Nová korpusová mluvnice češtiny

Klára Osolsobě
osolsobe@phil.muni.cz

Abstrakt

V roce 2010 spatřila světlo bohemistického světa nová mluvnice češtiny, která „je prvním gramatickým popisem naší mateřštiny, který není založen jenom na povědomí autorů o jazyce kolem nás, ale na studiu rozsáhlých souborů reálných promluv a textů.“ Jde o publikaci Václava Cvrčka a kol., která nese název Mluvnice současné češtiny.
Tato kniha se řadí ke stručnějším (má 353 s.) přehledům české gramatiky, které byly vydány v posledních dvaceti letech. Vedle příruček jako jsou např. Čeština, řeč a jazyk a Příruční mluvnice češtiny nabízí uživatelům do jisté míry srovnatelné (nikoliv totožné) informace o češtině.
Cílem textu je krátké představení ideového základu korpusově založeného popisu jazyka, který Mluvnice současné češtiny slibuje nabídnout. Zaměříme se na stěžejní kapitoly věnované morfologii a tvoření slov. Upozorníme na pozitiva i negativa velkého počtu statistik zaměřeným na problémy spjaté s variantností flexe. Ve druhé části se zaměříme na vybraná sporná místa, a to především na případy, kdy korpusy nebyly dostatečně využity. Našim cílem je demonstrovat, jak lze použít korpusy i korpusové nástroje k tomu, abychom odpověděli na otázky, které při četbě MSČ vyvstanou a jejichž zodpovězení zůstali autoři MSČ čtenářům dlužni. V korpusech lze řadu takových odpovědí najít a není to někdy tak úplně složité.

Úvod

Korpus jako zdroj observací fungování jazyka je možné zkoumat a) objektivně měřitelnými metodami a b) opakovaně, tudíž s možnou kontrolou/zpětným ověřením různých tvrzení. Podíváme se, jak výzkum založený na korpusech ČNK, a sice SYN2005, ORAL2006, PMP, BMK (srv. více http://ucnk.ff.cuni.cz) prezentovaný MSČ přispívá k obrazu mateřštiny. Zaměříme se na kapitoly věnované Morfologii (autor Václav Cvrček) a Tvoření slov (autor Michal Šulc).

Statistiky v MSČ pozitiva a problémy

Již v Předmluvě se autoři hlásí k tomu, že „ ... se nesnaží popisovat jazyk, jak by měl vypadat, ale jak skutečně vypadá“. Prostor interpretativní složky popisu je tak omezen na výčty řazené podle frekvence a na zprostředkování statistik.
V kap. 7. Morfologie se uvádí velké množství statistik založených na příslušných korpusech. (statistiky jednotlivých slovních druhů, flektivních typů, variantních koncovek). Obdobné (nikoliv totožné) statistiky pro češtinu sice k dispozici jsou, vycházely ovšem z nesrovnatelně menších dat (korpusy, s nimiž pracovala M. Těšitelová aj.) a od doby jejich vzniku/publikace nás dělí více než čtvrt století vývoje češtiny. Pozitivně lze hodnotit fakt, že statistiky neuvádějí pouhá absolutní čísla, ale šestistupňovou škálu, jejíž pomocí lze docílit objektivního srovnání dat získaných z různých korpusů. Správně se opakovaně poukazuje na závislost výsledků statistik týkajících se flektivních vlastností pojatých obecně (tedy jednotlivých flektivních typů) na povaze jednotlivých lexémů.
V propagačně zaměřených částech MSČ se tvrdí, že tato mluvnice jako první vychází ze studia rozsáhlých souborů reálných promluv a textů, avšak v případě mluvených komunikátů se vyznačuje naprostou nereprezentativností, což pak vede k jistým zkresleným tvrzením týkajícím se zejména jazykové situace na Moravě a ve Slezsku.

Zobecnění pozorování masových dat jako cíl korpusového výzkumu jazyka

Statistické údaje ovšem mohou a měly by pomoci k formulaci zobecnitelných závěrů. Vezmeme-li v úvahu, že korpusy (alespoň ty psané) představují dosud nevídanou základnu pro takováto zobecnění, podívejme se, jak byly autory MSČ k tomuto účelu využity.
Jako příklad poslouží srovnání téměř doslovně se opakujících vágních tvrzení týkajících se distribuce *-e/-ě* v koncovkách české substantivní flexe.
Na s. 174 se uvádí, že „Ke vzoru *duše* patří feminina s koncovkou *-e* někdy psanou *-ě* ...“, takřka stejná formulace se objeví na s. 188 a 189 (vzor *moře* a *kuře*). V kapitolách věnovaných vzoru *soudce* (s. 160n.) a *píseň* (s. 178n.) je z textu patrné, že i u těchto vzorů se vyskytuje dvojí možná grafická realizace *e/ě*.
Naše otázka podnícená mimo jiné výše uvedenými vágními formulacemi zní: Je psaní *-e/-ě* ve flektivních koncovkách popsatelné obecně platnými pravidly? Tuto otázku chci v rámci přednášky věnované zahraničním studentům češtiny položit ze dvou důvodů: 1) studenti bohemistiky (rodilí mluvčí) na ni odpověď hledali s jistými obtížemi (to může, i když ne nutně, svědčit o tom, že jde o složitý problém) a 2) odpověď na tuto otázku komplikuje (protože jde opravdu o odpověď komplikovanou) řešení některých oblastí počítačového zpracování přirozeného jazyka (konkrétně češtiny), což je oblast, která nás dlouhodobě odborně zajímá.
Korpusový lingvista by měl hledat odpověď na otázky v korpusech. Podívejme se, jak lze postupovat.
V prvním kroku můžeme vyhledat všechna substantiva taková, že končí na *-e* ne na *-ě*.
Dále můžeme vytvořit a prohlížet frekvenční seznam nalezených tvarů. Uvádíme pouze jeho část.

```
word: ##
roce        84640
době        49588
práce       44324
případě     39985
Praze       32046
země        30660
straně      25232
peníze      25080
situace     23352
světě       22571
informace   20872
místě       20286
konce       19190
dne         18580
policie     18445
komise      15731
základě     15563
ruce        15369
Evropě      14459
unie        14077
organizace  13761
republice   13752
soutěže     12733
funkce      12476
akce        12456
městě       12243
dítě        11534
ředitele    11029
muže        10972
```

```
životě      10831
televize    10727
měsíce      10519
nemocnice   10276
Brně        10205
```

Výsledkem tohoto pozorování může být hypotéza, že distribuce *-e/-ě* je vázána na předchozí grafém, přičemž můžeme vidět, že v naprosté většině případů jde o konsonant. Další postup může být takový, že se podíváme na možné kombinace jednotlivých souhláskových grafémů následovaných -e/-ě.

Výsledky shrneme do následující tabulky

	celkem (-e/-ě) lemmat	lemmat s tvary –e	lemmat s tvary -ě
.***b[eě]**	311	81	234
.***c[eě]**	8068	8068	0
.***č[eě]**	1195	1195	0
.***d[eě]**	707	231	497
.***ď[eě]**	0	0	0
.***f[eě]**	56	34	22
.***g[eě]**	34	34	0
.***h[eě]+ch[eě]**	88	88	0
.***j[eě]**	382	382	0
.***k[eě]**	35	35	0
.***l[eě]**	1634	1634	0
.***m[eě]**	274	140	140
.***n[eě]**	2809	729	2108
.***ň[eě]**	0	0	0
.***p[eě]**	140	74	66
.***r[eě]**	400	400	0
.***ř[eě]**	1230	1230	0
.***s[eě]**	1177	1177	0
.***š[eě]**	486	486	0
.***t[eě]**	1514	483	1056
.***ť[eě]**	0	0	0
.***v[eě]**	906	127	792
.***z[eě]**	761	761	0
.***ž[eě]**	214	214	0

Podíváme-li se na výsledky v předchozí tabulce, můžeme tvrdit, že :

1. Existují grafémy, za kterými se v češtině nepíše v koncovkách (zakončeních) substantiv ani *-e*, ani *-ě*. Jsou jimi *ď, ť, ň*.
2. Existují grafémy, za kterými se v češtině píše v koncovkách (zakončeních) substantiv vždy pouze *-e*. Jsou jimi *c, č, g, h, j, k, l, r, ř, s, š, z, ž*.
3. Existují grafémy, za kterými se v češtině píše v koncovkách (zakončeních) substantiv buď *-e* nebo *-ě*. Jsou jimi *b, d, f, m, n, p, t, v.*
4. Existují grafémy, za kterými se v češtině píše v koncovkách (zakončeních) substantiv buď *-e* nebo *-ě,* a to u téhož lemmatu. Plyne to z toho, že počet všech lemmat není vždy totožný se součtem lemmat, u nichž je buď jedna, nebo druhá varianta. Dle sledovaného korpusu jsou jimi *b, d, m, n, t, v.*

V dalším kroku si tedy budeme všímat pouze lemmat, jejichž tvary končí na -*e*, nebo -*ě*, před nimiž předchází [bdfmnptv]. Zopakujeme výše uvedený postup a vyhledáme v korpusu všechna substantiva, která končí na [bdfmnptv][eě]. Podívejme se alespoň na ta nejfrekventovanější.

word:	lemma:	##
době	doba	49588
případě	případ	39985
země	země	30660
straně	strana	25232
světě	svět	22571
místě	místo	20286
dne	den	18580
základě	základ	15563
Evropě	Evropa	14459
městě	město	12243
dítě	dítě	11534
životě	život	10831
Brně	Brno	10205
řadě	řada	9743
polovině	polovina	9233
cestě	cesta	8894
podstatě	podstata	8862
podobě	podoba	8740
sítě	síť	8504
vládě	vláda	8382
pane	pan	8194
daně	daň	7824
domě	dům	7520
týdne	týden	7497
skupině	skupina	5635
létě	léto	5423
minutě	minuta	5386
hodnotě	hodnota	5273
zbraně	zbraň	4953
Ostravě	Ostrava	4739
formě	forma	4680
většině	většina	4637
koně	kůň	4618
Moravě	Morava	4569
Bosně	Bosna	4565
hlavě	hlava	4534
Prostějově	Prostějov	4441
změně	změna	4424
firmě	firma	4334
půdě	půda	4283
církve	církev	4261
vodě	voda	4255
rodině	rodina	4230
úrovně	úroveň	4123
Země	země	4027
Moskvě	Moskva	3906
přípravě	příprava	3855
výrobě	výroba	3843
dítěte	dítě	3760
ceně	cena	3705
krve	krev	3679
návštěvě	návštěva	3665
scéně	scéna	3633
letiště	letiště	3490

```
závodě      závod       3463
Pane        Pan         3463
bytě        byt         3437
třídě       třída       3426
dohodě      dohoda      3404
přírodě     příroda     3389
```

Na základě pozorování dat můžeme říci, že ačkoliv se v uvedeném seznamu vyskytují substantiva většiny vzorů (*doba/žena, případ/hrad, země/růže, místo/město, dítě/kuře, daň/píseň, pan/pán, kůň/muž, letiště/moře* ...), v MSČ se příslušné vágní formulace stran distribuce grafému –e/-ě týkaly pouze vzorů *duše, moře, kuře, soudce* a *píseň.* Zdá se tudíž, že bychom případné obtíže měli hledat právě u těchto vzorů. Jak lze dále postupovat. Můžeme zjistit, která slova z výše uvedeného seznamu patří k uvedeným vzorům. V následující tabulce uvedeme příklady založené na korpusovém šetření.

	soudce	duše	píseň	moře	kuře
b[eě]	Vosolsobě	0	0	nebe	hrabě
d[eě]	-	hýždě	lodě	?rande	hádě
f[eě]	-	0	0	kafe	0
m[eě]	-	země	země	sémě	0
n[eě]	Bechyně	kuchyně	daně	poledne	štěně
p[eě]	-	koupě	0	kanape	doupě
t[eě]	-	kleště	sítě	letiště/?karate	dítě
v[eě]	-	0	církve	0	0

Na jeho základě můžeme formulovat následující tvrzení:

1) Substantiva skloňovaná podle vzorů *soudce, růže, kuře* mají (na základě korpusových dokladů) po grafémech [bd(f)mnpt(v)] koncovku *-e* vždy realizovanou jako grafické *-ě*.
2) Substantiva skloňovaná podle vzoru *píseň* mají (na základě korpusových dokladů) po grafémech [dnt] koncovku *-e* vždy realizovanou jako grafické *-ě*.
3) Substantiva skloňovaná podle vzoru *moře* mají (na základě korpusových dokladů) po grafému [t] koncovku *-e* vždy realizovanou jako grafické *-ě*, přičemž jde vždy o sufix *-iště*.

V dalším kroku se tedy budeme zabývat jednak substantivy skloňovanými podle vzoru *píseň*, která končí na [bfmpv], jednak substantivy skloňovanými podle vzoru *moře*, která končí na [bfmpvdnt]. Z korpusu získáme jejich seznamy.

```
lemma:      ##
církev      4565
krev        3707
láhev       1268
větev       1237
lahev       504
rakev       464
pánev       463
mrkev       277
ploutev     192
koroptev    154
broskev     150
konev       93
tykev       85
podešev     43
brukev      42
krokev      39
```

```
korouhev    33
ředkev      28
plástev     23
Cerekev     20
podoustev   8
vikev       8
štoudev     7
Chrudim     6
Ponikev     6
euroláhev   3
houžev      3
hnědozem    2
dratev      2
Vlašim      2
Býkev       2
Hořátev     1
šedozem     1
pseudocírkev      1

lemma:      ##
nebe        3675
poledne     2195
odpoledne   1811
Labe        1073
kafe        690
dopoledne   612
rande       397
kanape      104
sémě        48
plémě       32
símě        22
```

Na základě výše uvedených dat můžeme říci, že:

1. Ke vzoru *píseň* patří skupina substantiv zakončených na *-ev*, u nichž se koncovka *-ě* dy realizuje jako grafické *e*.
2. Ke vzoru *píseň* patří několik málo substantiv zakončených na *-m* u nichž se koncovka *-e* vždy realizuje jako grafické *ě*.
3. Substantiva zakončená na [bfmpvdnt] patřící ke vzoru *moře* mají s výjimkou derivátů na *-ště* a skupiny substantiv *sémě, plémě, símě* koncovku *-e* realizovanou jako grafické *-e*.
4. Jde o poměrně malý počet substantiv. Nicméně se většinou jedná o substantiva poměrně frekventovaná.
5. Můžeme je tudíž definovat výčtem, přičemž s ohledem na rozsah korpusu můžeme předpokládat relativní úplnost výčtu frekventovaných jednotek.
6. Vzhledem k tomu, že distribuce variant je alespoň u vzorů *píseň* a *moře* vázána nikoliv na distribuci danou grafickým okolím, ale na jednotlivé skupiny lexému, je třeba připustit, že v češtině existují u některých vzorů dvě varianty koncovek *-e/-ě* a že tyto varianty nejsou grafickými variantami v témže smyslu, jako jsou jimi varianty *-e/-ě* u jiných vzorů.

Závěr

Cílem textu je prakticky ukázat, že ačkoliv nová korpusová mluvnice češtiny nezahrnula řadu informací, které lze z korpusů vyčíst, není jejím vydáním možnost využívat korpusy i nadále jako zdroje observací jazyka nikterak potlačena.

Na základě pozorování dat získaných z korpusů je možné odpovídat na otázky, které před zvídavými čtenáři MSČ mohou vyvstat. Dopátrat se žádoucích odpovědí není vždy snadné, je ovšem třeba si uvědomit, že nejsou-li útrapy cesty cílem, mohou být jeho součástí. A tak jako se cestou zejména díky překonávání společných překážek dozvídáme hodně o tom, s kým cestujeme, tak se i cestou korpusového výzkumu lze dozvědět hodně o jazyce, kterým se „probíjíme".
Chtěla bych všem studentům češtiny, které korpusové cestování po češtině s češtinou neodradilo, popřát šťastnou cestu a hodně trpělivosti při překonávání překážek. A hlavně to, aby si nakonec řekli, že čeština za to stojí!

Literatura

Cvrček, V. a kol: Mluvnice současné češtiny 1 – Jak se píše a jak se mluví. Praha: Karolinum, 2010.
Čermák, F.; Schmidtová, V.: Český národní korpus – základní charakteristika a širší souvislosti. Národní knihovna, 15, 2004, č. 3, s. 152-168.
Osolsobě, K.: Recenze: František Čermák – Renata Blatná: Jak využívat Český národní korpus. Studijní příručka. Sas 68, 2007, s. 147-151.
Osolsobě, K.: Syntetické futurum v češtině – gramatiky, slovníky, korpusy, In: Přednášky a besedy z XL. běhu LŠSS, Brno 2007, s. 131-144.
Osolsobě, K.: Značkování gramatických kategorií v korpusech ČNK a jejich zachycení v gramatice a ve slovníku (syntetické futurum, stupňování adjektiv, neurčité číslovky a příslovce míry). In Štícha, F.: Grammar & Corpora / Gramatika a korpus 2007. Academia: Praha, 2008, s. 407-416.
Kosek, P., Křístek, M., Osolsobě, K., Vojtová, J., Ziková, M.: První korpusová mluvnice češtiny:
Václav Cvrček a kolektiv autorů: Mluvnice současné češtiny 1 – Jak se píše a jak se mluví. Praha: Karolinum, 2010. 354 s., Naše řeč 2/94, 2011, s. 149-160.
Slovník spisovného jazyka českého (SSJČ), Praha 1958-1971, reprint 1989.
Slovník spisovné češtiny pro školu a veřejnost (SSČ), Praha 1978[1], 1994[2].

Odkazy na „online" zdroje

Český národní korpus - SYN2000/SYN2005/SYN2006PUB/SYN. Ústav Českého národního korpusu FF UK, Praha 2000. K vyhledání on-line <http://ucnk.ff.cuni.cz>. (http://ucnk.ff.cuni.cz/bonito/)
Rychlý, P.: Bonito – grafické uživatelské rozhraní systému Manatee, Verze 1.49. 1998-2003. K vyhledání on-line http://ucnk.ff.cuni.cz/bonito/