# zmraženo

## Vztah NovaMorf a dosavadního brněnského a pražského systému

V tomto textu nastíníme vztahy mezi značkováním morfologických kategorií a hodnot v návrhu NovaMorf a 1) v upravené verzi brněnského systému (Jakubíček, Kovář, Šmerk, 2011) a 2) v pražském tagsetu (https://wiki.korpus.cz/doku.php/seznamy:tagy). Pokusíme se formou komentovaných tabulkových přehledů porovnat vztah mezi kategoriemi i hodnotami. Vztah návrhu NovaMorf k současné podobě značkování pozičním tagsetem užívaným na pražských pracovištích (ÚJAL MFF UK, ÚČNK FF UK, ÚTKL FF UK) viz též Osolsobě a kol. 2017.
V textu nebudeme podrobně probírat vše, co se týká otázky změn v tokenizaci a lemmatizaci navrhované v NovaMorf. Změny v lemmatizaci, jichž se zde okrajově dotkneme, se týkají zavedení tzv. vícenásobného lemmatu, lemmatizace tvarů derivovaných negujícím prefixem *ne-*, lemmatizace tvarů komparativu/superlativu od supletivních základů, lemmatizace tvarů komparativu/superlativu s nejasným vztahem ke tvaru pozitivu a lemmatizace tvarů komparativu/superlativu, jejichž formální pozitiv není adjektivum/adverbium, ale číslovka. Více o všech těchto změnách viz příslušné kapitoly.
V tomto textu pomineme veškeré technické problémy, což neznamená, že bychom si jich nebyli vědomi.[1] K vzájemné konverzi značek dosavadního pražského systému na brněnský viz Pořízka, Schäfer, 2009.

### Slovní druh (POS/k/1. pozice)

Tato kategorie existuje ve všech třech tagsetech a její hodnoty si rámcově odpovídají. Kromě deseti tradičních slovních druhů (substantiva: N/k1/N, adjektiva: A/k2/A, zájmena: P/k3/P, číslovky: C/k4/C, slovesa: V/k5/V, příslovce: D/k6/D, předložky: R/k7/R, spojky: J/k8/J, částice: T/k9/T a citoslovce: I/k0/I), u nichž existují dílčí neshody ve značkování na rovině slovníku,[2] a interpunkce značkované ve všech třech tagsetech jakožto slovní druh (Z/kI/Z), obsahuje návrh NovaMorf tři nové slovní druhy: cizí slovo, afixový segment a agregát [FSG]. Pro nově navržené slovní druhy existuje/existovala v brněnském tagsetu klasifikace, na kterou by bylo možné při snaze sjednotit datové zdroje ve smyslu návrhu NovaMorf navázat. Návrh NovaMorf přebírá z pražského systému označování slov, která nejsou v morfologickém slovníku POS=X.
V tabulce 1 naznačíme vzájemné korespondence jednotlivých tagsetů pokud jde o široce pojaté označení slovnědruhové platnosti analyzovaných slovních tvarů a v poznámkách vysvětlíme případy neshod.

### Tabulka 1

| | NovaMorf | Brno | Praha |
|---|---|---|---|

[1] Máme na mysli tu skutečnost, že nad brněnským i pražským slovníkem (daty popisujícími vlastnosti jazyka, které se odrážejí v lemmatizaci a tagování) pracuje v současnosti řada aplikací. Každý větší zásah do dat nutně spouští potřebu učinit odpovídající zásahy na mnoha dalších místech, víceméně všude, kde se s daty pracuje. Takové zásahy mohou vést k dočasnému znefunkčnění nástrojů na datech závislých, popřípadě mohou vést ke kolapsu těchto nástrojů. I v případě, že technická stránka celého problému bude natolik náročná, že nedojde ke shodě na sjednocení v popisu dat, měl by být popis co nejkompatibilnější. Také se domníváme, že už z hlediska širšího okruhu uživatelů nástrojů spojených s oběma systémy i systémem novým je žádoucí mít k dispozici informace o jejich vzájemných vztazích, společných rysech a rozdílech. Takové informace totiž do značné míry nejsou širší veřejnosti dostupné.
[2] K neshodám na rovině slovníku odkazujeme v jednotlivých kapitolách věnovaných slovním druhům.

| Slovní druh | **POS** | **k** | **1. pozice** |
|---|---|---|---|
| **substantivum** | **N** | k**1** | **N** |
| **adjektivum** | **A** | k**2** | **A** |
| **zájmeno** | **P** | k**3** | **P** |
| **číslovka**[3] | **C** | k**4**[4] | **C** |
| **sloveso** | **V**[5] | k**5** | **V** |
| **příslovce** | **D** | k**6** | **D** |
| **předložka** | **R** | k**7** | **R** |
| **spojky** | **J**[6] | k**8** | **J** |
| **částice** | **T** | k**9** | **T** |

[3] Mezi daty, která mají v návrhu NovaMorf POS=C a těmi, který brněnský systém značkuje jako k4 a pražský jako C, jsou dosti velké rozdíly (viz níže). Na tomto místě chceme poznamenat, že zachování slovního druhu číslovek, jehož delimitace je založena převážně na sémanticky motivovaných kritériích (morfologicky jde o slova s adjektivní, zájmennou i substantivní flexí, nebo o nesklonné výrazy), je dobré mít, přestože například v UD je značkování slov, která NovaMorf navrhuje klasifikovat jako číslovky, blízké brněnskému systému (viz níže a v kapitole věnované UD). Dobrým důvodem je například využití morfologického značkování nástrojem Sketch Engine v lexikografické praxi. Slovní profily číslovek mohou být při tvorbě slovníku většího rozsahu stejně důležitým objektem výzkumu jako slovní profily základových autosémantických slovních druhů (viz Benko, 2016 : 84).

[4] V brněnském systému mají tag=k4.* pouze číslovky určité základní (*jeden, dva, tři, ...* včetně těch se substantivní flexí – *sto, tisíc, milion, ...*), násobné (*první, druhý, třetí, ...*) a některé vztažené k celku (*dvoje, patery, čtverý, ...*). Dále mají značku k4.* slova (*málo, mnoho, pár, ...*) a zájmenné číslovky (*tolik, kolik, ...*), viz níže. Zásadně se brněnský systém liší ve značkování násobných číslovek. Slova derivovaná pravidelně od číslovek základních určitých postfixem *krát* jsou interpretována jako adverbia, mají značku k6.*. Podobně i další deriváty od základů číslovek určitých, tedy slova na .**ina, .**ice, .**násobný, .**násob, .**násobně*. Naopak některé deriváty adverbiální povahy *po.*é* (*potřetí, pošesté, ...*) číslovkovou interpretaci mají. Důvodem tohoto řešení je mimo jiné i to, že slova od číslovkových základů se chovají jako substantiva, adjektiva, popřípadě adverbia, takže pro aplikace zaměřené na syntaktickou analýzu, které pracují s výsledky automatického morfologického značkování (s tagy), není třeba přetěžovat množstvím pravidel zohledňujících sémantické kritérium vymezení slovního druhu číslovek.

[5] Jako **POS=V & SUB=b & VRB=K** jsou v NovaMorf značkovány tvary *bych, bys, by, ...*, viz níže. Návrh NovaMorf se tak liší od návrhu in Jakubíček a kol. 2011, podle nějž tyto tvary mají mít **k9zY** náhradou za dřívější **kY.*mC.***. Viz níže.

[6] Jako **POS=J & SUB=, & VRB=K** jsou v NovaMorf značkovány tvary *aby*, *kdyby*. Tvary *abych, abys, ...*, *kdybych, kdybys* jsou značkovány jako agregáty tvořené spojkou a prézentním tvarem slovesa *být*, viz níže. Návrh NovaMorf se tak liší od návrhu in Jakubíček a kol. 2011, podle nějž tyto tvary mají mít **k8zY** náhradou za dřívější **kY.*mC.***. Viz níže.

| | | | |
|---|---|---|---|
| **citoslovce** | **I** | k**0** | **I** |
| **interpunkce** | **Z** | k**I** | **Z** |
| **cizí slovo** | **F** | **---**[7] | **---** |
| **afixový segment** | **S** | **---**[8] | **---** |
| **agregát** | **G** | **---**[9] | **---** |
| **neznámé slovo** | **X** | **k?** | **X** |

**Poddruh (SUB a DEI/[zxyt]/2. pozice)**

Kategorie SUB je v návrhu NovaMorf relevantní pro všechny slovní druhy kromě předložek, citoslovcí, cizích slov a neznámých slov [RTIFX]. V případě nutnosti by nebyl problém hodnoty dodefinovat. 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 (Jakubíček a kol., 2011) obsaženy atributy k subklasifikaci frekvenčních charakteristik a stylových charakteristik, viz níže. Tyto čtyři subkategorizační atributy [zxyt] jsou relevantní pro všechny slovní druhy kromě předložek, částic a interjekcí **k[790]**. Atributy **x** a **y** se mohou kombinovat v jedné značce u zájmen, číslovek a zájmenných adverbií. Také subklasifikace zájmen, číslovek a adverbií na 2. pozici pražského systému popisuje poměrně značně heterogenní jevy. Z toho důvodu tabulkové přehledy pro tyto tři slovní druhy oddělíme zvlášť a porovnání značkování zájmen, číslovek a adverbií zahrneme níže do kapitoly věnované hodnotě kategorie SUB a druhé subklasifikační kategorie DEI v návrhu NovaMorf v porovnání s brněnským a pražským systémem značkování.

[7] Pro tuto kategorii sice neexistuje v brněnském systému adekvátní značkování. Přesto je v textu (Jakubíček a kol. 2011, s. 35) uvedena tabulka s převodem mezi značkami Google Universal Tagset a značkami brněnského systému, v níž značce X definované jako „other, **foreign words**, typos, abbreviations“ odpovídá značka k0, která je ovšem všude jinde v textu charakterizována jako značka slovního druhu **citoslovce/interjections**. V textu je značkám k9 a k0 věnován samostatný odstavec, v němž se hovoří o obtížné desambiguaci neohebných slovních druhů. Je ovšem třeba poznamenat, že v historii brněnského systému značka pro cizí slova v minulosti existovala. Šlo o značkování Korpusu soukromé korespondence (dále **ksk**), pro které byla vytvořena varianta morfologického analyzátoru *ajka* (Hlaváčková, Sedláček, 2006). V rámci ruční desambiguace založené na datech označkovaných touto variantou (viz Hladká a kol., 2005) byly zavedeny tagy pro cizí slova (více Osolsobě, 2006). Některým frekventovaněji užitým anglickým, francouzským, německým, slovenským, ruským aj. slovům v textech jsou přiřazeny následující značky: [tag="<anglicky>"], [tag="<nemecky>"], [tag="<francouzsky>"], [tag="<jiny_jazyk>"].

[8] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování.

[9] Značku pro nově zavedený slovní druh přiřazený ke slovnímu tvaru, který „vznikl z více slov (většinou různých slovních druhů) a určení jeho slovního druhu je problematické (příklady: *zač*, *oň*, *byls*)“ lze v brněnském tagsetu alespoň z části konstruovat. Tomuto slovnímu druhu odpovídá kombinace **k.**, kde za **“.”** doplníme hodnotu pro slovní druh slova s volným morfem *-s* (příklady: ***hlavous/k1*** *ani nepohnul,* ***zdrávas/k2*** *Maria,* ***kteréhos/k3*** *zabil,* ***druhýs/k4*** *nebyl,* ***šels/k5*** *tam,* ***tams/k6*** *nešel,* ***žes/k8*** *ho neviděl*) a značka bude obsahovat atribut subklasifikace **z=S**, viz níže.

V tabulce 2a uvedeme hodnoty kategorie SUB, x a kategorií značkovaných na 2. pozici pražského systému v závislosti na slovním druhu substantiv **N/k1/N** a adjektiv **A/k2/A**.

Tabulka 2a

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| Poddruh/ sublasifikace typu | **SUB** | **x** | **2. pozice** |
| **POS=N/k1** | **SUB=V**[10] | **---**[11] | **N** |
| **POS=N/k1** | **SUB=C**[12] | **---**[13] | **N** |
| **POS=N/k1** | **SUB=0**[14] | **---**[15] | **N** |
| **POS=N/k1** | **---**[16] | **xP**[17] | **N** |

[10] Jako POS=N a SUB=V budou značkována substantiva tvořená od sloves (s opěrným tvarem shodným s trpným příčestím) pravidelně a neomezeně produktivně sufixy ***-n-í/-t-í***.

[11] V brněnském tagsetu ve starší verzi a technicky patrně i nyní lze dogenerovat značkování substantiv slovesných na *ní/tí* tvořených pravidelně od kmene shodného s kmenem pro tvoření *n/t* příčestí. Původně označoval derivační historii slova v brněnském tagsetu atribut **r**, který nabýval hodnot **D**. Byl odstraněn a je dostupný ze samostatně budované databáze, v níž jsou uloženy informace o derivaci (viz více Jakubíček a kol., 2011, s. 34). V případě převodu by šlo pouze o úpravy slovníku (viz poznámky v kapitolách věnovaných substantivům a adjektivům). V derivačním analyzátoru *Derivancze* mají substantiva na *ní/tí* značku ***k1verb*** (viz více https://nlp.fi.muni.cz/projects/derivancze/index.cgi a také Pala, Šmerk, 2015).

[12] Jako POS=N & SUB=C budou značkovány výrazy substantivní povahy pravidelně odvozené od základů číslovek určitých a kompozita s prvním členem číslovkovým (*trojhvězda, pětiboj, šestihran, čtyřstěn, osmiválec, ...*). Konkrétně feminina na ***-ka/-ička/-ovka*** (například *jednička, čtyřka, čtverka, pětka, desítka, stovka, tisícovka*), maskulina na ***-ák*** (například *prvák, prvňák, druhák, třeťák, čtvrťák, čtvrták, páťák, pěťák, ...*), maskulina na ***-(n)ík*** (například *dvojník, troník, trojník*), neutra na ***-če*** (názvy mláďat z vícečetných vrhů/porodů, například *dvojče, trojče, vícerče, ...*) a konečně názvy jubileí pluralia tantum feminina na ***-iny*** (*padesátiny, šestnáctiny, ...*), která jsou ovšem homonymní s pl. tvary dílových číslovek, které budou podle návrhu NovaMorf značkovány jako POS=C & SUB=h & DEI=U a které bude třeba desambiguovat. (K jednotlivým lemmatům i k problémům desambiguace viz kapitola věnovaná substantivům.)

[13] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Domníváme se ovšem, že nejde o kategorii problematickou. V případě úprav slovníků bude nutné vycházet z toho, že substantivum s tímto poddruhem bude obsahovat číslovkový kořen. Těchto kořenů je omezený počet a možnosti automatické detekce jsou dosti dobré.

[14] SUB=0 budou mít substantiva, která nemají SUB=[VC].

[15] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Jde ovšem jen o doplňkovou množinu těch substantiv, která nemají SUB=[VC]. Převod by tudíž, pokud by se řešil komplexně, neměl činit potíže.

[16] Slovo ***půl*** (spolu se slovy *polovic, čtvrt*) je v návrhu NovaMorf značkováno jako číslovka dílová určitá (**POS=C & SUB=h & DEI=U**).

[17] Jde o značku slova *půl*. V korpusu czTenTen12 je lemma *půl* značkováno buď jako substantivum (**k1xPqP**), nebo jako adverbium (**k6eAd1qP**). Navíc ještě existuje lemma *půle* (substantivum s tvary podle vzoru *růže*).

| | | | |
|---|---|---|---|
| **POS=N/k1** | **---**[18] | **xF**[19] | **N** |
| **POS=A/k2** | **SUB=U**[20] | **---**[21] | **U** |
| **POS=A/k2** | **SUB=G**[22] | **---**[23] | **G** |
| **POS=A/k2** | **SUB=M**[24] | **---**[25] | **M** |

[18] Pro tuto kategorii neexistuje v návrhu NovaMorf adekvátní značkování. Domníváme se ovšem, že doplňování substantiv typu Novákovi(c), Hlaváčovi(c), Petkevičovi(c), Šimandlovi(c), Sváškovi(c) atd. by mohlo být spíše kontraproduktivní. V czTenTen12 není značka tag="k1.*xF.*" použita ani jednou a doklady mající ve značce **gR** (podle Jakubíček a kol., 2011 je **xF** "dědicem" **gR**) nesvědčí o tom, že by existovala nějaká rozumná desambiguace, která by odlišila posesivní adjektiva od substantivizovaných názvů skupin tvořících rodinu/příbuzné, viz níže.

[19] Značka je dědictvím dřívější značky kategorie rodu (**gR**) pro substantivizovaná posesiva typu *Novákovi*. Substantivizace adjektiv je širší problém, týká se řady produktivních typů adjektiv. V kapitole věnované adjektivům se popisují konkrétní návrhy řešení projektu NovaMorf. Návrh NovaMorf souzní s tvrzením: „Je třeba, aby pro každé slovo/slovní tvar mající obě interpretace (substantivní i adjektivní), existovala přísně ověřená korpusová evidence o výskytu obou případů. V opačném případě je třeba vybrat jedinou možnost." (Jakubíček a kol., 2011, s. 31) Slovníky není třeba v tomto směru masivně přegenerovávat a přetěžovat desambiguaci.

[20] SUB=U mají adjektiva pravidelně tvořená od substantiv (v podstatě pojmenování mužských a ženských osob) pravidelně a produktivně sufixy ***-ův/-in***, viz více kapitola Adjektiva.

[21] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Vzhledem k tomu, že tvary adjektiv produktivního slovotvorného typu tvořených od životných maskulin a od feminin označujících živé bytosti sufixy ***-ův/-in*** byly v pražském i brněnském slovníku rozgenerovány automaticky od příslušných substantiv, domníváme se, že informaci o poddruhu by nebylo nesnadné do brněnského slovníku doplnit, a to tím spíše, že ve starší verzi tagsetu (Jakubíček a kol., 2011, s. 40) se v tabulce uvádí značky **_,hF** a **_,hM**, které derivační historii zachycují. Derivační charakteristiky byly přesunutu do samostatné databáze, nicméně k dispozici jsou. V případě převodu by bylo třeba sjednotit značkování slovníku, viz příslušná pasáž v kapitole Adjektiva. V derivačním analyzátoru *Derivancze* mají adjektiva na ***-ův/-in*** značku ***k2pos*** (viz více https://nlp.fi.muni.cz/projects/derivancze/index.cgi a také Pala, Šmerk, 2015).

[22] SUB=G mají adjektiva pravidelně tvořená od sloves (s opěrným tvarem 3. osoby plurálu prézentu) pravidelně a produktivně sufixy ***-(ou)-c-í/-(í)-c-í***, viz více v kapitole Adjektiva.

[23] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Vzhledem k tomu, že tvary adjektiv produktivního slovotvorného typu tvořených od sloves sufixy ***-(ou)-c-í/-(í)-c-í*** byly v pražském i brněnském slovníku rozgenerovány automaticky od příslušných sloves, domníváme se, že informaci o poddruhu by nebylo nesnadné do brněnského slovníku doplnit. Původně označoval derivační historii slova v brněnském tagsetu atribut **r**, který nabýval hodnot **D**. Byl odstraněn a je dostupný ze samostatně budované databáze, v níž jsou uloženy informace o derivaci (viz více Jakubíček a kol., 2011, s. 34).V případě převodu by bylo třeba sjednotit značkování slovníku, viz příslušná pasáž v kapitole Adjektiva. V derivačním analyzátoru *Derivancze* mají adjektiva na *oucí/ící* značku ***k2proc*** (viz více https://nlp.fi.muni.cz/projects/derivancze/index.cgi a také Pala, Šmerk, 2015).

[24] SUB=M mají adjektiva pravidelně tvořená od sloves (s opěrným tvarem činného příčestí) pravidelně a produktivně sufixy ***-š-í/-vš-í***, viz více v kapitole Adjektiva.

[25] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Vzhledem k tomu, že tvary adjektiv produktivního slovotvorného typu tvořených od sloves sufixy ***-š-í/-vš-í*** byly v pražském i brněnském slovníku rozgenerovány automaticky od příslušných sloves, domníváme se, že informaci o poddruhu by nebylo nesnadné do brněnského slovníku doplnit. Původně označoval derivační historii slova v brněnském tagsetu atribut **r**, který nabýval hodnot **D**. Byl odstraněn a je dostupný ze samostatně budované databáze, v níž jsou uloženy informace o derivaci (viz více Jakubíček a kol., 2011, s. 34). V případě převodu by bylo třeba sjednotit značkování slovníku, viz příslušná pasáž v kapitole Adjektiva. V derivačním analyzátoru *Derivancze* mají adjektiva na *ší/vší* značku ***k2rakt*** (viz více https://nlp.fi.muni.cz/projects/derivancze/index.cgi a také Pala, Šmerk, 2015).

| | | | |
|---|---|---|---|
| **POS=A/k2** | **SUB=V**[26] | **---**[27] | **A (C)**[28] |
| **POS=A/k2** | **SUB=C**[29] | **---**[30] | **A** |
| **POS=A/k2** | **SUB=0**[31] | **---**[32] | **A** |

V tabulce 2b uvedeme korespondence hodnot kategorie SUB a atributů [xyz] v závislosti na slovním druhu spojek **J/k8/J**, a a **G/---/---**.

Tabulka 2b

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| Poddruh/ subklasifikace typu | **SUB** | **[zxy]** | **2. pozice** |
| **POS=J/k8** | **SUB=^** | **xC** | **^ spojka souřadicí** |
| **POS=J/k8** | **SUB=,**<br><br>***abych…, kdybych…*** **jsou agregáty** | **xS** | **, spojka podřadicí** **(vč. „aby“ a „kdyby“ ve všech tvarech)** |

[26] SUB=V mají adjektiva pravidelně tvořená od sloves (s opěrným tvarem shodným s trpným příčestím) pravidelně a produktivně sufixy ***-n-ý/-t-ý*** a dále (s opěrným tvarem činného příčestí) pravidelně a produktivně sufixem ***-teln-ý***, viz více kapitola Adjektiva.

[27] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Vzhledem k tomu, že tvary adjektiv produktivního slovotvorného typu tvořených od sloves sufixy ***-n-ý/-t-ý*** byly v pražském i brněnském slovníku rozgenerovány automaticky od příslušných sloves, domníváme se, že informaci o poddruhu by nebylo nesnadné do brněnského slovníku doplnit. Původně označoval derivační historii slova v brněnském tagsetu atribut **r**, který nabýval hodnot **D**. Byl odstraněn a je dostupný ze samostatně budované databáze, v níž jsou uloženy informace o derivaci (viz více Jakubíček a kol., 2011, s. 34). V případě převodu by bylo třeba sjednotit značkování slovníku, viz příslušná pasáž v kapitole věnované adjektivům a také doplnit značku u adjektiv na .**telný*. V případě úprav brněnského slovníku bude třeba zrevidovat přegenerování (viz více k tomuto tématu Jakubíček a kol., 2011, s. 32; Hájková, 2013). V derivačním analyzátoru *Derivancze* mají adjektiva na *ný/tý* značku ***k2rpas*****,** krátké tvary – pasivní participia na *n/t* mají zanačku ***k2pas*** (viz více https://nlp.fi.muni.cz/projects/derivancze/index.cgi a také Pala, Šmerk, 2015).

[28] Jmenné tvary adjektiv, mezi nimi i některé případy krátkých příčestí trpných (viz kapitola věnovaná Adjektivům), byly značkovány na 2. pozici jako C (adjektivum, jmenný tvar).

[29] SUB=C budou mít adjektiva (deriváty tvořené sufixy ***-ový, -itý,*** jako *dvojkový, dvojitý*, … a zejména **kompozita** jako *dvojsečný, dvojstranný, dvouhodinový, tříletý, čtyřprocentní*) z číslovkových kořenů s výjimkou adjektivně skloňovaných číslovek řadových (*první, druhý, třetí, pátý, stý, …*), číslovek vztažených k celku (*paterý, šesteý, desateý, …*) a kompozitních číslovek násobných (*dvojnásobný, osminásobný, dvacetinásobný, …*). Podrobnější rozbor viz kapitola věnovaná číslovkám.

[30] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Domníváme se ovšem, že nejde o kategorii problematickou. V případě úprav slovníků bude nutné vycházet z toho, že adjektivum s tímto poddruhem bude obsahovat číslovkový kořen. Těchto kořenů je omezený počet a možnosti automatické detekce jsou dosti dobré.

[31] SUB=0 budou mít adjektiva, která nemají SUB=[UGMVC].

[32] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Jde ovšem jen o doplňkovou množinu těch adjektiv, která nemají SUB=[UGMVC]. Převod by tudíž, pokud by se řešil komplexně, neměl činit potíže.

| | | | |
|---|---|---|---|
| | **spojkového typu s první složkou POS=J & SUB=,** | | |
| **POS=J/k.** | **SUB=*[33]** | **---[34]** | *** slovo *krát* (slovní druh: spojka)** |
| **POS=G/k.** | **SUB[35]** | **z=s[36]** | **(např. [ps][37])** |

**Vztah návrhu NovaMorf k brněnskému tagsetu (hodnoty SUB a DEI společné zájmenům, zájmenným číslovkám a zájmenným adverbiím)**

U zájmen, zájmenných číslovek[38] a zájmenných příslovcí[39] 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, kterou jsme nazvali 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.
Přehled vzájemných vztahů hodnot kategorií SUB a DEI u zájmen, zájmenných číslovek a zájmenných adverbií je patrný ze samostatných tabulek.
Tabulky 3a, 3b, 4a, 4b, 5a, 5b naznačují názorně vztah mezi hodnotami SUB a DEI v návrhu NovaMorf a subklasifikacemi atributů [xy] u k[346] v brněnském tagsetu. Tabulky 3c, 4c a 5c zaznamenávají značkování zájmen, číslovek a adverbií v pražském systému.

---

[33] V návrhu NovaMorf mají mít samostatnou značku (*) matematické operace (*plus, minus/mínus, krát, děleno* – neplést s *děleno* jako jmenným tvarem přídavného jména *dělený*, případně s adjektivním tvarem trpného příčestí od slovesa *dělit*). Předpokládáme, že jde o technickou záležitost, kterou by bylo možné společně řešit.

[34] V brněnském systému adekvátní značka není navržena. Je ovšem možné, že se s nějakým systémem značkování obdobných jevů počítá.

[35] U agregátů nemluvíme o poddruhu, protože dědí poddruhy všech svých složek. Typologie agregátů se značí pomocí kategorie Typ agregátu (AGR), viz kapitola věnovaná agregátům.

[36] Subklasifikaci **xS** mají v brněnském systému slovní tvary s volným morfémem *s* za sponu nebo auxiliár *být*. Ty také tvoří většinu jednotek (otevřená množina), které by podle návrhu NovaMorf měly mít POS=G. Kromě nich se v návrhu NovaMorf počítá se zájmennými agregáty, spřežkami tvořenými spojením předložky + krátkého tvaru zájmen *on* nebo předložky + zkráceného zájmena *co(pak)* (*-č, -čpak*), tedy **lc="(oň|proň|doň|zaň|naň|veň)|((nač|zač|oč|več)|(nač|zač|oč|več)pak)"**. Tyto tvary jsou v brněnském systému značkovány buď jako adverbia, nebo nejsou rozpoznány (jsou jim pak patrně na základě guesseru přiřazeny interpretace značně různorodé a budící rozpaky). Bylo by dobré koordinovat jejich doplnění do slovníku s jejich značkováním, které by umožnilo převod mezi oběma systémy. Jedná se o slova, jejichž správná desambiguace by mohla vylepšit i výsledky automatické syntaktické analýzy a aplikací na ní závisejících (například rozpoznání relativního *več, načpak, ...* jako tvarů, před nimiž předchází příslušná interpunkce.

[37] V pražském slovníku nebyly agregáty značkovány. Např. slovesné tvary s volným morfem *-s* zde mají bez odlišení tvarů bez tohoto morfému na druhé pozici značku [ps]. Mají ovšem vyznačen odpovídající význam osoby na 8. pozici. Totéž platí i pro některá další slova s volným morfem *-s* (např. spojku *žes* atd.).

[38] 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.

[39] Zájmenná adverbia mají zájmenné kořeny a zájmenné funkce.

V tabulce 3a je naznačen vztah mezi hodnotami SUB a DEI u zájmen (**POS=P**) v návrhu NovaMorf.
V tabulce 3a jsou uvedeny reprezentativní příklady, kompletní informaci o značkování zájmen obsahuje kapitola věnovaná zájmenům.

**Tabulka 3a (NovaMorf)**

| **POS=P** | **Určitá DEI=U** | **Neurčitá DEI=N** | **Záporná DEI=Z** | **Tázací DEI=T** | **Vztažná DEI=V** | **Reflexivní DEI=S** | **Ukazovací DEI=D** |
|---|---|---|---|---|---|---|---|
| **Osobní/ SUB=o** | ***já, ty, ...*** | **----** | **----** | **----** | **----** | ***se, ...*** | **----** |
| **Substantivní/ SUB=N** | **----** | ***někdo, ...*** | ***nikdo, ...*** | ***kdo, ...*** | ***jenž, ...*** | **----** | **----** |
| **Přivlastňovací/ SUB=U** | ***můj, ...*** | ***něčí, ...*** | ***ničí, ...*** | ***čí, ...*** | ***jehož, ...*** | ***svůj, ...*** | **----** |
| **Vymezovací/ ostatní/ SUB=v** | ***týž, jiný, sám, každý, všechen, ...*** | ***nějaký, některý, ...*** | ***nijaký, žádný, nijeden, ...*** | ***jaký, který, ...*** | **----** | **----** | ***ten, takový, onen, ...*** |

V tabulce 3b je naznačen vztah mezi odpovídajícími hodnotami atributů u zájmen (**k3.***) v brněnském tagsetu. Tagsety zachycují odpovídající si jevy obdobně. Rozdíly jsou u zájmen tázacích a vztažných. Návrh NovaMorf počítá u vztažných zájmen pouze se zájmeny *jenž, an,* zatímco v brněnském systému figurují mezi vztažnými zájmeny i zájmena tázací (homonyma: <***Koho***> *máš na mysli?* × *Volím toho,* <***koho***> *znám.*). Ta sice plní funkce relativních spojovacích výrazů, nicméně od desambiguace obou funkcí na rovině morfologické analýzy návrh NovaMorf upouští (nejde o morfologickou, ale o syntaktickou analýzu). Další rozdíly jsou patrně na úrovni slovníku (například *oba, obě* je v brněnském i pražském slovníku značkovány jako číslovka, *jiný, jediný* jako adjektivum atd.). Domníváme se, že oba tagsety jsou v zásadě kompatibilní.

**Tabulka 3b (Brno)**

| **k3** | **neurčuje se** | **I Indeterminate y=I** | **N Negative y =N** | **Q Interrogative y=Q** | **R Relative y=R** | **F Reflexive y =F** | **vynechán** |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **P Personal x= P** | ***já, ty, …*** | | | | | ***se*** | |
| **vynechán** | | ***někdo některý, ...*** | ***nikdo, žádný, ...*** | ***kdo, který, ...*** | ***kdo, který, ...*** | | |
| **O Possessive x=O** | ***můj, tvůj, ...*** | | | | | ***svůj, ...*** | |
| **D Demonstrative x=D** | | | | | | | ***ten, takový, onen, ...*** |
| **T Delimitative x=T** | ***týž, sám, každý, všechen, ...*** | | | | | | |

V tabulce 3c je přehled značkování zájmen na 2. pozici v pražském systému.

**Tabulka 3c (Praha)**

| **1. & 2. pozice** | **tvary příklady** | **popis** |
|---|---|---|
| P0 | *naň* | spřežka předložka+osobní zájmeno *on* |
| P1 | *jehož* | vztažné zájmeno *jehož* |
| P4 | *jaký, který* | tázací zájmeno *který, jaký, čí, jakýpak, kterýpak, čípak, kterýž, jakýž, jakýže, ...* |
| P5 | *něj* | osobní zájmeno *on* tvary po předložce (*n-*) |
| P6 | *sebe* | zvratné zájmeno tvary *sebe, sobě, sebou* |
| P7 | *se, si* | zvratné zájmeno tvary *se, si, ses, sis* |

| | | |
|---|---|---|
| P8 | *svůj* | přivlastňovací zvratné zájmeno *svůj*[40] |
| P9 | *něhož* | vztažné zájmeno *jehož* tvary po předložce (*n*-) |
| PD | *tento* | ukazovací zájmena *ten, tento, takový, tenhle, onen, týž, tentýž, takovýto, takovýhle, tenhleten, toť, tamten, taký, tamhleten, tadyten, tuhleten* |
| PE | *což* | vztažné zájmeno *což* |
| PH | *mě* | krátké (příklonné) tvary osobních zájmen *mi, mě, ti, tě, ji, je, ...* |
| PJ | *jenž* | vztažné zájmeno *jenž* |
| PK | *kdo* | vztažné/tázací zájmeno *kdo, kdopak, kdožpak, kdož, kdos* |
| PL | *všechen* | zájmena vymezovací (limitativa) *všechno, všecek, sám, samý, veškerý* |
| PP | *ty* | osobní zájmena *já (my), ty (vy), on* tvar *tys* |
| PQ | *co* | vztažné/tázací zájmeno *co, copak, cožpak, cos, což* |
| PS | *můj* | přivlastňovací osobní zájmena *můj, tvůj, jeho, náš, váš* |
| PW | *nic* | záporná zájmena *nic, žádný, nikdo, pranic, nijaký, pražádný, nižádný* |
| PY | *oč* | spřežka vztažné/tázací zájmeno předložka+*č* (*oč, nač, zač, več, ...*) |

[40] Pražský systém značkuje samostatně stojící tvary *nesvůj, nesvá, nesvé* jako slovní druh adjektivum, detailní určení slovního druhu má toto „adjektivum" společné s tvary *tentam, totam*. Tvar *tatam* je označkován XX (neznámé slovo). (Mezi doklady z korpusu SYN2000 jsme nalezli doklad *... kdo se chytne **nesvé** hvězdy ...*, kde by snad mohlo jít o negaci posesivního zájmena *svůj*, nikoli o frazeologismus. (Doklad pochází z beletristického textu – Vaculíkova románu „Jak se dělá chlapec". Z kontextu lze předpokládat význam *chytne se **cizí – té, která není jeho –** hvězdy*, nikoliv *chytne se hvězdy, která **není ve své kůži**.*)

| | | |
|---|---|---|
| PZ | *nějaký, něco* | neurčitá zájmena *některý, něco, nějaký, někdo, jakýsi, jakýkoli, jakýkoliv, cosi, cokoliv, málokdo, kdosi, kdokoli*<br>*kterýkoli, leccos, kdokoliv, ničí, kterýkoliv, všelijaký, kdekdo, málokterý, leckdo, leckterý, něčí, ledacos, kdejaký, kterýsi, jakýs, kdeco, máloco, čísi, takýs*[41], *bůhvíjaký, ledajaký, bůhvíco, lecjaký, všelicos, kdovíjaký, lecco, kdekterý, kdože, kdovíco, ledasco, ký, ledaco, ledaskdo, nevímjaký, bůhvíkdo, kdovíkdo, všelico, čertvíkdo, čertvíco, číkoliv, nevímkdo, číkoli, nevímčí, ledakdo, kdovíčí, zřídkakdo, ledakterý, čertvíjaký, všelikerý* |

Ve skupině osobních zájmen (tag=PP.*) mají zvláštní značku tvary zájmena *on* po předložce (tvary *něho, němu, něj, ně, něm, ním, ni, ní, nich, nimi* mají tag=P5.*) a krátké (příklonné) tvary osobních (tag=PH.*) i zvratných (tag=P7.*) zájmen. Lemmatem krátkých i dlouhých tvarů zájmen osobních jsou příslušné nominativní tvary. Lemmatem krátkých i dlouhých tvarů zvratného osobního zájmena (tag=P6.*) je tvar *se*.
Tvar zájmena *ty* s nesamostatným morfémem *-s* za 2. os. pomocného slovesa *být* (*tys*) má stejnou značku jako ostatní tvary zájmena *ty.* Ve značce je uvedena kategorie osoby, ta ovšem je uvedena v tomto případě proto, že jde o zájmeno 2. osoby, nikoli proto, že jde o tvar s nesamostatným *-s* za tvar 2. osoby pomocného slovesa *být.* Lemma je *ty* a ne *tys*, takže se ztrácí informace o tom, že jde o spřežku s tvarem slovesa *být* na úrovni lemmatu, stejně je řešena lemmatizace tvarů *ses*, *sis*, které mají lemma *se*. Problematicky se řeší lemmatizace a značkování tvarů *kdos...* (viz kapitola věnovaná zájmenům).
Zvláštní značku mají zájmenné spřežky předložka + tvar zájmena *on* (tvary *naň, zaň, proň, doň* mají tag=P0.*), chybí tvar *oň*, který má značku X (neznámé slovo). Důvodem samostatného označení těchto tvarů je jejich problematická lemmatizace, která je v rámci projektu NovaMorf vyřešena zavedením slovního druhu agregát (POS=G.*), viz více v příslušné kapitole. V rámci značky se uvádí pádová platnost příslušného tvaru. Lemma je tvar sám.
Tázací a vztažná zájmena mají několik různých značek. Zájmena *kdo, co* a některá od nich odvozená jsou jakožto substantivní zájmena bezrodá označkována samostatně (tag=P[KQ].*), jiná nikoliv (záporná – tag=PZ.* a neurčitá – tag=PW.*).
Tvary zájmena *jenž* na straně jedné a tvary téhož zájmena po předložce (tvary na *n-* např. o *němž,* s *nímž, ...*) na straně druhé jsou označkovány samostatně dvěma různými značkami (tag=P[J9].*).
Tvary zájmen *kdo, co, který, taký, jaký* + *-s* jsou v pražském slovníku lemmatizovány a označkovány velmi nejednotně (viz podrobně kapitola Zájmena). Tvary zájmen *kterýs, jakýs, takýs* jsou lemmatizovány jako zkrácené tvary zájmen (*kterýs<kterýsi, jakýs<jakýsi, takýs<takýsi*), což není vždy správně (viz výše).

[41] Tvar *takýs* může být tvarem *taký+s*, kde *-s* zastupuje tvar 2. os. pomocného slovesa *být.* (V korpusu SYN v7 jsou ovšem pouze doklady na užití ve frazeologickém spojení *jakýs takýs*). Zájmeno *jakýs* se ve významu *jaký+ -s* za 2. os. pomocného slovesa *být* okrajově objevuje.
Nesoustavnosti se vyskytují i ve značkování tvarů odvozených od zájmena *který*. Značku tag=PZ.* mají tvary *kterýs, kterás, kterés, kteréhos, kterému, kterýms*, ale tvary *kterous, kterýhos, kterejchs, kterýchžs* nejsou rozpoznány automatickou morfologickou analýzou (mají značku tag=X.*).

Vztažné *seč* (tvar zájmena *co*) je označkováno jako XX. Ostatní spřežky předložka+*č* za zájmeno *co* (*nač, oč, več, zač* ) mají samostatnou značku (tag=PY.*), v níž se ovšem neuvádí pádová platnost tvaru.
Samostatně je lemmatizován tvar *toť* lemma *toť* (tag=PD.*). Tvary *tenť, totoť* jsou ovšem označeny X@, stejně jsou označeny tvary osobních zájmen *jáť, onť*. Ve značce se u tvaru *toť* uvádí pouze to, že jde o ukazovací zájmeno. Význam *-ť* (částice kladená jako příklonka u některých slov) se neznačkuje. V tagsetu existuje zvláštní značka pro archaické tvary sloves na *-ť*, takže se vnucuje otázka, proč nejsou analogické případy řešeny jednotně, tedy proč nemají zájmena s připojeným *-ť* samostatnou značku analogicky jako slovesa.
Zájmena *týž, tentýž* jsou řazena mezi ukazovací (tag=PD.*). Vzhledem k tomu, že v tagsetu existuje samostatná značka pro kategorii limitativ (tag=PL.*), není důvod zařazení zájmen *týž, tentýž* mezi zájmena ukazovací bez dalšího upřesnění průhledný.
Totalizátory *každý, všeliký* a vymezovací, *samotný/samoten* a limitativní zájmeno (alterátor) *jiný* s adjektivní flexí jsou značkovány jako adjektiva.

V tabulce 4a je naznačen vztah mezi hodnotami SUB a DEI u číslovek (se zvláštním zřetelem k zájmenným číslovkám) v návrhu NovaMorf. V tabulce jsou uvedeny reprezentativní příklady, kompletní informaci o značkování číslovek obsahuje kapitola věnovaná číslovkám.

**Tabulka 4a (NovaMorf)**

| **POS=C** | **Určitá**<br>**DEI=U** | **Neurčitá**<br>**DEI=N** | **Tázací**<br>**DEI=T** | **Ukazovací**<br>**DEI=D** |
|---|---|---|---|---|
| **Základní**<br>**SUB=z** | ***jeden, dva, sto, ...*** | ***několik*** | ***kolik*** | ***tolik*** |
| **Řadové**<br>**SUB=r** | ***první, druhý, stý, ...*** | ***několikátý*** | ***kolikátý*** | ***tolikátý*** |
| **Násobné**<br>**SUB=n** | ***jednou, dvakrát, stonásobný, stonásobně, ...*** | ***několikrát*** | ***kolikrát*** | ***tolikrát*** |
| **Dílové**<br>**SUB=h** | ***půl, čtvrt, polovina, třetina, čtvrtina, …*** | ***několikátina*** | ***kolikátina*** | |

| | | | | |
|---|---|---|---|---|
| **Vztažené k celku SUB=u** | ***dvé, patero, dvoje, patery, dvojí, paterý, dvojice, pětice, ...*** | ***několikatero, několikatery, několikaterý, několikátice*** | ***kolikatero, kolikatery, kolikaterý, kolikátice*** | ***tolikatero, tolikatery, tolikaterý, tolikátice*** |

V tabulce 4b níže je naznačen vztah mezi odpovídajícími hodnotami atributů u číslovek (**k4.*** a někdy i k6.*, popř. dalšími) v brněnském tagsetu. Tagsety zachycují odpovídající si jevy obdobně. Rozdíly jsou u tzv. zájmenných číslovek *tolik, kolik, několik* a pravidelných derivátů od uvedených zájmenně číslovkových základů. Opět se objevuje dvojí interpretace tázací a vztažná u lemmat *kolik, kolikátý*. Ta sice plní funkce relativních spojovacích výrazů, nicméně od desambiguace obou funkcí na rovině morfologické analýzy návrh NovaMorf upouští. Další rozdíly jsou v přegenerování vícera interpretací například u lemmatu *tolik* a *kolikátý.* Lemma *několikátý* má interpretaci adjektivní. Další drobné rozdíly existují patrně na úrovni slovníku (například *oba, obě* je v brněnském slovníku značkováno jako číslovka). Domníváme se, že oba tagsety jsou v zásadě kompatibilní. Návrh NovaMorf je propracovanější v tom smyslu, že se jednak snaží zjednodušit interpretace obtížné k desambiguaci (funkční rozdíl mezi tázacími a relativními zájmeny a zájmennými číslovkami), jednak díky zařazení významu **ukazovací** (zájmeno, číslovka, příslovce) na stejnou úroveň (DEI), jako jsou zařazeny významy **tázací** a **neurčitý** brání vzniku **„střetu zájmu“** (více rozumně nedesambiguovatelných interpretací), který je patrný v brněnském tagsetu, viz níže tabulka (**oranžově podbarvené**).
Brněnský systém netaguje deriváty tvořené sufixoidem ***-krát*** od základů číslovek určitých jako **k4.***, nýbrž pouze jako **k6.***. Kompozita tvořená od základů číslovek určitých druhým členem ***-násobný*** jsou tagována jako **k2.*** (adjektiva). Podobně **názvy zlomků** od základů číslovek určitých (**dílové číslovky**) a názvy tvořené od základů číslovek určitých sufixem *-ice* (**číslovky vztažené k celku**) jsou tagovány jako **k1.*** (substantiva). V těchto rysech se brněnský systém odlišuje od návrhu NovaMorf.

**Tabulka 4b (Brno)**

| **k[421]** | **vynechán** | **I Indeterminate y=I** | **N Negative y=N** | **Q Interrogative y=Q** | **R Relative y=R** | **vynechán** |
|---|---|---|---|---|---|---|
| **Základní x=C** | ***Jeden, dva, sto*** | | ***nijeden, padesátitisíce, statisíce***[42] | | | |
| **Řadové x=O** | ***první, druhý, stý*** | ***několikátý, kolikátý,*** | | ***kolikátý*** | ***kolikátý*** | ***tolikátý*** |

[42] Zařazení dvou lemmat (*padesátitisíce* a *statisíce*) je patrně řešením ad hoc. V návrhu NovaMorf je *nejeden* neurčitá číslovka základní (POS=C & SUB=z & DEI=N), *nijeden* je záporné zájmeno vymezovací/ostatní (POS=P & SUB=v & DEI=Z), *nejednou* je adverbium číslovkové (POS=D & SUB=C) a *nijednou* je adverbium číslovkové (POS=D & SUB=C).

| | | | | | | |
|---|---|---|---|---|---|---|
| **x=R** | ***dvoje, patery, dvojí, paterý, ...*** | ***kolikatero, kolikatery, kolikaterý,*** | | | | |
| **x=D** | | | | | | ***tolik***[43] |
| **vynechán** | | ***tolik, několik*** | | ***kolik*** | ***kolik*** | |

V tabulce 4c je přehled značkování číslovek na 2. pozici v pražském systému.

**Tabulka 4c**

| **1.&2. pozice** | **tvary** | **popis** |
|---|---|---|
| C= | *1* | arabská čísla |
| C} | *XIV* | římská čísla |
| Ca | *mnoho* | tvary „číslovky“ *mnoh-o,-a, ...* |
| Cd | *čtverý* | druhové číslovky *dvojí, obojí, trojí,* a další tvořené sufixem *-erý,* |
| Ch | *jedny* | druhová číslovka *jedny* |
| Cj | *čtvero* | úhrnné číslovky *dvé, obé, tré,* a další tvořené sufixem *-ero* |
| Ck | *čtvery* | souborové číslovky *dvoje, oboje, troje,* a další tvořené sufixem *-ery* |
| Cl | *tři* | základní číslovky *jeden, dva, oba, tři, čtyři* |
| Cn | *pět* | základní číslovky *pět* a výše |
| Co | *tolikrát* | číslovka zájmenná ukazovací násobná *tolikrát* |

[43] Lemma je interpretováno i jako adverbium.

| | | |
|---|---|---|
| Cr | *druhý* | číslovky řadové |
| Cu | *kolikrát* | číslovka zájmenná tázací násobná *kolikrát* |
| Cv | *sedmkrát* | číslovky určité násobné . *-*krát* |
| Cw | *nejeden* | *nejeden* |
| Cy | *desetina* | číslovky dílové vyjadřující určitý počet . *-*ina* |
| Cz | *kolikátý* | číslovka zájmenná tázací/vztažná řadová *kolikátý* |

Oba systémy (brněnský i pražský tagset) neberou v úvahu zájmenné použití některých tvarů číslovek. Jedná se jednak o tvary, které automatická morfologická analýza derivuje od lemmatu *jeden*. Tvar *jednou* je interpretován jako číslovka základní, násobná, nebo jako příslovce. Další tvary jsou interpretovány jednak jako tvary číslovky základní, nebo druhové (tvary plurálového subparadigmatu *jedny, jedněch, ...*). Brněnský systém uvádí navíc možnost interpretovat tvar *jednou* jako částici. Zájmenné použití není zohledněno ani v jednom z obou analyzátorů.
Spřežky předložka + tvar *jeden* (*pojednou, najednou, zajedno*) jsou oběma systémy interpretovány jako příslovce. Brněnský systém navíc u tvaru *najednou* uvádí možnost interpretovat jej jako částici.
Tvary slova *druhý* jsou v pražském systému označeny jako slovní druh adjektivum. Tvar *druhdy* se chápe jako příslovce. Tvary *podruhé* interpretuje jako číslovku násobnou, tvar *zaprvé* jako příslovce, tvar *zadruhé* není rozpoznán (tag=XX.*), tvary *zatřetí, začtvrté* rovněž (tag=X@.*).
V brněnském systému jsou tvary *zaprvé, zadruhé, ...* označkovány jako číslovky, tvary *poprvé, podruhé, ...* jako příslovce.
Tvary *napoprvé, napodruhé, ...* jsou označeny jako příslovce v obou systémech. Tvary *dvojmo, trojmo, dvojitě, trojitě* interpretují oba systémy shodně jako adverbia. Tvary *dvojitý, trojitý* jako adjektiva.
Značkování slovních tvarů *mnoho, více, nejvíce* je do jisté míry z lingvistického hlediska zjednodušující v obou systémech (více k tomuto viz Osolsobě, 2008).
Tvar *mnohý* je analyzován oběma systémy jako slovní druh adjektivum. Brněnský systém navíc nabízí možnost zpodstatnělého adjektiva.

V tabulce 5a je naznačen vztah mezi hodnotami SUB a DEI u zájmenných adverbií v návrhu NovaMorf. V tabulce jsou uvedeny reprezentativní příklady, kompletní informaci o značkování zájmenných adverbií obsahuje kapitola věnovaná příslovcím.

**Tabulka 5a (NovaMorf)**

| POS=D | Určitá DEI=U | Neurčitá DEI=N | Záporná DEI=Z | Tázací DEI=T | Ukazovací DEI=D |
|---|---|---|---|---|---|
| Zájmenná SUB=P | *pokaždé, všude, vždy, jinde, jindy, jinak, ...* | *někde, někdy, nějak, ...* | *nikde, nikdy, nijak, ...* | *kde, kdy, jak, ...* | *tam, tehdy, tak, ...* |

**Tabulka 5b (Brno)**

| k6 | vynechán | I Indeterminate y=I | N Negative y=N | Q Interrogative y=Q | R Relative y=R | vynechán |
|---|---|---|---|---|---|---|
| D Demonstrative x=D | | | | | | *Tam, tehdy, tak, zde, sem, …, tak, takhle, takto,* |
| T Delimitative x=T | *pokaždé, všude, vždy, ...* | | | | | |
| vynechán | | *někde* | *nikde* | *kde* | *přičemž, načež, pročež, seč, pokud, kdežto, začež, kdežt'* | |

Návrh NovaMorf zájmenná adverbia SUB=P, tj. místní (*kudy, tudy, odkud, kamkoli, odkudkoli, nikudy, nikam...*), časová (*kdy, kdykoli, nikdy…, pokaždé, jedině, jedinkrát(e), po(v)obakakrát(e)*) a způsobová (*jak, všelijak, jakkoli, nijak, ...*), ani další sémantické třídy adverbií v rámci kategorie SUB nerozlišuje. V brněnském tagsetu se rozlišují u lemmat označkovaných jako adverbia (**k6.***) hodnoty atributu **x demonstrativní (D)** a **delimitativní (T)**, podobně jako u **k3.***, kde atribut **x** nabývá hodnot **[POTD]**, a u **k4.***, kde atribut **x** nabývá hodnot **[CORD]**, přičemž toto rozlišení odpovídá do jisté míry návrhu NovaMorf, jak je vidět z tabulky 5b. Hodnoty **neurčité, negativní, tázací,** jsou v obou tagsetech uvedeny, ve slovníku se u adverbií realizují podobně v obou systémech. Hodnotu **vztažné** má i u adverbií brněnský systém (viz tabulka 5b). V návrhu NovaMorf nebudou žádné výrazy značkovány jako

zájmenná vztažná adverbia.[44] Původně vztažné zájmenné příslovečné spřežky *přičemž, načež, začež, pročež, natož, protož, tož, ...* a snad i archaické *očež, počemž*, a také *seč, kdežť, když, pokud, potud* jsou užívány výhradně jako spojovací výrazy, jsou tudíž řazeny mezi spojky POS=J, viz kapitola Spojky.[45]
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.

V tabulce 5c je přehled značkování adverbií na 2. pozici v pražském systému.

**Tabulka 5c**

| POS&SUBPOS | tvary – příklady | popis |
|---|---|---|
| Db | *nahoru* | všechna příslovce, která nelze stupňovat |
| Dg | *rychle* | příslovce, která lze stupňovat |
| D! | | zastaralý |

Návrh NovaMorf počítá s další subklasifikací nezájmenných adverbií na rovině kategorie SUB, která se s brněnským i pražským systémem stýká jen zcela okrajově. Tuto klasifikaci názorně ukazuje tabulka 5d (viz více kapitola věnovaná adverbiím).

**Tabulka 5d**

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| Poddruh/ subklasifikace typu | **SUB** | **t** | **2. pozice** |

[44] V kapitole věnované zájmenům je podrobněji popsáno zdůvodnění nezavedení homonymních tázacích a vztažných zájmen do slovníku. Stejné principy platí i pro zájmenná adverbia.
[45] Stejnou funkci spojovacích vztažných výrazů ovšem mohou plnit i agregáty se zájmennou složkou (*nač, zač, več, oč, cos, očs, ...*, tedy POS=(R&P|P&V|R&P&V) AGR=[cPG]). Ty by pak měly mít DEI=T, a to jak v případech jako <***Nač***> *se ptáš? /* <***Cos***> *říkal? /* <***Načs***> *narážel?*, tak v případech jako *To,* <***nač***> *nikdo z nás nebyl připraven, byla moje reakce. / Ta z fotky,* <***cos***> *našla u Kitty v ložnici. / Kadára sháním, abych mu od Tebe vyřídil,* <***očs***> *mě žádal. ...*. Viz kapitoly věnované agregátům a zájmenům.

| | | | |
|---|---|---|---|
| **POS=D/k6** | **SUB=s**[46] | **---**[47] | **---** |
| **POS=D/k6** | **SUB=C**[48] | **t=Q**[49] | **----** |
| **POS=D/k6** | **SUB=V**[50] | **---**[51] | **----** |
| **POS=D/k6** | **SUB=0**[52] | **---**[53] | **----** |

## Značkování jmenných kategorií rodu (GEN/g), čísla (NUM/n) a pádu (CAS/c) v návrhu NovaMorf v porovnání s brněnským a pražským systémem

Ve značkování jmenných kategorií neexistují ve srovnávaných systémech výrazné rozdíly. Shoda panuje jak v kategoriích/atributech, tak v hodnotách, které nabývají. Kategorii rodu a čísla mají ve značce uvedeny tvary sloves,[54] které prostřednictvím těchto kategorií vyjadřují

[46] Značku budou mít adverbia tvořená ustrnutím předložkových pádů adjektiv (nikoli zájmen, ta budou zařazena mezi zájmenná adverbia) a substantiv (nikoli číslovek určitých a číslovky neurčité *mnoho*, viz podrobně kapitola Číslovky). Pomocí této značky se zachycuje derivační historie (vznik příslovečné spřežky). Návrh směřuje k tomu, aby bylo možné propojit spřežky typu *natvrdo* s dvouslovnými adverbiálními výrazy typu *na tvrdo*, viz více kapitola věnovaná adverbiím.

[47] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování.

[48] Návrh NovaMorf u adverbií na rovině SUB značkuje vztah k vyjádření množství (slovnědruhový přesah s kategorií číslovek). Je to návrh restriktivní. SUB=C budou mít adverbia derivovaná pravidelně z adjektiv, která budou mít SUB=C (například adverbia jako *trojitě, dvojkově, čtyřprocentně* viz více kapitoly věnované číslovkám a adjektivům) a adverbiální užití slovnědruhově přesažných číslovek (*mnoho*) a substantiv (*málo, moc*). Pomocí SUB=C je řešen statut **vybraných** měrových adverbií (*hodně*) a měrových adverbií druhého a třetího stupně, která nelze vztáhnout k jednoznačnému tvaru pozitivu (*více, méně*). Jako POS=D & SUB=C **nebudou** označkovány ustrnulé tvary substantivního původu označující množství jako *trochu, trošku, kapku, trošinku, kapičku, trošičku, …* Některé z nich jsou označkovány v současnosti jak jako adverbia (tag=Db.*), tak jako substantiva. Podobné (např. *spoustu*) pouze jako substantivum. Návrh NovaMorf směřuje ke zjednodušení desambiguace, pro kterou neexistuje všeobecná shoda, a k tomu, aby značkování slov různých slovních druhů, která plní funkci kvantifikátorů, nezatěžovalo automatickou morfologickou analýzu problémy, které na rovinu morfologie striktně vzato nepatří. Tato slova budou buď substantiva (například: *moře, hromada, kupa, spousta, …*), nebo substantiva i adverbia, tedy substantiva, pokud se skloňují a rozvíjejí jméno a adverbia, pokud v ustrnulém tvaru rozvíjejí sloveso: *dej mi* <***trochu/N***> *vody a já ti* <***trochu/D***> *pomůžu*).

[49] V brněnském slovníku jsou jako tag=”k6.*tQ.*” (adverbia míry) značkována kupříkladu lemmata *tak, hodně, velmi, moc, daleko, málo, trochu, dost, zcela, příliš, docela, velice, mnohem, většinou, takhle, takto, pomalu,* ***tolik****, trošku, …* Návrh NovaMorf je v případě POS=D & SUB=C restriktivní a zahrnuje lemmata, jejichž výběr působí méně nesourodě. Ke kritice lingvistické adekvátnosti sémantické klasifikace adverbií v brněnském systému, viz Hvězdová 1999.

[50] POS=D & SUB=V budou mít adverbia derivovaná pravidelně z adjektiv, která budou mít SUB=V (například: *nepřejícně, pohnutě, nepokrytě, pochopitelně, …* viz více kapitola Adjektiva).

[51] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Existovalo v něm ovšem značkování derivační historie slova. Původně byla značkována atributem **r**, který nabýval hodnot **D**. Byl odstraněn a je dostupný ze samostatně budované databáze, v níž jsou uloženy informace o derivaci (viz více Jakubíček a kol., 2011, s. 34).V případě převodu by bylo třeba sjednotit značkování slovníku.

[52] POS=D & SUB=0 mají adverbia, která nemají SUB=[PsCV].

[53] Pro tuto kategorii neexistuje v brněnském systému adekvátní značkování. Jde ovšem jen o doplňkovou množinu těch adverbií, která nemají SUB=[PsCV]. Převod by tudíž, pokud by se řešil komplexně, neměl činit potíže.

[54] Jediný problém představuje sjednocení tvarů *by, aby, kdyby, …*, neboť v brněnském návrhu (Jakubíček a kol., 2011) se, alespoň se tak zdá, nepočítá s tím, že by ve značkách byly uvedeny atributy **p** (osoba) a **n** (číslo) s

gramatickou shodu. Kategorii pádu mají ve značce uvedenu také prepozice v závislosti na tom, s kterým pádem jména se prepozice pojí.
Tabulky 6, 7, 8 ukazují na shody všech tří systémů.

**Tabulka 6**

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| **jmenný rod** | **GEN** | **g** | **3. pozice** |
| **POS=[NAPC]/k[1234]** | **GEN=M** | **gM** | **M** |
| **POS=[NAPC]/k[1234]** | **GEN=I** | **gI** | **I** |
| **POS=[NAPC]/k[1234]** | **GEN=F** | **gF** | **F** |
| **POS=[NAPC]/k[1234]** | **GEN=N** | **gN** | **N** |
| **POS=** | **GEN=[MIFN]** | **----** | **X**[55] |
| **POS=** | **[GEN=F & NUM=S] \| [GEN=N & NUM=P]** | **----** | **H**[56] |
| **POS=** | **[GEN=[FI] & NUM=P]** | **----** | **T**[57] |
| **POS=** | **[GEN=[MI] & NUM=S]** | **----** | **Y**[58] |

Kromě tradičních hodnot zachovává návrh NovaMorf již zavedenou praxi pražského i brněnského systému a rozlišujeme dva mužské rody. Na rozdíl od pražského systému nepřipouští částečně sdružené hodnoty (proměnné) pro případy, které nelze na základě žádné indicie ani širšího kontextu rozhodnout; disambiguaci řeší buď disjunkcí hodnot (nedisambiguuje), nebo jednoznačnou disambiguací podle rodové hierarchie (M>F>I>N), viz více v kapitole Substantiva. Brněnský systém se sdruženými hodnotami explicitně nepracuje. Přesto v desambiguační praxi užívá pravidla, která zjednodušují tuto praxi a odpovídají tak jiným způsobem řešení (viz více Jakubíček a kol., 2011) na potřebu vyhnout se složité desambiguaci.

příslušnými hodnotami. Vzhledem k flexibilitě brněnského systému by ovšem dodání těchto informací do značky neměl být problém. K tomuto bodu viz níže a též příslušná kapitola Spojky a partie o kondicionálu v kapitole Morfologické kategorie a jejich hodnoty.

[55] Sdružená hodnota pro libovolný rod.

[56] Ženský (v singuláru) nebo střední (v plurálu).

[57] Mužský neživotný nebo ženský (obojí v plurálu).

[58] Mužský životný nebo neživotný.

**Tabulka 7**

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| **číslo** | **NUM** | **n** | **4. pozice** |
| **POS=[NAPCV]/k[12345]** | **NUM=S** | **nS** | **S** |
| **POS=[NAPCV]/k[12345]** | **NUM=P** | **nP** | **P** |
| **POS=[NAPCV]/k[12345]** | **NUM=[SP]** | **----** | **X**[59] |

**Tabulka 8**

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| **pád** | **CAS** | **c** | **5. pozice** |
| **POS=[NAPCR]/k[12347]** | **CAS=[1234567]** | **g[1234567]** | **[1234567-]** |
| **POS=[NAPCR]/k[12347]** | **CAS=[1234567]** | **----** | **X**[60] |
| **POS=[NAPCR]/k[12347]** | **CAS=[24] \| [17] \| ...]**[61] | **----** | **----** |

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

Pro slovesa je v návrhu NovaMorf relevantní kategorie poddruh (SUB), vid (ASP), osoba (PER) a číslo (NUM viz výše), slovesný tvar (VRB) a pro krátké tvary n/t-ových příčestí[62] také

[59] Sdružená hodnota pro libovolné číslo.

[60] Sdružená hodnota pro libovolný pád.

[61] Disjunkce hodnot pro nedesambiguovatelné kombinace pádu, například a: kolísání mezi 2. a 4. pádem (ve větě *Užívá **stavení***, kde *stavení* lze interpretovat jako 2. pád (*užívá hezkého **stavení***), nebo jako 4. pád (*užívá hezké **stavení***), viz více v kapitole Substantiva.

[62] Ta jsou v návrhu NovaMorf značkována jako POS=A. Také v brněnském systému je snaha sjednotit značkování těchto tvarů s jim odpovídajícími dlouhými tvary adjektiv (viz Jakubíček a kol., 2011, s. 33). Interpretace krátkých tvarů (pasivních příčestí) byla totiž v brněnském slovníku masivně přegenerována, a to tak, že krátké tvary měly jak interpretaci k5, tak k2.

kategorie jmenný tvar (NOM),[63] viz více v příslušných kapitolách věnovaných adjektivům a slovesům.
Poddruh sloves má v návrhu NovaMorf dvě hodnoty. SUB=b pomocná (*být*,[64] *bývat*, *bývávat*) a SUB=0 (všechna ostatní), viz více v kapitole věnované morfologickým kategoriím a hodnotám.
Ani brněnský, ani pražský systém subklasifikaci sloves nemají. Vzhledem k tomu, že návrh NovaMorf je značně restriktivní, domníváme se, že sjednocení obou systémů by nemělo narazit na vážné překážky.
Hodnoty kategorií vid, osoba i číslo a slovesný tvar mají své protějšky v brněnském i pražském systému, takže sjednocení by nemělo narážet na významné překážky. Tabulky 9 a 10 ukazují korespondence mezi návrhem NovaMorf a brněnským i pražským systémem.

**Tabulka 9**

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| **vid** | **ASP**[65] | **a**[66] | **16. pozice** |
| **POS=V/k5** | **ASP=[DNO-]** | **a[PI]/aB**[67] | **[PIB-]** |

Tabulka 10

| | NovaMorf | Brno | Praha |
|---|---|---|---|
| **osoba** | **PER** | **p** | **8. pozice** |

[63] 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 lemma (například {*schopný, schopen*}, {*hrdý, hrd*}, {*ukrytý, ukryt*}, {*pokáraný, pokárán*}). Krátké (jmenné tvary) budou mít NOM=J, dlouhé tvary NOM=0. Tato kategorie bude relevantní pro adjektiva (zejména adjektiva tvořená ze sloves), zájmena ({*sám, samý*}), číslovky (tvary jmenné a složené u číslovek typu *devatero/devateré*) a pro adverbia (ustrnulé jmenné tvary adjektiv v adverbiální funkci po předložce, například *za studena*). Tato kategorie nemá sice obdobu v brněnském systému, přesto je její ideové východisko kompatibilní s brněnským systémem.

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

[65] Stejně jako dosavadní morfologické systémy nezavádí ani NovaMorf další hodnoty kategorie pro iterativní slovesa, i když jsou, pokud je to možné, při generování slovníku pravidelně doplňována. Ve valenčním slovníku VALLEX (viz http://ufal.mff.cuni.cz/vallex), ale i v systému užívaném na Slovensku pro značkování SNK, jsou vidové dvojice zpracovány jako jedno slovníkové heslo. **V morfologickém slovníku se takto nepostupuje: členy vidové dvojice se považují za dvě různá slova.** Nově se zavádí, že kromě sloves se vid určuje u deverbativních přídavných jmen (těch, která mají SUB=V), např. *spící,* , *usnuvší,* i u deverbativních podstatných jmen, např. *dělání* (ASP=N), *vydělání* (ASP=D),

[66] V brněnském systému je vid kategorií **relevantní pouze pro slovesa**. Doplnit v rámci sjednocení se systémem NovaMorf kategorii vidu s příslušnou hodnotou i k substantivům a adjektivům, která jsou pravidelně generována od sloves, by patrně nemělo představovat vážný problém, protože jde o údaj doplnitelný z derivační historie lemmatu/tvaru.

[67] Brněnský systém nemá hodnotu pro obouvidová slovesa, viz Jakubíček a kol., 2011, s. 33. U těch sloves, která měla dříve aB, je nyní a[PI] a hodnota se desambiguuje. Návrh NovaMorf na rozdíl od brněnského systému a v souladu s pražským má pro obouvidová slovesa hodnotu ASP=O.

| | | | |
|---|---|---|---|
| **POS=V/k5** | **PER=[123-][68]** | **p[123]** | **[123-]** |

Návrh NovaMorf kopíruje do značné míry brněnský systém značkování slovesných subparadigmat (kategorie slovesného módu značkovaná atributem **m**). Odpovídající významy se v dosavadním pražském systému značkovaly na 2., 9. a 12. pozici. (Podrobné porovnání obou systému i s ohledem na to, jak se oba systémy zračí v tagsetu používaném k tagování Slovenského národního korpusu viz Osolsobě, 2007.) Korespondence mezi návrhem NovaMorf a brněnským systémem jsou patrné z tabulky 11.

**Tabulka 11**

| **slovesný tvar/mód** | NovaMorf | Brno | Praha | Praha | Praha |
|---|---|---|---|---|---|
| | **VRB** | **m** | **2. pozice** | **9. pozice** | **12. pozice** |
| **POS=V/k5** | **VRB=F** | **mF** | **f** | **-** | **-** |
| **POS=V/k5** | **VRB=P** | **mI** | **B** | **A** | **PX** |
| **POS=V/k5** | **VRB=I** | **mR** | **i** | **-** | **-** |
| **POS=V/k5** | **VRB=L** | **mA** | **pq** | **A** | **RX** |
| **POS=V/k[Y89][69]** | **VRB=K[70]** | **mC\|zY** | **c** | **-** | **-** |

[68] V návrhu NovaMorf se počítá s tím, že PER=2 budou mít i agregáty s volným morfem *s* za tvar auxiliáru/ spony slovesa *být*, a také osobní a přivlastňovací zájmena. Osobní zájmena mají v brněnském systému atribut p a hodnotu příslušné osoby vyznačenu. U tvarů, které mají zS v brněnském systému atribut p ve značce chybí. Vzhledem k tomu, že jde vždy o druhou osobu singuláru, by sjednocení nemělo představovat náročný problém. K řešení rozporů v kategorii čísla u agregátů (například .... <***kterýmis>*** *umyla tu hlavu, aby ses zbavila vší ...* , kde *kterými* je v plurálu a morf *-s* zastupuje slovesný tvar v singuláru), viz více v kapitole věnované agregátům.

[69] V brněnském systému byly dříve tvary *bych, bys, by, bychom, byste, abych, ..., kdybych, ...* značkovány jednotně, buď jako k5.*mC.* (*by*) nebo jako kY.*mC.* (*aby, kdyby*). V textu Jakubíček a kol., 2011 se uvádí, že tento stav má být zjednodušen, a to tak, že tvary *aby, kdyby*, ... budou značkovány jako podřadící spojky (k8zY) a tvary *by, ...* jako částice (k9zY). Ze specifikace atributem zY bude patrné, že jde o kondicionál. Na značkování kategorie osoby i čísla se rezignuje.

[70] VRB=K bude mít kondicionálový slovesný tvar *by* a rovněž kondicionálová složka slovesného agregátu *bych, bys, bychom, byste,* ale i další substandardní tvary *bysem, byjsem, bysi, byjsi, bysme, byjsme, byjste*. Ty budou značkovány jako POS=V & SUB=b. VRB=K budou mít dále spojky *aby*, *kdyby* a substandardní *dyby* a rovněž i) spojková složka spojkového agregátu *abych, abys, abychom, abyste* a také substandardní tvary *abysem, abyjsem, abysi, abyjsi, abysme, abyjsme, abyjste*; ii) spojková složka spojkového agregátu *kdybych, kdybys, kdybychom, kdybyste* a také substandardní tvary (*k*)*dybysem*, (*k*)*dybyjsem*, (*k*)*dybysi*, (*k*)*dybyjsi*, (*k*)*dybysme*, (*k*)*dybyjsme*, (*k*)*dybyjste*, *dybych, dybys, dybychom, dybyste*. Ty budou značkovány jako POS=J & SUB=,. Ke značkování kategorie osoby a čísla (u víceznačného tvaru .**by*) a k lemmatizaci a tokenizaci viz příslušné kapitoly.

| | | | | | |
|---|---|---|---|---|---|
| **POS=V/k5** | **VRB=p** | **mS** | **e** | **-** | **-** |
| **POS=V/k5** | **VRB=m** | **mD** | **m** | **-** | **-** |
| **POS=V/k5** | **VRB=B**[71] | **mB**[72] | **B** | **-** | **F** |
| **POS=A/k[25]**[73] | **VRB=T**[74] | **mN/---** | **s** | **P** | **H**[75] |

## Značkování kategorie negace v návrhu NovaMorf v porovnání s brněnským a pražským systémem

Značkování kategorie NEGACE v návrhu NovaMorf počítá se zachováním stávající praxe (lemmatizace tvaru derivovaného negativním prefixem *ne-* tvarem bez tohoto prefixu) pouze u sloves.[76] U adjektiv a adverbií i dalších slovních druhů[77] tuto praxi navrhujeme změnit.
Je-li POS=[AD] & lc=ne.*, přičemž počáteční řetězec *ne-* má negující význam, pak bude lemmatem tvar s *ne-* a hodnota NEG=N (*nedobrý, nerad, nešťastný, nepřítelův, nešťastně,*

[71] Tuto značku měly v pražském systému jednak tvary *budu, budeš, …*, jednak tvary **syntetického futura** sloves typu *jít, jet, …* (*pojedu,..., půjdu, …*). K doplnění slovníku i s ohledem na problémy s desambiguací viz Osolsobě, 2007 a také kapitola věnovaná slovesům.

[72] V brněnském systému mají tuto značku pouze tvary *budu, budeš, …* Na tvary syntetického futura nebere brněnský systém na rovině tagu zřetel (ve značce je uvedeno **mI**). Problém je (v obou systémech) nevyřešen u lemmatizace homonymních dvojic typu: *Pak otřu, <**pomažu**> olivovým olejem, … × Dáda se vzbudil a <**pomažeme**> pro Klárku, …*, v obou případech je uvedeno lemma *pomazat*, což je ve druhém případě lingvisticky neadekvátní. Správné lemma mají tvary slovesa *jít* (*půjdu*), *jet* (*pojedu*), *běžet* (*poběžím*), …, informace o tom, že jde o budoucí čas, lze získat právě kombinací lemmatu a značky (**[lc="po.*" & lemma!="po.*" & tag="k5.*mI.*"]**). Uvedeným dotazem lze z korpusu czTenTen12 získat seznam 13 lemmat. Máme za to, že v uvedeném korpusu výskyt slovesných tvarů syntetického futura představuje mnohem větší počet lemmat. Náš seznam totiž čítá přes 100 lemmat a byl získán z korpusových dat a z internetu v době, kdy velké webové korpusy ještě nebyly k dispozici. Naše analýzy ukazovaly, že tvoření futura tímto způsobem není v češtině ničím neobvyklým, o čemž svědčí nejrůznější aktualizace, jako např. *… potom rozhodnout jako s kerou to <**popečeš**>*. … ve významu „s kým se budeš spolčovat“ (doklad pochází z Brněnského mluveného korpusu **-bmk**).

[73] Podle stavu značkování tvarů participií n/t-ových v korpusu czTenTen12 je slovník, s nímž pracuje použitý tagger masivně přegenerovaný. Desiderata formulovaná v článku Jakubíček a kol., 2011 i v Hájková, 2014, nenašla kulantně řečeno adekvátní odezvu. Idea zahrnout tvary pasivních příčestí pod lemma adjektivizovaného tvaru a značkovat je jako slovní druh adjektiv má oporu i ve značkování v derivačním analyzátoru *Derivancze* (více Pala, Šmerk, 2015), v němž dlouhé tvary mají značku ***k2rpas*** a krátké/jmenné ***k2pas***.

[74] Hodnotu VRB=T mají krátké tvary pasivních příčestí, které podle návrhu NovaMorf jsou lemmatizovány v naprosté většině případů tvarem dlouhým (.*[nt]ý) a značkovány jako POS=A & SUB=V & NOM=J, viz více kapitola Adjektiva.

[75] Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

[76] Není to ovšem jediné možné řešení, neboť například Slovenský národní korpus (SNK) lemmatizuje tvary sloves s prefixem *ne-* tvarem negativního infinitivu. Toto řešení má dobrý důvod pro aplikace zaměřené na tvorbu slovníku, které využívají výsledků automatické morfologické analýzy (například Sketch Engine), neboť slovní profily dvojic lišících se v prefixu *ne-* bývají z hlediska lexikografického popisu odlišné (viz Benko, 2016). I tento fakt byl motivací námi navrhované změny. Z hlediska potřeb pravidlové desambiguace (obecnost pravidel) je ovšem dobré zachovat u sloves lemmatizaci negovaných tvarů lemmatem bez negujícího prefixu.

[77] Máme na mysli například deverbativní substantiva jako *nepřítel, neplavec, neřidič, neplatič, nemakač, …,*ale i další (spojení typu *hlava nehlava*). U zájmen, zájmenných adverbií a patrně i některých číslovek, která mají DEI=Z (záporné), je NEG=-, takže například *žádný* i *nižádný* se nebudou lišit značkou na rovině NEG. Naopak číslovky *nejeden, nemnoho* budou mít NEG=N.

*nepřátelsky, ...*). Nemá-li počáteční řetězec negující význam, pak bude pochopitelně lemmatem tvar s *ne-* a hodnota NEG=A (jednak adjektiva a adverbia typu *neurotický*, *neteřin, neurotizovaně, neurotizujícně, ...,* jednak **negativa tantum**, přičemž máme na zřeteli, že vymezení této kategorie výčtem na rovině morfologického slovníku není triviální[78]).
V brněnském systému je atribut **e[AN]** NEGACE relevantní pro adjektiva, slovesa a adverbia **k[256]**.
Domníváme se, že sjednocení nestojí v cestě významné překážky.

**Značkování kategorie stupeň v návrhu NovaMorf v porovnání s brněnským a pražským 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], viz podrobněji kapitoly věnované adjektivům a adverbiím.
Jedná se o změnu oproti dosavadní pražské praxi, která považovala některé typy adjektiv a adverbií za explicitně nestupňovatelné (tag=Db.*), jiné za explicitně stupňovatelné (tag=Dg.*).
Tato změna zcela odpovídá praxi brněnského systému, je jí inspirována.
Návrh NovaMorf navíc zavádí hodnotu **DEG=s** pro deriváty typu *sebe* + komparativ.
Takto tvořená slova jsou v brněnském systému dosud značkována atributem **d** s hodnotou **1**. V pražském systému je praxe nekonzistentní.
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 (viz více v příslušných kapitolách).
Domníváme se, že sjednocení nestojí v cestě významné překážky.

**Značkování kategorie Zkratka (ABR) v návrhu NovaMorf v porovnání s brněnským a pražským 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, viz více pasáž v kapitole Morfologické kategorie a jejich hodnoty.
V brněnském systému zkratky byly dříve značkovány na rovině slovního druhu jako **kA**, v návrhu Jakubíček a kol., 2011 je uvedena hodnota **A (zkratka)** u obecného subklasifikačního atributu **z**.
V pražském systému existují na 2. pozici značky pro hodnoty zkratka jako substantivum (;), adjektivum (.), číslovka (3), sloveso (~), adverbium (!).
Domníváme se, že sjednocení nestojí v cestě významné překážky.

**Subklasifikace vlastní brněnskému systému**

**Statistické charakteristiky**

[78] Podrobněji viz příslušné kapitoly věnované adjektivům a příslovcím.

Brněnský systém (Jakubíček a kol., 2011) pracuje s atributem **statistická charakteristika** ~ nabývajícím hodnoty frekvence na škále **0-9 (~[0123456789])**. Nakolik je toto značkování v praxi zahrnuto, není z dostupných publikací patrné.

**Subklasifikace interpunkce**
V novém revidovaném tagsetu (Jakubíček a kol., 2011) se uvádí seznam hodnot pro subklasifikaci **interpunkce (kI)**. Nakolik je toto značkování v praxi zahrnuto, není z dostupných publikací patrné.
I návrh NovaMorf bude mít vlastní klasifikační praxi pro interpunkci (je pojata do širší kategorie Symboly), viz více v příslušné kapitole.

**Subklasifikace vlastní pražskému systému: značkování posesivního rodu a čísla (6. a 7. pozice)**
Pražský systém zavádí ve značkování kategorii posesivní rod (6. pozice) a posesivní číslo (7. pozice). Je relevantní pro posesivní zájmena a adjektiva derivovaná sufixy *-ův, -in*. Označuje rod a číslo (ne u všech adjektiv a zájmen) osoby/osob, jíž/jimž se přivlastňuje. V případě adjektiv mají tedy všechna adjektiva na *-ův* vyplněnu hodnotu *mužský životný (M)* a všechna na *–in* hodnotu *ženský (F)*, hodnota posesivního čísla se nevyplňuje. U posesivních zájmen mají hodnotu rodu i čísla vyplněna pouze zájmena 3. osoby, přičemž se (z důvodů zjednodušení disambiguace) používá sdružených hodnot. Zájmena 1. a 2. osoby mají uvedenu pouze hodnotu čísla.
V návrhu NovaMorf se od značkování posesivního rodu a čísla ustupuje.

**Subklasifikace vlastní návrhu NovaMorf bez zřetelné opory v odpovídajících značkách v brněnském a pražském systému**

Na tomto místě shrnujeme to, co je uvedeno v kapitole Morfologické kategorie a jejich hodnoty.
Návrh NovaMorf pracuje s těmito kategoriemi:

1. Slovní druh – POS (viz výše)
2. Poddruh – SUB (viz výše)
3. Deixe – DEI (viz výše)
4. Vid – ASP (viz výše)
5. Zkratka – ABR (viz výše)
6. Rod – GEN (viz výše)
7. Číslo – NUM (viz výše)
8. Pád – CAS (viz výše)
9. Osoba – PER (viz výše)
10. Stupeň – DEG (viz výše)
11. Negace – NEG (viz výše)
12. Slovesný tvar – VRB (viz výše)
13. Jmenný tvar přídavných jmen – NOM (viz výše)
14. Typ agregátu – AGR (viz výše a rovněž v samostatné kapitole věnované agregátům)
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 mají jenom GMU, výjimkou jsou adverbia, neboť u nich se koncept FMU využívá u variantnosti ve stupňování).

V brněnském systému není žádná explicitní opora pro tagování vlastností, které má zohledňovat značkování GMU a FMU (viz podrobně Hlaváčová, 2009, nejnověji Hlaváčová, 2017). 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 automatické morfologické analýzy, protože se netýká interpretace na rovině morfologické analýzy, ale interpretace na rovině jiné (nemusí přitom jít jen o rovinu jazykové kultury).[79] Navržená klasifikace GMU a FMU neodpovídá tudíž hodnocení stylistických/stylových variant ve stávajícím pražském systému, přestože se tuto klasifikaci snaží nahradit.
Subklasifikace pomocí atributu **w (**stylistický příznak) prošla až do nového revidovaného tagsetu (Jakubíček a kol., 2011) beze změn, přestože omezenost takové klasifikace je všeobecně známa (viz Osolsobě, 2006, poznámka o převzetí hodnot atributu **w** ze SSJČ). Domníváme se ovšem, že v návrhu na jednotnou klasifikaci variant by bylo možné zohlednit některé postřehy obsažené v disertaci P. Šmerka (Šmerk, 2010), popřípadě značku ***var***, kterou disponuje derivační analyzátor ***Derivancze*** (více Pala, Šmerk, 2015, s. 519).
V případě sjednocování obou systému bude třeba postupovat ve vzájemné koordinaci, protože půjde o velmi složitý systém.[80]

[79] 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 metatextové užití). Domníváme se, že pro výzkum zaměřený na uvedené jevy v jazyce lze kombinovat 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.

[80] Motivací pro zavedení kategorie GMU a FMU je snaha o realizaci zlatého pravidla morfologie (Hlaváčová, 2009). Jde o pravidlo, které se týká speciálního případu víceznačnosti. Tento případ není problematický pro desambiguaci při analýze textu nástroji NLP. Problém desambiguace se totiž netýká jistého typu víceznačného přiřazení, které představuje problém zejména v aplikacích zaměřených na syntézu textu (např. strojový překlad). Jedná se o případy, kdy více různým slovním tvarům je přiřazena jedna a táž interpretace na rovině lemmatu a tagu. Například slovní tvary pro lexém s významem „malé okno“ v instrumentálu singuláru mohou být realizovány následujícími (celkem osmi) textovými slovy: ***okénky, okýnky, vokénky, vokýnky, okénkama, okýnkama, vokénkama, vokýnkama***. S využitím dosavadního pražského tagsetu (totéž platí i pro tagset brněnský) lze jednotlivé tvary popsat tagy lišícími se hodnotou na pozici 15 (varianta, stylový příznak), v brněnském systému je k dispozici atribut **w**. Projdeme-li ovšem nabízené hodnoty 15. pozice (popřípadě velmi podobnou nabídku u atributu **w** v brněnském systému**)**, zjistíme, že není možné odpovídajícím způsobem rozdíly uvedených tvarů **jednoznačně** popsat tak, aby se **každý z osmi tvarů lišil dvojicí lemma+tag**. Navíc neexistuje jednotná a dodržovaná instrukce, kterou by se řídila lemmatizace variantních tvarů odpovídajících jednomu lexému. Bez takové instrukce a bez rozumněji navrženého tagsetu jsou výsledky automatického zpracování jazyka nejednoznačné, což může vadit zejména při některých aplikacích využívajících automatickou syntézu. Např. při strojovém překladu se na základě jedné dvojice lemma+tag vytvoří více slovních tvarů a bez přesnějšího popisu, který by je rozlišoval, není možno stanovit kritéria pro správný výběr jednoho.
Požadavek jednoznačné interpretace lemmatu a tagu (pracovně nazvaný „zlaté pravidlo morfologie“) má být zajištěn právě zavedením kategorie GMU a FMU v návrhu NovaMorf.

## Závěr

Máme za to, že brněnský systém je v zásadě kompatibilní s návrhem NovaMorf. Rovněž pražský systém, který se (spolu s brněnským) stal výchozím bodem celého projektu NovaMorf, je s ním v souladu.

Tento text pokládáme za otevření diskuse o praktických řešeních, o kterých lze s ohledem na nástroje, které jsou na obou slovnících závislé, v budoucnosti rozumně uvažovat.

## Bibliografie


Benko V. (2016): *Tvorba webových korpusov a ich využitie v lexikografii*. Bratislava, FF UK. Disertační práce.

Hajič J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha, Karolinum.

Hajič J. – Hlaváčová J. (2016): *MorfFlex CZ*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

Hájková V. (2014): *Analýza jmenných tvarů adjektiv a pasivních příčestí ve slovníku morfologického analyzátoru* ajka. Brno, FF MU. (Nepublikovaná bakalářská práce pod vedením K. Osolsobě.)

Hladká Z. a kol. (2005): *Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS.* Brno, Masarykova univerzita. 68 s. Masarykova univerzita.

Hlaváčková D. – Sedláček R. (2006): Morfologické značkování korpusu soukromé korespondence. In *Varia XIV*. 1. vyd. Bratislava, Slovenská jazykovedná spoločnosť pri SAV, s. 371–379, 453 s.

Hlaváčková D. – Osolsobě K. (2008): Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. In: Kopřivová M., Waclawičová M. (eds.), *Čeština v mluveném korpusu*. 1. vyd. Praha, Nakladatelství Lidové noviny/ Ústav Českého národního korpusu, s. 105–114.

Hlaváčková D. (2013): Korpusové zpracování korespondenčních textů: morfologické značkování. In: Hladká Z. a kol. (eds.), *Soukromá korespondence jako lingvistický pramen*. Vyd. 1. Brno, Masarykova univerzita. s. 19–31.

Hlaváčová J. (2009): *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. Praha, UK. (Disertační práce.) Dostupná z: <http://utkl.ff.cuni.cz/phpBB3/viewtopic.php?f=11&t=1>

Hlaváčová J. (2017): Golden Rule of Morphology and Variants of Word forms. *Jazykovedný časopis*, 2017, roč. 68, č. 2, s. 136–144.

Hvězdová B. (1999): *Tvoření adverbií paradigmaticky odvozených od adjektiv na materiálu ČNK*. Brno, FF MU. (Diplomová práce pod vedením K. Osolsobě.)

Jakubíček M. – Kovář V. – Šmerk P. (2011): Czech Morphological Tagset Revisited. In: Horák A., Rychlý P. (eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing 2011*. Brno, Tribun EU, s. 29–42. (https://nlp.fi.muni.cz/raslan/raslan11.pdf)

Osolsobě K. (1996): *Algoritmický popis české formální morfologie a strojový slovník češtiny*. Brno, FF MU. Disertační práce.

Osolsobě K. (2006): Korpus soukromé korespondence (KSK) z hlediska morfologického značkování. *Linguistica Brunensia*, Brno, Masarykova Univerzita, A 54, č. 1, s. 187–201.

Osolsobě K. (2007): Popis gramatických významů (hodnot) jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů (Tagging of Verb Forms in Czech (Slovak) Corpora). *Linguistica Brunensia*, Brno, Masarykova Univerzita, A 55, No 1, s. 201–218.

Osolsobě K. (2007): Syntetické futurum v češtině – gramatiky, slovníky, korpusy. In: *Přednášky a besedy ze XL. běhu LŠSS*. 1. vyd. Brno, Masarykova univerzita, s. 131–144.

Osolsobě K. (2008): Značkování a status některých gramatických kategorií v ČNK (syntetické futurum, stupňování adjektiv, neurčité číslovky a příslovce míry). In: *Grammar & Corpora / Gramatika a korpus 2007*. 1. vyd. Praha, Academia, s. 407–416.

Osolsobě K. – Hlaváčová J. – Petkevič V. – Šimandl J. – Svášek M. (2017): Nová automatická morfologická analýza češtiny. *Naše řeč*, AV ČR, Ústav pro jazyk český, roč. 2017, č. 4, s. 225–234.

Pala K. – Šmerk P. (2015): Derivancze — Derivational Analyzer of Czech. In: Král P., Matoušek V. (eds.), *TSD 2015: Text, Speech, and Dialogue*. Berlin – Heidelberg, Springer Verlag, s. 515–523. Dostupný z: <https://link.springer.com/content/pdf/10.1007%2F978-3-319-24033-6_58.pdf>.

Pořízka P. – Schäfer M. (2009): MorphCon – A Software for Conversion of Czech Morphological Tagsets. In: Levická K., Garabík R. (eds.), *NLP, Corpus linguistics, Corpus Based Grammar Research*. Brno, Tribun, s. 292–301.

Šmerk P. (2010): *Towards Computational Morphological Analysis of Czech.* Brno, FF MU. Disertační práce.

Šmerk P. (2011): A New Data Format for Czech Morphological Analysis. In: Sojka P., Horák A. (eds.), *Proceedings of the Fourth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2010*. Brno, Tribun EU, s. 3–8. (https://nlp.fi.muni.cz/raslan/raslan10.pdf)