NovaMorf: konec dlouhého období konvergencí a divergencí ve zpracování české morfologie

Klára Osolsobě[1] and Jaroslava Hlaváčová[2] *

[1] Ústav českého jazyka, FF MU, Brno
[2] Ústav formální a aplikované lingvistiky, MFF UK, Praha

(…… and 'tis not hard, I think,
For men so old as we to keep the peace.
Romeo and Juliet, I, 2)

1 Úvod

V tomto textu chceme nastínit shody a rozdíly dvou tagsetů užívaných k automatické morfologické analýze češtiny. Ukážeme, nakolik původně nezáměrná, ale časem udržovaná dvojkolenost tzv. pražského a tzv. brněnského systému může býti v dohledné době překonána díky projektu NovaMorf. Budeme se zabývat vztahy mezi značkováním morfologických kategorií a hodnot v návrhu NovaMorf v porovnání s oběma staršími systémy. Při posuzování brněnského systému vycházíme z článku [11]. Poznatky o pražském systému zakládáme na popisu pražského pozičního tagsetu (viz `//ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html`) a na monografii Jana Hajiče [2]. Naším cílem bude ukázat, kterak zkušenosti s užíváním obou systémů vyústily ve snahu inspirovat se pozitivy a vyhnout se neúspěšným řešením na obou stranách (srv. [15] . Ke vzájemné konverzi značek dosavadního pražského systému na brněnský viz [18].

2 Tagset návrhu NovaMorf

Návrh NovaMorf pracuje s těmito kategoriemi:

1. Slovní druh – POS
2. Poddruh – SUB
3. Deixe – DEI
4. Vid – ASP
5. Zkratka – ABR

* Text byl vytvořen za podpory těchto projektů: LM2015044 Český národní korpus, LM2010013 LINDAT-Clarin a MUNI/A/1061/2018 Čeština v jednotě synchronie a diachronie - 2019.

6. Rod – GEN
7. Číslo – NUM
8. Pád – CAS
9. Osoba – PER
10. Stupeň – DEG
11. Negace – NEG
12. Slovesný tvar – VRB
13. Jmenný tvar (přídavných jmen) – NOM
14. Typ agregátu – AGR
15. Globální mutace – GMU (slouží k zachycení variantnosti ve všech tvarech paradigmatu, tj. ve všech tvarech spadajících pod variantní lemma v konceptu vícenásobného lemmatu)
16. Flektivní mutace – FMU (slouží k zaznamenání varianty, která má stejné lemma a tag, FMU mají tedy pouze ohebné slovní druhy, neohebné mohou mít jenom GMU, výjimkou jsou adverbia, neboť u nich se koncept FMU využívá u variantnosti ve stupňování).

V následujícím textu nejprve stručně naznačíme (odd. 3), kterak se v návrhu NovaMorf zachází s variantností morfologického systému češtiny, která není ani v brněnském, ani v pražském systému dostatečně zohledněna. Poté se budeme v sekci 4 podrobně věnovat kategoriím návrhu NovaMorf pod zorným úhlem inspirací brněnským systémem.

3 Klasifikace návrhu NovaMorf bez zřetelné opory v odpovídajících značkách v brněnském a pražském systému

Ani v brněnském, ani v pražském systému není žádná explicitní opora pro tagování vlastností, které má zohledňovat značkování GMU a FMU (viz podrobně [5], nejnověji [6]). Navržený systém mutací odpovídá rozsáhlému množství variant, a to jak ortografických, tak hláskoslovných, morfologických a v neposlední řadě i stylových. Cílem je odstranit případy, kdy více různých tvarů dosud charakterizuje stejná kombinace lemmatu a značky (požadavek jednoznačného popisu), a co nejúplněji popsat varianty stejného typu stejně (požadavek konzistentnosti popisu). Cílem není hodnotící klasifikace. Údaj o tom, jak se ta která varianta má k dosavadní kodifikaci či k interpretacím variet národního jazyka, nemá být vložen do slovníku, protože se netýká interpretace na rovině morfologické analýzy, ale interpretace na rovině jiné (nemusí přitom jít jen o rovinu jazykové kultury)[3]. Navržená klasifikace GMU a FMU neodpovídá tudíž hodnocení stylis-

[3] Užívá-li se substandardních tvarů v beletristické části obecného korpusu řady SYN, jde o jiný stylový příznak a jiný vztah ke kodifikaci, než když jsou tytéž varianty užity v lokální publicistice, v interview, soukromé korespondenci, neřkuli jako příklady v odborném textu (mnoho okrajových jazykových variant doložených v některém z korpusů řady SYN pochází z lingvistických textů, jde tedy o metatextová užití). Domníváme se, že pro výzkum zaměřený na uvedené jevy v jazyce lze kombinovat

tických/stylových variant ve stávajících systémech (v pražském systému 15. pozice, v brněnském atribut w), přestože se tuto klasifikaci snaží nahradit. Subklasifikace pomocí atributu w (stylistický příznak) prošla až do nového revidovaného tagsetu ([11]) beze změn, přestože omezenost takové klasifikace je všeobecně známa (viz [13], poznámka o převzetí hodnot atributu w ze SSJČ). V návrhu na jednotnou klasifikaci variant byly ovšem zohledněny zkušenosti založené na užití značky var, kterou disponuje derivační analyzátor Derivancze (více [17]).

4 Jednotlivé kategorie návrhu NovaMorf se stálým zřetelem k inspiracím brněnským systémem

4.1 Slovní druh (POS)

Kategorie slovního druhu v brněnském atributivním tagsetu k, v pražském pozičním první pozice, v podstatě kopíruje tradiční pojetí deseti slovních druhů. Tato kategorie existuje v obou tagsetech a její hodnoty si rámcově odpovídají. Návrh NovaMorf přidává tři nové slovní druhy: cizí slovo (F), afixový segment (S) a agregát (G). Návrh navazuje na dlouhodobě deklarovanou potřebu ošetřit značkování cizojazyčných úseků textu (cizí slovo), značkování samostatných grafických jednotek (afixový segment), které tvoří slovnědruhovou jednotku (výrazy oddělené pomlčkou, které tvoří samostatné tokeny, např. *česko - německý, tří - až čtyřpokojový, Ho - Či - Min*) a značkování grafických jednotek, které reprezentují několik slovních tvarů (agregát jako např. *přišels → přišel jsi, očs → o co jsi*). Pro nově navržené slovní druhy existuje/existovala v brněnském tagsetu klasifikace, na kterou bylo možné navázat. Krátké cizojazyčné úseky textů v souvislém českém textu byly v Korpusu soukromé korespondence ([4]) značkovány samostatnými tagy ([13]) doplněnými do brněnského tagsetu upraveného pro potřeby tagování a následné manuální desambiguace (srv. [8], [7]). Idea zvláštní značky pro cizí slova rovněž figuruje v návrhu Jaroslavy Hlaváčové ([5]). Podobně v KSK existoval zvláštní atribut "z" s hodnotou S pro tvary s volným morfem *–s* za auxiliár *být* (např. *přišels, kams, žes, cos*, … viz [13]).

V nejnovější verzi pražského morfologického slovníku MorfFlex [1] se všechny tyto tři nové slovní druhy již implementují, i když ne vždy přesně v této podobě.

4.2 Dvě subklasifikační kategorie SUB a DEI

Kategorie SUB má v návrhu NovaMorf nahradit heterogenní subklasifikaci druhé pozice pražského systému. V brněnském tagsetu je subkategorizace řešena pomocí čtyř různých atributů (obecné subklasifikace atributem z a subklasifikačních typů pomocí atributů x, y, t). Kromě toho jsou ještě v návrhu [11] obsaženy atributy k subklasifikaci frekvenčních charakteristik a stylových charakteristik.

strukturní značkování (typ textu) a navržené značkování variant pomocí mutací. Jakákoliv další interpretace užité varianty je možná až na základě posouzení širšího kontextu (typu textu), a tudíž dalece přesahuje rámec automatické morfologické analýzy.

Atributy x a y se mohou kombinovat v jedné značce u zájmen, číslovek a zájmenných adverbií. Tento rys brněnského tagsetu inspiroval v návrhu NovaMorf vytvoření dvou subklasifikačních kategorií SUB a DEI.

U zájmen, zájmenných číslovek[4] a zájmenných příslovcí[5] se v návrhu NovaMorf rozlišují dva poddruhy. Dělení zájmen, zájmenných číslovek (*kolik, tolik, několik*) a zájmenných příslovcí v návrhu NovaMorf bere v úvahu dvojí podstatu užívaných hodnot. Např. zájmeno *něčí* je zároveň přivlastňovací i neurčité a vyjadřování obou významů v rámci jediného poddruhu by znepřehlednilo klasifikaci. Z tohoto důvodu byla v rámci NovaMorf vytvořena kategorie nazvaná deixe. Inspirací byl brněnský systém, v němž dosud existuje dvojí (dva atributy x, y ve značce) subklasifikace zájmen a dokonce trojatributové značkování zájmenných adverbií (specifikace adverbií atributy x, z, t). Hodnoty obou kategorií, SUB i DEI, se samozřejmě mohou kombinovat (proto byly zavedeny), ovšem ne zcela libovolně. V brněnském systému se navíc oproti návrhu NovaMorf realizuje tagování sémantických tříd adverbií pod atributem t, který nabývá hodnot jako stavové (S), modální (D), času (T), přípustky (A), příčiny (C), místa (L), způsobu (M), míry (Q). Hodnoty se netýkají jen zájmenných adverbií, ale jsou uváděny u všech adverbií. V tomto bodě je brněnský tagset bohatší. Do návrhu NovaMorf klasifikace brněnského systému přijata nebyla, a to s ohledem na fakt, že rozlišení tříd adverbií se týká spíše sémantické než morfologické klasifikace, a není ani dosti dobře formalizovatelné, více viz [9].

4.3 Značkování kategorií relevantních pro slovesa (POS=V/k5) v návrhu NovaMorf v porovnání s brněnským systémem

Porovnání pražského a brněnského tagsetu ve vztahu ke slovesnému tvarosloví zahrnuje studie Kláry Osolsobě [14]. Zde se ukazuje, jakým způsobem těží z výhod obou tagsetů návrh tagsetu pro značkování slovenštiny vytvořený v rámci projektu Slovenského národného korpusu. Pro slovesa je v návrhu NovaMorf relevantní kategorie poddruh (SUB), vid (ASP), osoba (PER) a číslo (NUM), slovesný tvar (VRB) a pro krátké tvary n/t-ových příčestí[6] také kategorie jmenný tvar (NOM)[7]. Poddruh sloves má v návrhu NovaMorf dvě hodnoty. SUB=b pro

[4] Slovo *kolik* můžeme z dobrých důvodů pokládat za základní číslovku i tázací zájmeno a slovo *kolikátý* za řadovou číslovku a tázací zájmeno. Podobně slovo *tolik* můžeme z dobrých důvodů pokládat za základní číslovku i za ukazovací zájmeno a slovo *tolikátý* za řadovou číslovku a ukazovací zájmeno. Zájmennými číslovkami se vyjadřuje vztah k množství. Plní stejné funkce (i syntaktické) jako zájmena.

[5] Zájmenná adverbia mají zájmenné kořeny a zástupné funkce podobné zájmenným.

[6] Ta jsou v návrhu NovaMorf značkována jako POS=A (adjektiva). Také v brněnském systému je snaha sjednotit značkování těchto tvarů s jim odpovídajícími dlouhými tvary adjektiv (viz [11]). Interpretace krátkých tvarů (pasivních příčestí) byla totiž v brněnském slovníku masivně přegenerována, a to tak, že krátké tvary měly jak interpretaci k5, tak k2 (viz [10]).

[7] Kategorie NOM ve vztahu ke slovesům je zavedena jako prostředek pro řešení společné lemmatizace krátkých (jmenných) a dlouhých (složených) tvarů adjektiv a tvarů slovesných příčestí trpných v krátké i dlouhé podobě pod několikanásobné

pomocná slovesa (*být*[8], *bývat*) a SUB=0 pro všechna ostatní. Hodnoty kategorií vid, osoba i číslo a slovesný tvar mají své protějšky v brněnském systému. Brněnské pojetí slovesných tvarových subsystémů (atribut slovesný tvar/mód) inspirovalo silně pojetí popisu systému slovesných forem v návrhu NovaMorf.

4.4 Značkování kategorie stupeň v návrhu NovaMorf v porovnání s brněnským systémem

Všechna adjektiva i adverbia v pozitivu bez ohledu na sémantické rysy, které bývají uváděny jako rozhodující pro pravidelné (víceméně paradigmatické) odvozování tvarů komparativu sufixy *-í/-ší/-[eě]jší* nebo slovnědruhovými charakteristikami *-e/-ě/-eji/-ěji* a superlativu prefixem *nej-*, mají mít podle návrhu NovaMorf hodnotu DEG=1. Pravidelně i nepravidelně odvozené tvary budou mít hodnotu komparativu, resp. superlativu DEG=[23], k nim budou patřit i pravidelně tvořené tvary typu *sebe*+komparativ, pro něž NovaMorf zavádí hodnotu DEG=s. Jedná se o změnu dosavadní praxe pražského systému, která považovala některé typy adjektiv a adverbií za explicitně nestupňovatelné (tag=Db.*). Navrhovaná změna zcela odpovídá praxi brněnského systému.

Návrh NovaMorf počítá rovněž s některými změnami v lemmatizaci. U adjektiv a adverbií druhého a třetího stupně od supletivních kmenů a u adjektiv a adverbií bez jednoznačného vztahu ke tvaru pozitivu navrhuje lemmatizaci tvarem komparativu. Tato změna odlišuje NovaMorf od pražského a brněnského systému, neboť v obou byl komparativ a superlativ důsledně (v některých případech násilně a chybně[9]) lemmatizován základním tvarem pozitivu.

4.5 Značkování kategorie Zkratka (ABR) v návrhu NovaMorf v porovnání s brněnským systémem

Tato kategorie je v návrhu NovaMorf relevantní pro všechny slovní druhy, má pouze jednu hodnotu, a to ABR=+, kteroužto hodnotu dostávají zkratky, ostatní slovní tvary nemají tuto hodnotu definovanou, mají tedy ABR=-. Zkratka může být libovolný slovní druh. V brněnském systému zkratky byly dříve značkovány na rovině slovního druhu jako kA, v návrhu [11] je uvedena hodnota A (zkratka) u obecného subklasifikačního atributu z.

---

lemma (například *schopný, schopen*, *hrdý, hrd*, *ukrytý, ukryt*, *pokáraný, pokárán*). Krátké (jmenné tvary) budou mít NOM=J, dlouhé tvary NOM=0. Tato kategorie je relevantní pro adjektiva (zejména adjektiva tvořená ze sloves), zájmena (*sám, samý*), číslovky (tvary jmenné a složené u číslovek typu *devatero/devaterý*). Tato kategorie nemá sice obdobu v brněnském systému, přesto je její ideové východisko kompatibilní s brněnským systémem.

[8] Včetně kondicionálových tvarů *bych, bys, by, bychom*, ...

[9] Srov. lemmatizaci adverbia *dřív/dříve* adverbiem *brzy*.

5 Závěr

V textu jsme stručně představili návrh nového tagsetu projektu NovaMorf se stálým zřetelem ke genezi změn, kterými nově navržený tagset reaguje na zkušenosti se značkováním korpusů současné češtiny. Ty byly získány už více než čtvrt století trvajícím užíváním nástrojů automatické morfologické analýzy opřené o pražský tagset ([2,3]) a brněnský tagset ([16,12]). Představený návrh vychází především z analýzy výhod a nevýhod obou systémů a staví na výsledcích disertační práce Jaroslavy Hlaváčové ([5]). Podrobnější popis celého návrhu je připraven k publikaci (Hlaváčová, Křivan, Osolsobě, Petkevič, Svášek, Šimandl, zatím bez názvu, předpokládané vydání 2019).

Domníváme se, že brněnský systém, u jehož zrodu od počátku stál jubilant, přispěl k návrhu NovaMorf v řadě bodů. Na tomto místě a na závěr tohoto textu bychom rády poděkovaly panu profesorovi Karlu Palovi za mnoho podnětů a za stálou pozornost, kterou automatické morfologické analýze češtiny věnoval, a předkládáme jeho kritickému duchu výsledky snah, které mu snad nebudou proti mysli.

References

1. J. Hajič and J. Hlaváčová. MorfFlex CZ, 2013. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
2. J. Hajič. *Unification Morphology Grammar*. PhD thesis, Charles University, 1994.
3. J. Hajič. *Disambiguation of Rich Inflection - Computational Morphology of Czech*. Charles University, 2004.
4. Z. Hladká and kol. *Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS*. Brno: Masarykova univerzita, 2005.
5. J. Hlaváčová. *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. PhD thesis, FF UK, 2009.
6. J. Hlaváčová. Golden rule of morphology and variants of wordforms. *Jazykovedný časopis / Journal of Linguistics*, 68(2):136–144, 2017.
7. D. Hlaváčková and K. Osolsobě. *Čeština v mluveném korpusu*, chapter Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. Nakladatelství Lidové noviny/ Ústav Českého národního korpusu, Praha, 1 edition, 2008.
8. D. Hlaváčková and Sedláček R. Morfologické značkování korpusu soukromé korespondence. In *Varia XIV*, pages 371–379. Bratislava: Slovenská jazykovedná spoločnosť pri SAV, 2006.
9. B. Hvězdová. Tvoření adverbií paradigmaticky odvozených od adjektv na materiálu Čnk. Master's thesis, FF MU : Brno, 1999.
10. V. Hájková. Analýza jmenných tvarů adjektiv a pasivních příčestí ve slovníku morfologického analyzátoru ajka. Bakalářská práce, Masaryk University, Brno, 2014.
11. M. Jakubíček, V. Kovář, and P. Šmerk. Czech morphological tagset revisited. In *The 5th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2011, Karlova Studanka, Czech Republic, December 2-4, 2011.*, pages 29–43, 2011.

12. K. Osolsobě. *Algoritmický popis české formální morfologie a strojový slovník češtiny.* PhD thesis, Brno: Masaryk University, 1996.
13. K. Osolsobě. Korpus soukromé korespondence (ksk) z hlediska morfologického značkování. *Linguistica Brunensia*, A 54(1):187–201, 2006.
14. K. Osolsobě. Popis gramatických významů (hodnot) jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů. *Linguistica Brunensia*, A 55(1):201–218, 2007.
15. K. Osolsobě, J. Hlaváčová, V. Petkevič, M. Svášek, and J. Šimandl. Nová automatická morfologická analýza češtiny. *Naše řeč*, 100(4/2017):225–234, 2017.
16. K. Osolsobě and K. Pala. Czech stem dictionary. czech stem dictionary. In *Sborník prací filozofické fakulty brněnské univerzity*, pages 51–60. Masarykova univerzita, Brno, 1993.
17. K. Pala and P. Šmerk. Derivancze—derivational analyzer of czech. In *International conference on text, speech, and dialogue*, pages 515–523. Springer, 2015.
18. P. Pořízka and M. Schäfer. Morphcon–a software for conversion of czech morphological tagsets. *NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 292–301, 2009.