Gramatika a korpus II plin032

Středa: 7.30-9.00 G13

27. 4. Rozbor dú.

Číslovky jako slovní druh a jejich vyhledávání

Úvod

Automatická morfologická analýza sloužící ke značkování jazykových korpusů má jako každá interpretace jistá omezení. Z principu není úkolem automatické morfologické analýzy řešit otázku, jak, tedy kolika způsoby, a proč právě tak a tolika způsoby se interpretují gramatické významy analyzovaných jednotek. Autoři slovníků automatických morfologických analyzátorů přejímali výsledky práce teoretiků/ lingvistů, kteří zmíněné otázky prezentují v dílech (slovnících a gramatikách), z nichž tvůrci morfologických analyzátorů čerpali data pro strojové slovníky, s nimiž pracují nástroje automatické morfologické analýzy. Lingvista, který po více než 20 letech užívání automatických morfologických analyzátorů stojí před úkolem rekonstrukce slovníku automatického morfologického analyzátoru, by měl stanovit a následně dodržovat průhledná kritéria přiřazování lingvistických interpretací k jednotkám, které do slovníku byly/budou zařazeny, ať už je přejímá odjinud, nebo zavádí interpretace vlastní.

Slovnědruhová platnost číslovek a měrových adverbií

V příspěvku *Značkování a status některých gramatických kategorií v ČNK (syntetické futurum, stupňování adjektiv, neurčité číslovky a příslovce míry)* (Osolsobě 2008) jsme se zabývali slovnědruhovou interpretací a problémy lemmatizace tvarů měrových adverbií v platnosti číslovek, a sice tvary *mnoho, moc, hodně, (nej)víc(e).*

Následující tabulka ukazuje přehled interpretací tvarů *hodně, moc, mnoho, málo, víc(e), mén(ě)/míň* v českých výkladových slovnících (SSČ, SSJČ).

Tabulka 1

SSČ	SSJČ
hodně/přísl.	hodně/přísl.
mnoho/přísl./ ve spoj. s počít. předmětem v plat. čísl. neurč.	mnoho/přísl. a čísl. neurč.
moc/přísl. v plat. čísl. neurč. hovor.	moc/přísl. a čísl. neurč.
málo/přísl./ ve spoj. s počít. předmětem	málo/přísl. a čísl. neurč.

v plat. čísl. neurč.	
víc/více/přísl. v plat. čísl. neurč.	víc/více/přísl. a čísl. neurč.
nejvíc/nejvíce/přísl.	ODKAZ k více

Interpretace v SYN2010

1) Tvar *hodně* je interpretován vždy pouze jako stupňovatelné adverbium. Interpretace odpovídá interpretaci SSJČ, která ignoruje užití ve významu číslovky, viz *hodně/ moc/mnoho/ málo/ víc(e)/ mén(ě) lidí*.
2) Tvary *mnoho, moc* jsou interpretovány buď jako adverbium nestupňovatelné, nebo jako číslovka, včetně tvarů *mnoha.* Interpretace odpovídá SSJČ.[1]
3) Tvar *málo* interpretuje buď jako adverbium stupňovatelné, nebo jako číslovku (vč. tvaru *mála*). Iinterpretace odpovídá SSJČ. Lemma *málo* je interpretováno též jako substantivum.
4) Tvary *více/víc* i *méně/míň* interpretuje jako II. stupeň adverbií *hodně* a *málo,* nebo jako číslovku (v souladu se SSJČ). V případě interpretace jako číslovky je lemmatem tvar sám.
5) Tvary *nejvíce/nejvíc* i *nejméně/nejmíň* interpretuje výhradně jako III. stupeň adverbií *hodně* a *málo*.[2]

Interpretaci kategorie pádu u tvarů *mnoha/mála* ponecháme na tomto místě stranou, neboť se domníváme, že nejde o problém rozgenerování na rovině slovníku, nýbrž o problém desambiguace.

Následující obrázky nabízejí přehled lemmatizace a tagování jednotek s frekvencemi. V záhlaví uvádíme dotaz, jímž byla data získána z korpusu SYN2010.

Obrázek 3

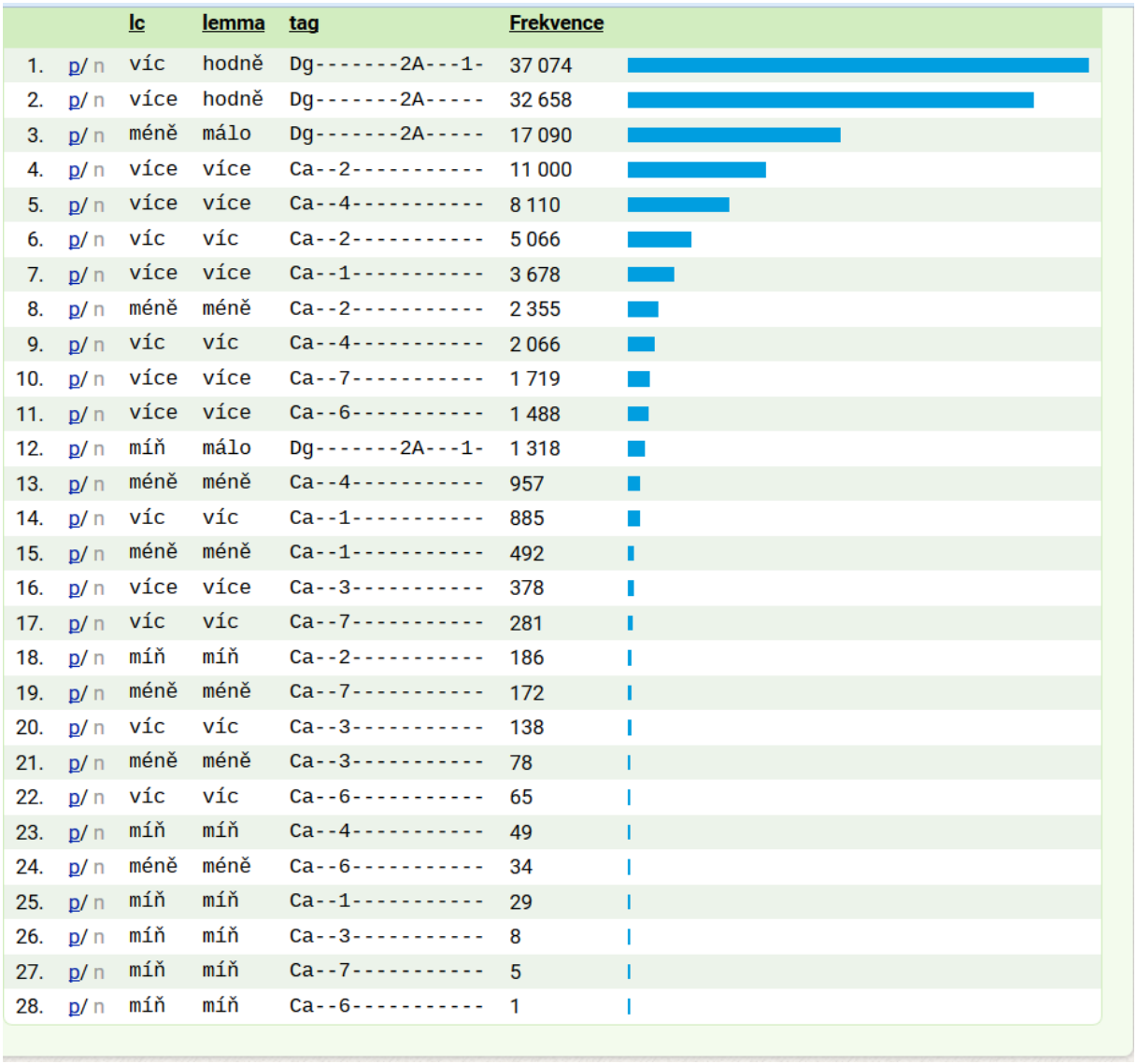
[lc="(hodně)|(mnoh[oa])|(mál[oa])|(moc)"]

[1] Interpretace jakožto nestupňovatelných adverbií řeší problém nemožné lemmatizace: Je lemmatem tvarů *víc(e)/nejvíc(e)* tvar *mnoho, moc* nebo *hodně*?
[2] K interpretacím v ‚*brněnském systému*' srov. Osolsobě 2008.

		lc	lemma	tag	Frekvence
1.	p/ n	moc	moc	Db----------	42 183
2.	p/ n	hodně	hodně	Dg-------1A-----	29 348
3.	p/ n	málo	málo	Dg-------1A-----	13 830
4.	p/ n	mnoho	mnoho	Ca--4----------	12 545
5.	p/ n	mnoho	mnoho	Ca--1----------	12 321
6.	p/ n	mnoha	mnoho	Ca--2----------	6 034
7.	p/ n	mnoha	mnoho	Ca--6----------	5 233
8.	p/ n	mnoha	mnoho	Ca--7----------	2 817
9.	p/ n	málo	málo	Ca--2---------1-	2 534
10.	p/ n	moc	moc	NNFS4-----A-----	2 374
11.	p/ n	moc	moc	Ca--4----------	1 833
12.	p/ n	mála	málo	Ca--2----------	1 226
13.	p/ n	moc	moc	NNFS1-----A-----	1 134
14.	p/ n	mnoha	mnoho	Ca--3----------	1 032
15.	p/ n	moc	moc	Ca--1----------	832
16.	p/ n	mála	málo	NNNS2-----A-----	314
17.	p/ n	málo	málo	Ca--7----------	289
18.	p/ n	moc	moc	Ca--3----------	254
19.	p/ n	málo	málo	Ca--6----------	199
20.	p/ n	málo	málo	Ca--3----------	126
21.	p/ n	mnoho	mnoho	Db----------	120
22.	p/ n	moc	moc	Ca--7----------	64
23.	p/ n	málo	málo	Ca--4----------	20
24.	p/ n	málo	málo	Ca--1----------	18
25.	p/ n	málo	málo	NNNS4-----A-----	9
26.	p/ n	moc	Moc	NNMS1-----A-----	8
27.	p/ n	moc	moc	Ca--6----------	6

Obrázek 4

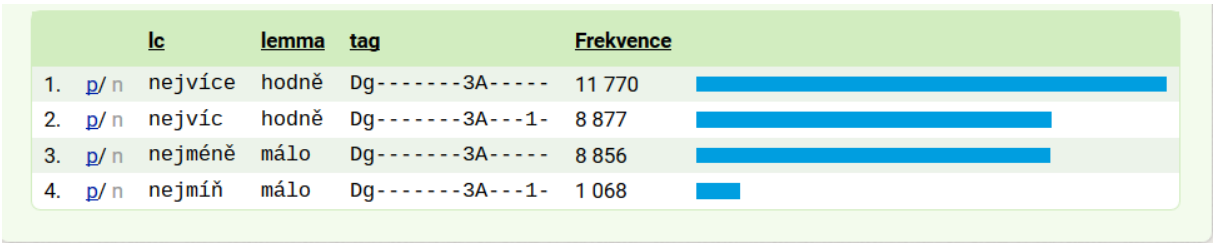
[lc="(více)|(víc)|(méně)|(míň)"]



		lc	lemma	tag	Frekvence
1.	p/ n	víc	hodně	Dg-------2A---1-	37 074
2.	p/ n	více	hodně	Dg-------2A-----	32 658
3.	p/ n	méně	málo	Dg-------2A-----	17 090
4.	p/ n	více	více	Ca--2----------	11 000
5.	p/ n	více	více	Ca--4----------	8 110
6.	p/ n	víc	víc	Ca--2----------	5 066
7.	p/ n	více	více	Ca--1----------	3 678
8.	p/ n	méně	méně	Ca--2----------	2 355
9.	p/ n	víc	víc	Ca--4----------	2 066
10.	p/ n	více	více	Ca--7----------	1 719
11.	p/ n	více	více	Ca--6----------	1 488
12.	p/ n	míň	málo	Dg-------2A---1-	1 318
13.	p/ n	méně	méně	Ca--4----------	957
14.	p/ n	víc	víc	Ca--1----------	885
15.	p/ n	méně	méně	Ca--1----------	492
16.	p/ n	více	více	Ca--3----------	378
17.	p/ n	víc	víc	Ca--7----------	281
18.	p/ n	míň	míň	Ca--2----------	186
19.	p/ n	méně	méně	Ca--7----------	172
20.	p/ n	víc	víc	Ca--3----------	138
21.	p/ n	méně	méně	Ca--3----------	78
22.	p/ n	víc	víc	Ca--6----------	65
23.	p/ n	míň	míň	Ca--4----------	49
24.	p/ n	méně	méně	Ca--6----------	34
25.	p/ n	míň	míň	Ca--1----------	29
26.	p/ n	míň	míň	Ca--3----------	8
27.	p/ n	míň	míň	Ca--7----------	5
28.	p/ n	míň	míň	Ca--6----------	1

Obrázek 5

[lc="(nejvíce)|(nejvíc)|(nejméně)|(nejmíň)"]



		lc	lemma	tag	Frekvence
1.	p/ n	nejvíce	hodně	Dg-------3A-----	11 770
2.	p/ n	nejvíc	hodně	Dg-------3A---1-	8 877
3.	p/ n	nejméně	málo	Dg-------3A-----	8 856
4.	p/ n	nejmíň	málo	Dg-------3A---1-	1 068

Kritika zvoleného řešení s ohledem na požadavek konzistentní automatické morfologické analýzy

Je naprosto zřejmé, že uvedená interpretace pražského systému vychází ze slovníku, který přebírá data ze SSJČ. V čem je hlavní problém takového přístupu? Tištěné slovníky jsou určeny uživatelům (většinově v případě češtiny rodilým mluvčím), kteří umějí s ambiguitními interpretacemi doplněnými ilustračními příklady úspěšně operovat. Přeneseme-li tyto ambiguity do morfologických slovníků užívaných k automatické morfologické analýze, nastává problém, jak se bude s interpretacemi nabízenými k desambiguaci dále pracovat. Jakákoliv desambiguace vyžaduje přesná pravidla. Ruční desambiguace dat by patrně nemusela činit vážné problémy. Není známo nic o tom, na jakém stupni vývoje je pravidlová desambiguace. Tuto otázku však můžeme pro naše potřeby ponechat stranou, protože se na této úrovni chceme zabývat výhradně slovníkem morfologického analyzátoru.

Pokud morfologický slovník zahrnuje více interpretací, pak by mělo jít pouze o interpretace, jejichž zjednoznačněním se do morfologického tagování vnese nějaká podstatná kvalita. Jinak je vícero interpretací pouze další obtíž kladená do cesty kvalitnímu značkování. Měli bychom si nejdříve položit otázku, zda je vůbec třeba rozlišovat dvojí slovnědruhovou platnost uvedených výrazů označujících množství.

Slovní druh číslovek je primárně vymezen na základě sémantického kritéria. Z tohoto hlediska představují výrazy označující neurčité (nepřeveditelné na výraz s významem určitého počtu) množství „šedou zónu“ s nesnadno vymezitelnými hranicemi. Vezmeme-li v úvahu slovnědruhové přesahy zájmen a zájmenných číslovek a příslovcí, pak bychom se nemuseli bránit myšlence zavedení slovnědruhové interpretace přesně ohraničeného slovnědruhového přesahu.

Návrh na zjednodušení a zprůhlednění dosavadní interpretace

Podíváme-li se na výše uvedené jednotky, tak pouze dvě z nich (*mnoho* a *málo*) rozlišují užití ve funkci číslovky formálně, a to nikoli tvary původní jmenné flexe (dochované v ustrnulých adverbializovaných tvarech *mnohem, namnoze)*, nýbrž paradigmatem složeným ze dvou tvarů, které odpovídá typu skloňování číslovek od 5 výše (více Komárek 2006). Proto je smysluplné pouze tvary *mnoho/mnoha* a *málo/mála* analyzovat jako tvary se značkou slovní druh číslovka, přičemž tvary *mnoho/málo* je třeba desambiguovat (mohou mít též platnost adverbia a odpovídající značku). Pokud bychom rezignovali na toto rozlišení, bylo by třeba rezignovat i na rozlišení dalších slovnědruhově závislých kategorií (tedy v tomto případě pádu).

Budeme-li chtít zachytit slovnědruhovou platnost číslovky u výrazu *moc*, pak je třeba zvážit, zda existuje nějaké zdůvodnění, proč tak činíme pouze u tohoto a žádného dalšího výrazu (např. *hodně*, ale třeba i dalších *?tuze*)? Budeme-li poctivě hledat odpověď, pak narazíme na celou řadu (patrně nějakou otevřenou množinu) výrazů, které ve stejných kontextech plní stejné funkce (např. *trochu* + sb. / verb.). Ostatně *moc* je od původu rovněž takovým substantivem! Jestliže žádné zdůvodnění pro výjimečné postavení adverbia *moc* neexistuje, je otázka, zda zachovat status quo, nebo jej měnit. S ohledem na fakt, že počet jednotek (adverbií, substantiv) s významem neurčitého množství tvoří patrně otevřenou množinu lemmat, pokládáme za vhodnější neomezovat další výzkum v této oblasti tím, že bychom interpretace ‚měrovosti' explicitně uváděli v morfologické značce. Pokud tak učiníme, je třeba opět explicitně v popisu morfologických značek, který má každý uživatel k dispozici, uvést, že se jedná o technické řešení problému a že seznam obsahuje jenom lemmata uvedená ve výčtu.

Jako nekonzistentní se totiž jeví také řešení víceré slovnědruhové interpretace tvarů *víc(e)* a *méně/míň*, nikoli však *nejvíc(e)* a *nejméně/nejmíň*. Chápeme-li tvoření tvarů komparativu a superlativu jako slovotvornou modifikaci, pak je přijetí lemmatizace tvarů komparativu a superlativu tvarem pozitivu pouze přijetím zavedené lexikografické konvence. Tuto konvenci porušuje lemmatizace u interpretace tvarů *víc(e)* a *méně/míň* jako číslovek. V obou případech jde o nepravidelné tvoření (od supletivních kořenů) a je pouze otázkou konvence, ke kterému z možných pozitivů měrových adverbií tvary interpretované jako adverbia budeme vztahovat (*hodně/ moc / mnoho / víc pracovat, málo/ trochu/ méně/míň pracovat*). Jde-li o pouhou konvenci, kterou v případě číslovkové platnosti porušujeme, je otázka, zda není více dobrých důvodů tuto konvenci za přesně definovaných podmínek porušit i u dalších jednotek s platností adverbií. V případě číslovkového užití je třeba sjednotit interpretaci tvarů komparativu a superlativu. Můžeme též uvažovat o řešení, které by tvary *(nej)víc(e)* a *(nej)mén(ě)/(nej)míň* interpretovalo pouze jako „příslovečné číslovky (nerozlišují žádné jmenné kategorie – tedy pád).[3]

[3] Technickým řešením je i lemmatizace a značkování (pos interpretace) tvarů *rád/(nej)raději* v současné podobě *‚pražského systému'*. V tomto případě jsme se přimlouvali za změnu stávajícího technického řešení.

Domácí úkol:

Viděli jsme, že pokud I. stupeň může být vyjádřen více než jedním tvarem (hodně/mnoho/moc – více – nejvíce), pak je desambiguace nemožná.

Opačný problém nastane v případě některých adjektiv od adverbií (konkrétně dřívější, hořejší, dolejší), k nimž existují tvary s prefixem nej-, nicméně je otázka, zda jde o II. stupeň, nebo o adjektivum I. stupeň.

1. Podívejte se na značkování a lemmatizaci těchto tvarů do korpusů.
2. Podívejte se do výkladových slovníků, jak jsou v nich tyto tvary charakterisovány.
3. Uvažujte o lepším taggování, než je stávající.