Využití velkých korpusů pro morfematickou analýzu českých slovesných předpon

Jaroslava Hlaváčová

Karlova Univerzita, Matematicko-fyzikální fakulta, Praha
hlavacova@ufal.mff.cuni.cz

Abstrakt

Prefixace je pro češtinu jedním ze základních slovotvorných prostředků. Zejména pro slovesa představuje velmi produktivní způsob, jakým lze modifikovat jejich význam a vid. V současném příspěvku ukážeme, jak lze korpusová data použít k analýze vícenásobných předpon vyskytujících se u českých sloves. Ve druhé části se zabýváme předponami, které se používají k modifikaci sloves přejatých z cizích jazyků.
Klíčová slova: slovesná předpona; česká morfologie; prefixace

1 Motivace

V textu budeme prezentovat výsledky první fáze projektu, který má za cíl celkovou morfematickou analýzu češtiny. Zatím jsme se zabývali především slovesy a jejich předponami. Analýzy jsme prováděli na základě velkých českých korpusů, především korpusu Omnia Bohemica II (Benko 2014), a korpusů řady SYN (SYN).

Prvotním východiskem byl však Retrográdní morfematický slovník češtiny (Slavíčková 1975). V dalším textu se na něj odkazujeme jako na morfematický slovník.

Zjišťovali jsme, jaké předpony se mohou kombinovat a v jakém pořadí. České sloveso může mít až 4 předpony (*pře-u-s-po-řádat*), ale existují pravidla pro jejich řetězení - zdaleka ne všechny kombinace jsou možné.

Dobrá znalost kombinovatelnosti předpon má velký potenciál pro zlepšení odhadování morfologických vlastností slov, která nejsou (dosud) zachycena v morfologických slovnících, tedy především slov nových a příležitostných.

2 Slovesné předpony v češtině[1]

České slovesné předpony lze třídit z mnoha hledisek. Pro účely morfematické analýzy jsme dospěli k následujícímu rozdělení:

- „základní“ slovesné předpony
- dlouhé předpony
- speciální předpony
- předpona *ne-*

[1] V odborné literatuře se za slovesné předpony považují občas jen předpony, které uvádíme dále jako „základní“ slovesné předpony, viz např. (Uher 1987). Zde pod termínem slovesné předpony máme na mysli obecně předpony, které se vyskytují u sloves.

- víceslabičné předpony
- cizí předpony

Jednotlivé skupiny nemusí být nutně disjunktní - některé předpony patří do více skupin. Příklad viz dále.

2.1 Základní předpony

Do první skupiny patří 20 předpon: *vy-, za-, z-, po-, roz-, na-, u-, o-, s-, od-, pro-, pře-, při-, do-, v-, pod-, ob-, vz-, před-, nad-*, přičemž všechny, které končí na souhlásku, mohou mít ještě vokalizovanou variantu (např. *ve-, roze-*), a předpony začínající samohláskou *o-* mohou mít v hovorové češtině protetické *v* (*vo-, vod-, vode-, vob-, vobe-*). Vokalizované varianty předpon, které se tím mohou stát dvojslabičnými, však nepočítáme do skupiny víceslabičných předpon (skupina 5).

2.2 Dlouhé předpony

Druhá skupina obsahuje předpony *ná-, zá-, pů-, dů-, prů-, pří-, vý-, zů-, ú-*. Většina sloves začínajících dlouhou předponou je odvozená z podstatného nebo přídavného jména, např. *zálohovat* od *záloha*, *úředničit* od *úředník*, *průsvitnět* od *průsvitný*. Existuje jen několik málo takových sloves, která jsou „původní“ (např. *následovat*, *důvěřovat*, *zůstávat*). Předpona *zů-* by se mohla řadit i do třetí skupiny předpon speciálních, protože se může vázat pouze s jediným kmenem, a to *stav/stáv* (*zůstat*, *zůstávat* a několik dalších odvozenin).

2.3 Speciální předpony

Třetí skupina obsahuje předpony, které se mohou připojit pouze k omezené množině kořenů. Jsou to:

bez-/beze- s kořenem *peč*, většinou vystupující v kombinaci s další předponou (*za-bez-pečit*)

(v)ot-/(v)ote- s kořenem *vř/vír/víř* ((*v)o-tevřít, (v)ot(e)-vírat*)

pa- s kořeny *děl* a *běr* (*pa-dělat*, *pa-běrkovat*)

zů- s kořenem *stav/stáv* - viz též výše.

2.4 Předpona *ne-*

Předpona *ne-* má funkci negace a až na pár výjimek (např. *ne-návidět*) se nepovažuje za „povinnou“ součást slovesa. Jinak tomu je ovšem v případech, kdy se předpona *ne-* vyskytuje ve slovese jako druhá, např. ve slovesech *z-ne-pokojit*, *z-ne-hybnit, roz-ne-moci*.

2.5 Víceslabičné předpony

Prefixaci pomocí víceslabičných předpon už lze považovat za jiný typ odvozování, totiž skládáním. Týká se předpon *polo-, samo-, spolu-, znovu-, sebe-*. Někteří lingvisté je neuznávají za typické předpony. Na rozdíl od dosud uvedených předpon (s výjimkou předpony *ne-*) mají velmi vyhraněný význam, kterým mohou modifikovat význam velkého množství sloves a obecně i slova jiného slovního druhu. Stačí, aby jejich význam nebyl v protikladu k významu původního slovesa. Více se těmito předponami nyní zabývat nebudeme.

Mezi víceslabičné předpony nepočítáme vokalizované varianty předpon zařazených do ostatních skupin (tedy ne *beze-, roze-* ani další).

2.6 Cizí předpony

Mezi cizí předpony patří *re-, de-, a-, in-, im-, ab-, ante-, per-. meta-* a řada dalších. Vyskytují se výhradně ve slovesech, která jsou přímo odvozena z cizího slova, např. *reklamovat*, *superponovat*, mají tedy i cizí kmeny (Esvan 2007). Tyto předpony jsou součástí cizích slov a pro naše současné potřeby je považujeme za součást kořene. Jinými slovy se nezabýváme morfematickou analýzou cizích (přejatých) slov.

3 Segmentace sloves

Naším cílem bylo rozdělit každé zadané sloveso na prefixové segmenty a zbytek (*po-vy-skočit*).

Slovesa jsme získali z morfologicky označkovaného korpusu SYN2015. Soubor všech sloves měl 20 553 položek.

Algoritmus byl jednoduchý - testovali jsme, zda se na začátku slovesa vyskytuje řetězec, který je shodný s některou předponou. Seznam předpon byl předem daný - využili jsme data z morfematického slovníku, kde je seznam předpon uvedený v dodatcích. V první fázi jsme pracovali pouze s českými předponami (kategorie 1 - 4), včetně jejich variant s protetickým *v* a vokalizovaných předpon (viz sekci 2).

Pokud se začátek slovesa shodoval s některou předponou ze seznamu, odtrhli jsme ho a rekurzivně jsme pokračovali v odtrhávání potenciálních předpon tak dlouho, dokud bylo co odtrhávat. Zbytek slovesa jsme nazvali pahýl. K tomu, abychom odtržené řetězce mohli prohlásit za skutečné předpony, je nutné, aby pahýl byl buď samostatným slovesem (např. pro segmentaci slovesa *při-jet*), nebo patřil do seznamu slovesných pahýlů (*při-cházet*). Pokud pahýl tyto podmínky nesplňuje, byla segmentace špatná. Např. ne všechna *s* na začátku slovesa jsou předponami: segmentace *s-kočit* není dobře, neboť pahýl *kočit* není ani slovesem, ani slovesným pahýlem.

První alternativu jsme testovali pomocí morfologického analyzátoru MorphoDita (MorphoDita, Straková et al. 2014). Jestliže morfologická analýza rozpoznala pahýl jako sloveso (jinými slovy jestliže pahýl je jako sloveso součástí morfologického slovníku), byla segmentace slovesa pravděpodobně správná.

Seznam slovesných pahýlů jsme získali opět z morfematického slovníku (Slavíčková 1975) pomocí jednoduchých pravidel nad segmentovanými slovesy.

Postupné odtrhávání předpon může vést k vícenásobným analýzám, z nichž jen některé mohou být správné. Např. pro sloveso *podotýkat* jsme dostali následující automatické segmentace, přičemž pouze ta druhá je správná:

pod - o - týkat

po - do - týkat

Posloupnost předpon *pod-* a *o-* v tomto pořadí se totiž u českých sloves nevyskytuje.

Poměrně velké množství takových vícenásobných segmentací nás vedlo ke zkoumání

kombinovatelnosti předpon. Ukázalo se, že některé předpony se mohou řetězit, zatímco jiné ne.

4 Vícenásobné předpony českých sloves

Zaměřili jsme se tedy na ta slovesa, která mají alespoň dvě předpony, a z těchto dvojic jsme vytvořili tabulku jejich možných kombinací - viz tabulka 1. Základem byla výše zmíněná analýza sloves z korpusu SYN2015, ale dohledávali jsme příklady i ve větším korpusu Omnia Bohemica II i přímo pomocí vyhledávače Google na internetu.

Sloupce i řádky tabulky obsahují typické slovesné předpony, dlouhé předpony, speciální předpony a předponu *ne-* (tedy kategorie 1 až 4 z rozdělení uvedeném v sekci 2 o slovesných předponách). Údaje v řádcích představují předpony vyskytující se jako první, předpony ve sloupcích vystupují jako druhé. Existenci slovesa s danou dvojicí předpon znázorňuje znak + v příslušném políčku.

Přestože řádky i sloupce obsahují stejné množiny předpon, tabulka není symetrická, neboť pochopitelně záleží na pořadí předpon. Tak např. dvojice předpon *po-ob-* má v příslušné buňce +, neboť existuje sloveso s touto posloupností předpon (*po-ob-veselit*), opačné pořadí *ob-po-* má ale buňku prázdnou, neboť sloveso s touto posloupností předpon jsme nenalezli.

Pro větší názornost jsme navíc tabulku přeuspořádali podle četnosti možných kombinací. V horní části tabulky se nacházejí předpony, které se vyskytují jako první v nejvíce kombinacích. Součet kombinací je uveden v posledním sloupci. Podobně v levé části tabulky (ve sloupcích nejvíce nalevo) jsou předpony, které se nejčastěji vyskytují v kombinacích jako druhé, součty jsou v posledním řádku.

Z Tabulky 1 je možno vyčíst několik zajímavých pozorování.

4.1 Samostatné předpony

Spodní část tabulky s prázdnými buňkami se týká předpon, za kterými už žádná další slovesná předpona stát nemůže. Jsou to předpony z kategorie 2 a 3, tedy dlouhé a speciální předpony. Ze základních předpon (kategorie 1) sem spadá pouze předpona *ob-*.

Výsledek je u katergorií 2 a 3 očekávatelný a v souladu s charakteristikami uvedenými výše. Dlouhé předpony jsou u sloves spíše výjimečné. Ty, které se vážou k několika starým původním slovesům, nemohou za sebe "pustit" žádnou další předponu. Z faktu, že ani slovesa vzniklá z podstatných (*úředničit*) nebo přídavných jmen (*výtvarničit*), nemají nikdy za dlouhou předponou žádnou další, by mohlo vyplývat, že ani podstatná a přídavná jména uvozená dlouhou předponou nemohou být vícepředponová. Tuto hypotézu bude třeba ověřit.

Přítomnost speciálních předpon v této skupině také není překvapující. Z toho, že se tyto předpony vážou jen s několika málo kořeny, přímo vyplývá, že nebudou příliš "ochotné" se připojovat k dalším předponám.

	po	s	z	vy	na	u	ná	pro	zá	ú	o	ob	roz	vz	do	pod	pře	pří	v	za	před	při	pů	vý	nad	od	ot	prů	ne	bez	dů	zů	#
za	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			30
do	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+					28
na	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+					28
z	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+		28
po	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+					+	27
vy	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+		+	+	+			+					+					20
pře	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+			+	+													17
od	+	+	+	+	+	+	+	+	+	+	+			+				+	+		+	+											16
pro	+	+	+	+	+	+	+		+	+	+	+	+	+			+			+													15
při	+	+	+	+	+	+	+			+	+	+				+	+		+	+													14
roz	+	+	+	+	+	+	+	+	+	+	+			+	+														+				14
u	+	+	+	+	+			+	+							+	+	+		+	+			+						+			14
před	+	+	+	+	+			+	+			+	+		+		+					+											12
o	+	+	+			+										+			+										+		+		8
s	+	+	+					+		+			+					+															7
ne	+						+								+																		3
pod	+			+	+																												3
nad			+			+																											2
v	+																																1
vz	+																																1
bez																																	0
dů																																	0
ná																																	0
ob																																	0
ot																																	0
prů																																	0
pří																																	0
pů																																	0
vý																																	0
zá																																	0
zů																																	0
ú																																	0
#	19	15	15	14	13	13	12	12	12	12	10	10	10	10	9	9	9	9	9	9	7	7	6	6	5	5	5	5	4	3	2	1	

Tabulka 1: Tabulka kombinovatelnosti českých slovesných předpon.
Znaménko + v buňce [x,y] znamená, že existuje sloveso s dvojicí předpon x- y- (v tomto pořadí).

Překvapivá je ale předpona *ob-*, která se může vázat k mnoha různým kořenům. Tato předpona má totiž také poměrně úzce vymezený význam. Může však modifikovat velké množství sloves - jen v morfematickém slovníku je 193 sloves s předponou *ob-/obe-*. Jestliže tedy není speciální v tom smyslu, že má jen velmi omezenou množinu kořenů, ke kterým se připojuje, mohli bychom očekávat, že se bude moci připojit i ke slovesům s prefixem. To však možné není, jak vyplývá z tabulky. Z toho můžeme udělat závěr, že specifický význam této předpony se neslučuje s žádnou další předponou, která by následovala po ní.

Předpony z této skupiny se ovšem mohou objevit jako druhé. Viz např. předpona *bez-*, která může mít před sebou předpony *za-* (*za-bez-pečit*), *z-* (*z-bez-mocnit*) a *u-* (*u-bez-pečit*).

Jako (částečně) samostatné předpony bychom mohli označit i ty předpony, které se vyskytují v tabulce těsně nad těmi s nulovou frekvencí na prvním místě. Nemají sice nulovou schopnost vázat se s dalšími, ale jejich ochota přibírat další předpony je velmi omezená. Jde o předpony *v-*, *vz-*, *nad-* a *pod-*. Počet předpon, které mohou za nimi následovat, je malý (1 až 3).

Předponu *ne-* na prvním místě nebudeme dále komentovat, protože bychom se dostali do řešení rozsáhlejšího problému, kdy lze počáteční *ne-* od slovesa odtrhnout, a kdy ne. V odpovědích na tuto otázku nepanuje všeobecná shoda.

4.2 Stupňovací předpony

Řádky vyznačené v tabulce šedou barvou se týkají předpon, které se používají pro tzv. „stupňování sloves“ (viz např. Hlaváčová & Nedoluzhko 2013). Jde o předpony *roz-, po-, za-, na-, vy-* a *u-*.

Tyto předpony mají tu vlastnost, že spolu s reflexivní částicí *se* nebo *si* mohou změnit nikoli samotný význam původního nedokonavého slovesa, ale stupeň jeho intenzity. Není podstatné, zda původní sloveso již mělo, nebo nemělo předponu, je možné je připojit zleva téměř vždy, a to vždy se stejným intenzifikačním významem. Do skupiny těchto předpon by bylo možno přidat i předponu *do-* (v tabulce vyznačeno světlejší šedou), která je též velmi univerzální v tom smyslu, že ji lze připojit téměř k libovolnému slovesu, a má význam dokončení nějaké činnosti nebo akce. Předpona *do-* však, na rozdíl od ostatních stupňovacích předpon, nevyžaduje reflexivní částici.

Rozdílné součty možných kombinací na koncích řádků pramení z toho, že intenzifikační význam těchto předpon, ač velmi univerzální, se nepoužívá příliš často, takže se nám nepodařilo nalézt příklady pro všechny kombinace. Často se totiž jedná pouze o příležitostná slova vzniklá pro speciální okamžité použití, takže jejich zachycení v textech, potažmo v korpusech, je nepravděpodobné.

Některé kombinace jsou dokonce téměř vyloučené z důvodu „nepěknosti“. Např. dvojice předpon *u-u-* nemá v tabulce záznam, přestože je možné tuto kombinaci pomocí stupňování utvořit. Příkladem může být sloveso *u-u-smívat (se)*. Příklad z Internetu: *Můžu se uusmívat k smrti, rozdat se naporcovaná na kousky, obětovat se a stejně je jedna hodná ženská prostě zatraceně málo.*[2]

4.3 Předpona *z-*

Mezi častými prvními předponami se objevuje i předpona *z-*, která patří i mezi nejčastější slovesné předpony obecně, hned po *vy-* a *za-*. Zajímavé je její běžné použití ve spojení s předponami dlouhými a s cizími slovesy.

Předpona *z-* vyskytující se ve slovese před dlouhou předponou má většinou stejný význam - změnu stavu (*zprůhlednit* = učinit průhledným), případně zvýšení míry (*zpříjemnit* = učinit příjemnějším). Týká se to především těch sloves, která vznikla z podstatných nebo přídavných jmen (viz 2.2).

Ve spojení s cizím slovesem (viz též dále) je *z-* nejčastější předpona, která se používá ke zdůraznění dokonavého vidu přejatého slovesa. Většina přejatých sloves utvořených z cizího kořene přidáním české koncovky je obouvidá. Týká se to zejména nejčastější koncovky *-ovat*. Teoreticky tedy není třeba použít prefixaci pro vytvoření dokonavého významu daného slovesa. Přesto se to velmi často děje. Předpona *z-* je nejčastějším kandidátem na toto zdůraznění.

Pro srovnání uvádíme příklady použití dvojice *redukovat*, *zredukovat* z korpusu Omnia Bohemica II:

[2] http://lumenn.blog.cz/1411/jak-mi-bylo-najednou-strasne-smutno-ze-sveta [27/06/2018]

… nežádoucí účinek se <dá redukovat> tím , že léky nebudete brát ve stejný čas …
… celá záležitost se <dá zredukovat> pouze na věc určitých kulturních a sociálních norem …

Je zřejmé, že význam obou sloves - s předponou i bez ní - je totožný. Ve druhém příkladě je však zdůrazněna (v tomto případě možná zbytečně) dokonavost.

Dalšími příklady jsou slovesa *z-reprodukovat, z-improvizovat*.[3]

5 Cizí slovesa s českými předponami

Dosud jsme se věnovali českým slovesům obecně. Zajímavé je se též podívat speciálně na slovesa přejatá z cizího jazyka - říkejme jim pro jednoduchost cizí slovesa.

Cizí slovesa jsme vyčlenili ze základního souboru sloves podle pravidel vycházejících z ortografických vlastností typických sloves přejatých z cizích jazyků - nejčastěji se jedná o angličtinu nebo latinu. Sloveso označujeme jako cizí, pokud je splněna alespoň jedna z následujících podmínek:

- má cizí předponu (kategorie 6, seznam cizích předpon jsme získali z přílohy morfematického slovníku),
- obsahuje znak *g*, *x* nebo *f* (až na pár výjimek typu *foukat*, *doufat* a několika dalších),
- pahýl po odtržení předpon (českých i cizích) začíná na *a*, *e* nebo *i*,
- měkké *i* následuje po tvrdé souhlásce *h*, *r* nebo *k*,
- obsahuje dvouhlásku *ie*, *io*, *ia*, *iu*, *uo*, *ua* nebo *ue*.

Ze souboru všech sloves získaných z korpusu SYN2015 jsme podle uvedených pravidel vybrali necelé 2000 cizích sloves. Z celkového počtu všech sloves tedy tvoří ta cizí necelou desetinu. Segmentace cizích sloves probíhala stejně, jak je popsáno v sekci 3. Zajímala nás distribuce českých předpon, proto jsme pro segmentaci použili stejnou množinu předpon, tedy jen české. Cizí slovesa s cizími předponami považujeme za bezprefixová - z celkového počtu jich byly zhruba tři pětiny.

Nejčastější českou předponou stojící před cizím slovesem (více než 200 výskytů) je předpona *z-*, o níž jsme se již zmiňovali v sekci 45. Následující předpony *za-* a *vy-* mají frekvenci pouze poloviční. Předpona *za-* je občas použita ve svém stupňovacím významu (*za-relaxovat si, za-experimentovat si, za-flirtovat si*). Jeden příklad stupňovaného slovesa jsme našli s předponou *po-* (*po-diskutovat si*), také s předponou *u-* (*u-fetovat se*). Pokud se ostatní stupňovací předpony ve spojení s cizími slovesy vyskytují, mají jiný význam. Některé se nevyskytly vůbec - např. předpona *roz-*.

Ani vícenásobné předpony nejsou ve spojení s cizími slovesy běžné, ale zcela vyloučeny nejsou. Příkladem je sloveso *při-na-trefit*.

Na závěr připojíme analýzu sloves, která jsou odvozena od cizího vlastního jména *Google*, mají tedy kmen *googl*. Vyskytují se i ortografické varianty *gůgl, gúgl*, ale těmi se zabývat nebudeme.

V korpusu Omnia Bohemica II jsme nalezli skoro 100 tisíc takových slov (7,7 per milion). Po

[3] Ke zdůraznění dokonavého vidu se používají i jiné předpony, ale *z-* je, zejména ve spojení s cizími slovesy, nejčastější.

vyloučení řetězců obsahujících nealfanumerické znaky (např. různé typy internetových adres, které jsou v korpusu tokenizované jako jeden celistvý řetězec - token) jsme dostali soubor 479 různých řetězců, které jsme mohli považovat za slova. Nás zajímala především slovesa.

Nejfrekventovanější jsou slovesa vytvořená pomocí přípony *-it* a *-ovat*, totiž *googlit* a *googlovat*. Obě mají zcela totožný význam - vyhledávat něco na internetu, v poslední době se dokonce používají jako (částečné) synonymum k běžnému českému slovesu *hledat*. Pro vytvoření dokonavého vidu se i u takto vytvořených (původně zřejmě příležitostných sloves) používá tradiční postup - prefixace. V tomto konkrétním případě se využívá jako bezpříznaková předpona *vy-*. Důvodem je zřejmě převzetí stejné předpony, kterou se mění vid již zmíněného českého slovesa *hledat*.

Vyskytují se ale i jiné předpony, jak ukazuje tabulka 2. V prvním sloupci jsou uvedeny předpony, které se vyskytly v našem souboru před slovesy *googlit* a *googlovat*. Jejich počty jsou uvedeny v dalších dvou sloupcích. Poslední sloupec s nadpisem "Ostatní" uvádí počet ostatních slov s příslušnou předponou a kořenem *googl*, která se vyskytla v našem souboru. Byla zde podstatná a přídavná jména, ale i překlepy, špatně utvořená slova, občas i slova s nečeskou koncovkou, nejčastěji slovenskou. Čísla uvádíme proto, že i takto deformovaná slova dokládají frekvenci užití.

Z tabulky je zřejmé, že výběr předpon pro každé z uvedených sloves není rovnoměrný, a zřejmě ani ustálený. U předpony *pře-* je všech 5 výskytů z 3. sloupce zastoupeno tvarem trpného rodu *přegooglováno*. Trpný rod se porůznu vyskytuje i u ostatních předpon.

Poměrně často se vyskytují i slovesné odvozeniny - častěji substantiva - *googlování*, *googlení*, ale i přídavné jméno *googlovaný*, *progooglený* i *progooglovaný, vygooglovatelný* apod.

Jen pro zajímavost uvedeme nová podstatná jména vytvořená pomocí přípon, která označují člověka, který se zabývá vyhledáváním na internetu (*googlením* nebo *googlováním*), nebo i přeneseně obecně pracuje často s počítačem. Jsou to: *googl-ič, googl-ák, googl-er(ka), googl-áč, googl-ist(k)a, googl-ík*.

Celkově můžeme shrnout, že výskyt nového přejatého slova, které se rychle stalo součástí české slovní zásoby, vyvolává potřebu vytvářet další nová slova odvozená, a to pomocí předpon i přípon.

	-googlit	*-googlovat*	Ostatní
vy-	6560	4319	480
za-	949	548	46
po-	140	34	5
pro-	81	183	21
do-	69	27	0
na-	58	62	2
pře-	3	5	5
od-	0	2	0

Tabulka 2: Tabulka předpon a jejich frekvencí v korpusu Omnia Bohemica II pro základová slovesa googlit a googlovat. Sloupec s nadpisem "Ostatní" obsahuje počet všech ostatních slov, která mají danou předponu.

6 Závěr, možné využití

Morfologické značkování korpusů spoléhá především na využívání rozsáhlých morfologických slovníků. Přesto bývá ve velkých korpusech určité procento slov nerozpoznaných. Většinu tvoří vlastní cizí jména, která však lze s velkou úspěšností v jednojazyčném českém textu identifikovat podle počátečního velkého písmene. Zbytek nerozpoznaných slov tvoří hlavně slova příležitostná a překlepy.

Příležitostným slovům lidé většinou rozumějí, i když se s nimi dosud nesetkali, neboť jsou vytvářena pomocí tradičních slovotvorných postupů z morfémů, které většinou znají. Často se nová slova tvoří přidáním české předpony nebo přípony k cizímu základu. Přesnou morfematickou analýzou lze napomoci taková slova spolehlivě rozpoznat a určit jejich morfologické vlastnosti.

Pochopitelně je ještě třeba rozšířit analýzu na ostatní slovní druhy, především podstatná a přídavná jména, a posléze i na přípony. Vzhledem k rozsáhlému systému českých přípon bude následující práce mnohem náročnější. Bez existence velkých jazykových korpusů by byla téměř nemožná.

7 Literatura

Benko, V. (2014). *Aranea: Yet Another Family of (Comparable) Web Corpora*. In: P. Sojka, A. Horák, I. Kopeček, K. Pala (eds.) TSD 2014. LNCS, vol. 8655, Springer International Publishing Switzerland, pp. 257-264.

Esvan, F. (2007). *Vidová morfologie českého slovesa*. Nakl. Lidové noviny, ÚČNK.

Hlaváčová, J., Nedoluzhko, A.. (2013). Intensifying verb prefix patterns in Czech and Russian. In *TSD 2013. Proceedings, volume 8082 of LNCS*, pp. 303–310, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

MorphoDita. Accessed at: http://ufal.mff.cuni.cz/morphodita [20/06/2018].

Slavíčková, E. (1975). Retrográdní morfematický slovník češtiny. Academia.

Straková, J., Straka, M. & Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 13-18, Baltimore, Maryland, Association for Computational Linguistics.

SYN. Ústav Českého národního korpusu FF UK, Praha. Accessed at: http://www.korpus.cz [25/06/2018].

Uher, F. (1987). *Slovesné předpony*. Brno: Univerzita Jana Evangelisty Purkyně.

Poděkování

Příspěvek vznikl v rámci řešení projektů

- GA16-18177S “DerInfMorph - An Integrated Approach to Derivational and Inflectional Morphology of Czech” podporovaný Grantovou agenturou České republiky.
- “LINDAT/CLARIN - Výzkumná infrastruktura pro jazykové technologie - rozšíření repozitáře a výpočetní kapacity”, grant č. CZ.02.1.01/0.0/0.0/16_013/0001781 poskytnutý MŠMT ČR.