sciendo
DOI 10.2478/jazcas-2019-0066

IMPROVING NOMINALIZED ADJECTIVES TAGGING

KLÁRA OSOLSOBĚ[1] – HANA ŽIŽKOVÁ[1]
[1]Faculty of Arts, Masaryk University, Brno, Czech Republic

OSOLSOBĚ, Klára – ŽIŽKOVÁ, Hana: Improving nominalized adjectives tagging. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 370 – 379.

Abstract: Part of speech transitions represent an interesting issue in terms of Automatic Morphological Analysis (AMA). In these cases, two parts of speech have to be considered: initial and final. However, their automatic recognition is complicated by the same form. This article presents the results of a corpus study aimed at mapping nominalized adjectives tagging with a focus on detecting candidates for nominalization among frequent adjectives. Analysis of the data obtained from the ČNK SYN v5 corpus shows different reasons for incorrect tagging. Taking into account these reasons, we propose three solutions for the improvement nominalized adjectives tagging.

Keywords: nominalized adjectives, automatic morphological analysis, disambiguation, corpus, tagging

1 INTRODUCTION

The division of vocabulary units into parts of speech is crucial for a systematic description of the language. Traditionally, in the classification of the part of speech, the synthesis of three criteria is based on formal, syntactic and semantic. However, in the case of natural language processing, we can only proceed from the form of the analyzed unit. This is because the automatic morphological analysis, which assigns units to part of speech, works mostly with the formal criterion of determining the part of speech. The syntactic and semantic criterion is sometimes used in disambiguation, but the rules are often difficult to formalize. In the case of part of speech transitions, the form is identical, and for this reason, the tagging is challenging.

There are three types of part of speech transitions [1]:

1. The initial and final part of speech is non-flexible. For example conjunctions → particle: *Prší, **ale** svítí slunce.* (conjunction) 'It's raining but the sun is shining.' vs. *To **ale** prší!* (particle) 'But it rains!'.

2. The initial part of speech is flexible, the final part of speech is non-flexible. For example noun → adverb: *Zadíval se na **modro** vod.* (noun) 'He looked at the blue of waters.' vs. *Obarvil látku na **modro**.* (adverb) 'He dyed the fabric in blue.'.

3. Both the initial and the final part of speech are flexible. For example adjective → noun: *Petr je **nemocný**.* (adjective) 'Peter is sick.' vs. ***Nemocný** se uzdravil.* (noun) 'The sick recovered.'.

This paper focuses on the third type of part of speech transition: nominalized adjectives.

Nominalized adjectives, sometimes also called syntactic nouns, have the same form and inflexion as adjectives, but syntactically they behave as a noun [2]. In this article, however, we do not distinguish nouns with adjectival inflexion (e.g. *mluvčí* 'speaker') and nominalized adjectives (e.g. *popravčí* 'executioner'). We refer to all analyzed units as nominalized adjectives because both groups have the same adjectival inflexion and the same syntactic distribution of nouns.

2 APPROACH

We carried out a corpus study with the intention of mapping how the nominalized adjectives are tagged and which units can be included in the group of nominalised adjectives. We chose to use the largest available corpus at the time, SYN v5 ČNK (3,836 billion words)[3].

We proceeded in several steps. First, all possible endings of the nominalised adjectives were defined using *Slovník afixů užívaných v češtině* [4] and available Czech grammar books ([1], [2], [9], [10], [11]). CQL queries were then formulated to obtain the lists of nouns and adjectives with defined endings. These were compared, and the accuracy of the tagging was evaluated. The first 600 most frequent adjectives were checked for nominalized adjectives. Subsequently, we tried to find a key, how to classify analysed data so that the classification is relevant for automatic morphological analysis. The frequency of use, the context, and the occurrence in dictionaries were taken into account. Also, the assignment to a semantic group was taken into account.

After applying the listed steps, 319 nominalized adjectives were selected and subjected to a detailed analysis. The tagging of all selected units was observed in context. If the unit had been tagged incorrectly, we were curious about why this error occurred and whether it was possible to set a rule that could be used to tag part of speech correctly/properly. We have focused on the most frequent collocation of analysed units.

In the study we did not intentionally include zoological and botanical terms (*vrubozobí* – Anseriformes, *blanokřídlí* – Hymenoptera etc.). We did not follow up the proper nouns, we only focused on the common nouns. Only the positive forms of adjectives were taken into consideration. Also a relatively large and open group of nouns type *Kladenští* 'Kladno inhabitants' was left aside [2].

3 FINDINGS

Analysis of the data shows that errors in the tagging of nominalized adjectives are due to two reasons in particular: inaccuracies in the morphological dictionary and erroneous disambiguation.

3.1 Inaccuracies in the morphological dictionary

There seem to be four types of inaccuracies in the morphological dictionary. There are some nouns (e.g. *šipkovaná*; 'treasure hunt') and adjectives (e.g. *hokejbalová* 'hockeyball', *jatečné* 'slaughter') which have been entered incorrectly as both POS = N and POS = A. Some adjectives (e.g. *basiliánský* 'basilian', *stehová* 'stitched') have been incorrectly entered as POS = N. We also found that many units which have only one interpretation, POS=A, are actually nominalized adjectives and can be used as a noun or an adjective (e.g. *vyučující* 'teacher', *popravčí* 'executioner') depending on the context. Similarly, other units only have the POS = N interpretation, but they can also be an adjective, POS=A (e.g. *košíková* 'basketball').

3.2 Erroneous disambiguation

Erroneous disambiguation leads to incorrect tagging as a noun instead of an agreeing postnominal or prenominal adjective. In the case of an agreeing postnominal adjective, such as *švihák lázeňský*; 'spa dude' (215 occurencies), we recorded 140 cases tagged incorrectly. Table 1 shows similar examples with incorrect disambiguation of other agreeing postnominal adjectives.

Problémem ale může být nedostatečné **pojistné**/pojistné/N **krytí** nebo nepřizpůsobitelnost parametrů pojištění (...) (SYN v5) 'However, the problem may be insufficient insurance coverage or non-adaptability of insurance parameters (…)'
(...) stal se ze mě **švihák lázeňský**/lázeňský/N. (SYN v5) (…) 'I became a spa dude.'
Dalším jídlem, které porotě předložily, bylo **kuřecí**/kuřecí/N **prsíčko** se špenátovou fáší (...) (SYN v5) 'Another meal presented to the jury was a chicken breast with spinach (...)'
Mám **hovorné**/hovorné/N **prodavače** rád. (SYN v5) 'I like talkative salespeople.'
V listopadu jsem pozvána do poroty další **taneční**/taneční/N **soutěže** Miss Belly dance, už se moc těším. (SYN v5) 'In November I was invited to the jury of another Miss Belly dance competition, I am looking forward to it.'

Tab. 1. Examples of erroneous disambiguation

We recorded erroneous disambiguation in cases where the unit precedes a proper noun:

Se závěrečným hvizdem **rozhodčího**/rozhodčí/A **Samka** tak vypukla na novopackém stadionu obrovská radost (...) (SYN v5) 'With the final whistle of the referee Samko so broke out at the stadium in Nová Paka great joy (...)'

Sousedé se jednou sešli v hospodě U Švejka, **hostinský**/hostinský/A **Petr Spittank** rozdal noty a 14. ročník dětských radovánek byl na světě. (SYN v5) The neighbors once met at the U Svejka pub, the innkeeper Petr Spittank gave out notes and the 14th year of children's fun was born.
Vzpomínky na natáčení má i jeho **příbuzná**/příbuzný/A **Hana Ševčíková**. (SYN v5) 'Also his relative Hana Ševčíková has memories of the shooting.'
(...) ve spolupráci s naší redakcí připravily výherní akci o půlroční **předplatné**/předplatné/A **MF DNES**. (SYN v5) '(…) in cooperation with our editorial team, they prepared the winning event for a six-month subscription to MF DNES.'
Že se děti nemůžou dočkat prázdnin konstatovala i její **třídní**/třídní/A **Eva Oherová**. (SYN v5) 'Even her class teacher Eva Oherová stated that children could not wait for the holidays.'

Tab. 2. Examples of units preceded by a proper noun

We also noticed the erroneous disambiguation if the unit was preceded by the lemmas *pán* 'mister' and *paní* 'missis':

Po hodině hledání ve skladu jim **pan vedoucí**/vedoucí/A přišel říci, že jejich pohovku nemohou najít (...) (SYN v5) 'After an hour of searching in the warehouse, (Mr.) supervisor came to tell them they couldn't find their sofa (...)'
Ani na to jim obezřetná **paní domácí**/domácí/A neskočila. (SYN v5) 'Even the prudent (Mrs.) landlady did not get to it.'
Po jeho odjezdu mně **paní představená**/představený/A citovala jeden z jejich rozhovorů. (SYN v5) 'After his departure, (Mrs.) Lady Superior quoted me one of their interviews.'
Pan vrátný/vrátný/A zakryl rukou sluchátko a řek mi, že to volá divadlo Šumperk. (SYN v5) '(Mr.) porter covered the handset with his hand and told me it calls the Šumperk theater.'
Tentokrát je nebohý **pan účetní**/účetní/A po smrti a stojí frontu před nebeskou bránou. (SYN v5) 'This time, the poor (Mr.) accountant is dead and faces the front of the heavenly gate.'

Tab. 3. Examples of units preceded by lemmas *pan* 'mister' and *paní* 'missis'

In rare cases, it seemed that the lemma and tag were incorrect:

Třídní/Třídeň/NNFS7-----A----- se zatvářila jako jeptiška, sepjala ruce a spustila (...) (SYN v5) 'The class teacher looked like a nun, clasped her hands and started (...)'

Nemám ani tušení, jaká **nemocenská**/nemocenské/N by mne čekala v případě onemocnění. (SYN v5) 'I have no idea what kind of sickness benefit awaits me in case of illness.'

Tab. 4. Examples of units with incorrect lemma and tag

4 SOLUTIONS

We propose three solutions for improving nominalized adjectives tagging: remove the inaccuracies from the morphological dictionary; add the obtained data described below to the Multiword Expressions Lexical Database (LEMUR) (see below); and apply our findings for disambiguation.

4.1 Removing the inaccuracies from the morphological dictionary

We believe that refinement of the data in the morphological dictionary [5] used for the ČNK corpora will lead to a more precise automatic morphological analysis. Below are proposals for adding analysed data to a morphological dictionary or for clarifying the interpretation of existing data.

We believe that refinement of the data in the morphological dictionary [5] used for the ČNK corpora will lead to a more precise automatic morphological analysis. Below are proposals for adding analysed data to a morphological dictionary or for clarifying the interpretation of existing data.

1) Only nouns, POS=N

The analysis showed that five units are nouns, even though they are listed in the morphological dictionary as both noun and adjective: *bytná* 'landlady', *bytný* 'landlord', *číhaná* 'lurking', *přísedící* 'associate', *šipkovaná* 'treasure hunt'.

2) Only adjectives, POS=A

The analysis showed that 35 units are adjectives (Appendix 1), even though they are listed in the morphological dictionary as both, noun and adjective.

We propose that units that represent school grades, *výborná* 'excellent', *výtečná* 'very good', *chvalitebná* 'good', *dobrá* 'satisfactory', *dostatečná* 'poor', *nedostatečná* 'failure' should be considered as adjectives. We think that from the contexts one can see the ellipsis of a noun. Within this semantic group, tagging will be unified and improved. We are aware of the problematic nature of this proposal. Ultimately, however, a pragmatic view of improved automatic tagging prevailed along with the most consistent tagging. By removing six units from a group of nouns, the consistency of the tagging within one semantic group will be preserved. The automatic part of speech tagging will be greatly improved, because it will not have to deal with the disambiguation, which is quite complicated especially in the case of frequent expressions as *výborná* 'excelent' and *dobrá* 'good'.

3) Nouns and adjectives POS=N, POS=A

The analysis showed that 50 units currently have only one interpretation (POS=N or POS=A). However, they can be both adjective and/or a noun (Appendix 2). In addition to the POS=N interpretation, it is also necessary to add a grammatical gender. [8]

4.2 Adding data to the Multiword Expressions Lexical Database

The analysis showed just how diverse the group of nominalized adjectives are. Although we tried to find different ways of characterization that could be generalized, it turned out to be almost impossible. The analysis confirmed that nominalized adjectives occur predominantly as one part of speech in certain contexts. Whatever this seems to be trivial, knowing the relevant collocations can greatly improve automatic morphological tagging.

The Multiword Expressions Lexical Database, LEMUR, ([6], [7]) was created by the Institute of Theoretical and Computational Linguistics, Charles University and the Institute of the Czech National Corpus FF UK, and is used in the disambiguation of corpora of the Czech National Corpus. Larger the database is, better result in tagging can be reached.

We will demonstrate our approach on lemmas, which can be both a noun or an adjective and belong to the semantic group of agentive nouns.

1) Units preceded by lemmas *pan* 'mister' and *paní* 'missis' and followed by a proper noun

- lemma *pan* 'mister'

hostinský, kantýnský, lázeňský, nadřízený, obžalovaný, odsouzený, podřízený, představený, vrátný

innkeeper, canteenman, spamaster, superior, defendant, convicted, subordinate, superior, porter

- lemma *paní* 'missis'

hostinská, kantýnská, lázeňská, nadřízená, obžalovaná, odsouzená, podřízená, představená, vrátná, zubatá

innkeeper, canteenlady, spamaster, superior, defendant, convicted, subordinate, superior, porter, Death

- lemmata *pan* i *paní*

The lemma *pan* or *paní* can help the gender disambiguation of units listed below because they have very often homonymous form.

cestující, domácí, dozorčí, duchovní, pokladní, produkční, provozní, radní, recepční, rozhodčí, spolubydlící, třídní, účetní, vedoucí, vrchní, výčepní

passenger, landlord, landlady, supervisor, clergyman, cashier, production manager, operating, councilor, receptionist, referee, roommate, class teacher, accountant, leader, waiter, bartender, barmaid

2) Collocations

The collocation overview does not aim to list all collocations, but to list those that can help with automatic tagging.

We are aware of the fact that some of the collocations below, e.g. *rozhodčí smlouva;* 'arbitration agreement' may also occur in the opposite part of speech classification than we have stated: *Rada města schválila* ***rozhodčí*** *smlouvu*. (SYN v5) 'The City Council approved the arbitration agreement.' vs. *Fotbalový* ***rozhodčí*** *smlouvu nepodepsal.* (our own example) 'The football referee did not sign the contract.'. However, we believe that similar contexts are really rare as we did not find examples in the corpus.

We believe that by applying relatively frequent phrases and collocations, the results of the automatic tagging will improve rather than deteriorate. Specifically, the collocation *rozhodčí smlouvu* 'arbitration agreement / referee contract' occurs in the SYN v5 corpus 108 times, all occurrences are erroneously marked as ***rozhodčí*** / *rozhodčí* / N; 'referee' ***smlouvu*** / smlouva / N; 'agreement'. In this case, there would be a 100% improvement. If we take into account the lemmas *rozhodčí* and *smlouva*, it is found in SYN corpus v5 614 times, of which the *rozhodčí* is only 34 times correctly tagged as an adjective. Even in this case, there would be a significant improvement in labelling.

The specific improvement of the tagging will differ from the unit to unit. Complete list of collocations states Žižková [8].

Unit	**Noun**	**Adjective**
dozorčí 'supervisor / supervisory'	**operační** dozorčí 'operational supervisor'	dozorčí **rada**, dozorčí **orgán**, dozorčí **služba**, dozorčí **komise**, dozorčí **útvar**, dozorčí **důstojník**, dozorčí **úřad; '**supervisory board, supervisory body, supervisory service, supervisory commission, supervisory unit, supervisory officer, supervisory authority'
lázeňský 'spa'		**švihák** lázeňský, lázeňský **dům** 'spa dude, spa house'
předsedající 'chairman / presiding'	předsedající **schůze**, předsedající **zasedání**; 'chairman of the meeting, chairman of the session'	předsedající **země**, předsedající **soudce**, předsedající **stát** 'presiding country, presiding judge, presiding state'
Představená 'Lady Superior / presented'	**matka** představená, představená **kláštera**, představená **řádu**; 'Mother Superior, Superior of the Monastery, Superior of the Order'	
radní 'councilor / town hall'	radní **kraje**, radní **města** 'district councilor, city councilor'	radní **věž** 'town hall tower'

recepční; 'receptionist / reception'	recepční **hotelu**, recepční **kempu**, recepční **autokempu**, recepční **penzionu**; 'hotel receptionist, camp receptionist, campsite receptionist, guesthouse receptionist'	recepční **služba**, recepční **pult**, recepční **estetika** 'reception service, reception desk, reception aesthetics'
rozhodčí 'referee / arbitration'	**hlavní** rozhodčí, **pomezní** rozhodčí, **čárový** rozhodčí 'chief referee, sideline referee, line referee'	rozhodčí **soud**, rozhodčí **nález**, rozhodčí **senát**, rozhodčí **výbor**, rozhodčí **tribunál**, rozhodčí **orgán**, rozhodčí **panel**, rozhodčí **řád**, rozhodčí **sbor**, rozhodčí **spis**, rozhodčí **výrok**, rozhodčí **institut**, rozhodčí **proces**, rozhodčí **soudce** 'arbitration tribunal, arbitration report, arbitration senat, arbitration committee, arbitration tribunal, arbitration body, arbitration board, arbitration code, arbitration board, arbitration records, arbitration statement, arbitration institutes, arbitration process, arbitration judge'

Tab. 5. Examples of collocations

4.3 Taking into account when disambiguating

In order to neutralise gender differences, masculine in plural is more frequent in the semantic group of agentive nouns (*dozorčí* 'supervisor', *přísedící* 'associate', *lázeňský* 'spamaster' etc.) and designation of persons having a certain quality (*dospělý* 'adult', *trpící* 'suffering' etc.). Only the unit *pokojská* 'maid' is more frequent in plural in feminine. We propose to take into account the neutralization of gender differences in disambiguation of units listed in Appendix 3.

5 CONCLUSION

The results of this investigation show that there is a way how to improve the nominalized adjectives tagging.

Thanks to the selected CQL queries and subsequent manual searches, we compiled a list of 319 terms that we considered to be a possible nominalized adjective.

The detailed analysis of nominalized adjectives showed that the part of speech is not always tagged properly. There are two reasons for the erroneous tagging: inaccuracies in the morphological dictionary used in the ČNK corpora, and the disambiguation errors. So three solutions for improving nominalized adjectives tagging were proposed.

The first proposal involves removing inaccuracies from the morphological dictionary. We proposed a change of interpretation to noun for 5 units to POS=N, then for 37 units change to adjective, POS=A, for 51 units we recommended a change to noun, POS=N, and adjective, POS=A, interpretation. The second proposal foresees the extension of the LEMUR database to the proposed collocations collected for 147 units. Thirdly, findings on disambiguation were formulated for 89 units.

The analysis shows how diverse and hard it is to tag properly a group of expressions that are subject to the part of speech transition. Nevertheless, we believe that the proposed solutions will at least partially improve the automatic part of speech tagging.

References

[1] Dokulil, M. et al. (1986). Mluvnice češtiny 1. Fonetika. Fonologie. Morfonologie a morfemika. Tvoření slov. Praha, Academia.

[2] Štícha, F. (2013). Akademická gramatika spisovné češtiny. Praha, Academia.

[3] Křen, M. et al. (2017). Korpus SYN, version 5 as of 24 April 2017. Praha, Ústav Českého národního korpusu FF UK. Accessible at `http://www.korpus.cz`.

[4] Šimandl, J. (ed.). (2018). Slovník afixů užívaných v češtině [online]. Praha, Karolinum [cit. 2018-08-24]. Accessible at `https://www.slovnikafixu.cz`.

[5] Hajič, J., and Hlaváčová, J. (2013). MorfFlex Praha: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

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

[7] Hnátková, M., T. Jelínek, M. Kopřivová, V. Petkevič, A. Rosen, H. Skoumalová, and Vondřička, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. Korpus – gramatika – axiologie, Univerzita Hradec Králové a ÚJČ AV ČR, 17, pages 3–22.

[8] Žižková, H. (2019). Slovnědruhové přechody jako problém automatické morfologické analýzy. Disertační práce. FF MU.

[9] Komárek, M. (1986). Mluvnice češtiny 2. Tvarosloví. Praha, Academia.

[10] Karlík, P., Nekula, M., and Rusínová, Z. (1995). Příruční mluvnice češtiny. Praha, Lidové noviny.

[11] Štícha, F. (2018). Velká akademická gramatika spisovné češtiny. Praha, Academia.

Appendix 1

barská, basiliánský, černá, divoká, dobrá, dostatečná, hokejbalová, chvalitebná, instrinsický, jatečné, kopulové, lutrový, maltézský, novellovský, oscilátorové, panamerická, panský, pětimiliardová, pětimiliardový, podrostové, poloninský, safesová, safesový, samodruhá, skopová, stehová, tajná, tvůrčí, umpirová, umpirový, verbovní, výborná, výtečná, vyvolený, zákolanská

‘bar, basilian, black, wild, good, sufficient, hockeyball, very good, instrinsic, slaughter, dome, low-wines, maltese, novel, oscillator, panamerican, manor, five

billion, undergrowth, polonin, safe, pregnant, mutton, stitched, secret, creative, umpire, recruiter, excellent, exquisite, chosen, from Zákolany'

Appendix 2
biovepřové, dančí, demonstrující, dojíždějící, dospívající, dostřelná, handicapovaný, hendikepovaný, košíková, kupující, místní, mrtvý, nakupující, nastávající, obviněný, oddávající, podezřelý, pohřešovaný, pokojská, pokojský, popravčí, postižený poškozený, prodávající, protestující, prvotrestaný, přednášející, předsedající, přespolní, přihlížející, příchozí, režný, sázející, sloužící, soutěžící, stávkující, startující, studující, tonoucí, trpící, trvalá, účinkující, umírající, volající, vystupující, vyšetřující, vyučující, zavražděný, zraněný, zúčastněný

'biopork, (of) fallow deer, demonstrating, commuter, teenage, firing, handicapped, basketball, buyer, local, dead, shopper, wife-to-be, accused, wedding registrar, suspect, missing, maid, chambermaid, executioner, handicapped, injured, seller, protesting, first punished, lecturer, presiding, cross-country, onlooker, incoming, rye, betting, serving, contestant, striking, starting, studying, drowning, suffering, permanent, acting, dying, calling, performer, investigating, teacher, murdered, injured, involved'

Appendix 3
cestující, demonstrující, dojíždějící, domácí, dospělá, dospívající, dozorčí, duchovní, handicapovaná, hendikepovaná, hostinská, kantýnská, kolemjdoucí, kupující, lázeňská, místní, mrtvá, nadřízená, nakupující, nastávající, obviněná, obžalovaná, oddávající, odsouzená, okolojedoucí, pocestná, poddaná, podezřelá, podřízená, pohřešovaná, pokladní, postižená, postupující, poškozená, pracující, prodávající, produkční, protestující, protijdoucí, provozní, přednášející, předsedající, představená, přespolní, příbuzná, přihlížející, příchozí, přísedící, radní, recepční, rozhodčí, sázející, sloužící, služebná, soutěžící, spolubydlící, spolucestující, stávkující, strážná, studující, tonoucí, trpící, třídní, účetní, účinkující, umírající, vedoucí, věřící, volající, vrátná, vrchní, výčepní, vystupující, vyšetřující, vyučující, zavražděná, zraněná, zúčastněná

'passenger, demonstrating, commuter, landlord, landlady, adult, teen, supervisor, clergyman, handicapped, handicapped, innkeeper, canteenlady, passerby, buyer, spamaster, local, dead, superior, shopper, wife-to-be, accused, indicted, wedding registrar, convicted, bystanders, wayfarer, subject, suspect, subordinate, missing, cashier, disabled, advancing, injured, working, seller, production manager, protesting, oncoming, operating, lecturer, chairman, superior, non-resident, relative, onlooker, incoming, associate, councilor, receptionist, referee, betting, serving, maid, contestant, roommate, fellow-traveller, striking, guard, studying, drowning, suffering, class teacher, accountant, performer, dying, leader, believer, calling, porter, waiter, bartender, performer, investigating, teacher, murdered, injured, involved'