPavlov parser (Burtsev et al. 2018) due to its superior accuracy. For our computations, we omitted non-dictionary words, given that DeepPavlov utilizes a neural network for lemmatization, which invariably generates a significant volume of unknown 'lemmas'. For the filtration process, we employed the OpenCorpora (Bocharov et al. 2011) lexicon, implemented in the Pymorphy2 library.

Collection	Number of sentences	Number of words	% of sentences	% of words
Ph.D. and doctoral thesis	26,764,667	560,907,459	14.45%	16.06%
General news	42,572,120	819,591,304	22.98%	23.47%
Thematic news	20,679,206	413,693,321	11.16%	11.85%
Fiction texts	57,230,401	799,596,093	30.89%	22.90%
Wikipedia	25,932,220	544,161,541	14.00%	15.58%
Official texts	12,082,916	354,259,132	6.52%	10.14%
Total	185,261,530	3,492,208,850	100.00%	100.00%

Tab. 1. Size of the text collections

3.2 Results of experiments

Two distinct representations of the dictionary are proposed, each contingent on its specific application. The first is geared toward Russian language acquisition. For this purpose, we have selected the 80%-quantile of the most frequent combinations of verbs and nouns in the nominative case. These combinations encapsulate the most practical elements of verbal governance for a student to master. The remaining combinations, though less common, comprise up to 4% of usage for the verbs under consideration.

A portion of the dictionary is exhibited in Tab. 2, where "-" and prepositions followed by grammatical cases represent a syntactic connection between a noun phrase (NP) and a prepositional phrase (PP), respectively, and a verb. The figures

represent the proportion of such syntactic patterns for a specific verb. The nominative case is not represented in Tab. 2; consequently, the total for a single row may not equal 80%.

verb	NP/PP	%	NP/PP	%	NP/PP	%
организовать 'to organize'	-Acc	37.9	*в* 'in' -Loc	10.4	-Ins	9.6
существовать 'to exist'	*в* 'in' Loc	15.2	*на* 'at' -Loc	4.7		
выполнять 'to execute'	-Acc	69.8				
разработать 'to develop'	-Acc	32.9	-Ins	10.5	*в* 'in' -Loc	8.3
рассмотреть 'to consider/to view'	-Acc	53.0	*в* 'in' -Loc	8.5		
обладать 'to possess'	-Ins	72.4				
поднять 'to lift'	-Acc	58.5	*на* 'at' -Loc	6.0	*в* 'in' -Loc	3.3
использоваться 'to be used'	*для* 'for' -Gen	20.9	*в* 'in' -Loc	16.5	-Ins	4.7
закрыть 'to close'	-Acc	44.7	-Ins	6.9	*в* 'in' -Loc	4.4
пользоваться 'to use'	-Ins	61.2				

Tab. 2. Examples of verbal government in the human-readable format

Our proposed methodology facilitates the compiling of a dictionary comprising 17,367 verbs. Furthermore, it includes 1,510 verbs wherein only the nominative case is attested, i.e., instances where the nominative case is utilized in over 80% of all co-occurrences. A few examples of such verbs are *засориться* 'to clog', *обесточиваться* 'to de-energize', *настать* 'to come' (e.g. the time has come), *расцениваться* 'to apprise', and *прозвенеть* 'to ring out'.

It is important to note that we have only utilized combinations that are represented in our corpus at least twice; there are examples with a frequency of less than 1 instance per million, which means they are extremely rare in the data. They necessitate a more flexible threshold and/or the need for more comprehensive data for their effective handling. Two further filtering criteria that we applied include the requirement that a specific verb-noun pattern must have a minimum frequency of 4–5% among all patterns for that verb and that the number of verbal arguments is limited to four or five.

3.3 Evaluation of results

To evaluate recall, we compared our results with three hundred of the most frequently used verbs in the Active Dictionary of Russian (AD; Apresjan et al. 2014–2017), which consists of 1,073 verbs, while our dictionary provides 9,045 verbs. Depending on frequency, 78–90% of combinations from the AD are found in our dictionary (the lower the frequency, the higher the recall), and only 55% from our dictionary are found in the AD. The AD provides an average of 2.28 combinations per verb compared to 19.32 in ours, but the former includes semantic relations such as purpose and direction introduced with dependent clauses, which are omitted in our dictionary. The recall can be seen as rather low, however, however our methodology identifies statistically significant discrepancies in the usage of aspectual pairs that are consistent with the findings of previous research (Janda et al. 2013). Adopting a more lenient threshold for noun frequencies could potentially facilitate the ex- traction of these patterns as well, thus the recall will be higher.

To assess the precision, we manually examined the 300 most commonly used verbs and discovered that our method commits less than 5% errors, with one stipulation. We considered noun phrases indicating time, frequency, logical inference, and similar concepts as appropriate usage. This can be observed in the phrase *прийти вечером* 'to arrive in the evening.' The word *вечером* can be categorized as a Noun.Ins or an Adverb, contingent upon the adopted theoretical framework. Such instances straddle the boundary between a proper noun phrases and adverbial ones.

After deeper analysis, we discovered that many errors in combinations pertained to the dative case in noun phrases – 17 out of top 100 verbs were errorneously associated (4 times) or disassociated (13 times) with the dative. The second most common errors were in the instrumental case (6 errors among top 100 verbs). Some of these errors can be ascribed to the parser's preferences in creation of syntactic dependencies in case of unconventional word order. For instance, a word in the instrumental case might be erroneously linked to a regular verb at a close distance, rather than to the auxiliary verb it should actually connect with. This could be attributed with the lower threshold value; however, we are faced with the classic precision VS recall conundrum, which should be addressed in regards to the purposes, which we discuss in the conclusion.

4 CONCLUSION

In this study we introduced a method for the automated construction of a dictionary of verbal governance. This dictionary includes a compilation of verbs accompanied by details about governed prepositional and non-prepositional phrases. A corpus of 3.5 billion tokens was employed to accumulate representative data for the construction of this dictionary. The efficacy of the method hinges on the quality of the syntactic parser utilized.

The approach generates approximately 5% non-attested errors for non-prepositional and roughly 3% for prepositional syntactic groups. It's worth noting that many contemporary NLP tools demonstrate a comparable level of accuracy and are still successfully employed in practical applications.

However, our project proves particularly beneficial for learners of the Russian language, as verbal governance is a common classroom topic. In this respect, the 5% error rate could be considered unsuitable for learners of the Russian language, as the resource we offer could potentially lead them astray. To circumvent this limitation, we have devised a condensed version of the dictionary that only includes the most prevalent combinations. This streamlined version boasts fewer errors and is more user-friendly. The application of additional filters in the future could serve to further refine these results.

Furthermore, verbal governance is intrinsically tied to verbal semantics, which we did not delve into in this study. We believe that our work lays the groundwork for future investigations by providing relatively unambiguous data for further exploration. For this project, we collected all noun phrases connected to the verb directly or through prepositions, along with their grammatical information. Clustering nouns into semantic groups is a matter for our future research. Another issue we in- tend to address in our work is the distinction between verbal arguments, which are semantically linked to a specific verb (e.g., to read a book), and adjuncts, which apply to an entire class of verbs (e.g., *to read at the table*). The resulting dictionary will undergo proofreading and will be made accessible in the Git repository of the project: https://github.com/klyshinsky/Slovko-2023.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to Prof. Olga Lyashevskaya for her insightful discussions on Russian syntax, as well as to ChatGPT for proofreading this paper.

References

Apresyan, Y. D. et al. (2014–2017). Active Dictionary of Russian [Activnyj slovar' russkogo yazyka]. Moscow: Yazyki slavjanskoj kultury, vol. 1–3.

Biryuk, O. L., Gusev, V. Y., and Kalinina, E. Y. (2008). Dictionary of Russian Abstract Nouns' Verbal Collocability [Slovar' glagol′noj sochetaemosti nepredmetnyh imen russkogo yazyka] Accessible at: http://dict.ruslang.ru/abstr_noun.php?.

Bocharov, V. V., and Granoskiy, D. V. (2011). Software for Collaborative Work on Morphological Tagging of a Corpus [Programmnoe obespechenie dlya kollektivnoj raboty nad morfologicheskoj razmetkoj korpusa]. Proc. of "Corpus Linguistics – 2011" (27-29 June 2011, Saint-Petersburg), p. 348.

Burtsev, M., Seliverstov, A., Airapetyan, R. et al. (2018). DeepPavlov: Open-Source Library for Dialogue Systems. Proc. of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 1–6.

Denisov, P. N., and Morkovkin, V. V. (1983). Combinatorial Dictionary of the Russian Words [Slovar' sochetaemosti slov russkogo yazyka]. Moscow: Russkij yazyk, 2nd ed., 688 p.

Klyshinsky. E. S., Lukashevich, N. Y., and Kobozeva, I. M. (2018). Creating a corpus of syntactic co-occurrences for Russian. Computational Linguistics and Intellectual Technologies. Proc. of "Dialogue 2018" (30 May–2 June 2018, Moscow), pages 305–316.

Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., and Yangarber, R. (2015). CoCoCo: Online Extraction of Russian Multiword Expressions. The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria), pages 43–45.

Kormacheva, D., Pivovarova, L., and Kopotev, M. (2014). Automatic Collocations Extraction and Classification of Automatically Obtained Bigrams. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations, Tübingen, 2014, pages 27–33.

Lyashevskaya, O., and Kashkin, E. (2015). FrameBank: a database of Russian lexical constructions. Proc. of Analysis of Images, Social Networks and Texts. 4th Int. Conf., AIST 2015, pages 337–348.

Meľčuk, I. A., and Zholkovsky, A. K. (2016). Explanatory Combinatorial Dictionary of Contemporary Russian [Tolkovo-kombinatornyj slovar' sovremennogo russkogo yazyka]. Moscow: LRC Publishing House, 2nd ed.

Prokopovich, N. N., Deribas, L. A., and Prokopovich E. N. (1981). Nominal and verb government in modern Russian language [Imennoe i glagol′noe upravlenie s ovremennom russkom yz̃yke]. Moscow: Russkij yazyk, 2nd ed.

Richardson, K. R. (2007). Case and Aspect in Slavic. Oxford, OUP, 288 p.

sciendo
DOI 10.2478/jazcas-2023-0036

THROUGH DERIVATIONAL RELATIONS TO VALENCY OF NON-VERBAL PREDICATES IN THE NOMVALLEX LEXICON

VERONIKA KOLÁŘOVÁ – VÁCLAVA KETTNEROVÁ
– JIŘÍ MÍROVSKÝ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic

KOLÁŘOVÁ, Veronika – KETTNEROVÁ, Václava – MÍROVSKÝ, Jiří: Through Derivational Relations to Valency of Non-verbal Predicates in the NomVallex Lexicon. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 182 – 192.

Abstract: NomVallex is a manually annotated valency lexicon of Czech nouns and adjectives that enables a comparison of valency properties of derivationally related lexical units. We present new developments in how the lexicon facilitates research into changes in valency across part-of-speech categories and derivational types. In particular, it provides links from derived lexical units to their base lexical units and also allows to search and display a base lexical unit together with all lexical units directly derived from it. Using an automatic procedure, any difference in valency between two derivationally related lexical units is specified. As a case study, focusing on nouns and adjectives directly or indirectly motivated by verbs, the facilities provided by the lexicon are used to show differences in what ways the particular deverbal derivatives representing various derivational types express the valency complementation standing in the base verbal construction in the subject position.

Keywords: adjectives, derivational relation, derivational type, nouns, valency behavior, valency lexicon

1 INTRODUCTION

Derivational relations in Czech are in focus of both theoretical studies (e.g. Ševčíková 2021) and projects aimed at automatic modeling of word-formation relations, for example DeriNet (Vidra et al. 2019).[1] Deverbal derivatives such as deverbal nouns and adjectives are often endowed with valency and they usually share some of the valency properties with their verbal base (the phenomenon referred to as the so-called argument inheritance, e.g. Booij 2007). In this paper, we deal with valency of nouns and adjectives directly or indirectly motivated by verbs (namely deverbal nouns, deverbal adjectives and nouns derived from deverbal adjectives); these deverbal derivatives represent those non-verbal predicates that typically denote actions, abstract results of actions or a quality and as such they are likely to be valent. In order to compare their valency behavior, we introduce the way the research can be facilitated by a lexicographic software.

[1] Accessible at https://ufal.mff.cuni.cz/derinet.

Valency properties of non-verbal predicates are covered in several Czech valency lexicons, first in a printed dictionary compiled by Svozilová, Prouzová and Jirsová (2005), second in two electronic valency lexicons, PDT-Vallex (Urešová et al. 2021) and NomVallex (Kolářová – Vernerová 2022). Out of these lexical resources, only NomVallex aims to systematically capture derivational relations between non-verbal predicates, making it possible to verify the hypothesis of argument inheritance (Sect. 2). First, we introduce the way NomVallex enables comparison of valency properties of derivationally related words, including both the annotation scheme (Sect. 3) and its visualisation (Sect. 4). Second, we present how the language material and facilities provided by NomVallex can be exploited in research into changes in valency across part-of-speech categories and derivational types (Sect. 5).

2 THE NOMVALLEX LEXICON

NomVallex is a manually created valency lexicon of Czech nouns and adjectives, adopting the theoretical framework of the Functional Generative Description (FGD) as its theoretical basis. Its newest version, NomVallex 2.0 (available in an electronic form, both as publicly available web-pages[2] and as downloadable and machine readable data; Kolářová – Vernerová – Klímová 2022), comprises 1,027 lexical units contained in 570 lexemes. As for derivational categories, it covers deverbal and deadjectival nouns, and deverbal, denominal, deadjectival or primary adjectives.

NomVallex adopts and further modifies, where necessary, the annotation scheme of the valency lexicon of Czech verbs VALLEX (Lopatková et al. 2022) (Fig. 1). The lexicon entry contains a lexeme, an abstract unit associating lexical forms with their lexical units (LUs), i.e., word senses. Each lexical unit is described by synonyms (the `synon` attribute) and assigned its id (e.g. blu-n-nadšenost-2 for the noun *enthusiasm* in Fig. 2). Aspectual counterparts formed by affixation, such as *vyzývání*^{impf} – *vyzvání*^{pf} ‘appealing’ or *ohrožovaný*^{impf} – *ohrožený*^{pf} ‘threatened’, are treated within a single lexeme; the aspectual properties are captured in the superscript of lemmas representing lexemes by the abbreviations *impf*, *pf* or *biasp*. Nouns or adjectives that do not express aspect are assigned the flag *no-aspect*, e.g. *výzva*$^{no\text{-}aspect}$ ‘appeal’.

The lexicon applies the valency theory of the FGD (Panevová 1980): valency properties of a lexical unit are captured in a valency frame, modeled as a sequence of valency slots, each supplemented with a list of morphemic forms. The following types of complementations may be a part of valency frames: obligatory or optional actants (i.e., ACTor, PATient, ADDRessee, EFFect, and ORIGin, e.g. *Petrova*ACT

[2] Accessible at https://ufal.mff.cuni.cz/nomvallex.

*výzva k pomoci*PAT ‘Peter’s appeal for help’, *prodejný mládeži*ADDR ‘marketable to the youth’, *odvolatelný z funkce*ORIG ‘dismissible from the post’), and obligatory free modifications, especially those with the meaning of direction (e.g. *muž povolaný do armády*DIR3 ‘a man drafted into the army’). In NomVallex, valency properties of a lexical unit are documented by examples from the Czech National Corpus (the `examplerich` attribute).[3]

In order to make it possible to study the relationship between valency behavior of base words and their derivatives, lexical units of nouns and adjectives in NomVallex are linked to their respective base lexical units (contained either in NomVallex itself or, in case of verbs, in the VALLEX lexicon, Sect. 3.1), linking together up to three parts-of-speech (i.e., noun–verb, adjective–verb, noun–adjective, and noun–adjective–verb). NomVallex aims to provide language material and lexicographic software allowing for linguistic research into various phenomena related to noun and adjectival valency, for example systemic (regular) and non-systemic (irregular) valency behavior (Sect. 3.2), including phenomena related to derivational type specificity (Sect. 3.3), and thus it employs facilities enabling to perform complex searches and comparisons (Sect. 4 and 5).

3 DERIVATIONAL RELATIONS IN NOMVALLEX

NomVallex describes derivational relations in several manually or automatically processed attributes (Sect. 3.1–3.3).[4]

3.1 Interlinking derivationally related lexical units

Derivationally related lexical units of nouns and adjectives are linked to each other (or to their base verbs in VALLEX) by means of two attributes, keeping both directions, namely:
(i) the attribute `derivedFrom` provides a link from a particular LU to its base LU;
(ii) the attribute `derivedLUs` captures links to all LUs derived from the base LU.

3.2 Automatic comparison of valency frames

Each lexical unit of an adjective or a noun (both deverbal and deadjectival) with a link to its respective base LU provided in the `derivedFrom` attribute is automatically supplemented with information on differences between valency frames of the two LUs; namely, the number and types of valency complementations and their morphemic forms are automatically compared. The changes (if any) are specified in the `valdiff` attribute.

[3] Accessible at https://www.korpus.cz/.

[4] In the following sections, we introduce recent developments in the data to be a part of the future published version; it concerns especially implementation of the attribute `derivedLUs` (Sect. 3.1), new labels for derivational types, including the numbers of lexical units representing them (Tab. 1 in Sect. 3.3), and the visualization part (Sect. 4).

3.3 Derivational types of nouns and adjectives

Each LU of an adjective or a noun is assigned a label indicating its derivational type (filled in the attribute `type`), see Tab. 1. The label provides the information on both part-of-speech membership of the LU (whether it is a noun (N) or an adjective (A)) and its derivational base (whether it is deverbal (DV), deadjectival (DA), denominal (DN) or primary (P)). Further, if the LU is directly or indirectly motivated by a verb, the label contains a number used to differentiate derivational history of the LU, reflecting especially the suffix by which the direct derivative was derived from the base verb. In the labels for deverbal nouns, number 1 is used for nouns ending in *-ní/-tí*, such as *vnímání* 'perceiving', number 2 indicates nouns derived from verbs by various suffixes, e.g. *-ka*, such as *námitka* 'objection', or by the zero suffix, such as *vjem* 'perception'. In the labels for deverbal adjectives, for instance number 5 marks adjectives derived from verbs by the suffix *-telný*, such as *vnímatelný* 'perceptible' (for more details see Kolářová – Vernerová – Klímová 2021).[5] Moreover, the labels for nouns are supplemented by the term most often used for their derivational type, such as *stem* or *root* deverbal nouns, or by the segment they typically end in (e.g. the nouns labeled as N-DA-3-lost are derived from deverbal adjectives of type 3 and they mostly end in *-lost*, e.g. *závislost* 'dependence').[6] Nouns derived from adjectives other than the deverbal ones get the label N-DA-O.

Part-of-speech category	Derivational category	Derivational type	Example	Lexical units	Lexemes
Nouns	deverbal	N-DV-1-stem	*vnímání* 'perceiving'	331	162
		N-DV-2-root	*vjem* 'perception'	185	91
	deadjectival	N-DA-1-cnost	*nemohoucnost* 'weakness'	3	174
		N-DA-3-lost	*závislost* 'dependence'	29	
		N-DA-4-1-nt-ost	*žádanost* 'demand' *použitost* 'state of usage'	31	
		N-DA-4-2-nt-ost	*nadšenost* 'enthusiasm' *dojatost* 'emotion'	11	
		N-DA-5-telnost	*vnímatelnost* 'perceptibility'	26	
		N-DA-6-vn-ost	*vnímavost* 'perceptiveness' *poslušnost* 'obedience'	70	
		N-DA-O	*žádostivost* 'desirousness', *nedůtklivost* 'touchiness', *hrdost* 'pride'	96	

[5] In NomVallex, all Czech deverbal derivatives with adjectival inflection are regarded as deverbal adjectives, no matter whether they denote an action (e.g. *porota rozhodující o cenách* 'a jury deciding the awards', *člověk přeživší havárii* 'a man surviving the crash'), a property (e.g. *rozhodující okamžik* 'the decisive moment') or an object (e.g. *můj známý* 'an acquaintance of mine', *přeživší havárie* 'a survivor of the crash').

[6] In Czech, no nouns are derived from adjectival types A-DV-2, A-DV-7 and A-DV-8, so no types N-DA-2, N-DA-7 and N-DA-8 are reflected in Tab. 1.

Adjectives	deverbal	A-DV-1	*nemohoucí* ‘not able’	11	133
		A-DV-2	*přeživší* ‘having survived’	5	
		A-DV-3	*závislý* ‘dependent’	30	
		A-DV-4-1	*žádaný* ‘desired’	47	
		A-DV-4-2	*nadšený* ‘enthusiastic’	11	
		A-DV-5	*vnímatelný* ‘perceptible’	31	
		A-DV-6	*vnímavý* ‘perceptive’	66	
		A-DV-7	*zasunovací – zasouvací* ‘sliding’	1	
		A-DV-8	*přeživší* ‘survivor’	5	
	denominal	A-DN	*žádostivý* ‘desirous’	28	14
	deadjectival	A-DA	*nedůtklivý* ‘touchy’	6	6
	primary	A-P	*hrdý* ‘proud’	62	28
Total				1,085	608

Tab. 1. Derivational types of nouns and adjectives in NomVallex

4 VISUALIZATION OF DERIVATIONAL RELATIONS

The NomVallex data can be searched at its web-pages (see footnote 2) or using the ‘vallex-like lexicons search tool’, called the Blue Search Engine (BlueSE),[7] which makes it possible to visualize all releases of the NomVallex data as well as their working version. Both the search tools allow for formulating complex queries based on a wide range of criteria, for example (a) derivational type of the noun or adjective (e.g. stem vs. root nouns), (b) its aspectual characteristics, (c) types of its valency complementations and their morphemic forms (including their distribution depending on the type of the word and/or the type of the complementation itself, individually and in combinations), and (d) the relation of the noun or the adjective to its base LU including the differences in valency behavior.

The BlueSE tool currently enables to visualize not only the list of individual LUs satisfying the criteria laid down in the query, but it also provides a facility that allows users to search and display a base LU together with all LUs directly derived from it, so that the research into the valency phenomena related to derivational type specificity is facilitated. The base LU can be represented by a verb (from VALLEX), an adjective or a noun. The LUs directly derived from the base LU are listed in the attribute `derivedLUs` (Sect. 3.1) and simplified entries of the particular derived LUs are then sketched out beside the base LU to enable the user to look over them and compare them. Two results of such a search are presented here:

(i) a verbal base LU, i.e., the verb *nadchnout se* ‘become enthusiastic’, and the LUs directly derived from it, i.e., the adjective *nadchnutý-nadšený-2* ‘enthusiastic’, and the noun *nadchnutí (se)* ‘becoming enthusiastic’ (Fig. 1);

[7] Accessible at https://quest.ms.mff.cuni.cz/vallex/.

(ii) an adjectival base LU, i.e, the adjective *nadchnutý-nadšený-2* ‘enthusiastic’, and the LU directly derived from it, i.e, the noun *nadšenost-2* ‘enthusiasm’ (Fig. 2).

As exemplified, LUs directly or even indirectly motivated by the verb *nadchnout se* ‘become enthusiastic’ are easily obtainable via BlueSE. The fact that all the derivatives are assigned their derivational type and that they have the information on changes in their valency provided in the `valdiff` attribute opens the possibility of the systematical study of the changes in their valency structure brought about in different types of derivational processes.

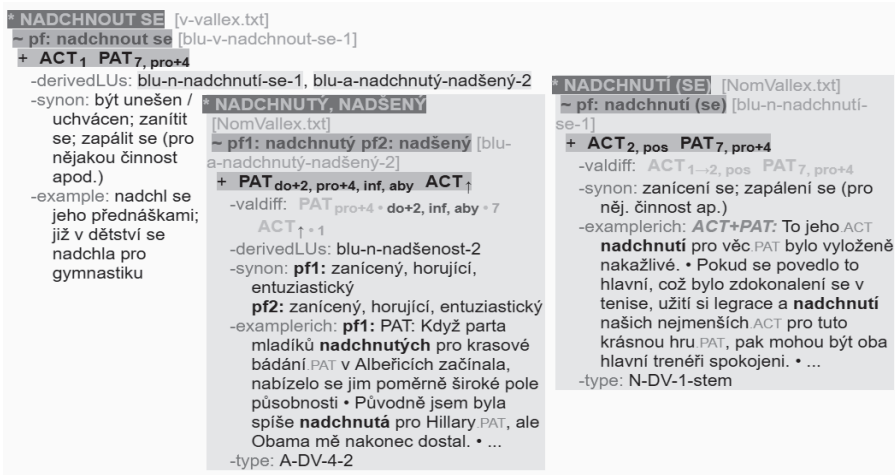

Fig. 1. Visualization of the LUs derived from the verb *nadchnout se* ‘become enthusiastic’

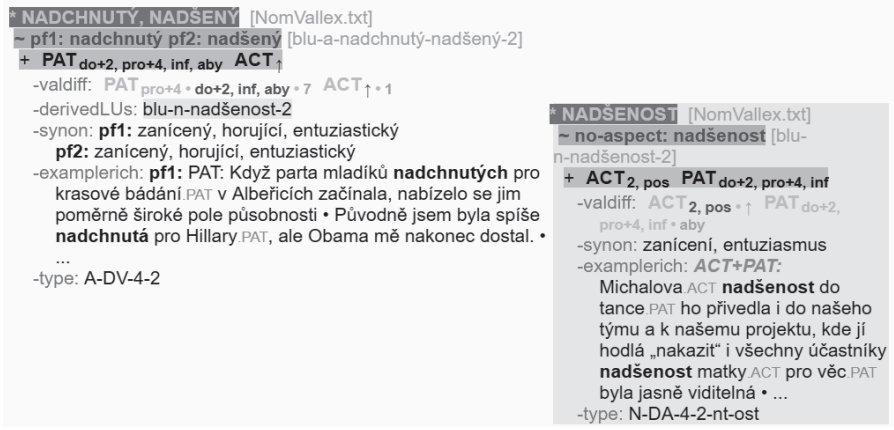

Fig. 2. Visualization of the LU derived from the adjective *nadchnutý-nadšený-2* ‘enthusiastic’

5 A CASE STUDY: THE VALENCY COMPLEMENTATION IN THE SUBJECT POSITION AND ITS CORRELATES ACROSS DERIVATIONAL TYPES

To illustrate the possibility of the systematical study of changes in valency structure across part-of-speech categories and derivational types, we focus on the valency complementation expressed in the base verbal structure in the subject position and on its correlates in noun and adjectival constructions, presenting changes in their structural configuration.

For example, the verb *vnímat* 'perceive', precisely its respective LU in example (1), forms a derivational base for several nouns and adjectives. First, its direct derivatives are represented by the deverbal nouns *vnímání* 'perceiving' and *vjem* 'perception', and by the deverbal adjectives *vnímatelný* 'perceptible' and *vnímavý* 'perceptive'.[8] Second, its indirect derivatives are exemplified by the deadjectival nouns *vnímatelnost* 'perceptibility' and *vnímavost* 'perceptiveness', directly derived from the adjectives *vnímatelný* 'perceptible' and *vnímavý* 'perceptive', respectively. The individual derivational relations are specified in Tab. 2, together with simplified valency frames of the derivatives.[9] The annotation maintains correspondences between individual valency complementations in valency frames across derivationally related lexical units, which allows for their comparison.

A close examination of the valency frames reveals differences in morphemic forms the particular derivational types use to express the valency complementation Actor standing in the base verbal construction of the verb *vnímat* 'perceive' in the subject position, see ACT_{Nom} in the valency frame of the verb in Tab. 2 and example (1).

(1) *muž*ACT-Nom *vnímá vysoký zvuk*PAT-Acc
'a man perceives a high sound'

(i) The Actor of the deverbal nouns *vnímání* 'perceiving' and *vjem* 'perception' can take on three forms, namely prepositionless instrumental (Ins), see (2), prepositionless genitive (Gen), see (3), or a possessive form (Poss), see (4).[10]

(2) *vnímání vysokého zvuku*PAT-Gen *mužem*ACT-Ins
'perceiving of a high sound by a man'

[8] According to Ševčíková (2021), the direction of motivation in pairs of Czech suffixless nouns (a part of the type N-DV-2-root) and verbs may be denominal in some cases (i.e., *vjem* 'perception' > *vnímat* 'perceive') rather than deverbal (i.e., *vnímat* 'perceive' > *vjem* 'perception').

[9] In the valency frames in Tab. 2, the abbreviation *cont* stands for dependent content clauses regardless of their complementatizers.

[10] The Actor of some other deverbal nouns, for example *žádost* 'request', can be expressed by the prepositional phrase *od* 'from'+Gen (see e.g. Kolářová – Vernerová – Verner 2019).

(3) *vjem muže*ACT-Gen, *že zvuk je*PAT-cont *vysoký*
‘perception of a man that the sound is high’
(4) *mužův*ACT-Poss *vjem vysokého zvuku*PAT-Gen
‘man’s perception of a high sound’

(ii) It is typical of adjectival valency structures, unlike the verbal and noun ones, that one valency complementation of the adjective is systematically elided from the surface and thus cannot be expressed on the surface as a modification of the adjective. Instead, it refers to its antecedent which is expressed outside the adjectival structure either as the noun governing the adjective, see (5), or as the subject of the copula verb the adjective forms a predicate with, see (6), cf. Kettnerová – Kolářová (2023).[11] In valency frames of adjectives, this valency complementation is marked by an upward arrow (see adjectival valency frames in Tab. 2). This sign is also used in (5–6) and (8) to pinpoint the antecedents of the systematically elided adjectival valency complementations.

(5) *muž*↑ *vnímavý k vysokému zvuku*PAT-k+Dat
‘a man perceptive about/of a high sound’
(6) *muž*↑ *je vnímavý k vysokému zvuku*PAT-k+Dat
‘a man is perceptive about/of a high sound’

Verb	Direct derivatives		Indirect derivatives		Valency frame
	Type	Lemma	Type	Lemma	
vnímat ‘perceive’					ACT_{Nom} $PAT_{Acc,cont}$
	N-DV-1 -stem	*vnímání* ‘perceiving’			$ACT_{Gen,Ins,Poss}$ $PAT_{Gen,Poss,cont}$
	N-DV-2 -root	*vjem* ‘perception’			$ACT_{Gen,Ins,Poss}$ $PAT_{Gen,Poss,cont}$
	A-DV-5	*vnímatelný* ‘perceptible’			$ACT_{Ins,pro+Acc}$ $PAT_{\uparrow}$
			N-DA-5 -telnost	*vnímatelnost* ‘perceptibility’	$ACT_{Ins,pro+Acc}$ $PAT_{Gen,Poss}$
	A-DV-6	*vnímavý* ‘perceptive’			$ACT_{\uparrow}$ PAT_{k+Dat}
			N-DA-6 -vn-ost	*vnímavost* ‘perceptiveness’	$ACT_{Gen,Poss}$ PAT_{k+Dat}

Tab. 2. Derivatives of the verb *vnímat* ‘perceive’

[11] In Kolářová – Vernerová (2022), this phenomenon was referred to as non-canonical realization of adjectival valency.

While the adjective *vnímavý* ‘perceptive’ systematically elides the Actor from the surface, see (5) and its valency frame provided in Tab. 2, reflecting hypothetically the active structure of its base verb in (1), the adjective *vnímatelný* ‘perceptible’ manifests a regular ellipsis of the Patient, see (8) and the valency frame in Tab. 2, mirroring rather the passive structure of the base verb, as illustrated in (7). The Actor of the adjective *vnímatelný* ‘perceptible’ is then either expressed by the form corresponding to the form of the Actor in verbal passive constructions (i.e., by Ins) or by the prepositional phrase *pro* ‘for’+Acc, see (8).[12]

(7) *vysoký zvuk*PAT-Nom *je/může být vnímán mužem*ACT-Ins
‘a high sound is/can be perceived by a man’
(8) *vysoký zvuk*↑ *vnímatelný mužem*ACT-Ins/*pro muže*ACT-pro+Acc
‘a high sound perceptible by a man/to a man’

(iii) In valency constructions of deadjectival nouns, valency complementations that are systematically elided in their base adjectival structures are “reactivated”, being expressed on the surface as an adnominal modification, typically in the form of Gen or Poss, see the Actor of the noun *vnímavost* ‘perceptiveness’ in (9–10) and its valency frame given in Tab. 2.[13] In contrast, the form of the Actor of the noun *vnímatelnost* ‘perceptibility’, derived from the adjective reflecting the passive structure of the base verbal construction, cf. (7–8), remains the same as in the adjectival construction, i.e., it is either Ins or sometimes a prepositional phrase, in this case *pro* ‘for’+Acc, see (11–12).

(9) *vnímavost muže*ACT-Gen *k vysokému zvuku*PAT-k+Dat
‘perceptiveness of a man to a high sound’
(10) *mužova*ACT-Poss *vnímavost k vysokému zvuku*PAT-k+Dat
‘man’s perceptiveness of/to a high sound’
(11) *vnímatelnost zvuku*PAT-Gen *mužem*ACT-Ins/*pro muže*ACT-pro+Acc
‘perceptibility of a sound by a man/to a man’
(12) *jeho*PAT-Poss *vnímatelnost mužem*ACT-Ins/*pro muže*ACT-pro+Acc
‘its perceptibility by a man/to a man’

Our examination of valency frames of direct and indirect derivatives of the verb *vnímat* ‘perceive’ has revealed a wide range of ways to express the subject of the

[12] The prepositional phrase *pro* ‘for’+Acc is not evidenced in the valency frame of the verb *vnímat* ‘perceive’, but it is attested in valency frames of adjectives of the given type, e.g. *ta instrukce není pro mě pochopitelná* ‘the instruction is not understandable to/for me’.

[13] It should be stressed that the “reactivation” concerns all the valency complementations that are subject to the systemic ellipsis regardless of their type, see examples (11–12) illustrating constructions of the noun *vnímatelnost* ‘perceptibility’ in which PAT is reactivated.

base verb, showing not only differences in morphemic forms the particular noun derivational types use to express it but also the specific ways it is expressed in adjectival constructions. We assume that the valency constructions of the verb *vnímat* 'perceive' and its derivatives illustrate typical changes in valency structures of derivationally related deverbal derivatives and we suggest that the changes are connected with particular derivational types the derivatives represent.

6 CONCLUSION

We have introduced the way how the NomVallex lexicon facilitates the research into changes in valency across part-of-speech categories and particular derivational types. Linking derivationally related words opens the possibility of the systematical study of (ir)regularity in their valency behavior. We have specified means for capturing derivational relations among verbs, nouns and adjectives, and described possible visualization of them. As a case study, we have examined valency frames of nouns and adjectives directly and indirectly motivated by the verb *vnímat* 'perceive', concentrating on the valency complementation expressed in the base verbal structure in the subject position and on its correlates in noun and adjectival constructions. We suggest that the differences in their structural configuration depend on respective derivational types represented by the deverbal derivatives.

ACKNOWLEDGEMENTS

The research reported in the paper was supported by the Czech Science Foundation under the project 22-20927S. The work described herein has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Booij, G. (2007). The Grammar of Words. Oxford: Oxford University Press, 345 p.

Kettnerová, V., and Kolářová, V. (2023). K reciprocitě adjektiv v češtině. Slovo a slovesnost, 84(3), pages 179–200.

Kolářová, V., and Vernerová, A. (2022). NomVallex: A Valency Lexicon of Czech Nouns and Adjectives. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1344–1352, European Language Resources Association, Marseille, France.

Kolářová, V., Vernerová, A., and Klímová, J. (2021). Systemic and non-systemic valency behavior of Czech deverbal adjectives. Jazykovedný časopis, 72(2), pages 371–382.

Kolářová, V., Vernerová, A., and Klímová, J. (2022). NomVallex 2.0., LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL),

Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-4663.

Kolářová, V., Vernerová, A., and Verner, J. (2019). Non-systemic valency behavior of Czech deverbal nouns based on the NomVallex lexicon. Jazykovedný časopis, 70(2), pages 424–433.

Lopatková, M., Kettnerová, V., Mírovský, J., Vernerová, A., Bejček, E., and Žabokrtský, Z. (2022). VALLEX 4.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-4756.

Panevová, J. (1980). Formy a funkce ve stavbě české věty. Praha: Academia, 222 p.

Svozilová, N., Prouzová, H., and Jirsová, A. (2005). Slovník slovesných, substantivních a adjektivních vazeb a spojení. Praha: Academia, 579 p.

Ševčíková, M. (2021). Action nouns vs. nouns as bases for denominal verbs in Czech: A case study on directionality in derivation. Word Structure, 14(1), pages 97–128.

Urešová, Z. et al. (2021). PDT-Vallex: Czech Valency lexicon linked to treebanks 4.0 (PDT-Vallex 4.0), LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-3499.

Vidra, J. et al. (2019). DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 81–89, Prague.

sciendo
DOI 10.2478/jazcas-2023-0037

LINEAR DEPENDENCY SEGMENTS IN FOREIGN LANGUAGE ACQUISITION: SYNTACTIC COMPLEXITY ANALYSIS IN CZECH LEARNERS' TEXTS

MICHAELA NOGOLOVÁ – MICHAELA HANUŠKOVÁ
– MIROSLAV KUBÁT – RADEK ČECH
Department of Czech Language, Faculty of Arts, University of Ostrava,
Ostrava, Czech Republic

NOGOLOVÁ, Michaela – HANUŠKOVÁ, Michaela – KUBÁT, Miroslav – ČECH, Radek: Linear Dependency Segments in Foreign Language Acquisition: Syntactic Complexity Analysis in Czech Learners' Texts. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 193 – 203.

Abstract: The paper discusses a new way to measure syntactic complexity in foreign language acquisition. It is based on a recently proposed syntactic unit called linear dependency segment (LDS), the longest possible sequence of words belonging to the same clause where all linear neighbours are also syntactic neighbours. The dataset comprises 5,721 Czech texts from the CzeSL-SGT learner corpus covering five CEFR proficiency levels (A1–C1). The study covers two analyses. First, the development of the average clause length in terms of LDS and the average LDS length in the number of words across the latter language proficiency levels. Second, we consider the differences between Slavic and non-Slavic speakers. The results show an increasing tendency of the average clause length measured in LDS while the average clause length measured in words is decreasing. Results also show statistically significant differences between Slavic and non-Slavic speakers in most cases. Our results indicate that using LDS may be a useful unit of syntactic complexity measure in foreign language acquisition research.

Keywords: foreign language acquisition, dependency grammar, linear dependency segment, syntactic complexity, Czech language

1 INTRODUCTION

Syntactic complexity has long been an interest in writing in the second language acquisition domain. Over the years, the complexity of syntactic structures has become a valuable indicator of language development, both in first language acquisition and any other foreign language (FL) acquisition (see Crossley – McNamara 2014; Yang et al. 2015). However, in the last decades, the traditional syntactic complexity measures (such as average length of clause or sentence, subordinate clause per clause, or T-unit (Hunt 1965) per sentence) have been faced with critique for their lack of linguistic background, problematic use for all language proficiency levels, and vague definition of syntactic complexity itself (see e.g. Biber et al. 2020; Kuiken 2022; Ouyang et al. 2022). Recent research has focused on finding alternative ways to measure syntactic

structures, particularly those considering the dependency structure of clauses or sentences, e.g. mean dependency distances (MDD; e.g. Jiang – Ouyang 2018; Ouyang et al. 2022) and linear dependency segment (LDS; Mačutek et al. 2021). The shift towards measurements based on dependency grammar also reflects a shift towards a deeper connection between linguistics and cognitive sciences.

This research focuses on evaluating FL writing with a focus on LDS. We explore the development of average LDS length and average clause length in LDS. The language material comes from the Czech learner corpus CzeSL-SGT, a part of the Czech National Corpus (Šebesta et al. 2014). It consists of 5,721 texts on A1–C1 language proficiency levels according to The Common European Framework of Reference for Languages (CEFR).

2 LANGUAGE MATERIAL AND METHODOLOGY

2.1 Language material

The language material used in the current study comes from the Czech National Corpus. It is a collection of selected texts from the CzeSL-SGT learner corpus (Šebesta et al. 2014). The corpus comprises 8,617 texts authored by 1,965 non-native Czech speakers of all language proficiency levels defined by CEFR. The corpus contains metadata on both authors and texts. In our research, we utilise data on the learner's language proficiency level, their first language (L1), and the length of the text. To ensure the accuracy of our analysis, we excluded texts with unclear or unknown proficiency levels, as well as texts assigned to the C2 level, because only one text is tagged to this category. Furthermore, we removed texts shorter than 55 words because the standard-length requirement for passing a written exam is typically around 50–60 words. We used the L1 information to categorise learners into Slavic and non-Slavic groups. In summary, our corpus consists of 5,721 texts that cover five CEFR proficiency levels. Additionally, we compare the results with the reference corpus (REF-CZ), consisting of texts written by Czech native speakers. The data come from the SKRIPT2012 corpus (Šebesta et al. 2013). Specifically, we use texts written by fourth-grade high school students because of their comparability with the CzeSL-SGT corpus regarding authorship. For sample details, see Tab. 1.

level	number of texts	number of texts Slavic L1	number of texts non-Slavic L1
A1	1,854	1,364	490
A2	1,738	1,157	581
B1	1,313	833	480
B2	702	497	205
C1	114	78	36
REF-CZ	87	-	-

Tab. 1. Number of texts in each group of the analysed sample

2.2 Linear dependency segments (LDS) and data processing

Mačutek et al. (2021) defined the linear dependency segment as follows: "[...] the longest possible sequence of words (belonging to the same clause) in which all linear neighbours (i.e., words adjacent in a sentence) are also syntactic neighbours [...]". In detail, every clause is created by a predicate and other directly or indirectly dependent words. The only exception occurs when the dependent word is another predicate. In that case, the syntactic relationship between word and other predicate presents a boundary between two clauses. LDS then is created only within a clause. For illustration, clauses and LDS determination in sentence (A) are presented in Fig. 1. The circles indicate particular clauses, and the squares then individual LDS.

(A) *Petr má psa, který hodně kouše.*
'Petr has a dog that bites a lot.'

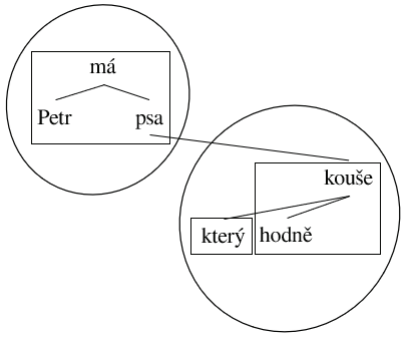

Fig. 1. Visualization of clauses and LDS determination in a sentence (A)

The sentence (A) has two clauses because two predicates – *má* ('has') and *kouše* ('bites') – are present. One LDS creates the first clause, and the second clause makes two LDS. The first word of the sentence (*Petr*) is directly dependent on the word *má.* It can be also seen that the second word (*má*) is directly connected to the word *psa*. These three words are neighbours in linear clause ordering, so creating one LDS. The third word (*psa*) is also related to the word *kouše*. However, these two words are not adjacent in the sentence, and they are not in the same clause. Therefore, they cannot be in the same LDS. The fourth word (*který*) is directly connected to the word *kouše.* These two words are not next to each other in the clause word order, so the word *který* creates one LDS. The last LDS of the second clause is created by the words *hodně* and *kouše* because they are directly connected as adjacent in the clause. As such, LDS captures a clause's linear sentence ordering and dependency structure.

The current study aims to analyse syntactic structures development of FL writing focusing on LDS. We use two indices:

(i) Average clause length measured in the number of LDS (ACL),
(ii) Average LDS length measured in the number of words (ALDSL).

As the LDS reflects both the dependency structure and word order in clause, the value of ACL and ALDSL can be in accordance with the clause complexity. For illustration, Fig. 2 and 3 show the dependency relationships of sentences (B) and (C). The squares represent individual LDS.

(B) *Petr má psa.*
'Petr has a dog.'

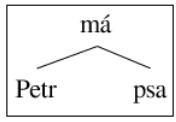

Fig. 2. Visualization of dependency relationships and LDS in a sentence (B)

(C) *Můj dobrý kamarád Petr má doma velkého psa.*
'My good friend Petr has a big dog at home.'

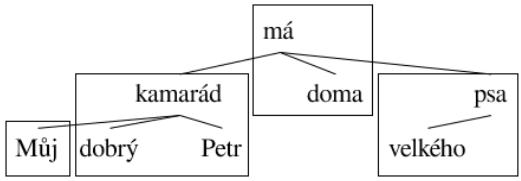

Fig. 3. Visualization of dependency relationships and LDS in a sentence (C)

Both sentences (B) and (C) contain only one clause. However, when one compares these two sentences, more complex syntactic structures are present in the longer one. The distance between two dependent words is also bigger. As the words that create an LDS must be directly connected as adjacent in the sentence word order, more complex syntactic structures will lead to more LDS within the clause and a shorter average length of LDS. The sentence (B) has an average clause length of 1 (1 LDS / 1 clause = 1), with an average LDS length of 3 (3 words / 1 LDS = 3). In contrast, sentence (C) has an average clause length of 4 (4 LDS / 1 clause = 4) and an average LDS length of 2 (1+3+2+2 = 8 words; 8 words / 4 LDS = 2). Therefore, we assume that development

towards a higher level of language proficiency will correspond with an increase in the number of LDS in a clause and a decrease in LDS length.

There are various approaches to processing the syntactic structure of a sentence, and these approaches influence the created annotation scheme used for sentence processing. This paper uses Surface Syntactic Universal Dependencies (SUD; Gerdes et al. 2018) for annotation. SUD is built upon the Universal Dependencies (UD; Zeman et al. 2022) which aims to provide a versatile annotation system for a wide range of languages. While UD exhibits a greater focus on semantic aspects, SUD, in contrast, adopts a more syntactically oriented approach, with auxiliaries and prepositions holding a superior position rather than being subordinate to content words, as seen in the case of UD. In conducting our analysis, we initially utilised UDPipe 2.0. (Straka 2018) to process all texts. Then, we used Grew software (Guillaume 2021) for UD to SUD conversion. The selection of SUD was motivated by our objective to conduct a syntactic analysis, making its more syntactically-oriented perspective highly suitable for our study.

The analyses were performed by the following steps. First, all texts were processed separately for each proficiency level. Second, the mean of clauses, LDS length, and standard deviations (sd) were calculated. Third, we used the Mann-Whitney test (Mann – Whitney 1947) with a significance level of $\alpha = 0.05$ to test statistical differences between pairs of proficiency levels. This statistical test was chosen due to the non-normal distribution of the data. We also performed the same test to compare groups of Slavic and non-Slavic learners at each proficiency level.

3 RESULTS

3.1 Development of ACL and ALDSL across the language proficiency levels

The results presented in Tab. 2 show the values of ACL and ALDSL for examined language proficiency levels and the values obtained from the reference corpus. We can see the increased tendency across all language proficiency levels towards the value obtained from native speakers when first focused on ACL. Fig. 4 gives a more detailed description of the obtained values. These results support our hypothesis that the ACL increases as the language proficiency level increases. ALDSL values in Tab. 2 and Fig. 5 show a descending tendency towards the value obtained from native speakers. These results also support our hypothesis that the ALDSL value decreases with increasing language proficiency level.

Additionally, the gap in syntactic abilities measured by ACL and ALDSL between learners and native speakers diminishes as language proficiency increases. Standard deviation (sd) values indicate consistent variability of the results across different proficiency levels in both ACL and ALDSL analysis.

	ACL		ALDSL	
level	mean	sd	mean	sd
A1	2.522	0.556	2.145	0.201
A2	2.572	0.522	2.118	0.192
B1	2.674	0.553	2.096	0.178
B2	2.788	0.556	2.085	0.181
C1	2.997	0.633	2.049	0.176
REF-CZ	3.216	0.573	1.935	0.097

Tab. 2. The mean values of ACL, ALDSL and their sd

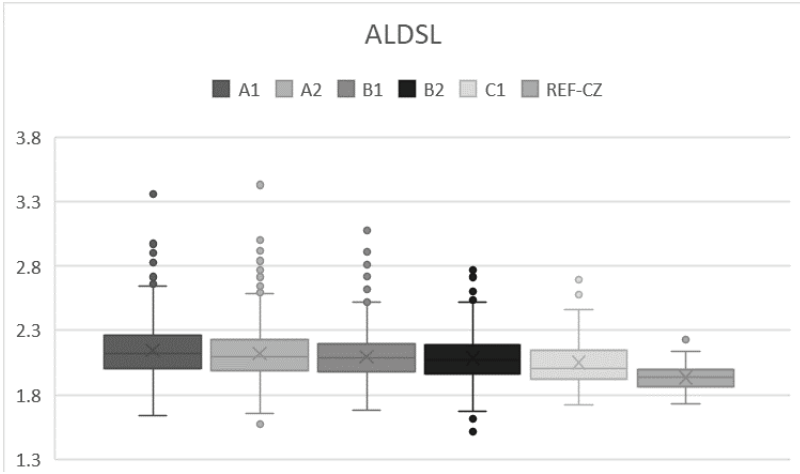

Fig. 4. ACL values from texts at A1–REF-CZ

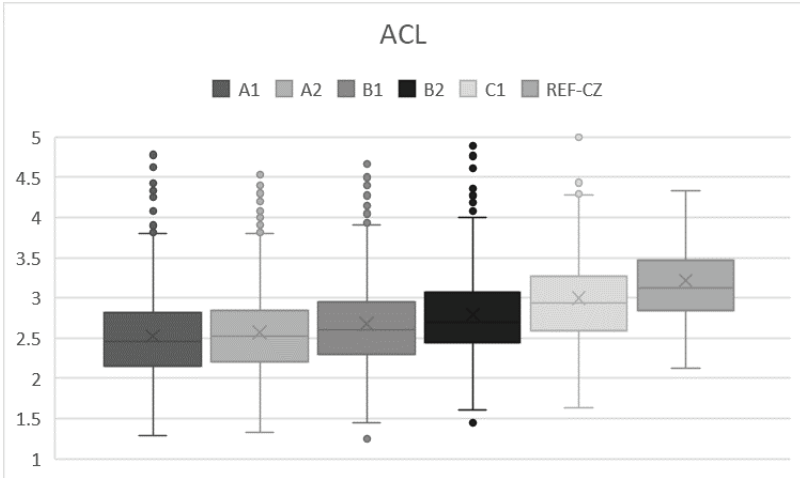

Fig. 5. ALDSL values from texts at A1–REF-CZ

The differences between pairs of levels were statistically tested. In the case of ACL, there are significant differences between all levels. These findings suggest that average clause length is relevant in FL syntactic development (for results, see Tab. 3).

ACL	A1	A2	B1	B2	C1
A2	**0.001**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	**<0.001**		
C1	**<0.001**	**<0.001**	**<0.001**	**0.007**	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.005**

Tab. 3. Statistical tests' results between each pair of the A1–REF-CZ values (ACL)

Concerning the average length of LDS, the statistically significant differences are detected between all pairs of levels except between B1 and B2 (for more details, see Tab. 4). Further research is needed to determine whether similar results from texts at B1 and B2 level represent random fluctuations or hint at some trend.

ALDSL	A1	A2	B1	B2	C1
A2	**<0.001**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	0.128		
C1	**<0.001**	**<0.001**	**0.030**	**0.019**	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.007**

Tab. 4. Statistical tests' results between each pair of the A1–REF-CZ values (ALDSL)

3.2 Differences between Slavic and non-Slavic groups

Given that Czech is a Slavic language, with many similarities in language syntactic structures between all languages in the group, it is reasonable to assume that learners with Slavic L1 will have higher values of ACL and lower ones in the case of ALDSL. To test this hypothesis, texts at each proficiency level were divided into Slavic and non-Slavic groups, and their respective values were compared. As can be seen in Tab. 5 and Fig. 6, the data confirmed the assumption, mainly regarding ACL development. Up to the C1 level, statistically significant differences were found between these two groups, indicating that Slavic learners possess a clear advantage due to their L1 background.

	ACL_S		ACL_N		statistical tests
level	mean	sd	mean	sd	p-value
A1	2.573	0.549	2.384	0.552	**<0.001**
A2	2.632	0.519	2.456	0.508	**<0.001**
B1	2.744	0.580	2.556	0.482	**<0.001**
B2	2.851	0.553	2.637	0.536	**<0.001**
C1	3.035	0.591	2.914	0.718	0.345

Tab. 5. The ACL means and sd for Slavic and non-Slavic groups and results of statistical tests

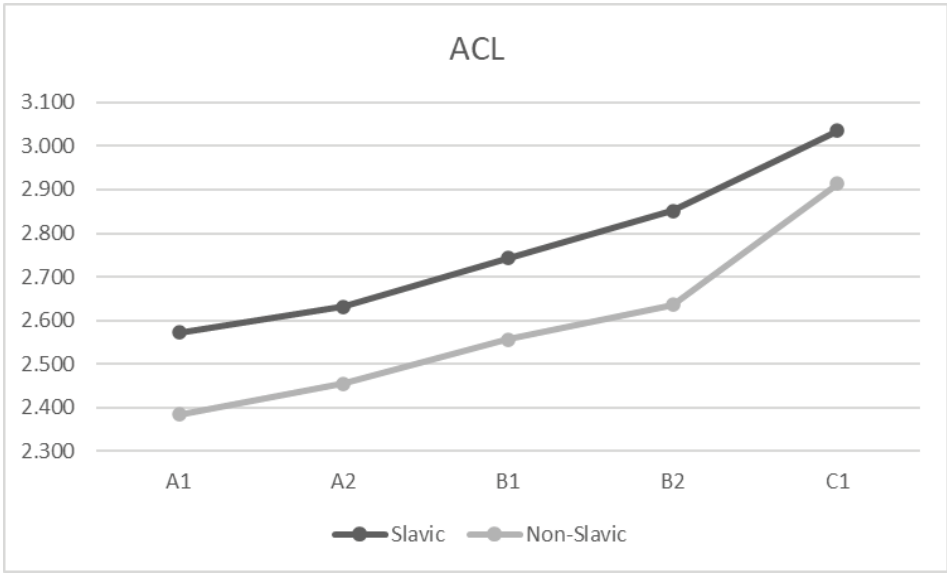

Fig. 6. The ACL means for Slavic and non-Slavic groups at all language proficiency levels

Regarding ALDSL, results in Tab. 6 and Fig. 7 demonstrate that the mean values for the Slavic groups are generally lower than those for non-Slavic counterparts across all proficiency levels. However, statistically significant differences between the two groups are observed only at levels A1 and C1.

	ALDSL_S		ALDSL_N		statistical tests
level	mean	sd	mean	sd	p-value
A1	2.131	0.184	2.182	0.237	**<0.001**
A2	2.111	0.185	2.133	0.204	0.094
B1	2.092	0.171	2.102	0.189	0.870
B2	2.081	0.178	2.092	0.187	0.425
C1	2.023	0.167	2.107	0.185	**0.013**

Tab. 6. The ALDSL means and sd for Slavic and non-Slavic groups and results of statistical tests

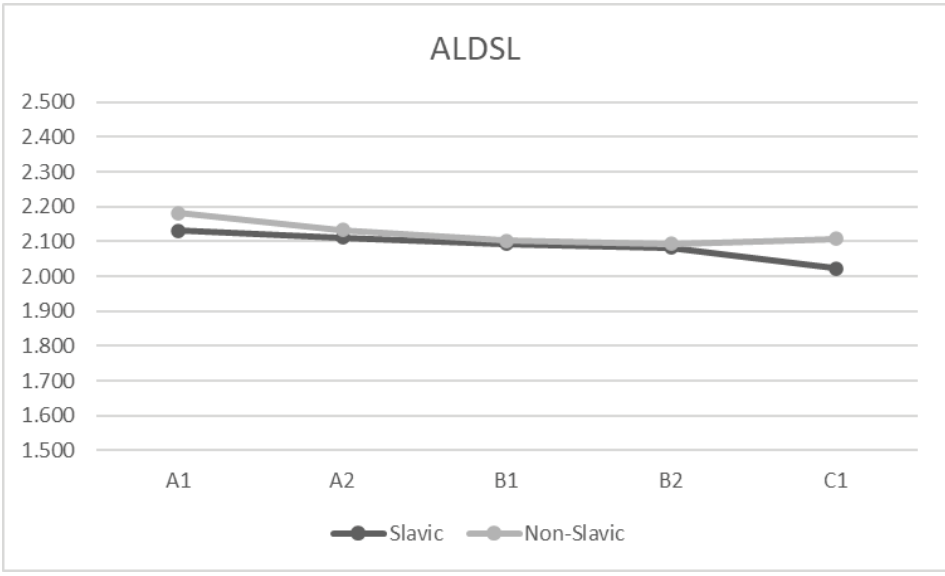

Fig. 7. The ALDSL means for Slavic and non-Slavic groups at all language proficiency levels

4 CONCLUSION

This study focused on linear dependency segments (LDS) in Czech texts written by non-native speakers. We aimed to explore the potential of LDS for measuring syntactic complexity in foreign language acquisition research. The results showed that the linear dependency segments could be useful for measuring syntactic complexity.

The analysis of average clause length based on the number of LDS revealed an increasing tendency towards native speakers across all levels of language proficiency. Furthermore, the data also confirmed our expectation that Slavic L1 speakers have longer clauses on average than their non-Slavic counterparts.

According to the average length of LDS measured in words, the higher the proficiency level, the shorter the LDS. LDS lengths for texts written by native Slavic speakers and native non-Slavic speakers are generally similar, except for the A1 and C1 levels. The differences between these two groups of learners are not as apparent as we had expected.

Since this is a pioneer study examining linear dependency segments in foreign language acquisition, further research must be done to confirm our results. It is important to perform research in other languages and different contexts, such as spoken language or different writing genres. Additionally, further exploration of the relationship between linear dependency segments and other measures of syntactic complexity could provide a deeper understanding of language acquisition.

ACKNOWLEDGEMENTS

The research is supported by Grant SGS08/FF/2023, University of Ostrava.

References

Biber, D., Gray, D., Staples, S., and Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predicative measurement. Journal of English for Academic Purposes, 46.

Crossley, A., S., and McNamara, S., D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. Journal of Second Language Writing, 26, pages 66–79.

Gerdes, K., Guillaume, B., Kahane, S., Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Universal Dependencies Workshop.

Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In Demonstrations – 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3. Champaign, IL, USA: NCTE.

Jiang, J., and Ouyang, J. (2018). Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition. In J. Jingyang – H. Liu (eds.): Quantitative Analysis of Dependency Structures. De Gruyter Mouton, pages 167–190.

Kuiken, F. (2022). Linguistic complexity in second language acquisition. Linguistic Vanguard.

Mačutek, J., Čech, R., and Courtin, M. (2021). The Menzerath-Altmann law in syntactic structure revisited. In Quasy, SyntaxFest 2021: Proceedings of the Second Workshop on Quantitative Syntax (March 21 – 25, 2022). Sofia: Association for Computational Linguistics, pages 65–73.

Mann, H. B., and Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, pages 50–60.

Ouyang, J., Jiang, J., and Liu, H. (2022). Dependency distance measures in assessing L2 writing proficiency. Assessing Writing, 51.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (October 31 – November 1, 2018). Brussels: Association for Computational Linguistics, pages 197–207.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M., and Rosen, A. (2014). CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014. Praha. Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Šebesta, K., Goláňová, H., Jelínek, T., Jelínková, B., Křen, M., Letafková, J., Procházka, P., and Skoumalová, H. (2013). SKRIPT2012: akviziční korpus psané češtiny – přepisy písemných prací žáků základních a středních škol v ČR. Praha, Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Yang, W., Lu, X., and Weigle, C., S. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. Journal of Second Language Writing, 28, pages 53–67.

Zeman, D., et al. (2022). Universal Dependencies 2.10, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Matematicko-fyzikální fakulta, Univerzita Karlova. Accessible at: http://hdl.handle.net/11234/1-4758.

sciendo
DOI 10.2478/jazcas-2023-0038

DIFFERENCES IN SPOKEN LANGUAGE PROCESSING IN GENERAL CORPORA (ORAL, ORTOFON) AND IN A SPECIALIZED CORPUS (DIALEKT) AND THEIR REFLECTION IN THE MAPKA APPLICATION

MARTINA WACLAWIČOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University,
Prague, Czech Republic

WACLAWIČOVÁ, Martina: Differences in Spoken Language Processing in General Corpora (ORAL, ORTOFON) and in a Specialized Corpus (DIALEKT) and Their Reflection in the Mapka Application. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 204 – 213.

Abstract: ORAL and ORTOFON, general corpora of the spoken Czech language, capture authentic and prototypical informal spoken language. DIALEKT, a specialized corpus, represents traditional regional dialects of the Czech language. Since the corpora's goals and the nature of the captured language data differ, different data collection methods were required. It concerns not only the choice of speakers, but the whole communication situation. Samples chosen from these three corpora are included in the Mapka application and reflect the distinct character of the corpora. The ORAL and ORTOFON samples show general spoken language in various informal situations and capture a wide range of speakers. The DIALEKT samples represent traditional regional dialects spoken by chosen types of speakers in a semiformal situation of guided interview.

Keywords: corpus design, Czech language, spoken language, dialect, map application

1 SPOKEN DATA IN THE CZECH NATIONAL CORPUS

The Czech national corpus covers a wide range of heterogeneous language data, both written and spoken. This text will concentrate on spoken language data and its processing. There are two main groups of spoken corpora in the Czech national corpus – general and specialized. Both are considered synchronic, however, they cover a time range from the middle of the 20th century to the present. General corpora on the one hand and specialized corpora on the other hand require different methods of data collection, since there are differences among a number of circumstances, in the character of language data and the purposes of the corpora.

The oldest general copora are represented by three basic corpora of the ORAL series: ORAL2006, ORAL2008 and ORAL2013 or by the ORAL corpus (also known as ORALv1), that unifies the corpora ORAL2006, ORAL2008, ORAL2013 and the unpublished data ORAL-Z. The two newest general spoken corpora collect informal spoken data (ORTOFON), as well as formal spoken data (ORATOR). Among the specialized corpora there are two corpora covering spoken language in

the cities – PMK (Prague spoken corpus) and BMK (Brno spoken corpus), two learner corpora: SCHOLA2010 (a school lessons corpus) and LINDSEI_CZ (a learner corpus of spontaneous spoken English) and two corpora od speeches: ParlCorp (a corpus of speeches delivered in the Lower Chamber of the Czech Parliament) and SPEECHES (a corpus of official presidential speeches). The newest addition to the group of specialized corpora is the DIALECT corpus (a dialectal corpus). This text will focus on three corpora – ORAL, ORTOFON and DIALECT and also the newest version of the Mapka application including samples chosen from these corpora.

2 GENERAL SPOKEN CORPORA: ORAL AND ORTOFON

2.1 ORAL – corpus description and composition

The ORAL corpus (Kopřivová – Lukeš – Komrsková – Poukarová – Waclawičová – Benešová – Křen 2017) was created by merging data from the already published corpora ORAL2006, ORAL2008 and ORAL2013 and by adding the previously unpublished ORAL-Z section. The overall size of the corpus is 5,368,391 words, and the recordings have a duration of 582 hours. Part of the transcripts (the data from the corpora ORAL2013 and ORAL-Z) is linked to the audio. The main goal of the ORAL corpus was to capture authentic and prototypical informal spoken language, and therefore it consists of the transcribed recordings of predominantly informal conversations. They take place between native speakers of Czech originating from all regions of the Czech Republic. The corpus is lemmatized and morphologically tagged, whereas the same type of morphological tagging was used for the contemporary written corpora.

There is only one level of transcription based on folkloristic methods. Its aim is to capture the relevant features of the spoken language and, simultaneously, to be close to the common spelling due to lemmatization and morphological tagging.

2.2 ORAL – collecting data

The recordings for the ORAL corpus were made between the years 2002 and 2011. The recording environment and speakers were chosen with respect to the focus on authentic informal language. The recordings and their transcriptions were made mostly by students of linguistics from various Czech universities. Recordings were made predominantly in a home environment, partly in restaurants, cafes or outdoors etc. All the speakers knew each other very well, mostly they were family members or friends. The speakers were informed about the purpose of recording after the recording had been made, in order that they behaved naturally and their speech was maximally informal and unaffected by the fact of recording. The only person who was always aware of the fact that a recording had been made was the person who made the recording. Nevertheless even these speakers act naturally.

The ORAL-Z section additionally contains several recordings of semi-formal and formal situations. The semi-formal recordings capture communication between speakers, while one of them might represent an institution, e.g. job interview, conversation at the office, in the shop, in the doctor's office, etc. These speeches are partly prepared and contain typical formulas or patterns. Formal recordings cover prepared speeches, e.g. lectures or orations.

2.3 ORTOFON – corpus description and composition

The ORTOFON corpus (Kopřivová – Komrsková – Lukeš – Poukarová – Škarpová 2017) presents spontaneous spoken language used in informal situations between speakers who know each other, as well as the corpora of the ORAL series. The ORTOFON corpus is composed of 332 recordings of 624 different speakers from 2012–2017, and the transcriptions contain 1,014,786 orthographic words. They were carried out on two tiers: the orthographic tier and the simplified phonetic tier (Kopřivová – Goláňová – Klimešová – Komrsková – Lukeš 2014). Together with the DIALEKT corpus, it is one of the first two spoken corpora of the Czech language with a multi-tier transcription. The transcriptions are linked to the corresponding audio tracks. The corpus is lemmatized and morphologically tagged according to the ORAL corpus.

ORTOFON was in its first version designed to be fully balanced regarding all the basic sociolinguistic speaker categories, i.e., gender, age group, level of education and region of childhood residence. The second version of the corpus is not balanced in any way, but, nevertheless, its value lies in the twice enlarged amount of collected material.

2.4 ORTOFON – collecting data

The language data processed in the Ortofon corpus were collected in accordance with the method used in the preparation of the corpora from the ORAL series. The recordings and their transcriptions were made by students of and also by a number of external collaborators of the Institute of the Czech National Corpus. The speakers were chosen in order to achieve the maximum possible variability with regard to the four basic sociolinguistic categories: gender, age, level of education and the dialectal region in which the speaker spent the majority of the first 15 years of life. As well as in the ORAL series, the speakers knew each other well, they were mostly friends or family members, and recordings were taken in a private, especially home environment, therefore their speech was natural and informal.

3 SPECIALIZED CORPUS OF THE CZECH DIALECTS: DIALEKT

3.1 Corpus description and composition

The DIALEKT corpus (Goláňová – Waclawičová – Lukeš 2021) is a specialized corpus of the traditional regional dialects of the Czech language. The sound

recordings were made in the period from the end of the 1950s until the present in all dialectal regions of the Czech Republic, and, additionally, several recordings were taken in the Czech language islands in Poland. The second version of the corpus contains more than 220,000 words and almost 28 hours of recordings of 291 speakers (Goláňová – Waclawičová 2019).

The transcription of the recordings was carried out on two tiers: the orthographic tier and the dialectological tier, which is based on the Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak (Dialektologická komise České akademie věd a umění 1951). The orthographic tier is as much as possible compatible with the corresponding orthographic tier used in the ORTOFON corpus. As well as in the ORTOFON corpus, the transcriptions are linked to the audio tracks, and the corpus is lemmatized and morphologically tagged.

3.2 Collecting data

The language material contained in the DIALEKT corpus originates from multiple sources, mainly dialectological or folklore research or student diploma thesis. The corpus is composed of two main time strata defined by the time of recording acquisition. The older stratum contains probes that originate from the period from the end of the 1950s until the 1980s. It is partly composed of language material acquired by the Department of Dialectology of the Institute of the Czech Language of the Academy of Sciences of the Czech Republic, v. v. i., published in the appendix to the Czech language atlas (Balhar et al. 2011). Beside this, recordings made by individuals, mostly dialectologists, are included. Most of the data has already been published elsewhere. The newer stratum of the corpus covers the period from the 1990s until the present. It is composed of the language data provided by dialectal researchers or students of various Czech university faculties. A considerable part of the dialect probes was also made by the Institute of the Czech National Corpus.

The method of data collection is based on the principles traditionally used in Czech dialectology. The main goal of the corpus is to capture the oldest state of traditional territorial dialects and as many archaic dialectal elements as possible. For this reason, the speakers of the recordings were exclusively chosen from the members of the oldest generation, specifically at the age over 60. They lived their whole life predominantly in the same location in rural areas and only rarely relocated. And in addition to this, their ancestors had been living in the same location for generations. The speakers mainly worked in the agricultural sector or practiced a craft.

The speech of dialectal speakers has a semiformal character. According to the method used in dialectology, the explorators (i.e., interviewers) made the recordings with the informants (i.e., dialect speakers) in the form of so-called guided interviews. The explorators asked prepared questions, giving the informants space to lead

a monologue on a selected topic. The topics were usually related to the traditional rural life and the world in the time of the informant's youth and were therefore connected to agriculture, crafts, folklore, local customs and traditions – Christmas and Easter, regional cuisine, local tales, historical events such as the Second World War, etc. These topics led the speaker to use older, especially rural lexical dialectisms, and hand in hand with that dialectisms of the other language levels also occurred. Despite the partly formal character of the method of guided interviews, the recordings were made in a private domestic environment, in a friendly and informal atmosphere.

4 MAPKA APPLICATION

The Mapka application (Goláňová – Waclawičová – Pejcha 2020) is an interactive map-based application combining many functions and intending to serve both linguists and the general public. The primary aim was to create a supplement for the spoken corpora of the Czech National Corpus that would integrate data from these corpora with a map-based interface. In the first two versions it served as a supplement for the DIALEKT corpus (Goláňová – Waclawičová 2021). Data from ORAL and ORTOFON were added recently this year.

The application displays various types of territorial division – above all the dialect-based territorial division of the Czech-speaking language territory (i.e., 10 regions,[1] including the Bohemian borderland and the Moravian and Silesian borderland). Besides this, the boundaries of the Czech Republic's administrative units are included in the application. The basic background map can be switched to the relief map which shows natural boundaries. Thanks to this, dialectal, administrative and natural boundaries can be compared and their relations can be revealed. Mapka contains a unique representation of the historical Bohemian-Moravian border and the Moravian-Silesian border, as they were formed at the end of the 12th century and stabilized in the 14th century, as well as the German language islands in the Czech Republic.

The Mapka application provides an option of displaying series of overviews of typical dialectal features pertaining to the regions of different levels – the main three regions of the Czech Republic (Bohemia, Moravia and Silesia), eight basic dialect regions and several dialect areas or types. The descriptions are mainly focused on phonological and morphological features, as they are fundamental for the dialect division.

As a supplement to spoken corpora, Mapka displays municipality networks that have certain connections to three corpora: DIALEKT, ORAL and ORTOFON. For

[1] Central Bohemia, Northeast Bohemia, West Bohemia, South Bohemia, Bohemian-Moravian transient region, Central Moravia, East Moravia, Silesia, Bohemian borderland, Moravian and Silesian boderland.

example, there is a network of municipalities where the recordings were produced or where the speakers come from. Users can also search for municipalities/parts of municipalities in the Czech Republic, plot these points on a map, and create their own map.

5 SAMPLES OF SPOKEN LANGUAGE IN THE MAPKA APPLICATION

5.1 General characteristic of the samples

The newest version of the Mapka application offers many samples of spoken language chosen from three corpora – DIALEKT, ORAL and ORTOFON. Samples of authentic dialectal discourses chosen from the DIALEKT corpus illustrate all the main dialectal regions and some of the dialect areas or types. The samples were chosen in order to demonstrate the most typical dialectal features of the given regions. For the main dialect regions, one sample was selected from the older stratum of dialect material and one from the newer one. The samples consist of an audio recording and its two transcripts, dialectological and orthographic. Each sample is accompanied by an analysis that describes relevant dialect features (phonological, morphological, syntactic and lexical) occurring in the sample. Samples chosen from the ORAL corpus demonstrate prototypical informal speech in all the main dialectal regions. For each region, there are two chosen samples – in most cases, one from a city, one from a village. Samples contain an audio recording and a transcript. Samples chosen from the ORTOFON corpus represent spontaneous informal speech in all the main dialectal regions as well, but in this instance only one sample is chosen for the region. Samples consist of an audio recording and orthographic transcription. For both ORAL and ORTOFON, speakers were chosen so that they originated from the region where the recording was made. On the other hand, samples intend to show variability of speakers in age, education or occupation.

5.2 ORTOFON sample

Bohemian borderland, municipallity Liberec, recording made in 2013; speaker 1: male, 66 years old, born in 1947, current occupation – retired, the longest occupation – room painter; speaker 2: female, 53 years old, born in 1960, occupation – automatic and semi-automatic line operator, the longest occupation – saleswoman; both of them – place of residence in childhood/the longest residence/current residence – Liberec, Bohemian borderland (orthographic transcription[2]):

1 tady už jedem přes Německo [ne]?
2 [přes Německo]
1 no

[2] Pause punctuation: .. pause, . partition. [ ] both speakers speak simultaneously. * unfinished word. + interrupted sentence.

dlouhá pauza
2 no .. to nepřečtu
1 to já taky ne . to je hrozně malinkým písmenkem .. přes ten most jsme jeli .. už do toho d .. do Německa .. přes Německo*
2 no a t .. počkej .. jo . tady to končí . to už jsme přijížděli a tajhle už je Oybin*
1 jo .. tadyten .. + [vláček]
2 + [vláček]
1 co tam je odstavenej
2 tajhle jseš [v uniformě]
1 [no .. no] .. no .. no tady už jsme vystupovali pod tadyto už je Oybin
2 hmm
1 pod Oybinem .. to je hezký ta . úzkokolejka ten vláček že jo?
2 je
1 no . tady to je taky .. [tady to už jsme šli] +
2 [tady jsme šli těma] stezkama mezi skalama
1 + tady už jsme šli nahoru
2 tam jsem si koupila ten .. krystal . ten kámen .. [co]
1 [hmm]
2 ten čistí
1 no jo tady už jsme nahoře .. na Oybině

‘1 here we go through Germany [don’t we]?
2 [via Germany]
1 yeah
long pause
2 well .. I can’t read it
1 I can’t either . those are very small letters .. we’ve crossed the bridge .. we’re going t* .. to Germany .. over Germany
2, and t* .. wait .. yeah . it ends here . we’ve already arrived and there’s Oybin
1 yeah .. this .. + [small train]
2 + [small train]
1 that’s parked there
2 there you are [in uniform]
1 [well .. well] .. well .. well we’ve already got off here under here is Oybin
2 hmm
1 under Oybin .. that’s a nice one .. narrow gauge that small train isn’t it?
2 it is
1 well . here it is too .. [here we went] +
2 [here we went through the] paths between the rocks
1 + here we went up
2 there I bought the .. the crystal . the rock .. [that]

1 [hmm]
2 the one that cleans
1 well yeah we're up here .. on Oybin'

5.3 DIALEKT sample

The West Bohemian dialect region, municipality Tymákov, recording made in 2008; speaker: male, 72 years old, born in 1936, the longest occupations – farmer and worker, place of residence in childhood/the longest residence/current residence – Tymákov, the West Bohemian dialect region (dialectological transcription[3]):

a tenkrát kouřili i [nekuřácí, gdiš] ([hmm].) tu bili Američaňí. protože ti rozdávali [cigareti]... ([hmm].) anebo si zapálil a párkrát popotach a uš to zahoďil. a to víš, ve válce [tadi bili] ([hmm].) tabaťerki (hmm.), to bili jako líski [na] ([hmm].) @, cigareti. a fšichňi ti kuřácí mňeli málo, ([hmm].) [balili] si <PR cigareta> i třeba z dubovího [listí a] ([hmm].) fšelijak. a najednou bila hojnost! ((smích) hmm.) ti se mohli prosťe ukouřit fšichňí. f tej dobje. (hmm.) no. tagže, to [bilo @] ([jasní].), pro ti kuřāki to bilo úžasní, no. (pousmání) ti bi <SM bejvali>... ti kouřili fšichňí jako to, řikam, to kouřili i nekuřácí. diš potom přišli Američaňí. to bilo cigaret šude, ([hmm].) [habaďej], no. ale to bilo krātkou [dobu, pokut] ([hmm].) tadi bili tuti. (hmm.) gdiš vodešli, tag bil utrum. (hmm.)

'and back then, even [non-smokers smoked. when] ([hmm]) the Americans were here. because they were handing out [cigarettes] ([hmm]) or you'd light up and take a few puffs and then you'd throw it away. and you know, in the war [there were] ([hmm]) snuffboxes (hmm), they were like tickets [for] ([hmm]) @ cigarettes. and all these smokers had a few ([hmm]), [they] [packed] <PR cigarette> even from oak [leaves and] ([hmm]) all sorts of things. and suddenly there was an abundance! ([laughs] hmm) you could just smoke them all. at that time. (hmm) well. so it [was @] ([clear]) for those smokers it was amazing, well. (chuckles) you would <SM used to be>... they all smoked like, I say, even non-smokers smoked. when the Americans came. cigarettes were everywhere, ([hmm]) [plenty], well. but it was a short time, till] ([hmm]) they were here. (hmm) when they left, it was the end. (hmm)'

5.4 Differences in spoken corpora and their reflection in samples

The Mapka application is intended to capture and show the variability of spoken language in the Czech-speaking language territory. The DIALEKT samples primarily capture dialects and show a maximum of territorial differences. The ORAL and ORTOFON samples show the everyday spoken Czech language with its current

[3] [ ] both speakers speake simultaneously. ( ) speech of explorator. @ hesitation. <PR> lapsus linguae. <SM> pronounced with a laugh.

territorial differences, as well as differences caused by speaker variability – gender, age, education, and so on. Differences between samples could also be taken as differences between monological and dialogical discourse or between urban and rural speech as well.

The character of the samples reflects the differences in methods of data collection. Observing samples in 4.2, it is obvious that there are many significant differences in using language. All ORAL and ORTOFON samples are highly informal and related to the situation, hence they include a number of contact and deictic expressions. The speeches are unprepared and therefore contain phenomena such as hesitation, pauses, and unfinished words or phrases. DIALEKT samples have more monological character; they are less related to the situation of the recording, and thus contain fewer contact or deictic expressions. However, they are unprepared as well, so hesitation, pauses, and unfinished words and phrases occur.

It is up to application users to accept and comprehend the differences in corpora goals and the resulting differences in sample character. We hope that the Mapka application and samples drawn from the spoken corpora will help researchers, schools, and the general public investigate various aspects of spoken language in all of its complexities.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2023044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

Balhar, J. et al. (2011). Český jazykový atlas. Dodatky. Praha: Academia.

Dialektologická komise České akademie věd a umění (1951). Pravidla pro vědecký přepis dialektických zápisů českých a slovenských. Praha: Česká akademie věd a umění.

Goláňová, H., Waclawičová, M., and Pejcha, J. (2021). Mapka: Mapová aplikace pro korpusy mluvené češtiny. Version 1.2. Praha: ÚČNK FF UK. Accessible at: http://korpus.cz/mapka.

Goláňová, H. (2015). A new dialect corpus: DIALEKT. In K. Gajdošová – A. Žáková (eds.): Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography). Lüdenscheid: RAM-Verlag, pages 36–44.

Goláňová, H., and Waclawičová, M. (2021). Mapka: A Map Application for Working with Corpora of Spoken Czech. Jazykovedný časopis, 72(2), pages 502–509.

Goláňová, H., and Waclawičová, M. (2019). The DIALEKT corpus and its possibilities. Jazykovedný časopis, 70(2), pages 336–344.

Goláňová, H., Waclawičová, M., and Lukeš, D. (2021). DIALEKT: nářeční korpus, version 2 from 23. 12. 2021. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.

Kopřivová, M., Goláňová, H., Klimešová, P., Komrsková, Z., and Lukeš, D. (2014). Multi-tier Transcription of Informal Spoken Czech: The ORTOFON Corpus Approach. In Complex Visibles Out There. Olomouc: Univerzita Palackého v Olomouci, pages 529–544.

Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017). ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.

Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., and Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, version 1 from 2. 6. 2017. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.

sciendo
DOI 10.2478/jazcas-2023-0039

MORPHOSYNTACTIC ANNOTATION IN UNIVERSAL DEPENDENCIES FOR OLD CZECH

DANIEL ZEMAN[1] – PAVEL KOSEK[2]
– MARTIN BŘEZINA[2] – JIŘÍ PERGLER[3]
[1] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[2] Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[3] Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic

ZEMAN, Daniel – KOSEK, Pavel – BŘEZINA, Martin – PERGLER, Jiří: Morphosyntactic Annotation in Universal Dependencies for Old Czech. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 214 – 222.

Abstract: We describe the first steps in preparation of a treebank of 14th-century Czech in the framework of Universal Dependencies. The Dresden and Olomouc versions of the Gospel of Matthew have been selected for this pilot study, which also involves modification of the annotation guidelines for phenomena that occur in Old Czech but not in Modern Czech. We describe some of these modifications in the paper. In addition, we provide some interesting observations about applicability of a Modern Czech parser to the Old Czech data.

Keywords: morphology, dependency syntax, universal dependencies, Old Czech language

1 INTRODUCTION

Universal Dependencies (UD)[1] (de Marneffe et al. 2021; Zeman 2015) is a project aiming at developing morphosyntactic annotation guidelines applicable to all human languages, and creating treebanks annotated according to the guidelines. In the course of the last nine years, UD grew from just 10 languages in release 1.0 to 245 treebanks of 141 languages from 30 families in release 2.12 (May 2023), and is now a de facto standard for morphological and syntactic annotation. With well over 500 contributors, UD has also become a worldwide research community.

Besides covering many modern languages, dialects and genres, UD also contains data for a considerable number of classical languages, such as Akkadian, Sanskrit or Latin, and for some historical varieties, such as Old French and Old East Slavic.

The Slavic genus is particularly well represented, with most languages having a decent UD treebank, and some of them having more than one treebank (Tab. 1).

[1] https://universaldependencies.org/

This includes two historical languages that have their own ISO 639 code: Old Church Slavonic and Old East Slavic.

Belarusian	305
Bulgarian	156
Croatian	199
Czech	2,227
Old Church Slavonic	199
Old East Slavic	333
Polish	499
Pomak	87
Russian	1,832
Serbian	97
Slovak	106
Slovenian	297
Ukrainian	123
Upper Sorbian	11

Tab. 1. Slavic languages in UD 2.12 with treebank sizes (× 1,000 words)

Treebank	**Train**	**Dev**	**Test**
CAC	473	11	11
CLTT	14	11	11
FicTree	134	17	17
PDT	1,173	159	174
PUD			19
Total	**1,794**	**198**	**232**

Tab. 2. Czech treebanks in UD 2.12 with data splits and sizes (× 1,000 words)

In the present paper we describe the first steps towards creating a UD treebank of the oldest preserved stage of Czech. With five treebanks of various genres and more than two million words (Tab. 2), Czech is the best represented Slavic language in UD, yet all the available data represents the modern language, spanning roughly the period 1971–2016. Morphosyntactic diversity in the Czech treebanks can be largely attributed to different genres; diachronic variation is primarily lexical. In contrast, the oldest Czech texts from around 1300 contain grammatical features that later vanished from the language. It would be interesting and useful to be able to compare them both with modern Czech and with other languages within a unified annotation framework.

Note that unlike some other historical languages, there is no ISO 639 code for Old Czech,[2] so technically it must be treated as one of the varieties of the single

[2] Also unlike Old Church Slavonic, which is older (it was used on the Czech territory since the 9th century) but distinct from Czech.

Czech language. This can be also seen as an advantage, as in the future, the continuous development of the language from its earliest stages to modern times can be studied and documented under one set of language-specific guidelines.

A large body of digitized Old Czech texts is available in the diachronic part of the Czech National Corpus[3] and in Old Czech Text Bank (Staročeská textová banka, STB);[4] in its current version, the latter contains almost 7 million tokens. The texts have been transcribed into modernized orthography while preserving their linguistic features. This has the double advantage of making the contents more easily accessible and standardizing the spelling. The transcription was done manually, as it often involves disambiguation where the unsettled orthography did not capture pronunciation unequivocally. For example, what is sometimes considered the oldest preserved Czech sentence,[5] originally spelled as (1), is transcribed as (2). A possible translation to modern Czech is shown in (3):[6]

(1) *Pauel dal geſt ploſcouıcıh zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma bogucea aſedlatu.*
(2) *Pavel dal jest Ploskovicích zem'u, Vlach dal jest Dolas zem'u bogu i sv'atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.*
(3) *Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.*
(4) 'Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata.'

The texts in the above sources are not annotated any further. Although lemmas, part-of-speech tags and morphological features are available for part of the Old Czech vocabulary, context-based disambiguation of individual tokens is still pending.

2 DATA SELECTION AND PREPROCESSING

For the pilot UD annotation, we chose two versions of the Gospel of Matthew from the oldest Czech translations of the Bible: the "Dresden Bible" (dated ca. 1365) and the "Olomouc Bible" (1417). Besides having two parallel Old Czech translations of the same Latin source, we can also compare the annotated corpus to other

[3] https://www.korpus.cz/kontext/query?corpname=diakorp_v6

[4] https://korpus.vokabular.ujc.cas.cz/

[5] A note from early 13th century on the (much older and written in Latin) charter of the Litoměřice chapter. Státní oblastní archiv v Litoměřicích, fond litoměřické kapituly (Litoměřice, Czechia), sign. R2, 1v. Editor Černá, Alena M. The status as the oldest Czech sentence has been challenged by Dittmann (2012).

[6] Precise wording is debatable; the word *dušník* is not used in Modern Czech unless as a historical term. Our goal is to illustrate morphosyntactic and phonological changes.

treebanks in UD. As many as 16 languages in UD 2.12 contain some biblical material. Out of these, at least[7] 6 contain fragments of the Gospel of Matthew: Ancient Greek, Gothic, Latin, Old Church Slavonic, Romanian, and Yoruba.

The source text is segmented to chapters and verses but not to sentences. A verse often corresponds to a sentence, but sometimes the relation is 1:N or N:1. We used UDPipe 2[8] (Straka 2018), trained on UD Czech PDT 2.6, to obtain an initial tokenization and sentence segmentation. The segmentation was then manually corrected and frozen for any subsequent processing.

The whole corpus (Dresden and Olomouc versions of the Gospel) comprises 2,447 sentences and 44,574 tokens.

Besides tokenization and segmentation, UDPipe also provided initial annotation on the morphological and syntactic layers. Its accuracy is, naturally, relatively low, given the significant differences between the data UDPipe was trained on (news from the early 1990s) and the target text. The intended workflow is bootstrapping: After manual correction of the initial segment of the data, the corrected part will be used to re-train the parser, which will then pre-annotate the next segment, hopefully with fewer errors, thus sparing more effort of the human annotators. The process will be repeated until the whole text is annotated and verified by humans. The existing UD annotation guidelines for Czech will be gradually adapted to the specifics of Old Czech during the process. The main adaptation steps described below have already been done during the first round, nevertheless, further adjustments during the later stages cannot be excluded.

3 LEMMATIZATION

Lemmatization of the old language is a complex issue, as the language was not regulated in any way. Therefore, lemmatization does not only involve picking a canonical form of a lexeme, such as the infinitive of a verb, but sometimes also normalization of several variants of a morpheme (including the stem), e.g. *Křtitel – Křstitel – Krstitel* 'Baptist'.

Moreover, it would be beneficial to be able to match old and modern forms of the same lexeme despite the changes they underwent over the centuries; this can be achieved if the old forms are annotated with the corresponding modern lemma. Not only can a human user take advantage of this, it is also the lemma that a parser with a Modern-Czech model knows. For example, the infinitive in Old Czech typically ends in *-ti* and while this form may occasionally occur in current texts, it is considered archaic, whereas the canonical form (and lemma of the verb) ends in *-t*. A Modern

[7] As identified by "Ref=MATT" in the data; this annotation is optional in UD, some other treebanks may thus have verses from Matthew without being counted here.

[8] https://ufal.mff.cuni.cz/udpipe/2

Czech parser may stand a chance of guessing the modernized lemma of the Old Czech verbs, but it will never predict the archaic infinitive.

However, there is a downside. When studying Old Czech without comparing it to newer stages, the modernized lemma seems inappropriate and may not be preferred by the users. There are also theoretical questions about which changes count as variants of one lexeme (e.g. if the conjugation class changes, should we say it is a different lexeme, hence a separate lemma?) Some words have fallen out of use and their modern form is not attested, even if we can estimate how it would look like, following phonological evolution of the language. For all these reasons, we maintain two lemmas for each word: the modernized one, and a canonical form as expected around the year 1300. Though in UDPipe experiments, we only evaluate the modern lemma.

4 MORPHOLOGICAL FEATURES

While in Modern Czech the past tense is formed periphrastically, using the l-participle and a finite auxiliary *být* 'be', Old Czech had finite simple past forms called *imperfect* and *aorist*. These were also attested in Old Church Slavonic and they have survived in a few modern Slavic languages, such as Bulgarian or Upper Sorbian. We thus use morphological features that are used in UD for these languages. For imperfect, we use `Tense=Imp` (we cannot use the `Aspect` feature because it would clash with the lexical aspect that has developed in Slavic languages), for aorist we use `Tense=Past` (together with `VerbForm=Fin`, meaning the finite verb, while the l-participle is tagged `Tense=Past` and `VerbForm=Part`). Sigmatic and asigmatic forms of the aorist are distinguished using the language-specific feature `Variant`.

Other than that, the morphological features already defined for Modern Czech were sufficient, although some of them occur in new combinations. The dual number (`Number=Dual`) is in Modern Czech used only for certain forms, mostly adjectives and nouns related to paired body organs, while in Old Czech it can occur almost everywhere where singular or plural can.

Animacy of masculine nouns in the 14th century did not yet play the role it plays in the grammar today, yet we tentatively annotate it to stay consistent with Modern Czech datasets.

Present (simultaneous) and past (anterior) converbs (`VerbForm=Conv`; also called gerunds) still exist in modern data, although they are very rare and archaic; in Old Czech their frequency is much higher but the annotation is analogous. However, the neuter singular form, which nowadays concurs with feminines, was identical to masculines in Old Czech. We also briefly considered adding the `Case` feature, as there are claims (Gebauer 1898, p. 83, § 35) that some of the converb forms correspond to the accusative, but we abandoned the idea both for consistency with modern data and for inability to reliably assign case values to some of the forms.

5 TAGGING RESULTS

In this section we report on the accuracy of the UDPipe models in the initial stages of the project. First of all, it is important to note that there are currently two versions of the UDPipe tool available: 1.2 and 2.0. Version 2.0 is known to perform significantly better; however, it is only available as a web service with pre-trained models. If we want to train a model on our own data, we have to downgrade to UDPipe 1.2, which is available as a standalone, trainable tool (but there are pre-trained models for it as well).

Currently available pre-trained models for UDPipe 1.2 are based on UD release 2.5; for UDPipe 2.0, one can choose between UD releases 2.6 and 2.10.[9] The differences between the UD releases should not be large and the size of the training data should be stable; nevertheless, there may be corrections of annotation errors, meaning that the resulting models are not the same. UDPipe did not have access to any dedicated morphological dictionary in our experiments, only to the training data.

Pre-trained models correspond to the four Czech UD treebanks that have designated training data (see Tab. 2). All four are automatic conversions from the PDT annotation style. Besides varying sizes of the treebanks (bigger is better), the genre also plays a role. None of the Czech treebanks contains biblical texts (which would be the best match, even if in a modern variety of the language). The largest treebank, PDT (Hajič et al. 2020), is based on newspapers and journals from the early 1990s. CAC (Vidová Hladká et al. 2008) is non-fiction from 1971–1985. CLTT (Kríž and Hladká 2018) is small and focused on the legal domain, containing the Accounting Act. Finally, FicTree (Jelínek 2017) contains fiction from 1991–2007; genre-wise, this treebank is probably most similar to our target data.

There are various sources of divergence between the trained models and the Old Czech data. First, the vocabulary is quite different in the news (or any other training genre) vs. in the Bible. Even parsing a Modern Czech Bible translation would be difficult because of this. Second, as mentioned above, some of the old words have fallen out of use and the parser cannot know them. Third, for words that survived to modern days, even though the old orthography is cleaned, transcribed and unified, many word forms still differ from their modern counterparts because their old pronunciation differs: *sě* vs. *se* ‘oneself’ (the reflexive marker), *viece* vs. *více* ‘more’ etc. Fourth, the morphological differences described above mean that some old forms do not exist any more (imperfect such as *bieše* ‘was’; aorist such as *vecě* ‘said’, *jide* ‘went’; most forms of the dual) or are much less frequent than in Old Czech (converbs such as *řka* ‘saying’, *přistúpiv* ‘having approached’).

[9] Only models on UD 2.6 were available at the time when we ran UDPipe 2.0.

We have manually verified the first five chapters of the Dresden version of Matthew, which amounts[10] to 148 sentences and 2,665 words. This data can be used to evaluate the accuracy of the initial model, trained only on Modern Czech. All five chapters can be used when training a model to preprocess the next segment of the corpus. However, if we want to evaluate a re-trained model on gold-standard Old Czech data, we need to split the dataset into training and test part. In such experiments, we reserve chapters 1–4 (86 sentences, 1,669 words) for training and chapter 5 (62 sentences, 996 words) for testing.

Since the present annotation is manually checked only at the morphological layer, we report accuracy of lemmatization and tagging (separately the main part-of-speech category and the remaining morphological features); we do not study the accuracy of dependency relations yet.

UDPipe 2 Model	**PDT**	**CAC**	**CLTT**	**FicTree**
Lemma	99.17	98.95	99.30	99.21
UPOS	99.30	99.54	99.49	98.69
Features	97.70	97.07	95.16	96.80

Tab. 3. For comparison, we show in-domain accuracy of UD 2.6 pre-trained models. Each model is evaluated on the test data from the corresponding treebank.

UDPipe 1.2 Model	**PDT**	**CAC**	**CLTT**	**FicTree**
(Modern) lemma	**69.68**	68.67	51.20	66.97
UPOS	**76.71**	74.00	55.82	70.58
Features	**54.82**	52.71	38.55	48.19

Tab. 4. Tagging Dresden Matthew chapter 5 (that is, all results are on the same test set). UDPipe 1.2 models pre-trained on UD 2.5.

UDPipe 1.2 Model	**PDT**	**FicTree**	**DMt1–4**	**Fic+Mt**
(Modern) lemma	69.68	66.97	67.27	**78.41**
UPOS	76.71	70.58	74.90	**85.44**
Features	54.82	48.19	58.84	**64.86**

Tab. 5. Tagging Dresden Matthew chapter 5 (that is, all results are on the same test set). First two columns are UDPipe 1.2 models pre-trained on UD 2.5, repeated from Tab. 4. The third column is a model trained on Dresden Matthew chapters 1–4. The last column is a combined model trained on chapters 1–4 and on FicTree from UD 2.10.

The lemmatization and tagging scores on chapter 5 of the Gospel of Matthew from the Dresden Bible are shown in Tab. 4 and 5. For comparison, Tab. 3 shows what scores one can expect when the pre-trained models are applied to Modern Czech data. Note however that these numbers are not directly comparable even

[10] Chapter 1 is not complete, the genealogy in the beginning was omitted.

among themselves, as each model was evaluated on different test set (namely on the test set from the treebank from which the model's training data were taken).

The DMt1–4 column in Tab. 5 demonstrates how important it is to train on data from the same domain and same stage of the language: Despite the fact that the training data is ridiculously small (less than 2K words), the results are not much worse (and for Features they are even better) than the Modern Czech model trained on over 1M words from PDT.

When combined with a larger Modern Czech corpus, the four chapters of Matthew provide for a model that is much better than any of the other models in isolation (a jump of 9–10 percent points). The results also proved that FicTree is, out of the Modern Czech treebanks, the best fit for our biblical data; when we combined Matthew with PDT, which is ten times larger than FicTree, the negative effect of the out-of-domain data prevailed and the scores were worse than with FicTree.

6 CONCLUSION

We have described the initial steps in order to create a UD-style annotated treebank of the Old Czech biblical texts. Some peculiarities of the historical language were discussed, a bootstrapping approach with a Modern Czech parser was proposed and first experiments evaluated. In the next phase we will address the syntactic layer. As soon as the syntactic annotation is ready, we intend to publish the treebank in a future release of Universal Dependencies.

ACKNOWLEDGMENTS

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation (GAČR), and The Grammar and Lexicon of Czech III – 2023 (MUNI/A/1249/2022).

The work uses data and tools provided by the research infrastructure LINDAT/CLARIAH-CZ (https://lindat.cz/), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2023062).

The authors are grateful for the input provided by the colleagues from Masaryk University and the Czech Language Institute, in particular Klára Osolsobě, Olga Navrátilová, Kateřina Granátová, Martina Ježová, Linda Rudenka, Radek Čech, Jana Zdeňková and Ondřej Svoboda.

References

Dittmann, R. (2012). Problém tzv. nejstarší české věty. Bohemica Olomucensia, 4(1), pages 26–36. Accessible at: https://bohemica.actavia.cz/pdfs/boh/2012/01/03.pdf.

Gebauer, J. (1898). Historická mluvnice jazyka českého. Díl III., II. Tvarosloví. Časování. Wien: F. Tempský. 508 p. Accessible at: https://kramerius5.nkp.cz/view/uuid:223066b0-6f7b-11eb-9f97-005056827e51?page=uuid:2bd40f78-a469-43e9-bc4b-2b579e00865c.

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., and Štěpánková, B. (2020). Prague Dependency Treebank – Consolidated 1.0. In Proceedings of LREC, pages 5208–5218.

Jelínek, T. (2017). FicTree: a Manually Annotated Treebank of Czech Fiction. In Proceedings of ITAT, pages 181–185.

Kríž, V., and Hladká, B. (2018). Czech Legal Text Treebank 2.0. In Proceedings of LREC, pages 4501–4505.

Marneffe de, M.-C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pages 255–308.

Staročeská textová banka [online]. (2006–2023). Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Verze dat 1.1.22 [cit. 26. 3. 2023]. Accessible at: https://vokabular.ujc.cas.cz/banka.aspx?idz=STB.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Bruxelles: Association for Computational Linguistics, pages 197–207.

Vidová Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., and Raab, J. (2008). The Czech Academic Corpus 2.0 Guide. In PBML 89, pages 41–96.

Vokabulář webový: webové hnízdo pramenů k poznání historické češtiny (2006–2023). [online]. Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Verze dat 1.1.22 [cit. 26. 3. 2023]. Accessible at: https://vokabular.ujc.cas.cz/.

Zeman, D. (2015). Slavic Languages in Universal Dependencies. In Natural Language Processing, Corpus Linguistics, E-learning, pages 151–163, Lüdenscheid: RAM-Verlag.

CORPUS BUILDING

sciendo
DOI 10.2478/jazcas-2023-0040

THE EFFECT OF (HISTORICAL) LANGUAGE VARIATION ON THE EAST SLAVIC LECTS LEMMATISERS PERFORMANCE

ILIA AFANASEV – OLGA LYASHEVSKAYA
– STEFAN REBRIKOV – YANA SHISHKINA
– IGOR TROFIMOV – NATALIA VLASOVA
Independent researchers

AFANASEV, Ilia – LYASHEVSKAYA, Olga – REBRIKOV, Stefan – SHISHKINA, Yana – TROFIMOV, Igor – VLASOVA, Natalia: The Effect of (Historical) Language Variation on the East Slavic Lects Lemmatisers Performance. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 225 – 233.

Abstract: The need to develop tools for historical and regional variations is becoming more urgent in natural language processing. In this paper, we present two candidate systems for lemmatising historical East Slavic lects (Late Old East Slavic and Middle Russian), as well as modern regional East Slavic lects (Belogornoje and Megra): BERT-based end-to-end pipeline with language-specific heuristics and sequence-to-sequence BART-based encoder-decoder. To evaluate their predictions, we use accuracy score and string similarity measures, such as Levenshtein distance. The BERT-based model is more suitable for the regional data, achieving 85% accuracy score, and only 74% on the historical data. BART-based model climbs up to 92.6% accuracy score on the historical data, yet gets only 80% on the regional data. We provide an error analysis and discuss ways to enhance models, such as dictionary lookup and spellchecker.

Keywords: East Slavic, language variation, lemmatisation, dialectology, historical linguistics, historical NLP

1 INTRODUCTION

Lemmatisation is an NLP task that includes predicting the dictionary form of a token, a lemma. Lemmatisation is crucial for the linguistic downstream tasks, such as dictionary compilation, or grammar description (Straka – Straková 2017; Bergmanis – Goldwater 2018; Kanerva et al. 2021), especially for historical and regional lects (Berdičevskis et al. 2016; de Graaf et al. 2022). However, the impact of language variation on the performance of lemmatisers is still understudied.

In this paper, we examine how two types of language variation, historical and regional, challenge two types of lemmatisers: a transformer-based end-to-end

pipeline system with heuristics that enhance its performance on a standard modern language (Anastasyev 2020), and a sequence-to-sequence transformer-based lemmatiser (Lewis et al. 2020).

We study two types of variation that may affect the given lemmatisers' performance. Historical material consists of Late Old East Slavic and Middle Russian texts. The northern and southern Russian territorial lects material represents the regional data. We hypothesise the following:

H1: Prioritising rare words (and, subsequently, non-productive word-formation models) enhances the performance of the language-specific pipeline model lemmatiser module over the regional data.

H2: Variation within the training historical data helps the sequence-to-sequence transformer-based lemmatiser to demonstrate better results on the historical evaluation subset. In addition, we are going to examine whether preliminary morphological tagging enhances the results of the sequence-to-sequence lemmatiser.

Section 2 discusses the previous research on the topics of East Slavic NLP and lemmatiser development current trends. In Section 3 we describe the training datasets and the test data. Section 4 presents the models used in the study. In Section 5 we report the conducted experiments and discuss their results. Section 6 wraps up the current research and sketches its further direction.

2 RELATED WORK

The most widely used lemmatisers are multilingual, with the underlying inference engines being language-agnostic (Straka – Straková 2017; Bergmanis – Goldwater 2018; Kanerva et al. 2021). These tools generally utilize sequence-to-sequence architecture (Sutskever et al. 2014; Cho et al. 2014) implemented via encoder-decoder transformers such as BART (Lewis et al. 2020).

The current trends stimulate the researchers of particular languages and language groups to develop the lemmatisers for a particular lect or a set of lects (Anastasyev 2020; Fernández 2020). The scholars employ external resources such as dictionaries (Milintsevich – Sirts 2021), thus significantly improving the models' efficiency for extremely low-resourced lects (de Graaf et al. 2022).

East Slavic NLP for a long period concentrated around standard Russian (Anastasyev 2020) and its historical varieties (Berdičevskis et al. 2016; Lyashevskaya – Penkova 2021; Pedrazzini – Eckhoff 2021). However, the particular texts that are the focus of this study were digitalised fairly recently and were not tagged (Novokhatko 2012; Kusmina – Filippova 2013).

The interest in standard Ukrainian (Omelianchuk et al. 2020; Omelianchuk et al. 2021) and standard Belarusian (Shishkina – Lyashevskaya 2021) sparked in recent years. However, there is hardly any information on attempts to develop NLP tools for smaller modern East Slavic lects, despite, for instance, Saratov dialectological school's

active study of the Belogornoje and Megra lects and the development of a corpus and dialectal dictionaries (Kruchkova – Goldin 2011; Kruchkova – Goldin 2015).

3 DATA

The first training dataset consists of different standard Modern Russian texts balanced by genres (news, poetry, fiction, social media, the latter taken from the Taiga corpus (Shavrina – Shapovalova 2017)), orthography (premodern/modern), and periods (the 1700s, 1800s, and 1900s–2020s): overall, ca. 2,000,000 tokens.

The second training dataset consists of the Late Old East Slavic and Middle Russian legal texts from 1400 to 1700 taken from different collections (Russian History Library 1875; Rozysknyje dela o Fedore Shaklovitom i jego soobshchnikakh 1893; Likhachov 1954; Cherepnin 1961; Ankhimiuk 2000). The overall size of this collection is ca. 923,000 tokens.

The regional test data consist of two groups of East Slavic lects, northern Megra (Vologda Region, Russia), and southern Belogornoje (Saratov Region, Russia). The texts spelt close to the actual pronunciation of the native speakers are taken from the Saratov dialectological corpus (Kruchkova – Goldin 2011). Together they form a dataset of ca. 8,000 tokens.

The historical test data consist of a single set of documents, Besobrasow's archive (Novokhatko 2012; Kusmina – Filippova 2013). These are legal texts of the latter half of the 1600s. The overall size of these texts achieves ca. 437,000 tokens.

4 METHOD

In this research, we compare two approaches to the lemmatisation of historical and regional East Slavic varieties.

The first approach is a robust end-to-end pipeline model, which performs morphological tagging, lemmatisation, and dependency parsing. We use qbic (Anastasyev 2020), a state-of-the-art ensemble for Russian that makes use of a pre-trained RuBERT model. We modify the ensemble with a RuBERT-large model; lemmatisation making use of morphological classification; dictionary lookup (a mapping of word with its part-of-speech to its lemma); correcting symbol sequences forbidden in Russian; and orthographical normalisation. We refer to this model as Rubic.

The second approach focuses on sequence-to-sequence lemmatisation that can use morphological information from data. We use only the BART-large pre-trained model without additional enhancements.[1]

[1] Source code is available at https://github.com/The-One-Who-Speaks-and-Depicts/transformer-lemmatiser; models are available at https://huggingface.co/djulian13/bart-large-Modern-Russian-lemmatisation and https://huggingface.co/djulian13/bart-large-Middle-Russian-lemmatisation.

The training generates four models: Rubic for modern data, BART-large for modern data, Rubic for historical data, BART-large for historical data. We use some heuristics for augmenting the modern training data: capitalisation, quotation marks-to-guillemets conversion, and jo-fication (this heuristic transforms e into ë in tokens when they are interchangeable by the standard Russian rules). These additional heuristics allow to add contexts for words from rare grammatical paradigms.

We test the models on historical and regional data. Initially, the conditions for Rubic and BART-large test on regional data are not equal. BART-large lemmatiser is trained with the use of morphological information, which the regional data lack. To solve this issue, we train the morphological tagger (Scherrer 2021) on standard Russian data. It achieves the 87% F1-score, and we use it to produce a silver (more or less reliable, but not reliably checked by humans) morphological tagging for the regional data.

For an initial evaluation, we use an accuracy score. We also implement Levenshtein (Levenshtein 1966), Damerau-Levenshtein (Damerau 1964), and Jaro-Winkler (Jaro 1989; Winkler 1990) distances, a method that allows us to get fine details (Lyashevskaya – Afanasev, in print), to study Rubic and BART-large performance on the regional data.

The analysis aims at understanding the key errors and the explanation of models' performance, both in comparison and on their own. The following discussion of the results is required to get an outline for future work on both approaches.

5 RESULTS AND ANALYSIS

The four models achieve an accuracy score from 85 to 99% (depending on the exact subset) on both modern and historical data. Two series of experiments are conducted. The first series of experiments include testing the Rubic and BART-large models, fine-tuned on the historical data, on Besobrasow's archive. The second series includes testing the Rubic and BART-large, fine-tuned on the modern data, on regional East Slavic lects material.

5.1 Experiment 1

Tab. 1 presents the results of the experiments on the Besobrasow's archive.

Dataset	Besobrasow's archive
Rubic (token normalisation)	73.8
BART-large	85.0
BART-large (token normalisation)	**92.6**

Tab. 1. The accuracy score, %, of the historical East Slavic datasets lemmatisation by Rubic and BART-large

Both models are sensitive to the orthographical variation: BART-large gets the 0.07 boost when the tokens are normalised. The sequence-to-sequence architecture performs better: the Rubic augmentations enhance its performance on standard Russian and not the historical material (Lyashevskaya et al. 2023).

BART-large performance is not perfect. The historical data are significantly more heterogeneous than the modern data, and it is harder for the model to predict non-standard inflexion because the chance to meet it in the training data is smaller. For comparison, an element -мѣрити is often replaced by a more frequent -мерети, in verbs like смѣрити 'measure', cf. умерети 'die'. Nouns and adjectives are also affected, cf. выклатка instead of выкладка 'facing' influenced by оплатка 'wafer', and друговъ instead of другой 'other', influenced by каковъ 'which'. There are also non-standard transformations, cf. еиц >> яйцо 'egg', for which BART-large correctly predicts the lemmatisation model, yet does not predict the in-root phonetic alternation producing the forms like ейцо. Sometimes, the specifics of the annotation schema cause errors. The earlier data use lemma и for a masculine 3rd person pronoun, and the later data uses the lemma онъ, which confuses the model, unaware of the East Slavic lects evolution.

5.2 Experiment 2

Next, we test the models' abilities to operate within the conditions of synchronic variation. The dataset presents an additional challenge for a sequence-to-sequence model. The morphological tags of the test data do not correspond to the training tag set, so we use silver tagging made with Scherrer (2021). We run the BART-large model separately on non-tagged and tagged data.

Tab. 2 and 3 demonstrate the results for the Megra and Belogornoje datasets.

Metrics	A	L	L: N	D-L	D-L: N	J-W	J-W: N
Rubic	**86.9**	**0.31**	**0.31**	**0.31**	**0.31**	**~98.0**	**~98.0**
BART-large	49.66	1.08	0.94	1.08	0.94	89.4	94.7
BART-large (pre-tagged data)	82.67	0.37	0.37	0.37	0.37	95.7	95.7

Tab. 2. The results of the Megra dataset lemmatisation by Rubic and BART-large: accuracy score (A), %; averaged string similarity measures (Levenshtein distance (L), Levenshtein distance, normalised (L:N), Damerau-Levenshtein distance (D-L), Damerau-Levenshtein distance, normalised (D-L: N); averaged Jaro-Winkler distance (J-W) and Jaro-Winkler distance, normalised (J-W: N)), %.

Metrics	A	L	L: N	D-L	D-L: N	J-W	J-W: N
Rubic	**87.8**	**0.25**	**0.25**	**0.25**	**0.25**	**98.2**	**98.2**
BART-large	52.97	1.02	0.85	1.02	0.85	91.3	96.1
BART-large (pre-tagged data)	84.89	0.29	0.29	0.29	0.29	97.9	97.9

Tab. 3. The results of the Belogornoje dataset lemmatisation by Rubic and BART-large: accuracy score (A), %; averaged string similarity measures (L, L: N, D-L, D-L: N); averaged Jaro-Winkler distances (J-W, J-W: N), %.

In this experiment BART-large performs worse than Rubic: the heuristics of the latter enable it to overcome some difficulties of the regional data lemmatisation, like ещё/еще contrast. Note that BART-large demonstrates dependency on preliminary morphological tagging: the model run on raw data falls behind the model run on tagged data by 30%.

Generally, the high results of string similarity measures show that each model captures the general principle of lemmatisation: the output, lemma, is usually similar to the token that is provided as a part of the input. The equal results of Levenshtein and Damerau-Levenshtein distances support the claim: they mean that there are no character transpositions in target/output pairs. The equal normalised and non-normalised string similarity measures for Rubic and BART-large, run on tagged data, demonstrate that the models generate sequences similar to the target and do not prioritise a particular inflexion type over the others. BART-large, run on raw data, on the other hand, often produces erroneous output.

There are some general issues that may explain why the models struggle with the regional data. Lemmata in gold data reflect a particular lect norm, not a standard norm. Thus, correctly – by standard means – predicted еще 'yet' becomes incorrect, when compared to ещё. The relatively high string similarity measures scores support the hypothesis of the relatively high rate of this kind of errors. Linguistic differences between modern standard Russian and regional East Slavic lects also negatively affect the models' performance. Models are not aware of some pronoun forms, and lemmatise, for instance, мни '1.SG.DAT' as мни and not я '1.SG.NOM'. Non-standard spelling, combined with non-productive paradigms, leads to errors: models predict робят instead of робёнок 'child-SG.NOM'. The gold data also contain multi-word lemmata which training data lacks, which also results in errata, cf. gold вот и and predicted и 'and'. These issues are well-known for the smaller lects lemmatisation: there is no consensus on what should be considered a lemma in such cases (de Graaf et al. 2022).

6 CONCLUSION

In this paper, we studied how the synchronic and diachronic variation in East Slavic languages impacts the lemmatisers' performance. We compared the lemmatiser module of a robust end-to-end BERT-based pipeline (Rubic) and the independent sequence-to-sequence BART-based model (BART-large). Different kinds of variation negatively affect the performance of both models, lowering their average accuracy score to 80–90% from more than 90% on the validation data. BART-large overcomes historical variation more easily, especially when provided with gold morphological tagging, which is supported by string similarity measures evidence. Rubic does not depend on morphological data and better overcomes the synchronic variation of the territorial East Slavic lects closely related to standard

Russian. Both models are confused by orthographical variation, non-standard paradigms, and difference of lemma choice in golden data.

Both models need enhancements. For BART-large, we could suggest implementing heuristics, in a similar fashion to Rubic's. Built-in lect identification may be of some help, normalising tokens before their lemmatisation. As symbol-by-symbol generation causes a lot of BART-large errors, this model may benefit from spellchecker or a dictionary lookup during the postprocessing. The biggest issue is the BART-large model being a stand-alone lemmatiser, it needs the morphological tagging complement to be efficient, in contrast with Rubic. The development of tagging tool should be the primary focus of future research. The same heuristics may be useful for Rubic, a model that is even more suitable for this kind of modification, as it already employs some heuristics. It may also benefit from other types of pre-trained language models, pushed up in its architecture.

The datasets augmentation approaches include incorporating irregular lemmatisation models, as well as the balancing of wordforms by the frequency of word changing paradigm. The modern dataset will benefit from the addition of a regional component.

References

Anastasyev, D. (2020). Exploring pretrained models for joint morphosyntactic parsing of Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue", 19, pages 1–12, Moscow, Russia.

Ankhimiuk, U. V. (2000). Soligalicheskije akty iz "Arkhiva Volynskikh". In A. V. Antonov (ed.): Russian Diplomatary. Moscow: Archeographical center, pages 25–42.

Berdičevskis, A., Eckhoff, H., and Gavrilova, T. (2016). The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. In Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog», pages 99–111, Moscow, Russia. RSSU.

Bergmanis, T., and Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Cherepnin, L. V. (1961). Akty feodal'nogo zemlievladenija i khozyajstwa XIV – XVI vekov (in 3 volumes). Moscow: USSR Academy of Sciences.

Cho, K., Merriënboer van, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), pages 171–176.

Fernández, L. G. (2020). A contribution to Old English lexicography. NOWELE / North-Western European Language Evolution, 73(2), pages 236–251.

Graaf de, E., Stopponi, S., Bos, J. K., Peels-Matthey, S., and Nissim, M. (2022). AGILe: The first lemmatizer for Ancient Greek inscriptions. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5334–5344, Marseille, France. European Language Resources Association.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Association, 84, pages 414–420.

Kanerva, J., and Ginter, F., and Salakoski, T. (2021). Universal lemmatizer: A sequence--to-sequence model for lemmatizing universal dependencies treebanks. Natural Language Engineering, 27(5), pages 545–574.

Kruchkova, O., and Goldin, V. (2011). Corpus of Russian dialect speech: concept and parameters of evaluation. In Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2011", pages 359–367, Moscow, Russia.

Kruchkova, O., and Goldin, V. (2015). The parameters of text processing for the Russian dialect corpus. In Proceedings of the international conference "Corpus linguistics — 2015", pages 307–314, Saint Petersburg, Russia.

Kuzmina, O. V., and Filippova, I. S. (2012). Arkhiv stol'nika Andreja Il'jicha Besobrasowa, vol. II. Moscow: Russian History Insitute of Russian Academy of Sciences, 877 p.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), pages 707–710.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Likhachov, D. S. (1954). Puteshestvija russkich poslov XVI – XVII vv. Statejnyje spiski. Moscow, Leningrad: USSR Academy of Sciences, 490 p.

Lyashevskaya, O., Afanasev, I., Rebrikov, S., Shishkina, Y., Suleymanova, E., Trofinov, I., and Vlasova, N. Disambiguation in context in the Russian National Corpus: 20 years later. In Proceedings of International Conference "Dialogue 2023", pages 1–12, Online.

Lyashevskaya, O., and Afanasev, I. (in print). String similarity measures for evaluating the lemmatisation in Old Church Slavonic. In Proceedings of International Conference on Historical Lexicography and Lexicology. La Rioja, Spain. Universidad de La Rioja.

Lyashevskaya, O., and Penkova, Y. (2021). Revised entries in the multi-volume edition and TEI encoding: a case of the historical dictionary of Russian. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, vol. II, pages 655–662. Komotini, Greece. Democritus University of Thrace.

Milintsevich, K., and Sirts, K. (2021). Enhancing sequence-to-sequence neural lemmatization with external resources. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3112–3122, Online. Association for Computational Linguistics.

Novokhatko, O. V. (2012). Arkhiv stol'nika Andreja Il'jicha Besobrasowa, vol. I. Moscow: Russian History Insitute of Rassian Academy of Sciences, 903 p.

Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 163–170, Seattle, WA, USA. Online. Association for Computational Linguistics.

Omelianchuk, K., Raheja, V., and Skurzhanskyi, O. (2021). Text Simplification by Tagging. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–25, Online. Association for Computational Linguistics.

Pedrazzini, N., and Eckhoff, H. M. (2021). OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data. Software Impacts, 8, pages 1–4.

Rozysknyje dela o Fedore Shaklovitom i jego soobshchnikakh, in 4 volumes (1893). Saint Petersburg: Arkheological Commission.

Russian History Library, volume II (1875). Saint Petersburg: Arkheological Comission, 351 p.

Scherrer, Y. (2021). Adaptation of morphosyntactic taggers: Cross-lectal and multilectal approaches. In M. Zampieri – P. Nakov (eds).: Similar languages, varieties, and dialects: A computational perspective. Studies in Natural Language Processing, Cambridge University Press, pages 138–166.

Shavrina, T., and Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. In Proceedings of the International Conference "CORPORA 2017", Saint-Petersbourg, Russia.

Shishkina, Y., and Lyashevskaya, O. (2021). Sculpting enhanced dependencies for Belarusian. In Revised Selected Papers of Analysis of Images, Social Networks and Texts: 10th International Conference (AIST 2021), pages 137–147. Tbilisi, Georgia.

Straka, M., and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014) Sequence to sequence learning with neural networks In Z. Ghahramani – M. Welling – C. Cortes – N. Lawrence – K. Q. Weinberger (eds.): Advances in neural information processing systems, vol. 27. Proceedings of NIPS 2014, pages 3104–3112, Montreal, Curran.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In Proceedings of the Section on Survey Research Methods, pages 354–359, Alexandria, VA. American Statistical Association.

Zaharova, K. F., and Orlova, V. G. (2004). Dialektnoe chlenenie russkogo yazyka. Moscow: URSS, 176 p.

sciendo
DOI 10.2478/jazcas-2023-0041

ANNOTATION OF ANALYTIC VERB FORMS IN CZECH – COMPLEX CASES

VLADIMÍR PETKEVIČ – HANA SKOUMALOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

PETKEVIČ, Vladimír – SKOUMALOVÁ, Hana: Annotation of Analytic Verb Forms in Czech – Complex Cases. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 234 – 243.

Abstract: The article deals with complex cases of determining the attribute verbtag, which contains the values of morphosyntactic categories of analytic verb forms. The latest corpora of contemporary written Czech from the SYN series are tagged with this attribute. In this paper, we focus on cases where it is difficult to identify values of verbtag categories. These include, e.g. the identification of the auxiliary verb *být* 'to be', recognition of the mood and tense of coordinated participles, or determining the number in compound forms in which the individual parts have a different morphological number. Some of the problems are of a theoretical nature, since it is not clear what the correct solution should be. Here we have arbitrarily opted for one option that was offered. Other problems are due to imperfections in the algorithms we use for annotation. The solution here is to improve these algorithms.

Keywords: analytic verb forms, verbtag, morphological and morphosyntactic annotation

1 INTRODUCTION

Prior to the creation of the SYN2020 corpus (Křen et al. 2020), only orthographic words were morphologically annotated in the SYN corpora of contemporary Czech, e.g. the form *rozhodnuto* 'decided' was marked as a passive participle, and there was no indication that this form was part of the passive *bylo rozhodnuto* 'was decided'. In addition, the tag for participles inappropriately indicated person (while the morphological tag is supposed to describe only the relevant form and participles do not express person). For some forms the tense was given, but related only to the isolated form (the past active participle thus had the past tense assigned in the tag, even though it was a part of the present conditional). For verbs the important morphological category of mood was not given. Information about tense, mood, and voice had to be elicited by the user through complex queries.

In order to be able to conveniently search for values of morphological categories expressed analytically, a new attribute was introduced: verbtag, which brings together all important grammatical categories of compound and simple verb forms and also expresses whether the verb form *být* 'to be' is auxiliary or main

verb. The attribute is assigned to all tokens in the corpus, but only for verbs and deverbal adjectives ending in *-ný/tý* (but not *-telný*) it takes on values other than "irrelevant". On the other hand, the original morphological tags are consistently stripped of the transferred categories, e.g. the person is no longer specified for participles.

All corpora of the SYN series since SYN2020, as well as ETALON2020, test and training corpus (Skoumalová 2021), are annotated with the verbtag attribute.

2 MORPHOLOGICAL CATEGORIES AND VALUES IN THE VERBTAG ATTRIBUTE

2.1 Introduction

Although in many languages of the world verb morphological categories are often expressed analytically (e.g. passive voice), in most corpora of these languages only the categories of isolated words in analytic verb complexes are tagged, rather than the properties of these complexes as such. Thus, for Czech, in which the categories of mood, voice, and tense are often expressed analytically, it is desirable to capture the properties of these complexes.

2.2 Morphological categories and their values in the verbtag attribute

In SYN2020 and subsequent corpora each word form is annotated with a 15-position tag and a 6-position verbtag (see Jelínek et al. 2021; Jelínek et al. 2022; Křivan et al. 2022 for details). While the tag expresses the values of grammatical categories of an isolated word form, the verbtag captures information about the grammatical categories of both simple (*napíšu* 'I will write') and compound (*budeme pracovat* '(we) will work', *nevěděli jsme* '(we) did not know') verb forms. It also distinguishes the auxiliary forms of the verb *být/bývat* 'to be/used to be' from the main verb forms including the copula.

The 6-position verbtag concentrates the values of the analytic verb complex at the main verb, i.e., it transfers the values of some grammatical categories from the forms of the auxiliary verb to the form of the main verb, exceptionally also from pronouns (if the person is expressed only by a personal pronoun). Modal and phrasal verbs are not marked, they are treated as main verbs. For example, in the sentence

(1) *Nikdy bys takový výsledek nebyla předpokládala.*
Never be.COND.2SG such result be.NEG.PSTP.SG.F expect.PSTP.SG.F
'You would never have expected such a result.'

we annotate the analytic verb complex *bys nebyla předpokládala* as past active conditional of the 2nd person singular. This information is specified in the verbtag of the main verb form *předpokládala* 'expect'; the forms *bys*, and *nebyla* are annotated as auxiliaries.

The following values are distinguished at the verbtag positions:
1st position – verb type: auxiliary verb (A) / main verb (V)
2nd position – verb form type: indicative (D) / conditional (C) / infinitive (F) / imperative (I) / transgressive (T) / freestanding passive participle, typically a verbal complement (O)
3rd position – voice: active (A) / passive (P) / deverbal adjective (p)
4th position – person: first (1) / second (2) / third (3)
5th position – number: singular (S) / plural (P) / polite form (v)
6th position – tense: pluperfect (Q) / past tense (R) / present (P) / present or future of biaspectual verbs (B) / future (F).

The advantage of the chosen concept of the verbtag attribute is that it is possible to find synonymous expressions of each tense, mood or voice with a single query, e.g. all forms of the past tense and past conditional can be found in the Corpus Query Language (CQL) by querying: [verbtag=".....R"].

Automatic annotation is performed by the hybrid LanGr+MorphoDiTa system in two phases, where LanGr is a rule-based component (Květoň 2006; Jelínek et al. 2011), and MorphoDiTa is a neural tagger (Straka et al. 2019). First, LanGr removes as many incorrect interpretations as possible, and then MorphoDiTa completes the disambiguation.

3 PROBLEMATIC ANNOTATION CASES

Automatic verbtag annotation has to deal with simple cases (the vast majority) and cases that are theoretically or practically more complex. The theoretical difficulties are mainly related to coordination, where one form of *být* 'be' connects a non-verbal (typically adjectival) complement coordinated with a passive participle, and thus it is not clear whether this form should be annotated as main verb (i.e., as a copula with non-verbal complements), or as auxiliary (with passive participles). Another problem is related to the coordination of passive participles: in this case, the second auxiliary form of the verb *být* is elided, but the passive participles do not agree in gender or number.

By practical problems we mean cases where the annotation fails, although it is clear what verbtag values should be assigned to the verb form.

3.1 Theoretically problematic cases

The form of the verb *být/bývat* has both a copulative and an auxiliary function

If the form of the verb *být/bývat* fulfils both a copulative and an auxiliary function, the annotation must cope with the ellipsis of the form of the verb *být/bývat* (generally, no tool for automatic ellipses detection was used for the annotation):

(2) *Lůžka s nastavitelnou výškou* *jsou*/VDA3PP//**wrongly:A-----** *polohovatelná*
Height-adjustable beds be.PRES.3PL positionable.ADJ.PL
a *elektricky* *ovládána*/VOP-P-//**wrongly:VDP3P**P
and electrically control.PASP.PL
'Height-adjustable beds are positionable and electrically controlled.'

(3) *... rostlinná společenstva* *budou*/A----- *mozaikovitě*
... plant communities be.FUT.3PL mosaically
uspořádána/VDP3PF *nebo vzájemně* *propletená*/--p---
arrange.PASP.PL or mutually intertwine.ADJ.PL
'... plant communities will be mosaically arranged or intertwined'

In sentence (2), the finite form of the verb *jsou* serves a dual function: (i) the copula in conjunction with the adjectival nominal predicate *jsou polohovatelná*, (ii) as an auxiliary verb in the passive construction *jsou ... ovládána* (due to the ellipsis of the second form of the verb *být*). As it is impossible to assign correct verbtags to all parts of the construction *jsou polohovatelná a ovládána*, we adopted the following rule: The form *jsou* is annotated as main verb, since the nominal part of the predicate (*polohovatelná*) is closer to it than the passive participle *ovládána*.

Differently – as an auxiliary verb – we annotate the form *budou* in sentence (3), where the passive participle *uspořádána* is closer to it than the deverbal adjective *propletená*. The general criterion is thus the mutual proximity of the respective forms.

The problem of the erroneous annotation in (2) will be solved by creating new, finer disambiguation rules for the LanGr system.

Determining the verb mood in the coordination of past participles

When coordinating past participles, it is often unclear what mood the second conjunct expresses: whether the construction is a coordination of participles expressing the same mood (typically a conditional), or whether the second participle already belongs to an indicative clause, especially if the second clause lacks a nominative subject:

(4) *... jako* *by*/A----- *také* *vystupovala*/VCA3SP
... as be.COND.3 also ascend.PSTP.SG.F
po Stromu světa *a* *byla*/VCA3SP//**wrongly:VDA3SR** *vysoko nad ním.*
on Tree of the World and be.PSTP.SG.F high above it.
'... as if she too ascended the Tree of the World and was high above it.'

The annotation error in (4) is to be eliminated by creating a new rule in the LanGr system that annotates the coordination, and thus also the agreement of participles in mood (conditional) and tense (present), if the subject of the second clause is omitted.

Person mismatch in coordination of past active participles

In structures with coordination of past active participles, it is often difficult to determine the person of the second participle if the subject is not lexically expressed in the second clause. Here only non-trivial semantic (i.e., difficult to algorithmize) reasoning can help. Compare sentences (5) and (6):

(5) *Mám jen okamžik na to, abych*/A----- *se zamyslela*/VCA1SP
I just have a moment for that so that be.COND.1SG REFL think.PSTP.SG.F
a hlavně se zhluboka nadechla/VCA1SP
and mainly REFL deeply inhale.PSTP.SG.F
'I just have a moment to think and, most importantly, take a deep breath.'

(6) *Na digitalis jsem*/A----- *samozřejmě myslel*/VDA1SR
On digitalis be.PRES.1SG of course think.PSTP.SG.M
a byl/VDA1SR *součástí toxikologického rozboru.*
and be.PSTP.SG.M part of toxicological analysis.
'Digitalis was obviously on my mind and was part of the toxicology analysis.'

While in sentence (5) both past participles are part of one clause, in sentence (6) the participles belong to different clauses that differ in person. By analysis of sentences in the corpus, we found that the problematic form is the copula *byl* in the position of the second conjunct. We can therefore set the rule that the coordinated participles of two main verbs always belong to one clause. No rules leading to 100% correct annotation can be formulated for the copula.

If the subject of the second clause is expressed, the situation is clear, and the automatic annotation is not flawed:

(7) *Když jsem byl malý, něco jsem*/A----- *napsal*/VDA1SR *a táta*
When I was small something be.PRES.1SG write.PSTP.SG.M and dad
vtipkoval/VDA3SR *a řekl*/VDA3SR *že to bylo trapné.*
joke.PSTP.SG.M and say.PSTP.SG.M that it was embarassing.
'When I was a kid, I wrote something and my dad joked and said it was embarrassing.'

In the clause *něco jsem napsal* 'I wrote something', the person of the participle (*napsal* 'wrote') is correctly determined as the first one; in the next clause, where the nominative subject (*táta* 'dad') is lexically expressed, the person of the participles (*vtipkoval* 'joked' and *řekl* 'said') is correctly determined as the third one.

Disagreement in the coordination of passive participles

In this type, two passive participles are connected by one form of the auxiliary *být*, the other form is elided, which makes it difficult to determine the values of the verbtag attribute. In the sentence:

(8) *V 16. a 17. století* *byl*/A----- *hrad silně* *opevněn*/VDP3SR
In the 16th and 17th centuries be.PSTP.**SG**.M castle heavily fortify.PASP.**SG**.M
a *vybudovány*/VDP3PR *další pevnostní objekty.*
and build.PASP.**PL**.M other fortress objects
'In the 16th and 17th centuries the castle was heavily fortified with other fortress structures built.'

The passive participles *opevněn* 'fortified' and *vybudovány* 'built' do not match in number because they belong to different clauses – the auxiliary form *byly* 'were' before the participle *vybudovány* is elided.

A similar mismatch occurs in constructions with the passive infinitive as in the sentence:

(9) *Použité zdroje* *musí* *být*/A----- *citovány*/VFP--- *a*
Used sources must.PRES.**SG/PL** be.INF cite.PASP.**PL.M** and
odkazováno/VFP---//**wrongly:VOP-S-** *na ně podle normy ISO 690.*
reference.PASP.**SG.N** to them according to norm ISO 690
'Used sources must be cited and referenced in accordance with ISO 690.'

Sentence (9) contains the coordination of two clauses with a passive infinitive, the second lacking the auxiliary form of the verb *být 'be'*, whose presence would have facilitated the annotation of the passive participle *odkazováno* 'referenced'. This participle does not agree with the first participle *citovány* 'cited' and is therefore incorrectly annotated as a complement (VOP-S-). However, the sentence is grammatical, albeit odd – the repetition of the verbs *musí být* 'must be' in the second clause (with changes in the word order) would help the clarity of the sentence.

The difficulties with reconstructing the ellipsis of the passive participle and the ellipsis of the auxiliary verb *být* are encountered in the sentence:

(10) *podrobný popis účelu,* *na který* *je*/A-----//**wrongly:VDA3SP** *půjčka*
detailed description of purpose for which be.PRES.**SG** loan.**SG**
nebo při možném souběhu *půjčky* *požadovány*/VDP3PP//**wrongly:VOP-P-**
or in possible concurrence loan.**PL** request.PASP.**PL**
'a detailed description of the purpose for which the loan or, in the case of possible overlapping loans, the loans are requested'

Here, the first clause lacks a singular passive participle (*požadována* 'requested') that would agree with the expressed auxiliary form *je* 'is' in number; the second clause, on the other hand, lacks an auxiliary form (*jsou* 'are') that would agree with the expressed participle *požadovány* 'requested'. The form *je* is incorrectly annotated as the main verb, the form of the participle *požadovány* is incorrectly annotated as a verbal complement, because the corresponding auxiliary form of the verb *být* (*jsou*) is not found.

The form of the auxiliary verb *být* in the future tense and the infinitive of the imperfective/biaspectual verb

In Czech, the future tense of imperfective or biaspectual verbs is formed analytically: by combining the auxiliary future form of the verb *být* and infinitive of an imperfective or biaspectual main verb, while perfective verbs express future tense simply by the present forms. The construction of analytic future tense is morphologically and syntactically ambiguous. In the sentence

(11) *Všiml si, že se chystá něco podotknout, ale pak se zřejmě rozhodla,*
He noticed that she is going to make a comment but then she perhaps decided
že nejchytřejší bude/VDA3SF/A----- *mlčet*/VFA---/VDA3SF
that smartest.NOM be.FUT.SG be-silent.INF
'He noticed that she was about to make a comment, but then she apparently decided that the smartest thing to do was to keep quiet.'
'He noticed that she was about to make a comment, but then she apparently decided that the smartest one would keep quiet.'

it is semantically very likely that the subject is the infinitive of the imperfective verb *mlčet* 'keep quiet', the form *bude* 'will be' is a copula, and the adjective *nejchytřejší* 'smartest' is a nominal part of the predicate; syntactically, however, the form *nejchytřejší* may have the function of a nominal subject and the construction *bude mlčet* 'will keep quiet' would have the function of the analytic future.

The disambiguation of these structures by linguistic rules can be improved by enlarging a group of adjectives that predominantly have the function of a nominal part of predicate in these constructions (e.g. *ideální* 'ideal', *lepší* 'better', *nejlepší* 'best', etc.).

Structures of addressing individuals politely

In Czech, there are analytic verb constructions expressing polite addressing of individual persons where there is a mismatch in number – the form of the auxiliary verb *být* is in plural and the form of the past or passive participle of the main verb is in singular. For example, in the sentence

(12) *Jak dobře jste*/A----- *ho znala*/VDA2**v**R
How well be.PRES.2**PL** him know.PSTP.**SG**.F
'How well did you know him?'

where the auxiliary form *jste* is in plural (tag=VB.**P**), the past participle *znala* in feminine singular (tag=VpF**S**), and the polite form is expressed by the value **v** in the verbtag.

Syntactically, sentence (12) is ambiguous, since the past participle *znala* 'knew' can also be the form of the neuter plural, which, however, agrees with the plural auxiliary *jste*.

Analytic forms where the number of the auxiliary verb does not match the number of the participle of the main verb

There are other types of constructions in Czech where the plural auxiliary form does not agree with the past active participle or the passive participle in singular and there is no polite addressing of individuals.

(13) *Poslouchal*/VDA2**v**R jste/A----- *někdo* *ráno rádio?*
Listen.PSTP.**SG**.M be.PRES.2**PL** somebody.**SG** morning radio
'Did anyone listen to the radio this morning?'

(14) ... *vždyť* *jsme*/A----- *každý* *vytížen*/VDP1**v**P *svými aktivitami.*
... after all be.PRES.1**PL** everyone.**SG** busy.PASP.**SG**.M own activities.
'I mean, we're all busy with our own activities.'

We annotate also these constructions as polite forms, but thanks to the mismatch between the subject and the auxiliary form in number, the annotation rules are not too complex. However, the pronominal subject (*někdo* 'someone', *každý* 'everyone') must be identified in the clause.

3.2 Practically problematic cases

Here we present cases where it is obvious how to annotate analytic forms, but automatic disambiguation fails.

Wrong POS disambiguation

Among the forms difficult to annotate there are three forms of the verb *být*:

- *si* – the non-standard form of *jsi* (2nd person singular present tense of the verb *být*) / reflexive particle
- *je* – 3rd person singular present tense of the verb *být*/personal pronoun in accusative singular or plural
- *buď* – 2nd person singular imperative of the verb *být*/conjunction 'either'.

An error in determining the POS will affect, of course, the selection of the correct verbtag. Disambiguation errors occur especially in more complex contexts where the parts of the analytic verb form are separated and there is syntactically complex linguistic material (punctuation, verbs, etc.) between them. The LanGr system therefore contains many sub-rules for disambiguating these forms.

Incorrect number and gender of participles

It often happens that the POS of ambiguous forms are disambiguated correctly, but their morphological properties are not. This happens, for example, in the case of systemic ambiguity in past and passive participles between feminine singular and neuter plural (*odešla* 'she/they.N left'). The sentence context is sometimes insufficient, as in

(15) *Bylo* *jich málo a spíše dávno,* *ale* *byla*/VDA3PR//**wrongly:VDA3SR**
be.PSTP.SG them few and rather long-ago but be.PSTP.PL.N//**wrongly:PSTP.SG.F**
'They were few and rather ancient, but they were.'

where the form *byla* 'were' is incorrectly identified as the feminine singular form.

If there is no subject in the clause, the disambiguation rules and MorphoDiTa fail, as they only use the context of the given sentence. There is no satisfactory solution here.

Misidentification of the auxiliary vs. main verb *být* in structurally complex sentences

It may happen that the parts of analytic verb forms are distant from one another and the sequence of words between them is complex. For example, in the sentence

(16) *Primátor Richard Svoboda byl*/A-----//**wrongly:VDA3SR** *sdělením,*
Mayor Richard Svoboda be.PSTP.SG.M by-announcement
že jméno jeho podřízeného figuruje v registru svazků, *nemile*
that name of his subordinate appears in the registers unpleasantly
překvapen/VDP3SR//**wrongly:VOP-S-**
surprised
'Mayor Richard Svoboda was unpleasantly surprised by the announcement that his subordinate's name appeared in the registers.'

The disambiguation system does not see the pair *byl ... překvapen* 'was … surprised' as a unit, since the two parts are separated by the nested clause *že ... svazků*. Proper annotation can be assured by creating additional LanGr rules.

4 CONCLUSION

Although we have shown a number of theoretical and practical problems associated with the annotation of analytic verb forms in this paper, the overall annotation accuracy (measured on the ETALON2020 corpus) is quite high. For verbtags computed over the whole corpus, it is 99.77%; if we compute the accuracy only on verbs, it is 98.57% (Jelínek et al. 2022).

In the future, in the interest of improved annotation, we will focus on

- more linguistically adequate solutions to the theoretical problems,
- further improving the annotation (disambiguation of the listed complex cases),
- identifying analytic verb forms as such, i.e., marking which components of an analytic verb form make the relevant unit (the software annotation system does not yet allow this).

ACKNOWLEDGEMENTS

The research was supported by the grants LM2023044 and GA-23-05240S.

References

Jelínek, T., and Petkevič, V. (2011). Systém jazykového značkování současné psané češtiny. In Korpusová lingvistika Praha 2011, sv. 3. Gramatika a značkování korpusů. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu, pages 154–170.

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., and Šindlerová, J. (2021). SYN2020: A New Corpus of Czech With an Innovated Annotation. In K. Ekštein – F. Párt – M. Konopík (eds.): Proceedings of the Text, Speech and Dialogue 24th International conference TSD 2021. Olomouc, Czech Republic, September 6–9, 2021. LNAI 12848. Springer Nature Switzerland AG 2021, pages 48–59. Accessible at: https://doi.org/10.1007/978-3-030-83527-9.

Jelínek, T., Petkevič, V., and Skoumalová, H. (2022). Hodnoty slovesných morfologických kategorií v korpusu SYN2020 – atribut verbtag. Časopis pro moderní filologii 104(1), pages 89–109.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2020). SYN2020: reprezentativní korpus psané češtiny. Praha: Ústav Českého národního korpusu FF UK. Accessible at: https://www.korpus.cz.

Křivan, J., and Šindlerová, J. (2022). Změny v morfologické anotaci korpusů řady SYN: nové možnosti zkoumání české gramatiky a lexikonu. Slovo a slovesnost, 83(2), pages 122–145.

Květoň, P. (2006). Rule-based Morphological Disambiguation. Ph.D. thesis. Praha: MFF UK.

Skoumalová, H. (2021). Etalon: manuálně anotovaný synchronní korpus českých textů. Praha: Ústav Českého národního korpusu FF UK. Accessible at: https://www.korpus.cz and http://hdl.handle.net/11234/1-3698.

Straka, M., Straková, J., and Hajič, J. (2019). Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In International Conference on Text, Speech, and Dialogue, Ljubljana: Springer, pages 137–150.

sciendo
DOI 10.2478/jazcas-2023-0042

CAPEKDRACOR: A NEW CONTRIBUTION TO THE EUROPEAN PROGRAMMABLE DRAMA CORPORA

PETR POŘÍZKA
Department of Czech Studies, Faculty of Arts, Palacký University,
Olomouc, Czech Republic

POŘÍZKA, Petr: CapekDraCor: A New Contribution to the European Programmable Drama Corpora. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 244 – 253.

Abstract: The aim of this paper is to present the new *CapekDraCor* corpus and the *DraCor* project with its research-oriented concept of a programmable corpora focused on quantitative analyses within the framework of computational literary studies. This digital platform extends the possibilities of large-scale drama analysis with a focus on the dramatic character(s). The basic operationalisation is the interaction within a dramatic configuration, i.e., the scenic co-presence of two speakers, from which network data are automatically extracted, both global networks of interactions of dramas and data characterising individual actors, i.e., literary characters.

The paper demonstrates the *CapekDraCor* corpus, a new contribution to the extensive *DraCor* database, and presents the way the data are processed with respect to their specific multi-layered structure. The corpus contains all the plays written by Karel and Josef Čapek and the data are processed in a standardized format based on XML and general TEI guidelines for processing drama with a defined basic drama tagset. *CapekDraCor* also uses the newly created EZdrama format for data processing, which works as an intermediate step from .txt to .xml file as a lightweight YAML-like markup language. A file in this format can be automatically converted into a DraCor-ready XML file with a TEI header.

The advantage of the programmable corpora concept is the possibility to use suitably structured data for drama research outside the *DraCor* platform and with other methods or tools for textual analysis. Simultaneously, this approach moves the researcher from the technical requirements of the analysis to operationalised computational analysis based on research questions and pre-prepared and flexible tools. *DraCor* is a unique open infrastructure (both in terms of data and tools) for the analysis of European drama, currently comprising 15 corpora in 10 different languages with a total of about 3,000 plays from a wide range of periods.

Keywords: data annotation, computational literary studies, corpus building, drama, DraCor, network analysis, quantitative analysis

1 INTRODUCTION AND STATE OF THE ART

The Čapek brothers and their plays represent a unique literary, linguistic and socio-cultural phenomenon, because of the interplay between literary fiction and strong moral, political and social commentary in their plays. Karel Čapek is a world-famous writer and his theatrical work undoubtedly represents one of the

most important parts of his entire oeuvre. His brother Josef, who co-wrote three plays and was himself the author of two other dramas, also contributed significantly to this work. In total, the dramatic work of the Čapek brothers comprises ten plays and has been discussed many times in the scholarly literature and analysed from various perspectives: theatrical, literary and linguistic. Philological analyses have thus far relied mainly on the introspection of the researcher and rather traditional, structuralist interpretative approaches without the use of computer text analysis. To this day, the plays by Josef have remained largely unexplored by researchers, including linguists and literary scholars with respect to quantitative analyses. The *CapekDraCor* corpus, which we present here, aims to change this situation and enable researchers to analyse these texts using modern methods of corpus and quantitative linguistics within the framework of Computational Literary Studies. Although the *Czech National Corpus* (CNC) already has a corpus of Karel Čapek's plays (Čermák et al. 2007) called *capek*, its processing is limited because it does not reflect their multilayered structure (the textual and metatextual parts are not separated), which can (and does) affect subsequent quantitative analyses and their results. See Fig. 1 below, where both the comments and the characters' labels preceding the dialogues, or even the initial cast list, are represented as text in the CNC *capek* corpus (Fig. 1):

Fig. 1. The Czech National Corpus and Karel Čapek's dramas: text and meta-text are mixed and not differentiated in the CNC *capek* corpus.

Thus, the use of modern and exact methods of computer text analysis requires the re-processing of Čapek's plays into a new annotated corpus that will reflect the specifics of the plays, including their multi-layered structure, which will enable the analysis not only of the text as a whole, but also of its sub-parts (by separating the dialogue layer from the metatextual parts). In addition, other structural elements of the text should be possible to use as a unit or parameter of analysis, and the literary

characters of the individual plays are especially the key in this respect. These possibilities are offered by the *DraCor* project focusing on European drama, which seeks to apply computational methods to digitized literary texts, in line with a similar COST Action on European novels (Schöch et al. 2018).

2 DRACOR AS A DIGITAL PLATFORM FOR THE DRAMAS

DraCor is an open platform as well as a growing network of textual resources for hosting, accessing, and analysing European dramas, which also includes tools designed for data mining. It currently includes 15 corpora in 10 different languages with a total of approximately 3,000 plays of a wide temporal range.

The *DraCor* infrastructure is based on the concept of "Programmable Corpora", a new term proposed in Fischer et al. (2019) for research-oriented corpora providing APIs, which is an important methodological concept that moves the researcher from the technical requirements of analysis to an operationalised computational analysis that is based on research questions and pre-prepared and flexible tools that enable such analysis. I quote: "Projects like DraCor seek to provide the digital literary studies with a reliable and extensible infrastructure so that the research community can focus on research questions. (…) Programmable Corpora facilitate the implementation of research questions around corpora. (…) and make it easier to decide on which level of the platform your own research process starts." (cf. Fischer et al. 2019). Another important conceptual term related to the DraCor base is vanilla corpora: "Similar to Paul Fièvre's collection 'Théâtre classique', these corpora are designed as vanilla corpora, which initially contain hardly any special markup beyond the necessary, but are freely available and can therefore be forked, enriched and expanded" (ibidem).[1] The necessary annotation therefore concerns specific aspects of dramatic texts.

2.1 Technical specification

DraCor data are processed in a standardised format based on XML and general TEI requirements or guidelines (P5 model) for drama processing with a defined bas ic drama tagset (TEI Performance Texts 2023), bu t texts are also available in other (different) formats to suit the needs of individual tools and analyses (CSV, JSON, GEXF, GraphML, RDF).

Additional corpora can be easily linked to the *DraCor* infrastructure, and all existing data mining and visualisation methods of the platform are easily applicable to newly added corpora. The only prerequisite is that they are encoded in TEI/XML (see

[1] This does not mean, however, that it is impossible to additionally annotate the texts linguistically if necessary: the Shakespearian corpus *ShakeDraCor* is even linguistically annotated (lemma and tag by Morphadorner and NUPOS English Tagset).

above) and freely available as a corpus hosted on GitHub that can be cloned and loaded directly into a structured database. *DraCor* relies on eXist-db as XML database to process TEI files and provide functions for researching the corpora. The frontend is built using *React* (https://reactjs.org/), which is responsive and easily extensible, with the focus on APIs. Other tools in the system are *SPARQL* (an application for eXist-db), *ezlinavis* (a didactic tool), *ShinyDraCor* (an R-based framework that allows interactive visuali s ations to be displayed in the browser, relying entirely on the *DraCor* API for data retrieval), and most recently *rdracor* (a new R interface for the *DraCor* API) (see https://dracor.org/ in the Tools section for more information).

3 CAPEKDRACOR CORPUS AND DATA PROCESSING

The *CapekDraCor* corpus as a new contribution to the *DraCor* corpora is composed of the plays written by Karel and Josef Čapek, namely (1) *Loupežník* (‘The Robber ’, 1920); (2) *R.U.R.* (1920); (3) *Věc Makropulos* (‘The Makropulos Affair’, 1922); (4) *Bílá nemoc* (‘The White Disease’, 1937); (5) *Matka* (‘The Mother’, 1938); (6) *Lásky hra osudná* (‘The Fateful Game of Love’, 1910); (7) *Ze života hmyzu* (‘The Insect Play’, 1921); (8) *Adam Stvořitel* (‘Adam the Creator’, 1927); (9) *Země mnoha jmen* (‘The Land of Many Names’, 1923); (10) *Gassirova loutna* (‘Gassir’s Lute’, 1923). As far as the authorship of the individual plays is concerned, Karel wrote plays Nos. 1–5, together with his brother Josef they wrote plays Nos. 6–8, and Josef is the author of plays Nos. 9–10. Critical editions of the plays of the Čapek brothers from the following editions were used as a textual basis:[2]

- Karel Čapek: *Spisy Karla Čapka: Dramata*, Československý spisovatel 1992/94 (editor Emanuel Macek); dramas: *Loupežník*, *R.U.R.*, *Věc Makropulos*, *Bílá nemoc*, *Matka;* with respect to the edition Česká knižnice, Brno: Host 2014 (editor Jiří Holý); dramas: *Loupežník*, *R.U.R.*, *Bílá nemoc*
- Karel Čapek – Josef Čapek: *Ze společné tvorby*, Praha: Československý spisovatel 1982 (editor Emanuel Macek); dramas: *Lásky hra osudná*, *Ze života hmyzu*, *Adam Stvořitel*
- Josef Čapek: *Beletrie 1*, Brno: Triáda 2011 (editor Daniel Vojtěch); dramas: *Země mnoha jmen*, *Gassirova loutna*

These plays were processed into standardised, commonly used and extended formats (.txt, .xml) and are thus not limited to use in the *DraCor* database, which will increase their subsequent usability by other researchers and other software tools (e.g. R and packages like *stylo*, *quadrama* etc.). The following categories of metadata were encoded (and can subsequently be used for analyses):

[2] All texts are freely available and copyright is not applicable (copyright protection expires 70 years from the author’s death). Other texts for the purposes of this project (with subsequent publication of the results: corpora and papers) were kindly provided free of charge by the publishing houses Host (with the consent of the editor J. Holý) and Triáda (also with the consent of the editor D. Vojtěch).

- external and textual: *author*, *title*, *year (written, printed, premiered)*, *language*, *genre*, *edition; cast list*;
- internal: text structure (*prologue/act/scene/epilogue*); layer distinction (*characters/dialogues/commentary/stage directions*); literary characters (*cast group* and *role* ; *gender/sex*).

The texts have not yet been annotated linguistically (lemma, part-of-speech category, morphological tags), but this variant is being considered for processing for the Czech National Corpus.

The final and most important format is undoubtedly the TEI/XML-encoded version of texts. The processing is carried out, however, through an intermediate stage – conversion to *EZdrama* format, which makes it easier and faster to create the required XML file. See the diagram below (Fig. 2):

Fig. 2. Individual stages of text processing for the *DraCor* database

EzDrama format is a lightweight Markdown-like (or YAML-like) markup language and a file in this format can be created either manually or semi-automatically through search–replace conversions using regular expressions. This phase is followed by automatic conversion to DraCor ready XML file with TEI header.

There will now be a demonstration of what the text in *EZdrama* format looks like before conversion to XML (syntax and example below). Several metasymbols and line breaks are used to encode the text, namely: [3]

means a structural part like an *act* (#) or *scene* (##); technically the number of nesting levels is not limited

$ means a new *stage direction* or *commentary* ; brackets () are also used and automatically converted to stage directions or comments – they do not require any special editing

@ means the text line contains the name (label) of the character preceding the dialogue; note: any other line without these special symbols will be treated as the direct speech of the last encoded speaker (@)

~ means that this and the following unmarked lines are in verse (poet ry)

Let us look at a short fragment of dialogue between Mimi and Fanka from the play *Loupežník* (The Robber), at the beginning of the second act (DĚJSTVÍ DRUHÉ), prelude (PŘEDEHRA) – Fig. 3:

[3] An overview of all metasymbols and more information about the EZdrama format can be found on the website https://github.com/dracor-org/ezdrama/blob/main/README.md.

```
#DĚJSTVÍ DRUHÉ
##PŘEDEHRA
$Loupežník se zavázanou hlavou a Doktor vycházejí z hospody U Beránka, sednou
si venku u stolu a mastí karty.
@MIMI (vyjde z vrat s nějakým šitím):
Proč nikoho neposílá? Co se s ním děje? Bože, slituj se nade mnou!
@FANKA (vyjde z vrat):
Nikam nechodila, Mimi!
@MIMI:
Já nikam nejdu. (Sedne si na lavičku.) Ach, kdyby už poslal zprávu, jak mu je!
```

Fig. 3. EZdrama format: lines are marked with special symbols at the beginning; an excerpt from the play *Loupežník* (The Robber)

Specific parts marked with special symbols are colour-coded for better readability (comments or stage directions in green, character names in blue, acts/scenes in red). The texts are automatically converted from this EZdrama format to the DraCor-ready XML file – cf. Fig. 4:

```
<div type="act">
  <head>DĚJSTVÍ DRUHÉ</head>
  <div type="scene">
    <head>PŘEDEHRA</head>
    <stage>Loupežník se zavázanou hlavou a Doktor vycházejí z hospody U
    Beránka, sednou si venku u stolu a mastí karty.</stage>
    <sp who="#mimi">
      <speaker>MIMI:</speaker>
      <stage>(vyjde z vrat s nějakým šitím)</stage>
      <p>Proč nikoho neposílá? Co se s ním děje? Bože, slituj se nade
      mnou!</p>
    </sp>
    <sp who="#fanka">
      <speaker>FANKA:</speaker>
      <stage>(vyjde z vrat)</stage>
      <p>Nikam nechodila, Mimi!</p>
    </sp>
    <sp who="#mimi">
      <speaker>MIMI:</speaker>
      <p>Já nikam nejdu. <stage>(Sedne si na lavičku.)</stage> Ach,
      kdyby už poslal zprávu, jak mu je!</p>
    </sp>
```

Fig. 4. DraCor-ready XML file after the automatic conversion from EZdrama format; an excerpt from the play *Loupežník* (The Robber)

The TEI header contains the necessary metatextual information, an overview of literary characters and their characteristics (cast list and role description), plus other necessary meta-information (see above).[4] The predominant way of processing plays in

[4] For more information about the TEI header see the website: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html, or https://teibyexample.org/tutorials/TBED02v00.htm.

the *DraCor* subcorpora (TEI/XML) (as shown in Fig. 4) is based on adequate segmentation of the dramas mainly through XML elements into different layers of text/meta-text without any additional linguistic annotation, or through XML attributes used to encode other aspects. The distinction is mainly related to spoken dialogues (<p> element for prose and <l> element for poetry) versus comments and stage directions (<stage> element); <sp> i s the element to mark up an individual speech in a dramatic text (the actual utterances) and <speaker> elements are used to mark up dramatic characters (producing the speech); element and *type* attribute tags with values *act*, *scene* (and others) are used to indicate the text structure. [5]

4 DIGITAL NETWORK ANALYSIS OF THE DRAMAS

The *DraCor* platform approach aims to extend the possibilities for large-scale drama analysis with a focus on the dramatic character(s). The literary figure is understood here as the central element of narrative (Todorov 1977; Fořt 2008). A network-analytic conceptualisation of dramatic interaction is used as the theoretical basis (Moretti 2011): the basic operationalisation is the *interaction* within a dramatic configuration, which is the scenic co-presence of two speakers (Trilcke 2015). Based on this notion of relationships, network data are automatically extracted, both global networks of ‚interactions' of the dramas (density, average degree, connectivity, etc.) and data characterising individual actors, i.e., literary figures (degree and various other centrality indices). The following picture from the *CapekDraCor* corpus illustrates the network of relationships between the characters in Karel Čapek's drama R.U.R. (Fig. 5):

As mentioned above, the role of the character in relation to other participants is of key importance regarding the interpretation of narrative. It is a co-presence network that is based only on the co-occurrence (or co-appearance) of the characters in each section (in this case, scene). The size of the bubble is the weighted degree of the node (i.e., the figure), which is the sum of the edge weights for the edges incident to that node.

*DraCor*system allows such networks to be generated based on various criteria – all characters, selected characters (e.g. male or female, only 'main'figures, etc.) and offers various metrics that can be used for further analysis. A major asset is the already quite robust database of plays by various European playwrights with a large time span, which can again be used for comparisons and contrastive analyses (via *DraCor*metrics). For illustrative examples of such comparative analyses see e.g. Skorinkin et al. (2018).

[5] For a more detailed description of the XML format related to dramas cf. https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html, https://teibyexample.org/tutorials/TBED05v00.htm2, or Base Tagset for Drama: https://tei-c.org/Vault/P4/doc/html/DR.html.

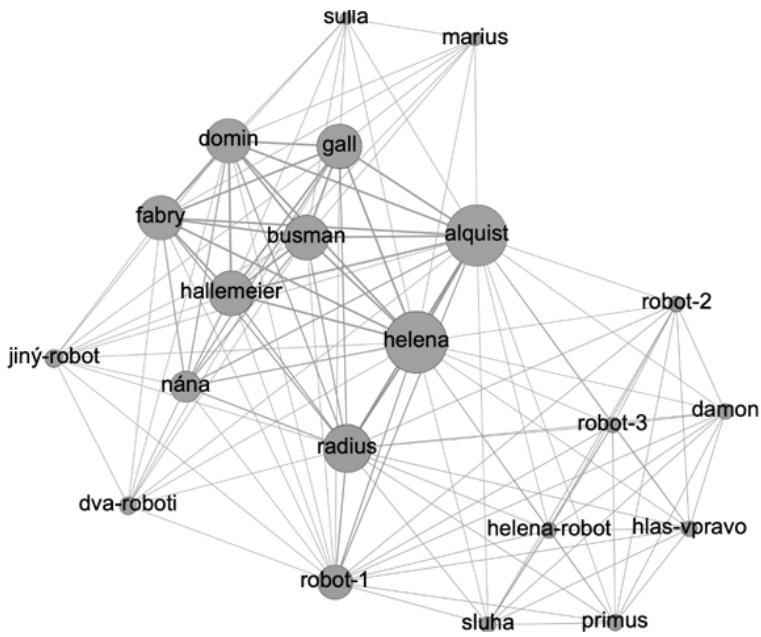

Fig. 5. *CapekDraCor*: co-occurrence network of characters from Karel Čapek's R.U.R.

5 CONCLUSIONS AND POSSIBILITIES OF USING PROGRAMMABLE CORPORA

Case studies (or examples) can be found in the collection of research papers on drama analysis using data provided through the *DraCor*database and using the *Dracor*tools on the project website.[6]

The *DraCor*infrastructure provides a unique, data-rich and extensive base not only for the analysis of the individual plays or authors, but also for comparative or contrastive analyses, for which other European playwrights and their works can be used (see the relevant *DraCor*subcorpora). It is thus possible to trace the development of selected aspects of European drama over time (plays from the 16th to 20th centuries) as well as other narratological aspects of different European playwrights, the complexity of dramas (i.e., the networks' complexity of relationships between characters) or sub-aspects such as the development of commentary sections, the ratio and characteristic of male and female figures, the role and contribution of characters to the narrative of the text, etc. An indisputable advantage with great research potential is the overall concept of programmable corpora and the possibility to use suitably structured data (open access source) for drama research outside the

[6] See the website: https://dracor.org/doc/research.

*DraCor*platform and with other methods or tools for textual analysis. The most illustrative example is the *QuaDramA*case study, which uses the *GerDraCor*corpus for illustrative drama analysisand demonstrates the flexibility and versatility of the programmable corpora approach (Reiter – Pagel 2022).

In quantitative research, analyses include both descriptive statistics and an exploratory quantitative approach (inferential statistics), in this case, for example, based on network analysis of dramas. The *CapekDraCor*corpus, as well as the entire *DraCor*infrastructure that we wanted to present in this text, are designed to meet these needs and objectives.

ACKNOWLEDGEMENTS

The research has been supported by the Program of the European Commission funded project CLS INFRA (https://clsinfra.io/) from the European Union's Horizon 2020 program (grant agreement No. 101004984).

References

Čapek, J. (2011). Beletrie 1. Praha: Triáda.

Čapek, K., and Čapek, J. (1982). Ze společné tvorby. Praha: Československý spisovatel.

Čapek, K. (1992/94). Spisy Karla Čapka: Dramata. Praha: Československý spisovatel.

Čapek, K. (2014). Tři hry (Loupežník, RUR, Bílá nemoc). Brno: Host.

Čermák, F. et al. (2007). Capek: korpus vlastních textů Karla Čapka. Praha: Ústav Českého národního korpusu. Accessible at: http://www.korpus.cz.

DraCor. Drama Corpora Project [database]. Accessible at: https://dracor.org/.

Reiter, N., and Pagel, J. (2022). Analysis of Dramatic Texts. Package DramaAnalysis. Accessible at: https://quadrama.github.io/DramaAnalysis/tutorial/3/index.html.

Fischer, F. et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University.

Fořt, B. (2008). Literární postava. Vývoj a aspekty naratologických zkoumání. Praha.

Herman, D. (2002). Story Logic. Problems and Possibilities of Narrative. University of Nebraska Press.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1, pages 127–165.

Moretti, F. (2011). Network Theory, Plot Analysis. Stanford Literary Lab Pamphlets 2. Accessible at: http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

QuaDramA. Quantitative Drama Analytics [Software]. Accessible at: https://github.com/quadrama.

React. The Library for Web and Native User Interfaces. Accessible at: https://reactjs.org/.

The Shakespearian corpus ShakeDraCor. Accessible at: https://dracor.org/shake.

Schöch, Ch. et al. (2018). Distant Reading for European Literary History. A COST Action [Poster]. In DH2018: Book of Abstracts / Libro de resúmenes. Mexico: Red de Humanidades Digitales A. C. Accessible at: https://dh2018.adho.org/en/?p=11345.

Skorinkin, D., Fischer, F., and Palchikov, G. (2018). Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies. In

Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference »Dialogue 2018«, pages 662–682, Moscow.

stylo: Stylometric Multivariate Analyses (version 0.7.4). [Software]. Accessible at: https://cran.r-project.org/web/packages/stylo/index.html.

TEI Performance Texts (2023). In TEI P5 – Guidelines for Electronic Text Encoding and Interchange (version 4.6.0). Accessible at: https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html.

Textometrie project: TXM (version 0.8.1) [Software]. Accessible at: http://textometrie.enslyon.fr.

The R Project for Statistical Computing: R (version 4.0.4) [Software]. Accessible at: https://www.r-project.org/.

Todorov, T. (1977). The Poetics of Prose. Ithaca – New York: Cornell university Press.

Trilcke, P., Fischer, F., and Kampkaspar, D. (2015). Digital Network Analysis of Dramatic Texts. In DH2015: »Global Digital Humanities«. University of Western Sydney.

sciendo
DOI 10.2478/jazcas-2023-0043

THE *INTERCORP* PARALLEL CORPUS WITH A UNIFORM ANNOTATION FOR ALL LANGUAGES

ALEXANDR ROSEN
Institute of the Czech National Corpus, Faculty of Arts,
Charles University, Prague, Czech Republic

ROSEN, Alexandr: The InterCorp Parallel Corpus with a Uniform Annotation for All Languages. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 254 – 265.

Abstract: Recently, the language-specific morphosyntactic annotation of *InterCorp*, a large multilingual parallel corpus, has been replaced by the language-uniform morphosyntactic and syntactic annotation following the guidelines of the Universal Dependencies project. Because the corpus is used predominantly by human users via a token-based concordancer, the CONLL-U format produced by the *UDPipe* parser has been extended by attributes such as lemma of the token's syntactic head or morphosyntactic categories of the content verb's auxiliary. We conclude that despite some theoretical and practical issues, the new annotation is a promising solution to the issue of mutually incompatible tagsets within a single corpus.

Keywords: parallel corpus, Universal Dependencies, multilinguality, syntactic annotation, language-universal categories

1 INTRODUCTION

InterCorp[1] is a multilingual parallel corpus centred around Czech (Čermák – Rosen 2012; Rosen 2016). Since 2009 the corpus has been available with language-specific morphosyntactic annotation. In a recent release of *InterCorp* (v.13ud), a uniform annotation scheme was introduced, based on the guidelines of the *Universal Dependencies* project[2] (UD, de Marneffe et al. 2021). Texts in the corpus were tagged and parsed by *UDPipe*,[3] one of the UD tools (Straka 2018): in addition to morphosyntactic categories, the annotation offers also syntactic functions and structure.

Most users access *InterCorp* via *KonText*,[4] the standard token-based search interface of the Czech National Corpus.[5] To make the UD annotation more useful in *KonText*, the

[1] https://wiki.korpus.cz/doku.php/en:cnk:intercorp
[2] https://universaldependencies.org/
[3] https://ufal.mff.cuni.cz/udpipe/2
[4] https://www.korpus.cz/kontext/
[5] https://www.korpus.cz/

output of *UDPipe* (in the CONLL-U format)[6] has been extended by other attributes. Such attributes help to query the corpus and to generate statistics more efficiently.

After presenting *InterCorp* and its language-specific linguistic annotation in Section 2, we describe the specifics of the UD annotation as it was modified and extended for the use in *KonText* in Section 3. In Section 4 we explain how the annotation can be used in corpus queries. Finally, we point out some issues, sketch plans and conclude in Section 5.

2 INTERCORP

2.1 The contents

InterCorp consists of "Core", including mainly fiction, and "Collections": political commentaries by *Project Syndicate*[7] and *VoxEurop*,[8] legal texts from the *Acquis Communautaire* corpus,[9] proceedings of the European Parliament from the *Europarl* corpus,[10] film subtitles from the *Open Subtitles* database[11] and translations of the Bible, all equipped with bibliographical data.

Like the *Universal Dependencies* project, *InterCorp* is the result of collaborative efforts: academic staff and students of linguistic departments, mostly from Charles University's Faculty of Arts, are responsible for the choice of texts in the Core and much of their processing. Texts in the Core are also proofread for typos, sentence segmentation and correct alignment. Collections are not checked that carefully, and – moreover – they do not retain all texts from the original source, especially those that do not have any Czech counterpart.

For every text there is a single Czech version (original or translation), aligned with one or more foreign-language versions. In release 13, the Core numbers 328 mil. words in foreign texts and 114 mil. in Czech texts. For Collections, it is 1,223 mil. words (foreign) and 90 mil. (Czech).[12] Many texts are available only in few languages, which means that the corpus is far from balanced. Fig. 1 shows the share of text types for each language. *InterCorp* v.13ud contains the same texts as *InterCorp* v.13, the difference is in the UD-based linguistic annotation. Out of a total of 41 languages, 36 are annotated by *UDPipe.*[13]

[6] https://universaldependencies.org/format.html

[7] http://www.project-syndicate.org/

[8] http://www.voxeurop.eu/

[9] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[10] http://www.statmt.org/europarl/

[11] http://www.opensubtitles.org/

[12] At the time of writing, *InterCorp* v.16 is close to release. It is planned as the last version with language-specific tags, to be followed by UD-annotated versions.

[13] The remaining 5 languages were not supported by *UDPipe* at the time v.13ud was getting ready for release. In this sense, the title (… FOR ALL LANGUAGES) is a misnomer, but we hope to remedy this in a future release.

2.2 Language-specific linguistic annotation

For linguistic annotation of a large multilingual corpus an opportunistic strategy of adopting available tools and annotation schemes is the only realistic option. Before UD had reached a practically useful stage, the only reasonable solutions were language specific. Tab.1 shows a sample result. The notational diversity may obscure the fact that many of the seemingly corresponding labels have mismatching denotations, due to different assumptions behind the designs of the tagsets rather than to cross-lingual differences. For example, the English tag IN is used both for prepositions and subordinating conjunctions, while all the other languages make this distinction. Or the Czech equivalent of a determiner is tagged as demonstrative pronoun (PD), undistinguished between attributive and substantive use, whereas its Polish counterpart is tagged as an adjective because it exhibits adjectival declension.

Incompatible tagsets can be mapped onto an interlingual taxonomy, as proposed e.g. by Rosen (2014), but once a uniform annotation scheme and the corresponding tools are available, a more viable solution is to apply language-universal annotation to all texts from scratch.

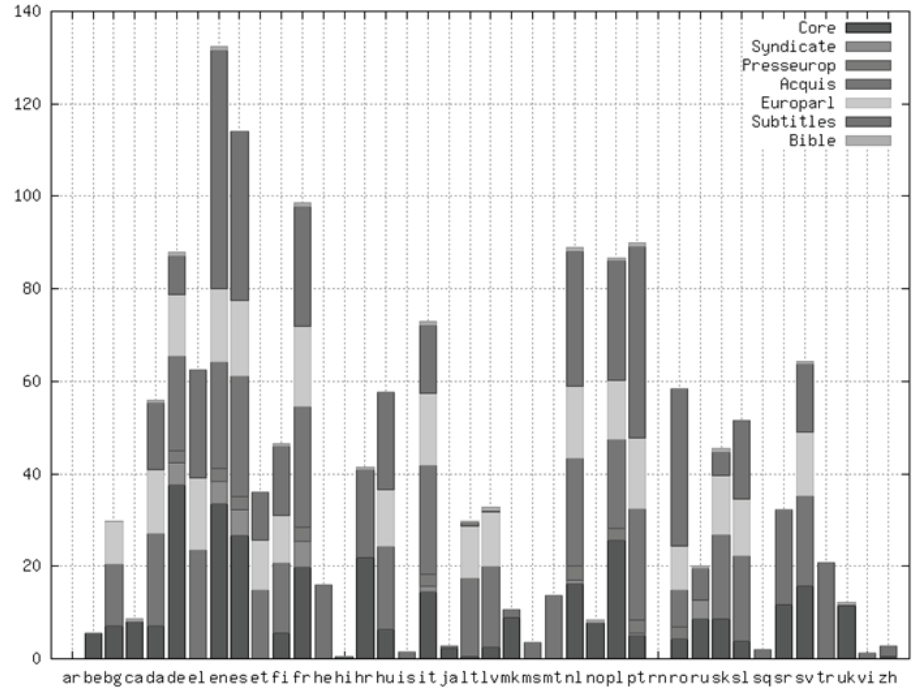

Fig. 1. *InterCorp v13*: languages and text types in mil. words[14]

[14] Language codes are given in ISO 639-1 except for Romany, represented as rn.

Lang	Preposition	Determiner	Adjective	Noun
bg	R	Pde-os-n	Ansi	Ncnsi
cs	RR-6	PDXP6	AAFP6---3A	NNFP6---A
de	APPR	ART	ADJA	NN
en	IN	DT	JJS	NNS
es	PREP	ART	NC	ADJ
et	P--s3		A-p-s3	Nc-s3
fr	PRP	DET:ART	ADJ	NOM
it	PRE	PRO:demo	NOM	ADJ
no	prep	det	adj	subst
pl	prep:loc:nwok	adj:sg:loc:m3:pos	adj:sg:loc:m3:pos	subst:sg:loc:m3
pt	SPS	DA0	NCFS	AQ0
ru	Sp-l	P--pl	Afp-plf	Ncmpln
sk	Eu6	PFip6	AAip6z	SSip6
sl	Sl	Pd-nsg	Agpfsg	Ncnsl

Tab. 1. A phrase such as *in the remotest suburbs*, annotated by language-specific tags[15]

2.3 *Universal Dependencies* as the obvious choice for consistent linguistic annotation

Language-universal annotation moves a part of the burden of enforcing consistency from the corpus onto the annotation scheme. As a result, the scheme and its application face the linguistic diversity of multilingual data.

Language-universal annotation assumes a single set of categories for all languages. It seems that this is what UD does. However, de Marneffe et al. (2021) and the Introduction to the UD Guidelines[16] describe the design of UD as a "very subtle compromise" between six competing criteria, only one of which concerns cross-lingual universals: "UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families." Yet, according to the guidelines, each language is described by the single universal set of categories. Only when they are too coarse, an additional language-specific distinction can be used. This is at odds at with the mainstream typological literature (Croft et al. 2017; Haspelmath 2007; Haspelmath 2010). The compromise is therefore less subtle than presented. However, given the practical goals of the project, the sacrifice of some typological neutrality seems to be justified.

As a more down-to-earth issue, some UD treebanks do not fully comply with the guidelines, either due to specific (mis)interpretation of the guidelines or to failures in applying the guidelines consistently during annotation. The tools trained on the treebanks also rely on the size of the treebanks, their domains and annotation

[15] Tags such as PDXP6 are truncated for the sake of brevity.

[16] https://universaldependencies.org/introduction.html

quality. This is the main reason why some results may be disappointing when compared with those of language-specific tools, often trained on larger data without syntactic annotation.[17] However, both the UD treebanks and the tools have seen significant improvements since the first UD-annotated release of *InterCorp*. An experiment with recent versions of *UDPipe* and Czech treebanks promises a rise in success rates by almost 10 points.[18]

The following section refer to the annotation guidelines on the UD project website (UD Guidelines[19]), including a detailed description of word types (Universal POS tags[20]), morphological categories (Universal features[21]) and syntactic functions (Universal Dependency Relations[22]).

3 THE HAPPY RELATIONSHIP BETWEEN *INTERCORP* AND *UNIVERSAL DEPENDENCIES*

3.1 Annotation adopted from the UD scheme:

- The UD word class and UD morphological categories are listed separately as values of the `upos`[23] and `feats`[24] attributes. Language-specific tags are values of the `xpos` attribute.
- Each word is assigned its syntactic function (`deprel` – see Section 3.5) and a pointer to its syntactic head in the dependency tree (`head`).

3.2 Modification of the UD annotation scheme in *InterCorp* v.13ud:

- Some common categories from the `feats` list are listed also as regular attributes: `case`, `number`, `gender`, `person`.
- Fused forms are represented as single tokens, with their components still accessible. For more details see Section 3.4 below.
- Each word is also annotated with references to important properties of its head (lemma, part of speech and morphological categories), see Section 3.6.

---

[17] In *InterCorp* v.13ud, annotation errors can be found even in the morphosyntactic annotation of Czech, a language with large and high-quality UD treebanks. E.g. a syncretic adjectival form ending in *–ní* after the preposition *na* 'on' (taking accusative or locative), does not agree in case (accusative or locative) with its immediately following noun head in 259 hits, (1 ipm), as in *na východní hranici* 'on the eastern border'. The same sequence is tagged correctly in 75,591 hits (292 ipm), which means that only 0.34% of such constructions are mistagged. However, there is no mistagged constructions like this in v.13.

[18] Daniel Zeman, p.c.

[19] https://universaldependencies.org/guidelines.html

[20] https://universaldependencies.org/u/pos/index.html

[21] https://universaldependencies.org/u/feat/index.html

[22] https://universaldependencies.org/u/dep/index.html

[23] For a list of the 17 UD word classes see also UD Parts of speech (in the *InterCorp* wiki (https://wiki.korpus.cz/doku.php/en:pojmy:ud#parts_of_speech).

[24] See Section 3.3 below.

- If a content word occurs with a function word (preposition, auxiliary verb, copula, subordinate conjunction, determiner, classifier), the content word includes some properties of the function word (see Section 3.7).[25]
- Annotations differ between languages in the number of categorial attributes and in links to function words.[26]

3.3 Morphological categories

Morphological categories are embedded under the `feats` attribute as a list of `<category_name>=<category_value>` pairs, separated by "|", as in (1).[27] Their choice and values are determined by a part of speech and language, but identical or comparable categories and their values have the same name in all languages.[28]

(1) `feats="Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing"`

3.4 Multi-part tokens

Some tokens, *fused words* in the UD parlance, consist of multiple parts, corresponding to different nodes in the syntactic structure. In English this concerns contractions such as *isn't* or *cannot*. Their orthographic form is preserved, the individual parts are separated only in the annotation (e.g. in `lemma`), with "|" as the separator, resulting in `lemma="be|not"`.

A part of the fused token, such as *n't* in English, may have a different form when occurring as a separate word: *not*. The Czech 2nd person singular auxiliary enclitic *s*, fused with the reflexive *se* into *ses*, corresponds to *jsi*. Both variants (*n't* and *not*, *s* and *jsi*) are represented in the annotation of *isn't* or *ses*: the `iword` attribute shows the original form – `iword="is|n't"` or `iword="se|s"`, while the `sword` attribute shows the unabbreviated version – `sword="is|not"` or `sword="se|jsi"`.[29] The example *Oč mu vlastně jde?* 'What's he really up to?' in Fig. 2 (from the *KonText* tree display of the concordance) concerns a fused form *oč*, consisting of the preposition *o* 'about' and an abbreviated form of the pronoun *co* 'what' preposition. The full form of the pronoun appears as the `sword` value after a click on the node for *č*.

[25] In UD, copula is a function word dependent of the nominal predicate. In *This was not a good moment in the history of English cuisine* the noun *moment* is the root node of the whole tree.

[26] See Description of the list of attributes in the *InterCorp* wiki (https://wiki.korpus.cz/doku.php/en:pojmy:ud#description_of_the_list_of_attributes).

[27] As a reviewer observed, the double use of the "=" operator is unfortunate. Within the feats list the operator means strict equality but after `feats` (a CQL attribute) it means membership in the following list of attribute-value pairs. While the latter use is given by CQL, the former is inherited from the CONLL-U format and could be changed in a future release.

[28] See Universal Features or Other categories in the *InterCorp* wiki (https://universaldependencies.org/u/feat/index.html or https://wiki.korpus.cz/doku.php/en:pojmy:ud#other_categories).

[29] It is this unabbreviated form which is available in the output of *UDPipe*.

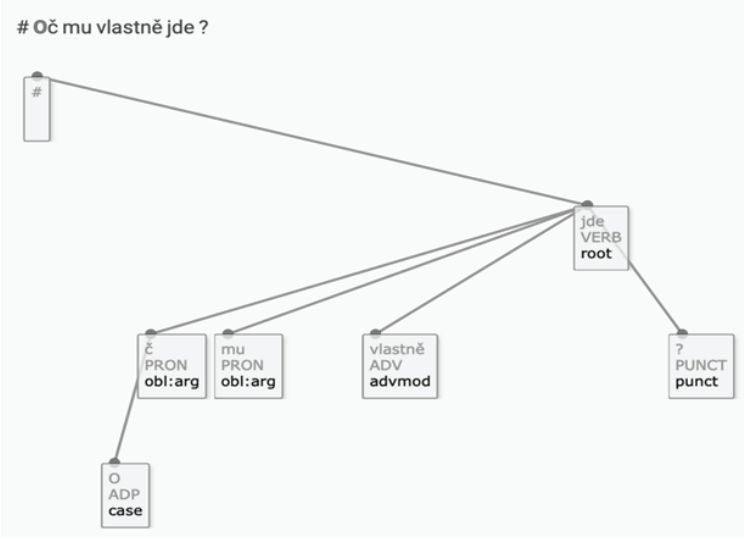

Fig. 2. Syntactic tree including a Czech fused form *oč*, consisting of *o* 'about' + *co* 'what'

3.5 Syntactic functions

Each token specifies its dependency relation (deprel) and a reference to its syntactic governor (head).[30] The 37 relations are cross-classified in terms of their relation to the head (core arguments, non-core dependents, dependents of a noun) and their structural categories (nominals, clauses, modifiers, function words), e.g. iobj (indirect object, a core nominal: *Can I buy* you *a drink?*), advcl (adverbial clause, a non-core clausal dependent: *The country will pay a heavy price* if the president's obsessions prevail for long.), amod (adjectival modifier, a modifier of a noun: *The* sustainable *future of humanity is at stake*.), or cop (copula, a function word as a non-core dependent: *Where's the rest of your luggage?*). Some deprels may have subtypes, e.g. acl:relcl indicates an attribute expressed by a relative clause.

3.6 References to syntactic heads

Each token specifies attributes referring to the token's head (its absolute position – head, or relative to the token – parent). Other attributes of the head include its lemma (p_lemma), POS (p_upos), morphological categories (p_feats), and syntactic function (p_deprel).[31]

[30] See Universal Dependency Relations (https://universaldependencies.org/u/dep/index.html) for an overview and detailed descriptions of syntactic functions used in UD, or Syntactic functions (https://wiki.korpus.cz/doku.php/en:pojmy:ud#syntactic_functions) as they are used in *InterCorp*.

[31] Cf. similar attributes in other syntactically annotated corpora available in the *KonText* browser (https://www.korpus.cz/kontext/query?corpname=syn2020 or https://lindat.mff.cuni.cz/services/kontext/corpora/corplist?keyword=ud).

3.7 References to function words

In UD, function words include auxiliaries, pre- and postpositions, conjunctions, determiners, copulas and classifiers. All function words depend on their content word host. Their types are specified by their deprel attribute.

It is often useful to interpret their properties as features of their content word heads.[32] This is why content words have attributes specifying properties of dependent function words. For example, the lemma of a preposition is shown by the attribute case_lemma, categories of an auxiliary by aux_feats, categories of a copula by cop_feats, word class of a determiner by det_upos, lemma of a marker by mark_lemma. The content word may also include a more detailed specification of the function word's type, e.g. aux_type="pass" or det_type="numgov".[33]

A single content word can govern several function words, e.g. three for *she would have been pleased*. The values, separated by "|", appear concatenated in a single content word attribute. The feats from multiple auxiliaries, concatenated into a single value may include some conflicting values or repetitions. For example, in the sentence *who would have guessed that*, the aux_feats of the content verb *guessed* are composed of the feats of the auxiliary verbs *would* (Mood=Ind|Person=3|Tense=Past|VerbForm=Fin) and *have* (VerbForm=Inf).[34]

3.8 Coordination

In UD, the first conjunct depends on the head of the coordination. Its deprel determines the syntactic function of the whole coordination. The other conjuncts always depend on the first conjunct. Their deprel is specified as conj. Conjunctions depend on the following conjunct with the deprel of cc.

A reference to the "effective head" is used to identify the head of the whole coordination: the e_id attribute refers to the effective head's absolute position, the eparent attribute to its position relative to the token.[35]

---

[32] Cf. the recent addition of auxiliary-related features to the Czech tagset (Jelínek et al. 2022). In addition to the head-related attributes (see Section 3.6), information about the head's function word dependents is sometimes useful but not easily accessible in *KonText.* The solution along the lines of *ibid.* could be an interpretation of *all* dependent function words (e.g. as analytical verb forms or prepositional phrases), stored in the content word head and accessible as another head-related attribute from the head's non-functional daughters.

[33] See https://universaldependencies.org/cs/dep/aux-pass.html, or https://universaldependencies.org/cs/dep/det-numgov.html.

[34] In the next UD-annotated release, categories from different function words concatenated in a single content word's list will be separated by two vertical bars.

[35] One more attribute, useful especially in the annotation of non-initial conjuncts, should specify its "real" deprel, i.e. the syntactic function of the whole coordination, specified in the first conjunct. This attribute (e_deprel) is due to appear in the next UD-annotated release.

4 QUERYING THE UD-ANNOTATED *INTERCORP*

4.1 Basic query

Parts of fused words should not be separated by a space, as in *is n't* or *był bym*, but entered as a whole. The same applies to the value of the `word` attribute in an advanced query. A query for *is* or *n't* will not show concordances including the form *isn't*.

4.2 Advanced query for lemmas and tags

Word class and other categories are listed separately as values of the attributes `upos` and `feats`. The values are identical for all languages. E.g. the proper query for proper names is `[upos="PROPN"]`.

Each of the other categories can be specified separately: the Czech form *radostmi* 'joy-PL-INS' is one of the answers to the query (2). The order of categories in the query is irrelevant. The value of `feats` can also be treated as a string of characters using regular expressions, as in (3). Here the order of categories in the query should correspond to their order in the corpus. Some of the categories in `feats` are listed also outside the list as categorial attributes. The query (4) gives the same result as the two queries above.

(2) `[upos="NOUN" & feats="Gender=Fem" & feats="Number=Plur" & feats="Case=Ins"]`

(3) `[upos="NOUN" & feats=".*Case=Ins.*Gender=Fem.*Number=Plur.*"]`

(4) `[upos="NOUN" & gender="Fem" & case="Ins" & number="Plur"]`

Categorial attributes can also be used to generate frequency lists or in global conditions,[36] e.g. to require morphological agreement of two or more words.[37] Such attributes appear on the light brown background in Attribute list by language[38] or in *KonText* in the lower part of the table shown in `View / Corpus-specific settings...`

For most languages the `xpos` attribute provides a "legacy" language-specific tag, usually identical to the tag used in the previous releases of *InterCorp*. For Czech, the queries (2)–(4) can also be entered as in (5).

(5) `[xpos="NNFP7.*"]`

[36] https://www.sketchengine.eu/documentation/cql-global-conditions/

[37] Note that for technical reasons the names of the categorial attributes are all in lower case, including names such as `VerbForm` (in `feats`), rendered as `verb_form`, or `NumType`, rendered as `num_type`. The attribute values, such as `Fem`, retain the initial upper case character, but are enclosed in double quotes, like other attribute values outside `feats`.

[38] https://wiki.korpus.cz/lib/exe/fetch.php/cnk:intercorp:ud_ic_atributy.pdf

Values of upos and of any category from the feats list can also be entered using the "Insert tag" function of *KonText*. A menu helps to select upos and/or feats categories. The choice is determined by their actual occurrence in the corpus and may reflect incorrect combinations.[39]

4.3 Query for syntactic functions

Syntactic function is specified by the deprel attribute (see Section 3.5 above). E.g. a query meant to show the occurrences of the verb *run* as the head of an adnominal clause is entered as [lemma="run" & deprel="acl"]. Results include examples such as *Everyone of the rabbits was seized by the instinct to run away.* or *Some people have the idea that rabbits spend a good deal of their time running away from foxes.* When querying a deprel that may have a subtype, a possible subtype should be taken into account, i.e. deprel="acl.*" is preferable to deprel="acl".

When the deprel in a query targets a coordinated structure, only the first conjunct is found. Non-initial conjuncts are marked as deprel="conj". To query the "true" deprel of a non-initial conjunct (annotated by deprel="conj"), the p_deprel attribute should be used instead. This attribute shows deprel of the token's head. For example, a query for all indirect objects, including all conjuncts, should use the disjunction operator "|" as in (6).[40] See Section 3.8 above for details.

(6) [deprel="obj" | deprel="conj" & p_deprel="obj"]

4.4 Display of syntactic structure

After clicking on the syntax tree icon preceding a concordance line, the syntactic structure of the sentence is displayed. For each node, the word form, upos and deprel of the token appears. More attributes appear after clicking on the node.

Fused words are split into multiple nodes. After clicking on such a node, its full form (the sword attribute) and the entire token (word) also appear. In the text line above the structure and in the structure, the relevant strings and nodes are highlighted under the cursor in parallel (see Fig. 2 above).

5 CONCLUSION

UD seems to be a good solution for a multilingual corpus when the goal is to avoid heterogeneous annotation. The impossibility of language-universal annotation facing the typological diversity of human languages does not seem to seriously damage the

[39] In the next UD-annotated release, the "Insert tag" function will be extended to cover also syntactic functions (see Section 4.3 below).

[40] We plan to avoid the need to specify queries in such a clumsy way in the new release. See footnote 35 above about the e_deprel attribute. This attribute would duplicate the deprel value of each token, except for non-initial conjuncts, where the value of the first conjunct's deprel would appear.

practical gains of a consistent annotation. It seems that some problems in the use of UD categories for language-specific phenomena are due to the authors' failure to follow the guidelines properly rather than to the inherent impossibility to use the categories cross-linguistically. There is an ample space for improvements in the size of some treebanks, quality of their annotation and the type of texts included. Also, the development of better tools has surely not reached the limits. Together with the progress in the acquisition of larger, more varied and better annotated data, the perspectives are promising, despite the still higher accuracy of some language-specific taggers.

Besides providing a solid foundation for comparative and other cross-lingual studies and applications, UD offers the bonus of syntactic annotation, turning *InterCorp* into a parallel treebank.

ACKNOWLEDGEMENTS

The research is part of the project *Czech National Corpus*, supported by the Ministry of Education, Youth and Sports of the Czech Republic as one of the *Large Research, Development and Innovation Infrastructures* (LM2023044).

References

Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic Typology Meets Universal Dependencies. In M. Dickinson – J. Hajič – S. Kübler – A. Przepiórkowski (eds.): Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), pages 63–75. Bloomington, Indiana University. Accessible at: http://ceur-ws.org/Vol-1779/05croft.pdf.

Čermák, F., and Rosen, A. (2012). The Case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 13(3), pages 411–427.

Haspelmath, M. (2007). Pre-established categories don't exist – Consequences for language description and typology. Linguistic Typology, 11, pages 119–132.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. Language, 86(3), pages 663–687. Accessible at: https://doi.org/10.1353/lan.2010.0021.

Jelínek, T., Petkevič, V., and Skoumalová, H. (2022). Hodnoty slovesných morfologických kategorií v korpusu SYN2020 – atribut verbtag. Časopis pro moderní filologii, 104(1), pages 89–109.

Marneffe de, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pages 255–308. Accessible at: https:doi.org/10.1162/coli_a_00402.

Rosen, A. (2014). A 3D taxonomy of word classes at work. In L. Veselovská – M. Janebová (eds.): Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure, Vol. 4, pages 575–590. Olomouc, Palacký University. Accessible at: http://olinco.upol.cz/assets/olinco-2014-proceedings.pdf.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In E. Gruszczyńska – A. Leńko-Szymańska (eds.): Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora, Vol. 1, pages 21–40. Instytut Lingwistyki Stosowanej. Accessible at: https:doi.org/10.13140/RG.2.1.2808.7444.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207. Accessible at: https:doi.org/10.18653/v1/K18-2020.

sciendo
DOI 10.2478/jazcas-2023-0044

MULTIPLE INTERPRETATION AND FRAGMENTED TEXTS WITHIN A HISTORICAL CORPUS: THE CASE OF OLD EAST SLAVIC VERNACULAR WRITING

DMITRI SITCHINAVA
Department for Slavic Linguistics, Institute of Slavistics, Potsdam University, Potsdam, Germany

SITCHINAVA, Dmitri: Multiple Interpretation and Fragmented Texts within a Historical Corpus: The Case of Old East Slavic Vernacular Writing. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 266 – 274.

Abstract: The paper presents the issue of fragmented and/or ambiguously interpreted texts within the corpora of Old East Slavic vernacular writing. One of these corpora, the corpus of the Old East Slavic birchbark letters, is already available, the other, comprising the texts of Old East Slavic inscriptions, is under preparation. Due to the fragmentary state of many birchbark and epigraphy texts, their lemmatization and grammatical tagging may be uncertain and multiple interpretations may coexist. Some lemmas survive only in fragments which are nevertheless relevant for the study of lexicon. The grammatical status of many fragments may be firmly established despite lacking lexical information. However the relevant data on these fragments is not available in the word indices and corpora that take into consideration only best-preserved word forms. In the paper, the representation and annotation of such word forms within the Old East Slavic vernacular corpora is presented, and relative frequencies of such phenomena within the birchbark letter corpus are shown, with some case studies showing the relevance of the annotation of fragmented forms. The existing approaches, namely for the classical epigraphy within the EpiDoc standard and in the Hittite syntactic treebanks, are also briefly presented and compared to the solution found within the Old East Slavic vernacular corpora.

Keywords: lacunae, epigraphy, fragmented text, historical corpus, birchbark letters annotation, lemmatization, Old East Slavic

1 INTRODUCTION

An "ideal" linguistic corpus, including historical ones, typically requires unambiguous annotation of each token (including the cases of homonymy), although in practice this demand is often neglected. Nevertheless there exist specific texts where ambiguity is not removable, either as a part of their inherent design (which is the case of puns and some poetic texts) or as a consequence of the fact they lack an unquestionable interpretation – which is the case of many historical texts in extinct languages. This may arise due to the fact that the surviving readings are distorted by a copyist and call for a critical reconstruction that cannot be established for sure; *hapax legomena* and other factors of a text in a dead language may also contribute

into the situation. But yet another factor that is under scrutiny in the present paper are short texts that survive either without contexts or as fragments. This is known to be a property of many inscriptions (epigraphy monuments) as well as messages on a dedicated carrier found in archeological excavations, be it tablets of clay, papyri, or East Slavic birchbark letters. The fragmented states of (con)texts, sentences, and tokens, and the ambiguity of interpretations related thereto, are a challenge for building a linguistic corpus.

The paper is organized as follows. In Section 2 we present the state of the art of dealing with fragmented words and broken sentences, using the examples of the ancient epigraphy databases and Hittite corpus. In Section 3 we elaborate on the solutions of these issues as presented in the corpora of the Old East Slavic epigraphy and developed by the author. In the final section we propose discussion of the results.

2 STATE OF THE ART

A standard for building electronic database of epigraphy (mainly of Antiquity, with such languages as Ancient Greek, Latin, or Coptics) and papyri is being developed as a dialect of the TEI encoding standard. Known as EpiDoc, it is described in (Elliott et al. 2007–2020) and is used for building online repositories and databases for such sources as the Greek inscriptions of Northern Black Sea,[1] papyri,[2] the Vindolanda tablets from Roman Britain and others. The structure of the EpiDoc format provides for marking the lost or illegible passages and their probable size, annotating separate symbols interpreted uncertainly, as well as providing alternative lemmatization and analysis. The tradition of classical epigraphy studies includes extensive reconstructions of the lacunae, based on the precedent texts, formulas, or context. These reconstructions are typically included, with their hypothetical status marked, into the representation of the inscription. The interface of these databases typically involves switching between text-only and XML formats.

The lemmatization and even tokenization issues, however, are seen as secondary ones for the EpiDoc project, albeit certainly fully implementable within the TEI format. A separate subsection in the EpiDoc guidelines is "Tagging lexical words in the text and/or linking to lemmata for purposes of indexing or search".[3] This section states that "explicit markup of words (tokenization) and identification of their dictionary headwords (lemmatization) are both optional. Many projects simply leave these features unmarked, or rely on automated processes in search software to detect word-breaks and link to lemmatizing tools such as Morpheus". In this guideline, the task of lemmatization is related only to building lexical indices, whereas morphological or syntactical annotation is not mentioned at all. The developers are invited to link the

[1] https://iospe.kcl.ac.uk/corpus/index.html
[2] https://papyri.info/
[3] https://epidoc.stoa.org/gl/latest/idx-wordslemmata.html

lemmas of their corpus to an online lexicographical database, where the morphological information may be specified. The issue of alternative lemmas is discussed: a token that can be related to multiple lemmas due to a lacune may be marked as such. In the following example given by the authors, the fragment preceding a lacune may instantiate in Greek both the noun ἔκφευξις 'escape' and the verb ἐκφεύγω 'to flee'.

<w lemma="ἔκφευξις ἐκφεύγω" part="I">ΕΚΦΕΥΞ</w>
<gap reason="lost" extent="unknown" unit="character"/>

The EpiDoc guidelines do not mention cases when the set of alternative lemmas is not defined (say, if one annotates a fragmented rare proper name or a potential *hapax legomenon*).

In the paper (Molina – Molin 2016) the problem of reconstruction and broken clauses is addressed within the context of a syntactic corpus of Hittite.[4] The bulk of cuneiform tablets written in Hittite survived in a fragmented state. Unlike the databases of the EpiDoc family, in the Hittite corpus linguistics is in the focus, and, particularly, the syntactic level. The authors distinguish between cases when a "broken clause" can still be analyzed to some extent with regard to its syntactic structure and cases when a token is devoid of surviving syntactic context and this is informative only with regard to lexical and morphological statistics. Overall, five degrees of "brokenness" are specified in the annotation of Molina and Molin's corpus. A clause, the minimal unit of the Hittite corpus, is defined manually and then split automatically into tokens. The fragments of word forms that lack a reconstruction are dealt by the tokenizer either as separate tokens in their own right or as a part of a larger token (although the authors do not state it explicitly). At least the authors quote an example word form […]-*šarraš* glossed as nominative singular of the lemma *X-sarra.* In the paper alternative lemmas or morphological interpretations are not discussed.

3 OLD EAST SLAVIC VERNACULAR CORPORA: PROBLEMS AND SOLUTIONS

3.1 Properties of fragmented and ambiguous texts in the Old East Slavic writing

The Old East Slavic language is an extinct ancestor of Belarusian, Russian, and Ukrainian, corresponding to the period since the first attestations in the 11th century until the conventional date of 1400. For some descriptive purposes this latter date may be shifted.

The monuments of the Old East Slavic epigraphy (mainly on churches' walls or on archaeological objects) as well as birchbark letters, the medieval Slavic documents

[4] http://hittitecorpus.ru/

written on tree bark which was a cheap writing material in Eastern Europe, are the major classes of East Slavic texts that share many challenges with the corpora of cuneiform languages, classical papyri or the epigraphy of Antiquity. They are a valuable linguistic and historical source of everyday spoken and dialectal language, to a larger extent than literary and ecclesiastical texts that survive in the medieval East Slavic books. As such they are described as *non-bookish* or *vernacular* Slavic writing (Franklin 2002), a quality not always shared by the Ancient archaeological written sources (e.g. the Vindolanda tablets reflect pretty standard Latin around AD 100).

Much of the birchbark letters and inscriptions feature physical lacunae (often only a small fragment survives) or are not well legible, with different interpretations available for some symbols. As far as birchbark letters are concerned, 67% qualified either as "fragments" or "small fragments", whereas only 28% survived in full (according to the database gramoty.ru, see the next section in more detail). Compared to the cuneiform tablets or Greek inscriptions, the vernacular East Slavic lacunae are reconstructed with more difficulty. Unlike the Akkadian or Hittite cuneiform, different copies of the same text are rarely found in the Slavic inscriptions or birchbark letters; the exceptions are rare and related mainly to ecclesiastical texts. Unlike the classical Greek inscriptions, the Old East Slavic epigraphy is less formulaic. The epigraphical texts often have multiple interpretations on many levels: for example, for the oldest short Cyrillic inscription found in Old Rus' not less than five different interpretations with different transcriptions were proposed, without a convincing academic consensus (Medynceva 2000, pp. 21–31). As the Old East Slavic texts were normally written without spaces or other signals of word division (*scriptio continua*), competing versions of word division often coexist for different text chunks. Note that in many classical Greek and Roman inscriptions, as opposed to the books, the word segmentation was in place.

The most comprehensive description of the language of the Old East Slavic birchbark letters, the bulk of which is excavated in Veliky Novgorod in what is now Northern Russia is presented in the book (Zalizniak 2004). The book comprises a grammar sketch and a commented publication of the majority of the texts. It is completed with lexical indices (direct and *a tergo*) with extensive lists of word forms and their grammatical analysis. These indices make feel the degree of uncertainty in linguistic description of these texts: the sign "(?)", manifesting this incertainty, occurs in the direct index 642 times, the tag "possible" 313 times, the tag "unknown grammatical form or lexeme" 190 times, three dots marking fragments are attested 501 times. The overall size of the birchbark corpus, by the time the book was published, equalled 15,300 tokens and 3,200 lexemes (Zalizniak 2004, p. 15).

Within the main text of this description fragmented word forms relevant for the linguistic description are mentioned that did not find way into this index – exactly due to their fragmented status. For example, the word form that looks like *...avita*

(possibly ...[*pr*]*avita*) (birchbark letter 111, about 1240s–1260s) might belong to different verbal lexemes, ending in *...praviti* 'to direct, to settle' (maybe with a prefix) or just in *...aviti* (such as *staviti* 'to put', *izbaviti* 'to liberate' etc.). But whatever this lexeme would be, it is positively a verb in dual number and in second person, either imperative or present. This fact is important for the history of the Old Novgorod dialect, as this example is the latest attestation of dual number in verb, superceded by plural in East Slavic (Zalizniak 2004, p. 518). However, this word form cannot be found in the index (as there is no lexeme to ascribe it in an alphabetic list), and if an electronic corpus of birchbark letters was annotated by this index, the form would never be retrieved. Another example is the possessive construction *Vorontsja Vojkov prja…* '(a thing) belonging to Voronets Vojko' from the letter 332 (1160s–1170s). Syntactically it is a typical early Slavic possessive construction: the first name of this person is marked by genitive, the second name by the attributive adjective, followed by a noun starting with *prja…* This possessed object cannot be positively identified (Zalizniak 2004, p. 158; p. 432). Again, this noun is absent from the index, and in a corpus annotated by it, the whole construction "genitive + possessive + noun" would not be searchable.

3.2. Treatment of the fragments and ambiguity in the vernacular corpora

Special historical databases describing not only the texts but the materiality (photographs and description of the carriers, archeological data, etc.) of the birchbark letters[5] and inscriptions[6] are being developed by a Moscow epigraphical team; see (Sitchinava – Dyshkant 2021). In the epigraphy database the option that enables multiple interpretations, transcriptions and translations of the same inscription is entrenched inherently in the design. For example, on the page of the aforementioned 10th-century inscription five transcriptions and five interpretations are listed.[7]

These databases do not feature linguistic annotation. However, both are synchronized with respective subcorpora included into the Russian National Corpus (RNC, ruscorpora.ru). These are the corpora of birchbark letters and epigraphy (the latter is under development, due to be published in 2023); the programming of these corpora is made by Timofey Arkhangelsky and Anton Dyshkant using the indices of (Zalizniak 2004) with manual post-correction and additional annotation. For birchbark letters, a separate syntactic treebank is annotated on the base of the RNC subcorpus by Olga Lyashevskaya and is published in the Universal Dependencies repository as an independent project.[8] This treebank also meets some of the challenges described in (Molina – Molin 2016), although differs from them in that syntactic dependencies are marked there rather than constituents (Lyashevskaya, in print).

[5] http://gramoty.ru

[6] http://epigraphica.ru

[7] http://epigraphica.ru/epigraphy/inscription/show/177

[8] https://universaldependencies.org/treebanks/orv_birchbark/index.html

The specific challenges of the vernacular corpora are dealt with using the following approaches (all the examples provided from the birchbark corpus). Note that similar solutions may be applied to other epigraphical material as well, and some are in fact already implemented in other corpora, such as (1) in EpiDoc or (2) in Molina and Molin's Hittite project (see examples in Section 2 above).

1. Multiple variants for lemmatization and/or grammatical tagging of the same token (including partly surviving tokens) are provided:

<w addr="003:1"><ana lex="поклонъ" gr="S,m,sg,nom"></ana><ana lex="поклонъ" gr="S,m,sg,acc"></ana>поклонъ</w>
<w addr="003:1"><ana lex="Григша" gr="S,persn,m,sg,gen"></ana><ana lex="Гришка" gr="S,persn,m,sg,gen"></ana>грикши</w>

Within the same birchbark letter № 3 alternative analyses of the formulaic *poklonъ* 'bow, greetings' and the proper name *Grikši* are proposed. The first can be construed either as accusative (cf. Latin *salutem* 'greeting.ACC') or nominative, whereas the second one can refer to multiple lemmas: it might be either *Grigša* (with a devoiced consonant) or *Griška* (with misplaced letters) in a normativized form.

2. "Partly surviving" lemmas (see examples above). Lemmas with three dots (initial) are used in the annotation of the corpus. All the grammatical and dictionary information with regards to them is fully searchable.

<w addr="(332a):2a"><ana lex="Воронець" gr="S,persn,m,sg,gen"></ana>вороньцѧ</w>
<w addr="(332a):2a"><ana lex="Воиковъ" gr="A,possess,sg,m,acc"></ana>въ[и]къвъ</w>
<w addr="(332a):2a"><ana lex="прѧ…" gr="S,m,sg,acc"></ana>прѧ---</w>
</line>

This is the example addressed above with the unidentified possessed object with a name starting with *prja-*. It may be, however, accepted for sure that is a masculine noun in accusative singular within a certain construction.

<w addr="001:1"><ana lex="…ии" gr="S,persn,m,sg,dat,fragment"></ana>…[и]ю</w>

Here is an example of the (further unspecified) proper name ending in *-ii*.

Rare are the lemmas with lacunae both in the beginning and in the end (the grammatical status of this adjectival toponym *...oster...* is extracted from the context where multiple place names are listed in repeating constructions).

<w addr="001:1"><ana lex="…остер…" gr="S,topn,n,sg,gen,fragment"></ana>-остер…</w>

3. Fragments where segmentation into words is unclear. A dedicated tag "unknown_segmentation" is added there and the original *scriptio continua* may be preserved in difficult cases.

<w addr="414:3"><ana lex="въ" gr="incorrect,PR,gvrn:loc,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana><ana lex="вѣсъ" gr="incorrect,S,m,sg,loc,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana><ana lex="взвѣсити" gr="incorrect,V,partcp,act,non past,sg,m,nom,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana>вовѣсѧ</w>

In this example, the analysis may be either in two tokens *vo věsja* 'in the weight' or in a single token *vzvěsja* 'having weighted'.

4. Fragments with uncertain lemmas are marked with a dedicated tag "lemma_unsure" (see the example in 3. above).

5. Fully reconstructed tokens (with all the letters lost) are annotated on a par with surviving tokens, if included in a published transcription. This is an example from the same letters with multiple place names; here *selo* 'village' is reconstructed with certainty, as this document is heavily formulaic. Such tokens are marked by a dedicated tag "reconstr".

<w addr="001:1"><ana lex="село" gr="S,n,sg,gen,reconstr"></ana>(села)</w>

6. Fragmented tokens. The tokens that have considerable lacunae and are reconstructed with high probability are marked by a dedicated tag "fragment".

<w addr="036:1"><ana lex="гривна" gr="S,f,fragment"></ana>(гри)<lbr/>вна</w>

In this example the word *grivna* (a currency) is split by line break and only the half of the word survived (although, in this case the partial reconstruction is made with a good degree of certainty).

In Tab. 1 we give absolute and relative frequencies of the described phenomena in the annotation of the birchbark corpus. In the last four lines of the table the instances are given where multiple interpretations are ascribed to the same token, from two to five.

The share of texts where a phenomenon is attested is calculated twice: by all the documents of the birchbark corpus and only by the "good" documents that are not

smallest fragments or texts without interpretation (that is they have a significant amount of interpreted texts). In the last column we give the average number of instances of a given phenomenon per each text where they are attested at least once. This parameter is introduced because of differences between fully preserved documents and texts with lacunae (where we tend to attest more that one phenomenon of the kind).

	instances	share of tokens	texts	share of all the texts	share of "good" texts	instances per text where found
Ending-only lemmas	127	0.54%	95	7.83%	8.64%	1.34
Beginning-only lemmas	91	0.39%	71	5.85%	6.46%	1.28
Unknown segmentation	20	0.09%	19	1.57%	1.73%	1.05
Uncertain lemma	784	3.36%	465	38.30%	42.31%	1.69
Reconstruction (fully lost tokens)	264	1.13%	158	13.01%	14.38%	1.67
Fragments	1,134	4.86%	510	42.01%	46.41%	2.22
2 variants	1,756	7.53%	746	61.45%	67.88%	2.35
3 variants	113	0.48%	102	8.40%	9.28%	1.11
4 variants	38	0.16%	34	2.80%	3.09%	1.12
5 variants	5	0.02%	5	0.41%	0.45%	1.00

Tab. 1. Fragmented and ambiguous tokens within the Birchbark letter corpus

4 DISCUSSION AND CONCLUSIONS

The main contribution of the proposed approach is the filling of the gaps in annotation, providing tagging for the tokens that allow for a lexical or grammatical interpretation but are devoid of it under traditional annotation approach. This solution allows to search for syntactic constructions, for morphemes and derivational types that are present in the text.

The indices in (Zalizniak 2004), including the one *a tergo,* lack "ending-only" lemmas, that are quite well marked grammatically and often keep derivational suffixes as well. 127 lemmas of this kind is now annotated and feature in the treebank (Lyashevskaya, in print) as syntactic nodes.

The phenomena related to the fragmented state and ambiguity of texts are attested on a large scale. Ambiguous tokens are represented at least once in two thirds of birchbark letters, "broken" tokens and uncertain lemmas in 40% of cases. If ambiguity or fragmented word form is attested once in a given text, the average overall number is between two to three attestations per text. Fully reconstructed tokens are found more rarely because of relatively less successful reconstruction of birchbark letters as compared to the inscriptions of Antiquity (see above).

Specific markup such as "uncertain lemmas" and "unknown segmentation" which is not represented in the most widespread epigraphical standard is useful for lexical and syntactical analysis.

The implementation of these approaches in birchbark and epigraphy corpora can be further elaborated in problematic passages within "conventional" livresque literature in extinct languages.

References

Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, Ch., Vanderbilt, S. et al. (2007–2020). EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Accessible at: https://epidoc.stoa.org/gl/latest/.

Franklin, S. (2002). Writing, Society and Culture in Early Rus, c. 950–1300. Cambridge: Cambridge University press, XVI + 325 p.

Lyashevskaya, O. (in print). Syntactic annotation of the Old Russian Birchbark Letters Corpus (submitted).

Medynceva, A. A. (2000). Gramotnost' v Drevnej Rusi [Literacy in Old Rus']. Moskva: Nauka, 293 p.

Molina, M., and Molin, A. (2016). Syntactic Annotation for a Hittite Corpus: Problems and Principles. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016 (no pagination).

Sitchinava, D., and Dyshkant, A. (2021). Integration of the Old East Slavic Epigraphical Databases, Corpora and Indices. In Scripta & E-Scripta 21, pages 95–108.

Zalizniak, A. (2004). Drevnenovgorodskij dialekt [Old Novgorod dialect]. Moskva: Jazyki slavjanskoj kul'tury, 879 p.

sciendo
DOI 10.2478/jazcas-2023-0045

LEMMATIZATION OF THE DIA1900 DIACHRONIC CORPUS

LUCIE BENEŠOVÁ[1] – KLÁRA PIVOŇKOVÁ[1,2]
– MARTIN STLUKA[1]

[1] Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic
[2] Department of Philosophy and History of Science, Faculty of Science, Charles University, Prague, Czech Republic

BENEŠOVÁ, Lucie – PIVOŇKOVÁ, Klára – STLUKA, Martin : Lemmatization of the DIA1900 Diachronic Corpus. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 275 – 284.

Abstract: This paper focuses on the process of lemmatization of the upcoming Czech diachronic corpus of the second half of the 19th century, DIA1900. The article describes different approaches to the corpus lemmatization of synchronic written, spoken and diachronic corpora within the Czech National Corpus project, including single- and multilevel lemmatization and available tools used to link the variants.

Keywords: Czech, diachronic corpus, disambiguation, lemmatization, morphological dictionary, variability

1 INTRODUCTION

The Czech National Corpus[1] (CNC) offers free access to corpora of various types. These include not only large synchronic corpora of written language (SYN series), corpora of spoken language (ORAL, ORTOFON, DIALEKT), but also various types of specialized corpora and the diachronic corpus DIAKORP. The aim of this paper is to describe the concept of lemmatization of another upcoming diachronic corpus, DIA1900.

2 CORPUS DIA1900 AND ITS SPECIFICS

The DIA1900 corpus comprises 103 full texts from the second half of the 19th century and contains a total of approx. 1 million word forms.[2] It is designed to be as representative as possible. It is genre-balanced (fiction/non-fiction/newspaper and magazines) in each decade, including consideration of the balance of length and number of texts. The corpus will be fully lemmatized and morphologically annotated.

[1] https://korpus.cz/
[2] At present (March 2023), this corpus is in its penultimate (disambiguation) phase.

Its aim is to faithfully capture historically relevant linguistic variability, and at the same time in such a way that variation will present as little of an obstacle to the user as possible. The second half of the 19th century represents a specific period in the development of the Czech language, involving several distinctive linguistic phenomena. The gradual expansion of the functional range of Czech in many disciplines was accompanied by the rise of linguistic variability (Kučera et al. 2019). This includes word forms that, although understandable from a contemporary perspective, are unpredictable. For these reasons, the decision was made not to use procedures and tools common in the creation of synchronic corpora when building the DIA1900 corpus (see Benešová et al. 2023).

3 MULTILEVEL LEMMATIZATION AND CORPORA OF DIFFERENT TYPES

Lemmatization, a part of the process of morphological annotation, consists in the assignment of a lemma to one word form. Before we focus on the actual description of the lemmatization of the DIA1900 corpus, it is necessary to mention the change introduced in the field of lemmatization by the latest corpora of the SYN series. Starting with the SYN2020 corpus, lemmatization in the CNC is two-level. In addition to the standard lemma, each form is assigned a sublemma attribute.[3] While a lemma can cover more variants of a single word (e.g. the lemma *filozofie* 'philosophy' represents both forms with *-z-* (*filozofie*) and with *-s-* (*filosofie*)), sublemmas define subsets of forms according to this variation (the sublemma *filozofie* represents only forms with *-z-* (*filozofie*), the sublemma *filosofie* only forms with *-s-* (*filosofie*) in all cases).

Different types of invariance are handled by these sublemmas in synchronic written corpora: e.g. orthographic variation; spelling variation and alternations (*citron*/*citrón* 'lemon') and variants of other types; irregular forms and special categories of gradations (*dobře/lépe/líp* 'well/better'); nominal forms of adjectives and others.

Multilevel lemmatization is an excellent option for solving a number of linguistic dilemmas related to the creation of corpora of different types. The lemmatization of diachronic data, as well as e.g. spoken data processing,[4] is far more challenging than in the case of written synchronic language due to their greater linguistic variability.

The linguistic material of spoken and diachronic corpora could be treated using the aforementioned multilevel lemmatization, and this would be a viable solution. However, at the time the lemmatization concept for the DIA1900 corpus was developing, a two-level lemmatization as implemented in e.g. the SYN2020 corpus

[3] For details, cf. https://wiki.korpus.cz/doku.php/cnk:syn2020:lemmatizace.

[4] The concept of lemma is adapted in these corpora so that all forms realized in diverse graphical and sound variants can be found by the user. Within the spoken language corpora, for example, the lemma *tenhleten* 'this' includes a total of 105 word forms; in Nsg. neuter the following forms occur: *tohleto, todnecto, todleto*, *todlecto*, *todlencto*, *tohlencto*, *tohlento*, *toleto*, *tohlensto*, *todlensto*.

(cf. Hlaváčová et al. 2019) was not yet an option.[5] In the context of diachronic corpora, the need to interlink variant sets of paradigms was even stronger than in synchronic corpora, given that it is a solution that can make historically symptomatic variants available to the user. It was necessary to develop a concept that would, even with a single-level lemmatization, provide a versatile model of variant mapping that also takes into account the typology of the linguistic phenomena reflected in the variants.

4 VARIATION AND FORM OF THE LEMMA IN THE DIA1900 CORPUS

4.1 Morphological dictionary

A morphological dictionary (MD) is a collection of individual word forms with lemmas and morphological tags. A synchronous morphological dictionary is used in the automatic creation process of synchronic corpora.

Although we are aware of the alternative approaches to compiling diachronic corpora[6] (for example, with regard to the distribution of automatic processes used in text preparation and lemmatization), the concept of the DIA1900 corpus reckoned with the completely new creation and generation of a large custom-made morphological dictionary. The resulting MD contains 23.5 million word forms, and the lexical material originates from various sources including 19th-century dictionaries and texts. The tag basis of this MD is the Prague (Hajič's[7]) tagset. The original tagset was modified in several positions for the specific purposes of the DIA1900 corpus (a detailed description of the morphological dictionary is beyond the scope of this paper – for more details cf. Benešová et al. in 2023). This allowed the use of lemmatization not only for the simple preservation of period variability, but also for linking variants with synchronic standard lemmas so that even forms unpredictable to contemporary speakers could be made available. The linking of variants concentrated mainly on phenomena that are not related to word formation and that are not included in the contemporary dictionary.[8]

The MD was tailored to the chosen concept of lemmatization. The creation of the MD was motivated by a desire to capture the wide variability of 19th-century language, which could not be adequately treated by any of the existing synchronic tools, i.e., without a large error rate.

---

[5] The concept of two-level lemmatization of the SYN series corpora was developed later and independently of the concept of lemmatization of the DIA1900 corpus.

[6] Cf. e.g. Hana et al. 2012; Kieraś et al 2018.

[7] Cf. Hajič 2004.

[8] It was not the aim in this respect to cross the word-formation and lexical boundary; lemmas such as *propůjčník,* 'lender', *zakazovatel* 'forbidder' or *vodychudý* 'water-poor' (all obsolete) will not be linked in any way to potential equivalents that are part of synchronic language usage, since such a link would be unreliable and rather confusing without a detailed lexicographical analysis.

Variation is captured in the MD both at the level of forms and at the level of lemmas.[9] The form-level variation is characterized by the existence of overabundance (cf. Thornton 2019) within the same paradigm (the synchronic aspect of standard or non-standard variation is not relevant). A typical case of form-level variability is e.g. the form of the word *tráva* 'grass' for INS sg. (*trávou*/ *travou*).[10]

Lemma-level variability refers to all word forms related to a single paradigm. The problem of how to treat the variability of this type is ever-present in diachronic corpora due to the spread of linguistic changes over time. The MD framework reflects 2 types of this variability, which are implemented in the corpus by the attribute of lemma, respectively hyperlemma[11] and the Search Suggestions Tool (hereafter SST) (see Section 4.2 for details).

The MD also includes the generated morphological categories in a different form and scope in comparison with contemporary Czech. For example, for the past and present transgressive it is necessary to assume the 19th-century interpretation that each of the forms of these transgressives is capable of expressing any gender and number.[12] Such a broader conception of certain morphological categories allows us to decide in favour of function over form in specific cases (e.g. within a morphological category, the form of the transgressive *dělaje* 'doing' can be feminine in the sg.).

The MD enabled us to identify linguistic phenomena typical of an earlier phase of Czech, and the modified tagset allowed us to mark categories evaluated as relevant and worthy of specification.[13] Providing a detailed description of all the tagset specifics of the DIA1900 corpus goes beyond the scope of this paper, but at least two examples are worth mentioning: marking the proprial character in the 6th position and marking the affixed particle in the 15th position of the tagset.[14]

The success rate of MD when first applied to the corpus text material was 97%, and further modifications increased this score.

---

[9] Within the context of the cited paper by Hlaváčová et al. (2019), we can refer to inflectional and paradigmatic variability.

[10] Quantity is the most frequent source of variation in the 19th century, both at the level of forms and at the level of lemmas.

[11] On the concept of hyperlemma, cf. https://wiki.korpus.cz/doku.php/pojmy:lemma; e.g. the forms *bejt* and *být* have the lemma *být* 'be' assigned throughout their paradigms.

[12] The usage of past and present transgressive was considerably divergent in the 19th century due to conflicting tendencies, one of which led to grammaticalization in the language and the other to their rehabilitation in the state that existed about two centuries earlier.

[13] For a detailed description of the tagset values of the DIA1900 corpus specifics cf. Benešová et al. (2023).

[14] The usage of the affixed particle -ť, -tě, -(i)ž is a frequent linguistic phenomenon in the 19th century (***Bylotě** to jeho přání* 'it **was** his wish'). This has been reflected in the MD – the forms with the affixed particles have been generated to all verbal and adjective paradigms, e.g. *bylotě být* VpNS------AA--**T**I.

4.2 Linking variants

In order to link variants and to facilitate the search, three procedures were reserved in the lemmatization process of the DIA1900 corpus:

1. Assigning an individual form, set of forms, or the whole paradigm to a synchronically standard lemma. One lemma has more than one paradigm in the MD in the following cases:

(a) Regular vowel variations of a vocalic nature (*ú>ou, ý>ej*, prothetic *v* and, to a limited extent, *é>í*) are handled in this way – these variations are assigned to the lemma that represents the synchronic standardized variant. Thanks to this procedure, they are both preserved and traceable.

(b) This procedure is also used in cases which, within the context of linguistic and lexicographic convention, appeared to be a user-friendly solution: comparative and superlative forms for adjectives and adverbs are assigned to the positive, nominal forms of adjectives to the standard adjectival lemma, and the negative paradigm for verbs is assigned to the affirmative lemma[15] (see Tab. 1, Section 7).

2. The second approach is to draw the user's attention to the relationship between the lemmas via the search engine user interface, more specifically by the SST. This tool offers variant lemmas during the process of setting a corpus query. Typically, these are variants in the domain of spelling, or for example variants resulting from consonant alternation and phenomena related to variations in vocalic quantity,[16] manifested in the whole paradigm[17] (see Tab. 1, Section 5).

3. The third technique used is the assignment of more than one lemma to a single form. These are both cases of aggregates that contain more lemmas in a single token (*uvěřils* > *uvěřit_být* 'you believed > to believe_to be') and so-called "multi-lemma", i.e., capturing form-level homonymy in semantically equivalent variants where the lemma cannot be determined from the context. Although this procedure is not directly related to the user's query input, it provides the user with information about the existence of variability that might be unintuitive to the contemporary native speaker (see Tab. 1, Sections 2 and 3; see more detail in Section 5).

---

[15] The area of negation signalled by the prefix of *ne-* 'not' was generally treated as follows in the lemmatization: negative verb forms were assigned to the affirmative lemma and have a marked negation in the 11th position of the tag. For adjectives, the forms with negation were given the negative lemma and have negation marked in the 11th position of the tag. Nouns beginning in the prefix *ne-* also have a lemma in *ne-*, but due to more advanced lexicalization, negation is no longer marked in the 11th position of the tag.

[16] Lemmas that are non-standard from today's point of view, and hence harder to predict, will be offered during query entry even if the corpus does not contain a synchronically standard lemma. E.g. the DIA1900 corpus does not contain the *užívatel* 'user' lemma, but we do find here the forms of the *užívatel* 'user' lemma. The existence of a lemma with an irregular quantity is pointed out to the SST when the synchronously standard variant *uživatel* 'user' is entered.

[17] Where the deviation is restricted to a single form (e.g. for alternation in the durative participle of verbs, for quantity in single forms in paradigms where the alternation of quantity is standard), the variant form is assigned as a doublet (procedure 1.b, see Tab. 1, Section 6).

LIST OF EXAMPLES			
	Token	**Lemma**	**SST**
I. *regular changes of vocalic character*	*kdož jste vy podlého* ***smejšlení*** 'who are you of a mean **mind**'	smýšlení 'mind, opinion'	no
	smýšlení *o tom rozdílné panuje* 'there is a difference of **opinion** about it'		
	nalézáme jména: [...] štika, kapr, ***vokoun*** 'we find the names: pike, carp, **perch**'	okoun 'perch'	no
	nelze rozeznati, je-li ***okoun*** *mlíčný* 'it is not possible to distinguish if the **perch** has a milt'		
	nepřítele ***nejouhlavnějšího*** *pouštíš tak na svobodu* 'you're letting your **main** enemy go free'	úhlavní 'main'	no
	na [...] Pinovi spočívalo podezření co ***úhlavním*** *spachateli* 'Pino was suspected of being the **main** perpetrator'		
II. *form-level homonymy for lemmas that are synonymous and predictable to the contemporary speaker*	*Černá* ***země*** *ssaje krve potoky* 'the black **earth** sucks streams of blood'	země 'earth'	no
	pálená ***zem*** *s popelem se roztrousila* 'the scorched **earth** with ashes was scattered'	zem 'earth'	no
	zasype je ***zemí*** 'he'll cover them with **earth**'	země/zem 'earth'	no
III. *form-level homonymy for lemmas that are synonymous and include a variant that is potentially unpredictable to the contemporary speaker*	*umrtvuje v sobě ruch* ***duchovný*** 'he mortifies **spiritual** activity in itself'	duchovný 'spiritual, clerical'	yes
	duchovní *ruch poměrně živěji v Evropě [se] osvědčoval* '**spiritual** activity proved to be relatively more lively in Europe'	duchovní 'spiritual, clerical'	yes
	duchovní *hodnostáři* '**clerical** dignitaries'	duchovný/duchovní 'spiritual, clerical'	yes
IV. *phenomena covering equivalent variants (typically spelling) that are potentially unpredictable to contemporary native speakers*	*Tento* ***způsob*** *sázení* 'this **way** of planting'	způsob 'way'	yes
	dosavadní ***způsob*** *panování* 'the existing **way** of ruling'	způsob 'way'	yes
	Černá péra [...] můžeme dle ***methody*** *této úplně vybíliti* 'black feathers can be completely bleached out using this **method**'	methoda 'method'	yes
	Návodův a ***metod*** *ke cvičením zpěvu zná každý* 'everyone knows the instructions and **methods** to practice singing'	metoda 'method'	yes

LIST OF EXAMPLES			
	Token	**Lemma**	**SST**
V. *phenomena related to e.g. vocalic quantity or consonant alternation – the whole paradigm*	*když dosavádní **užívatel** dobrovolně zákazu [...] se podrobí* 'when an existing **user** voluntarily submits to a ban'	užívatel 'user'	yes
	*Mezi Rakouskem a Bavory je uzavřeno **usnešení*** 'a **resolution** is concluded between Austria and Bavaria'	usnešení 'resolution'	yes
	*aby se **usnesení** o peticích odložilo* 'that the **resolution** on the petitions be postponed'	usnesení 'resolution'	yes
VI. *phenomena related to e.g. consonant alternation or quantity – single form*	*Bylť jest [...] ze země **vypovězen*** 'he was **expelled** from the country'	vypovědět 'expel'	no
	*ze země **vypověděna** byla* 'she was **expelled** from the country'		
	*Měřice **hráchu*** 'scoops of **peas**'	hrách 'peas'	no
VII. *negation, gradation, nominal adjectives*	*nutili otroky své ku […] **nepravostem*** 'they forced their slaves into **iniquity**'	nepravost 'iniquity'	no
	*vše připraveno bylo rukou [...] **nešťastných** spolubratrů* 'it was all set up by the **unfortunate** fellows'	nešťastný 'unfortunate'	no
	*vina **nepochybně** padá na manželku* 'the blame **undoubtedly** falls on the wife'	nepochybně 'undoubtedly'	no
	*Ploutev […] **má** načervenalý kraj* 'the fin **has** a reddish edge'	mít 'to have'	no
	*jednotlivý víc **nemá** místa* 'the individual **has no** place anymore'		
	*krmiti ji musíme hojně a **dobře*** 'we have to feed her **well** and abundantly'	dobře 'well'	no
	*jeho zboží [...] **nejlépe** se doporučuje* 'his goods are **best** recommended'		
	*Zůstals **věren** učení* 'you have remained **faithful** to the teachings'	věrný 'faithful'	no
	*Vítej mi, **věrný** strážce* 'welcome, **faithful** guardian'		

Tab. 1. Methods of capturing variability in the DIA1900 corpus

5 DISAMBIGUATION

After the MD was applied to the texts, the next step was the creation of an etalon corpus with 500 thousand tokens (the size of the etalon corpus corresponds to the needs of later automatic annotation, as well as to the personnel and time constraints).[18] The process of lemmatization and manual disambiguation consists not only in the selection of a single-valued lemma for a given form, but often also in deciding on the assignment of a multi-lemma.

5.1 Single-lemma choice

Concerning the ambiguous forms, we distinguish between homonymy that can be disambiguated by the context and homonymy where our decision can only be based on knowledge of the language system of a given time period and is necessarily inherently associated with uncertainty. In other words, we are able to resolve the homonymy when we have sufficient relevant evidence to make a decision, thanks to a sufficiently comprehensive context and semantics. Based on the context that identifies the meaning of a given token, we are able to assign an adequate lemma and tag, e.g. in the context *valně se liší od* ***divákův*** 'largely differs from **spectators**' the form *divákův* is the noun *divák* 'spectator', not the possessive adjective *divákův* 'spectator's'.

The problem of homonymy arises when the language system involves equivalent variants within the same paradigm and the context cannot be used to specify them. At the same time, it is necessary to comply with the requirement that all categories that are part of a tag are unambiguously specified. As a result, a decision in favour of one of the variants is necessary. Therefore, specific disambiguation rules regarding these contentious solutions have been developed, taking into account the assumed language and grammar of the 19th century.[19] Even so, in many cases it is not possible to make a clear-cut decision with certainty. This is the case, for example, with the case homonymy of the genitive and accusative (*nesnadno si představiti* ***člověka*** *tak lhostejného* 'it is not easy to imagine a **man** so indifferent').[20]

5.2 Multi-lemma choice

However, the specific rules adopted, together with the technical solution of lemmatization, also allow us to assign multi-lemmas (each with their own tag) to a single token. As an example, we can use the form *křeče* 'cramp, spasm' in the sentence *stonala 15 let na žaludkové* ***křeče*** 'she suffered from stomach **cramps** for 15 years', where we have the possibility of assigning both a lemma and a tag for the

[18] The application of the MD resulted in 43.45% of unambiguous tokens and 56.55% of ambiguous tokens that need to be manually disambiguated.

[19] E.g. Pravidla (1902) and (1913), and Jungmann (1835–1839), PSJČ (1935–1957).

[20] In 19th-century Czech, there are some linguistic phenomena which are no longer known in contemporary Czech and which do not make the decision easier, e.g. the genitive of negation.

feminine and the masculine to a given token, i.e., *křeč* ‘cramp’ N-FP4-----A----- and *křeč* ‘cramp’ N-IP4-----A-----). However, assigning multiple lemmas to a single token also has limitations. If we disambiguate the form of the adjective ***žaludkové*** ‘stomach’ in the above sentence, we need to choose either the masculine or the feminine (due to the adjectival character of the lemma, the forms of both genders belong to the same lemma). In these cases, we choose the gender variant of the adjective that is more prevalent in the text with the controlling noun in unambiguous contexts. If we do not have contextual support, we take into account the linguistic usage of the 19th century or the contemporary usage.

6 CONCLUSION

Building a diachronic corpus is always a complex challenge. The DIA1900 corpus represents one possible solution to such a task. In principle, the concept of lemmatization of the DIA1900 corpus offers the user access to historically characteristic variability, either through directly assigned synchronously standard lemmas or through the SST when entering a corpus query. Thanks to the chosen measures in the processing of the DIA1900, in particular the creation of a manually disambiguated etalon that can be used as training data for the tagger (automatic annotation), a solution has been provided that will enable easier automatic processing of other corpora in future, even using multilevel lemmatization.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2023044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

Benešová, L., Kučera, K., Najbrtová, K., Pivoňková, K., and Stluka, M. (2023). Korpus DIA1900: jeho koncepce a vytváření. Časopis pro moderní filologii, 105, pages 121–140.

Hajič, J. (2004). Disambiguation of Rich Inflection: Computational Morphology of Czech. Praha: Karolinum.

Hana, J., Lehečka, B., Feldmann, A., Černá, A., and Oliva, K. (2011). Building a Corpus of Old Czech. In Proceedings of the Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects Workshop associated with the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 9–15.

Hlaváčová, J., Mikulová, M., Štěpánková, B., and Hajič, J. (2019). Modifications of the Czech morphologigal dictionary for consistent corpus annotation. Journal of Linguistics, 70(2), pages 380–389.

Jungmann, J. (1835–1839). Slovník česko-německý. Praha. Accessible at: https://vokabular.ujc.cas.cz/moduly/slovniky/.

Kieraś, W., and Woliński, M. (2018). Manually Annotated Corpus of Polish Texts Published between 1830 and 1918. In Proceedings of the Eleventh International Conference on Language Recources and Evaluation (LREC 2018). Miyazaki, Japan: ELRA, pages 3854–3859.

lemma. https://wiki.korpus.cz/doku.php/pojmy:lemma (Last visited March 24, 2023).

lemmatizace. https://wiki.korpus.cz/doku.php/cnk:syn2020:lemmatizace (Last visited March 24, 2023)

Kučera, K., Najbrtová, K., Pivoňková, K., Řehořková, A., and Stluka, M. (2019). Korpus českého jazyka 2. poloviny 19. století. Časopis pro moderní filologii, 101, pages 92–97.

Příruční slovník jazyka českého (PSJČ) (1935–1957). Accessible at: https://bara.ujc.cas.cz/psjc/.

Pravidla hledící k českému pravopisu a tvarosloví s abecedním seznamem slov a tvarů (1902). Praha: Školní knihosklad.

Pravidla českého pravopisu s abecedním seznamem slov a tvarů: Jediné c.k. ministerstvem kultu a vyučování schválené vydání (1913). Praha: Školní knihosklad.

Thornton, A. M. (2019). Overabundance in Morphology. In Oxford Research Encyclopedia of Linguistics. Oxford University Press.

NATURAL LANGUAGE PROCESSING AND DIGITAL HUMANITIES

sciendo
DOI 10.2478/jazcas-2023-0046

USE OF COMPUTER AND CORPUS TOOLS IN THE RESEARCH OF A 19TH CENTURY GERMAN-LANGUAGE MANUSCRIPT BOOK OF NOTES AND EXTRACTS

MARTIN BRAXATORIS[1] – ANITA BRAXATORISOVÁ[2]
[1] Institute of Slovak Literature, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Department of German Studies, Faculty of Arts, University of Ss. Cyril and Methodius, Trnava, Slovakia

BRAXATORIS, Martin – BRAXATORISOVÁ, Anita: Use of Computer and Corpus Tools in the Research of a 19th Century German-language Manuscript Book of Notes and Extracts. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 287 – 300.

Abstract: The study explores the possibilities of using computer and corpus tools in the interpretation of texts of the genre of book of notes and extracts; these are documents consisting of extracts and modified excerpts from contemporary press and literature, records of the author's own thoughts, etc. Samuel Ferjenčík's manuscript is a German-language document by a Slovak author intended for private use; cited or adapted passages are usually given without any reference to the source. The paper introduces the problems of automatic identification of the source base, which relate to the application of OCR and content similarity detection tools. It discusses the results of text matching, which revealed several manipulations of source texts, especially substitutions, indicating attitudes and priority problems in the author's thought-world. It further interprets the results of the use of the Sketch Engine corpus manager tools by which the frequency of occurrence of key terms and their collocability were investigated, paying special attention to substituted words. The paper is an example of the application of computer and corpus-linguistics methods to the interpretation of literary texts, which is represented by a number of current studies in the field of digital humanities. The proposed approaches are applicable to research on other books of notes and extracts, topical in the context of research trends related to egodocuments, as well as to textual research on monuments of other genres.

Keywords: egodocument, digitisation, source database, content similarity detection, corpus analysis

1 INTRODUCTION

This study focuses on the possibilities of using computer and corpus tools in the interpretation of texts belonging to a specific genre, such as a book of notes and extracts: a document consisting of extracts and modified excerpts from contemporary press and literature, records of the author's own thoughts, etc. The examined work of

Samuel Ferjenčík (1793–1855)[1] contains 908 entries on 155 manuscript pages and represents a German-language document of the Slovak author, which was intended for private use.[2] The original assumption that it was primarily Ferjenčík's original work was refuted by a linguistic (mainly stylistic) analysis of the text and its subsequent textual-critical treatment. It was found that these were notes from Ferjenčík's own reading, serving to capture ideas and formulations that he considered impressive and worthy of recording; given their persuasive power, he might have used them, for example, when arguing before state and church dignitaries, at church assemblies and conventions, and when composing his own writings, articles, and sermons. The genre of monument can be characterised as a special type of egodocument[3] in which the personality of the author is constantly present through the application of a specific selection and modification key. The document can be described as an important key to the attitudes of its author in the 1840s. The research of the manuscript acquires particular significance in connection with the controversies in the author's biography: on the one hand, active advocacy for the linguistic and national rights of Slovaks in Hungary, including participation in the delegation of the Slovak Petition to the Throne in 1842, and on the other hand, speaking out from patriotic Hungarian positions against the Slovak movement in the revolutionary years of 1848/1849.

Given the nature of the genre of book of extracts, intertextual correspondence with identified source documents is of key importance in researching the monument. Corpus-based research on intertextuality already has its own tradition (e.g. Teubert 2010, pp. 199–214; Stubbs 2015; Mao 2015; Hohl Trillini – Quassdorf 2010; Visser – Duthie – Lawrence – Reed 2018; Chollier 2014; Burns – Brofos – Chaudhuri – Li – Dexter 2021; Milička – Cvrček – Lukeš 2022). However, the present study focuses on the specific situation when the source base of a single literary monument is being reconstructed, a large corpus containing potential source texts does not exist, and its

[1] Samuel Ferjenčík was a Slovak Evangelical priest, publicist, meteorologist and pomologist, known for his association with Ján Kollár and his acquaintance with Johann Wolfgang Goethe. From 1827 until his death he lived in Jelšava, where he served as a senior deputy and later as a senior priest of Gemer. He was actively involved in religious, scientific, and journalistic writing, as well as in organizing church and school life, making him one of the most influential intellectual figures in the contemporary Slovak context. In 1842, he co-presented the Slovak Petition to the Throne, protesting against the national situation in the Kingdom of Hungary. He opposed the Slovak revolutionary movement of 1848–1849.

[2] The broader project research also included the processing of several other documents related to the author's life and work.

[3] The term "egodocument" was introduced in the 1950s by the Dutch historian Jacques Presser. He used it to refer to writings with an authorial "ego" as "writing and describing subject with a continuous presence in the text", which "intentionally or unintentionally discloses, or hides itself" (Baggerman – Dekker 2018, p. 93; Dekker 2002, p. 7; see Presser 1958; Presser 1969). However, egodocuments are understood variously in different national contexts, disciplines, and authorial approaches. Their definitions and understandings encompass diverse expressions of subjectivity and personhood in historical documents (on conceptions, see Depkat 2019 and the resources listed there).

creation specifically for these purposes would be like using a sledgehammer to crack a nut. The absence of such a corpus is therefore compensated by a combination of manual search, content similarity detection tools, and a smaller corpus of individual work using the Sketch Engine manager.[4]

2 MANUAL AND AUTOMATIC RECONSTRUCTION OF THE SOURCE BASE

Research on the document is complicated by the fact that, with a few exceptions, cited or adapted passages are given in the manuscript without any reference to their sources. This causes problems in the reconstruction of the source base, which are connected with the identification of potential source documents and with the results of their character recognition and transcription, editing, and further processing, as well as with the retrieval and localisation of strings from the text in the corpus of potential source texts. The research has shown that the vast majority of extracts and adapted passages are based on pre-texts published between 1840 and 1842 or earlier, but exceptionally as late as 1846 or 1848. Among the source documents, the relevant volumes of the *Allgemeine Zeitung* and *Allgemeine Kirchenzeitung* (including their supplements), several books, and other magazine titles appear very frequently.

The implementation of the research began with the creation of a digital facsimile of the manuscript and its diplomatic transcription. Due to the transcriber's preferred working procedures, the transcription was performed manually with a multistage check: 1. comparison with the automatic transcription using Transkribus, model German handwriting M1 (in this case with a roughly 10% error rate of the machine transcription; however, due to the previous manual transcription, no correction of the results was performed in Transkribus), which revealed mainly minor errors of the transcriber; and 2. verification against the identified source base.[5] The automatic transcription facilitated the navigation of the manuscript even for collaborators with limited experience working with 19th-century German-language manuscripts.

Based on the transcription of the monument, various character strings from individual records were searched, initially primarily in the Google Books database (usually by searching for exact phrases in quotation marks). The search was complicated by issues related to the results of character recognition and transcription of potential source documents (misjoining of text in adjacent columns into a single line, separation of continuous text in adjacent lines, word splitting, incorrect transcription of individual characters), for which multiple potentially suitable strings had to be repeatedly extracted and searched for in most records.

[4] http://www.sketchengine.eu/

[5] We would like to express our gratitude to Dr. habil. László V. Szabó, PhD., for his careful and highly professional transcription of the monument.

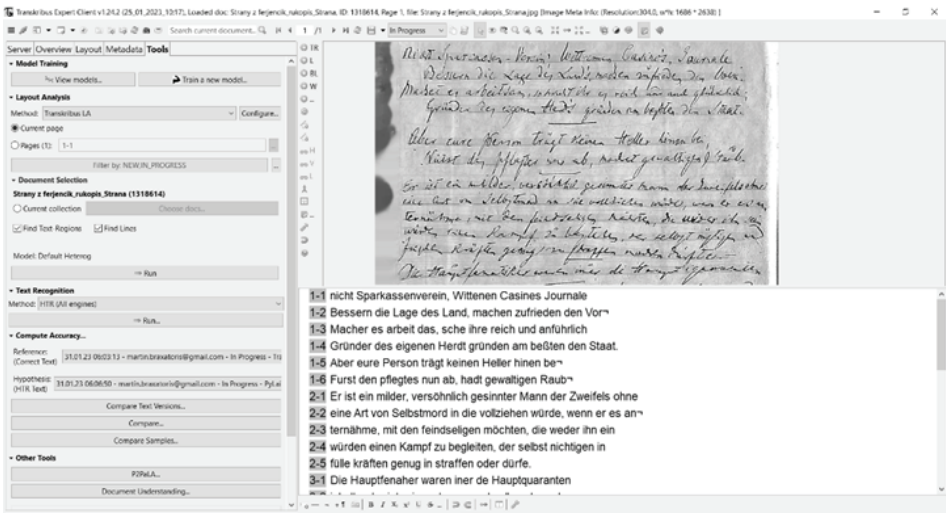

Fig. 1. Automatic transcription of a manuscript fragment in Transkribus (German handwriting M1)

The processing of the majority of the text showed that most of the identified items (around 80%) were digitised by the Bavarian State Library, although a significant part was digitised by other memory institutions (e.g. the National Library of Israel, the University of Bern, the University of California, University Library in Bratislava). In view of this knowledge, the search for the source base of the remaining manuscript text was also performed by a full-text search in the database of the Munich Digitisation Centre of the Bavarian State Library[6] (MDZ/BSB), which discovered sources of other manuscript entries, pre-texts of identified source documents, alternative editions of texts, etc.[7]

For certain documents where the source text was assumed to be present, but the full-text search did not reveal a match, character recognition in facsimiles of potential German-language printed preprints was used via the Transkribus platform, model ONB _Newseye_GT_M1+ German Fraktur 19th–20th century. This method proved to be highly reliable; however, its application was limited by the number of pages that can be processed in the software free of charge (500) and by the necessity to edit text regions, lines, and recognised text.

In parallel with the ongoing textual-critical research, the possibilities of using available content similarity detection (i.e., anti-plagiarism) tools were verified to

[6] https://digitale-sammlungen.de

[7] Additionally, several other databases were searched, including DiFMOE – Digital Forum Central and Eastern Europe (http://difmoe.eu) and others.

determine the methodological possibilities of processing other texts of the given genre. Books of notes and extracts represent a relatively common genre, which has a significant potential in terms of current trends in the research of egodocuments, or historical documents that reveal the personality of their author. In the future, content similarity detection tools could radically simplify the identification of their source base and the detection of interference with the source text, which is the basis for identifying forms of adapting textual sources. Since the tools mentioned above work with certain databases or corpora that vary in focus and robustness, the search for suitable software has encountered the issue that the available anti-plagiarism programs index a wide variety of databases of theses, scientific studies, etc. of contemporary texts but not the digital collections of the Bavarian State Library or the Google Books database. The success of the application of these tools would be enhanced by the availability of the source base in the archive.org database, which is indexed by some of the tools tested, but specifically the volumes of the *Allgemeine Zeitung* and the *Allgemeine Kirchenzeitung* (1840–1842) and other source documents are absent from it. Therefore, the free versions of the available anti-plagiarism tools proved to be only partially effective for our purposes. For example, the plag.sk tool detected only 1% similarity to other texts, while the results of its application, however, made it possible to supplement the source base with a certain number of previously undetected items.

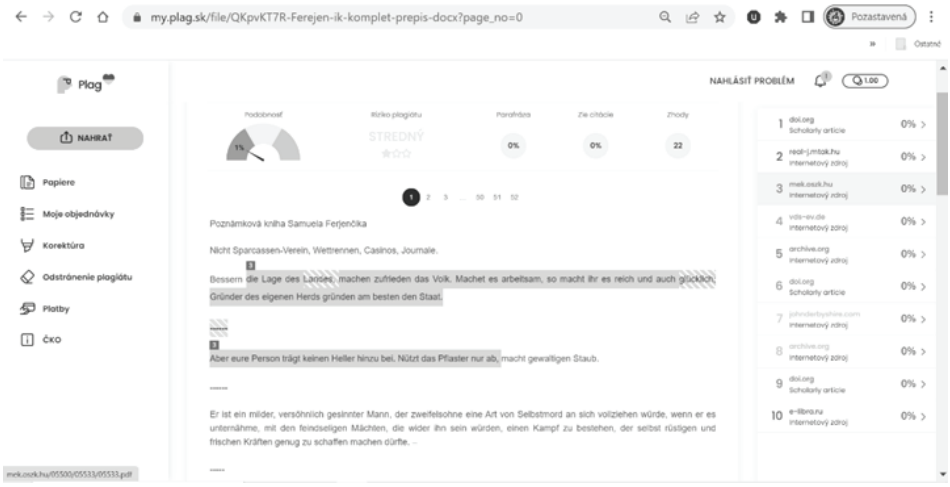

Fig. 2. Example of a text match detected by Plag.sk

We excluded several other applications from the range of tested tools due to the severely limited functionality of their trial versions. However, some of the commercial applications with a limited scope of free text (10 pages and 1 page, respectively)

proved to be promising and useful in completing the source text base. For example, the PlagAware application revealed a 5.19% similarity with other texts over a 10-page area, while the PlagiarismCheck.org application revealed a 31.79% similarity with other sources (with references) over a 1-page area. However, the percentage of success is not indicative of the effectiveness of the programs in question as it strongly depends on the choice of the parts of the text that have been uploaded for processing.[8]

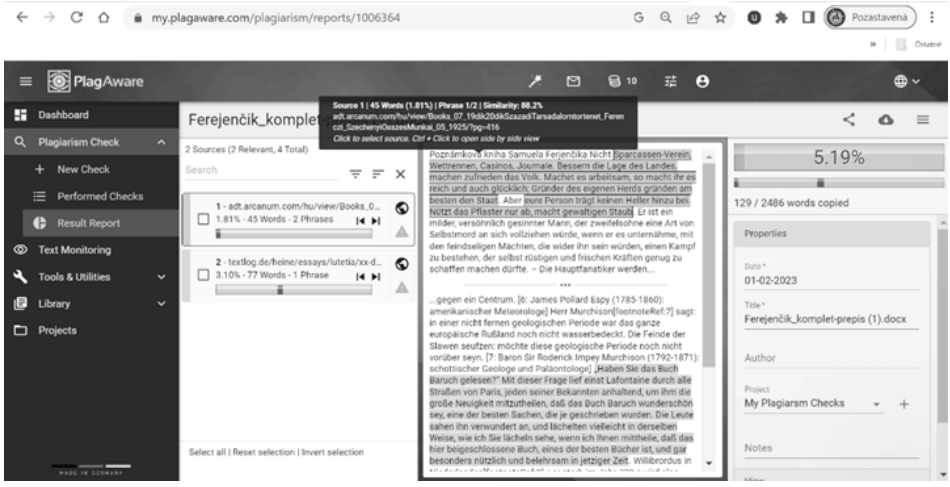

Fig. 3. Example of text matches detected by PlagAware

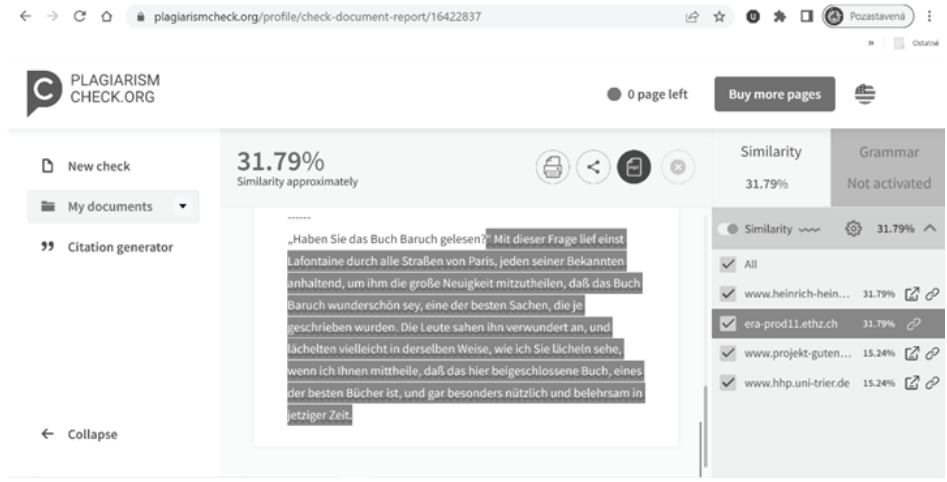

Fig. 4. Example of a text match detected by PlagiarismCheck.org

[8] Even if one set of test pages is used for comparison, the results are not comparable as the results vary substantially when a different set is chosen.

Further considerations were directed towards the possibilities of adapting corpora of existing anti-plagiarism tools, in particular to indexing resources in the archives of the MDZ/BSB, etc. In this context, we asked the Centre for Scientific and Technical Information of the Slovak Republic (CVTI SR) for assistance, and they provided trial testing of the text overlap with the thesis index of the Central Register of Final Theses (CRZP) corpus, Internet sources, Wikipedia, and the Slov-Lex database through the Antiplagiarism System (APS).[9] Testing revealed a 2.06% similarity with other texts[10] and allowed for limited supplementation of the source base with individual entries. Currently, the usefulness of this system is limited by its focus on different types of sources, the unavailability of some source documents in indexed databases, and the restrictions on expanding the APS corpus with items unrelated to the originality check of theses and dissertations. Technically, the indexing of related databases would be complicated by their security settings, but it would be possible to upload text files of potential source documents to the system, based on the textual similarity that can be reliably identified. For example, the MDZ/BSB sends digitised documents with OCR results in TXT format upon individual request for access to a particular file; in our case, however, the range of potential source documents is very wide[11] and their use requires further editing (including removing word splitting, etc.). The realisation of the possibility of processing the content of the library's digital archive by means of the APS crawler is a matter of potential future research collaboration, for which the involvement of several institutions, including the MDZ/BSB, is required.

3 TEXT COMPARISON

The text matching was performed in a combined way: manually and automatically. During the textological research, we have identified further possible limitations of automatic comparison through a specialised corpus: 1. to determine a potential source base, it must be thoroughly mapped over a wide area of different sources. Manual or automatic search for duplicates through a search engine or in many databases is therefore irreplaceable. This is even more important when not all the items have been digitised and made accessible by a single memory institution. 2. With major interventions in the text of the prelims, not to mention interlingual translation, textual parallels become virtually unidentifiable by means other than classical reading. The aforementioned circumstance limits the usefulness of automatic duplicate detection in research of texts of other genres as well. 3. Particular

[9] We would like to express our gratitude to CVTI SR for the exceptionally responsive cooperation in the implementation of the research objectives.

[10] However, the percentage cannot be directly compared with the compliance rate quantified by other tools, as the methodology for determining the number of characters is different.

[11] This virtually precludes comparing the manuscript transcription with individual files by simply checking the similarity of the text files.

correspondences should always be searched for and interpreted in the light of their nature and the context of the source documents. Interpretive procedures and other research methods depend on the document under study, its pre-texts, and the nature of the textual matches.

Fig. 5. Example of text matches detected by APS CVTI SR

In the case of Samuel Ferjenčík's book of notes and extracts, it has proved desirable that the interpretative procedures focus primarily on a contrastive grasp of its text and its sources. The textual comparison revealed several interventions in the pre-texts, such as omissions, insertions, substitutions, and contextual transpositions. Some of them point to priority themes in the author's thought-world and to the anchoring of his views. Substitutions of peoples, countries, persons, etc., which are represented by the examples in the Tab. 1, seem to be the most interpretatively productive.

in the source document	**in the manuscript**
Frankreich 'France'	*Ungarn* 'Royal Hungary'
Frankreich 'France'	*Magyaren* 'Hungarians'
Griechen 'Greece'	*Magyaren* 'Hungarians'

althellenischen ‘Old Greek’	*altmagyarisch* ‘Old Hungarian’
Griechenland ‘Greece’	*Magyarenland* ‘Hungarian country’
der reformirten Schweiz ‘reformed Switzerland’	*Lande* (Ungarn) ‘country’ (Royal Hungary)
radical ‘radical’	*magyarisch* ‘Hungarian’
Syrien ‘Syria’	*Ungarn* ‘Royal Hungary’
deutsch ‘German’	*slawisch* ‘Slavic’
spanisch ‘Spanish’	*slawisch* ‘Slavic’
deutsche Volkstamme ‘German tribes’	*Herren [?] Ungarns* ‘lords of Hungary’
Franzose ‘French’	*Magyar* ‘Hungarian’
Alianz mit Frankreich in Spanien ‘Aliance with France in Spain’	*Ungarn* ‘Royal Hungary’
Frankreich ‘France’	*Magyaren* ‘Hungarians’
Spanien ‘Spain’	*Slaven* ‘Slavs’
national-spanisch ‘national-Spanish’	*national-slavisch* ‘national-Slavic’
französisch Staatsmechanismus ‘French state mechanism’	*Magyarismus* ‘Hungarianism’
Demoiselle Fanny Eisler	*N.N.* (here apparently Károly Zay)
Lamennais	*K.* (here apparently Lajos Kossuth)
Mensch ‘human’	*Glaubensgenosse* ‘fellow believer’
diese und hene ‘this and that’	*die Magyaren, die Slawen* ‘the Magyars, the Slavs’
Conventikel, ‘Conventicle’, *Mysticismus* ‘mysticism’	*Magyarisirung* ‘Hungarianization’
Tendenzen, die jetzt mit aller Kraft die christliche Bevölkerung des türkischen Reichs ‘Tendencies which are now vigorously attacking the Christian population of the Turkish empire’	*Aufsatz über den Panslawismus* ‘essay on Panslavism’
gewissen Ständen ‘certain estates’	*Dominicanern* ‘Dominicans’
Eine diesen Behörden ‘one of these authorities’	*Dominicanern* ‘Dominicans’

Tab. 1. Selected substitutions revealed by text matching

The above substitutions are easily detectable through an automatic comparison of the texts.[12] In interpreting them, we proceeded on the assumption that the expressions present in the author’s manuscript are more revealing of his thought-world than the original expressions in pre-texts. Indeed, he often selected compelling formulations (regardless of their thematic load) and inserted entities for which he

[12] The limitations of this approach are shown by the correspondence of the content of the introductory passage of Ferjenčík’s later publication (Ferjenčík 1846, p. 214) with a part of his manuscript (Ferjenčík s. a., p. 1) containing an excerpt from Heinrich Heine (Heine s. a., p. 212). While the manuscript and its draft are written in German, the 1846 article, in which the substitution of the *Book of Baruch* for *Slovenskje národňje novini* and of the present for the homeland occurs, is written in contemporary vernacular Czech.

sought a persuasively powerful statement. In certain cases, however, the substitutions may be indicative of negative or positive connotations that were associated in his thought-world with particular countries, nations, etc.: *Frankreich* 'France', *Franzose* 'French', *französisch* 'French', or *Spanien* 'Spain', *spanisch* 'Spanish'. In addition to the significant adaptations, such as those documented in the Tab. 1, we have noted several less significant changes, notably various simplifications of the source texts, which indicate the omission of what was not essential for the author (in terms of the considerations, parallels, or analogies being pursued). However, abbreviations such as *evang[elisch]* 'evangelical', *Xten* instead of *Christen* 'Christians'; *Xtenthum*; *Xtentum* instead of *Christenthum* 'Christianity', *Xtenheit* instead of *Christenheit* 'Christendom; Christianitas' also occur in the manuscript, which are related to the high frequency of the expressions associated with their stable place in the author's thought-world and are therefore worthy of interpretive attention.

4 CORPUS ANALYSIS

We created a corpus of Ferjenčík's notebook in Sketch Engine, in which we detected the most frequently used words, the frequency of occurrence of selected terms, and their collocability. Considering the thematic parameters of the identified substitutions and contextual transpositions (see Tab. 1), we focused on expressions that reveal the author's attitudes towards the national relations of Hungarians and Slavs (Slovaks) in Hungary. Alongside these, the manuscript also frequently thematised ecclesiastical conditions and inter-confessional (Protestant–Catholic) relations, the reflection of which on the surface of the monument deserves more detailed research.

The corpus of Samuel Ferjenčík's manuscript notebook was created through the digitisation and transcription of the original manuscript. It consists of a single document, which contains 45,549 tokens, 37,290 words, and 1,016 sentences.

Using the tools Wordlist, Wordlist noun and Wordlist adjective, we identified the most frequent words, and also specifically nouns and adjectives. The first 20 most frequent adjectives include, in addition to context- and topic-unspecific adjectives, the adjectives *magyarisch* 'Hungarian', *slawisch*/*slavisch* 'Slavic', *deutsch* 'German', *ungarisch* 'Hungarian', *christlich* 'Christian', *protestantisch* 'protestant', and *religiös* 'religious'. Among the most frequent nouns, a number of words are related to national, historical, religious and ecclesiastical themes: 1. *Kirche* 'church', 2. *Zeit* 'time', 3. *Volk* 'nation/folk', 4. *Sprache* 'language', 5. *Ungarn* 'Hungary', 6. *Geist* 'spirit', 5. *Mensch* 'people', 6. *Baum* 'tree', 7. *Gott* 'god', 8. *Magyar* 'Hungarian', 9. *Erde* 'earth'', 10. *Glaube* 'faith', 11. *Land* 'country', 12. *Leben* 'life', 13. *Slaven* 'Slavs', 14. *Jahr* 'year', 15. *Welt* 'world', 16. *Mann* 'man', 17. *Slawe* 'Slav', 18. *Gemeinde* 'community', 19. *Nationalität* 'nationality', and 20. *Wort* 'word'.[13]

[13] In several cases, the words reflect the author's interest in pomology.

We compared the results of using the Wordlist tool (after abstraction from synsemantic words) with the output of the Keywords single-word tool (*magyarisch* 'Hungarian', *magyar* 'Hungarian', *Slaven* 'Slavs', *slavisch* 'Slavic', *slawisch* 'Slavic', *magyarisiren* 'magyarization', *magyarismus* 'magyarismus', *magyarisirung* 'magyarization', *dominicaner* 'Dominican', as well as some proper names, numeralia and demonstrative pronouns) and with the phrases extracted by the Keywords multi-word term tool (*magyarische Sprache* 'Hungarian language', *protestantische Kirche* 'Protestant church', *slavisches Volk* 'Slavic folk', *slawische Sprache* 'Slavic language', *slavische Gemeinde* 'Slavic community', *ungarische Nationalität* 'Hungarian nationality', *gemeinsames Vaterland* 'common fatherland', *nationale Bestrebung* 'national efforts', *griechische Kirche* 'Greek church', *kirchliche Angelegenheit* 'ecclesiastical matter', *ungarische Sprache* 'Hungarian language'). On the basis of formal and semantic correlation, we identified four main thematic groups represented in the text. These are national, historical, religious and ecclesiastical themes, which the author dealt with repeatedly in his other works. It turned out that the identified thematically significant words largely matched the keywords revealed by the intertextual research aimed at detecting manipulations of the source text. The word frequency shows, that state political problems, national issues (especially in relation to the Hungarian, Slovak and generally Slavic community), as well as confessional issues (related to Protestantism, Catholicism and church life) had a prominent place in the author's thought world. Another relevant topic is pomology.

Next, we focused on the co-occurrences of the identified nouns and adjectives, given the assumption that they might be indicative of the author's evaluative attitudes towards nations and nationalities. Among the names of nationalities, *Slawen/Slaven* is used mainly in contexts in which individual groups of Slavs are identified from among the surrounding non-Slavs or other Slavic inhabitants (*türkische* 'Turkish', *graubärtig* 'grey bearded', *protestantische* 'protestant', *geborene* 'nee', *gebildeten* 'educated' *Slawen*); Slavs as a whole are referred to only sporadically.[14] Through the Word Sketch Difference function, we found that the adjective *slawisch/slavisch* 'Slavic' is mainly associated with words such as *Gemeinde* 'parish', *Lande* 'country', *Theilen* 'parts', *Katheder* 'cathedral', *Anstalt* 'institution', and *Nachbar* 'neighbour' are mainly associated with the adjective *slavisch/slawisch*, *Völkerstamm* 'tribe', *Litteratur* 'literature', *Bevöl-kerung* 'population', *Gottesdienst* 'worship', *Adel* 'nobility', *Grundlage* 'foundation', *Kind* 'child', *Herz* 'heart', *Nationalität*

[14] *Hier muß aber das Geständniß abgelegt werden, daß die bittersten Slavenfeinde, die grimmigsten Magyaromanen, abtrünnige Slaven sind, und daß die Mehrzahl der katholischen Magyaren, weit entfernt von dieser Wuth des jungen Magyarenthums, in den mild menschlichen Geleise, daß die Regierung vorgezeichnet einzulenken. wünscht* 'Here, however, the confession must be made that the bitterest enemies of the Slavs, the fiercest Magyaromaniacs, are apostate Slavs, and that the majority of Catholic Magyars, far from this rage of young Magyarism, wish to turn into the mildly human path.'

'nationality', *Liturgie* 'liturgy', resp. *Geist* 'spirit', *Wort* 'word', *Angelegenheit* 'matter', *Sitte* 'custom', *Grund* 'reason', *Kraft* 'force', *Volksbildung* 'popular education', *Gebrauch* 'use', *Stadt* 'city', *Volkslehrer* 'popular teacher', *Kleidung* 'clothing', *Volksstamm* 'tribe', *Milde* 'mildness', *Überrest* 'remnant', *Propaganda* 'propaganda', *Universalmonarchie* 'universal monarchy', *Benennung* 'naming', *Modulation* 'modulation', *Nation* 'nation'. The following units are particularly associated with the adjective *magyarisch*: *Sprache* 'language', *Journalistik* 'journalism', *Predigten* 'sermons', *Zeitungen* 'newspapers', *Aussprache* 'pronunciation', *Vocabeln* 'vocabularies', *Comitate* 'county', *Übertreibungen* 'exaggerations', *Superstition* 'superstition', *Litteratur* 'literature', *Publicum* 'audience', *Zeitschriften* 'journals', *Rednern* 'speakers', *Gottesdienst* 'worship', *Adel* 'nobility', *Patrioten* 'patriots', *nationalen Fanatismus* 'national fanaticism', *Journale* 'journals', *Zweig* 'branch', *Wesen* 'being', *Geistlichen* 'clergy', *Nationalität* 'nationality', *Kirche* 'church'.

Phrases with the variants *Slawe/Slave* (73 occurrences in total) or *slavisch/slawisch* (the same 73 occurrences in total) refer to two levels of society: a) meso-level, e.g. *slawische/slavische Gemeinden* 'community'; b) macro-level *Städte* 'cities', *Volk* 'folk', *Nation* 'nation'. At the individual meso-levels, certain groups of persons (Hungarianised Slavic children, Slavic teachers learning Hungarian), institutions (Slavic departments), etc. are also mentioned in the text, such as *Slawische propaganda* 'Slavic propaganda' or *Universalmonarchie* 'universal monarchy'. The words *Ungar* 'Hungarian' (10 occurrences) and *Magyar* 'Hungarian' (49 occurrences) are used mainly as contextual antonyms to *Slawe/Slave* 'Slavs/Slav'; however, elsewhere they function as cohyponyms, e.g. *Ungarn von Magyaren, Slaven und Deutschen bewohnt* 'Hungary inhabited by Magyars, Slavs, and Germans'. The noun Magyar refers especially to Catholic or zealous Hungarians (*katholische Magyaren*; *eifrige Magyaren*), to Hungarians with specific characteristics such as inconsiderateness (*Zumutung*), playfulness (*Übermuth*), haste (*Eile*), vanity (*Eitelkeit*), zeal (*Eifer*), interests (*Interesse*), and language (*Sprache*). The attributes *bidergesinnte* 'biederminded', *gebildete* 'educated', *schlechte* 'bad', *alte* 'old' are associated with the noun *Ungar* 'Hungarian'. Hungary is referred to by two terms, namely *Ungarn* (77x), but occasionally by the semantically specific *Magyarenland* 'country of Magyars, i.e. the asserted form of Hungary as a Hungarian nation-state'. The noun *Ungarn* 'Hungary' is mainly used in geographical, political, and historical contexts. In two places the author singles out a group of Magyaromaniacs (*Magyaromanen*) whom he criticizes for denouncing the renegades and denying the need to learn the Slavic (Slovak) language, while condemning the people who learn it.

The corpus research did not confirm the assumption that the co-ocurrences reflect the author's evaluative attitudes towards nationalities. In the case of Hungarians, however, typical characteristics are more often thematised, such as

recklessness, playfulness, hastiness, vanity, zeal (negative qualities), but also positive qualities as bravery and educatedness. In the case of the Slovaks or Slavs, these are not thematic characteristics, but their community is rather materially distinguished among the other nations of Europe and the multi-ethnic Hungary.

5 CONCLUSION

Samuel Ferjenčík's manuscript book of notes and extracts has been subjected to multi-level interdisciplinary research. The text of the monument was digitized and the Transkribus platform was used for its processing in addition to manual transcription. Furthermore, the reconstruction of the document's source base was performed, in which the detection of text matches in potential source documents through automated content similarity detection tools played a non-negligible role. In the next research perspective, we plan to focus on customising the corpora of these tools, which have the potential to significantly facilitate the research of intertextual relations. The same approach can be used in future research on other literary monuments and historical documents. In the analysis of the work, we relied on the results of textual comparison, which revealed substitutions and other interventions in the wording of the source texts. These interventions speak directly about the author's thought and value world and stimulate further research on the monument in the corresponding point of view. The corpus research of the monument through the Sketch Engine manager has created a starting point for further interpretive research of the monument through the quantitatively based identification of priority themes and issues and, in certain perspectives, evaluative attitudes towards them. The detection of textual similarities and the results of the corpus analysis do not replace in-depth reading of the work and its qualitative research, but they undoubtedly facilitate its textual processing (reconstruction of the source base) and form the basis for further interpretative research.

ACKNOWLEDGEMENTS

The paper is an output of the project VEGA 2/0136/21 Literary historical, cultural historical and editorial study of a manuscript of Samuel Ferjenčík's book of recordings.

References

Baggerman, A., and Dekker, R. (2018). Jacques Presser, Egodocuments and the Personal Turn in Historiography. The European Journal of Life Writing, vol. VII, pages 90–110.

Burns, P. J., Brofos, J. A., Li, K., Chaudhuri, P., and Dexter, J. P. (2021). Profiling of Intertextuality in Latin Literature Using Word Embeddings. In NAACL: Proceedings of the

2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4900–4907. Accessible at: https://aclanthology.org/2021.naacl-main.389.

Chollier, C. (2014). Textual Semantics and Literature: Corpus, Texts, Translation. Signata. Annales des sémiotiques / Annals of Semiotics [Online], vol. 5, pages 77–99.

Dekker, R. (2002). Introduction. In R. Dekker (ed.): Egodocuments and History: Autobiographical Writing in its Context since the Middle Ages. Hilversum: Verloren, pages 7–20.

Depkat, V. (2019). 2.8 Ego-documents. In M. Wagner-Egelhaaf (ed.): Handbook of Autobiography/Autofiction. Volume I: Theory and Concepts. Berlin/Boston: Walter de Gruyter, pages 262–267.

Ferjenčík, S. (1846). XXVIII. Slowo lásky ku Břetislawským Nowinám. In Hlasowé o potřebě jednoty spisowného jazyka pro Čechy, Morawany a Slowáky. Nákladem Českého museum. W Praze: W kommissí u Kronbergra i Řiwnáče, pages 214–216. München, Bayerische Staatsbibliothek – L.rel. 2632 x.

Ferjenčík, S. (s. a.). Poznámová kniha Samuela Ferjenčíka, manuscript, 155 p. Archive of the Congregation of the Evangelical Church of the Augsburg Confession in Jelšava, Slovakia.

Heine, H. (s. a.). Paris, den 1. October 1840. In: Heinrich Heine's Sämmtliche Werke. Sechster Band: Französische Zustände. Berlin/Leipzig: Th. Knaur Nachf, pages 212–214. National Library of Israel, Google Books ID ojfNEsjFpWUC.

Hohl Trillini, R., and Quassdorf, S. (2010). A 'Key to All Quotations'? A Corpus-Based Parameter Model of Intertextuality. Literary and Linguistic Computing, 25(3), pages 269–286.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), pages 7–36.

Milička, J., Cvrček, V., and Lukeš, D. (2022). Unpacking lexical intertextuality: Vocabulary shared among texts. In M. Yamazaki – H. Sanada – R. Köhler – S. Embleton – R. Vulanović, R. – E. S. Wheeler (eds.): Quantitative Approaches to Universality and Individuality in Language. Berlin/Boston: Walter de Gruyter GmbH, pages 101–116.

Mao, Z. (2015). Intertextuality in institutional talks: a corpus-assisted study of interactions between spokespersons and journalists. Thesis. Birmingham: University of Birmingham, 260 p.

Presser, J. (1969). Uit het werk van J. Presser. Amsterdam: Athenaeum-Polak/Van Gennep, 332 p.

Presser, J. (1985). Memoires als geschiedbron. In Winkler Prins Encyclopedie VIII. Amsterdam: Elsevier, pages 208–210.

Stubbs, M. (2015). Computer-Assisted Methods of Analyzing Textual and Intertextual Competence. In D. Tannen – H. E. Hamilton – D. Schiffrin (eds.): The Handbook of Discourse Analysis. John Wiley & Sons, pages 486–504.

Teubert, W. (2010). Meaning, discourse and society. Cambridge: Cambridge University Press, 300 p.

Visser, J., Duthie, R., Lawrence, J., and Reed, C. (2018). Intertextual correspondence for integrating corpora. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA), pages 3511–3517, Miyazaki, Japan.

sciendo
DOI 10.2478/jazcas-2023-0047

LEXICAL DIVERSITY AND LANGUAGE IMPAIRMENT

NATALIIA ČASNOCHOVÁ ZOZUK
Department of Informatics, Faculty of Natural Sciences and Informatics,
Constantine the Philosopher University in Nitra, Nitra, Slovakia

ČASNOCHOVÁ ZOZUK, Nataliia: Lexical Diversity and Language Impairment. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 301 – 309.

Abstract: The development of artificial intelligence tools has seen an enormous growth recently. Linguistic artificial intelligence tools are being successfully applied in the field of speech analysis and discourse. In our study, we used automatic NLP tools to detect differences in picture description in the discourse of people diagnosed with Alzheimer's disease (AD), Mild Cognitive Impairment (MCI) and healthy people. A measure of lexical diversity was used to compare discourse complexity. Transcripts of recordings of the probands within the EWA project were used in the study. From the multiple comparisons, we found that there is a statistically significant difference between healthy people and people suffering from MCI and AD. Our results indicate that healthy people have more lexical diversity than people suffering from MCI and AD – a more diverse vocabulary in spontaneous speech, in our case, when describing a picture.

Keywords: natural language processing, Alzheimer's disease, spontaneous speech, picture description, lexical diversity

1 INTRODUCTION

Nowadays, there is an increasing incidence of civilizational diseases in society affecting the activity of the brain and its cognitive functions, including language and speech. These diseases are included under the collective name – neurodegenerative diseases (Buckner 2004). One of the most serious is Alzheimer's disease (AD). The sooner the diagnosis is made, the sooner methods and means (medicines, therapies) can be applied, which can slow down or even stop further worsening of the condition (Klimova et al. 2015). To determine the correct diagnosis, financially demanding invasive methods are often used, such as MRI examinations or cerebrospinal fluid punctures. However, the symptoms of neurodegenerative diseases are also manifested in the manner and quality of speech of the person impaired, which can be detected by non-invasive methods.

Speech impairment in Alzheimer's disease primarily occurs as a result of a decline in the semantic and pragmatic level of language processing (Ferris – Farlow 2013). Based on the decline in the level of language processing, several language-oriented research methods have been developed to assess the language deficits of people with AD.

One of these methods is the description of a picture that contains several topics (Mueller et al. 2018). A person describes not only the objects and activities, but also emotional states and social ties. Tasks such as describing pictures are often used in scientific research (Lindsay-Troeger – Koenig 2021, Szatloczki et al. 2015). Lindsay et al. (2021) used natural language processing (NLP) methods to extract specific semantic, syntactic, and other linguistic features in healthy people and people with AD, and based on the difference in parameters (language features), they trained a model that classified healthy and impaired people with AD. Frase et al. (2016) used linguistic features to identify AD in narrative speech. They showed some accuracy in the automatic identification of Alzheimer's disease from short speech discourses that were created during the picture description task and revealed significant linguistic features of the speech of healthy and impaired individuals. Jarrold et al. (2014) evaluated the ability of a trained classifier to diagnose dementia subtypes based on spontaneous speech. The findings of Ahmed et al. (2013) indicate that the level of lexical and semantic content and syntactic complexity of the language and speech best describe or reveal the degree of language impairment.

The aim of our study is to compare healthy people and people suffering from MCI and AD based on the lexical diversity of their spontaneous speech discourse when describing a picture.

The study is divided as follows: the following subsections briefly describe the state of the art of the examined issue in Slovakia and define the concept of lexical diversity. The second section is devoted to the research itself, in which we describe methods and procedures. In the third section, we present the results. The research findings are interpreted within the discussion and the final section contains the conclusion of the study.

1.1 Alzheimer's disease research in Slovakia

The EWA[1] (Early Warning of Alzheimer's) research project has been implemented in Slovakia since 2020, the aim of which is to develop a mobile application that would be able, from a person's speech, to detect the presence of early AD symptoms and other neurodegenerative diseases such as Mild Cognitive Impairment (MCI), Parkinson's disease and others. The MCI is the first stage of an incipient neurodegenerative disease, where roughly 25% of cases transform into AD within 5 years. The symptoms of MCI distinguish healthy people from people who evince symptoms of some cognitive problems.

In the EWA project, two types of tasks involving the description of pictures are used to record human speech. In the first type of task, the focus is on appellation of objects or activities that are shown in the picture displayed on a mobile phone. A person has to name what she/he sees using one single word. In the second task, the focus is on a more complex picture description, i.e., the picture contains more

[1] https://www.projektewa.sk/

persons, activities, objects, and relationships. Her/his task is to describe the whole scene of the picture in as much detail as possible. As part of the project, the participants described 65 different images, while over a thousand healthy and over two hundred diagnosed people were recorded, and several tens of thousands of recordings were obtained. However, in our study we will focus only on one more complex picture and its description by a smaller sample of participants.

1.2 Lexical diversity

When cognitive functions decline, language expression or speech discourse is simplified to so-called flat speech, in which linguistic complexity decreases. Complexity is a basic characteristic of a text, depending on many qualitative and quantitative parameters. The latter are the subject of NLP research, as we can determine and quantify them using automatic NLP tools. Within language complexity, we recognize grammatical and lexical complexity. One of the measures for assessing lexical complexity is called lexical diversity which is the subject of our study. From the beginning, lexical diversity (LD) was defined as a ratio of the type and token of the words TTR (Type Token Ratio) (Templin 1957; Johnson 1944), i.e., the total number of unique words (types) is divided by the total number of words (tokens). The closer this ratio is to 1, the greater the lexical diversity of the text. Basically, lexical diversity is the range of unique words used in a text or in speech relative to the overall range of the words in the given text or speech. A larger range corresponds to a higher diversity (Baese – Berk 2021; Durán 2004). This measure is also used as a measure of second language proficiency (Cumming et al. 2005) or vocabulary knowledge (Zareva et al. 2005; Yu 2010), but also as a warning signal or sign of the onset of Alzheimer‘s disease (Garrard et al. 2005; van Velzen – Garrard 2008).

Lexical diversity is calculated according to the following formula:

$$TTR = V/N,$$

where V is the number of unique words and N is the number of all words.

This measure has been proven to be the most suitable for the purposes of our research, because TTR is the most used index of the lexical diversity of a text (Hardie – McEnery 2006).

2 METHODOLOGY

We have no knowledge that similar research has been conducted in Slovakia, expect those mentioned previously. There exists neither research, nor study focusing on lexical complexity as an indicator for detecting neurodegenerative diseases. Therefore, it was necessary to determine which linguistic features (parameters) of language utterance will be investigated and also to define or select participants from the EWA project.

2.1 Participants and materials

For our research, we used the database of texts obtained in the EWA project. Although, Alzheimer‘s disease manifests itself mainly in the elderly population, in the

EWA project, the age of 50+ was chosen as an inclusion criterion. This is due to the fact that the project's task is to investigate early symptoms of the disease, which begin to manifest themselves even at a younger age. We divided our participants into three groups– diagnosed AD people, diagnosed MCI people, and healthy people. People diagnosed with AD and MCI were recruited for the project from specialised medical facilities. Healthy people were recruited through advertising media, magazine advertisements or retirement homes. The participants were informed about the purpose of the project and agreed to provide personal data and speech recordings for scientific purposes.

The inclusion criterion for demonstrating a cognitively healthy mind was the achievement of a specified score in the Montreal cognitive assessment (MoCA) test. Due to the correlation of the occurrence of AD with older age, the average age in the AD group was up to 78 years, while in the group of healthy persons it was only 65 years. A decline in cognitive functions is a natural accompanying phenomenon of human ageing. In order to assess the symptoms of the disease independently of age, balanced groups with approximately the same age means were created. As a result, we included 44 people in the AD group, 57 people in the MCI group, and 204 people in the healthy group.

2.2 Instrument

We used a specific suitable tool from one of the libraries of the Python programming language for the texts obtained from the probands' spontaneous speeches. It was the Natural Language Toolkit (NLTK) library for tokenization, lemmatization, and other tasks related to natural language processing. The lemmatizer developed by LINDAT/CLARIN (the Czech national node of the pan-European research infrastructure CLARIN) with the slovak-snk-ud model was applied from this library. Statistical methods were applied to the obtained values to determine the significance of the differences found.

3 RESULTS

Based on the Mean as well as the Mean Rank (Tab. 1), the differences in lexical diversity between healthy people and people suffering from MCI and AD are visible below.

Diagnosis	N	LD Mean	LD Std.Dev.	LD Std.Err	LD -95,00%	LD 95,00%	LD Sum of Ranks	LD Mean Rank
AD	44	0.75	0.11	0.02	0.72	0.78	29742.00	675.95
MCI	57	0.71	0.11	0.01	0.68	0.73	32523.50	570.59
Healthy person	204	0.66	0.10	0.00	0.65	0.66	370649.50	447.10

Tab. 1. Lexical diversity – mean

In the case of the AD and healthy groups (Tab. 2), we identified significant deviations from normality based on the Shapiro-Wilk W test.

Diagnosis	N	W	p
AD	44	0.98	0.50
MCI	57	0.98	0.49
Healthy person	204	0.92	0.00

Tab. 2. Shapiro-Wilk W test – results

Due to deviations from normality, we will use the non-parametric Levene test (for homogeneity of variances) to test the equality of variances. We reject the null hypothesis of equality of variances stating that there is no statistically significant difference in the variances of the lexical diversity between the three examined groups (Tab. 3).

	MS Effect	MS Error	F	p
Lexical density	0.01101	0.00249	4.42221	0.01226

Tab. 3. Levene test – results

Due to the violation of the assumptions of normality and equality of variances, we use the Kruskal-Wallis test to test the global null hypothesis. Based on the results ($H(2, N = 930) = 39.622$, $p = 0.0000$) we reject the global null hypothesis at the significance level of 0.001, which claims that there is no statistically significant difference between the groups in lexical diversity. After rejecting the global hypothesis, we were interested in groups between which there exists a statistically significant difference. From the multiple comparisons (Multiple comparisons of mean ranks for all groups) we identified two homogeneous groups (MCI, AD) and (Healthy persons) as well as statistically significant differences between healthy persons and persons suffering from MCI and AD (Tab. 4).

Diagnosis	LD Mean	LD Mean Rank	1	2
Healthy p.	0.65808	447.10		****
MCI	0.70575	570.59	****	
AD	0.75083	675.95	****	
Note: **** - Homogenous Groups, $p > 0.05$				

Tab. 4. Multiple comparisons – results

4 DISCUSSION

Although speech impairment is a secondary symptom of AD, many studies (e.g. Bucks et al. 2000; Kavé – Goral 2016; Kavé – Goral 2018) have shown that the

decline in language skills occurs relatively early in people diagnosed with Alzheimer's disease and can serve as a sensitive indicator of the gravity and progression of the disease over time.

It has been shown that the level of lexical diversity is statistically significant for assessing the health of a person's cognitive abilities. In accordance with Kavé and Goral (2018), we also confirmed that the ratio of type and token, in our case, unique and all words, is significantly influenced by the total number of words in utterance. Previous studies (e.g. Bucks et al. 2000; Kavé – Goral 2016) have found that, in general, lexical diversity is lower within the utterance of AD sufferers than healthy people. However, this phenomenon was not specifically confirmed in our study. We believe it is caused by the diagnosed persons describing the picture very briefly. The average number of words used by people diagnosed with MCI was approximately 95 words, compared to only 50 words for those diagnosed with AD. Healthy people used an average of around 120 words, which is a statistically significant difference compared to AD people. It resulted in the finding that the lexical diversity of people diagnosed with AD or MCI is higher compared to healthy people. When using fewer words, the ratio of unique words to all words increases, pointing out that a higher value of lexical diversity in our case does not mean a more complex and rich expression. Here is an example of a picture and the transcription of discourse of the probands of each examined group (AD, MCI, and healthy people).

AD: *"no neviem prečo tam do toho klepe či búcha do toho svetla tam a ach je chlapček zase berie si z oného banán ale sa mu šmykla asi stolička neviem či nespadne tam tam je ešte nejaké.."* ***(37 slov zo 64 slov)***

[AD: 'Well, I don't know why he's knocking or banging on that light over there, and oh, there's the boy again, he's taking that banana from the other one, but maybe his chair slipped, I don't know if he will fall, there there's another one there...' (37 words out of 64 words)]

MCI: *"no v kuchyni decko nejaké tam niečo pustil vodu voda do drezu vyteká z drezu voda vonku vidím tu ďalšie na kraji ešte mačku nejakú mačičku a dotyčný pán rozbil bola buchol do svetla varechou a zase na kuchynskom pulte tam je nejaké nejaký hrniec tiež niečo vyteká vonku nejaká omáčka alebo také niečo..."* ***(54 slov zo 106 slov)***

[MCI: 'Well, in the kitchen, a child has poured water into the sink, water is flowing out of the sink, outside, I see another cat on the side, and the man in question broke it, hit the light with a cooking pot, and there is a pot on the kitchen counter, something is also leaking outside, some kind of sauce or something like that...' (54 words out of 106 words)]

Healthy person: *"chlapec stojí na stoličke naťahuje sa za banánom stolička sa mu prevracia asi padne z vodovodu tečie voda do umývadla vyteká von pozerá sa tam kocúr na to zboku otec má v ruke varechu zdvihol ju chcel trafiť muchu ale rozbil lampu ktorá je visiaca majú tam dve dve police jedna je otvorená polovica dverí sú tam priečky medzitým tam je fľaška ktorá…"* ***(63 slov zo 138 slov)***

[Healthy person: "a boy is standing on a chair, he is reaching for a banana, his chair is tipping over, he is about to fall from the water tap, water is flowing into the sink, it is flowing out, there is a cat looking at it from the side, the father has a cooking pot in his hand, he raised it, he wanted to hit a fly, but he broke the lamp that is hanging. there are two two shelves one half of the door is open there are partitions meanwhile there is a bottle which..." (63 words out of 138 words)]

5 CONCLUSION

Investigating the complexity of human speech may benefit the automatic detection of Alzheimer's Disease symptoms through speech pattern analysis. Differences at the lexical level between the speech of a person diagnosed with AD and the speech of a healthy person can be captured and quantified. However, it is necessary to know which lexical parameter is suitable for a specific task of speech analysis. It was evident that the lexical diversity of AD or MCI people is higher for a short speech utterance describing a picture, which, however, does not represent the richness of the speech utterance. This is an interesting and scientifically significant finding.

ACKNOWLEDGEMENTS

During the research work, we were able to use the data obtained within the EWA project, for which we are very grateful to the researchers of this project.

References

Ahmed, S., Haigh, A. M., Jager de, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain: a journal of neurology, 136(12), pages 3727–3737.

Baese-Berk, M. M., Drake, S., Foster, K., Lee, D., Staggs, C., and Wright, J. M. (2021). Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions. Front. Psychol. Vol. 12. Accesibble at: https://doi.org/10.3389/fpsyg.2021.661415.

Buckner, R. L. (2004). Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. Neuron, 44, pages 195–208.

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology, 14(1), pages 71–91.

Covington, M. A., and McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). Journal of Quantitative Linguistics, 17, pages 94–100.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., and Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. ETS Res. Rep. Ser. 2005, pages 1–77.

Durán, P., Malvern, D., Richards, B., and Chipere, N. (2004). Developmental trends in lexical diversity. Applied Linguistics, 25(2), pages 220–242.

Fergadiotis, G., Wright, H. H., and Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. Journal of Speech, Language, and Hearing Research, 58(3), pages 1–13.

Ferris, S., and Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. Clin. Interv. Aging 8, pages 1007–1014.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. Journal of Alzheimer's Disease, 49(2), pages 407–422.

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128, pages 250–260.

Hardie, A., and McEnery, T. (2006). Statistics. In K. Brown (ed.).: Encyclopedia of Language and Linguistics, 2nd edition. Amsterdam: Elsevier, pages 138–146.

Jarrold, W., Peintner, B., Wilkins D., Vergryi D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 27–37. Baltimore, Maryland, USA. Association for Computational Linguistics.

Kavé, G., and Dassa, A. (2018). Severity of Alzheimer's disease and language features in picture descriptions. Aphasiology, 32(1), pages 27–40.

Johnson, W. (1944). Studies in language behavior: a program of research. Psychol. Monogr., 56, pages 1–15.

Kavé, G., and Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology, 38(9), pages 958–966.

Klimova, B., Maresova, P., Valis, M., Hort, J., and Kuca, K. (2015). Alzheimer's disease and language impairments: social intervention and medical treatment. Clin. Interv. Aging, 10, pages 1401–1407.

Lindsay, H., Tröger, J., and König, A. (2021). Language Impairment in Alzheimer's Disease–Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. Front. Aging Neurosci. 13. Accesibble at: https://doi.org/10.3389/fnagi.2021.642033.

Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. J. Clin. Exp. Neuropsychol., 40(9), pages 917–939.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front. Aging Neurosci., 20(7). Accessible at: https://doi.org/10.3389/fnagi.2015.00195.

Templin, M. C. (1957). Certain Language Skills in Children; Their Development and Interrelationships. Minneapolis, MN. University of Minnesota Press.

Velzen van, M., and Garrard, P. (2008). From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. Interdiscipl. Sci. Rev., 33, pages 278–286.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. Appl. Linguist., 31, pages 236–259.

Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: variable sensitivity. Stud. Second Lang. Acquisit., 27, pages 567–595.

sciendo
DOI 10.2478/jazcas-2023-0048

TEXT VECTORIZATION TECHNIQUES BASED ON WORDNET

DÁVID DRŽÍK– KIRSTEN ŠTEFLOVIČ
Department of Informatics, Faculty of Natural Sciences and Informatics,
Constantine the Philosopher University in Nitra, Nitra, Slovakia

DRŽÍK, Dávid – ŠTEFLOVIČ, Kirsten: Text Vectorization Techniques Based on Wordnet. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 310 – 322.

Abstract: The utilization of text vectorization techniques has become essential for numerous classification tasks in present-day natural language processing. Word embedding methods commonly used today, such as Word2Vec, GloVe, etc., are based on the semantic similarity of words. WordNet, as a lexical database of words, provides a rich source of semantic information. In our article, we propose a text vectorization technique using extended text data with the data augmentation method, specifically by replacing words with their synonyms obtained from WordNet. The results obtained from text classification tasks using multiple classifiers demonstrate that expanding the corpus with this method leads to improved vector representations of words.

Keywords: word embedding, Word2Vec, Glove, synsets, text data augmentation, semantic similarity

1 INTRODUCTION

The use of text vectorization techniques is nowadays a necessity for many classification tasks in the field of natural language processing. Modern word embedding models such as Word2Vec, Doc2Vec, Glove, etc. are based on the semantic similarity of words. On the other hand, the semantic relations of individual words are also available in the form of synsets. Synsets capture information about words that are similar in meaning. Unlike embedding models, these relationships are not created automatically by a computational algorithm, but by human annotators.

In the article, we propose a text vectorization method based on synsets. This method will represent an improvement in the training of modern word embedding models. Semantically similar words will not be automatically inferred only from the corpus, but we will also use information about semantically similar words from synsets.

We hypothesize that incorporating synset information can help improve the quality of word vectors by capturing more semantic relationships between words. Synsets will thus provide models with additional context. However, the question how to include these synsets in word embedding models, remains. For individual word embedding models, there are already used libraries that ensure training as well

as the actual use of the created vectors. Our effort is to offer the simplest possible intervention in the process of training embedding models using the available corpus.

With our approach, we want to use information about synsets in the training stage of these models. Synsets will therefore be used to modify and especially supplement the corpus used for model training. In each sentence of the corpus, we will analyse its individual words. For each word of the selected sentence, we find synonyms using synsets. And we will add new sentences to the corpus, in which the analysed word will be replaced by a synonym. For example, for the sentence "Dad is cooking lunch." and the word "dad" we add the following sentences:

- Dad is cooking lunch.
- Daddy is cooking lunch.
- Papa is cooking lunch.
- Father is cooking lunch.
- Male parent is cooking lunch.

We apply a similar procedure for the other words „cooking" and „lunch".

There are several approaches to compare the success of the proposed method. Nazir et al. (2022) evaluate success using datasets wordsim-353 and Lexsim-999, which contain calculated similarities of selected pairs. Calculating the similarities between two words is often presented in educational examples focused on word vectors. However, from a practical point of view, word vectors are used in different practical tasks. Classification is one of them. For this reason, we will verify the suitability of our method by using the trained methods to create vectors in the classification task. Subsequently, we will evaluate the performance measures of the classification itself, i.e., we will not evaluate the word vectors themselves directly, but we will use them to solve the classification task and evaluate the success of the classification.

We will proceed according to the following methodology (Fig. 1):

1. Preprocessing of the corpus for training Word2Vec and GloVe models (corpus 1).
2. Creation of the second corpus using synsets (corpus 2).
3. Training Word2Vec (Skip-gram, CBOW) and GloVe with two separate corpora – in this way, 6 word embedding models will be created:
   - **Word2Vec Skip-gram (C1)** – Word2Vec Skip-gram trained using Corpus 1
   - **Word2Vec CBOW (C1)** – Word2Vec CBOW trained using Corpus 1
   - **GloVe (C1)** – GloVe trained using Corpus 1
   - **Word2Vec Skip-gram (C2)** – Word2Vec Skip-gram trained using Corpus 2
   - **Word2Vec CBOW (C2)** – Word2Vec CBOW trained using Corpus 2
   - **GloVe (C2)** – GloVe trained using Corpus 2.
4. Preprocessing of the dataset for the classification task.
5. Creation of input sequences of vectors to classifiers from the dataset.
6. Creating and verifying the quality of classification models using k-fold validation.
7. Identification and comparison of the performance of the created models.

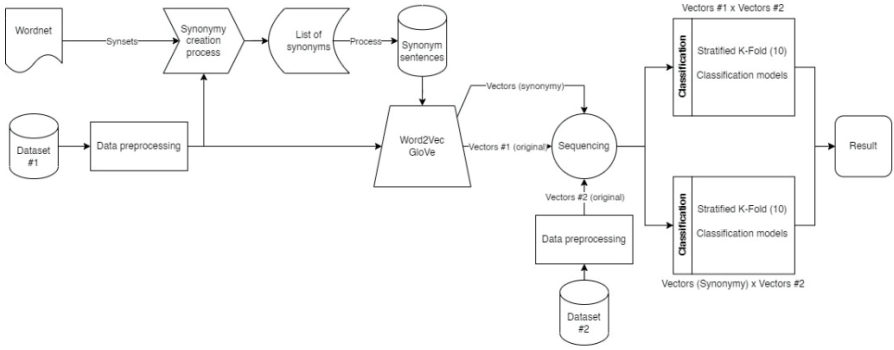

Fig. 1. Workflow of our proposed method

The article has the following structure. The second part summarizes the current state of research. The text corpus, but also the classification dataset and their preprocessing, the creation of the second corpus using synsets and the used word embedding models are described in the third part. The most important results are summarized in the fourth section. Discussion and conclusions form the content of the last part of the article.

2 RELATED WORK

Data augmentation is a technique used to increase the size and diversity of a training dataset by creating new examples from existing ones through various operations such as random rotations, flips, zooms, and translations (Pellicer – Ferreira – Costa 2023). The goal of data augmentation is to enhance the performance and robustness of machine learning models by exposing them to a wider range of variations and situations. Data augmentation has been successfully applied in various domains such as computer vision, natural language processing, and speech recognition. Moreover, data augmentation can also help mitigate overfitting by introducing more variation into the training data and preventing the model from memorizing the training examples.

There are few popular techniques for text data augmentation:

- **Back Translation**: Translating sentences to a foreign language and then back to the original language to create new variations of the original text.
- **Synonym replacement**: Replacing words in a sentence with their synonyms to increase the diversity of training data.
- **Paraphrasing**: Restating the same idea using different words and structures to generate new variations of a sentence or passage.
- **Random Insertion**: Adding randomly chosen words to a sentence to create new variations and increase the length of the sentence.

- **Random Swap**: Swapping two words in a sentence to create new variations while preserving the overall structure of the sentence.
- **Random Deletion**: Removing randomly chosen words from a sentence to create new variations and increase the focus on the most important words.

Data augmentation techniques, such as easy data augmentation (EDA), have been proposed to boost the performance of text classification tasks Wei et al. (2019). EDA consists of four of these simple techniques including synonym replacement, random insertion, random swap, and random deletion. Through experiments on five text classification tasks, EDA has shown to improve performance for both convolutional and recurrent neural networks. EDA also demonstrated particularly strong results for smaller datasets. On the other hand, Haralabopoulos et al. (2021) propose reverse classification of each augmented example on antonyms and negation.

Traditional synonym recommendations often include ill-suited suggestions for the writer's specific contexts. To address this issue, Glenski et al. (2021) proposed a simple approach for contextual synonym recommendation. This method combines existing human-curated thesauri, such as WordNet (Miller 1995), with pre-trained language models. Wang et al. (2015) is using the same thesauri to enhance computational behavioral analysis using social media text. To further improve the effectiveness of data augmentation through synonym replacement, Kobayashi (2018) proposes a novel method called contextual augmentation using BERT transformer (2019).

Marivate et al. (2020) explores different text augmentation methods on three datasets. The goal is to provide insights on making choices for classification use cases. Word2Vec-based augmentation works well without a formal synonym model. Mixup further improves performance and reduces overfitting.

3 MATERIALS AND METHODS

3.1 Datasets and their preprocessing

In our article, we used two freely available data files. The first text dataset or corpus (Risdal 2016) contains 143,000 English articles from 15 American journals. Before using this text, it is necessary to preprocess it as precisely as possible. We started by tokenizing the text into sentences using the NLTK library (Bird – Klein – Loper 2009). This dataset contains a total of 704,357 sentences. Subsequently, we converted the text in each sentence to lowercase letters, cleaned the text from hyperlinks, from white characters and from special characters that are not numbers or letters. We split the sentences into words. We then removed words that contained numbers as well as numbers themselves, ensuring that only words with linguistic meaning appeared in the dictionary. We have also removed stop words that appear too often in the text but have no meaning in themselves. And the last step of text

preprocessing was lemmatization using WordNetLemmatizer, which reduced the number of words in the dictionary, since they can appear in the dictionary in different forms, but carry the same meaning. For further work, we did not use all this text, but only a part of it, namely the first 10,000 sentences.

The second dataset, WELFake (Verma – Agrawal – Amorim – Prodan 2021), contains a total of 72,134 articles correctly labelled as fake news (35,028) and real news (37,106). We also preprocessed this file in the same way as the first one. We will use this classification dataset to verify the quality of the vector model.

3.2 Creating a list of synonyms from wordnet synsets

In this article, we want to experiment with word vectors and see if replacing words with their synonyms improves their quality. Therefore, we need to create a list of synonyms for the words in our examined text.

To create a list of synonyms, we used the sentences from the first data set, because in this text we will replace words with their synonyms, so it does not make sense to look for synonyms for words that we will not need later. The process of creating synonyms was as follows. We gradually went through the individual words of all preprocessed sentences. Then, for each word, we searched for synsets (list of synonyms) using the function synsets through the WordNet library (Miller 1995) from the NLTK library (Bird – Klein – Loper 2009). If no synonyms were found for the specified word in the WordNet dictionary, we tried to lemmatize the word again, this time using the WordNet morphy function, which tries to find its lemmatized form, which is listed in the WordNet dictionary. If the word is found, a list of synonyms for the modified word is searched.

After finding a synset for our word, we went through each synonym in turn and filtered the list based on semantic similarity (by using path similarity, result value is between 0 and 1, where a value of 1 indicates maximum similarity) to our word. If this similarity was higher than 0.5, we kept this synonym, otherwise we removed this synonym from the synset. In the end, we lemmatized all valid synonyms and stored a list with a maximum length of 5 synonyms per word. We saved the list as a binary file using the Pickle library (Rossum 1999).

3.3 Replacing words in the original sentences with their synonyms

After creating a list of synonyms (a word and a list of synonyms for this word), of which there were 9,704, we decided to gradually replace each word for which we register a list of synonyms in all sentences. We used a technique known as Data Augmentation – replacement by synonyms (Wei – Zou 2019). In our experiment, we go through all words in all sentences and check for each word whether it contains a list of synonyms. If so, then we will duplicate the whole sentence and replace the one specific word with each synonym for this word. As a result, we will have several times more sentences than the original ones. Just to give an idea, from the original

10,000 sentences, we created 175,354 sentences with this method, which is more than 17 times.

3.4 Creation of word vectors

Word vectors is one of the ways to represent words using numbers. They are n-dimensional vectors of real numbers, which are placed in the vector space in such a way that they capture the meaning or relationships between individual words. In our article, we used word models Word2Vec and GloVe to create word vectors. Using the Gensim (Řehurek – Sojka 2010) and GloVe (Pennington – Socher – Manning 2014) libraries, we implemented these 3 vector models in Python by creating the corresponding models.

Word2Vec

The Word2Vec model works on the principle that the very meaning of a word is represented by its context, i.e., the words that come before and after it. As a result, this model creates one multidimensional vector for each word. Syntactic or semantic relations between words are preserved, and the vector distances of individual words correspond to the human idea of the relation between words (Wei – Zou 2019; Mikolov – Chen – Corrado – Dean 2013).

The creators of the Word2Vec model themselves state two basic architectures: CBOW (Continuous Bag-of-Words) and Skip-gram. CBOW focuses on predicting a target word based on its context words within a context window, while Skip-gram focuses on predicting context words based on the target word within a context window. Both architectures are actually just simple neural networks with input, hidden and output layers. The result of the network training is not the output of the output layer, but the weights of the hidden layer, which represent the word vectors. A Softmax activation function is used for the output layer, but Word2Vec also uses algorithmic improvements such as hierarchical Softmax and negative sampling to reduce the computational complexity (Řehurek – Sojka 2010; Mikolov – Chen – Corrado – Dean 2013).

In Python, we used the Word2Vec model implementation using the Gensim library (Řehurek – Sojka 2010). The input to the model was preprocessed sentences, we set the size of the dimension to 150 and the size of the context window to 7. The reason for choosing these values was our previous research, where we had been investigating the appropriate setting of these parameters. We used both the skip-gram and CBOW architectures, so we created two models for unedited sentences and two models for edited sentences (words in sentences replaced synonyms). By training the model defined in this way, we obtain word vectors for all available words. We saved the individual word vectors and their corresponding words using the Gensim library (Řehurek – Sojka 2010) and the KeyedVectors class in the form of a complex dictionary in a binary file.

GloVe (Global Vectors)

The GloVe (Global Vectors) model is another important model for creating word vectors, which is similar to the Word2Vec model. Its advantage is that it can consider the statistical occurrences of words in the context in which they appear, thus capturing finer semantic relationships between words. The GloVe model is based on matrix factorization of the word context matrix. That is, first a large matrix is created that contains information about the common occurrence of words in context. For each word, its frequency of occurrence in a context in the corpus is calculated. Then this matrix is factorized to obtain a lower-dimensional matrix (Pennington – Socher – Manning 2014). We also implemented the GloVe model in Python using the GloVe library (Pennington – Socher – Manning 2014), with all parameter settings identical to those of the Word2Vec model, except for a learning_ rate of 0.05 and a number of iterations of 30.

3.5 Verification of the quality of the vector model on the classification task

It is generally stated that intrinsic and extrinsic evaluations are used to evaluate the performance of language models. The basic difference between them is that with Intrinsic evaluations, the performance of the system is evaluated directly on the specific task for which it was designed. For example, if a language model is designed to generate coherent sentences, such an evaluation would include measuring the coherence of the sentences generated. On the other hand, extrinsic evaluation involves evaluating the system's performance on a subsequent task for which the system was not specifically designed. For example, if a language model is designed to generate coherent sentences, extrinsic evaluation would involve measuring the model's ability to improve the performance of a downstream task such as machine translation or sentiment analysis (Zhai – Tan – Choi 2016; Shi – Zheng – Guo – Zhu – Qu 2018).

Since in our article we want to find out whether word vectors created on modified sentences (words replaced by their synonyms) or on original sentences are of better quality, we will use extrinsic evaluation. In particular, we will solve the classification of fake news using our created word vectors and we will compare how well we can classify fake news with vectors created over the original and modified sentences.

3.6 Creating sequences of word vectors

Even before creating the sequences of word vectors, we need to modify the classification dataset of fake news. We have already done the basic preprocessing, but this dataset also contains words for which we have not created word vectors. Such words will not help us in the classification of fake news and thus also in verifying the correctness of our word vectors, so we will remove them. After the removal, we noticed that some records remained completely empty of content, or the

number of words in the records decreased. To improve the quality of our results, we removed 5% of extreme entries from the top and bottom based on the number of words in each entry.

Another problem we had to solve was that the number of words in each record was not the same. Some entries had few words and some had too many words. Since we will want to create sequences of word vectors for each word in the record, and one vector has dimension 150, with very many words we would end up with a sequence of huge dimension. Therefore, we decided to take only the first 10 words from each record to classify fake news.

Creating the sequences was already easy because we went through the individual records and replaced the words in them with their word vectors. The result is a list of sequences, where one sequence consists of ten vectors of dimension 150.

3.7 Classification of fake news

Since our classification dataset contains too many records, we decided to continue working only with the first 15,000 records. And since we wanted to have a balanced dataset, in terms of the balance of false and true message classes, we used the undersampling method to balance the dataset by randomly reducing the number of entries from the larger class to the number of entries from the smaller class. As a result, we will work with 7,265 records for each of the classes. Thus, the input to the classification model will be a list of sequences of word vectors and the corresponding indication of whether it is a false message or not.

To classify fake news, we use several different classifiers from the Scikit-learn library (Shi – Zheng – Guo – Zhu – Qu 2018), which we will introduce in the next section. For classification, we used the Stratified K-fold cross-validation method, which works on the principle that the entire set of data is divided into the same number of K sets, one of which is used for testing and the rest are used for training the classification model, while the entire process on K-repeats times. Moreover, all sets will contain the same representation of individual classes.

In our evaluation we set k = 10, so the dataset was split into 10 subsamples. To assess the performance of the model in each part of the cross-validation, we used the basic performance measure metrics of the model: accuracy, precision, recall and F1 score. After obtaining the results from all parts of the cross-validation, we averaged the results of the individual metrics and thus obtained the average results of the correctness of the classification of fake news.

4 RESULTS AND DISCUSSION

After creating the Word2Vec and GloVe models, we used these models to create word vectors for fake news classification. Word2Vec and Glove were trained on a text corpus of English articles with a dimension of 150 and a window size of

7. According to the methodology presented in the Introduction section, we created classification models for identifying fake news based on word vectors. The classification models were created from the WELFake dataset presented in Section 3.1. The following classification algorithms were used to create classification models:

- Logistic Regression,
- SGD Classifier,
- K-Neighbours,
- BernoulliNB,
- SVC.

All algorithms are well known in the community. For the comparability of the results, we used them with basic settings without hyperparameter optimization. We used Stratified k-fold validation for evaluating our classification models. The evaluation metrics in our experiments is the classification accuracy. Accuracy is the ratio of correct predictions to the total number of samples and is computed as (1):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} ¿¿ \quad (1)$$

where TP represents the number of True Positive results, FP represents the number of False Positive results, TN represents the number of True Negative results, and FN represents the number of False Negative results. To evaluate, analyse and describe the results of our model we also calculated precision (2), recall (3) and F1 score (4) as follows:

$$precision = \frac{TP}{TP + FP} ¿¿ \quad (2)$$

$$recall = \frac{TP}{TP + FN} ¿¿ \quad (3)$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} ¿¿ \quad (4)$$

Precision is the ratio of correctly predicted positive observations of the total predicted positive observations. Recall shows the ratio of correctly predicted positive observations to all observations in the actual class. F1 score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

In our results, we present the average values of the stratified k-fold measurements. In the case of fake news classification models created from vectors generated by Word2Vec Skip-gram models (Tab. 1), better results can be seen in the accuracy metric for all used classification methods in the case of the C2 corpus. So it

means that we can achieve better results with our proposed method of preparing Word2Vec models using synsets (corpus C2). In the table (Tab. 1), for the sake of clarity, we have marked only pairs of performance measures in which we did not achieve an increase in the case of the C2 corpus. From the results, it can be seen that for most performance measures, the results improved in the case of applying our method (corpus C2). In the case of recall, the results worsened only for the K Neighbors classification method, in the case of precision, again for the BernoulliNB and SVC classification methods. Here it is interesting that the Precision results decreased, but the Recall values increased quite significantly. Also, for this reason, an increase in the results for the F1 score was also observed.

Classification method	Word2Vec Skip-gram (C1)				Word2Vec Skip-gram (C2)			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Logistic Regression	80.95	80.24	82.19	81.19	81.77	80.81	83.34	82.05
SGD Classifier	79.99	77.87	84.76	80.86	81.54	79.40	85.37	82.18
K Neighbours	72.94	74.50	**69.94**	**72.11**	75.15	83.84	**62.37**	**71.43**
BernoulliNB	79.37	**75.69**	86.65	80.78	80.91	**72.60**	99.34	83.88
SVC	81.99	**80.41**	84.66	82.47	84.91	**79.12**	94.88	86.28

Tab. 1. Average k-fold values for selected performance measures of models created using Word2Vec Skip-gram C1 and C2

In the case of models created over vectors from Word2Vec CBOW (Tab. 2), the average values of the Accuracy metric and also the F1 Score improved in all classification models. The results for the Precision and Recall metrics are mixed.

Classification method	Word2Vec CBOW (C1)				Word2Vec CBOW (C2)			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Logistic Regression	79.80	78.10	**82.90**	80.41	81.42	80.73	**82.57**	81.63
SGD Classifier	76.80	74.92	**83.81**	77.95	77.87	77.21	**79.19**	78.15
K Neighbours	67.77	**67.31**	66.36	67.31	70.74	**65.09**	89.90	75.44
BernoulliNB	74.14	**73.91**	74.72	74.29	80.67	**72.54**	98.75	83.64
SVC	75.81	74.10	79.42	76.66	84.20	80.35	90.59	85.15

Tab. 2. Average k-fold values for selected performance measures of models created using Word2Vec CBOW C1 and C2

Perhaps the best results in favor of our proposed method of corpus editing according to synsets were recorded for GloVe (Tab. 3). In addition to the average

Recall value for the SGDClassifier classifier, an improvement in results was observed in all classification methods in favor of GloVe for the C2 corpus. At the same time, the largest increase in the value of Accuracy was recorded here in all monitored models and classification methods (value increase by 12.53 for the BernoulliNB classifier, from 67.42 to 79.95). Also, the largest increase in the value of F1 Score was observed here in all monitored models and classification methods (increase in value by 14.96 for the BernoulliNB classifier, from 68.06 to 83.02).

Classification method	GloVe (C1)				GloVe (C2)			
	Acc.	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**
Logistic Regression	76.62	74.81	80.29	77.45	80.56	79.91	81.68	80.77
SGD Classifier	76.60	72.80	**85.81**	78.53	80.05	78.95	**82.00**	80.42
K Neighbors	66.11	66.99	63.57	65.22	70.36	73.46	64.46	68.29
BernoulliNB	67.42	66.80	69.40	68.06	79.95	72.03	97.98	83.02
SVC	73.75	71.78	78.34	74.90	83.90	79.64	91.11	84.98

Tab. 3. Average k-fold values for selected performance measures of models created using GloVe C1 and C2

By comparing the average values of the monitored performance measures, we pointed out the fact that by using our method it is possible to improve the Word2Vec and GloVe models. We evaluated their improvement for subsequent classification tasks. Our method significantly improved the GloVe model.

In addition to presenting our improvement of Word2Vec and GloVe models using synsets, in the article we also wanted to present a methodology for verifying the suitability of methods for improving these models. From the analysis of the available literature, we discovered the use of the methodology presented in the article (Nazir et al. 2022), by calculating the similarity of pairs of words. This approach is justified in the case of using the Word2Vec and GloVe models for the identification of semantically similar words. However, we think that in practical applications these models will be used mainly for the preparation of word vectors for classification tasks. For this reason, we wanted to evaluate our improvements to the Word2Vec and GloVe models based on improved performance measures in classification tasks.

5 CONCLUSION

New methods of representing word vectors (Word2Vec, GloVe) represent a significant leap forward in the ability to analyze relationships between words, sentences, and documents. Even though natural language processing is becoming something of a black box that usually works well, we have pointed out other possibilities for improving models for word embedding with our proposed method.

ACKNOWLEDGEMENTS

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and Slovak Academy of Sciences under Contract VEGA-1/0490/22, also supported by the Slovak Research and Development Agency under the contract no. APVV-18-0473.

References

Bird, S., Klein, E., and Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, 1st ed. Beijing; Cambridge Mass. O'Reilly Media.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, May 24, 2019. Accessed: Mar. 26, 2023. [Online]. Accessible at: http://arxiv.org/abs/1810.04805.

Glenski, M., Sealy, W. I., Miller, K., and Arendt, D. (2021). Improving Synonym Recommendation Using Sentence Context. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pages 74–78. Accessible at: doi 10.18653/v1/2021.sustainlp-1.9.

Haralabopoulos, G., Torres, M. T., Anagnostopoulos, I., and McAuley, D. (2021). Text data augmentations: Permutation, antonyms and negation. Expert Systems with Applications, Vol. 177. Accessible at: doi 10.1016/j.eswa.2021.114769.

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 452–457. New Orleans, Louisiana. Accessible at: doi 10.18653/v1/N18-2072.

Marivate, V., and Sefara, T. (20202). Improving Short Text Classification Through Global Augmentation Methods. In Machine Learning and Knowledge Extraction, pages 385–399. Accessible at: doi 10.1007/978-3-030-57321-8_21.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv. Accessible at: doi 10.48550/arXiv.1301.3781.

Miller, G. A. (1995). WordNet: A lexical database for English, Commun. ACM, 38(11), pages 39–41. Accessible at: doi 10.1145/219717.219748.

Nazir, S, Asif, M., Sahi, S. A., Ahmad, S., Ghadi, Y. Y., and Aziz, M. H. (2022). Toward the Development of Large-Scale Word Embedding for Low-Resourced Language. IEEE Access, Vol. 10, pages 54091–54097. Accessible at: doi 10.1109/ACCESS.2022.3173259.

Pellicer, L. F. A. O., Ferreira, T. M., and Costa, A. H. R. (2023). Data augmentation techniques in natural language processing. Applied Soft Computing, Vol. 132. Accessible at: doi 10.1016/j.asoc.2022.109803.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543. Doha, Qatar. Accessible at: doi 10.3115/v1/D14-1162.

Risdal, M. (2016). Getting Real about Fake News, Kaggle. Accessible at: doi 10.34740/KAGGLE/DSV/911.

Rossum, G. V. (1999). Python Library Reference. To Excel Inc.

Řehurek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, pages 45–50. Accessible at: doi 10.13140/2.1.2393.1847.

Shi, Y., Zheng, Y., Guo, K., Zhu, L., and Qu, Y. (2018). Intrinsic or Extrinsic Evaluation: An Overview of Word Embedding Evaluation. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pages 1255–1262. Singapore. Accessible at: doi 10.1109/ICDMW.2018.00179.

Verma, P. K., Agrawal, P., Amorim, I., and Prodan, R. (2021). WELFake: Word Embedding Over Linguistic Features for Fake News Detection. IEEE Trans. Comput. Soc. Syst., 8(4), pages 881–893. Accessible at: doi 10.1109/TCSS.2021.3068519.

Wang, W. Y., and Yang, D. (2015). That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557–2563. Lisbon, Portugal. Accessible at: doi 10.18653/v1/D15-1306.

Wei, J., and Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv, Aug. 25, 2019. Accessible at: doi 10.48550/arXiv.1901.11196.

Zhai, M., Tan, J., and Choi, J. (2016). Intrinsic and Extrinsic Evaluations of Word Embeddings. AAAI, 30(1). Accessible at: doi 10.1609/aaai.v30i1.9959.

sciendo
DOI 10.2478/jazcas-2023-0049

SLOVAK LANGUAGE MODELS FOR BASIC PREPROCESSING TASKS IN PYTHON

DANIEL HLÁDEK – MAROŠ HARAHUS
– JÁN STAŠ – MATÚŠ PLEVA
Faculty of Electrical Engineering and Informatics, Technical University,
Košice, Slovakia

HLÁDEK, Daniel – HARAHUS, Maroš – STAŠ, Ján – PLEVA, Matúš: Slovak Language Models for Basic Preprocessing Tasks in Python. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 323 – 332.

Abstract: We propose a Slovak language model for the spaCy library in Python. These models are easy-to-use for basic natural language processing tasks in a single package. The package contains several components for basic preprocessing tasks, such as tokenization, sentence boundary detection, syntactic parsing, lemmatization, named entity recognition, morphology analysis, and word vectors. It is based on the state-of-the-art monolingual SlovakBERT model. Named entity recognition is trained on a separate, publicly available WikiAnn database. The other statistical classifiers use a Slovak Dependency Treebank corpus. Morphological tags are compatible with the conventions of the Slovak National Corpus. The part of speech tags use conventions of the Universal Dependencies framework. We trained a separate word vector model on a web-based corpus. The training uses fastText with Floret modification. We present a series of experiments that confirm that the model performs similarly to other languages for all tasks. Training scripts and data are publicly available.

Keywords: Slovak language model, transformers, natural language processing

1 INTRODUCTION

Recently, several models and language resources have appeared for the Slovak language. Thanks to the work of the Slovak National Corpus, a corpus for determining parts of speech (PoS), lemmas (LEM), or dependencies (DEP) (Gajdošová 2016) is available. This dataset was later converted to the Universal Dependency format (Zeman 2017). There are several pre-trained models with the support for the Slovak language, based on the transformer architecture (Devlin 2018; Pikuliak et al. 2022). These new models achieve results comparable to those in other languages. We can say that the basic tasks of natural language processing, such as recognition of parts of speech, are solved and usable for Slovak.

Despite these advances, the available models are fragmented. There are separate models for each task, such as named entity recognition (NER) or part of speech tagging. Compiling a more complex sequence of preprocessing of the Slovak text is

still very laborious and requires detailed knowledge of the available sources. This complicates the development and research of more complex natural language processing tasks, such as dialogue systems, information retrieval, or automatic question answering.

Our models are trained transparently on publicly available data. The training scripts and model download links are available on GitHub.[1] This will allow the feedback collection, easy correction of possible errors, and gradual improvement in the event of the availability of better data or better modeling methods. We use state-of-the-art pretrained models which improve performance, even with smaller training data. This model follows our previous effort (Harahus 2022).

2 STATE OF THE ART

2.1 Available datasets

The most important thing for a statistical model is a good dataset. There are several corpora with morpho-syntactic annotations and named entities available for the Slovak language.

The Slovak Categorized News Corpus (Hládek 2014) is an automatically tagged collection of newspaper articles. This corpus has been meticulously categorized into distinct themes and topics, enabling its use in various natural language processing tasks, such as text classification, sentiment analysis, and topic modeling.

MULTEXT-East (Erjavec 2012) is a multilingual dataset based on the novel "1984" by George Orwell, designed for a variety of natural language processing tasks. It includes resources for several Central and Eastern European languages, such as Bulgarian, Czech, Estonian, Hungarian, Polish, Romanian, Serbian, Slovak, and Slovene. The corpus is annotated with part-of-speech tags, morphological information (MORPH), and lemmatization, making it a valuable resource for researchers and developers working on multilingual natural language processing tasks.

The Slovak Dependency Treebank (Gajdošová 2016) is a syntactically annotated corpus of the Slovak language. The corpus was prepared by the Slovak National Corpus at the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences and its annotation follows the guidelines of the Prague Dependency Treebank (PDT) (Hajič 2017) in Czech. The texts in the treebank come from different genres and domains, such as news articles, literature, and websites. The annotations in the treebank contain information on part-of-speech tags, morphological features, and dependency relations between words in a sentence. This corpus was later converted into Universal Dependencies by Daniel Zeman. The publication (Zeman

[1] https://github.com/hladek/spacy-skmodel

2017) describes the conversion of the syntactically annotated part of the Slovak National Corpus into the Universal Dependencies annotation scheme. The authors selected sentences for which two human annotators agreed 100% on the analysis. The data set is divided into a training set and a test set.

One of the most common datasets for NER training is **WikiAnn**. Cross-lingual Name Tagging and Linking for 282 Languages. Automatic extraction from Wikipedia (WikiAnn) is a large cross-lingual named entity recognition (NER) dataset derived from Wikipedia (Pan 2017). It covers 282 languages, making it a valuable resource for multilingual NLP tasks, especially in the context of NER. The dataset is designed to identify and classify named entities, such as persons, organizations, and places, in a wide range of languages. The annotations in the dataset are in IOB format, which stands for Inside-Outside-Beginning. In this format, each word in a sentence is tagged with a tag that indicates whether it is inside an entity, at the beginning of an entity, or outside of any entity.

2.2 Slovak tools

Tool name	Tasks	Architecture	Language
NLP Cube	SENT, PoS, MORPH, DEP, LEM	CNN	Python
Dl4dp	PoS, DEP, LEM	LSTM	Python
UD Pipe	PoS, MORPH, DEP, LEM	LSTM	C++, Python
Stanza	PoS, MORPH	CNN, LSTM	Python

Tab. 1. Tools for basic natural language preprocessing with Slovak support

There are several tools for preliminary NLP processing of the Slovak language. They are based on convolutional neural networks or long short-term memory without pretraining. Most of them offer a Python interface. They do not contain word vectors and named entity recognition. The tools are summarized in Tab. 1.

NLP-Cube is an end-to-end Natural Language Processing framework that performs sentence splitting, tokenization, compound word expansion, lemmatization, tagging, and parsing. It takes raw-text as input and annotates it, generating a CoNLL-U2 format file. It is written in Python and based entirely on convolutional neural networks built in DyNET2. NLP-Cube is open-source with support for languages included in the UD Treebanks (Boros 2018).

Dl4dp[2] is a Python NLP library that provides tools for morphological tagging, lemmatization, and dependency analysis. It is mainly designed for English but can also work with other languages that have training data in universal dependencies. Universal Dependencies is a framework for consistent annotation of grammar across languages.

[2] https://github.com/peterbednar/dl4dp

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization, and dependency analysis of CoNLL-U files. It is not language dependent and can work with any language that has annotated data in CoNLL-U format. CoNLL-U is a format for consistent annotation of grammar across languages. UDPipe is available as a library for C++, Python and R (Straka 2016).

Stanza is an NLP library for Python languages. It supports many human languages. It provides pre-trained models for various NLP tasks, such as part-of-speech tagging, named entity recognition, dependency analysis, and sentiment analysis. It also includes tokenization features, sentence splitting, and multi-word token expansion (Qui 2020).

2.3 SpaCy framework

SpaCy[3] features state-of-the-art models for tasks such as named entity recognition, part-of-speech tagging, and dependency parsing, as well as pre-trained models for several languages. With its efficient processing speed and intuitive API, spaCy makes it easy foS developers to incorporate NLP capabilities into their applications. SpaCy supports several languages, including German, Spanish, French, Italian, Portuguese, Dutch, Greek, Norwegian, Lithuanian, Danish, Finnish, Hungarian, Romanian, Slovenian, and Swedish. The classification models are based on a deep learning approach known as transformers (Devlin 2017), which is used in state-of-the-art natural language processing (NLP) models. In the paper (Ye 2019) authors investigate the use of multitask learning and word embeddings for named entity recognition in informal language texts. The authors use spaCy and several reference datasets for their experiments.

3 THE PROPOSED ARCHITECTURE

Training a custom model in spaCy involves defining the pipeline components, specifying the training data, and selecting the hyperparameters for training. The training pipeline is a sequence of components. Components are applied to a text document to extract useful information from it. The default pipeline components include tokenization, part-of-speech tagging, lemmatization, dependency parsing, and named entity recognition. Besides that, there is a word vector component that is available upon request.

The tokenizer takes a text string and splits it into individual tokens. The tokenizer in spaCy is rule-based; it splits the text based on whitespace and punctuation marks and processes special cases such as abbreviations and acronym words.

Part-of-speech tagging, lemmatization, and dependency parsing are performed using the first neural model trained on the UD Slovak Dependency treebank (Zeman

[3] https://spacy.io/

2017). Named entity recognition is performed using another neural model, trained on a WikiAnn (Pan 2017) corpus. Both neural models use a pretrained Slovak model based on a RoBERTa architecture (Liu et al. 2019), called SlovakBERT (Pikuliak et al. 2022). The SlovakBERT model is trained on a large corpus of Slovak text, leveraging the power of the transformer architecture and the masked language modeling objective.

3.1 Word vectors

Word vectors represent each word in the dictionary as a high-dimensional vector, where words that are semantically similar to each other are close to each other in vector space (Mikolov 2013). The word vectors module uses a modification of the version of fastText (Joulin 2016). called Floret. This modification uses Bloom inserts to create compact vector tables of word information. Floret vectors are trained using a technique called "sliced sparse coding", which helps to reduce the dimensionality of the vectors while maintaining their precision. This makes them much smaller and faster to compute than traditional nested words, which can be important for applications where computational resources are limited, such as mobile devices or embedded systems.

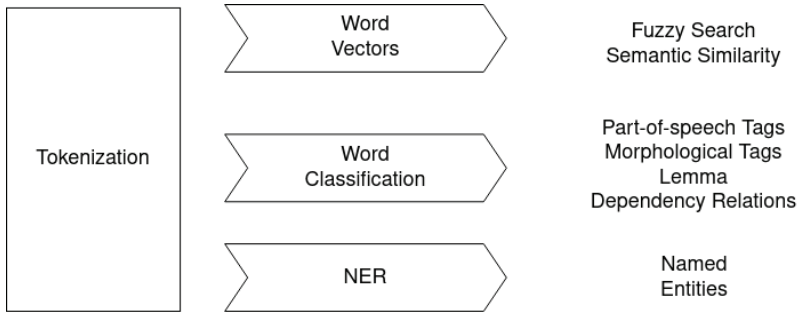

Fig. 1. The proposed model architecture

3.2 Word classification model

This model uses the Transformers architecture (SlovakBERT) and is independent of other models. It performs:

1. POS tagging (into UD tags);
2. morphological analysis (into SNK tags);
3. dependency parsing (into UD tags);
4. lemmatization.

The role of the PoS tagger is to take the output of the tokenizer and assign a PoS tag to each token based on its context in the sentence. It uses contextual information such as the surrounding words and the grammatical structure of the

sentence. It tries to predict the most likely PoS tag for each token. The POS tags follow the conventions of the Universal Dependencies corpora. The output is a sequence of token objects that have been annotated with the corresponding PoS tags (Partalidou 2019). Morphological analysis refers to the more precise analysis of a specific word form. We use the Slovak National Corpus tagset[4] for this purpose. The output of a dependency parser is a tree structure that represents the dependency relations between words in a sentence (Colic 2019). The dependency tags are the same as in the training corpus.

3.3 Named entity recognizer

The second neural model performs named entity recognition. This model uses the transformer architecture and is independent of other models. The model is trained separately because the training corpus is different and has a different distribution license.

The output of the named entity recognizer is a document that contains the identified named entities and their corresponding tags. A named entity is a non-overlapping span in the document that contains one or multiple words. The named entity recognizer assigns three different tags to each word – beginning, inside, outside (BIO). The tags mark whether a word is the beginning of the named entity (B), is inside it (O), or does not belong to a named entity (O). Each word in the input sequence has exactly one tag assigned; thus, the named entities cannot overlap.

4 TRAINING AND EVALUATION

SpaCy offers several different pre-trained model architectures that are optimized for different natural language processing tasks. The „small" model architecture is a basic NLP pipeline that includes tokenization, part-of-speech tagging, and dependency parsing. The „medium" model architecture adds named entity recognition to the basic pipeline of the small model. The „large" model architecture includes all the components of the medium model as well as a deep neural network that is trained on a large corpus of textual data. Our models contain PoS, DEP, and NER taggers, so they are marked as „medium".

4.1 The training process

We trained three statistical models for spaCy:

1. Word vectors – modified fastText on custom web corpus;
2. Word classifier – SlovakBERT with Slovak Dependency Treebank UD corpus;
3. NER recognizer – SlovakBERT with WikiAnn NER corpus.

[4] https://www.juls.savba.sk/~vladob/resources/20180403_ukl_cheatsheet_0.pdf

The workstation used for this experiment was a standalone Ubuntu-Linux PC with a 2.67 GHz Core i7920 processor, 32GB of RAM, and 2 NVIDIA GeForce 1080 graphics cards with 12 GB of VRAM each.

In the first step, we trained word vectors. The word vectors were trained on a large Slovak web-based corpus using a modified fastText tool.[5] We gathered many the Slovak web pages from the Internet. The text was extracted using jusText[6] and tokenized by a rule-based tokenizer.[7] Redundancy was lowered by detecting duplicities on the level of paragraphs. The training corpus has over 4 billion tokens. We trained the model with the recommended hyperparameters: method CBOW/floret, dimension 300, minn 4, maxn 5, hash count 2, bucket 50000.

Then we trained the word classifier and named entity recognizer separately. For both models, we used pretrained SlovakBERT. Hyperparameters were the same for both models: learning rate 1e-5, update freq. 3, dropout 0.1, max. steps 2000, optimizer Adam.

During training we monitored these metrics:

- Part-of-speech accuracy (PoS) and morphological analysis accuracy (MORPH);
- Lemmatizer Accuracy (LEM);
- Unlabeled dependencies score (UAS): which measures the percentage of tokens whose head is correctly predicted;
- Labeled dependencies (LAS): measures the percentage of tokens whose head and dependency label are correctly predicted;
- Named entity recognition precision, recall, and F1.

4.2 The evaluation

The scores were calculated on a development set, left over from the training. The development set consists of 10% of the total data. Results of word classification training are displayed in Tab. 2. We added results for English, German, and Polish languages for comparison. Results for other languages were recorded from the spaCy webpage for "core_medium" models.

Language	PoS	MORPH	LEM	DEP_UAS	DEP_LAS
English	0.97	n-a	n-a	0.92	0.90
German	0.92	0.92	0.98	0.93	0.91
Polish	0.98	0.90	0.94	0.89	0.81
Slovak	0.92	0.94	0.86	0.92	0.89

Tab. 2. Results of the Slovak word classification, compared with other languages

[5] https://github.com/explosion/floret

[6] https://github.com/miso-belica/jusText

[7] https://github.com/hladek/slovak-lexer

The results of the named entity recognition model training are in separate Tab. 3, where the F1, precision, and recall scores for each language are presented.

Language	F1	Precision	Recall
EN	0.85	0.85	0.85
DE	0.84	0.84	0.83
PL	0.83	0.83	0.82
SK	0.89	0.89	0.89

Tab. 3. Results of the named entity recognition

The tables show that the results are comparable with other languages. The only significant difference drop of precision was for the Slovak lemmatization when compared to Polish and German. The reason requires further investigation.

5 CONCLUSION

In this work, we solve the problem of fragmentation and provide a complete pipeline for the preprocessing of the Slovak language using state-of-the-art models. Our approach uses proven means – the Python programming language and the spaCy library with transformers architecture. With one command,[8] it is possible to install a Slovak language preprocessing tool for tokenization, recognition of parts of speech and lemmas, parsing, and recognition of named entities. The experiments have shown that the results are comparable with other languages. This will facilitate access to NLP technologies even for beginners and programmers unfamiliar with the Slovak NLP. The spaCy library already provides support for several other languages, so it is possible to easily add support for the Slovak language as well as for existing applications.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under the project VEGA 2/0165/21, and by the Slovak Research and Development Agency under the projects APVV-22-0261, APVV-SK-TW-21-0002, and APVV-22-0414.

References

Boroş, T., Dumitrescu, S. D., and Burtica, R. (2018). NLP-Cube: End-to-end raw text processing with neural networks. In Proc. of the CoNLL 2018 Shared Task: Multilingual

[8] pip install https://files.kemt.fei.tuke.sk/models/spacy/sk_core_web_md-3.4.1.tar.gz

Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, pages 171–179. Accessible at: https://aclanthology.org/K18-2017.pdf.

Colic, N., and Rinaldi, F. (2019). Improving spaCy dependency annotation and PoS tagging web service using independent NER services. Genomics Inform., 17(2) e21. Accessible at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6808626/.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL HLT, Minneapolis, Minnesota, pages 4171–4186. Accessible at: https://aclanthology.org/N19-1423.pdf.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages Language Resources and Evaluation, 46(1), pages 131–142. Accessible at: https://www.jstor.org/stable/41486069.

Gajdošová, K., Šimková, M. et al. (2016). Slovak dependency treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: https://lindat.cz/repository/xmlui/handle/11234/1-1822.

Hajič, J., Hajičová, E., Mikulová, M., and Mírovský, J. (2017). Prague dependency treebank. In Handbook of Linguistic Annotation, pages 555–594.

Harahus, M, Juhár, J., and Hládek D. (2022). Morphological annotation of the Slovak language in the Spacy library with the pretraining. 32nd International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE, 2022. Accessible at: doi 10.1109/RADIOELEKTRONIKA54537.2022.9764935.

Hládek, D., Staš, J., and Juhár, J. (2014). The Slovak Categorized News Corpus. In LREC, pages 1705–1708. Accessible at: https://aclanthology.org/L14-1517/.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. In Proc. of EACL: Volume 2, Short Papers, Valencia, Spain, pages 427–431. Accessible at: https://aclanthology.org/E17-2068.pdf.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR, arXiv preprint, arXiv:1907.11692. Accessible at: https://arxiv.org/pdf/1907.11692.pdf.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR, arXiv preprint, arXiv:1301.3781. Accessible at: https://arxiv.org/pdf/1301.3781.pdf.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In Proc. of ACL: Volume 1, Long Papers, Vancouver, Canada, pages 1946–1958. Accessible at: https://aclanthology.org/P17-1178.pdf.

Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., and Diamantaras, K. (2019). Design and implementation of an open source Greek PoS tagger and entity recognizer using spaCy. In Proc. of WI'19: IEEE/WIC/ACM International Conference on Web Intelligence, Thessaloniki, Greece, pages 337–341. Accessible at: https://dl.acm.org/doi/10.1145/3350546.3352543.

Pikuliak, M., Grivalský, Š., Konôpka, M., Blšták, M., Tamajka, M., Bachratý, V., Šimko, M., Balážik, P., Trnka, M., and Uhlárik, F. (2022). SlovakBERT: Slovak masked language model. In Proc. of EMNLP, Abu Dhabi, United Arab Emirates, pages 7156–7168. Accessible at: https://aclanthology.org/2022.findings-emnlp.530.pdf.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, D. Ch. (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proc. of ACL: System Demonstrations, Online, pages 101–108. Accessible at: https://aclanthology.org/2020.acl-demos.14.pdf.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, PoS tagging and parsing. In Proc. of LREC, Portorož, Slovenia, pages 4290–4297. Accessible at: https://aclanthology.org/L16-1680.pdf.

Ye, W., Li, B., Xie, R., Sheng, Z., Chen, L., and Zhang, S. (2019). Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. arXiv preprint arXiv:1906.08931.

Zeman, D. (2017). Slovak dependency treebank in universal dependencies. Jazykovedný časopis, 68(2), pages 385–395. Accessible at: https://sciendo.com/article/10.1515/jazcas-2017-0048.

sciendo
DOI 10.2478/jazcas-2023-0050

ANOPHONE: AN ANNOTATION TOOL FOR PHONEMES AND L2 ANNOTATION SYSTEMS FOR CZECH

RICHARD HOLAJ[1] – PETR POŘÍZKA[2]
[1] Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

HOLAJ, Richard – POŘÍZKA, Petr: ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 333 – 344.

Abstract: The goal of this text is the presentation of the ANOPHONE annotation system, which allows for the management and annotation of speech data to develop a tool for the automatic transcription of speech of non-native speakers of Czech. This system is currently designed for annotations on the segmental level of recordings of non-native speakers of Czech, with the aim to train automatic speech recognition (ASR) models used in this tool.

After an introductory section that discusses the use of technology in pronunciation teaching and mentions some of the e-learning applications for teaching the pronunciation of second languages (L2), we address both general and more specific aspects of speech data annotation to train ASR models and mention attributive and synthetic segmental systems of speech data annotation for Czech as L2. We also briefly introduce the annotation system of non-native speakers of Czech called BV1, which is used for testing the ANOPHONE tool. The main part of this text focuses on presenting the annotation tool itself, while the conclusion describes the experience of testing the speech data annotation tool using BV1 annotation system for Czech as L2.

Keywords: annotation, ANOPHONE, pronunciation, L2, ASR, Czech, e-learning

1 THE USES OF TECHNOLOGY IN PRONUNCIATION TEACHING

The importance of e-learning has been growing dramatically in recent years, as has been its use in foreign language teaching (Blake 2011, cf. also Holaj 2018, Holaj – Pořízka 2021). One of the challenges is in tools that would help with the issues of teaching pronunciation to non-native speakers, especially in providing individualized feedback and knowledge verification, as these elements are extremely time- and skill-intensive for teachers. Technology presents itself here as the ideal solution. Derwing and Munro (2015) come to the same conclusion, stating that such tools should be able to identify and automatically evaluate pronunciation errors that have a negative effect on the intelligibility and accessibility of speech production, or provide the user with exercises with feedback and monitor their progress through practice. This is referred to as computer-assisted pronunciation training (CAPT) and appears to be the most promising method for improving the level of teaching

pronunciation to non-native speakers (cf. Levis 2007). Tools that meet these requirements involve the use of machine learning technology.

Although there already exist several tools that address Czech pronunciation learning, notably CzechME (Holaj et al. 2021) and ProCzeFor (Veroňková et al. 2022), none of these tools provide an automatic feedback on pronunciation. It should be mentioned that feedback on Czech L2 pronunciation is provided by the Czech version of the Duolingo application, but it works on the basis of native speaker speech recognition tools and its output is heavily biased (cf. Nováčková 2017, 2018). The situation is better in the teaching of, for example, English pronunciation, where several advanced tools exist, notably ELSA Speak (Elsa Corp.), which focuses only on pronunciation teaching, but even this state-of-the-art application is prone to significant errors in feedback (Becker – Edalatishams 2019). This may be due to the fact that essentially all automatic feedback tools for native speaker pronunciation only work with segments that occur in the pronunciation of native speakers of a given language (Holaj – Pořízka 2021).

Such annotation is time consuming and places specific demands on the annotators. In this paper, we want to show that this process can also be greatly facilitated with the use of technology, specifically by using the ANOPHONE annotation tool we have developed, which reflects these specific requirements. In the following section, we first briefly describe the issue of speech annotation and then use the example of several speech annotation systems for non-native speakers of Czech to show what such annotation can look like. In the last part we will introduce the ANOPHONE tool itself and show how it can facilitate the process of annotating the speech of non-native speakers using one of the described systems.

A solution to this problem could be a speech recognition system trained on data from non-native speakers, which would take into account speech segments that do not occur in the pronunciation of native speakers. However, such a system, which is already under development (Holaj 2023), requires a huge amount of annotated audio data (recordings). Such annotation is time-consuming and places specific demands on the annotators. In this paper, we want to show that this process can be greatly facilitated by using technology, specifically by using the ANOPHONE annotation tool, which reflects these specific requirements. In the following section, we first briefly describe the issue of speech annotation and then, using the example of several systems for annotating the speech of non-native speakers of Czech, we show what such annotation can look like. In the last section, we introduce the ANOPHONE tool itself and show how it can facilitate the process of annotating the speech of non-native speakers using one of the mentioned L2 Czech systems.

2 SPEECH DATA ANNOTATION AND L2 ANNOTATION SYSTEM FOR CZECH

Compared to typical linguistic annotation, speech annotation is different in that it uses audio data rather than text data. An essential criterion for speech annotation is

deciding what the purpose of the annotation is and what aspects or features of the audio recordings we want to capture or annotate. In most cases, sequential annotation is used, which annotates speech as a sequence of phonemes (or speech sounds), syllables, words or other sound segments. This design also includes all transcription-based annotations of recordings. The essential feature of sequential annotation is that the recording is annotated as a sequence of tags (or labels). This implicitly presumes that the recording itself is also a sequence of segments, with each such segment corresponding to one label in the recording's annotation. The order of labels (or tags) and speech segments mutually corresponds in the vast majority of cases. However, aligning sequential annotations represents somewhat of a problem, as the individual segments (speech sounds, syllables, words) are not of the same length. Thus, the labels correspond to sections of different lengths, and a simple sequential annotation alone does not provide information on how to align these labels with the recording. Currently, this issue is most often resolved by using a programme which uses the annotation, such as the CTC (Graves et al. 2006) algorithm. Thanks to this algorithm, there is no need to laboriously include the manual alignment in the annotation, and all that is required is merely to indicate the sequence of the labels (tags) themselves, regardless of their length and alignment with the recording. This is the approach we chose for our ASR application.

In the context of segmental annotation, it is necessary to mention the form of the labels for the individual segments. The most common are simple synthetic labels (tags) that have no internal structure and cannot be further divided. Each segment is then assigned exactly one of these labels. Within this approach, one can imagine, for example, a simple annotation of speech sounds where each speech segment (corresponding speech sound) is assigned exactly one of the possible labels (the one that corresponds to the given speech sound or is closest to it).

In addition to the sequential synthetic label system, there are, for example, attributive or positional systems: an example of an attributive annotation of spoken language are the variant annotations AV1 and AV2 (see Holaj – Pořízka 2021 for more), which we tested in the first phase of the project, but based on the results of the trained models, we decided to change the annotation and replace it with a synthetic one called BV1 (Holaj 2023). An illustrative comparison of a few selected labels for annotations AV1, AV2 and BV1 is shown in Tab. 1.

AV1	AV2	BV1
e::a	a	a
k:z1	g	g
t:vM	t:vM	tʲ

Tab. 1. Illustrative comparison of a few selected annotation labels of AV1, AV2 and BV1 annotation systems

The new annotation system BV1 is a sequence of synthetic tags (labels) without an internal structure. A recording (sound segment) is always assigned a sequence of these

tags separated by a space and corresponding to the sequence of speech sounds in the recording. This system meets the requirements for simplicity (the annotation corresponds only to the identifier of the respective speech sound), using as few tags as necessary (only sounds selected by a user are part of the inventory) and extensibility (the annotation can be extended with new tags and it does not interfere with other labels).

For the purpose of clarity, these labels were divided into several groups based on two criteria. The primary criterion is the type of speech sound corresponding to the label. Along those lines, we divide the tags into vowels, consonants and other tags (such as aspiration, quantity, etc.), which always refer to the previous segment. The secondary criterion indicates whether this is a speech sound from the standard Czech inventory or a non-standard speech sound. The individual tags available in BV1 are described in a work on a tool for automatic speech transcription of non-native speakers of Czech (Holaj 2023).

The standard procedure for annotation in cases where the pronounced speech sound is not among the available tags is to choose the perceptually closest speech sound. If the pronunciation is perceived as acceptable (i.e., it is not a fundamental error in terms of intelligibility), standard speech sounds are preferred if they are roughly similar. Non-standard speech sounds are preferred in cases of unacceptable pronunciation.

A clear advantage of this system is its simple extensibility. Without affecting the already existing tags, additional segments can be added into this system on the basis of new findings, which has already occurred several times over the course of data annotation with BV1. This makes it easy, for example, to add segments specific to certain nationalities as they are included among the data being annotated. The described system was used to test our annotation tool ANOPHONE.

3 ANOPHONE AND SPEECH DATA ANNOTATION

In order to create a tool based on ASR technology, it is first necessary to collect enough speech data from non-native speakers (in our case Czech data) and subsequently annotate it. This annotated data is then trained via neural networks and used to create a speech recognition model. For this annotation, we developed the ANOPHONE application,[1] which fulfils a dual purpose. Firstly, it serves as an online database of recordings of non-native speakers and a repository of their annotations with the ability to filter data using metadata. The second purpose is the annotation of these recordings itself through the annotation tasks that the tool allows us to create; i.e., we can add the task for the annotation of speech sounds or another task for the annotation of stressed and unstressed syllables etc. Although this annotation is manual, ANOPHONE provides a user-friendly interface and makes the entire annotation process simpler and faster minimizing annotation errors.

[1] https://anophone.evetech.cz

The application is divided into an administrator section and a user section. The user section allows annotators to browse the recordings database, including metadata and information of the final annotation for the individual annotation tasks and recordings. These recordings can also be filtered by metadata using regular expressions.

Fig. 1 shows the most important part of the application, the annotation interface itself (with the already mentioned BV1 annotation):

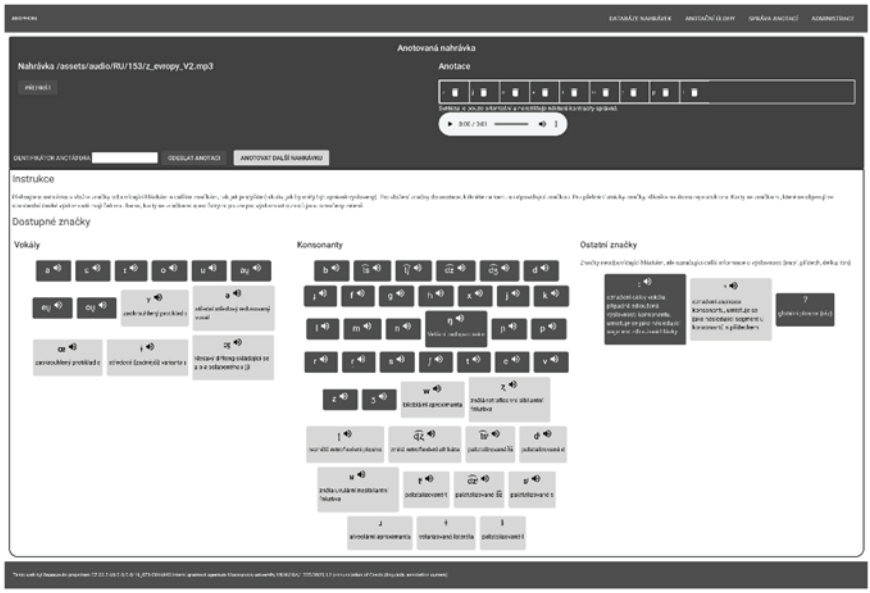

Fig. 1. ANOPHONE: Interface for annotation of recordings
Explanatory notes: anotace = annotation; (anotovaná) nahrávka = (annotated) recording; přehrát = to play; identifikátor anotátora = annotator's identifier; odeslat anotaci = to send an annotation; anotovat další nahrávku = to annotate another recording; dostupné značky = available labels; vokály = vowels; konsonanty = consonants; ostatní značky = other labels.

The lower part of the interface contains annotation instructions and the individual (available) tags. These tags (or labels) are divided into three sections: vowels, consonants and other tags (labels). Standard speech sounds (L2 language inventory) are marked in green (here in light gray) and non-standard sounds are in purple (here in dark grey). In most cases, the labels are accompanied by sample recordings (see speaker icon) or a description of the speech sound to facilitate orientation during annotation (to simplify the annotation process).

Both the recording and the current annotation are located at the top of the interface. The upper left corner contains the name, the recording's path within the source data and a button for playing back the recording being annotated. The right

side contains the current sequence of labels in the annotation. After clicking on a label card, the selected label is added to the end of the sequence corresponding to the annotation of the current segment. These labels can also be removed from the sequence by clicking on the bin icon. Below the current annotation is the experimental option to play back the entire sequence of tags using simple speech synthesis. In the lower left part, there is a box for entering the annotator's identifier (the annotation cannot be saved without it being filled) and buttons for saving the annotation and for moving on to annotating the next recording.

Recordings for annotation are automatically selected form the provided database based on the number of already existing annotations for a given segment within the currently selected annotation task. A random recording is always selected from the recordings with the fewest annotations. A single recording can be used in multiple annotation tasks, while the annotations are always saved for the corresponding annotation task. Before beginning the annotation, the user chooses from among the available annotation tasks.

ANOPHONE also allows (within the administrator section) to edit available labels, add recordings or create new annotation tasks. The editing window essentially allows us to define all aspects of the annotation task. Specifically, we can select a subset (or all recordings) from the entire database of recordings based on set parameters, which will be annotated as part of the annotation task, specify instructions and define labels available for annotation (labels can optionally be divided into groups named by the user). The colour of the label card and the text colour of the given label can also be defined. An example of annotation task administration can be seen in Fig. 2. For a specific annotation task, labels are selected from a database of all available labels, which can also be edited by the administrator (see Fig. 3). In this case, the administrator sets the label identifier, name, description, and path to the recording.

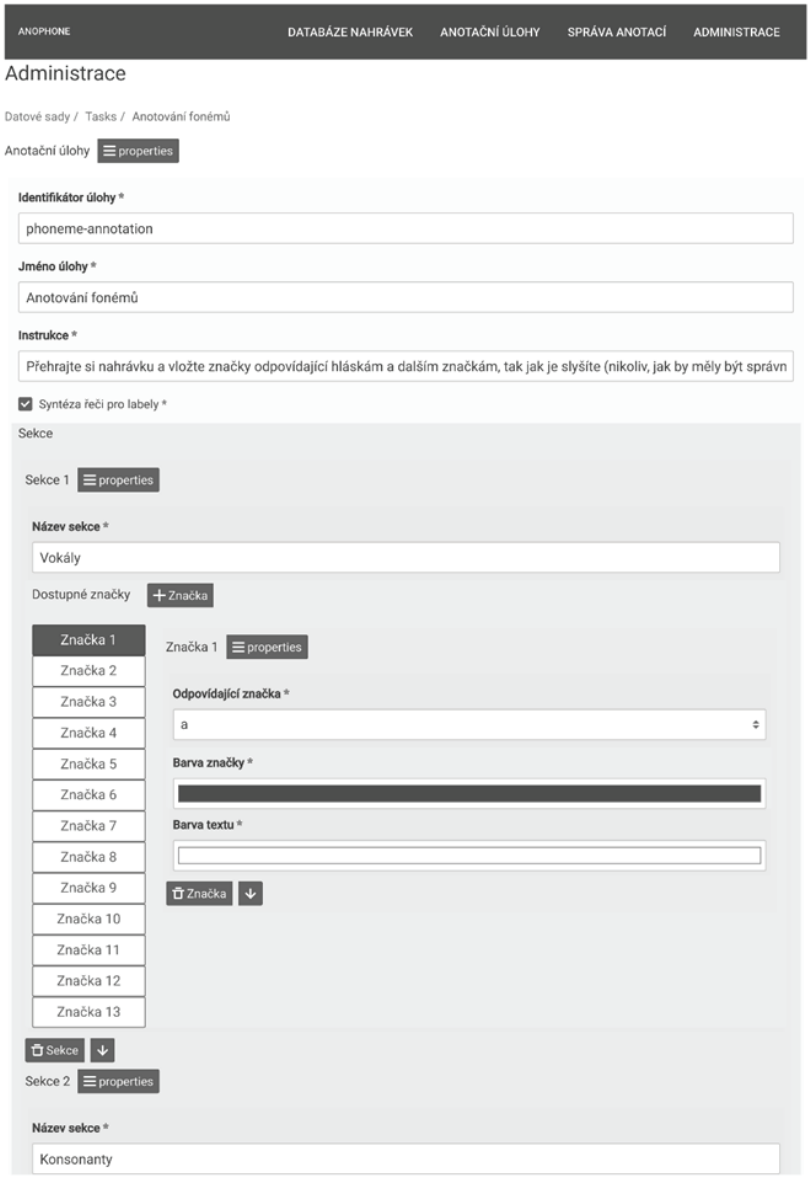

Fig. 2. ANOPHONE: Project administration
Explanatory notes: **Header labels** – databáze nahrávek = database of recordings; anotační úlohy = annotation tasks; správa anotací = annotation administration.
Interface – administrace = administration; identifikátor/jméno úlohy = task identifier/name; instrukce = instructions; název sekce = section name; (odpovídající) značka = (corresponding) label; barva značky/textu = label/text colour.

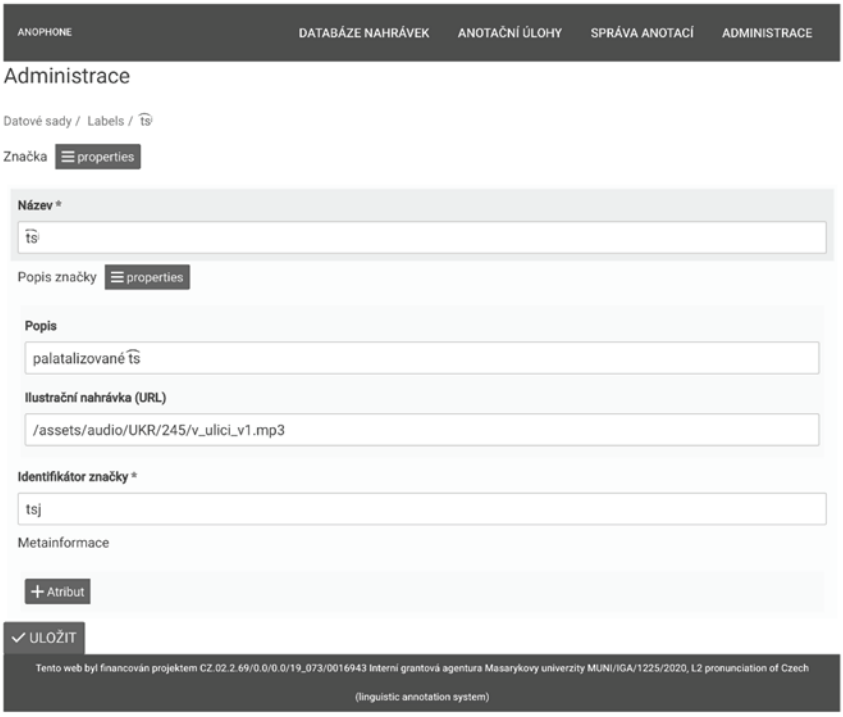

Fig. 3. ANOPHONE: Tag (label) administration
Explanatory notes: administrace = administration; datové sady = datasets; značka = label (tag); název = name; popis = description; ilustrační nahrávka = sample recording; identifikátor značky = tag identifier; metainformace = metadata.

The final relevant function of ANOPHONE available in the administration section is annotation management. In addition to browsing recordings and manual simplified annotation, it also allows for viewing all assigned annotations for a given recording within all annotation tasks. This part of the application also provides functions for setting and editing the final annotation, by selecting from different annotations of the same segment (speech sound, word, phrase), which is then displayed in the recording database. Annotations are grouped by recording and can again be filtered by meta-information. The interface of this screen is shown in Fig. 4.

Fig. 4. ANOPHONE: Annotation management
Explanatory notes: filtrování nahrávek = filtering of recordings; atributy = attributes; finální anotace = final annotation.

3.1 Data Annotation Process Using ANOPHONE

During the annotation process using BV1 tagset, it is necessary to create appropriate data structures in ANOPHONE for the individual recordings (complete with metadata) and import them into the tool along with the recordings themselves. Furthermore, one needs to create data for individual labels of BV1: vowels, consonants, and other tags.

The subsequent annotation of the data prepared in the system typically occurs in two rounds. The first, „draft" version of the annotation, where there are several annotations from different annotators for each recording (segment), is followed by the second phase, in which the final annotation of the recording is determined (and optionally modified) from the available annotations. These resulting annotations are subsequently exported from the ANOPHONE system and, using Python scripts (created by the authors of this text), converted to files named after the individual recordings containing the corresponding annotation. The annotation files processed in this way along with the recordings (both files share the same name and differ only in the file extension: *.label* for annotations and *.wav* for recordings) were subsequently used as input for machine learning (ASR model training). The output is always a recording and a text file containing the relevant annotation (with labelss separated by spaces). This format matches the input data requirements of the speech recognition (ASR) library Persephone (Adams et al. 2018), which we used to train all speech recognition models to create the tool for automatically transcribing the speech of non-native speaker of Czech. (A comparison of the results of the individual models is presented in Holaj – Pořízka 2021 and Holaj 2023.)

Based on a comparison of the fully manual annotation process using a text editor (and the AV1/AV2 attributive tagsets, see Holaj – Pořízka 2021) with the annotation process using ANOPHONE, it can be said that ANOPHONE provides a significant simplification and acceleration of the annotation process. Moreover, compared to text editor-based annotation, typing errors in the annotation are minimized. Thus, the tool proved to be very successful in testing the annotation process of recording non-native speakers of Czech with BV1 annotation.

4 CONCLUSION

The main goal of this contribution was to present the ANOPHONE annotation system, a flexible tool that allows for sound segments to be uploaded into a database for further processing, i.e., speech data annotation. It also serves as a database of recordings with their metadata. Additionally, it enables not just data annotation, but also creating or defining sets of annotation tags or modifying them to suit the users' needs and applying these tags onto data (sound segments) through a browser-based and freely available interface. ANOPHONE also allows users to manage multiple annotation tasks for available recordings or to create new

annotation tasks, independently of the existing ones. This language-agnostic (i.e., relatively universal) tool can thus be used by other researchers for other, similarly oriented projects, helping them to improve and speed up the process of annotating speech data.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic MUNI/IGA/1225/2020 "L2 Pronunciation system (Linguistic Annotation System)".

References

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In LREC 2018 (Language Resources and Evaluation Conference), Japan, pages 3356–3365.

ANOPHONE application. Accessible at: https://anophone.evetech.cz.

Arora, V., Lahiri, A., and Reetz, H. (2018). Phonological feature-based speech recognition system for pronunciation training in non-native language learning. The Journal of the Acoustical Society of America, 143(1), pages 98–108. Accessible at: doi 10.1121/1.5017834.

Becker, K., and Edalatishams, I. (2019). ELSA Speak – Accent Reduction [Review]. In Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference, Iowa State University, pages 434–438.

Blake, R. (2011). Current Trends in Online Language Learning. Annual Review of Applied Linguistics, 31, pages 19–35. Accessible at: doi 10.1017/ S026719051100002X.

Derwing, T., and Munro, M. (2015). Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research. John Benjamins. Accessible at: doi 10.1075/ lllt.42.

Duolingo Inc. The free, fun, and effective way to learn a language! Accessible at: https://www.duolingo.com/.

ELSA Corp. Meet ELSA – Your personal AI-powered English speaking coach. Accessible at: https://elsaspeak.com/.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (ICML '06), New York, USA, pages 369–376. Accessible at: doi 10.1145/1143844.1143891.

Holaj, R. (2018). Počítačová podpora osvojování české ortoepické výslovnosti. Master thesis. Masarykova univerzita, Brno. Accessible at: https://is.muni.cz/th/k5tej/.

Holaj, R. (2023). Nástroj pro automatickou transkripci řeči nerodilých mluvčích češtiny. Dissertation. Masarykova univerzita, Brno.

Holaj, R., et al. (2021). CzechME. Accessible at: https://www.phil.muni.cz/en/research/publishing-and-editorial-activities-of-the-faculty/overview-of-publishing-and-scientific-activities/1778138.

Holaj, R., and Pořízka, P. (2021). L2 Czech Annotation for Automatic Feedback on Pronunciation. Jazykovedný časopis, 72(2), pages 510–519. Accessible at: doi 10.2478/jazcas-2021-0047.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. Annual Review of Applied Linguistics, 27, pages 184–202. Accessible at: doi 10.1017/S0267190508070098.

Nováčková, S. (2017). Nové aplikace. Unpublished. Masarykova univerzita, Brno.

Nováčková, S. (2018). Návrh cvičení k výuce češtiny pro cizince přes mlearning. Accessible at: https://is.muni.cz/th/m0q42/.

Veroňková, et al. (2022). ProCzeFor. Accessible at: https://proczefor.cz.

sciendo
DOI 10.2478/jazcas-2023-0051

DISTRACTOR GENERATION FOR LEXICAL QUESTIONS USING LEARNER CORPUS DATA

NIKITA LOGIN
Independent researcher

LOGIN, Nikita: Distractor Generation for Lexical Questions Using Learner Corpus Data. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 345 – 356.

Abstract: Learner corpora with error annotation can serve as a source of data for automated question generation (QG) for language testing. In case of multiple choice gap-fill lexical questions, this process involves two steps. The first step is to extract sentences with lexical corrections from the learner corpus. The second step, which is the focus of this paper, is to generate distractors for the retrieved questions. The presented approach (called DisSelector) is based on supervised learning on specially annotated learner corpus data. For each sentence a list of distractor candidates was retrieved. Then, each candidate was manually labelled as a plausible or implausible distractor. The derived set of examples was additionally filtered by a set of lexical and grammatical rules and then split into training and testing subsets in 4:1 ratio. Several classification models, including classical machine learning algorithms and gradient boosting implementations, were trained on the data. Word and sentence vectors from language models together with corpus word frequencies were used as input features for the classifiers. The highest F1-score (0.72) was attained by a XGBoost model. Various configurations of DisSelector showed improvements over the unsupervised baseline in both automatic and expert evaluation. DisSelector was integrated into an open-source language testing platform LangExBank as a microservice with a REST API.

Keywords: distractor generation, learner corpora, automated question generation

1 INTRODUCTION

According to (Granger 2008), learner corpora can be extremely useful for e-learning. One possible use of learner corpora is to serve as a source of data for exercise generation, reducing the time needed for test item preparation. In the described system LangExBank[1] learner corpus REALEC (Vinogradova – Lyashevskaya 2022) was implemented as a source of text for gap-fill exercises. One of the main challenges in automated test creation is distractor generation (DG below), as the system usually needs to provide multiple independent outputs for one input. This work is aimed at solving this challenge by using a combination of semantic-based and supervised learning-based methods.

[1] https://github.com/lcl-hse/LangExBank

2 RELATED WORK

According to the review of Kurdi et al. (2020), general and language gap-fill exercises are among the most popular types of automatically generated questions. However, few works concern DG for this type of questions – while QG can be quite easy (an algorithm needs only to insert a gap at a certain place in the input sentence), DG can be tricky as the generated outputs need to be both not *too good* and not *too bad* in terms of language.

In Kumar et al. (2015) gap-fill questions in the domain of high school-level Biology were generated in a three-stage approach. First, topic-related sentences were selected by a NN-based topic model, then a gap was placed by an SVM trained on crowdsourced data and, finally, distractors were selected using the factors of semantic similarity according to a Word2Vec model (Mikolov et al. 2013), syntactic similarity (Dice Coefficient), and contextual fit (language model output probability). In 48.69% of cases, generated distractors were considered appropriate by student evaluators.

Jiang and Lee (2017) present an unsupervised DG approach for Chinese gap-fill lexical questions. Distractor candidates were proposed on the basis of Part-of-Speech (POS) tag and the same difficulty level, and either spelling similarity, word co-occurrence (as attested by PMI) or semantic similarity (as attested by Word2Vec). Then candidates appearing in author-curated Wikipedia corpus in contexts similar to that of the input sentence were filtered out. The maximum plausibility score (46.6%) in human evaluation was obtained by a combination of POS tag-based and similarity-based distractor generation techniques.

In Sakaguchi et al. paper (2013) learner corpus data was used to predict distractors for gap-fill exercises with English verbs. The system implemented a classification approach: distractor occurrences in corpus mistakes were considered positive examples, word lemmas and syntactic dependency types were used as features, and the verb vocabulary was used as the distractor candidate set. The system produced only one distractor for a given input.

3 SYSTEM OVERVIEW

Our system includes two major components: the LangExBank testing platform and Testmaker – the tool for extracting test questions (Vinogradova 2019).

3.1 Data Source: REALEC Corpus

REALEC (Russian Error-Annotated Learner English Corpus) is a corpus consisting of essays written by students of HSE University in the format of IELTS writing exam. As of May 2022, it contained 19,079 texts, including 9,494 argumentative essays, 9,584 graph descriptions and 1 short text serving as an example of annotation schema. The size of corpus is 4,925,478 words with 60,001

unique types. The interface and annotation format of corpus is based on the BRAT (Stenetorp et al. 2012) engine. REALEC annotation includes error span selection, classification (based on a broad hierarchical system of tags) and correction.

3.2 Testmaker

Testmaker can be applied both on LangExBank platform and as a separate Python package. It downloads a selected subset of REALEC data and creates a set of gap-fill exercises for the error tags specified by the user. During the test generation, Testmaker creates gaps in places of spans of specified error types, while replacing all other errors with annotator-suggested corrections. Each output sentence contains only one gap, and the corrections are used as the right answers. Testmaker provides different options of exercise generation: a user can add to the test item two sentences from the source text surrounding the sentence with the error span. A user can also decide in which manner exercises should be generated: an exercise from one sentence or an exercise for each mistake in the sentence.

3.3 LangExBank testing platform

LangExBank is a platform for online testing that supports two types of exercises: corpus-based gap-fill questions and IELTS-like Listening, Reading, and Writing exercises. It is freely available from the Github repository and includes a subplatform for reference material.

4 DATASET

The dataset for distractor generation consisted of 6,804 lexical (formed from errors with the *Choice of lexical item* annotation tag) gap-fill exercises retrieved by Testmaker. Contexts of surrounding sentences were included, and multiple gap-fill items were generated from one sentence. After retrieval, each item was equipped with a set of distractor candidates – words from lexical error spans where *Answer* was also used as a correction. This resulted in d 66,647 triplets consisting of a gap-fill Sentence (*S*), a right answer (*Ans*) and a candidate distractor (*d*).

The constructed dataset was filtered according to the following rules:

1. *Ans* and *d* should appear in vocabulary of both Word2Vec and Brown Corpus (Francis – Kucera 1979).
2. *Ans* and *d* should have at least one common possible POS tag in Brown Corpus vocabulary.
3. Both *Ans* and *d* should not bear an auxiliary POS tag (CC, CD, DT, EX, IN, LS, MD, PT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WO, WP$ or WRB) in context of the sentence.
4. The original correction in sentence should not be a *than* to *that* or vice versa.
5. *d* should not be a one-syllable adjective after a comparative marker.
6. *d* should not be an auxiliary verb.

7. *d* should not copy the word immediately preceding or succeeding the gap.
8. *d* should not be a negation of *Ans* or vice versa.

After filtering, 12,679 triplets remained, from which the first 3,003 were annotated manually – each example was assigned a binary label indicating whether *d* was a plausible distractor. Additionally, 76 gap-fill items from the source set were saved for the final evaluation.

NLTK (Bird et al. 2009) package was used for contextual POS tagging and for obtaining the data from the Brown Corpus, as it is fast on big amounts of input examples and is therefore suitable for production environments. A Word2Vec model trained on non-normalized 2021 Wikipedia dump from NLPL Repository (Fares et al. 2017) was used for filtering and later was included in the feature extraction process. The Word2Vec implementation from Gensim (Řehůřek – Sojka 2010) package was used. The process of dataset construction is displayed in Fig. 1.

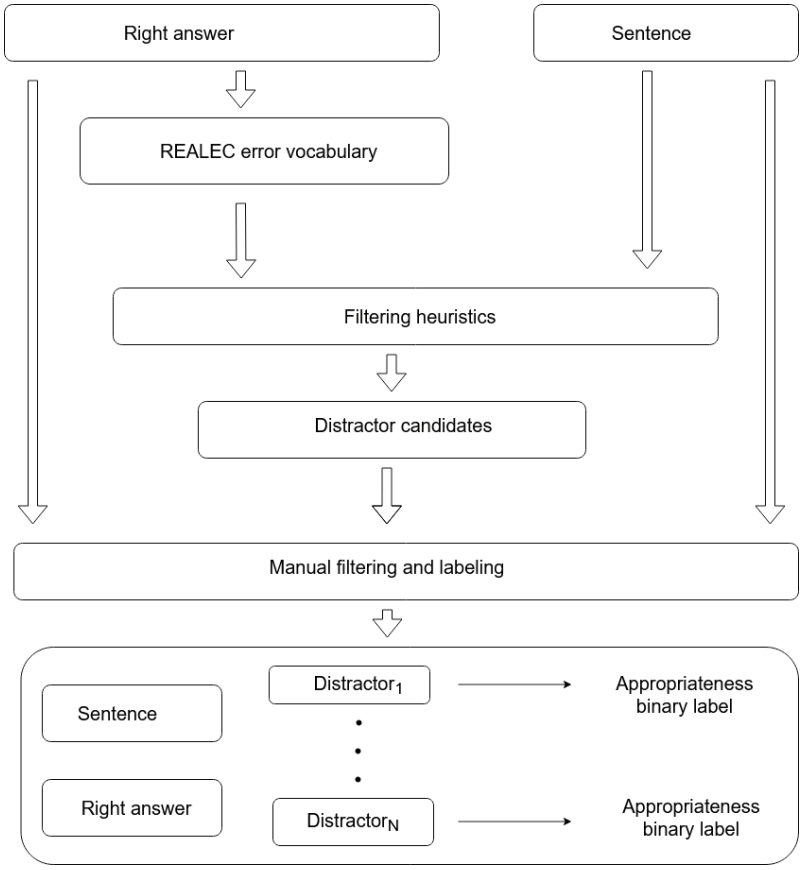

Fig. 1. Dataset construction pipeline

5 MANUAL LABELLING

3,003 distractors from 852 sentences were manually labelled as plausible or implausible. During labelling 69 sentences with 165 distractors were filtered out as implausible due to one of the following factors:

- The original correction was of a properly used word.
- Both *Ans* and *d* were plausible in context of *S* (1).
- *Ans* was an auxiliary word not filtered out automatically.
- *S* had an unclear meaning (2).
- The error consisted in the use of a full form instead of an abbreviation or vice versa.

Examples (1) – (2) illustrate some of the reasons for exclusion (error and correction are **in bold font**, error is ~~crossed out~~).

(1) *Sport centers and sporting goods* ***~~shops~~ stores*** *can reduce prices for sport equipment or for club cards to do different sports.*

(2) *On top of that, parents should tell or show to what can* ***~~prevent~~ lead*** *such activities.*

Example (3) illustrates how the decision whether a distractor was plausible was taken during annotation. The correction (right answer) is **bold**, "too bad" (marginal) distractors are ~~crossed out,~~ and "too good" (correct) distractors are underlined:

(3) *The bar chart illustrates the average amount of minutes which were spent by English people doing sports activities in 2012. At first glance, we can* _____ *that the most active part of the English population in that period was men, while the least active was women. The men aged from 16 to 24 years showed enormous activity, which reached a peak at the point of 282.1 minutes in 2012.*

a. ***see***
b. admit
c. say
d. watch
e. conduct
f. find
~~g. face~~
h. observe

After manual filtering, the labelled dataset consisted of 2,837 examples[2] (1,339 appropriate and 1,498 inappropriate) that corresponded to 781 sentences (1.71 appropriate and 1.92 inappropriate distractors per sentence on average).

[2] One sentence (corresponding to one example) was filtered out during dataset processing by feature extraction algorithm, as the correction word did not appear in corpus in any other contexts.

6 METHOD

As the dataset was relatively small, DG was viewed as a classification problem, and gradient boosting and traditional machine learning approaches were used. The DG pipeline was implemented in three stages: first, our algorithm proposes distractor options for the given sentences, then input examples (consisting of *S*, *d* and *Ans*) are transformed into feature vectors, and, finally, plausible variants are filtered by a supervised binary classifier. During training, distractor candidates from the labelled dataset were used, and distractor suggestion in inference mode is described in Section 6.3.

6.1 Feature Extraction

For the feature extraction, we used a concatenation of Word2Vec and BERT (Devlin et al. 2019) embeddings and corpus frequency counts (Fig. 2). BERT-base-cased model (Devlin et al. 2019) available from Transformers package (Wolf et al. 2020) was used for extracting BERT embeddings. The full set of features included:

- BERT vector of [MASK] token in place of the error span;
- Averaged BERT vector of *S* with gap replaced by *Ans*;
- Word2Vec vector of *Ans*;
- Word2Vec vector of *d*;
- Frequency of *Ans* in corpus corrections;
- Frequency of *d* misuses instead of *Ans* in corpus;
- Frequency of *Ans* in the whole corpus;
- Frequency of *d* in the whole corpus.

6.2 Distractor Classification

The following models were trained on extracted features: K-Nearest Neighbors, SVM, Decision Tree, Logistic and Ridge Regression, Random Forest, Naive Bayesian Classifier, Gradient Boosting, AdaBoost, CatBoost (Dorogush et al. 2018), LightGBM (Ke et al. 2017) and XGBoost (Chen – Guestrin 2016). For all models we used their implementations from Scikit-Learn (Pedregosa et al. 2011) package, except for XGBoost CatBoost and LightGBM which were taken from the eponymous Python packages.

The labelled set was split into training and testing parts in 4:1 ratio using source sentence indices (and not example indices), so no example with *S* from training set would be encountered in testing.

6.3 Distractor Suggestion on Inference

For the inference mode, additional distractor candidates were extracted from Word2Vec. First, cosine similarities between the given word from vocabulary and all the other words that could have the same POS tag were calculated. Then, this set of words was filtered according to rules described in Section 4.

Calculating cosines between these many vectors is computationally expensive, which is not suitable for a production setting. To optimize inference speed, we saved extracted options vocabulary to a file. The option sets were sorted in the following way: first came the words from corpus (sorted by conditional frequency of appearing in corpus error spans, *F(Ans|d)*), and then came words from the Word2Vec model (sorted by cosine similarity). At inference our algorithm extracts top *K* (parameter provided by user) distractor candidates from the described set. The full inference pipeline is shown in Fig. 3.

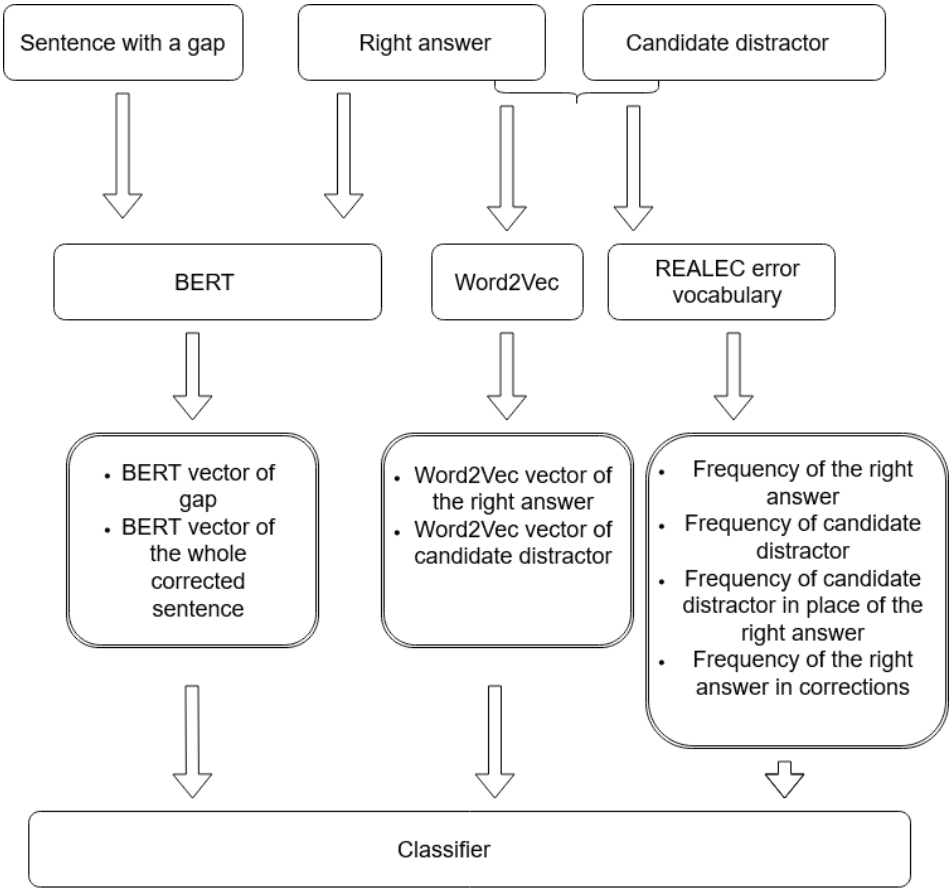

Fig. 2. Feature extraction pipeline

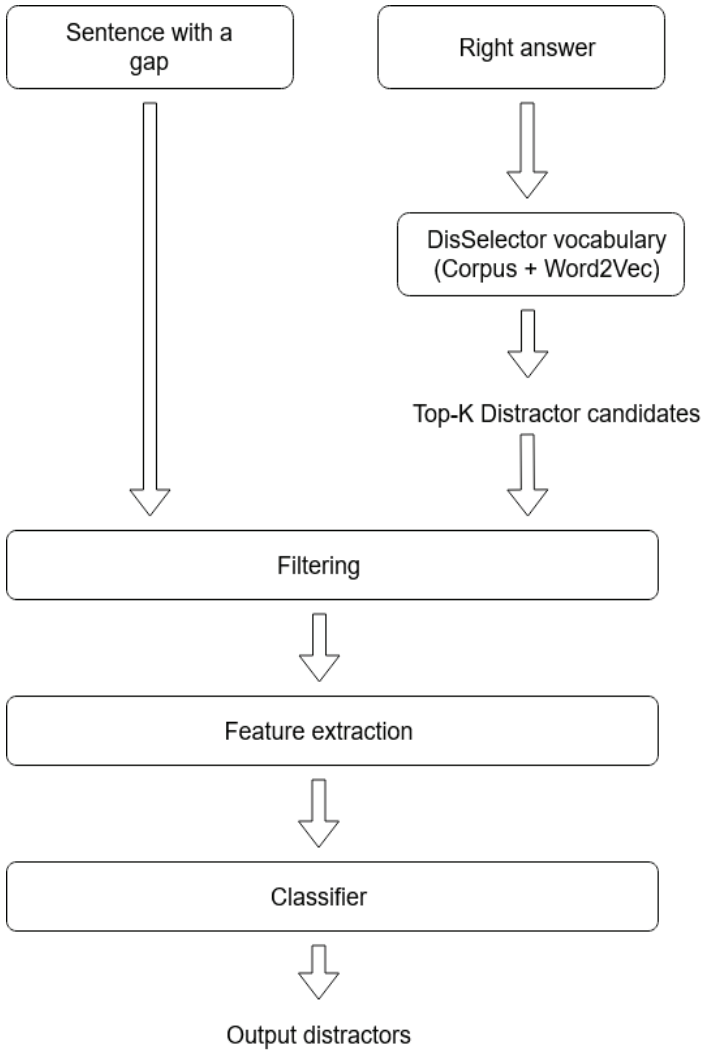

Fig. 3. Inference pipeline

7 EVALUATION

7.1 Automatic evaluation

Distractor classification component was evaluated automatically. As a baseline, we used models trained on the same data, but without manual labelling – only the

original correction was labelled *Appropriate*, as in (Sakaguchi et al. 2013). Class imbalance was encountered in the baseline setup (81.07% examples of *Inappropriate* class), while this was not so for the manually labelled set (52.78% *Inappropriate* examples). The author has also tried training the classifiers only on embeddings (*VecsOnly*), only on frequency features (*FreqsOnly*), and by individually excluding each of the features (*FeatDrop*). Models trained on the whole set of features received *AllFeats* postfix.

7.2 Human evaluation

The inference quality of the whole pipeline was evaluated manually. As a baseline, the author used distractor suggestion component without a classifier on top. The annotator (an EFL expert) was asked to label each of the suggested distractors as "Too good", "Too bad" or "Appropriate". The author used different values of *K* (from 3 to 20) for top classification models (from each of the *AllFeats*, *VecsOnly*, *FreqsOnly* and *FeatDrop* settings) in terms of F1-Score.

8 RESULTS

Results of the automatic evaluation for the best models from each of the settings (*AllFeats*, *VecsOnly*, *FreqsOnly*, *FeatDrop* and baseline) are presented in Tab. 1. The best F1-score (0.72) was achieved by a XGBoost model trained on all features. The model trained exclusively on frequency features shows performance comparable to models trained on word and sentence vectors. Baseline models show far lower performance, which illustrates the benefit of manual annotation of distractor plausibleness.

Model	Setting	Dropped feature	F1-Score	Accuracy	Precision	Recall
XGBoost	AllFeats	—	0.72	0.72	0.69	0.75
Random Forest	FreqsOnly	word and sentence vectors	0.69	0.71	0.70	0.67
CatBoost	VecsOnly	Word frequencies	0.72	0.72	0.68	0.76
CatBoost	FeatDrop	Word2Vec vector of *Ans*	0.73	0.73	0.70	0.75
Gradient Boosting	AllFeats (**unlabelled baseline**)	—	0.43	0.83	0.53	0.36

Tab. 1. Top results of automatic evaluation of distractor classifiers (with baseline)

The top results of the manual evaluation are presented in Tab. 2.[3] The highest percentage of *Appropriate* distractors is demonstrated by a XGBoost model trained on all features, with K=3. There is a visible trend that the quality of distractors is decreasing with the increase of *K*.

Model	Setting	Dropped feature	K	% Appropriate	% Too bad	% Too good	N Appropriate	N distractors
XGBoost	AllFeats	—	3	59.87	27.63	12.5	1.4605	2.5132
CatBoost	VecsOnly	—	3	57.46	27.19	12.72	1.3158	2.2763
XGBoost	AllFeats	—	4	56.14	32.68	11.18	1.7368	3.2368
CatBoost	FeatDrop	Word2Vec vector of *Ans*	3	55.04	30.48	14.47	1.4605	2.6711
CatBoost	VecsOnly	—	4	53.62	31.47	12.28	1.5789	2.9079
XGBoost	All Feats	—	5	53.55	35.77	10.68	1.9342	3.8947
Baseline		—	3	53.07	33.33	13.6	1.5921	3

Tab. 2. Top results of manual evaluation of the whole pipeline

XGBoost model trained on all features with K=4 was selected as the default for DisSelector (as it had the best quality among models producing no less than 3 distractors on average). Its output is presented in Example (4) (with the same highlighting as in Example 3):

(4) *Manufactured goods take one quarter of the pie chart. The second chart _____ that food products and manufactured goods are the most popular goods transported by road. However, food products have the largest proportion (30%).*

a. ***shows***
b. *illustrates*
c. ~~*gives*~~
d. *pictures*
e. *presents*

9 CONCLUSION

This paper describes an implemented solution for generating distractors for gap-fill lexical questions sourced from annotated errors in the learner corpus. After manual evaluation, XGBoost classifier with 4 input distractor candidates was chosen as the default distractor generation method. Manual annotation of distractor dataset has resulted in a considerable performance improvement over the baseline with no additional annotation. This improvement is expected to increase with the enlargement

[3] The percentages do not always exactly add up to 100 as the number of generated distractors may vary among sentences. The final values were calculated as averages of percentages in each sentence.

and quality enhancement (by using labels from multiple independent annotators) of the labelled distractor dataset.

ACKNOWLEDGEMENTS

The research has been supported by the "Development of second language acquisition models according to existing theories using experiment automation methods on REALEC and another learner corpora data" project (TA-102).

The author would like to thank Olga Vinogradova and Anna Viklova.

Notes:

The demo of DisSelector is accessible at http://langexbank-fikl.ru/distractor_api/docs.

The source code for DisSelector is accessible at https://github.com/nicklogin/DistractorSelector.

The full (containing all results) versions of Tab. 1 and 2 are accessible at https://github.com/nicklogin/DistractorSelector/tree/main/full_tables.

References

Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. Sebastopol, CA: O'Reilly Media, Inc, 509 p.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, San Francisco, CA. Accessible at: http://doi.acm.org/10.1145/2939672.2939785.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Accessible at: http://dx.doi.org/10.18653/v1/N19-1423.

Dorogush, A., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. ArXiV preprint, 7 p. Accessible at: https://arxiv.org/abs/1810.11363.

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st Nordic Conference on Computational Linguistics, pages 271–276, Gothenburg, Sweden. Accessible at: https://aclanthology.org/W17-0237/.

Francis, W., and Kucera, H. (1979). Brown Corpus. Providence, Rhode Island: Department of Linguistics, Brown University. Accessible at: http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM.

Granger, S. (2008). Learner Corpora. In A. Lüdeling – M. Kyto (eds.): Corpus Linguistics. An International Handbook. Volume 1. Berlin: Walter de Gruyter, pages 259–275.

Jiang, S., and Lee, J. (2017). Distractor Generation for Chinese Fill-in-the-blank Items. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 143–148, Copenhagen, Denmark. Accessible at: http://dx.doi.org/10.18653/v1/W17-5015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Internal Conference on Neural Information Processing Systems (NIPS 2017), pages 3149–3157, Long Beach, CA. Accessible at: https://dl.acm.org/doi/10.5555/3294996.3295074.

Kumar, G., Banchs, R., and D'Haro, L. (2015). Automatic fill-the-blank question generator for student self-assessment. In Proceedings of 2015 IEEE Frontiers in Education Conference (FIE), pages 1–3, El Paso, TX. Accessible at: https://doi.org/10.1109/FIE.2015.7344291.

Kurdi, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. International Journal of Artificial Intelligence in Education, 30(1), pages 121–204.

Liu, M., Rus, V., and Liu, L. (2018). Automatic Chinese Multiple Choice Question Generation Using Mixed Similarity Strategy. IEEE Transactions on Learning Technologies, 11(2), pages 193–202.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiV preprint, 12 p. Accessible at: https://arxiv.org/abs/1301.3781.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pages 2824–2830.

Řehůřek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 46–50, Malta. Accessible at: http://dx.doi.org/10.13140/2.1.2393.1847.

Sakaguchi, K., Arase, Y., and Komachi, M. (2013). Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 238–242, Sofia, Bulgaria. Accessible at: https://aclanthology.org/P13-2043/.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107, Avignon, France. Accessible at: https://aclanthology.org/E12-2021/.

Vinogradova, O. (2019). To automated generation of test questions on the basis of error annotations in EFL essays: A time-saving tool? In S. Götz – J. Mukherjee (eds.): Learner Corpora and Language Teaching. Volume 29. Amsterdam, Netherlands: John Benjamins, pages 29–48.

Vinogradova, O., and Lyashevskaya, O. (2022). Review of Practices of Collecting and Annotating Texts in the Learner Corpus REALEC. In P. Sojka – A. Horák – I. Kopeček – K. Pala (eds.): Text, Speech and Dialogue. 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings. Cham, Switzerland: Springer Nature Switzerland AG, pages 77–88.

sciendo
DOI 10.2478/jazcas-2023-0052

IS IT POSSIBLE TO RE-EDUCATE ROBERTA? EXPERT-DRIVEN MACHINE LEARNING FOR PUNCTUATION CORRECTION

JAKUB MACHURA[1] – HANA ŽIŽKOVÁ[1]
– ADAM FRÉMUND[2] – JAN ŠVEC[2]
[1] Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

MACHURA, Jakub – ŽIŽKOVÁ, Hana – FRÉMUND, Adam – ŠVEC, Jan: Is it Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 357 – 368.

Abstract: Although Czech rule-based tools for automatic punctuation insertion rely on extensive grammar and achieve respectable precision, the pre-trained Transformers outperform rule-based systems in precision and recall (Machura et al. 2022). The Czech pre-trained RoBERTa model achieves excellent results, yet a certain level of phenomena is ignored, and the model partially makes errors. This paper aims to investigate whether it is possible to retrain the RoBERTa language model to increase the number of sentence commas the model correctly detects. We have chosen a very specific and narrow type of sentence comma, namely the sentence comma delimiting vocative phrases, which is clearly defined in the grammar and is very often omitted by writers. The chosen approaches were further tested and evaluated on different types of texts.

Keywords: comma, Czech, vocative, machine learning, RoBERTa

1 INTRODUCTION

Punctuation, along with other graphical features, is a key element for the correct understanding of any text. Nunberg (1990) describes punctuation as a linguistic subsystem of written language that identifies different text categories. Among the punctuation marks, the sentence comma generally receives the biggest attention. Its use is most often described in grammar, because it is the most frequent punctuation mark across various languages (see Švec et al. 2021; Chordia 2021).

Two tasks usually deal with the automatic insertion of a sentence comma: (1) punctuation restoration in speech transcripts from automatic speech recognition (ASR), where the restored punctuation dramatically improves the readability of the recognised transcript, and (2) grammatical error correction in written texts where the comma is missing or redundant.

For the first task, defining the beginning and the end of the sentence unit is necessary. A system should handle capitalisation to identify the beginning and insert a period or question mark to indicate the end of the sentence. The comma itself then serves as the intra-sentence mark that helps to structure the sentence on syntactic and semantic levels. The second task usually does not have to deal with the insertion of terminal punctuation marks, as these usually already appear in the texts. It is obvious both tasks have different starting positions, and we do not consider it fair to compare the results of tools that insert commas in texts for different tasks.

Nevertheless, the BERT-based method for ASR proposed in (Švec et al. 2021) had promising results, so we decided to train a new model targeted only on the comma correction task (hereinafter referred to as **RoBERTa-commas**). The new model's results were presented in (Machura et al. 2022). The precision was slightly higher (96.1%) than the best rule-based system for Czech (95.1%). But the new model completely outperforms the recall (number of correctly found commas) – 89.5% versus 58.8%. Although the results of the new model are excellent, for the purposes of a reliable language corrector, we were interested to find out whether we would be able to modify the model in any way further – to find and correct the errors it makes and to identify the 10% of commas that the model ignored. The goal of this paper is to see whether we can expand the number of commas found by modifying the learning process. Two new models – **Extra-Fine-tuning** and **Mixed-data-fine-tuning** – will be compared with the **RoBERTa-commas** model.

2 APPROACH

Writing commas in Czech is traditionally part of the grammar, and reference books extensively describe the rules for writing commas. Nunberg (1990) recognises two main classes of commas. He sees a difference between a comma which separates structures at the same level (the separator comma), and a comma which delimits the boundaries between syntactic structures at different levels (the delimiter comma). To define all possible places where a comma should be inserted in the Czech text, we have created a typology of the comma insertion places described in detail (Machura et al. 2022).

In our previous research, we used a corpus of newspaper articles with 87,379 commas for further analysis. All commas were intentionally deleted, and the **RoBERTa-commas** model reinserted commas back into the texts. From the texts, we selected a sample of about 10,000 sentences that did not contain a comma, or a comma was inserted in a different place compared to original newspaper articles. These sentences are currently being further classified according to the established typology. As a result, we can more clearly identify types of commas that are ignored or incorrectly inserted by the model.

When we find a comma that the model did not insert into the text, the first question is why. Is it a type that was not included in the training data, and therefore the model could not learn it? And if so, can we influence the training process in any way – e.g. by making the training data larger or by extra fine-tuning? Or is there a possibility to do postprocessing and fix the commas using formal linguistic rules?

Based on these questions, we decided to choose a very narrow class of commas, namely commas around vocatives ("vocative comma"), for the following reasons:

1. it is a very common error among a large number of writers[1]; 2. writing of commas around vocative in Czech has very fixed rules without any exceptions (Internet Language Reference Book 2023).

We wondered whether it was possible to improve the language model by supplying it with specific sentences containing constructions that were previously evaluated incorrectly by the model.

2.1 Vocative

Traditionally, the vocative is understood as a case that significantly differs from others. Morphologically, it behaves like other cases in Czech – the vocative suffix is attached to the right periphery of the noun. However, the vocative is not governed by any other clause element. In the vast majority of cases, it is a noun phrase that is not part of the valence frame. Its position in the clause is almost arbitrary – at the beginning: *Pane, vyslyš nás.* 'Lord, hear us.'; in the middle: *Prosím, pane, vyslyš nás.* 'Please, Lord, hear us.'; and at the end: *Vyslyš nás, pane.* 'Hear us, Lord.' Some theories propose the idea that the vocative stands completely outside the syntactic structure of the clause – the evidence is, e.g. the vocative's invisibility to clitics that compulsorily occupy the second position: ***Karle, se schovej!* × *Karle, schovej se!*** 'Charles, hide!' (Karlík 2017).

Some suffixes of masculine singular (e.g. *Bože* 'God' etc.) and feminine singular (e.g. *ženo* 'wife') are unique only to the vocative. However, most vocative suffix forms are shared with other cases, most often with the nominative case (e.g. *paní ministryně* 'Madame Minister' etc.). Therefore, the presence of a comma around the vocative is crucial for automatic morphological analysis if the vocative is to be correctly tagged. This is also one of the reasons why rule-based systems dependent on morphological analysis cannot detect vocative commas reliably.

[1] For instance, a probe of the web corpus Araneum Bohemicum Maius (Benko 2015) confirmed the error rate of the vocative forms for the noun *slečna* 'Miss' and the proper noun *Adam*. The corpus contains 635 occurrences of the vocative *Slečno* at the beginning of a sentence of which 33 instances (5%) are missing a comma (e.g. *Slečno nemáte tam něco ostřejšího?* 'Miss do you have anything sharper?'). The form *Adame* is found 393 times in the corpus, with at least 59 occurrences (approx. 15%) being vocatives without a missing comma.

3 ROBERTA-COMMAS TRAINING

We used our pre-trained model on a collection of web data processed in our web mining tool (Švec et al. 2014), CommonCrawl data[2] and texts collected from the Czech Wikipedia. We followed all the training steps mentioned in (Liu et al. 2019) for pre-training the Czech RoBERTa model (Lehečka et al. 2021). As suggested, we used dynamic masking of the tokens. This strategy generates the masking pattern every time a sequence is fed to the model. Also, the pre-training procedure does not use Next Sentence Prediction loss in contrast with the BERT model (Devlin 2019). For tokenizing the input sequence, we used Byte-Pair Encoding (BPE), introduced in (Radford et al. 2019), with a subword vocabulary containing 50K units. This implementation of BPE uses bytes instead of Unicode characters as the base subword units.

We used the ADAM optimisation with linear warmup up to learning rate $4 \cdot 10^{-4}$ for the first 24K steps followed by linear decay to 0. We pre-trained the model for 500K steps as described in (Lehečka 2021).

For this experiment, we proposed the RoBERTa model extended by a few extra classification layers (Švec et al. 2021). An input sequence is preprocessed using the tokenizer, and special tokens are added. Next, we use the Czech pre-trained RoBERTa model with output dimension d=768. The last hidden states are transformed by four regular densely-connected neural network layers. Three of these layers use the element-wise ReLU activation function, and the last layer uses the softmax activation function. The last layer output defines whether the comma should be placed right after the current token. The overall scheme of the proposed neural network architecture is depicted in Fig. 1.

As the training data set used for fine-tuning, we used 10 GB of raw text extracted from the Czech CommonCrawl data set. Because RoBERTa's output is related to input tokens (not words), we assigned the target label ("," for comma, "0" for no comma) to each token of the word, which is followed by a comma (as shown at Fig. 1). In the prediction phase, it is necessary to combine the predictions for the partial tokens into word-level predictions using a per-word pooling. We use a simple average pooling algorithm to obtain the word-level predictions. As the model defines this experiment as a two classes classification per each token, we use standard categorical cross-entropy loss. For optimisation, we use the ADAM optimisation algorithm. The parameters of the whole network (**RoBERTa-commas**) – consisting of the RoBERTa and the classification layers – were updated during fine-tuning.

The epoch size of **RoBERTa-commas** equals 10K sequences, the batch size equals 45, and the number of epochs equals 25, 50 and 75. During the fine-tuning, we use a linear learning rate decay with values starting at 10^{-4} and ending at 10^{-5}. The maximum sequence length is set to 128.

[2] https://commoncrawl.org/

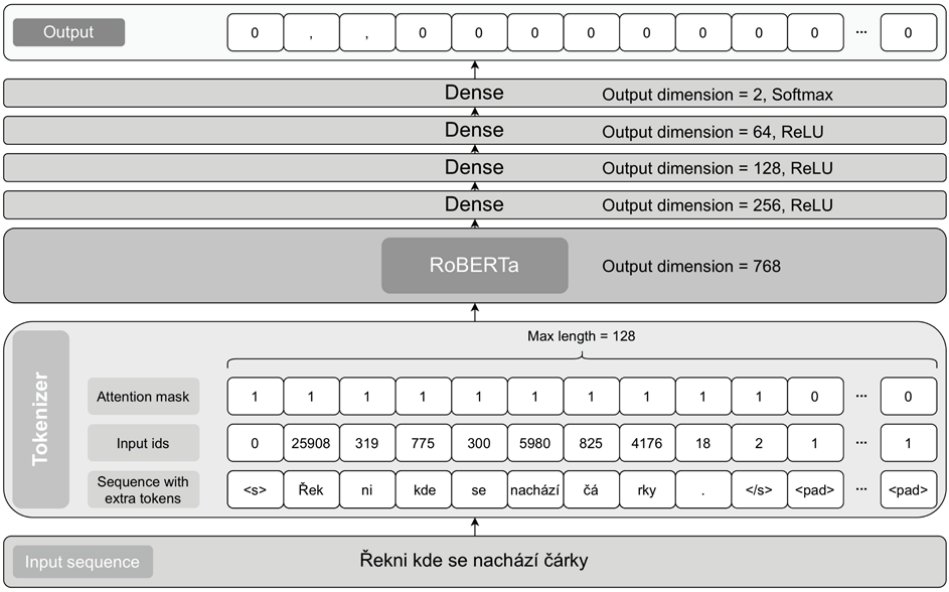

Fig. 1. Proposed neural network architecture

Then, the analysis of the incorrectly evaluated sentences followed (on approximately 10K sentences), and a sample of 89 sentences that did not contain a comma around the vocative was selected. The language model is quite successful regarding proper nouns and nouns. Nevertheless, the model did not insert some commas, mostly for appellative vocatives:

Kluci já jsem delegace* 'Guys* I am a delegation'.

To chce klid vládo, že ano?* 'Take it easy* government, won't you?'

There were also instances where the model inserted a comma in the wrong place.

> *Paní ministryně*, kariérně prošla Senát.* 'Mrs. Minister*, has had a career in the Senate.'

We assume the model made this error because the form *paní ministryně* 'Mrs. Minister' appeared very often as a vocative in the training data. Therefore, we were interested in identifying the most frequent appellative Czech vocative.

To further improve the **RoBERTa-commas** model after examining the phrases that had been improperly processed, we decided to build a small extended text corpus that purposely contained vocative phrases. The following section goes into further detail about this corpus.

Firstly, we attempted to improve the **RoBERTa-commas** model that had already been fine-tuned using the corpus of vocative phrases. Thus, we used a model that was fine-tuned on 10 GB of raw text data, and then we ran a quick extending

(second) fine-tuning process (utilizing the corpus of vocative phrases) to adjust the weights of the RoBERTa-commas model. This approach will hereafter be referred to as **Extra-Fine-tuning**.

Second, we tried a different approach by merging the original raw text corpus with the corpus of vocative phrases. Then, using the newly obtained mixed data, we replaced the initial fine-tuning procedure on raw text data with this mixed data. This approach will be referred to as the **Mixed-data-fine-tuning**. Both approaches are described in more detail in the following sections.

3.1 Retraining data collection

We performed a corpus analysis on the csTenTen17 corpus (Suchomel 2018), which is a web corpus containing 10.5 billion words. It turned out that the appellative vocative in Czech consists most often of one to three words (in the case of four or more consecutive vocative word forms, these were enumerations, repeated exclamations, etc.). In the case of one-word vocative, the most frequent lemmas are *pan, bůh, doktor* 'mister, god, doctor'. If it is a two-word vocative, constructions such as *pane ministře, pane doktore, vážený pane* 'Mr. Minister, Mr. Doctor, Dear Sir' are most frequent. Feminine pronouns are less frequent. Furthermore, it turns out that appellative two-word vocatives often have a form with a personal pronoun: *ty vole, ty bláho, vy idiote, vy troubo* 'you dude, you fool, you idiot, you moron'. In the case of vocative containing three parts, the most frequent are *vážený pane přísedící, vážený pane místopředsedo, vážený pane předsedo* 'dear Mr. Chair, dear Mr. Vice-President, dear Mr. Chairman'.

Based on the analysis, we collected a set of 170,000 sentences with 82,381 vocatives using CQL queries.[3] The set contained sentences where there were one to three words in the vocative between punctuation marks. Proper nouns were filtered and removed. Since the previous set completed by the language model lacked punctuation for feminine vocatives (the type *paní ministryně* 'Mrs. Minister'), we tracked down sentences that explicitly contained the lemma *paní* 'madame'. A sample of selected sentences was as follows:
one-word vocative – 30,000 sentences
two-word vocative – 30,000 sentences (including approximately 18,000 sentences with vocative type *pane ministře* 'Mr. Minister', and 12,000 sentences with vocative type *paní ministryně* 'Mrs. Minister'
three-word vocative – 11,271 sentences
type *vy osle* 'you dude' pl. – 1,110 sentences
type *ty osle* 'you dude' sg. – 10,000 sentences

In each sentence, the vocative was separated by punctuation, and the position of the vocative in the structure varied.

[3] Such as [tag="kI.*"][tag=".*c5.*"][tag="kI.*"] and its modifications.

4 EXPERIMENTAL RESULTS

4.1 Findings

Both approaches (Extra-Fine-tuning and Mixed-data-fine-tuning) for learning the "vocative comma" type were first evaluated on a corpus of newspaper articles. As mentioned above, the sample of 89 sentences containing mainly appellative vocatives was used to analyse this specific type. **Extra-Fine-tuning** brought significantly better results. More than 60% of the vocative phrases in the sample were correctly delimited, whereas the **Mixed-data-fine-tuning** correctly delimited less than 40%.

Vocative – total 89	**Extra-Fine-tuning**	**MIXED-DATA-FINE-TUNING**
Correctly delimited	54	35
Ignored	22	38
Delimited only from one side	5	7
Semantically ambiguous	8	9

Tab. 1. Evaluation of both approaches (Extra-Fine-tuning and Mixed-data-fine-tuning) on a sample of 89 vocatives

The sample also contained semantically ambiguous clauses, and there was more than one option for inserting a comma.

E.g.
Ale jen generační, to není jasný, pane. 'But only generational, that's not clear, sir.'
vs.
Ale jen generační to není, jasný pane. 'But it's not just generational, kind sir.'

The data used to retrain the model also affected other types of sentence commas. The RoBERTa-commas model could not handle tokens that were connected with other punctuation marks in the text (the original training data contained only periods, commas and question marks):

Tedy představte si sloupec 70 cm vody (po ***stehna) kdyby*** *to spadlo naráz.* 'So imagine a column of 70 cm of water (up to your thighs) if it fell all at once.'

Ilf zmiňuje ve svých zápiscích ***„tvář jež*** *nejeví vyčerpanost z mentálního úsilí".* 'Ilf mentions in his notes "a face that does not show exhaustion from mental effort".'

However, the data used for retraining also contained other punctuation marks. As a result, during the **Extra-Fine-tuning**, the model learned to insert commas after tokens that were directly connected with another punctuation mark:

*Tedy představte si sloupec 70 cm vody (po **stehna), kdyby** to spadlo naráz.*

*Ilf zmiňuje ve svých zápiscích „**tvář, jež** nejeví vyčerpanost z mentálního úsilí“.*

And the **Mixed-data-fine-tuning** model no longer incorrectly inserts a comma in a sentence with the homonymous phrase *paní ministryně* ‘Mrs. Minister’ mentioned above:

Paní ministryně kariérně prošla Senát. ‘Mrs. Minister has had a career in the Senate.’

Apparently, retraining can improve the detection of a specific type if we intentionally control what the model has to learn. This small sample of vocatives, though, is a part of the corpus of newspaper articles that has 87,379 commas, and in general, we estimate that vocative is a very marginal phenomenon, less than 1% of all commas of text (Machura et al. 2022). Therefore, we were interested in how retraining one specific type affects the global metrics on the entire corpus, which is used for detailed analysis of the insertion of the automatic comma.

Evaluation of precision, recall and F1 score (the harmonic mean of precision and recall) for the **RoBERTa-commas**, **Extra-Fine-tuning** and **Mixed-data-fine-tuning** on the corpus of newspaper articles is presented in the following table.

Evaluation on the corpus of newspaper articles total number of commas: 87,379					
	Absolute number of commas inserted	**Number of commas correctly inserted**	**P [%]**	**R [%]**	**F1 [%]**
RoBERTa-commas	78,146	76,250	97.6	87.3	92.1
Extra-Fine-tuning	77,976	75,839	97.3	86.8	91.7
Mixed-data-fine-tuning	77,890	75,553	97.0	86.5	91.4

Tab. 2. Evaluation of precision, recall and F1 score

While the evaluation showed that both approaches improved “vocative comma” detection, it seems that the retraining data slightly degraded all global metrics. This may be caused partly by the fact that “vocative comma” is a very marginal phenomenon, and the retraining data was too large (more than 80K sentences) and distorted the natural distribution of all types of commas in the original training data. With this hypothesis in mind, we decided to evaluate both approaches on real texts of different natures.

4.2 Evaluation data sets

The same data as presented in (Kovář et al. 2016) were used to evaluate and compare the methods described above. These texts were prepared specifically for the automatic insertion of commas. Since the data are exactly the same, it is also possible to compare the current results with testing done in the past. In total, seven texts of different nature and styles were used, see Tab. 3.

Testing set	**# words**	**# commas**
Selected blogs	20,883	1,805
Internet Language Reference Book	3,039	417
Horoscopes 2015	57,101	5,101
Karel Čapek – selected novels	46,489	5,498
Simona Monyová – Ženu ani květinou	33,112	3,156
J. K. Rowling – Harry Potter 1 (translation)	74,783	7,461
Neil Gaiman – The Graveyard Book (translation)	55,444	5,573
Overall	**290,851**	**29,011**

Tab. 3. Statistics of the test data

Both newly trained models behaved differently than we expected (see Tab. 4). They significantly improved precision, with **Mixed-data-fine-tuning** outperforming **Extra-Fine-tuning**. This fact disproved our original hypothesis, as we predicted that precision would decrease based on the results from the corpus of newspaper articles (see Tab. 2).

Recall, on the other hand, decreased more for both models than we expected. For the **Extra-Fine-tuning** by 4.5 percentage points, for the **Mixed-data-fine-tuning** even by 11.5 percentage points. For lower-register texts (blogs and horoscopes), the **Mixed-data-fine-tuning** found more correct commas than the **Extra-Fine-tuning**, whereas, for fiction, the **Extra-Fine-tuning** won over the **Mixed-data-fine-tuning**.

A possible explanation might be that stylistically lower texts would not be rich in complex constructions, and writers use frequent connectors and basic sentence structures. On the other hand, fiction is full of direct speech and more complex sentence constructions, and this requires more frequent delimitation of parts of the text without the presence of connectors. The original model outperformed both retrained models in recall and F1 score. The absolute number of inserted commas clearly shows that RoBERTa-commas tried to insert far more commas than the two new models (Mixed-data-fine-tuning even inserted almost 4,000 fewer commas than RoBERTa-commas). Therefore, a logical explanation for the better precision is

offered: The fewer commas a model adds, the less chance it has to make a mistake. We deliberately did not compare our trained models with other approaches that deal with automatic insertion of commas because this would not answer our research question.

RESULTS	***ROBERTA-COMMAS***			***EXTRA-FINE-TUNING***			***MIXED-DATA-FINE-TUNING***		
	P [%]	**R [%]**	**F1 [%]**	**P [%]**	**R [%]**	**F1 [%]**	**P [%]**	**R [%]**	**F1 [%]**
SELECTED BLOGS	**95.5**	**88.3**	**91.8**	**97.5**	**83.3**	**89.8**	**97.6**	**84.8**	**90.8**
INTERNET LANGUAGE REFERENCE BOOK	**91.8**	**70.0**	**79.5**	**95.5**	**55.4**	**70.1**	**96.1**	**52.5**	**67.9**
HOROSCOPES 2015	**96.4**	**93.9**	**95.2**	**97.5**	**87.7**	**92.3**	**97.7**	**87.7**	**92.4**
KAREL ČAPEK – SELECTED NOVELS	**95.3**	**88.9**	**92.0**	**97.3**	**85.7**	**91.1**	**97.5**	**78.8**	**87.2**
SIMONA MONYOVÁ – ŽENU ANI KVĚTINOU	**95.8**	**93.1**	**94.4**	**97.5**	**88.6**	**92.8**	**98.3**	**86.3**	**91.9**
J.K. ROWLING – HARRY POTTER 1 (TRANSLATION)	**96.6**	**88.4**	**92.3**	**97.6**	**85.0**	**90.9**	**97.9**	**77.5**	**86.5**
NEIL GAIMAN – THE GRAVEYARD BOOK (TRANSLATION)	**96.8**	**87.2**	**91.8**	**98.0**	**82.1**	**89.4**	**98.6**	**64.3**	**77.8**
OVERALL PERFORMANCE	**96.1**	**89.5**	**92.7**	**97.6**	**84.9**	**90.8**	**97.9**	**78.0**	**86.9**
Absolute number of commas inserted	**27,000**			**25,244**			**23,125**		
Number of commas correctly inserted	**25,958**			**24,631**			**22,640**		

Tab. 4. Results: RoBERTa-commas, Extra-Fine-tuning and Mixed-data-fine-tuning

5 CONCLUSION

This paper aimed to find out whether it is possible to re-train the language model RoBERTa by providing example sentences in which the language model made errors. We performed a very narrowly focused study: we chose to detect commas in the Czech vocative. We deliberately chose an area that is not problematic in Czech, and writing commas, in this case, is fixed without exceptions. We analysed the types of Czech sentences with vocative, determined how such sentences look in an authentic text and developed a typology, according to which we collected 170,000 sentences with 82,381 vocatives from the csTenTen17 corpus using CQL queries. We used this set of sentences to retrain RoBERTa-commas.

If we precisely analyse the phenomenon that the RoBERTa-commas model ignores by means of retraining, we can detect a specific type. For retraining, we created a corpus that purposefully contained commas around appellative vocatives. We then tried to integrate this new corpus into the learning process: (1) as extra additional training (**Extra-Fine-tuning**), (2) we merged the specialised corpus with the original large corpus of training data (**Mixed-data-fine-tuning**). However, testing showed that in our case, deliberate modification of the original training data or extra fine-tuning significantly reduced recall and likely led to an increase in precision.

This result is probably caused by overfitting the neural network to one specific type of comma. The model weights are then adjusted to best cover the comma around the vocative, and all other types of commas could be overlooked. As the neural network is more like a black box, we cannot say with certainty that this is the case. But it could be a good idea to consider the distribution of the phenomena in the text and proportionally create the retraining data accordingly. The comma types distribution has already been roughly estimated in (Machura et al. 2022). Still, this estimate was made on a very small sample, and the distribution will need to be calculated on a larger sample of different types of texts. Otherwise, the new knowledge may overshadow the original functionality.

Our analysis showed that re-training RoBERTa is possible, but the structure of training data plays an important role.

ACKNOWLEDGEMENTS

A. Frémund and J. Švec were supported by the grant of Czech Science Foundation (GA CR), project No. GA22-27800S.

References

Benko, V. (2015). Araneum Bohemicum Maius, verze 15.04. Ústav Českého národního korpusu FF UK, Praha 2015. Accessible at: http://www.korpus.cz.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. /abs/1810.04805 Accessible at: https://doi.org/10.48550/arXiv.1810.04805.

Chordia, V. (2021). PunKtuator: A multilingual punctuation restoration system for spoken and written text. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pages 312–320. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2021.eacl-demos.37.

Internet Language Reference Book. (2008–2023). Praha: Ústav pro jazyk český AV ČR. Accessible at: https://prirucka.ujc.cas.cz/.

Karlík, P. (2017). Vokativ. In M. Nekula et al. (eds.): Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/search?action=listpub&search=vokativ.

Kovář, V. et al. (2016). Evaluation and improvements in punctuation detection for Czech. In P. Sojka et al. (eds.): Text, Speech, and Dialogue, pages 287–294. Springer International Publishing.

Lehečka, J. et al. (2021). Comparison of Czech Transformers on Text Classification Tasks. In L. Espinosa-Anke et al. (eds): Statistical Language and Speech Processing. SLSP 2021. Lecture Notes in Computer Science, vol. 13062. Springer. Accessible at: https://doi.org/10.1007/978-3-030-89579-2_3.

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv. /abs/1907.11692. Accessible at: https://doi.org/10.48550/arXiv.1907.11692.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts: Comparative Study. In P. Sojka et al. (eds): Text, Speech, and Dialogue. TSD 2022. Lecture Notes in Computer Science, vol 13502. Springer. Accessible at: https://doi.org/10.1007/978-3-031-16270-1_10.

Nunberg, G. (1990). The Linguistics of Punctuation. CSLI lecture notes. Cambridge University Press.

Radford, A. et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2018, pages 111–123.

Švec, J. et al. (2014) General framework for mining, processing and storing large amounts of electronic texts for language modelling purposes. Lang Resources & Evaluation 48, pages 227–248. Accessible at: https://doi.org/10.1007/s10579-013-9246-z.

Švec, J. et al. (2021). Transformer-based automatic punctuation prediction and word casing reconstruction of the ASR output. In K. Ekštein et al. (eds.): Text, Speech, and Dialogue, Springer International Publishing, pages 86–94.

sciendo
DOI 10.2478/jazcas-2023-0053

WHEN IS A CRISIS REALLY A CRISIS? USING NLP AND CORPUS LINGUISTIC METHODS TO REVEAL DIFFERENCES IN MIGRATION DISCOURSE ACROSS CZECH MEDIA

ONDŘEJ PEKÁČEK[1] – IRENE ELMEROT[2]
[1] Department of Sociology, Faculty of Social Sciences, Charles University, Prague, Czech Republic
[2] Department of Slavic and Baltic Studies, Finnish, Dutch and German, Faculty of Humanities, Stockholm University, Stockholm, Sweden

PEKÁČEK, Ondřej – ELMEROT, Irene: When Is a Crisis Really a Crisis? Using NLP and Corpus Linguistic Methods to Reveal Differences in Migration Discourse across Czech Media. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 369 – 380.

Abstract: This article presents an interdisciplinary analysis of discourses on refugees, asylum seekers, immigrants, and migrants (RASIM) in mainstream and alternative media in the Czech Republic. Using techniques from corpus linguistics (CL) and natural language processing (NLP) and drawing on insights from media sociology, we demonstrate the value of an interdisciplinary approach for conducting robust research that can inform policymakers and media practitioners. Our analysis of nearly one million documents from January 2015 to February 2023 reveals distinctive terms and phrases used by alternative media, highlighting the growing divergence between the mainstream and alternative media discourse and its intensity over different periods. These findings have implications for understanding the mobilization of anti-systemic groups, particularly those on the far right.

Keywords: European refugee crisis, Czech media, alternative journalism, corpus linguistics, media sociology, natural language processing, RASIM

1 BACKGROUND

Since 2015, Europe has experienced a significant influx of refugees, mainly from war-torn countries such as Syria, Iraq, and Afghanistan. The arrival of Ukrainian refugees following the Russian invasion in February 2022 prompted reflection on the lessons learned from the previous "crisis" and led to calls for comparative studies. At the same time, factors such as the prominence given to migration by different types of media and the nature of the discourse can influence public attitudes and whether migration is perceived as a crisis, which in turn affects support for populist right-wing parties (Strömbäck et al. 2021; Lecheler et al. 2019; Klawier et al. 2022). Previous studies on discourses surrounding refugees, asylum seekers, immigrants, and migrants (RASIM) focus on a few European countries, e.g. Germany and the United Kingdom (Eberl et al. 2018; Baker et al. 2008). However, there is a limited amount of scholarly attention on this topic in the Czech context,

which, as we argue, presents several compelling research opportunities. For example, the Czech media paid considerable attention to migration in and after 2015, despite the marginal number of asylum applications (Jelínková 2019). The situation changed significantly in February 2022, and by the end of that year, the country hosted the highest number of Ukrainian refugees per capita (Plevák 2023).

The Czech media landscape experienced significant changes in ownership concentration, political pressure, and technology in the 2000s. Online alternative[1] journalism has emerged, but often takes a hostile stance toward marginalized communities and liberal democracy (Holt et al. 2019; Štětka et al. 2021) and may spread disinformation about migration (Gregor – Mlejnková 2021). With only 34% of the Czech population trusting media (Newman et al. 2022), results may differ from Western European findings on media consumption and immigration attitudes (Kondor et al. 2022).

Scholars studying RASIM or ethnicity-based media coverage face challenges due to the large data volume; therefore, qualitative studies predominate (Seo – Kavakli 2022). In the Czech context, most studies either use qualitative methods (Průchová Hrůzová 2021) or small data sets from limited outlets (Kluknavská 2021), with an overreliance on manual content analysis (Esser et al. 2019). Using computational methods, such as clustering (Urbániková – Tkaczyk 2020) or corpus-assisted discourse analysis (Elmerot 2021, 2022), is rare. Natural language processing (NLP) tools have gained popularity in interdisciplinary studies, with landmark Törnbergs' (2016) combination of topic modeling and discourse analysis, and some Corpus Linguistics (CL) scholars advocate the utility of adding close reading to reveal the meaning behind automated results (Brookes – McEnery, 2019).

This study aims to analyse parts of the Czech media discourse on international migration in mainstream and alternative media, covering eight years from January 2015 to February 2023, focusing on three methods. The aim is to show when different types of media consider migration a crisis and to shed light on the role of the media in shaping migration representations and public opinion in the Czech Republic.

2 RESEARCH DESIGN

2.1 Background to the research questions

Firstly, we focus on the linguistic distinction between forced and voluntary migration in the Czech mainstream and alternative media. Word choice is crucial to understanding social group othering (Elmerot 2022), and word frequency analysis is

[1] We follow the definition by Holt et al. (2019, p. 862), which argues: "Alternative news media represent a proclaimed and/or (self-) perceived corrective, opposing the overall tendency of public discourse emanating from what is perceived as the dominant mainstream media in a given system."

an effective tool (Brouwer et al. 2017). We define 'refugee' as a label for people forced to leave their country for fear of persecution or harm, contrasting it with 'migrant,' who moves to improve their status (Douglas et al. 2019). Our first research question is:

RQ1: Between January 2015 and February 2023, what differences in usage of terms for voluntary versus forced migration are there between mainstream and alternative media?

Secondly, the power of individuals and organizations to shape media messages on migration is often emphasized in political communication literature, but studies rarely analyze their presence in the discourse (Boomgaarden – Vliegenthart 2009). Named Entity Recognition (NER) is a suitable method to find individuals and organizations, but rarely used with a migration focus (Nemes – Kiss 2021). Our second research question is:

RQ2: What are the differences in the presence of actors in the RASIM news coverage between mainstream and alternative media during periods of large refugee influxes in 2015 and 2022–2023?

Thirdly, collocation analysis is a CL tool that can provide insights into the linguistic patterns of discourse. Collocates frequently occur near the target word and can reveal significant associations between concepts (Stubbs 1995). Collocations have been widely used in RASIM studies, notably by Baker et al. (2008), who found eight consistent collocational categories. More recently, Zawadska-Paluektau (2023) analysed the portrayal of Ukrainian refugees in the Polish mainstream press during the first week of March 2022. She found that Ukrainians were more welcome in Poland than refugees from the Middle East and were more frequently referred to as war refugees. To better understand the framing (or "semantic preference") of crucial migration terms in the Czech media discourse, we examine seasonal collocations across media types and periods (Baker et al. 2008, pp. 278–286). Our final research question is:

RQ3: In the periods 2015–2016 vs. 2022–2023, what are the dominant collocates of the terms "refugee," "migrant," "immigrant," and "asylum seeker" in Czech mainstream and alternative media?

2.2 Corpus of Czech Media News on Migration

To obtain a comprehensive corpus of Czech migration-related news, we used the Newton Media Archive API to access full-length documents (including text articles and audio transcripts) with a Czech Boolean search adapted from Esser et al.

(2019).[2] The search was limited from 1st January 2015 to 28th February 2023 to capture at least 12 months of media coverage of each migration "crisis."[3]

The search parameters yield nearly one million documents from over four thousand Czech online and offline media (see Tab. 1). The total published content of all these media is 41 million documents. Our migration corpus thus represents a small fraction of the total media output. However, during the peak of interest in March 2022, migration was a topic in about 10 percent of all documents. Migration-related documents appear in all media sections, sometimes even in sports and technology.

Documents	Media	Sentences	Tokens
998,740	4,166	48,400,000	800,000,000

Tab. 1. Czech Migration News corpus, January 2015 to February 2023

2.3 Dataset with media type labels

To investigate the differences in migration discourse between different media types, we use the labels from the ONLINE2_NOW corpus of the Czech National Corpus (Cvrček et al. 2022). To adapt it to our research, we extended it by adding offline news according to the same key and then merged the categories "anti-system" and "political tabloid" (Cvrček – Fidler 2022, p. 268) into a broader category of alternative media, as some prominent media in the latter category (such as *Parlamentní Listy*) position themselves antagonistically to the mainstream (Štětka et al. 2021).

The media type and article datasets were merged, resulting in 2,735 labeled media. The resulting dataset of 971,000 documents with an identified media type includes approximately 189,000 alternative media documents (Tab. 2) and 342,000 mainstream media documents (Tab. 3).

Period	Documents	Media	Sentences	Tokens
whole	189,783	109	9.2M	156M
Jan.–Dec. 2015	14,016	27	0.7M	12.3M
Jan. 2016–22	161,220	90	7.6M	129.4M
Feb. 2022–2023	14,547	75	0.8M	14.2M

Tab. 2. Alternative Media sub-corpus, split by periods. M = million.

[2] The string: běženec* OR běženk* OR imigrant* OR migra* OR imigra* OR přistěhoval* OR uprchl* OR utečen* OR azylant*. (The English translation of *běženec,* the female form *běženkyně, uprchlík/uprchlice* and *utečenec,* are all 'refugee'. Both *imigrant* and *přistěhovalec* mean 'immigrant' in English, and *azylant* / *azylantka* 'male/female asylum seeker.') The only difference from the original string is the exclusion of stems related to integration, assimilation, and deportation, which yielded many irrelevant documents for our study.

[3] Our analysis workflow and additional visualizations are available in an open-source GitHub repository: https://github.com/opop999/media_discourse_research. Larger raw data files are available via the OSF repository: https://osf.io/j28v3.

Period	Documents	Media	Sentences	Tokens
whole	342,383	226	17M	276M
Jan.–Dec. 2015	48,145	109	1.8M	27.6M
Jan. 2016–22	222,858	222	9.6M	156.2M
Feb. 2022–2023	71,380	193	5.6M	92M

Tab. 3. Mainstream Media sub-corpus, split by periods. M = million.

Fig. 1 shows a distinct difference between the proportion of overall coverage devoted to migration in the mainstream (M = 4.3%, SD = 0.1%) and alternative media (M = 13.9%, SD = 2.5%). This contrast is most apparent in mid-2017, with alternative media focusing on migration in over 70% of their documents.

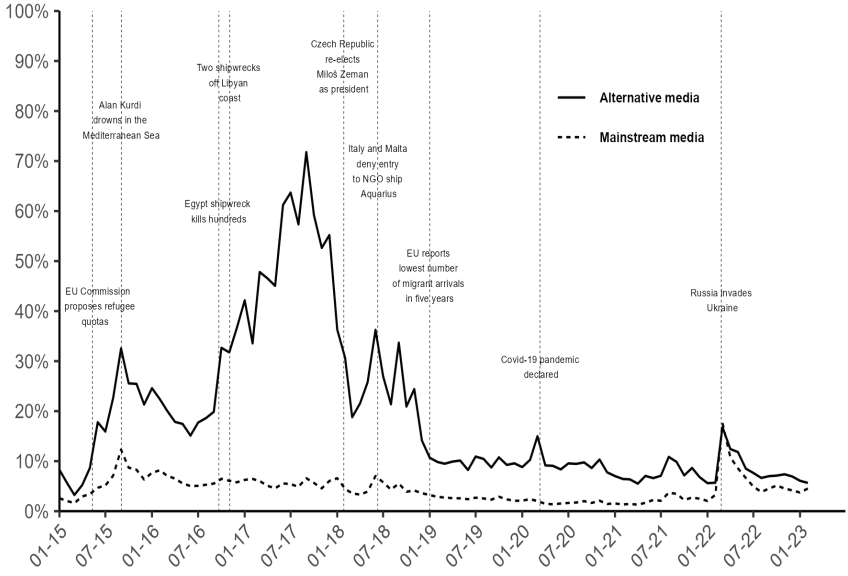

Fig. 1. The proportion of migration content across Czech media types

2.4 Data pre-processing

Czech is a highly inflected language with 13 word forms, making it challenging for automated natural language processing (Lenc – Hercig 2016). We apply several preprocessing steps to prepare the corpus for more computationally intensive analysis, including text cleaning and tokenization using R's UDPipe library (Wijffels 2023) and a pre-trained model based on the Czech PDT UD treebank 2.6 (Straka 2018). Lemmatization reduces words to their base forms, and a deep learning model generates embeddings to increase the accuracy of this process (Straka et al. 2019). Finally, part-of-speech tagging (POS) identifies the part of speech of each token.

3 ANALYSES AND FINDINGS

3.1 Word-frequency analysis of migration labels

To operationalize our voluntary versus forced migration study, we again used our migration string (see 2.2) and applied it to the lemmatized dataset. We then removed matching lemmas that appeared less than ten times in the entire corpus, resulting in 352 terms. Next, both authors manually reviewed each of these terms and categorized them as either "refugee term" (98 types), "migration term" (198 types), or "unknown" (56 types). We used this lexicon to examine their relative monthly frequency over the entire period, stratified by alternative media and mainstream categories.

Regarding RQ1, refugee terms were more frequently used in the mainstream media between February 2015 and May 2016, with a peak around the finding of the drowned boy Alan Kurdi (see Fig. 2). After that, migrant terms were used more frequently, with a peak around the start of the invasion of Ukraine in 2022. On the other hand, alternative media had a small peak of more refugee terms at the time of Alan Kurdi and again from March 2022 to mid-2022. Mainstream media used refugee terms for longer in both 2015 and 2022, while alternative media returned to using migrant terms more quickly.

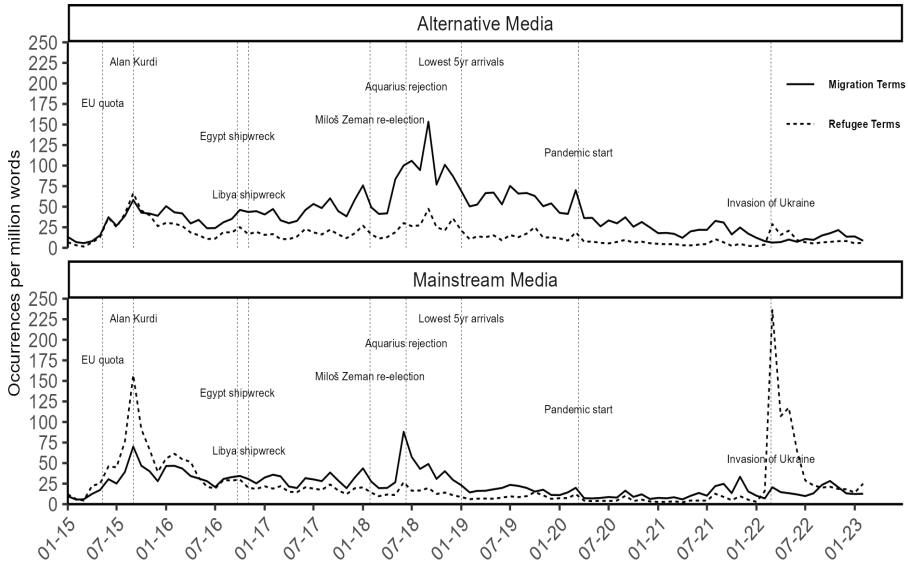

Fig. 2. Relative monthly frequencies of terms by media types

3.2 Named Entities in migration news coverage

We use Named Entity Recognition (NER) to identify and classify named individuals and organizations in our corpus, to the context influencing Czech migration discourse. The NameTag 2 model (Straková et al. 2019) identifies multiword entities from eight "parent" and forty-six "child" categories. Due to the inflectional nature of the Czech language, we use non-lemmatized text and apply stemming after analysis to obtain interpretable summaries of top entities in mainstream and alternative media. Our summary focuses on geographic locations, personal names, media organizations, and other institutions.

Regarding RQ2, the named entities of places and persons differ between media types. Five of the top thirty most frequent entities in alternative media in 2015 (see Fig. 3) are absent from mainstream media: "the West," NATO, China, the USA, and the Czech conservative party ODS. This result supports Cvrček and Fidler's (2022) conclusions that a Czech NATO exit was a part of the anti-systemic media discourse. Several entities have a higher frequency ranking in alternative media, such as Russia(n), ISIS, Ukraine, the USA, and Africa. Most entities unique to mainstream media are geographical. In the intermediate period (2016–2022), unique entities appear in the alternative media: Ukraine, NATO, "the West," and Africa. One unique entity in the mainstream media is Donald Trump, while the others are primarily geographical.

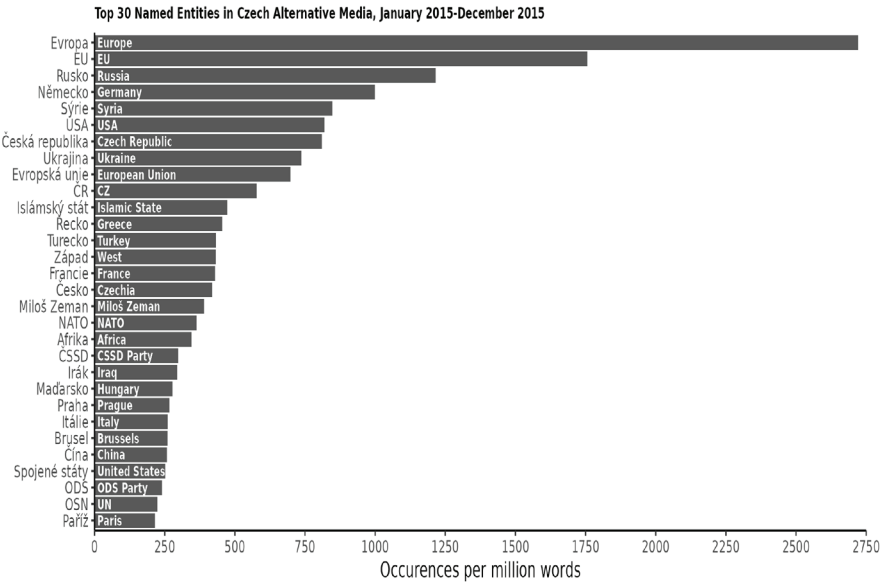

Fig. 3. Top named entities in alternative media, January to December 2015

In the final period (February 2022–2023), NATO is no longer unique in the alternative media. However, more unique entities in each media type may reveal a growing split between the two. Mainstream media refer to cities like Kharkiv and Mariupol, while alternative media refer to countries like Sweden and Turkey.

In summary, "the West" is one of alternative media's most frequent geographical and personal entities. Furthermore, Ukraine was among the top thirty most mentioned entities in alternative media already in 2015. A future study could thus further examine this pre-2022 invasion discourse better to understand the messaging of the prominent Kremlin-linked alternative media, similar to Cvrček and Fidler (2022). Overall, there are about twice as many unique entities in both media types during the last period (four to five in periods 1 and 2, eight and nine in period 3), pointing to a growing gap between the mainstream and alternative media types.

3.3 Collocations

To address RQ3, we compared the first period, 2015, with the third period, 2022–2023, using the collocates[4] of four lemmas: 'migrant', 'immigrant', 'refugee', and 'asylum seeker'. Here, we focus on the Provenance/transit/destination, Number, and Legality collocate groups, identified by Baker et al. (2008).

Provenance/transit/destination

In 2015, all four migration terms were associated with, e.g. the lemmata 'country,' 'Europe,' and 'state' in both media types. 'Migrant' and 'immigrant' were associated with 'border,' while 'refugee' was associated with 'Syria.' 'Migrant' was also associated with 'territory.' In alternative media, 'refugee,' 'migrant,' and 'immigrant' co-occurred with 'Africa.' In 2022–2023, only 'country' or 'land' (*země*) remained within the top thirty collocations for all four nouns in both media types. We observe a shift towards Ukraine and smaller geographical entities such as 'town' and 'region.' Europe now has a much looser association than in the first period. Africa appeared only in alternative media as a collocation of 'immigrant' and 'migrant.' Sweden emerged as a prominent geographical name in the same media type, especially for 'asylum seeker,' just as in the NER analysis.

Numbers and numerical lemmata

During the first period, the term 'million' co-occurs with 'refugee' in both media types, with mainstream collocations also including 'thousand.' However, neither numerical lemma is in the top thirty for 'asylum seeker.' Instead, we find 'number,' 'quota,' and 'contingent' in both media types. In the third period, 'quota' and 'contingent' are absent. 'Million' appears in alternative media with 'refugee.' In

[4] Like the korpus.cz's KonText tool, we use the logDice association measure, here with a context window of three words on each side.

contrast, ‘number’ appears in both media types around ‘migrant’ and ‘immigrant,’ but only in the mainstream media for ‘asylum seeker’ (see Fig. 4).

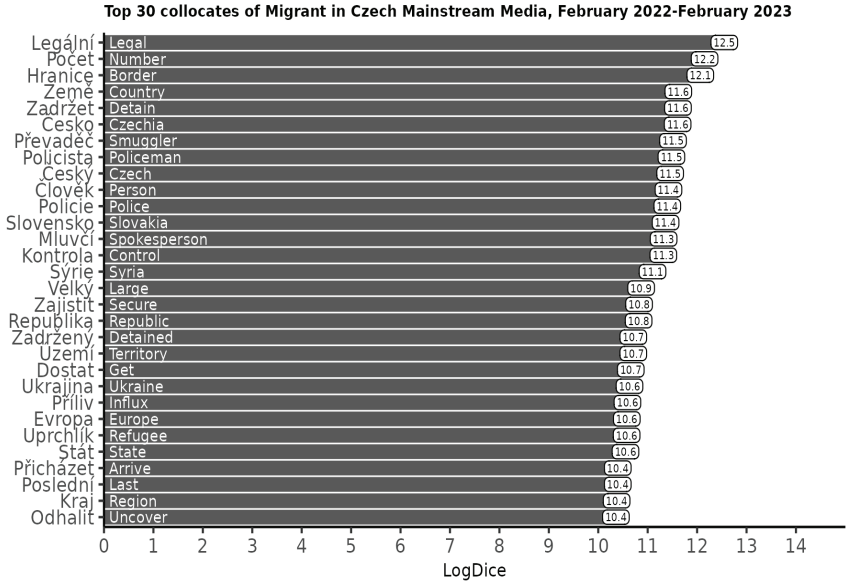

Fig. 4. Collocates with ‘migrant’, mainstream media, February 2022 to February 2023

Legality

In the first period, ‘illegal’ and ‘legal’ are found with ‘immigrant.’ At the same time, with ‘migrant,’ both collocations appear in alternative media but ‘legal’ only in the mainstream media. However, the non-lemmatized text may negate this as ‘illegal.’[5] In the third period, ‘illegal’ is the top collocate with ‘immigrant’ in alternative media and number 13 in mainstream media, but absent with ‘migrant.’

In response to RQ3, a radical shift in discourse concerning all four migration terms is visible between February 2022 and February 2023, with increased co-occurrence of Czech geographical names and neighboring countries. For example, the word ‘quota’ disappears from prominent discourses when Ukrainian refugees and asylum seekers arrive.

4 CONCLUSION

Our article examines differences in RASIM crisis discourse in nearly one million Czech alternative and mainstream media documents over eight years. Using CL and

[5] Czech negations may use the prefix ne-, which is removed when lemmatizing.

NLP methods informed by media sociology, we conclude that alternative media often use voluntary flight terminology and frequently refer to 'the West' as an actor.

In addition, Ukraine is continuously mentioned in a RASIM context in alternative media during 2015–2023. The gap between media types grows as more refugees arrive from Ukraine after February 24, 2022. Mainstream media then focus on Ukrainian cities affected by the war and Czech cities providing aid. At the same time, alternative media discuss international actors, e.g. 'the West,' Sweden, and Turkey. Language usage also changes, with 'illegal' collocating with 'immigrant' continuously but not with 'migrant' in 2022–2023.

Moreover, the 2022 crisis is portrayed as a reality, with increased mentions of Czech geographical names and neighboring countries in both media types compared to the more distant RASIM coverage of 2015–2016. These findings have implications for understanding far-right media communication and anti-immigrant sentiment and aim to inform policymakers and media practitioners in promoting responsible and informed migration reporting.

Our study highlights the need for further research on Czech migration discourse. For instance, verbal aspects connected to voluntary or forced migration could vary significantly between media types. Finally, linking social media data with news data and examining sources and co-occurrence origins in both media types could also shed light on how particular ideas spread in the public sphere.

ACKNOWLEDGEMENTS

The research has been supported by the *Gunvor och Josef Anérs Stiftelse* foundation, application number FB22-0088, and by the Specific University Research (SVV) grant no. 260 728.

References

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3), pages 273–306.

Boomgaarden, H. G., and Vliegenthart, R. (2009). How news content influences anti--immigration attitudes: Germany, 1993–2005. European Journal of Political Research, 48(4), pages 516–542.

Brookes, G., and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. Discourse Studies, 21(1), pages 3–21.

Brouwer, J., van der Woude, M., and Leun, van der, J. (2017). Framing migration and the process of crimmigration: A systematic analysis of the media representation of unauthorized immigrants in the Netherlands. European Journal of Criminology, 14(1), pages 100–119.

Cvrček, V., and Fidler, M. (2022). No Keyword is an Island: In search of covert associations. Corpora, 17(2), pages 259–290.

Cvrček, V., Jeziorský, T., and Henyš, J. (2022). ONLINE2_NOW: monitoring corpus of online Czech. Ústav Českého národního korpusu FF UK, Praha. Accessible at: http://www.korpus.cz.

Douglas, P., Cetron, M., and Spiegel, P. (2019). Definitions matter: Migrants, immigrants, asylum seekers and refugees. Journal of Travel Medicine, 26(2). Accessible at: https://doi.org/10.1093/jtm/taz005.

Eberl, J.-M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., Berganza, R., Boomgaarden, H. G., Schemer, C., and Strömbäck, J. (2018). The European media discourse on immigration and its effects: A literature review. Annals of the International Communication Association, 42(3), pages 207–223.

Elmerot, I. (2021). Income, nationality and subjectivity in media text. Jazykovedný časopis, 72(2), pages 667–678.

Elmerot, I. (2022). Constructing "Us" and "Them" through Conflicts – Muslims and Arabs in the News 1990–2018. In L. Filardo-Llamas – E. Morales-López – A. Floyd (eds.): Discursive Approaches to Sociopolitical Polarization and Conflict (1st ed.), pages 122–136. Routledge.

Esser, F., Stępińska, A., Pekáček, O., Seddone, A., Papathanassopoulos, S., Peicheva, D., Milojevic, A., Blassnig, S., and Engesser, S. (2019). Event-, Politics-, and Audience-Driven News: A Comparison of Populism in European Media Coverage in 2016 and 2017. In C. Reinemann – J. Stanyer – T. Aalberg – F. Esser – C. de Vreese (eds.), Communicating populism: Comparing actor perceptions, media coverage, and effects on citizens in Europe, pages 123–140. Routledge: Taylor & Francis Group.

Gregor, M., and Mlejnková, P. (2021). Facing Disinformation: Narratives and Manipulative Techniques Deployed in the Czech Republic. Politics in Central Europe, 17(3), pages 541–564.

Holt, K., Figenschou, T. U., and Frischlich, L. (2019). Key dimensions of alternative news media. Digital Journalism, 7(7), pages 860–869.

Jelínková, M. (2019). A Refugee Crisis Without Refugees: Policy and media discourse on refugees in the Czech Republic and its implications. Central European Journal of Public Policy, 13(1), pages 33–45.

Klawier, T., Prochazka, F., and Schweiger, W. (2022). Comparing Frame Repertoires of Mainstream and Right-Wing Alternative Media. Digital Journalism, 10(8), pages 1387–1408.

Kondor, K., Mihelj, S., Štětka, V., and Tóth, F. (2022). News consumption and immigration attitudes: A mixed methods approach. Journal of Ethnic and Migration Studies, 48(17), pages 4129–4148.

Lecheler, S., Matthes, J., and Boomgaarden, H. (2019). Setting the Agenda for Research on Media and Migration: State-of-the-Art and Directions for Future Research. Mass Communication and Society, 22(6), pages 691–707.

Nemes, L., and Kiss, A. (2021). Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic. Applied Sciences, 11(22). Accessible at: https://doi.org/10.3390/app112211017.

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., and Nielsen, R. K. (2022). Reuters Institute Digital News Report 2022.

Plevák, O. (2023, February 24). Czechia hosts most Ukrainian refugees per capita. https://www.euractiv.com/section/politics/news/czechia-hosts-most-ukrainian-refugees-per--capita/

Průchová Hrůzová, A. (2021). What is the image of refugees in Central European media? European Journal of Cultural Studies, 24(1), pages 240–258.

Seo, S., and Kavakli, S. B. (2022). Media representations of refugees, asylum seekers and immigrants: A meta-analysis of research. Annals of the International Communication Association, 46(3), pages 159–173.

Štětka, V., Mazák, J., and Vochocová, L. (2021). "Nobody Tells us what to Write about": The Disinformation Media Ecosystem and its Consumers in the Czech Republic. Javnost – The Public, 28(1), pages 90–109.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207.

Straka, M., Straková, J., and Hajič, J. (2019). Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing (arXiv:1908.07448). arXiv.

Straková, J., Straka, M., and Hajič, J. (2019). Neural Architectures for Nested NER through Linearization. Proceedings of the 57th Annual Meeting of the Assoc. for Computational Linguistics, pages 5326–5331.

Strömbäck, J., Meltzer, C. E., Eberl, J.-M., Schemer, C., and Boomgaarden, H. G. (2021). Media and Public Attitudes Toward Migration in Europe: A Comparative Approach. Routledge.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 2(1), pages 23–55.

Törnberg, A., and Törnberg, P. (2016). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. Discourse, Context & Media, 13, pages 132–142.

Urbániková, M., and Tkaczyk, M. (2020). Strangers ante portas: The framing of refugees and migrants in the Czech quality press. European Journal of Communication, 35(6), pages 580–596.

Wijffels, J. (2023). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the "UDPipe" "NLP" Toolkit. Accessible at: https://CRAN.R-project.org/package=udpipe.

Zawadzka-Paluektau, N. (2023). Ukrainian refugees in Polish press. Discourse & Communication, 17(1), pages 96–111.

sciendo
DOI 10.2478/jazcas-2023-0054

SLOVAK QUESTION ANSWERING DATASET BASED ON THE MACHINE TRANSLATION OF THE SQUAD V2.0

JÁN STAŠ[1] – DANIEL HLÁDEK[1] – TOMÁŠ KOCTÚR[2]
[1] Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
[2] Deutsche Telekom IT & Telecommunications Slovakia s.r.o., Košice, Slovakia

STAŠ, Ján – HLÁDEK, Daniel – KOCTÚR, Tomáš: Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 381 – 390.

Abstract: This paper describes the process of building the first large-scale machine-translated question answering dataset SQuAD-sk for the Slovak language. The dataset was automatically translated from the original English SQuAD v2.0 using the Marian neural machine translation together with the Helsinki-NLP Opus English-Slovak model. Moreover, we proposed an effective approach for the approximate search of the translated answer in the translated paragraph based on measuring their similarity using their word vectors. In this way, we obtained more than 92% of the translated questions and answers from the original English dataset. We then used this machine-translated dataset to train the Slovak question answering system by fine-tuning monolingual and multilingual BERT-based language models. The scores achieved by EM = 69.48% and F1 = 78.87% for the fine-tuned mBERT model show comparable results of question answering with recently published machine-translated SQuAD datasets for other European languages.

Keywords: language modeling, machine reading comprehension, machine translation, natural language processing, question answering

1 INTRODUCTION

In the field of natural language processing (NLP), understanding text and answering questions about this text is an important and challenging task for machines. Question answering (QA) is a conversational task in NLP that retrieves correct answers to questions asked by humans in natural language, which is useful for searching for an answer in a text document. Machine reading comprehension (MRC) is a type of QA in which a machine can read any text, understand it, and answer questions about that text. Therefore, the relationship between QA and MRC is very close and is often altered.

Recent methods for textual QA rely on large-scale human-annotated datasets. Existing QA datasets are primarily built in English. The most popular benchmark datasets for evaluating QA models is the Stanford Question Answering Dataset (SQuAD). It contains more than 100k pairs of questions and answers posed by crowd workers on a set of Wikipedia articles. The answers to each question are a text fragment from the corresponding reading paragraph of the article (Rajpurkar et al. 2016). The SQuAD v1.1 dataset was later

updated to the SQuAD v2.0, which includes additional 50k questions, labeled as unanswerable concerning the corresponding target text, and forces QA models to understand when a question cannot be answered given the context (Rajpurkar et al. 2018).

Building large-scale resources is labor-intensive and expensive for others, especially low-resourced languages. However, the unavailability of such datasets makes it challenging to train multilingual QA systems with performance comparable to the English one.

One solution is to apply machine translation (MT) to an existing QA dataset. The advantage is that it is fast and does not require human labor and finances. That is why we decided to create a QA dataset for the Slovak language based on machine translation of the original English SQuAD v2.0 dataset. The dataset will be publicly available at.[1]

The paper is organized as follows. In Section 2 we briefly describe current approaches to creating QA datasets using machine translation and summarize 25 existing non-English QA datasets that have been created in recent years. In Section 3, we describe in detail the proposed approach to create a machine-translated QA dataset for the Slovak language. Next, we fine-tuned several available monolingual and multilingual BERT (Bidirectional Encoder Representations from Transformers) language models on a machine-translated SQuAD-sk dataset and evaluate them. Finally, Section 5 summarizes the contribution of our work and concludes the paper with future directions.

2 STATE-OF-THE-ART

For the task of question answering, many SQuAD-like datasets have been created in the last ten years. Most of them were created only for English. By the end of 2022, we would have identified 25 machine-translated SQuAD datasets in 21 other languages. The primary data source was the English SQuAD in version 1.1, after 2019 also in version 2.0, containing a translation of unanswerable questions. Tab. 1 summarizes the basic information about them.

When creating MT datasets, authors encounter the following two problems:

- the errors caused by the machine translation;
- correct alignment of the translated answers in the translated paragraph context.

For accurate machine translation, authors often use Google Translate. In rare cases, other MT systems, such as DeepL for Finnish and Italian, Moses for Spanish and Danish, LINDAT for Czech, Yandex for Russian, and Marian for German were used.

In this case, each paragraph is thus automatically translated with all relevant questions and marked answers. Some of these triplets cannot be translated due to unknown characters or low lexical coverage of the MT system. Also, the MT system can some-times be unable to recognize named entities without context and transliterate them. Furthermore, translations of the questions can result in multiple translation variants (Lee et al. 2020).

[1] https://huggingface.co/TUKE-DeutscheTelekom

Due to the differences in nature and complexity between languages, the precision of machine translation varies greatly. Some languages produce better translation results than others. Machine translation works fine for most Latin and Western languages, but the accuracy of some under-resourced African or Asian languages is often poor.

Language	Title	Translation	v2.0	Size	Reference
Arabic	Arabic-SQuAD	Google Translate	N	60k	Mozannar et al. 2019
Brazilian Portuguese	BraSQuAD	Google Cloud API	Y	-	Esposito 2020[2]
Bengali	Bengali-SQuAD	Google Cloud API	Y	91k	Mayeesha et al. 2021
Czech	Czech-SQuAD	LINDAT Translator	Y	117k	Macková – Straka 2020
Danish	SQuAD-da	TAR Method/Moses	Y	156k	Carrino et al. 2020[3]
Finnish	SQuAD_v2_fi	DeepL	Y	-	Kylliäinen 2022[4]
French	SQuAD-fr	Google Translate	N	100k	Cattan et al. 2021
French	French-SQuAD	Google Translate	N	-	Kabbadj 2019[5]
German	German SQuAD	Marian NMT	N	-	Giagoulas 2020[6]
Hindi	Hindi SQuAD	Google Translate	N	18.5k	Gupta et al. 2019
Indonesian	SQuAD-id	Google Translate	-	-	Wikidepia 2021[7]
Italian	SQuAD-it	DeepL	N	60k	Croce et al. 2018
Korean	K-QuAD	Google Translate	N	77k	Lee et al. 2018
Latvian	SQuAD-lv	-	Y	-	Tilde 2019[8]
Persian	ParSQuAD	Google Translate	Y	95k	Abadani et al. 2021
Polish	PolAQ	Google Cloud API	N	14k	Jurkiewicz 2020[9]
Polish	SQuAD-pl	Google Translate	Y	-	Brodzik 2022[10]
Portuguese	SQuAD_v1.1_pt	Google Cloud API	N	100k	Carvalho 2019[11]
Portuguese	SQuAD_v2.0_pt	Google Cloud API	Y	-	Janiake 2020[12]
Russian	SQuAD-ru	Yandex Translate	N	-	Semyonov 2017[13]
Spanish	SQuAD-es-v1.1	TAR Method/Moses	N	87.5k	Carrino et al. 2020
Spanish	SQuAD-es-v2.0	TAR Method/Moses	Y	46k	Carrino et al. 2020
Swedish	SQuAD-v2-sv	Google Translate	Y	125k	Okazawa 2021[14]
Turkish	SQuAD (Tr)	Google Translate	N	-	Gemirter – Goularas 2021
Ukrainian	SQuAD-uk	Google Cloud API	N	30k	Tiutiunnyk – Dzomkin 2019

Tab. 1. Summary of available machine-translated SQuAD datasets

[2] https://github.com/piEsposito/br-quad-2.0
[3] https://github.com/ccasimiro88/TranslateAlignRetrieve/tree/multilingual/squads-tar/da
[4] https://github.com/ilmarikyl/SQuAD_v2_fi
[5] https://github.com/Alikabbadj/French-SQuAD
[6] https://github.com/agiagoulas/german_squad
[7] https://github.com/Wikidepia/SQuAD-id
[8] https://github.com/tilde-nlp/SQuAD-LV
[9] https://github.com/tjur/polaq_master_thesis
[10] https://github.com/brodzik/SQuAD-PL
[11] https://github.com/nunorc/squad-v1.1-pt
[12] https://github.com/cjaniake/squad_v2.0_pt
[13] https://github.com/lightforever/SQUAD_RU
[14] https://github.com/susumu2357/SQuAD_v2_sv

On the other hand, to verify whether the translation between the languages has been done correctly, it is approached either by laborious manual correction or by using back translation, which is the conversion of a translated text back to its original language. Using the back translation and then comparing the original document with the back translation, we can gauge how closely the translation mirrors the meaning of the original text. This also gives us feedback on how accurate the translation between languages is.

Back translation can also be used when creating MT datasets without any additional grammar correction of the translated questions and answers. An example of this approach is the Ukrainian SQuAD-uk or Polish SQuAD-pl dataset.

The second issue is the problem with alignment after translation. It occurs when the start and end points of the answers to the relevant questions are not correctly marked in a paragraph or if the links between words in the adjacent context are broken due to translation. Therefore, it is necessary to look for the correct translated answer in the paragraph context. This is usually done by:

- using special tags (for example quotation marks) that are inserted before the machine translation (may break the context);
- using attention of the MT system and choosing the words aligned to the source answer (very difficult);
- using an approximate search for the translated answer and measuring the minimal distance between translated answer and translated context, for example using Levenshtein distance or by cosine similarity between these two word vectors.

3 SLOVAK TRANSLATION OF THE SQUAD V2.0

3.1 Machine translation from English to Slovak

To translate the original English SQuAD v2.0 into Slovak, we used an efficient and freely available neural MT framework Marian (Junczys-Dowmunt et al. 2018) together with the Helsinki-NLP Opus English-Slovak model[15] (Tiedemann – Thottingal 2020). The BLEU score for this model was evaluated at 36.8.

The process of machine translation can be described as follows (Hládek et al. 2023):

- we loaded the original English SQuAD v2.0 in JSON file format and transformed it into a set of triplets: paragraphs, questions, and answers;
- for each paragraph, question, and answer in English, we obtained their translations into Slovak;
- we searched for the translated answer in the translated paragraph; the result is part of the paragraph that contains the translated answer;

[15] https://huggingface.co/Helsinki-NLP/opus-mt-en-sk

- in validation, we eliminated the triplets where the answer could not be found;
- we compiled the translated paragraphs, questions, and answer areas into a new Slovak SQuAD-sk dataset.

3.2 Proposed approach for the approximate search of the translated answer

The SQuAD dataset contains the exact position of the answer in the context of the paragraph. Marks for the start and end of the answer are not suitable for being inserted into the translated context because unknown and unexpected tags can disrupt the context and reduce the quality of the translation. They can also get lost in the translation process. Therefore, it is necessary to translate the answer separately and look for it again in the translated paragraph.

The main disadvantage is that the independent translation of the answer can be affected by the missing context since the answer that was translated on its own may differ from the answer in the translated context.

When searching for an answer in the paragraph context, we follow this procedure:

- exact search for the original answer;
- exact search for the translated answer;
- approximate search for the translated answer.

First, we searched for the exact occurrence of the original, untranslated answer in the translated paragraph. We started from the assumption that the most well-known named entities for persons, organizations and geographical locations will not be changed by machine translation. If we did not find the original answer, we tried to find a translated answer. This option handles cases where the translation of the answer is unambiguous and does not depend on the context. If we did not find the answer in this case either, then its translation was influenced by the context, and the independently translated form differed from the form occurring in the paragraph context. Since Slovak is a morphologically rich language, the answer in the translated context will likely contain word forms different from those of the independently translated answer.

Therefore, we proposed an approximate search that would identify the part of the answer that is semantically similar. We assume that a translation in context will be similar in meaning to a translation without context but may differ by using a different synonym or morphological form. We also assume that the number of words in the answer in context will be the same as in the case of the answer without context.

We evaluated the semantic similarity of two texts a_O, and a_P using word vectors. If the answer is one word, the distance between the word vectors will be smaller than the threshold T as follows:

$$D\left(V\left(a_o\right), V\left(a_p\right)\right) \leq T,$$

where V is the vector of texts a_O and ap, respectively, and D is the distance metric between the two word vectors. If the answer is multi-word, we choose the arithmetic mean of the word vectors of all words as the semantic representation.

As a distance metric between two word vectors, we chose cosine distance, which is the negative value of cosine similarity. Cosine similarity is defined as the cosine angle of two word vectors and is often used in determining semantic similarity.

We used the spaCy library (Honnibal et al. 2020) because it allows easy use of the word vector model and the method for approximate search using cosine similarity. We created a model of word vectors using fastText (Bojanowski et al. 2016) with the Floret[16] subword units. The word vector model was trained on a web corpus of Slovak written texts and created as part of the Slovak spaCy model available at.[17]

We searched in the paragraph context using a floating window. We set the size of the floating window to the same as the number of words in the translated answer. For the floating window, we calculated its average word vector. Next, we calculated the cosine distance between the mean word vectors of the translated answer and the floating window.

We selected the window that had the minimum distance from the translated answer. In this way, we found the translated answer in the context. We ignored distances greater than the constant T to limit spurious results. Only in a few cases could no region with a similarity less than T be found. We discarded these answers.

A visual inspection of the translated database revealed that most errors were caused by the translation itself, sometimes not all words were found, and at other times the answer did not grammatically match the question. These types of errors are usually caused by differences between the Slovak and English languages. The easiest way is to identify such grammatically incorrect answers and eliminate them. However, a deeper analysis would be needed in this area, which may be the subject of further research.

However, in most cases, the result was understandable and grammatically correct. The number of answers that could not be found by an approximate search in the paragraph context is very low (see Tab. 2).

Number of	**SQuAD v2.0**	**SQuAD-sk**	**% of original**
Documents	442	442	100
Paragraphs	19,035	18,931	99.45
Questions	130,319	120,239	92.27
Answers	86,821	79,978	92.12
Unanswerable questions	43,498	40,261	92.56

Tab. 2. Statistics on the English SQuAD v2.0 and the MT Slovak SQuAD-sk datasets

[16] https://github.com/explosion/floret

[17] https://github.com/hladek/spacy-skmodel

Note that the proposed approach is applicable to another language that has a trained word vector model and for which an MT model exists. The approach is independent of the MT system; any other translation model can be used.

4 EXPERIMENTAL RESULTS

The experiment was focused on the evaluation of the performance of several available monolingual and multilingual pretrained BERT models after fine-tuning on a machine-translated SQuAD-sk dataset in the task of automatically answering questions in Slovak. For this purpose, we created two evaluation datasets. The first consisted of a machine-translated English SQuAD development set into Slovak. The second consisted of 9,583 hand-annotated questions for 7,822 answers marked in 2,568 paragraphs, which were obtained from 940 Slovak documents published on Wikipedia. The annotation was performed by crowd workers, in a similar way as it was in the original English SQuAD.

We used the following types of pre-trained BERT models that support Slovak:

- our own monolingual Slovak RoBERTa-base cased model;
- monolingual SlovakBERT-base cased model (Pikuliak et al. 2022);
- multilingual mBERT-base cased model (Devlin et al. 2019);
- multilingual XLM-RoBERTa-base cased model (Conneau et al. 2020).

To evaluate the fine-tuned BERT models, we used the exact match (EM) and F1 scores. EM measures the percentage of predictions that match any reference answers at the token level. The F1 score is the harmonic mean of precision and recall, where precision measures the ratio of correct tokens in the prediction and recall rates the ratio of the correct tokens in the prediction to the correct response.

Training set	**Evaluation set**	**SK RoBERTa**		**SlovakBERT**		**mBERT**		**XLM-R**	
		EM	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
MT SQuAD-sk	machine-translated	52.50	58.73	56.02	62.78	56.02	63.02	56.56	62.91
MT SQuAD-sk	hand-annotated	48.16	57.75	62.97	71.14	69.48	**78.87**	**69.76**	78.03

Tab. 3. Results of the Slovak question answering

The experimental results of the Slovak question answering for selected BERT models fine-tuned on the SQuAD-sk dataset are summarized in Tab. 3. It should be noted that the number of parameters of pre-trained models was approximately the same in all cases.

As we can see from these results, the huge multilingual models outperform the monolingual models. This is because the multilingual models have a vocabulary several times larger than monolingual models and can therefore better capture the context of answers in paragraphs. To improve monolingual models, it would be necessary to significantly increase the size of the training data or the number of model parameters.

5 CONCLUSION

SQuAD-sk is the first machine-translated QA dataset in Slovak. The experimental results show the usefulness of the dataset for monolingual and multilingual QA. It will mainly contribute to the creation of new systems to generate answers to questions in natural language. The dataset will be published free of charge for scientific use.

In the future, we would like to focus more on errors and biases caused by machine translation to improve the SQuAD-sk dataset. Further research should focus on machine translation of other types of QA datasets that do not only contain factoid questions but also multiple-choice or causal questions.

ACKNOWLEDGEMENTS

This research was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, and the Slovak Academy of Sciences under the project VEGA 2/0165/21 funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic; and by the Slovak Research and Development Agency under the grants APVV-SK-TW-21-0002, APVV-22-0261, and APVV-22-0414.

Our special thanks go to Deutsche Telekom IT & Telecommunications Slovakia s.r.o. for fruitful cooperation and personal and financial support.

References

Abadani, N., Mozafani, J., Fatemi, A., Nematbakhsh, M., and Kazemi, A. (2021). ParSQuAD: Persian question answering dataset based on machine translation of SQuAD 2.0. International Journal of Web Research, 4(1), pages 34–46.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. Trans. of the ACL, Vol. 5, Cambridge, MA, pages 135–146. Accessible at: https://aclanthology.org/Q17-1010.pdf.

Carrino, C. P., Costa-Jussa, M. R., and Fonollosa, J. A. R. (2020). Automatic Spanish translation of the SQuAD dataset for multilingual question answering. In Proc. of LREC, Marseille, France, pages 5515–5523. Accessible at: https://arxiv.org/abs/1912.05200.

Cattan, O., Servan, C., and Rosset, S. (2021). On the usability of transformers-based models for a French question-answering task. In Proc. of RANLP, Varna, Bulgaria, pages 244–255. Accessible at: https://hal.archives-ouvertes.fr/hal-03336060/.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proc. of ACL, Online, pages 8440–8451. Accessible at: https://aclanthology.org/2020.acl-main.747.pdf.

Croce, D., Zelenanska, A, and Basili, R. (2018). Neural learning for question answering in Italian. In C. Ghidini – B. Magnini – A. Passerini – P. Traverso (eds): Advances in Artificial Intelligence, LNAI vol. 11298, Springer, Cham, pages 389–402. Accessible at: https://doi.org/10.1007/978-3-030-03840-3_29.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL, Minneapolis, Minnesota, pages 4171–4186. Accessible at: https://aclanthology.org/N19-1423/.

Germirter, C. B., and Goularas, D. (2021). A Turkish question answering system based on deep learning neural networks. Journal of Intelligent Systems: Theory and Applications, 4(2), pages 65–75. Accessible at: https://dergipark.org.tr/tr/download/article-file/1361881.

Gupta, D., Ekbal, A., and Bhattacharyya, P. (2019). A deep neural network framework for English Hindi question answering. ACM TALLIP, 19(2), Article No. 25, pages 1–22.

Hládek, D., Staš, J., Juhár, J., and Koctúr, T. (2023). Slovak dataset for multilingual question answering. IEEE Access, Vol. 11, pages 32869–32881. Accessible at: https://ieeexplore.ieee.org/document/10082887.

Honnibal, M., Montani, I., Landeghem van, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Accessible at: doi 10.5281/zenodo.1212303.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In Proc. of ACL, Melbourne, Australia, pages 116–121. Accessible at: https://aclanthology.org/P18-4020.pdf.

Lee, K., Yoon, K., Park, S., and Hwang, S. W. (2018). Semi-supervised training data generation for multilingual question answering. In Proc. of LREC, Miyazaki, Japan, pages 2758–2762. Accessible at: https://aclanthology.org/L18-1437.

Macková, K., and Straka, M. (2020). Reading comprehension in Czech via machine translation and cross-lingual transfer. In Proc. of TSD, Brno, Czech Republic, pages 171–179. Accessible at: https://arxiv.org/abs/2007.01667.

Mayeesha, T. T., Sarwar, A. Md., and Rahman, R. M. (2021). Deep learning based question answering in Bengali. Journal of Information and Telecommunication, 5(2), pages 145–178. Accessible at: https://doi.org/10.1080/24751839.2020.1833136.

Mozannar, H., El Hajal, K., Maamary, E., and Hajj, H. M. (2019). Neural Arabic question answering. In Proc. of WANLP, Florence, Italy, pages 108–118. Accessible at: https://arxiv.org/abs/1906.05394v1.

Pikuliak, M., Grivalský, Š., Konôpka, M., Blšták, M., Tamajka, M., Bachratý, V., Šimko, M., Balážik, P., Trnka, M., and Uhlárik, F. (2022). SlovakBERT: Slovak masked language model. In Proc. of EMNLP, Abu Dhabi, United Arab Emirates, pages 7156–7168. Accessible at: https://aclanthology.org/2022.findings-emnlp.530.pdf.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proc. of EMNLP, Austin, Texas, pages 2383–2392. Accessible at: https://aclanthology.org/2021.emnlp-main.530.pdf.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In Proc. of ACL, Melbourne, Australia, pages 784–789. Accessible at: https://aclanthology.org/P18-2124.pdf.

Tiedemann, J., and Thottingal, S. (2020). OPUS-MT – Building open translation services for the Worlds. In Proc. of EAMT, Lisboa, Portugal, pages 479–4810. Accessible at: https://aclanthology.org/2020.eamt-1.61.pdf.

Tiutiunnyk, S., and Dyomkin, V. (2019). Context-based question-answering system for the Ukrainian language. In Proc. of MS-AMLV, Lviv, Ukraine, pages 81–88. Accessible at: https://ceur-ws.org/Vol-2566/MS-AMLV-2019-paper17-p081.pdf.

sciendo
DOI 10.2478/jazcas-2023-0055

SYLLABIC CONSONANTS IN HISTORICAL CZECH AND HOW TO IDENTIFY THEM

MARKÉTA ZIKOVÁ – MARTIN BŘEZINA
– RADEK ČECH – PAVEL KOSEK
Department of Czech Language, Faculty of Arts, Masaryk University,
Brno, Czech Republic

ZIKOVÁ, Markéta – BŘEZINA, Martin – ČECH, Radek – KOSEK, Pavel: Syllabic Consonants in Historical Czech and How to Identify Them. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 391 – 400.

Abstract: The paper provides fine-grained evidence concerning the development of syllabic consonants /r l/ in Czech, that is only sketched in the existing literature. The evidence is based on an automatic parser that identifies potential syllable-projecting segments according to sonority. The parser was applied to six verse texts from the 14th–16th centuries, which show a strong tendency towards octosyllabicity. The data provided by the parser newly reveal that the shift from non-syllabic to syllabic /r l/ is position-dependent: word-medial non-syllabic strings C(r/l)C change more rapidly than non-syllabic word-final ones C(r/l)#. This finding is in line with a cross-linguistic observation that non-syllabic C(r/l)C are marked, hence they are regularly syllabified prior to less marked C(r/l)#.

Keywords: syllabic consonants, historical Czech, syllable markedness, Sonority Sequencing Principle

1 INTRODUCTION

The paper investigates syllabification algorithm throughout the history of Czech. We focus on sonorants /r l/ (henceforth R) whose syllable status varied considerably during the 14th–16th centuries, as reported in the literature (Gebauer 1963; Komárek 1982; Lamprecht et al. 1986). We examine three contexts shown in Tab. 1 in which R is C(onsonant)-adjacent, but not vowel-adjacent. According to the literature, there is an asymmetry between word-initial #RC on the one side and word-medial CRC and word-final CR# on the other: the former are syllabified uniformly in historical Czech, the latter variously. Word-initial #RC never project syllables; e.g. *lhal* ‘he lied’ and *rval* ‘he tore’ are both monosyllabic, and the number of syllables (σ) corresponds to the number of vowels. By contrast, word-final CR# underwent a change towards syllable-projecting structures. For example, final R of *nesl* ‘he carried’ and *Petr* ‘Peter’ received the same syllabic status as a preceding vowel, i.e., they became syllabic consonants; originally monosyllabic words *nesl* and *Petr* thus turned into bisyllabic ones. Finally, word-medial CRC vary between

syllabic and non-syllabic, depending on particular lexical items; cf. monosyllabic *slza* 'tear' or *krvi* 'blood.gen.pl' on the one hand (where only the vowels project syllables) and bisyllabic *vlna* 'wool' or *brzo* 'soon' on the other (where both the vowels and R are syllable-projecting). This lexical contrast has however been eliminated: all word-medial R eventually became syllabic.

#RC	lh**a**$_{\sigma}$l, rv**a**$_{\sigma}$l	=	lh**a**$_{\sigma}$l, rv**a**$_{\sigma}$l	'he lied, tore'
CR#	n**e**$_{\sigma}$sl, P**e**$_{\sigma}$tr	->	n**e**$_{\sigma}$s**l**$_{\sigma}$, P**e**$_{\sigma}$t**r**$_{\sigma}$	'he carried, Peter'
CRC	slz**a**$_{\sigma}$, krv**i**$_{\sigma}$	->	s**l**$_{\sigma}$z**a**$_{\sigma}$, k**r**$_{\sigma}$v**i**$_{\sigma}$	'tear, blood.gen.pl'
CRC	v**l**$_{\sigma}$n**a**$_{\sigma}$, b**r**$_{\sigma}$z**o**$_{\sigma}$	=	v**l**$_{\sigma}$n**a**$_{\sigma}$, b**r**$_{\sigma}$z**o**$_{\sigma}$	'wool, soon'

Tab. 1. Syllable structure of /r l/ in historical Czech (14th–16th century)

The aim of this paper is to verify the above-mentioned claims, made by the historical grammars. Syllabic consonants are quite easy to be detected in the contemporary language: we can simply ask native speakers how they syllabify words like *lhal, nesl* or *slza*. This method cannot, of course, be applied in investigating historical Czech because we rely on written records. Moreover, the investigation is complicated by the fact that the R-syllabicity is not marked consistently by any graphic means in the historical texts. Thus, we work with the idea that the syllabic structure of R can be seen well in syllable-based poetry.

The idea of examining poetry to learn about syllable structure is not new, it has been put forward already in the literature mentioned above, and more recently for example, in Scheer and Ziková (2017). However, to our knowledge, there is no empirical study that thoroughly examines behavior of non-vowel-adjacent R in historical verses, which was one of the motivations for this pilot study.

The paper is organized as follows. In Section 2, we introduce the Sonority Sequencing Principle governing syllabification according to sonority of phonological segments. Section 3 describes implementation of this principle into an automatic sonority parser that identifies potential syllable-projecting segments according to sonority. The parser enables us to pick up all the instances of potential syllabic R in the contexts CRC and CR#. In Section 4, we show and discuss the results we got by applying the parser to six syllabic verse texts from the 14th to the 16th century. In Section 5, the results of our research are discussed.

2 SONORITY SEQUENCING PRINCIPLE

In derivational approaches to phonology, syllabification algorithm is governed by the Sonority Sequencing Principle (Selkirk 1984; Clements 1990), according to which the syllable structure is derived in terms of sonority.

As for sonority, two major categories of segments are identified, i.e., vowels and consonants, the latter being further subcategorized into obstruents and sonorants.

These three sonority categories form a hierarchy, shown in Fig. 1, in which sonorants (R) are between the more sonorous vowels (V) and the less sonorous obstruents (T).

\+ V >> R >> T −

Fig. 1. Sonority hierarchy: V>>R>>T

The Sonority Sequencing Principle (SSP) postulates that the sonority of a syllable decreases from the nucleus towards the margins, i.e., to onset and coda. Thus, according to this principle, syllable nuclei are the sonority peaks of words. And since it is the nucleus that constitutes a syllable, the number of syllables of a word is equal to the number of (syllabic) nuclei in the word, and, transitively to the number of sonority peaks. These default sonority-syllable correspondences are summarized in Fig. 2.

number of syllables = number of nuclei = number of sonority peaks

Fig. 2. Sonority-syllable correspondences

From the perspective above, the diachronic evolution in which sonority peaks based on sonorants /r l/ (henceforth R-peaks) are syllabified as nuclei, can be understood as a path towards optimal syllabification. As illustrated in Fig. 3, the newly created bisyllabic structures *nesl* 'he carried' and *slza* 'tear' are fully in accordance with the sonority-syllable correspondences predicted by the SSP. Each of the sample words has two relative sonority peaks (shaded): one is created by a vowel and the other by a sonorant. Both the V-peak and R-peak project syllable nuclei (N_σ) regularly, yielding thus optimal bisyllabic structures.

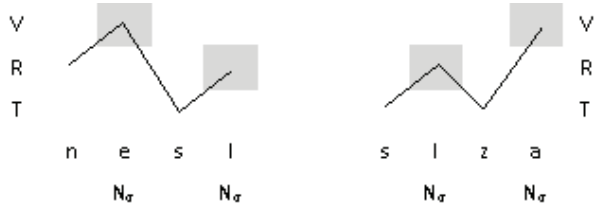

Fig. 3. Optimal syllabification: syllabic CR# and CRC

The R-peaks in CRC and CR#, which pattern with V-peaks in terms of syllabification, contrast with R-peaks in #RC. In a mono-syllabic word *lhal* 'he lied', for example, only the V-peak (occupied by *a*) projects the nucleus, but not the initial R-peak. Furthermore, a comparison of the monosyllabic *lhal* (on the left in Fig. 4) and the bisyllabic *udal* 'he provided' (on the right) reveals that V-peaks always project nuclei, even word-initially.

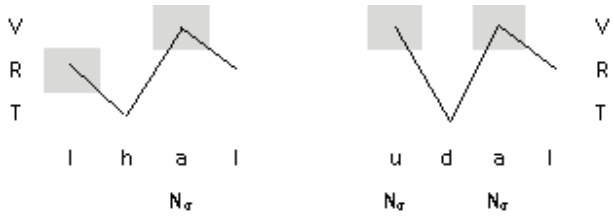

Fig. 4. Non-syllabic #RC: R-peak ≠ Nσ vs syllabic #VC: V-peak = Nσ

To sum up, the idea we pursue is that syllable structure of words is read off from their sonority profiles: the universal principle is that sonority peaks are syllabified as nuclei. While V-peaks are nuclei by default, R-peaks are syllabified upon language-specific parameters. To put it simply, all languages feature syllabic vowels, but only a subset of them have syllabic consonants.

In the certain stage of history of Czech, Czech had the parameter on syllabic consonants set positively: R-peaks do project syllable nuclei. However, the projection of a syllabic nucleus by an R is dependent on R's position (word-initial R-peaks never project syllables), and, also, the situation changes over time (in case of final and medial R-peaks). In the next section, we test these assumptions on a relatively large historic data sample.

3 AUTOMATIC SONORITY PARSER

To see the sonority peaks and their contribution to the syllable structure, we need to annotate words (and syllables) with respect to their sonority. For this purpose, we created an automatic sonority parser.[1] The parser has two main ingredients: the sonority hierarchy and the inventory of phonological segments classified according to the sonority scale.

The hierarchy embodied in the parser categorizes sounds into 7 sonority levels. As shown in Tab. 2, this fine-grained hierarchy, based on Parker (2011), identifies three subclasses of sonorants; the liquids /r l/ (that are our main concern in this paper), are in the 'middle', surrounded by the less sonorous nasals and the more sonorous glides. Thus, /r l/ are sonorants of level 4 in Tab. 2, and hence we refer to them as R_4 from now on.

[1] The parser is available at https://github.com/cechradek/analysis_of_syllables_in_old_czech/blob/main/02SLABIKA_sonoritni_profily.py.

[2] Line-by-line glosses: 'tear, elm'; 'to provide'; 'apple'; 'wool, brother'; 'elm, he carried'; 'wool, tear'; 'brother'.

son-level	phon-class	phon-subclass	example[2]
7	V	low and mid vowels	slz**a**, jil**e**m
6	V	high vowels	**u**dat**i**
5	R	glides	**j**ablko
4	R	liquids	v**l**na, b**r**at**r**
3	R	nasals	jile**m**, **n**esl
2	T	fricatives	**v**lna, **sl**z**a**
1	T	plosives	**b**rat**r**

Tab. 2. 7-point sonority hierarchy

The second ingredient the parser considers is a sonority-annotated set of segments that were part of the phonological system of the 14th–16th century. For convenience, the segmental inventory, compiled from the historical grammars (see Section 1 for the references), is shown in Tab. 3 (IPA annotated), and it is supplemented by the corresponding graphemes.[3] (We should add that the table displays only those graphemes that are recorded in the edited texts analyzed in this paper.)[4]

son-level	segments	graphemes
7	/a aː e eː o oː/	a á e ě é o ó
6	/i iː u uː/[4]	i y í ý u ú ů
5	/j/	j
4	/r l/	r r’ l l’
3	/m n ɲ/	m n ň
2	/f v s z ʃ ʒ ɾ x ɦ/	f v s z š ž ř ch h
1	/p b t d t͡s t͡ʃ c ɟ k g/	p b t d c č ť ď k g

Tab. 3. Segmental inventory of 14th–16th century

In Fig. 5, the outputs provided by the parser are illustrated. The diagram displays the sonority profile of a word *milosrdenstvie* ‘mercy’: the segments (on the horizontal axis) were mapped onto the 7-point sonority hierarchy (on the vertical axis).

[3] The segmental inventory underwent several changes in the 14th–16th century, which however did not result in reordering of segments with respect to the sonority levels. The consonant inventory was simplified to the extent that palatalized consonants merged with their plain counterparts, e.g. /nʲ/ merged with either /n/ or /ɲ/. Since the input and the output of this diachronic change (called depalatalization in the historical literature) are always on the same sonority level, only the output segments are involved in the parser. The similar strategy was used for the vocalic part of the parser: the reported historical merger of high vowels (of any length) /i y/ is represented by the output /i/.

[4] In addition to monophthongs, high vowels were involved in rising diphthongs /uo/ and /ie/, the latter having been a reflex of a Common Slavic long vowel, called *jať*; for details see Kosek and Ziková (2022). In terms of syllable structure, these diphthongs count as a single nucleus, similarly to monophthongs. That is exactly what is predicted by the fine-grained sonority scale: since high vowels sit lower in the sonority hierarchy than mid vowels, only the latter project the V-peaks in the diphthongs /uo/ and /ie/. See also Fig. 5.

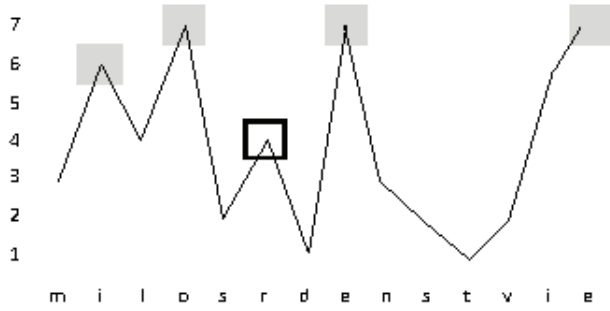

Fig. 5. Sonority profile of *milostrdenstvie* 'mercy'

For the given word (***milo****srd****e****ntsv****ie***) shown in the diagram, there are 5 vowels that correspond to 4 sonority peaks (shaded); all of them are either V_6-peaks or V_7-peaks. The difference in the number of the vocalic segments and the sonority peaks is due to the word-final bi-vocalic string *ie*: *e* is more sonorous than *i,* and therefore projects a single V_7-peak. This is in line with the fact that *ie* is one of the two rising diphthongs in OCz; see also footnote 4. In addition to the vocalic peaks, *milosrdenstvie* contains a consonantal peak (squared): it is a R_4-peak, projected by /r/. (Notice that there is yet another instance of the R_4 in the given word, i.e., /l/. Since it appears intervocalically, in this position it does not project a sonority peak.)

Diagrams generated by the parser (as the one above), are inputs for the analysis of syllable structure. As mentioned previously, the analysis pursues the idea that V-peaks are syllable nuclei by default, but the syllabification of consonantal peaks is parameterized. The parameter is assumed to be set so that the consonantal peaks of type R_4 can be syllabified as nuclei in a certain point in the history of Czech. This means that the word *milosrdenstvie* either could have five or four syllables in the examined historical period, depending on whether the /r/ was syllabified the same way as the V-peaks – or not.

To conclude, the sonority parser automatically identifies words with R_4-peaks that could project syllable nuclei in various stages in the history of Czech. To identify syllable-projecting R_4-peaks, i.e., syllabic instances of /r l/, we have been tracing the behavior the words with R_4-peaks show in the syllable-counting verses. The method used, as well as the results are described in the next section.

4 DATA FROM THE SONORITY PARSER

In the 14th–16th century, much of the poetry is syllable-counting; thus, verses have a regular number of syllables – and, therefore, a regular number of nuclei. In particular, the most common verse is octosyllable (Jakobson 1932). We thus follow this line of thinking: the appearance of a word like *milosrdenstvie* in the 8-peak verse

indicates that its R_4-peak /r/ counts as a verse unit, i.e., projects the syllable nucleus – to keep the octosyllabic rhythm.

We applied this method to a corpus of 15,837 verses extracted from six texts from the 14th–16th century; see the details listed in Tab. 4. The right-most column displays the total number of verses.[5]

source text	century	number of verses
Kunhutina modlitba [KunM] 'Kunhuta's Prayer'	14th	154
Alexandreis – Budějovický fragment [AlexB]	14th	342
Život Svaté Kateřiny [SvKat] 'The Life of Saint Catherine'	14th	3,518
Alexandreis – Svatovítský fragment [AlexSv]	15th	2,462
Hádání Prahy s Kutnou Horou [Had] 'Disputation between Praha and Kutná Hora'	15th	2,989
Instrukcí Šimona Lomnického z Budče [Lom] 'Instructions by Simon Lomnicky'	16th	6,372

Tab. 4. Analyzed texts

The texts were run through the parser, and the data provided by the parser confirm a strong tendency towards octosyllable. As shown in Tab. 5, the proportion of 8-peak verses, identified by the parser, does not decline below 85%.

source	total number of verses	proportion of 8-peak verses
KunM	154	93%
AlexB	342	95%
SvKat	3,518	89%
AlexSv	2,462	85%
Had	2,989	95%
Lom	6,372	97%

Tab. 5. Proportion of 8-peak verses

Given their octosyllabic nature, the selected texts are thus a good ground for verifying the assumptions of the historical grammars. In particular, we assume that: 1. #RC are never syllabic, 2. CR# turn from non-syllabic to syllabic, 3. CRC vary between syllabic and non-syllabic parsing.

[5] The analyzed corpus is built on the following critical editions: *Kunhutina modlitba* [KunM]; available at https://vokabular.ujc.cas.cz/moduly/edicni/seznam-edic/datace-asc/strana [14/03/ 2023]; Vážný, V. (1963). *Alexandreida* [AlexB; AlexSv]. Praha: Nakladatelství československé akademie věd; Hrabák, J., and Vážný, V. (1959). *Dvě legendy z doby Karlovy* [SvKat]. Praha: Nakladatelství Československé akademie věd; Daňhelka, J. (1952). *Husitské skladby Budyšínského rukopisu* [Had]. Praha: Orbis; Heřmanská, K. (2016), *Instrukcí aneb Krátké naučení hospodáři mladému Šimona Lomnického z Budče (edice a literárně historický rozbor)* [Lom]. Master Thesis, UK Praha.

The absolute non-syllabicity of #RC is clearly confirmed: we recorded 40 words with initial R_4-peaks and neither of them is involved in the 8-peak verse as a syllabic nucleus.

As for CR#, they predominantly occur in 9-peak verses: 190 vs 34 in 8-peak verses. Since 9-peak verses do exist in the analyzed texts, we might simply conclude that 9-peak verses with final R_4-peaks (found in words like *mysl* 'mind', *mistr* 'master' or *spadl* 'he fell') violate the octosyllabic rhythm. Under this assumption, thus, the 9-peak verses are indeed 9-syllabic, and the word-final syllabic /r l/ are projected as an extra syllable on top of the octosyllable. The second possible approach is to take 9-peak verses as an indicator of non-syllabicity of R_4-peaks. Under this approach, 9-peak verses follow the octosyllabic rhythm regularly, because the final R_4-peaks are not syllabified as nuclei. From this perspective, thus, the attested words like *mysl, mistr* and *spadl* are monosyllabic, and only the V-peaks project nuclei.

We favour the latter approach, thus, 9-peak verses contain non-syllabic R_4-peaks; there are two reasons for that. First, the analyzed texts have a strong tendency to – indeed – be octosyllabic (proven by the proportions in the Tab. 5 above). Moreover, 9-peak verses tend to contain R_4-peaks: in our corpus, more than 60% of 9-peak verses include R_4-peaks.

In sum, the considerable difference between syllabic and non-syllabic CR# (34 vs. 190) suggests that the postulated diachronic process resulting in syllabicity of CR# proceeded relatively slowly in the examined period of 14th–16th century and that most of the final R_4-peaks remained non-syllabified.[6]

Following the same logic, we classify CRC as syllabic or non-syllabic according to their distribution in 8-peak and 9-peak verses. In this case, the distributional discrepancy between the two classes is not as sharp as the one in CR#. This, however, is expected by the historical grammars: many instances of the syllabic CRC were inherited from Proto-Czech, hence syllabic CRC could appear even in the texts from the very beginning of the 14th century This expectation is validated by the data: one of the two earliest texts in our corpus, i.e., *Kunhutina modlitba*, contains 5 instances of syllabic CRC in 8-peak verses (***krm****iteľu* [v.16] 'feeder.voc.sg', ***prv****niemu* [v.67] 'first.dat.sg', *čt****vrt****ému* [v.70] 'forth.dat.sg', ***mrt****vých* [v.80] 'dead.gen.pl', ***krm****ě* [v.87] 'food.nom.sg').

Similarly to CR#, CRC are supposed to gravitate to syllabicity in the course of time – and that should be manifested in two things. First, we expect the original syllabic CRC to retain their syllabic status, hence appearing in 8-peak verses. This holds for all the four syllabic CRC-roots bolded above. They are involved in various word types and tokens that are distributed quite evenly throughout our corpus, and they predominantly appear in 8-peak verses.

[6] It is not surprising that 33 out of 34 instances of syllabic CR# are found in the texts from 15th and 16th century, but not in earlier texts from the 14th century. Though, even in the later texts, the non-syllabic CR# are still prevalent.

The second hypothesis is that non-syllabic CRC-roots should turn into syllabic roots. This diachronic change is well-documented by the root *krv* 'blood', recorded quite widely in our corpus. The data in Tab. 6 show an obvious shift in the distribution towards 8-peak verses, in which, as we assume, the root *krv* is syllabic.

source	8-peak verses	9-peak verses
KunM	0	1
AlexB	0	2
SvKat	1	7
AlexSv	2	7
Had	10	0
Lom	4	0

Tab. 6. Tokens with the root *krv*

To conclude, the data confirm the hypotheses concerning the development of syllabic R: word-initial #RC are not syllabic at all, word-medial CRC and word-final CR# are both non-syllabic or syllabic. Through the examined period of the 14th–16th century, there is a tendency towards R-syllabicity, which is, however, stronger in the word-medial context than in the word-final context.

5 CONCLUSIONS

In this paper, we investigated development of syllabic consonants in Czech. We followed a view that syllabic consonants – in accordance with vowels – are sonority peaks that project syllable nuclei. On this backdrop, we were tracking the sonority peaks formed by sonorants /r l/ and observed their behavior in syllable-counting verses: if they contribute to the regular octosyllabic rhythm, they are proven to be syllabic.

We created a sonority parser that automatically identifies /r l/ as sonority peaks in three contexts: #RC, CR#, and CRC. The parser was applied to 6 verse texts from 14th–16th century (all of the texts have a strong tendency towards the octosyllable). The data provided by the parser were analyzed according to the following criteria: 1. the octosyllable corresponds to 8 sonority peaks; 2. if one of the 8 peaks is /r/ or /l/, they are syllabic consonants; 3. if the octosyllabic verse has 9 sonority peaks and one of them is /r/ or /l/, then the /r/ or /l/ are not syllabic consonants. The synoptic picture of the data we obtained is shown in Tab. 7.

	non-syllabic	syllabic	tendency to R-syllabicity
CRC	158	968	high
CR#	190	34	low
#RC	40	0	none

Tab. 7. Syllabicity of R-peaks

The results of our pilot study reveal a distinction between CRC and CR# contexts and this observation, to our knowledge, has not yet been documented: CR# show a weaker tendency towards syllabicity than CRC. In other words, the change from non-syllabic to syllabic R proceeded more rapidly in word-medial position than in word-final position.

The revealed dynamic of the change of syllabic structure is in line with a cross-linguistic observation made by Cyran (2010): non-syllabic CRC are more marked typologically than non-syllabic CR#. In other words, the more marked non-syllabic CRC turn to regular syllabic structures more rapidly than the less marked non-syllabic CR#.

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation (grant No. 23-04719S *Development of Syllabic Sonorants in Czech*).

References

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston – M. Beckman (eds.): Papers in laboratory phonology 1: between the grammar and physics of speech. Cambridge: Cambridge University Press, pages 283–333.

Cyran, E. (2010). Complexity scales and licensing in phonology. Berlin, New York: De Gruyter Mouton.

Gebauer, J. (1963). Historická mluvnice jazyka českého I. Hláskosloví. 2nd ed. Praha: ČSAV.

Jakobson, R. (1932). Verš staročeský. In Československá vlastivěda 3 – Jazyk. Praha: Sfinx, pages 429–459.

Komárek, M. (1982). Nástin fonologického vývoje českého jazyka. Praha: SPN.

Kosek, P., and Ziková, M. (2022). Czech Vowel Fronting (Česká Přehláska). In M. L. Greenberg (ed.): Encyclopedia of Slavic Languages and Linguistics Online. Leiden: Brill. Accessible at: https://referenceworks.brillonline.com/browse/encyclopedia-of-slavic-languages-and-linguistics-online.

Lamprecht, A., Šlosar, D., and Bauer, J. (1986). Historická mluvnice češtiny. Praha: SPN.

Parker, S. (2011). Sonority. In M. Oostendorp et al. (eds.): The Blackwell Companion to Phonology. Vol. II: Suprasegmental and Prosodic Phonology. Oxford: Blackwell, pages 1160–1184.

Scheer, T., and Ziková, M. (2017). Branching onsets in Old Czech. In O. Mueller-Reichau – M. Guhl (eds.): Aspects of Slavic Linguistics: Formal Grammar, Lexicon and Communication. Berlin: De Gruyter, pages 285–309.

Selkirk, E. (1984). On the Major Class Features and Syllable Theory. In M. Aronoff – R. Oehrle (eds.): Language Sound Structure. Cambridge, Mass.: MIT Press, pages 107–136.

POKYNY PRE AUTOROV

Redakcia JAZYKOVEDNÉHO ČASOPISU uverejňuje príspevky **bez poplatku** za publikovanie.
Akceptované jazyky: všetky slovanské jazyky, angličtina, nemčina. Súčasťou vedeckej štúdie a odborného príspevku je abstrakt v angličtine (100 – 200 slov) a zoznam kľúčových slov v angličtine (3 – 8 slov).
Súčasťou vedeckej štúdie a odborného príspevku v inom ako slovenskom alebo českom jazyku je zhrnutie v slovenčine (400 – 600 slov) – preklad do slovenčiny zabezpečí redakcia.
Posudzovanie príspevkov: vedecké príspevky sú posudzované anonymne dvoma posudzovateľmi, ostatné príspevky jedným posudzovateľom. Autori dostávajú znenie posudkov bez mena posudzovateľa.
Technické a formálne zásady:

- Príspevky musia byť v elektronickej podobe (textový editor Microsoft Word, font Times New Roman, veľkosť písma 12 a riadkovanie 1,5). V prípade, že sa v texte vyskytujú zvláštne znaky, tabuľky, grafy a pod., je potrebné odovzdať príspevok aj vo verzii pdf alebo vytlačený.
- Pri mene a priezvisku autora je potrebné uviesť pracovisko.
- Text príspevku má byť zarovnaný len z ľavej strany, slová na konci riadku sa nerozdeľujú, tvrdý koniec riadku sa používa len na konci odseku.
- Odseky sa začínajú zarážkou.
- Kurzíva sa spravidla používa pri názvoch prác a pri uvádzaní príkladov.
- Polotučné písmo sa spravidla používa pri podnadpisoch a kľúčových pojmoch.
- Na literatúru sa v texte odkazuje priezviskom autora, rokom vydania a číslom strany (Horecký 1956, s. 95).
- Zoznam použitej literatúry sa uvádza na konci príspevku (nie v poznámkovom aparáte) v abecednom poradí. Ak obsahuje viac položiek jedného autora, tie sa radia chronologicky.

Bibliografické odkazy:

- knižná publikácia: ONDREJOVIČ, Slavomír (2008): *Jazyk, veda o jazyku, societa*. Bratislava: Veda, vydavateľstvo SAV. 204 s.
- slovník: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.) (2011): *Slovník súčasného slovenského jazyka. H – L.* [2. zv.]. Bratislava: Veda, vydavateľstvo SAV.
- štúdia v zborníku: ĎUROVIČ, Ľubomír (2000): Jazyk mesta a spisovné jazyky Slovákov. In: S. Ondrejovič (ed.): *Sociolinguistica Slovaca 5. Mesto a jeho jazyk*. Bratislava: Veda, vydavateľstvo SAV, s. 111 – 117.
- štúdia v časopise: DOLNÍK, Juraj (2009): Reálne vz. ideálne a spisovný jazyk. In: *Jazykovedný časopis*, roč. 60, č. 1, s. 3 – 12. DOI 10.2478/v10113-009-0001-3. [cit. DD-MM-RRRR].
- internetový zdroj: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV. Dostupné na: https://korpus.juls.savba.sk [cit. DD-MM-RRRR].

INSTRUCTION FOR AUTHORS

JOURNAL OF LINGUISTICS publishes articles **free of publication charges**.
Accepted languages: all Slavic languages, English, German. Scientific submissions should include a 100-200 word abstract in English and a list of key words in English (3-8 words).
Scientific articles in a language other than Slovak or Czech should contain a summary in Slovak (400-600 words) – translation into Slovak will be provided by the editor.
Reviewing process: scientific articles undergo a double-blind peer-review process and are reviewed by two reviewers, other articles by one reviewer. The authors are provided with the reviews without the name of the reviewer.
Technical and formal directions:

- Articles must be submitted in an electronic form (text editor Microsoft Word, 12-point Times New Roman font, and 1.5 line spacing). If the text contains special symbols, tables, diagrams, pictures etc. it is also necessary to submit a pdf or printed version.
- Contributions should contain the full name of the author(s), as well as his/her institutional affiliation(s).
- The text of the contribution should be flush left; words at the end of a line are not hyphenated; a hard return is used only at the end of a paragraph.
- Paragraphs should be indented.
- Italics is usually used for titles of works and for linguistic examples.
- Boldface is usually used for subtitles and key terms.
- References in the text (in parentheses) contain the surname of the author, the year of publication and the number(s) of the page(s): (Horecký 1956, p. 95).
- The list of references is placed at the end of the text (not in the notes) in alphabetical order. If there are several works by the same author, they are listed chronologically.

References:

- Monograph: ONDREJOVIČ, Slavomír (2008): *Jazyk, veda o jazyku, societa*. Bratislava: Veda, vydavateľstvo SAV. 204 p.
- Dictionary: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.) (2011): *Slovník súčasného slovenského jazyka. H – L.* [2. zv.]. Bratislava: Veda, vydavateľstvo SAV.
- Article in a collection: ĎUROVIČ, Ľubomír (2000): Jazyk mesta a spisovné jazyky Slovákov. In: S. Ondrejovič (ed.): *Sociolinguistica Slovaca 5. Mesto a jeho jazyk*. Bratislava: Veda, vydavateľstvo SAV, pp. 111–117.
- Article in a journal: DOLNÍK, Juraj (2009): Reálne vz. ideálne a spisovný jazyk. In: *Jazykovedný časopis*, Vol. 60, No. 1, pp. 3–12. DOI 10.2478/v10113-009-0001-3. [cit. DD-MM-RRRR].
- Internet source: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV. Available at: https://korpus.juls.savba.sk [cit. DD-MM-YEAR].

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Objednávky a predplatné prijíma/Orders and subscriptions are processed by:
SAP – Slovak Academic Press, s. r. o., Bazová 2, 821 08 Bratislava
e-mail: sap@sappress.sk

Registračné číslo 7044

Evidenčné číslo 3697/09

IČO vydavateľa 00 167 088

Ročné predplatné pre Slovensko/Annual subscription for Slovakia: 12 €, jednotlivé číslo 4 €
Časopis je v predaji v kníhkupectve Veda, Štefánikova 3, 811 06 Bratislava 1