Příklad pravidelných slovotvorných vzorců v automatickém zpracování češtiny a ruštiny

Jaroslava Hlaváčová, Anja Nedolužko
Ústav formální a aplikované lingvistiky, MFF UK, Univerzita Karlova v Praze
Malostranské náměstí 25, Praha 1

Abstrakt. *V češtině i v ruštině existuje množina předpon, jejichž připojením k nedokonavému slovesu a přidáním zvratného zájmena pozměníme význam původního slovesa vždy téměř stejným způsobem. Toho lze využít při automatickém rozpoznávání slovních tvarů, aniž by bylo třeba je ukládat do morfologických slovníků.*

1 Úvod

Při automatickém zpracování textů se většinou využívá morfologických slovníků, které obsahují "všechny" slovní tvary (dále jen slova) daného jazyka. Slůvko "všechny" je v uvozovkách proto, že umístit všechna slova do slovníku v praxi nelze, a to z mnoha důvodů. Jedním z nich je neshoda mezi uživateli jazyka, která slova do jazyka patří, a která ne – např. spory o slova přejatá z jiných jazyků, v dnešní době především z angličtiny. Stejně tak není možné zahrnout do slovníku všechna vlastní jména, tedy jména osob, zeměpisné názvy, názvy firem apod.

Dalším důvodem je přirozený vývoj jazyků, který neustále přináší nová a nová slova. Některá se uchytí natrvalo, jiná se objeví jen v několika málo textech, nezřídka jde jen o tzv. okazionalismy, tj. slova vytvořená pro jednu konkrétní příležitost. Jiná slova naopak ze slovní zásoby mizí, a přestože se s nimi můžeme ve starší literatuře setkat, časem přestávají být v běžné populaci srozumitelná. Patří taková slova ještě do současného jazyka, nebo ne?

Slovník tedy z principu nemůže obsahovat všechna slova nějakého jazyka. Při morfologické analýze textů, která je základním kamenem pro všechny další automatické jazykové aplikace, však je třeba rozpoznat co největší množství slov. Naštěstí jsou jazyky do značné míry pravidelné, čehož lze využít k vytvoření tzv. guesserů, tedy k hádání, co slovo, ač nepřítomno ve slovníku, znamená, a jaké má vlastnosti.

V našem příspěvku jsme se zaměřili na dva zástupce slovanských jazyků - češtinu a ruštinu. Zabýváme se zde jen jedním malým výsekem, a to několika slovesnými předponami, které pozměňují nedokonavá slovesa vždy stejným způsobem, takže je lze velmi snadno a spolehlivě analyzovat.

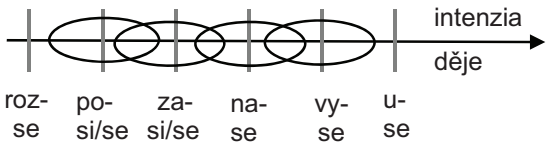
2 „Stupňování sloves" v češtině a ruštině

2.1 „Stupňovací" předpony

Většina českých a ruských nedokonavých sloves má schopnost spojovat se s některými speciálními předponami v kombinaci se zvratným zájmenem, přičemž takto vzniklé slovo modifikuje význam původního slovesa, a to podobným způsobem.

Jde o české předpony: *roz-, po-, za-, na-, vy-* a *u-* a jejich ruské ekvivalenty *раз*-(*рас*-), *по-, за-, на-, вы-* a *у-*. Pro ruštinu tuto řadu můžeme prodloužit ještě předponou *из-* /=z/.

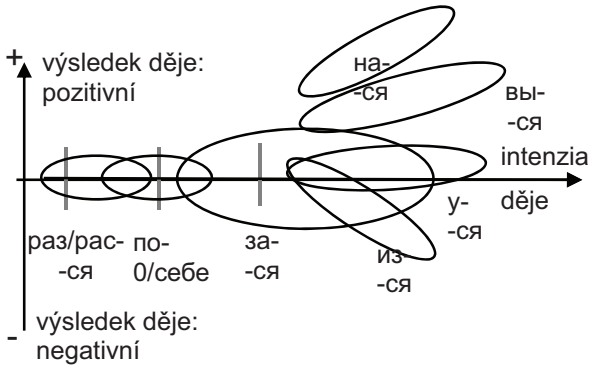
Jednotlivá prefigovaná zvratná slovesa v češtině je možno s určitou tolerancí uspořádat podle intenzity děje, jak ukazuje následující obrázek:



Krajní body tvoří předpony *roz-* a *u-*, uprostřed je podle intenzity posloupnost *po-, za-, na-* a *vy-* s vágním až překrývajícím se rozsahem.

Z tohoto důvodu nazýváme tento způsob tvoření s jistou nadsázkou pracovně „stupňování" intenzity slovesného děje.

Podobně, i když ne úplně stejně, vypadá situace s těmito předponami a zvratným „ся" v ruštině. Srov. následující obrázek::



První tři předpony bychom mohli schématicky zobrazit stejně jako pro češtinu, i když význam předpony *за-* už zasahuje „přes obvyklé hranice". Předpona *раз*-(*рас*-) navíc není umístěna tak striktně na samém začátku škály, je vágnější. Předpony *вы-, на-, у-* a *из-* však jsou již z hlediska intenzity děje na stejné úrovni a liší se sémantickými odstíny jako je např. kladný a negativní výsledek děje, což je v obrázku schematicky znázorněno svislou osou po levé straně.

Obrázek zobrazuje význam předpon zjednodušeně. Více o významu kombinací daných předpon se zvratným zájmenem je popsáno např. v [3].

V následujícím přehledu uvádíme významy jednotlivých předpon v kombinaci se zvratným zájmenem.

U každé dvojice předpon (pro češtinu a pro ruštinu) uvádíme slovotvorný vzorec: X značí nějaké nedokonavé sloveso, přidáním předpony a zvratného zájmena vytvoříme sloveso „stupňované“. Uvádíme stručně význam. Rozdíl mezi češtinou a ruštinou (pokud existuje) je uveden jen velmi vágně. Každou předponu ilustrujeme příkladem z obou jazyků, ruský příklad je přeložen do češtiny. České příklady pocházejí z korpusu SYN, ruské příklady byly nalezeny pomocí vyhledávače Google na internetu.

2.1.1 roz-X se, раз-X-ся

Význam: začít X
rusky + zesílit intenzitu X

Příklady:

Z jara se člověk nejdříve musí rozlétat.

Просто расслушаться надо, тогда почувствуешь
/= Musíš se prostě rozposlouchat, pak to pochopíš./

2.1.2 po-X se/si, по-X 0/себе

Význam: X nějakou dobu
česky + většinou spíš v klidu, užít si to
rusky + nejčastěji relativně krátce
Poznámka: Rozdíl 0/себе je stylistický, přítomnost komponentu себе (=sobě) je neobligatorní, přidává hovorový subjektivní odstín.

Příklady:

Pak vyslechl Aliciny kritické poznámky, pár minut si popřemítal, načež prohlásil, že všecko souhlasí.

Анатоль Максимович укрепился на своих ногах, еще немного пообдумывал ситуацию, а затем, очень внезапно, начал проявлять повышенную активность.
/=Anatol Maximovič se upevnil na svých nohou, ještě chvíli si popřemýšlel nad situací a pak najednou začal projevovat zvýšenou aktivitu./

2.1.3 za-X se/si, за-X-ся

Význam:X po delší dobu
česky + a užít si to
rusky + déle než obvykle, příliš dlouho

Příklady:

Loupil tady i loupežník Klempera a zaloupežil si tu i slavný lupič Babinský.

Избавляемся от вещичек, которые давно зависелись в нашем шкафчике!
/=Zbavujeme se věcí, které příliš dlouho visí (lit. zavisely si) v naší skříni!/

2.1.4 na-X se, на-X-ся

Význam: hodně X
rusky + úplně, často spíš s kladným výsledkem

Příklady:

Přestože profesor napsal tolik dopisů a tak na sebe upozorňoval, tolik se snažil a napředstavoval se v desítkách kanceláří, jeho stále hektičtější úsilí se bohužel nesetkalo s úspěchem.

Она руку привязала, чтобы не позвонить Новикову и не спросить, какими же это чудесными именами он Евгению Гордееву напредставлялся.
/= Uvázala si ruku, aby nezavolala Novikovovi a neřekla mu, jakými úžasnými jmény se napředstavoval Jevgeniji Gordejovovi./

2.1.5 vy-X se, вы-X-ся

Význam: hodně X
česky + s víceméně kladným výsledkem
rusky + až do vyčerpání

Příklady:

Francis Kennedy se tam uchýlil, aby se vytruchlil ze smrti své ženy.

Я много выстрадался с тех пор, как расстался с вами в Петербурге.
/=Hodně jsem si vytrpěl od té doby, kdy jsme se rozloučili v Petrohradě./

2.1.6 u-X se, y-X-ся

Význam: až do vyčerpání

Příklady:

Zemřel na těžkou bronchitidu. Prostě se ukašlal!

Всем смешно, а я уже весь укашлялся и усопливился...
/=Všichni se smějí, ale já jsem se už úplně ukašlal a usmrkal.../

2.1.7 ---, из- X-ся (jen rusky)

Význam: velká intenzita, často s víceméně negativním výsledkem

Příklady:

За 40 с лишним лет рисования он не изрисовался.
/= lit. Za 40 let malování se nevymaloval (ve smyslu nepřestal být dobrým malířem)/

2.2 „Stupňovaná slovesa“ v kontextu běžné slovní zásoby

Jak je vidět z předchozího přehledu a uvedených příkladů, situace v ruštině a češtině je velice podobná. Liší se jednak produktivitou v každém z jazyků, ale hlavně odstíny významů zkoumaných předpon. To ale bude předmětem dalšího, spíše lingvisticky zaměřeného výzkumu.

Některá slova utvořená podle výše popsaného způsobu, jsou běžnou součástí jazyka. Např. slovo *rozesmát se* znamená začít se smát, vyjadřuje tedy přesně to, co uvádíme v předchozím výčtu. Podobně je na tom jeho ruský ekvivalent *рассмеяться*.

Jiné složeniny tohoto typu jsou nezvyklé a vznikají skutečně jen příležitostně (viz většina výše uvedených příkladů). Existují však i taková slova, která jsou součástí běžného slovníku, mají však jiný význam. Příkladem je *usmát se*. Běžný význam je vyjádřen v SSJČ jako „úsměvem projevit radost" (např. *Hezky se na mě usmál.*) V našem stupňovaném významu nejde o úsměv, ale o smích. Následující příklad je vymyšlený, protože vyhledat kontext, ve kterém by dané slovo vystupovalo v jiném významu, je obtížné, navíc jde o velmi řídký jev.

Celý večer jsme se něčemu smáli, až jsme se skoro usmáli.

3 Využití v automatickém zpracování jazyků

3.1 Morfologické slovníky a guessery

Při automatickém zpracování přirozeného jazyka je třeba nejprve rozpoznat slova textů. K tomu slouží morfologická analýza, která každému slovu přiřadí nějaké morfologické charakteristiky. Nejčastěji to bývá základní slovní tvar (lemma), potom slovní druh a případně další hodnoty morfologických kategorií, jako je rod, číslo, pád, čas apod., v závislosti na slovním druhu. Tyto hodnoty bývají většinou vyjádřeny pomocí morfologické značky (tagu). Lemmatem sloves v češtině i v ruštině (ale i ve většině ostatních jazyků) bývá infinitiv.

Mnoho jazyků má dnes k dispozici morfologický slovník, který běžná slova obsahuje, takže ho lze využít pro jejich rozpoznání. Ostatní slova, která slovník neobsahuje[1] je třeba nějak odhadnout, označit pomocí nějaké heuristiky. Takovým nástrojům se říká guesser[2]. Obecně může guesser využívat nejrůznějších pravidel odpozorovaných z textů daného jazyka. Pozorování může být jednak "ruční", potom mluvíme o pravidlových guesserech, jednak automatické, směřující ke guesserům statistickým.

Naše pozorování, popsané v předešlých sekcích, se týká spíše pravidlového přístupu. Vzhledem k velké pravidelnosti tvoření stupňovaných sloves není třeba (a vlastně ani není možné) všechna zahrnovat do slovníku. Ta, která jsou běžnou součástí jazyka, ve slovnících většinou jsou (např. již dříve uvedený příklad *rozesmát se*), ostatní se mohou velmi spolehlivě rozpoznat.

3.2 Homonymie běžných a stupňovaných sloves

V případech homonymie (viz příklad slovesa *usmát se*) by samozřejmě bylo nejsprávnější analyzovat slovo obojím způsobem, tedy jako běžné sloveso týkající se úsměvu, i jako nejvyšší stupeň stupňování slovesa *smát se*. Vzhledem k nízkému výskytu tohoto významu to však neděláme[3].

Přesto existují určité příznaky, které by bylo možno při rozlišování významu v takových případech použít. Ve větách se stupňovaným slovesem často nebývá výslovně vyjádřeno jeho doplnění. V našem příkladě se člověk při úsměvu usměje na něco nebo na někoho, při smíchu jako takovém sloveso *usmát se* ztrácí svůj předmět. To platí o většině sloves, která mají ve svém nestupňovaném významu nějaké rozvití (ať už předmět, nebo příslovečné určení). Příkladem může být sloveso *vyhrabat se*. Ve svém běžném významu je toto zvratné sloveso rozvito nejčastěji nějakým příslovečným určením, jako v příkladě z korpusu SYN: *Vyhrabu se z peřin.* Jako stupňované sloveso by kontext mohl vypadat např. takto (příklad je opět vymyšlený): *Holčička se hrabala v písku celé odpoledne. Když se do sytosti vyhrabala, šla si hrát s panenkou.*

Rozpoznání stupňovaných sloves může usnadnit další intenzifikátor ve větě, jako *do sytosti, dosyta, úplně, k smrti, do (úplného) vyčerpání* a podobně, v ruštině *совсем, досыта, до смерти*.

3.3 Lemma stupňovaného slovesa

Na první pohled je otázka lemmatizace jednoduchá. Přidáním předpony vzniklo nové slovo, přiřaďme mu tedy jeho infinitiv i s předponou, stejně jako ostatním slovesům. Vzhledem k vysoké pravidelnosti se však na vytvoření stupňovaného slovesa můžeme dívat nikoli jako na slovotvorbu, jak by bylo z pohledu přísné lingvistiky přirozené, nýbrž jako na morfologii, čili na operaci tvaroslovnou[4]. Uvedené předpony totiž v těchto případech pouze modifikují význam základního slovesa, jak bylo ukázáno v přehledu výše.

Z toho důvodu nám připadá přirozenější zvolit jako lemma základní sloveso bez předpony. Tedy např. u slovesa *rozsténat se* by mělo být lemmatem základní sloveso *sténat*, nikoli *rozsténat*. Podobně pro ruštinu.

Kromě toho by se tím usnadnilo rozlišení mezi homonymními slovesy, jejichž příklady jsme uvedli v předchozí sekci. Ve zmíněném příkladu homonymie slovesa *usmát se* bychom dostali dvě lemmata – *smát se* pro stupňovaný význam, *usmát se* pro sloveso odkazující k úsměvu.

Toto značkování se také snáze využije při následných procesech automatického zpracování textů. Např. při automatickém překladu je velmi málo pravděpodobné, že se stupňované sloveso vyskytne v překladovém slovníku, neboť, jak jsme upozornili výše, jde většinou o příležitostná slova. Podle předpony se potom mohou vytvořit pravidla pro překlad. Např. česká předpona *roz-* při překladu do angličtiny by mohla být nahrazena výrazem *start (begin)to* + lemma.

Budeme-li tedy chtít aplikovat uvedené pravidlo na sloveso *rozsténat*, překladový slovník nám přeloží jeho lemma – sténat – moan, a dostaneme překlad *start (begin) to moan*:

James Stidham, který to vše sledoval z nosítek, se tiše rozsténal hrůzou.

James Stidham, who watched all this from a stretcher, began to moan quietly with horror.

Podobná pravidla by se dala vytvořit i pro ostatní uvedené předpony.

[1] Používá se zkratka OOV – out of vocabulary words

[2] Přes velké úsilí se nám nepodařilo zvolit vhodný český termín.

[3] Homonymie většinou odradí mluvčí od použití stupňovaného slovesa. Pokud se takové užití přesto vyskytne, bude to spíše v mluveném projevu.

[4] Podobné spory se vedou o zařazení stupňování přídavných jmen a příslovcí. Většinou se ale považuje za součást morfologie, i když některé vlastnosti jsou slovotvorné. Blíže viz práce [4], která pro nás byla inspirací pro termín „stupňování sloves".

3.4 Poznámky k morfologické analýze stupňovaných sloves

Právě představené poznatky o „stupňování sloves“ jsou zatím jen rozpracovaným tématem, nicméně tzv. předponový guesser už se v české morfologii používá, a to v projektu Morfo (viz [7]). Zde však ještě nebyla implementována lemmatizace tak, jak je popsaná v předchozí sekci. Slovesa, jakož i ostatní slova, jsou lemmatizována i se svými předponami.

Rozpoznávání stupňovaných sloves by bylo vhodné u všech sloves s předponami, aby se v případech homonymie mohlo rozhodnout, o které sloveso se jedná – zda má být předpona součástí lemmatu, či nikoli. Jelikož se ale stupňovaná slovesa objevují v textech jen zřídka, budou se zpočátku analyzovat jen ta slovesa, která nerozpozná morfologický analyzátor využívající morfologický slovník. Vzhledem k řídkosti výskytu je riziko chyby poměrně malé.

Kvůli malému počtu výskytů nemůžeme ani uvést žádné číselné charakteristiky zlepšení úspěšnosti morfologické analýzy, byly by zanedbatelné. Zpřesnění morfologických a slovotvorných charakteristik však vždy může přispět k přesnějšímu zpracování jazyka.

4 Závěr

V tomto článku jsme upozornili na existenci řady produktivních slovotvorných vzorců skládajících se z předpony (*roz-, po-, za-, na-, vy-, u-* v ruštině také *z-*) a zvratného zájmena. Na příkladech z češtiny a ruštiny jsme ukázali, že při aplikaci na téměř libovolné nedokonavé sloveso tyto vzorce mění jeho význam vždy podobným způsobem. Tato vlastnost má však zatím jenom charakter pozorování. Pro efektivní použití v NLP je třeba se podrobně věnovat některým formálně gramatickým i sémantickým vlastnostem těchto konstrukcí.

Z gramatických rysů je zajímavá tranzitivita. Zdá se, že při stupňování tranzitivního slovesa se jeho objekt generalizuje (srov. *читать /=číst/ - учитаться /=učíst se/*) a ve větě se většinou nevyjadřuje.

Ze sémantických rysů pravděpodobnost výskytů daných slovotvorných vzorců zvětšuje existence životního aktora, který je schopen daný děj řídit. U různých typů sloves mají analyzované slovotvorné vzorce různou míru produktivity. Také samotné přepony nejsou vždy stejně produktivní v češtině a ruštině, což může ovlivnit pravděpodobnost výskytu „našich“ významů při vyhledávání.

V dalším výzkumu se chceme věnovat i dalším jazykům, zejména slovanským, kde předpokládáme také existenci „stupňovaných sloves“.

Poděkování

Práce na tomto příspěvku byla podpořena z prostředků grantů GAČR č. P406/10/0875 a P406/12/0658.

Literatura

1. Ruský národní korpus, www.ruscorpora.ru
2. Český národní korpus SYN, www.ucnk.ff.cuni.cz
3. Малый академический словарь, МАС. Словарь русского языка в 4-х томах (М., Русский язык, 1999. Т. 1-4.
4. P. Karlík, Z. Hladká:. *Kam s ním? (Problém stupňování adjektiv).* In Život s morfémy. Sborník studií na počest Zdenky Rusínové, s. 73-93. Brno, MU v Brně, 2004.
5. Slovník spisovného jazyka českého. Praha: Academia, 1989
6. J. Hlaváčová: *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. Disertační práce. Praha 2009.
7. J. Hlaváčová , D. Kolovratník: *Morfologie češtiny znovu a lépe*. In: Informačné Technológie – Aplikácie a Teória. Zborník príspevkov, ITAT 2008, Copyright © PONT s.r.o., Seňa, Slovakia, ISBN 978-80-969184-8-5, pp. 43-47, 2008