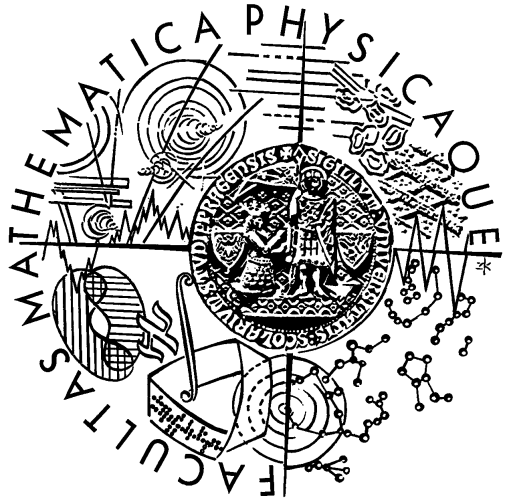
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Jan Ptáček

Generování vět z tektogramatických stromů Pražského závislostního korpusu



Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: Ing. Zdeněk Žabokrtský, Ph.D.

Studijní program: Informatika

Studijní obor: Softwarové systémy

Praha 2005

Poděkování

Chtěl bych poděkovat zejména vedoucímu mojí diplomové práce Zdeňkovi Žabokrtskému za podnětné nápady, pomoc při odstraňování chyb, za rady a za pařížský salát.

Použitý software

Pro práci s tektogramatickými stromy jsem užíval program TrEd Petra Pajase.

Pro dotazy do valenčního slovníku užívám knihovnu JHXML doc. Jana Hajiče.

Tvarosloví generuji pomocí morfologických nástrojů doc. Jana Hajiče.

Tato práce byla vysázena českou verzí programu LaTeX.

Kontakt

E-mail: *jan.ptacek@gmail.com*

Prohlášení

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 9. ledna 2006 Jan Ptáček

Obsah

1 Úvod **4**

1.1 Zadání 4

1.2 Cíl 4

1.3 Motivace 5

1.4 PDT 5

2 Fáze generování **7**

2.1 Rozklad úlohy na fáze 7

2.2 Preprocesing 8

2.3 Slovesa 10

2.4 Rekurze 13

2.4.1 Derivace 13

2.4.2 Valenční slovník 16

2.4.3 Volba obvyklé formy 16

2.5 Shoda 17

2.5.1 Algoritmus 17

2.5.2 Inicializace 17

2.5.3 Taxonomie 17

2.5.4 Poznámky 19

2.6 Složené tvary 19

2.7 Morfologie 20

2.7.1 Číslovky 20

2.7.2 Ostatní ohebné druhy 24
2.7.3 Nástroje 29
2.8 Uspořádání 31
2.9 Konektory 33
2.10 Postprocesing 35

3 Implementace a vyhodnocení **37**
3.1 Datové soubory 37
3.2 Vyhodnocení pomocí BLEU skóre 38
3.3 Rychlost 38

4 Závěr **40**

A Ukázky vygenerovaných vět **43**

B Ukázky výstupu generátoru číslovek **55**

Název práce: Generování vět z tektogramatických stromů Pražského závislostního korpusu

Autor: Jan Ptáček

Katedra (ústav): Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: Ing. Zdeněk Žabokrtský, Ph.D.

e-mail vedoucího diplomové práce: zabokrtsky@ufal.mff.cuni.cz

Abstrakt: Pražský závislostní korpus je sada digitalizovaných textů, opatřených značkami, které umožňují další elektronické zpracování. Jedním ze zápisů věty v korpusu je tektogramatický strom. V tomto stromě mají své uzly jen slova plnovýznamová – autosémantická. Významy, které byly původně vyjádřeny tvary pomocných sloves, spojkami, přímými pády, předložkami a interpunkcí, jsou zachyceny odpovídající značkou a svůj uzel nemají.

Věříme, že přeložit tektogramatický strom do jiného jazyka bude pro jeho vlastnosti jednodušší a dosáhneme tak v budoucnu lepších výsledků než současná řešení strojového překladu.

Autor této diplomové práce si kladl za cíl vytvořit program, který ze zadaného tektogramatického stromu odvodí odpovídající českou větu. Takový program najde využití při již zmíněném překladu přes tektogramatickou rovinu, ale i jako nástroj kontroly kvality stávajícího korpusu. V práci je popsán algoritmus řešení a jsou představeny a vyhodnoceny stávající výsledky.

Klíčová slova: strojový překlad, tektogramatický strom, generování vět, FGP

Title: Generation of sentences from Prague Dependency Treebank tectogrammatical trees

Author: Jan Ptáček

Department: Institute of Formal and Applied Linguistics

Supervisor: Ing. Zdeněk Žabokrtský, Ph.D.

Supervisor's e-mail address: zabokrtsky@ufal.mff.cuni.cz

Abstract: The Prague Dependency Treebank is a set of tagged sentences, well suited for futher machine processing. A tectogrammatical tree is one possible representation of a sentence in the corpora. Only autosemantic words – words with full meaning are depicted in corresponding tectogrammatical tree. Auxiliary verbs, conjunctions, grammatical cases and prepositions and their respective meanings are represented as properties of governing autosemantic words.

We believe that the task of translation of tectogrammatical tree into tree in another language will be a not so difficult one and that this approach will overcome existing MT systems.

Such a program will be of use in the above mentioned machine translation strategy. Another use-case is a quality assurance on existing corpora. The implementation and current results are presented and evaluated in this thesis.

Keywords: machine translation, tectogrammatical tree, sentence generating, FGD

Kapitola 1

Úvod

1.1 Zadání

V rámci projektu Pražský závislostní korpus jsou pro rozsáhlý vzorek českých vět vytvářeny tzv. tektogramatické stromy, které popisují strukturu věty a funkci jednotlivých jejích slov. Cílem této diplomové práce bude navrhnout a implementovat softwarový modul, který z tektogramatického stromu naopak vygeneruje úplnou větu. Bude třeba řešit mj. následující podúlohy:

- doplnění funkčních slov a interpunkce
- výpočet těch morfologických atributů, které jsou potřebné k určení konkrétního tvaru slova
- nalezení vhodného slovosledu
- vyhodnocení správnosti/přijatelnosti generovaných vět

1.2 Cíl

Nejobecnějším cílem je verifikace Funkčního generativního popisu (FGD) a jeho jedné implementace v podobě PDT. Od původního návrhu z roku 1967 prošel zápis tektogramatického stromu a zachycení jednotlivých jazykových skutečností mnoha změnami, další vývoj je jistý. Generátor je nástroj pro praktické ověření, že zvolený způsob zápisu věrně zachycuje reálné promluvy a zda případná další navrhovaná změna anotace nepřiřazuje dvěma významově rozdílným konstrukcím stejnou podkladovou strukturu.

Reálně dosažitelný praktický přínos leží ve strojovém překladu vět z cizího jazyka. Zdá se, že tektogramatické stromy téže věty ve dvouch různých jazycích jsou si podobné do té míry, že překlad by měl být proveditelný více méně pomocí slovníku. Zbytek úlohy řeší generátor z tektogramatického stromu do cílového jazyka. V našem případě do čestiny.

Přínosem je i možnost využít generátor pro kontrolu kvality stávajícího korpusu. Problémy v anotaci, na které jsme v průběhu práce na generátoru narazili, jsme průběžně reportovali vedoucímu.

1.3 Motivace

Osobní motivace vychází z mé časté práce s počítačem a z dlouhotrvajícího pocitu nepohodlí při jeho ovládání. Přiblížit to můžeme například na plánování cesty městskou hromadnou dopravou. Mluvená konverzace s člověkem, který ví, jakým autobusem jet, a zná odjezdové časy, je otázkou několika sekund. Při interakci se současnou výpočetní technikou stojí podobná výměna informací více námahy. S nástupem přenosných telefonů a jejich malých kláves se kvalita a rychlost uživatelského rozhraní ještě více zhoršuje. A přitom každý telefon má nutně i funkční mikrofon. Generování vět spadá do oblasti našeho zájmu, komunikace člověka a stroje.

1.4 PDT

Začínali jsme pracovat s PDT ve verzi 1.0, ale v době dokončení naší práce se blížilo vydání verze 2.0, domovskou stránku nové verze korpusu najdete na adrese:

`http://ufal.mff.cuni.cz/pdt2.0/`

Měli jsem přístup i k novým datům, prezentované výsledky se vztahují k PDT 2.0. Současně s novými daty se kompletuje i nová verze tektogramatického manuálu. Jedná se o srovnatelně důležitý počin. V době před PDT 2.0 byl autory v případě nutnosti odkázat na rámec jejich práce zmiňován anglický text [Sgall–1986], který rozhodně není vhodný k prvnímu seznámení čtenáře s oborem a terminologií. Nová verze manuálu [Tman–2005], byť pracovní, pro nás byla jako reference při dokončování práce nepostradatelná.

Stejně jako první verze obsahuje nový závislostní korpus data anotovaná na třech rovinách.

Morfologická rovina

Základní jednotkou morfologické roviny je věta, tedy posloupnost slovních tvarů, kterým je přiřazené morfologické lema a tag. Vygenerovat tuto posloupnost je naším cílem. V dalším textu značíme morfologickou rovinu jako rovinu M.

Analytická rovina

Větě na analytické rovině (rovina A) odpovídá strom. Hrany vyjadřují syntaktické funkce potomků ve vztahu k rodičům. Během procesu generování budeme většinu času budovat

syntaktické vztahy mezi uzly, které jsou v korpusu zachyceny analytickým stromem. Bohužel nemůžeme ukázat na nějaký okamžik v procesu a říci, že v tomto momentu máme v paměti počítače strom, jak ho definuje analytická rovina. Přineslo by nám to možnost vyhodnotit úspěšnost tohoto dílčího převodu.

Specifikace analytického stromu vyžaduje v zápisu věty informace, které nejsou nutné pro náš úkol. Z analytických funkcí pro nás hraje klíčovou roli subjekt. Když ale na základě diateze subjekt zvolíme, pouze si tuto skutečnost k němu připíšeme. Abychom naplnili definici analytického stromu, museli bychom přiřadit analytickou funkci i všem ostatním původně tektogramatickým uzlům. V dalším kroku i uzlům pro interpunkční znaménka. Myslíme si, že některé rysy zápisu analytických stromů jsou nadbytečné i pro zachycení jevů analytické roviny. Uzly pro interpunkci znovu vyjadřují informace o hranicích mezi klauzemi, které už strom obsahuje i bez nich.

Nejblíže k plnému zápisu jevů analytické roviny (ale ne přímo analytickým stromem) máme po dopočítání shody. Uzlům, jsou v té době přiřazené povrchové formy. Kroky, které následují po shodě, bychom už hodnotili spíše jako cestu k zápisu na morfologické rovině, od povrchových forem ke slovním tvarům. Rozdíl mezi pojmy povrchová forma a tvar slova můžeme ilustrovat nejlépe na složených tvarech sloves. Syntaktickému slovesu i po fázi shody stále přísluší jeden uzel. U něj je zaznamenána jeho povrchová forma – například reflexivní pasivum ve futuru. Když se poté uzel rozpadne, odvodíme od povrchové formy původního uzlu slovní tvary pro jednotlivé části jeho povrchové reprezentace.

Tektogramatická rovina

Věta na tektogramatické rovině je reprezentována také jako strom. Zaznamenány jsou pouze plnovýznamová slova. Jedná se o náš výchozí bod pro generování. Značíme jako rovinu T.

Kapitola 2

Fáze generování

2.1 Rozklad úlohy na fáze

K problému generování vět lze přistoupit z úhlů dvou vědních disciplín. Z pohledu statistiky nebo z pohledu bohemistické lingvistiky. První přístup k našemu zadání, ale v angličtině jako cílovém jazyce, zvolil například Petr Němec a Keith Hall.[1] My jsme si vybrali druhou možnost. Myslíme si, že důvody pro takové rozhodnutí jsou spíše subjektivní. Pro nás je důležité pochopení souvislostí, které analytický přístup může přinést.

Výběrem aparátu jsme jaksi automaticky provedli základní dekompozici problému. Bylo jasné, že je potřeba nějakým způsobem zpracovat následující fáze, bez kterých se nám zdá nemožné dojít od stromu k přijatelné české větě. Jedná se o určení povrchové formy, derivaci, přidání chybějících uzlů, určení slovosledu, shodu, spočítání tagů pro generátor, vygenerování slovních tvarů a konečně zploštění stromové struktury do posloupnosti slov.

Ukázalo se, že ústřední problém práce je určit přesnou návaznost vyjmenovaných kroků a jejich vzájemné prerekvizity. V prvním pokusu o implementaci jsme chtěli v každé fázi projít celý strom a provést odpovídající změny.[2] Nebyli jsme však schopni jednotlivé fáze seřadit do funkčního celku.

Zkusili jsme se tedy na problém podívat z pohledu menších jednotek, které můžeme ve stromě nalézt. Uvažovali jsme o sentenci, jednotlivých klauzích, o samotné slovesné frázi a konečně nejobecněji o větném členu. Poznali jsme, že k sentenci a klauzi se váže doplnění chybějících uzlů pro interpunkci a pro synsémantické spojovací výrazy. Že pro zpracování slovesné fráze i větného členu a jejich bezprostředních potomků už musí být provedena derivace hlavního uzlu. A také, že derivaci a určení povrchové formy od sebe nemůžeme oddělit.

Výsledkem je následující schéma. Zpracujeme fáze, které zůstaly samostatné. Poté v rekurzivním průchodu stromem provedeme pro sentence a klauze rozhodnutí o interpunkci, pro větné

[1] Zprávu o výsledcích ještě nemáme k dispozici.

[2] Jako inspirace nám zřejmě posloužila zmínka o II. a III. automatu v [Panevová–1980].

členy a slovesné fráze provedeme zároveň derivaci a výběr povrchové formy a pro slovesné fráze navíc i výběr diateze. Po rekurzivním průchodu následuje sekvence dalších izolovaných fází.

I předkládaná implementace má problémy s ne zcela vhodně zvoleným pořadím jednotlivých fází. V textu popis řadíme tak, jak si dnes po rozboru chybných případů generování myslíme, že odpovídající fáze mají následovat. Na odchylky od implementovaného stavu upozorníme.

2.2 Preprocesing

Vstupní strom si ve třech krocích předpřipravíme:

- přidání parenteze,
- převod t-lematu číslovek,
- obrácená závislost.

Každá z následujících tří změn má jiný důvod. U úprav parenteze si myslíme, že změna by se měla provést přímo v datech korpusu. T-lemata číslovek převádíme na jejich číselné hodnoty, protože to vyžaduje náš modul pro generování tvarů. V důsledku je to proto, že zápis pomocí číslic považujeme za výhodný pro strojový překlad. Jedině obrácení závislosti u nekontejnerových číslovek je první krok od tektogramatického zápisu k analytické rovině.

Parenteze

V korpusu se vyskytují případy, kdy všechny děti technického uzlu koordinace mají nastavený příznak is_parenthesis, ale sám uzel do takové vsuvky podle anotace nepatří. Důsledkem je pak věta, ve které je každý koordinovaný člen samostatně uzávorkovaný. Neznáme důvody, které k takové anotaci vedly. Myslíme si ale, že je z hlediska významu věty bezpečné nastavit příznak parenteze i pro uzel koordinace.

Číslovky

Číselné údaje zapsané slovem se snažíme převést na zápis číslicemi. Pro uzly se sémantickým slovním druhem adj.quant.def a n.quant.def voláme morfologickou analýzu a hledáme číselnou hodnotu uvedenou v komentáři k m-lematu, tj. za zpětným apostrofem. Pokud uspějeme, nastavíme nové t-lemma. Původní t-lema zachováváme v případech, které náš modul pro zpracování číslovek není schopný obsloužit.[1] Z převodu t-lematu proto vyloučíme římské číslice a číslovku s rysem úplnosti *oba*. Číselná hodnota v komentáři k m-lematu bohužel chybí

[1] Popis modulu v sekci 2.7.1.

u řadových číslovek. Některé zlomkové číslovky jsou anotovány t-lematem základní číslovky,[1] ostatní jsou nepřevedené. Ani u těchto nepřevedených se číselnou hodnotu nedozvíme.

Poté všechny číslovky zapsané číslicemi[2] rozdělíme do tří skupin. Současně pro ně určíme uzel, jehož množství počítají, a případně je sdružíme s jejich potomky, které modifikují jejich číselnou hodnotu. Označený počítaný uzel bude v pozdější fázi pro číslovku zdroj morfologických atributů, pokud do hry nevstoupí shoda. U uzlů s t-lematem zapsaným číslem tedy rozeznáváme následující typy:

i. datum
Číslovka typu datum je sémantické substantivum a její rodič je uzel s t-lematem *rok* nebo názvem měsíce. Tento typ definujeme, protože vyjádření data a letopočtu by jinak spadalo pod adjektivní typ a většinou by splňovalo i podmínky pro obrácenou závislost, což není žádoucí.

ii. kontejnerové číselné výrazy
V některých konstrukcích s číslovkami ještě není tektogramatická anotace důsledná a ve stromě je zachycená závislost z analytické roviny. Jedná se mimo jiné o následující t-lemata: *sto*, *tisíc*, *milion*, *miliarda* a o zlomkové číslovky. Kontejnerové číslovky detekujeme podle gramatému numertype = frac nebo podle t-lematu *100* a kladné mocniny od *1000*.

U nalezených výskytů se zajímáme o potomky s funktorem RSTR a MAT. První uzel materiálu je pro nás počítaný uzel. Nemusí ale být vždy přítomen.[3] Máme za to, že na taková místa by měl být na tektogramatické rovině doplněný nějaký substantivní uzel.

Potomci s číselným t-lematem a funktorem RSTR jsou další části původního víceslovného číselného výrazu. V analytickém stromu se k hodnotě celého výrazu dojde následujícím způsobem. Levé děti sečteme a součtem vynásobíme hodnotu rodiče. Pravé děti zachycují nižší řády. Postupujeme zleva, hodnoty sčítáme než narazíme na řádový uzel, přenásobíme, výsledek přičteme k hodnotě rodiče a pokračujeme až k poslednímu pravému uzlu s funktorem RSTR. V tektogramatickém stromě dojde při anotaci aktuálního větného členění ke ztrátě informace. Všechny části víceslovného výrazu mají zpravidla stejnou kontextovou zapojenost, takže dojde ke slití všech uzlů na jednu stranu rodiče. Smíchají se tak multiplikátory pro řád rodiče a pro další nižší řád. Původní hodnotu se prozatím snažíme odhadnout heuristikou. První uzel s funktorem RSTR považujeme za multiplikátor, ostatní přičítáme. Zpracované uzly s funktorem RSTR ze stromu odstraníme.

V některých pádech jsou i kontejnerové číslovky zaanotovány v adjektivní pozici. Tuto výjimku jsme neošetřili. Pokud takové uzly nemají vyplněný gramatém numertype a číselná hodnota je větší než čtyři, zafunguje obrácená závislost. *Třetina* a *čtvrtina* bude v takových případech vygenerována špatně, ale nemusí se to projevit, pokud generátor zvolí pád, který chybu zamaskuje.

[1] Definiční soubor pro anotaci uvádí *třetina*, *čtvrtina*, *pětina*, *šestina*, *desetina*, *setina*, *sedmdesátina*.
[2] Mimo číslovky s funktorem MAT, ty jsou částí složitějších výrazů, zpracujeme jejich rodiče.
[3] Například v *Jejichž počet se v Budapešti odhaduje na téměř deset tisíc.* v t-ln94204-141-p6s2.

iii. adjektivní číselné výrazy
Adjektivní typ pojmenováváme podle sémantického slovního druhu těchto číslovek. Sdružování nemá opodstatnění, pokud by byl uzel částí nějakého výrazu, proběhlo sdružování již na jeho rodiči typu kontejner. Počítaný uzel není třeba označovat, pro tyto případy v pozdější fázi zavedeme shodu.

iv. "nálepka"
Číslovky s funkcí nálepky stojí vždy za svým rodičem, mají funktor RSTR a jsou vždy vyjádřeny číslem. Neimplementujeme, protože nevíme, jak je rozeznat od adjektivního typu, který se ale chová odlišně.

v. nepřevedené číselné výrazy
S číslovkami, které se nám nepodařilo převést na číselný tvar, to dopadne také poměrně dobře. Změny v závislosti se z nich účastní jen zlomkové číslovky, a ty jsou anotovány jako kontejnerové.

Obrácená závislost

K číslovkám se váže i poslední změna. Pro adjektivní typ s t-lematem *kolik*, *mnoho* nebo s hodnotou větší než čtyři měníme směr závislosti. Číslovka se stává kontejnerem. Modelujeme tak chování číslovek, které za daných kontextových podmínek vstupují do shody s přísudkem místo počítaného materiálu. Jev se podle našeho pozorování týká prvního a čtvrtého pádu.

Předpokládali jsme, že není potřeba čekat až na určení povrchové formy, protože pro ostatní pády bude na povrchu nerozlišitelné, který člen je řídící. Mylný předpoklad jsme prozatím nahradili zavedením speciálního typu takzvané materiální shody,[1] která rozdíly realizuje.

2.3 Slovesa

Prozkoumejme, co určuje povrchovou formu pro sloveso. V této fázi již učiníme některá rozhodnutí o výsledné formě, ačkoliv je to příliš brzo, jak vysvětlíme později.

Vybrat povrchovou formu pro sémantická slovesa není pro počet faktorů, které výběr ovlivňují, jednoduché. Pokusíme se shrnout nejdůležitější rozhodnutí, které mluvčí – v našem případě generátor – musí učinit, i s jejich důsledky.

Určitý tvar nebo vyjádření infinitivem

Volba mezi infinitivem[2] nebo finitním tvarem určuje, zda se uzel bude vůči svému okolí chovat jako větný člen nebo klauze. V druhém případě musíme zajistit oddělení věty interpunkcí,

[1] Více v sekci 2.5.3.

[2] Máme na mysli pojem infinitiv ve vztahu k analytické rovině. Rozhodování o infinitivu jako o morfologické kategorii (například infinitiv jako součást složeného futura) přijde až později.

v případě podřadící věty se ujistit o přítomnosti spojovacího výrazu.

Myslíme si, že pro všechny slovesné uzly je možné zvolit vyjádření klauzí. Pro jednoduchost jsme měli v plánu generovat všechna sémantická slovesa ve tvaru verba finita. V praxi se ukázalo se, že to není možné. V tektogramatických stromech z PDT 2.0, které máme k dispozici, se odráží, zda mluvčí v předloze užil určitý nebo neurčitý tvar, ačkoliv se podle nás jedná už o volbu, která náleží k analytické rovině. Přiznejme ale, že náš silný předpoklad o možnosti vždy užít určitý tvar budeme muset zeslabit. Při zpětném zhodnocení se zdá, že mluvčí nemusí mít kompletní znalosti o ději, který popisuje, a to může vést právě k užití neurčitého tvaru. Pak i v hloubkovém zápisu nemohou být vyplněny všechny kategorie nutné pro vygenerování určitého tvaru.

Infinitiv v implementaci zvolíme pro uzly, které mají jiný slovesný způsob než imperativ a zároveň mají nevyplněný gramatémem času. Přestože oba rozhodující gramatémy nepochybně patří do tektogramatické anotace, živý mluvčí se určitě rozhoduje podle jiných kritérií.

Přechodníky nezpracováváme. Nevíme, zda je řadit k větným členům a upravit kód pro přidávání interpunkce, nebo zda je nutné přemýšlet o nich jako o samostatné skupině.

Diateze

Jako součást určení povrchové formy chápeme i volbu diateze. Mluvčí podle literatury[1] diateze užívá k vyjádření svých komunikačních záměrů, například pasivum se užívá za účelem potlačení aktora. Takové komunikační záměry vidíme v tektogramatickém stromě explicitně vyjádřeny u rezultativu a v případě dispoziční modality disp1. V ostatních případech se rozhodujeme podle náznaků, které ve stromě pro užití nějaké diateze svědčí, a podle schopností sloves jednotlivé diateze tvořit.

Jednotlivé typy diateze si definujeme čistě syntakticky, podle gramatického větného vzorce. Ne všechna slovesa jsou schopná utvořit každou z následujících diatezí. V takových případech existují dvojice sloves, které popisují identickou situaci z pohledu různých diatezí. I tomu přizpůsobíme volbu.

i. základní - (matka předělala loutku dětem z kašpárka na čerta)
Základní diateze je výchozí vztah mezi aktanty a analytickými funkcemi. Subjektem je aktor v nominativu. Infinitivy a sloveso s t-lematem *být* generujeme vždy v základní diatezi, u ostatních sloves užijeme základní diatezi pokud nejsou splněny podmínky pro nějakou sekundární diatezi. Takto definovanou primární diatezi nemusí být některá slovesa schopná utvořit. Vybereme pro ně diatezi jinou.

ii. diateze D1 - (*loutka se matkou předělala z kašpárka na čerta)
Reflexivní pasivum. Pacient je subjektem, následuje sloveso, zvratné *se* a aktor v sedmém nebo čtvrtém pádě.[2] Diatezi volíme pro slovesa, která podle valenčního slovníku neumí vy-

[1] Viz [Mluvnice II–1986, 171].

[2] Nevíme, na základě čeho pád vybírat. Jen v případě, že sloveso má tuto diatezi jako svou výchozí a základní diatezi s aktorem v nominativu netvoří, se pád dozvíme z valenčního slovníku.

jádřit subjekt v nominativu, v případě anotované dispoziční modality a pokud je aktor všeobecný či doplněný koreferující. Na příkladu vidíme, že pro výběr reflexivního pasiva platí kontextová podmínka. Pacient nesmí mít sémantický rys osobnosti, takovému pacientu by pak z věty nepřípustně plynula role činitele. Tuto kontextovou podmínku neimplementujeme.

iii. diateze D2 - (loutka byla matkou předělána z kašpárka na čerta)
Opisné pasivum. Pacient je subjektem, následuje pomocné sloveso *být*, participium trpné a případně aktor v sedmém pádě. Podmínky pro výběr opisného pasiva jsem konzultovali s prof. Panevovou. Došli jsme k závěru, že neuděláme velkou chybu, pokud místo opisného pasiva vždy užijeme reflexní pasivum. Například implementace kontextové podmínky zmíněné v předchozím odstavci si žádá klasifikaci substantiv z hlediska jejich sémantických rysů, kterou nemáme k dispozici.

iv. diateze D3 - (děti mají loutku předělánu od matky z kašpárka na čerta)
Rezultativ. Subjektem je adresát, následuje tvar pomocného slovesa *mít*, participium trpné, a aktor v předložkovém pádě *od+2*. Tuto diatezi volíme v případě anotovaného gramatému resultative = res1. U některých stromů není v takových případech přítomný adresát. Situaci řešíme zvolením diateze D2.

Další faktory

Následující gramatémy podle nás ovlivňují až výsledný tvar slovesa a nevyžadují, abychom na ně brali ohled při výše popsaném rozhodování:

- čas,
- slovesný vid,
- deontická modalita,
- slovesný způsob.

Víme, že mapování relativního času[1] na morfologické kategorie času není přímočaré. V současnosti ho tak přesto implementujeme.

K povrchové formě slovesa dojdeme v několika krocích. V této fázi rozhodujeme u každého sémantického slovesného uzlu mezi určitým tvarem nebo infinitivem. V rekurzivním průchodu stromem pak v rámci zpracování slovesné fráze vybereme diatezi. Ostatní faktory se uplatní, až budeme vytvářet analytické uzly pro jednotlivé části složených slovesných tvarů.

[1] V tektogramatickém stromě gramatém času vyjadřuje relativní časový údaj.
Více v [Čas a modalita–1971].

2.4 Rekurze

Sentence a klauze

V této fázi si v rekurzivním průchodu zaznamenáme, jaké uzly pro interpunkci budeme později vytvářet. Kritéria pro detekci sentencí a klauzí jsou následující:

Každý uzel přímo pod technickým kořenem věty je pro nás hlava sentence. Jiné uzly jsou hlavou sentence, pokud mají nastavený atribut větné modality sentmod.

Uzly sentencí mají zároveň i status klauze. Navíc je klauzí i každý uzel, který je syntaktickým slovesem a nebyla pro něj vybrána forma infinitivu. Podřadící klauze v implementaci vždy oddělujeme spojovacím výrazem. Pokud ve stromě uzel spojovacího výrazu není, poznamenáme si, že bude třeba ho dodat. M-lema takového výrazu odvozujeme od funktoru.

Slovesné fráze a ostatní větné členy

V rekurzivní průchodu všechny uzly postupně derivujeme a vybíráme jim povrchovou formu. Pro potomky buď nalezneme záznam ve valenčním slovníku nebo pro ně máme definovanou jejich obvyklou formu na základě funktoru a dalších kritérií. Volná doplnění realizujeme vždy jejich obvyklou formou. V případě slovesné fráze navíc vybíráme diatezi.

2.4.1 Derivace

Vzájemně si odpovídající uzly z tektogramatické a analytické roviny nemusí mít stejné lema. Odvození správného m-lematu se nazývá derivace. Rozdílnost lexikální jednotky nastává z několika důvodů.

Příčiny

Jednak se vybraná doplnění při anotaci zachycují se stejným t-lematem, a to jak v adverbiální, tak i v atributivní pozici. Jedná se o adjektiva a od nich odvozená deadjektivní adverbia anotovaná s adjektivním t-lematem. Do budoucna se počítá s analogickou anotací i pro deadverbiální adjektiva. Při překladu věty zpět až na rovinu *M* je třeba přizpůsobit lema pozici, ve které se uzel ve výsledné větě nachází.

Zadruhé se změnou v lematu vyjádří některé příznaky zachycené funktory nebo gramatémy indeftype a numertype. Například změnu *kdo* → *někdo* si vynucuje gramatém indeftype.[1] Dalším příkladem změny v m-lematu je vyjádření funktoru přináležitosti APP. V jistých kontextech je z hlediska spisovné kodifikace nevhodné užít genitivní formu, a je nutné od substantiva odvodit posesivní adjektivum: *klobouk otce* → *otcův klobouk*.

[1] Někdy proto nazývaný derivém. Pro seznam všech takovýchto změn viz [Razímová–2005].

Existují také zástupná t-lemata. Pokud se promítnou do výsledné věty, je třeba je nahradit odpovídajícími výrazy. Tak je tomu v případě všech osobních zájmen s t-lematem *#PersPron*.

Taxonomie

Změny, které mohou nastat, dělíme podle J. Kuryłowicze [Kuryłowicz–1936, 87-94] do skupin syntaktických a lexikálních derivací. Jako syntaktické označujeme takové derivace, kdy použité m-lema závisí na pozici ve stromě, jakou uzel zaujímá. Pokud je m-lema určeno na základě funktoru nebo gramatému, vidíme v tom změnu sémantiky slova a takovou derivaci považujeme za lexikální. Z Kuryłowitzova pohledu se nejspíše jedná o okrajové případy. U některých změn se postupně uplatňují oba principy, hovoříme pak o derivacích smíšeného typu.

Syntaktické derivace

Nutnost syntaktické derivace plyne ze syntaktické pozice uzlu ve větě. Syntaktická derivace zajišťuje přechody do adjektivní pozice u případů posesivity nevyjádřené genitivem, přechody do adverbiální pozice u základních a řadových číslovek a také deadjektivní adverbia anotovaná s adjektivním lematem a přípravu pro tutéž proceduru s deadverbiálními adjektivy.

i. n.denot na přivlastňovací adjektivum (otec → otcův)
Postupujeme tak, že k tvaroslovnému základu připojíme koncovku *-ův* či *-in* v závislosti na rodu. Tvaroslovný základ zjišťujeme dotazem na druhý a sedmý pád singuláru. Společnou část těchto slovních forem považujeme za tvaroslovný základ, který má finální skupinu (po měkčení v případě rodu ženského) ve vhodném tvaru pro tvoření posesiva. Toto řešení jsme upřednostnili před extrahováním tvaroslovného základu přímo ze slovníkového souboru dodávaného s generátorem. Nevytváříme tak duplicitní seznamy v paměti. Jiný způsob než dvě popsané metody jsme nenalezli. Předpokládáme, že informace o tvaroslovném základu bude v budoucnu součástí elektronického lexikonu, a tudíž přímo k dispozici.

ii. n.pron.indef na posesivní zájmeno (nikdo → ničí)
Tato derivace proběhne společně s lexikální derivací realizující gramatém indeftype.[1] Morfologické lema vyplývá z definičního souboru zmíněného na straně 16. Přiřazené m-lema je určeno i s přihlédnutím k případně zvolené posesivní formě.

iii. n.pron.def.pers na posesivní zájmeno osobní (#PersPron → její)
Všechny tvary jsou generátorem odvozovány od m-lemat *můj*, *tvůj* a *jeho*. Konkrétní m-lema je určeno gramatémem person.

[1] Oddíl 2.4.1, *kdo* → *nikdo*.

iv. n.pron.def.pers na posesivum zvratné (#PersPron → svůj)
Reflexiva identifikujeme podle hodnot relevantních gramatémů. U gramatémů gender, number, person a politeness je vyplněna hodnota inher, narozdíl od ostatních zájmen osobních a posesivních.

v. adj.denot na adverbium (hezký → hezky)
Informace potřebné k derivaci jsou specifikovány v seznamu příslovcí `adverbs-neg.txt`. Odtud čerpáme údaje o stupňovatelnosti a možnosti tvoření negace.

vi. adv.denot na adjektivum (dnes → dnešní)
Adjektiva deadverbiální jsou prozatím v treebanku reprezentována svým morfologickým lematem. Pro implementaci derivační rutiny nám chybí seznam cílových adjektiv odvozených z adverbií.

vii. číslovky v adverbiální pozici (tři → třikrát/potřetí)
Protože systém číslovek je podle Mluvnice [Mluvnice I–1986, 508] otevřený, prakticky nekonečný, ale využívá přitom jen několik málo základních prostředků, neřešíme generování jejich m-lemat v rámci lexikálních a syntaktických derivací pomocí výčtu v definičním souboru, ale samostatnou procedurou, používající právě několik základních tvarů.

Lexikální derivace

Pomocí lexikální derivace se v m-lematu vyjadřují semantické rysy zachycené gramatémy indeftype, numertype a funktory místního a časového doplnění. Gramatémy degcmp a negation nemají vliv na m-lema, tak jak ho vyžaduje morfologický generátor. Hodnoty těchto gramatémů se promítnou přímo do tagu, na jehož základě je vygenerován požadovaný tvar. Na t-lematu může zároveň proběhnout i více lexikálních změn z následujícího seznamu.

Pokud současně s lexikální derivací probíhá i některá ze syntaktických derivací, pak je vše upraveno najednou v rutině obsluhující primárně lexikální derivaci.

i. indeftype (kdo → někdo/kdosi/nikdo/každý/...)
Gramatém indeftype zachycuje sémantický rys neurčitosti, totalizace, záporu a interogativity u zájmen, adverbií a číslovek. Gramatém se zpravidla realizuje prefixem nebo sufixem. V případě hodnoty total i změnou kmene: *který* → *každý/všechen*.

ii. numertype (kolik → kolikátý/kolikery/kolikerý)
Zpracování určitých číslovek, jakožto souboru m-lemat co do počtu neomezeného, jsme vydělili do samostatné sekce 2.7.1. Souběžně s běžnými lexikálními derivacemi běží úpravy lematu číslovek *kolik* a *tolik*.

iii. místní doplnění (kde → kde/odkud/kudy/kam)
Při anotaci se vybraná místní doplnění převádějí na jednotné t-lema. Z doposud zpracovaných výrazů se zdá, že systém je pravidelný a od jednoho t-lematu existuje vždy celá řada m-lemat postupně se sémantikou LOC, DIR1, DIR2 a DIR3.

iv. časová doplnění (kdy → kdy/odkdy/dokdy/navždy)
Analogicky k místním doplněním se anotují i některá časová doplnění.

Všechny implementované lexikální derivace jsou inverzní ke změnám v t-lematu v průběhu vytváření tektogramatických stromů PDT. Tyto derivace jsou popsány ve strojem čitelné podobě v definičním souboru `conversion-rules.txt`. My jsme tuto informaci použili k sestrojení vlastní funkce pro zpětný převod. Kontextové podmínky vyžadované při anotaci respektujeme i v opačném směru.

Zpětné zobrazení popsané v souboru ale není jednoznačné. K mnohoznačnosti vedou případy, kdy výsledné m-lema závisí na funktoru,[1] na syntaktické pozici nebo na vybrané formě. Původní definiční soubor jsme upravili a přidali jsme kontextové podmínky, které vyřazují nevhodné kandidáty na užité m-lema.

2.4.2 Valenční slovník

Slovesný uzel nebo i obyčejný větný člen může vymezovat svým potomkům použitelné povrchové formy. Vždy se dotazujeme do valenčního slovníku, zda-li taková situace nenastala pro právě zpracovávaný uzel. Původně jsme pracovali s valenčním slovníkem sloves [VALLEX–2003]. Bohužel jsme nebyli moc úspěšní ve volbě správného rámce pro dané lema. Pak jsme zjistili, že slovesný tektogramatický uzel obsahuje v atributu val_frame.rf identifikátor rámce z konkurenčního slovníku – PDT vallexu. Ve PDT vallexu jsou rámce i pro ostatní sémantické slovní druhy. Ty už musíme hledat podle m-lematu a problém s volbou mezi případnými více rámci zůstává. V implementaci v takových případech volíme první rámec.

2.4.3 Volba obvyklé formy

Pokud s valenčním slovníkem neuspějeme, přiřadíme potomkovi povrchovou formu na základě jeho funktoru. Spíše výjimečné jsou ale případy, kdy povrchová forma závisí pouze na funktoru. Většinou existuje více možností vyjádření a jsou vázány kontextovými podmínkami, které neznáme. Velkou naději jsme vkládali například do podrobně rozebraných kontextových podmínek pro funktor MEANS v [Panevová–1980, 103]. Ukázalo se ale, že povrchová forma se zde stanovuje i podle sémantické třídy slovesa a klasifikace sloves do těchto tříd prozatím neexistuje. Prošli jsme tedy informace o jednotlivých funktorech/subfunktorech v tektogramatickém manuálu a přiřazenou formu jsme stanovili.

I slovesným uzlům v případě neúspěchu přiřadíme definovaný výchozí slovesný rámec. U všech sloves ještě musíme zajistit, že se podle zvolené diateze modifikuje přiřazený rámec.

[1] Například: *kdy* → *kdy*, *odkdy*, *dokdy* nebo *dokud*.

2.5 Shoda

2.5.1 Algoritmus

Při rekurzivním průchodu stromem jsme si průběžně zaznamenávali případy shody. Směr shody nemusí být stejný jako směr závislosti, pro podmět a přísudek platí, že přísudek přejímá morfologické kategorie od podmětu. Tento vztah si modelujeme formou šipky, která směřuje od zdroje morfologických atributů k příjemci.

Definujeme dvě množiny. Množinu uzlů F, které mají přiřazeny finální morfologické atributy a množinu šipek A – ještě nezpracované případy shody. Při inicializaci nastavíme morfologické atributy u uzlů, u kterých můžeme hodnoty odvodit z příslušných gramatémů. Nastavíme výchozí obsah obou množin a poté iterujeme.

V každém kroku iterace procházíme množinu šipek A. Pokud je počáteční uzel šipky v množině F, realizujeme shodu a cílový uzel zařadíme do F. Zpracovanou šipku z A odebereme. Pokračujeme dokud se množina F rozrůstá. Nakonec vypíšeme na chybový výstup seznam nezpracovaných šipek.

Tento jednoduchý postup mírně komplikuje koordinace a apozice.

2.5.2 Inicializace

Do množiny F na počátku umístíme následující uzly:

- syntaktická substantiva,
- uzly s t-lematy *#Gen*, *#Unsp*, *#Cor*,
- syntakticky zapojené infinitizované slovesné uzly,
- koordinace a apozice předchozích.

Je vidět, že ne všechny uzly z F disponují potřebnými gramatémy pro odvození jmenného rodu a čísla. U infinitivů, vedlejších vět a uzlů se zástupnými t-lematy, které se učastní shody podmětu s přísudkem, volíme převážně shodu jako singulár neutera nebo čerpáme z antecedentu. U apozice použijeme gramatémy prvního aponovaného členu. V případě koordinace je číslo automaticky množné, zbylé atributy převezmeme od prvního členu.

2.5.3 Taxonomie

Zastavme se u realizace šipky, u samotného přenosu morfologických atributů. Každá šipka má z doby svého vzniku přiřazený typ. Typ šipky určuje, které atributy se při shodě přenášejí. Rozeznáváme následující druhy shody:

i. subjektová shoda
Při zpracování slovesa byla určena jeho diateze a tím i uzel, který se stane subjektem. Tento uzel určuje osobu, číslo a rod výsledného tvaru slovesa.

ii. adjektivní shoda
Na syntaktická adjektiva pod rodičem substantivem aplikujeme adjektivní shodu. Přenášíme rod, číslo a pád. Z této shody jsou vyloučená posesiva.

iii. verbonominální shoda
Pokud je zpracovávané sloveso hlavou verbonominální klauze nebo pokud se jedná o *#EmpVerb*,[1] poté aplikujeme na děti s funktorem PAT shodu se subjektem. Technicky volíme řešení postupného přenesení atributů nejprve na sloveso (pomocí subjektové shody) a teprve ze slovesa na nominální část. Přenášíme jmenný rod a číslo, pádem je vždy nominativ.

iv. substantivní shoda
Pro syntaktická substantiva v rámci substantivního rodiče je určen tento typ shody, přenášející pouze pád.

v. posesivní shoda
Vlastní typ shody pro posesiva je dán potřebou nastavit morfologické atributy přivlastňovacího čísla a rodu. Výjimkou z pravidla je posesivum *svůj*, které tyto atributy ve svém tvarosloví nemanifestuje. Pro ostatní posesiva nastavujeme atribut přivlastňovacího čísla na hodnotu odvozenou od gramatému čísla. Atribut přivlastňovacího jmenného rodu má smysl pouze pro třetí osobu.

vi. koreferenční shoda
Pro všechny anotované gramatické koreference uplatníme shodu v rodě a čísle. Domníváme se, že takový postup má oporu v definici gramatické koreference, která říká, že antecedent byl určen na základě gramatických pravidel.

vii. doplňková shoda
Doplněk se v rámci možností, daných svým tradičním slovním druhem, shoduje se jménem, jehož okolnosti determinuje. Tento zdroj morfologických hodnot je v korpusu zapsán pomocí atributu compl.rf. Neimplementováno.

viii. materiální shoda
Naše řešení obrácené závislosti u některých číslovek[2] funguje dobře v případě shody podmětu s přísudkem. Je ale problémové, pokud číslovka není v rámci rodiče aktantem, ale volným doplněním. Pak může být vyjádřena i jiným pádem než nominativem nebo akuzativem, k obrácené dominaci nedojde a shoda počítaného s počítaným je druhu adjektivního. Přechodně definujeme pro funktor MAT shodu materiální, která v uvedených případech nahrazuje shodu

[1] Zdá se, že přiznat gramatém semantického slovního druhu i kvazikomplexním uzlům by pro nás bylo zjednodušením situace.
[2] Viz sekci 2.2

adjektivní. Definitivní řešení bude odložit realizaci obrácené závislosti až do okamžiku, kdy už je pád jmenné skupiny určen, i přes vyplývající praktické programátorské obtíže.

2.5.4 Poznámky

Při zpracování shody jsme mnoho jevů jednoduše zanedbali. Například vykání zachycené gramatémem politeness, shodu doplňku s podmětem nebo podrobná pravidla pro shodu koordinovaných podmětů.

Ač jsme v minulých odstavcích hovořili o tom, že uzlům připisujeme morfologické atributy, tyto atributy ve skutečnosti nepatří do zápisu na morfologické rovině. U většiny syntaktických slovních druhů se tyto hodnoty postupně beze změny přenesou až k morfologickému generátoru tvarů. Ale například u sloves, které vytvoří složené tvary, se tyto vypočítané hodnoty budou přerozdělovat mezi jednotlivé komponenty slovesného tvaru.

2.6 Složené tvary

Znovu projdeme celý strom a vytvoříme nové uzly pro složené povrchové formy. Jedná se o případy zvratného *se/si*, o pomocná slova složeného futura, préterita a vyjádření deontické modality, o kondicionál a uzly pro předložky. V případě tvarů budoucího a minulého času je situace poněkud zjednodušená tím, že generátor nyní podporuje pouze dvě varianty diateze – základní aktivum a pak reflexívní pasivum.

i. zvratné se/si

Doposud známe tři možnosti, jak se ve výsledné větě může objevit zvratné *se/si*. Buď je součástí rámce rodiče, pak mu náleží vlastní uzel s t-lematem *#PersPron*, které se derivuje na *se*.[1] Do výčtu této sekce patří zbylé dvě možnosti. Zaprvé si přítomnost částice *se* může vynutit zvolená diateze. Zadruhé může vyplynout z t-lematu slovesa.[2] V obou případech u vytvoření uzlu zaznamenáme, že má být ve výsledném slovosledu přemístěn na Wackernagelovu pozici.

ii. futurum

Tvar slovesa pro budoucí čas závisí na slovesném vidu. Pomocné sloveso *být* přidáváme u nedokonavých sloves a zároveň u plnovýznamové části vynutíme infinitiv. Případy aktivní diateze tak vygenerujeme správně, pasivum ale nebude mít správný tvar. Příznak negace přechází z plnovýznamové části na pomocné sloveso. V případě dokonavého vidu sloveso realizujeme v prézentu. Na rezultativ prozatím rezignujeme.

[1] Derivace proběhne na základě gramatémů nastavených na hodnotu inher.

[2] Jako například u slovesa *stávat_se*.

iii. préteritum
Pro první a druhou osobu se přidává pomocné *být* jako nejbližší levé dítě původního tektogramatického uzlu slovesa.

iv. deontmod
Modální slovesa se při pasivizaci nemění, takže můžeme bez ohledu na zvolenou diatezi přidat uzel pro pomocné modální sloveso. Lema se řídí anotovaným gramatémem deontmod. Morfologické atributy se i s příznakem negace uplatní na pomocném slovese, z plnovýznamového zůstane infinitiv.

v. verbmod
Podmiňovací způsob tvoříme přidáním klitiky *by*. Z původního slovesného uzlu (případně z přidaného modálního slovesa, pokud takové existuje) se stane příčestí minulé. V budoucnu by bylo vhodné spojit spojku *aby*, *kdyby* s případnou následující kondicionální klitikou, abychom se vyhnuly nežádoucí posloupnosti tvarů *aby bychom*.

vi. předložky
Další typ obohacování stromu o analytické uzly se odehrává na substantivech, a to v případě, že jim byla vybrána forma předložkového pádu.

2.7 Morfologie

Cílem této fáze převodu je opatřit každý ohebný uzel formou, slovním tvarem, který bude užit ve výsledné větě. Postup se liší pro číslovky, které se nám povedlo převést na t-lema psané číslicemi, a pro ostatní ohebné slovní druhy. Vnitřní strukturu číselných uzlů jsme řešili odděleně od běžné posloupnosti kroků, tj. derivace m-lematu, shoda, výpočet morfologického tagu a dotaz na formu. Důvodem je množství nepravidelností, které se uplatňují při odvození slovního tvaru složené číslovky. Nejprve popíšeme speciální zacházení s číslovkami, realizované v softwarovém modulu `Numerals.pm`. Popis zařazujeme v této sekci, protože odvození tvaru startuje právě v okamžiku, kdy známe morfologické atributy pro celou číslovkovou skupinu. Z teoretického hlediska ale provádíme v tomto jednom modulu pro číslovky derivaci, shodu i tvarosloví.

2.7.1 Číslovky

Specifikum číslovek se projevilo již při programování derivací. Záznamy pro číslovky v definičním souboru pro derivace, který používáme, nemohou být úplné. To nás vedlo k jinému přístupu ke zpracování číslovek určitých.

Naším základním předpokladem je, že dostatečně univerzálním vyjádřením jsou arabské číslice. Jako cíl jsme si tudíž stanovili vytvoření modulu, který z vstupního čísla a z dalších gramatémů uzlu vygeneruje číslovku zapsanou slovem. Pro většinu existujících anotovaných

vět bude tento modul také využitelný. Číslovky, které rozeznává morfologický analyzátor, mají v komentáři k m-lematu uvedenou svou číselnou hodnotu.[1]

Stejné zadání řešili autoři otevřeného systému epos,[2] který syntetizuje zvukovou podobu českých vět na základě vstupního psaného textu. Jejich sada přibližně šedesáti substitučních regulárních výrazů ale plně nepokrývá naše potřeby. Neumí například skloňovat ani utvářet tvary nutné pro uzly v adverbiální pozici. Na druhou stranu se umí vypořádat s desetinnými a zápornými čísly, které jsme my opomněli.

Z definičního souboru vyplynulo pro číslovky následující schéma 2.1. Označené tvary jsou zpracovány a jsme schopni je vygenerovat.

numertype			
basic	✓pět	patery	✓pětkrát
set	patery		
kind	paterý		
ord	✓pátý		✓popáté
frac	✓pětina		

Tabulka 2.1: Podporované tvary podle numerype

Vidíme, že gramatém numertype neurčuje m-lema jednoznačně. Tvar *patery* je vyžadován u pomnožného řídícího uzlu. Z dvojic *pět/pětkrát* a *pátý/popáté* se volí na základě syntaktické pozice. Poslední sloupec tabulky obsahuje m-lemata v adverbiální pozici.

Základ, ze kterého vycházíme, je sada m-lemat a tagů pro základní a řadové číslovky od jedné do dvaceti a všechny desítky. I v názvech z druhé desítky a samotných desítek lze pozorovat jisté pravidelnosti, výhodnější je ovšem zahrnout je mezi základní prvky. Navíc potřebujeme znát také m-lemata a tagy pro následující číslovky s významem kontejneru: *sto*, *tisíc*, *milion*, *miliarda*, *bilion*, *biliarda* a *trilion*. Uvedená data jsou zapsána v definičním souboru `cfg/numerals.csv`.

Záměrem bylo naprogramovat pravidla utváření společná pro všechny odvozené tvary. Poté pro jednotlivé typy číslovek vyjmenujeme nepravidelnosti a aplikujeme je v jednotlivých fázích odvození, abychom dostali v češtině obvyklý výraz.

Ideální počítání

Každé libovolně velké číslo rozložíme na části podle jednotlivých řádů a ty zpracujeme rekurzivně. Například u zadání '*333 111 222*' musíme spočítat miliony, tisíce a zbytek. Máme tedy tři podúlohy stejného typu jako původní úloha. Každou podúlohu vyřešíme tak, že spočítáme stovky a připojíme vyjádření posledního dvoučíslí. Zde mohou nastat dva případy. Buď se

[1] Řadové číslovky jsou bohužel výjimkou.
[2] Domovská stránka: `http://epos.ure.cas.cz`

jedná o číslovku z našeho základního souboru, tu vyjádříme odpovídajícím m-lematem. Nebo poslední dvoučíslí složíme z desítek a jednotek. Rozhodli jsme se dvoučíslí vždy generovat ve tvaru jednoslovné složené číslovky: *dvaatřicet* a nikoli *třicetdva*. Toto rozhodnutí zužitkujeme při generování například druhových číslovek, kde jiné vyjádření není možné.

Při zpětném zhodnocení vidíme, že výhradní použití jednoslovné složené číslovky zvláště pro počítání tisíců a dalších vyšších řádů zní nezvykle a bude potřeba implementovat standardní pořadí, tj. nejprve desítky, poté jednotky. Teprve nepravidelné případy, kdy je standarní pořadí vyloučeno, je vhodné vyjádřit složeným tvarem.

Kontejnerové číslovky máme definovány až k trilionu. Řády až po trilion jsou počítány po stovkách. Vyšší objemy budou vyjádřeny jako triliony trilionů trilionů ... Nemyslíme, že by se tak dělo často.

V souladu s komunikační maximou kvantity nevyjadřujeme počet u kontejnerových číslovek, jestliže se rovná jedné *(jednomu stu třinácti miliardám devíti stům jednomu milionu → sto třinácti miliardám devět set jednomu milionu)*.

Prostudujme nyní dvojici počet a počítaný předmět/kontejner. Morfologické kategorie rodu, čísla, pádu v nejobecnějším případě přebírá číslovka od počítaného předmětu. V úplnosti je tomu tak jen u číslovek s gramatémem numertype = ord. Případy, kdy je tato adjektivní shoda porušena, zpracováváme v rámci nepravidelností.

Nepravidelnosti

Rozeznáváme následující odchylky od popsaného schématu. Mnohé typy číslovek jsou utvářeny jako číslovky základní a odchylují se pouze tvarem koncové skupiny. U ostatních částí takovéto komplexní číslovky měníme hodnotu požadovaného atributu numertype na basic. Často bývá porušena shoda mezi počtem a uzlem počítaným nebo kontejnerové číslovky vyžadují konkrétní hodnoty morfologických kategorií. V definičním souboru jsou vypsána pravidla, za jakých podmínek ke změnám dochází. Změny v tvarosloví členíme do skupin podle toho, zda mají vliv na celou skupinu, pouze na vlastní počet nebo na počítanou entitu. Pořadí zápisu určuje prioritu změny. Pravidla jsme zjišťovali empiricky. Je pravděpodobé, že námi nalezenou množinu pravidel by bylo možné upravit do jednoduššího tvaru.

Pro daný atribut numertype se také v rámci nepravidelností specifikuje, zda je potřeba výslednému slovu předřadit prefix nebo k němu připojit sufix a zda se má na jednotlivé číselné skupiny aplikovat flexe.

i. základní číslovky
jedno sto/tisíc/milion ap.: → *jedna*.hidden = '1'
dvě stě: case = 1/4 → *dva*.gender='H' a *sto*.variant = '1'
o sto jedné: pokud není na konci výrazu → *sto*.case = 1
kontejnerové číslovky manifestují svůj rod
kontejnery odvozují svoje číslo od posledního dvoučíslí vyjadřovaného počtu → *sto jeden tisíc*
hodnota > 4 a case = 1/4 → kontejner.case = 2 (*pět milionů ap.*, *tisíc* má výjimku)

hodnota > 1 a case = 2 → počítané *tisíce*.sg.nominativ *(bez dvou tisíc)*
hodnota > 4 a case = 1/4 → počítané *tisíce*.sg.nominativ *(našel pět tisíc)*

pro vyjádření číslic ze základního souboru mimo kontejnerových platí:
číslovky odvozují svoje číslo od vyjadřovaného počtu (až na výjimku)
hodnota > 4 a case = 1/4 → generátor vyžaduje singulár

pro vyjádření poslední číslice jako první části jednoslovné složené číslovky se uplatní:
první_část.case = 1
první_část.gender = 'F' (až na výjimku)
hodnota = '2' → první_část.gender = 'M' *(dvaadvacet)*
první_část.sufix = 'a' *(dvaadvacet)* (až na výjimku)
hodnota = '1' → první_část.sufix = '' *(jednadvacet)*

ii. základní číslovky v adverbiální pozici (pětkrát)
Použijeme tvar vygenerovaný pro klasickou základní číslovku s počítaným předmětem v rodě mužském a připojíme sufix *krát*.

iii. řadové číslovky (pátý)
Tento typ má v definičním souboru vlastní sadu m-lemat a morfologických tagů. Ta se použije až u poslední skupiny v celém komplexním číslovkovém výrazu. Ostatní části i první část u jednoslovné složené číslovky jsou převedeny na základní typ v nominativu. Pokud výraz končí kontejnerovou číslovkou, tak se s ní asociovaný počet převádí na základní typ v genitivu *(pětistý)*.

I pro poslední skupinu ve výrazu platí maxima kvantity - nadbytečné výrazy nejsou vyjádřeny:
jedno stý/tisící/miliontý ap.: → *jedna*.hidden = '1'

iv. řadové číslovky v adverbiální pozici (popáté)
Číselná adverbia tvoříme od klasických řadových číslovek vygenerovaných s počítaným uzlem rodu ženského v šestém pádě. Připojujeme prefix *po*. Výjimku z tohoto pravidla tvoří vstupní hodnota '1', které předepisujeme tvar v definičním souboru.

v. zlomkové číslovky (pětina)
Pro vyšší řády znovu využijeme základní číslovky typu basic a zabýváme se pouze poslední skupinou. Definiční soubor pro všechny základní zlomkové číslovky uvádí m-lema *třetina*, ale definuje slovní základ, kterým se po vyskloňování nahradí řetězec *třet*. Tím se vyhneme problémům s neuplným slovníkem generátoru.

Pro samostatné použití mimo rámec naší práce je připravený skript `genNumeral.pl`. Bylo by zajímavé přibalit k němu minimální verzi morfologického serveru a nabídnout jej například programátorům systému epos.

Představené zpracování číslovek má i své zápory. V této verzi sice vytváříme analytické uzly i pro vnitřní kontejnerové číslovky, slouží však jen jako transportní objekt pro předávání

parametrů během rekurze a nezapojujeme je do překládaného analytického stromu. Rozvití číslovky jako například *tři celé tisíce* tak nejsme schopni realizovat. Produktem je nyní plochý řetězec, ne stromová struktura. Nejedná se dokonce ani o m-lemata, ale o finální povrchové tvary. Tento fakt vyvolává otázku, zda nepřesunout tuto úlohu dovnitř morfologického generátoru. V neposlední řadě je třeba analyzovat nepravidelnosti pro zbývající tři typy číslovek.

2.7.2 Ostatní ohebné druhy

Pro získaní finálního tvaru pro ostatní uzly použijeme morfologický generátor. Interface je přímočarý, specifikujeme lema a morfologický tag a obdržíme výsledný tvar. Háček se skrývá v tom, že některá m-lemata ještě nesplňují bezezbytku podmínku, kterou jsme na ně kladli v sekci o derivacích, tj. že se jedná o korektní vstup pro generátor. Ten totiž v případech homonymie v lematu a zároveň existujícího rozdílu v tvarosloví rozlišuje jednotlivá lemata pomocí číselného indexu, připojeného pomlčkou.

Určit morfologický tag znamená od syntaktického slovního druhu a od m-lematu odvodit tradiční slovní druh, tedy první a druhé místo pozičního tagu. Do zbývajících pozic pouze promítneme morfologické atributy spočítané v předchozích fázích převodu.

Při řešení tohoto problému často dotazujeme morfologický analyzátor. Vyrovnáváme se tak s neexistencí nějakého elektronického lexikonu, který by pro disambiguovaná lemata obsahoval údaje, které nejsou součástí tektogramatického zápisu, ale použité nástroje je vyžadují.

Na první pohled se zdá, že by šlo s výhodou využít možnosti uvádět v morfologickém tagu zástupný metaznak, hvězdičku. V programu této možnosti nevyužíváme. U první pozice v tagu považujeme za výhodnější tradiční slovní druh odvodit, omezíme tak případy záměn za homonymní lemata. Pro druhou pozici v tagu je užití hvězdičky omezeno na specifické případy, může pak ale ušetřit jeden dotaz na analyzátor oproti řešení, které jsme implementovali.

Slovní druh

Morfologický analyzátor nám vrátí seznam značek, které pro m-lema připadají v úvahu. Vyřadíme značky jmen, které nejsou v nominativu, značky pro neohebné slovní druhy, a filtrujeme přípustné hodnoty podle syntaktického slovního druhu:

- syntaktická substantiva → tradiční substantiva, zájmena, číslovky, adjektiva,
- syntaktická adjektiva → tradiční adjektiva, zájmena, číslovky,
- syntaktická adverbia → tradiční adverbia,
- syntaktická slovesa → tradiční slovesa.

Pokud stále existuje více kandidátů na tradiční slovní druh, zvolíme první v pořadí a ohlásíme varování. Taktéž v případě neexistence kandidáta.

Jemnější třídění slovních druhů

Dělení na slovní poddruhy je klasifikací podle mnoha hledisek. Slouží k informaci o slovotvorbě, k rozlišení na podskupiny podle typů flexe nebo na podskupiny nesoucí specifický významový rys. Postup přidělení se liší podle již určeného tradičního slovního druhu.

i. substantiva
Všechna tradiční substantiva mají na druhé pozici značku N.

ii. adjektiva
Výchozí hodnotu určíme pomocí analyzátoru. Pokud analýza není jednoznačná, bude výchozí hodnotou A. Analýza tak obslouží skupiny přídavných jmen, kde druhá pozice ukazuje na původ. Z vlastností uzlu se poté snažíme dovodit, zda není potřeba označit přítomnost některého z významových rysů. Kromě hodnot, které podrobněji popíšeme níže, se v korpusu vyskytují i uzly s podezřelou kombinací AO.[1]

A Hodnota označuje dlouhé tvary tvrdých přídavných jmen a měkká přídavná jména.

2 Označení pro první část adjektivního sousloví spojeného pomlčkou, jehož tvar vyjadřuje náležitost k většímu celku. Jako v sousloví *pedagogicko - psychologická poradna*. Určit kontextové podmínky pro přidělení této hodnoty je těžké. Nevíme, jak rozpoznat sousloví od prostého koordinovaného determinování.[2]

Generátor dvojici tag a lema zpracuje korektně, selhává pouze v případě negovaného tvaru.[3] A to jak při pokusu o negaci vyznačenou v tagu, tak v lematu.

Další komplikací je existence dvou tvarů pro některá přídavná jména. Například *anglo/anglicko*, *eko/ekologicko*, *rap/rapově*.

Přidělení značky 2 je prozatím neimplementováno.

C Jmenný tvar, nazývaný také krátký. Můžeme rozlišit tři podskupiny.

Do první skupiny patří pouze m-lema *rád*. Toto slovo tvoří jen krátké tvary a analyzátor mu správně přiřadí hodnotu C.

U druhé skupiny se krátký tvar lexikálně osamostatnil.[4] Podmníkou pro zařazení do skupiny je nemožnost záměny ve výpovědích, a to i když se odhlédne od stylistiky a případného vyjádření trvalosti vlastnosti. Nenalezli jsme kritérium určující užití krátkého tvaru. S rizikem chyby generujeme vždy dlouhý. Mezi tato adjektiva patří:

[1] Například uzel *tentam* s id m-ln94211-120-p5s10w4.
[2] Například *Stala se zdrojem nekontrolovatelných - pro armádu ne vždy výhodných - kontraktů.*
[3] Kupříkladu *nepolitícko-politické odbory.*
[4] [Mluvnice II–1986, 76]

hodný	*hoden*
mocný	*mocen*
prostý	*prost*
daleký	*dalek*
hotový	*hotov*
schopný	*schopen*
vědomý	*vědom*

Ostatní adjektiva patří do třetí skupiny. Některá krátký tvar netvoří vůbec. Při existenci obou tvarů je podle literatury možno/možné užít je k rozlišení, zda se jedná o vlastnost stálou nebo přechodnou. Krátké tvary korespondují s přechodností. Myslíme si, že použití jednoho z tvarů pro vyjádření trvanlivosti je v současném jazyce na ústupu a užívání krátkých tvarů je dáno hlavně stylem. Pro naše účely je důležité, že z tektogramatického stromu nevyčteme, zda-li se jedná o případ trvalý či přechodný a nemůže tudíž toto kritérium použít pro rozhodování, který tvar zvolit. Volíme tedy vždy tvar dlouhý.

G Značka je určena pro adjektiva odvozená od přechodníku přítomného. Toto označení geneze tedy závisí pouze na m-lematu a mělo by být přiřazeno analyzátorem. V našich testech ještě nad PDT 1.0 jsme v korpusu nalezli 218 výskytů lemat z celkového počtu 2301 výskytů tagu AG, které analyzátor neumí rozpoznat. Domníváme se, že je to způsobeno použitím odlehčené verze morfologické analýzy.

M Adjektiva utvořená z přechodníku minulého. Situace je obdobná jako u značky G.

U Přivlastňovací adjektiva. Máme za to, že koncovky *-ův*, *-in* v kombinaci se syntaktickým slovním druhem jsou spolehlivým indikátorem příslušnosti ke skupině přivlastňovacích přídavných jmen. V tektogramatické stromě by se žádné takové jméno nemělo vyskytovat. Všechna by měla vzniknout až při derivaci. Tam jim také současně s m-lemmatem připíšeme i tradiční slovní druh. Modul pro určení hodnoty na druhé pozici v tagu už provedenou volbu pouze respektuje.

iii. zájmena

Stejně jako u adjektiv nejprve zkusíme určit případy závislosti výhradně na m-lematu pomocí morfologické analýzy. Pokud výsledek není jednoznačný, nastavíme výchozí značku na 4 a posléze testujeme kontextové podmínky pro eventuální přidělení vhodnější značky.

H Krátká forma osobního zájmena, nutné užít, pokud se zájmeno nachází na Wackernagelově pozici. Prozatím neimplementováno.

5 Tvar osobního zájmena *on* vyžadovaný po předložce. Konzultujeme vybranou formu, pokud je předložková, m-lema *on-1* dostane tuto značku.

P Osobní zájmena. Značka určená pro m-lemata *já*, *ty* a *on-1*.

S Osobní zájmeno přivlastňovací. Značka pro derivovaná m-lemata *můj*, *tvůj* a *jeho*.

1 Vztažné zájmeno přivlastňovací. Značka pro derivované m-lema *jenž*.

8 Zvratné zájmeno přivlastňovací. Značka pro derivované m-lema *svůj-1*.

D Ukazovací zájmeno. Určí analyzátor.

4 Vztažné nebo tázací zájmeno *jaký*, *který*, *čí*. Určeno analyzátorem až na m-lema *jaký*, pro které analyzátor připouští i značku Z. Uplatní se výchozí hodnota 4, nastavená v případě nejednoznačnosti.

K Vztažné nebo tázací zájmeno *kdo*. Zanalyzováno v pořádku.

Q Vztažné nebo tázací zájmeno *co*, *copak*, *cožpak*. Přiděleno ručně na základě m-lematu.

Y Zájmeno *co* jako klitika po předložce. Neimplementováno.

9 Tvar vztažného zájmena *jenž*, *již* vyžadovaný po předložce. Neimplementováno.

J Tvar vztažného zájmena *jenž*, *již* v užití bez předložky. Určeno analyzátorem.

E Vztažné zájmeno *což*. Značku případně určí analyzátor, ale z anotace jako t-lema *co* není toto m-lema v současnosti nikdy odvozeno.

6 Dlouhý tvar zvratného zájmena *se*. Vybíráme pro předložkové vazby m-lematu *se*. Dlouhá forma náleží m-lematu *se* i pokud je částí ohniska, toto pravidlo neimplementujeme.

7 Tvar zvratného zájmena *se/si*. Vybíráme pro m-lema *se* bez předložky.

Z Neurčité zájmeno *některý*, *nějaký*, *cosi* a další. Značku určí analyzátor.

L Neurčité zájmeno *všechen* a *sám*. Značku určí analyzátor.

W Záporné zájmeno *nic*, *nijaký*, *nikdo* a další. Značku určí analyzátor.

O Zvlášní užití, *svůj*, *nesvůj*, *tentam*. Neimplementováno.

iv. číslovky

Číselné uzly, zapsané slovy i číslicemi zpracováváme zvlášť. Podklady pro tvarosloví číslovek jsou uvedeny v definičním souboru `cfg/numerals.csv` včetně značky pro druhou pozici v tagu. Z neurčitých číslovek zpracováváme v současnosti pouze číslovky s tagem Ca.

v. adverbia

Druhá pozice v tagu u adverbií vypovídá o tom, zda je příslovce stupňovatelné a negovatelné. Rozhodujeme podle přiděleného synktaktického slovního druhu. Značka b náleží adverbiím bez možnosti stupňování a negace. O této schopnosti víme buď přímo z tektogramatického stromu, konkrétně ze sémantického slovního druhu uzlu. Nebo v případě, že se jedná o adverbium deadjektivní, jsme při derivaci m-lematu určili nový syntaktický slovní druh i s informací o stupňovatelnosti a negovatelnosti podle seznamu `cfg/adverbs-neg.txt`.

vi. slovesa

Kromě možností vypsaných níže existují ještě značky q a t pro archaické tvary, o kterých jsme nenašli žádné podrobnější informace.

f Infinitiv. Značka je přiřazena hlavnímu uzlu klauze, pokud jsme pro něj zvolili infinitiv. Dále také značka náležím infinitivům, které jsou součástí složených slovesných tvarů.

- i Imperativ. Přidělujeme na základě slovesného způsobu. Bude potřeba tyto případy zpracovat i s přihlédnutím k osobě, pro realizaci opisných vyjádření typu *Ať pracují!*
- e Přechodník přítomný. Tvar pro vyjádření slovesného volného doplnění děje. Neimplementováno.
- m Přechodník minulý. Tvar pro vyjádření slovesného volného doplnění děje. Neimplementováno.
- B Prézentní forma. Užitá pro tvary přítomného času a prefixové tvary futura. Určujeme na základě přiřazeného atributu času.
- p Příčestí činné. Součást tvarů préterita a kondicionálu. Určujeme na základě přiřazeného atributu času.
- c Kondicionální klitika *být*. Vzniká jen jako součást složených slovesných tvarů, kde je rovnou přiřazena i značka pro druhou pozici v tagu.
- s Příčestí trpné. Bude vznikat jako součást opisného pasíva. Neimplementováno.

Tag

Nyní máme spočítané všechny údaje, které potřebujeme k sestavení morfologického tagu. Konkrétní atributy, které se do tagu promítají, závisí na tradičním slovním druhu. U zájmen je v případě osoby nastavené na hodnotu inher potřeba využít textové nebo gramatické koreference.

Indexované lema

Poslední překážkou na cestě ke slovnímu tvaru je případná homonymie v lematu a nutnost rozlišit tyto případy pro generátor pomocí číselného indexu. Částečným řešením je dotázat se morfologické analýzy, která v takových případech vrátí seznam lemat, která připadají v úvahu. Filtrováním podle tradičního slovního druhu můžeme vybrat správné lema i s indexem.[1] Řešení je částečné, protože nám nepomůže v případě homonymie uvnitř jednoho tradičního slovního druhu.[2] Pro slovesa můžeme rozhodnout i v takové situaci a to podle identifikátoru valenčního rámce. Z dostupných zdrojů se nám nepodařilo sestavit mapování rámců na indexovaná lemata, takže se zatím jedná o možnost teoretickou.

Při procházení seznamu vrácených lemat můžeme být přísní a uvažovat pouze lemata, která se shodují s m-lematem až na případný hledaný číslený index. Druhou možností je tuto kontrolu neprovádět. Ukazuje se, že lepší výsledky dostáváme pro benevolentnější filtr. Bez testu na přesnou shodu se nám podaří vygenerovat správné tvary i pro následující skupiny m-lemat, které bychom jinak neobdrželi.

[1] Například *růst-1* versus *růst-2*.
[2] Například *stát-1* versus *stát-2*.

i. variace v m-lematu
Zjistili jsme, že mezi lematy v korpusu a slovníkem generátoru jsou mimo číselné indexy i další drobné odchylky. Užitý generátor v takových případech umí tvořit slovní tvary pouze pro lemata v pravém sloupci.

sezona	*sezóna*
jakýkoli	*jakýkoliv*
alkoholismus	*alkoholizmus*

ii. rozdíly v negaci
Zápor je v korpusu pro některé uzly anotován jen v t-lematu.[1] Morfologické nástroje ale předpokládají kladný tvar lematu a případnou negaci zapsanou v tagu.

iii. vlastní jména
Generátor pochopitelně nemá ve slovníku záznam pro každé možné příjmení.[2] Když se ve slovotvorbě podaří postoupit o krok zpět, zvyšuje se šance na vygenerování správného tvaru.

Předkládaná řešení jsou dočasná. V některých případech jsou zdrojem chyb.
Je šéfem mocné vojenské komise při ÚV KS Číny.
**Je šéf mocné vojenské komise u ÚV Kus Číny.*
Zavedení jednoznačných identifikátorů t-lemat se podle našeho názoru v budoucnu nevyhneme.

Tvar

M-lema a tag předáme generátoru. Vrácený tvar ještě upravujeme s ohledem na výše zmíněné výjimky. Doplníme zpět zápor pomocí prefixu *ne-* tam, kde o něj m-lema v poslední fázi přišlo, a u vlastních jmen znovu zajistíme velké počáteční písmeno.

2.7.3 Nástroje

Používáme již naprogramované morfologické nástroje, k dispozici jsme měli jejich demoverze. Logické je využití generátoru tvarů, analyzátor se nám hodí pro zjištění informací, na jejichž základě tvoříme posesívní tvar nebo formulujeme dotaz pro generátor. Oba nástroje pracují se slovníkem. Jeho načtení je časově náročnější operace, takže s výhodou používáme serverové varianty, která startuje jen jednou. Pokud se skriptu nepodaří spojit se na nakonfigurovaných adresách a portech, pokusí se servery spustit lokálně.[3]

[1] Příkladem jsou t-lemata *nezaměstnaný*, *neustálý*.
[2] Například t-lema *Zeman*.
[3] Testováno v unixu a prostředí cygwin. Neodladěné pro active perl.

Fuzzy tagy

Tag na druhé až osmé pozici popisuje morfologické atributy, jejichž vyjádření není ve všech případech jednoznačné. Proto jsou definovány i hodnoty pro případy, kdy ze slovního tvaru nelze katerogii jednoznačně určit.

V otázce značek je generátor striktní a v případě mnohoznačnosti vyžaduje víceznačný kód. Situaci řešíme v modulu `MorphologyConnection.pm` postupným zkoušením stále obecnějších značek.

Například pro jmenný rod postupujeme podle následujícího uspořádání.

$$\begin{aligned}
&\mathsf{M} \to \mathsf{Y} \to \mathsf{Z} \to \mathsf{X} \\
&\mathsf{I} \to \mathsf{Y} \to \mathsf{Z} \to \mathsf{T} \to \mathsf{X} \\
&\mathsf{F} \to \mathsf{Q} \to \mathsf{T} \to \mathsf{H} \to \mathsf{X} \\
&\mathsf{N} \to \mathsf{Q} \to \mathsf{Z} \to \mathsf{H} \to \mathsf{X} \\
&\mathsf{-} \to \mathsf{X}
\end{aligned}$$

Dotazování by se dalo urychlit. Generátor akceptuje na vstupu i více tagů a vygeneruje tvar podle prvního použitelného. Tuto vlastnost nevyužíváme.

Příliš mnoho tvarů

Pro jeden tag někdy generátor vrací dva i více tvarů, pro m-lema *rodina* → *rodina/rodinka*. Obě varianty rozhodně nejsou v takových případech zaměnitelné. Bohužel pořadí, ve kterém generátor na jednom řádku výsledky uvede, se liší například i v jednotlivých pádech, takže volba použitého tvaru je v těchto případech v podstatě náhodná. Zdá se ale, že plná verze morfologických nástrojů vrací výsledky v deterministickém pořadí.

Rozšíření

Morfologický generátor pracuje se slovníkem. Pokud pro nějaké m-lema nenalezne záznam v aktuálním slovníku, generování tvaru končí neúspěšně. Při analýze problému výpočtu značky na druhé pozici v tagu u adjektiv jsme zjistili, že generátor neumí skloňovat přivlastňovací adjektiva odvozená od vlastních jmen. Přesněji, poradí si pouze s některými adjektivy tohoto typu, která má ve slovníku. Přitom koncovky jsou shodné pro celou třídu přivlastňovacích přídavných jmen. Případné rozšíření slovníku nikdy plně nepokryje celou třídu přivlastňovacích adjektiv.

Situaci řešíme úpravou kódu generátoru. Zavádíme pojem *analogy rule* s významem: slovo nebo třída slov se skloňuje podle definovaného vzoru. Pokud běžný postup neuspěje, projdou se sekvenčně všechna pravidla. První vyhovující se použije. Třída slov, pro které je pravidlo platné, je určena dvěma regulárními výrazy. Aplikují se jeden na t-lema a druhý na tag a pokud oba uspějí, pravidlo je vybráno. Pravidlo specifikuje, jaké m-lema se má použít jako vzor a zda-li je potřeba upravit tag specifikující požadovaný tvar.

Seznam pravidel se nachází v souboru `analogyRules.txt` v adresáři s morfologickými servery. Kódování češtiny tohoto souboru se musí shodovat s použitým slovníkem. Soubor s pravidly se zadává jako volitelný parametr z příkazové řádky.

Příklad záznamu pro přivlastňovací adjektiva z definičního souboru: `otcův otc ův ův$ ^AU`

Myslíme si, že pravidlo tvořící přivlastňovací tvar i s rizikem zanedbání nepravidelností je lepší než použití m-lematu, jak tomu je v případě, kdy slovník selže.

2.8 Uspořádání

V současné implementaci uspořádáváme uzly ve třech krocích. První skupina pravidel pro přesuny se uplatní ještě v přípravné fázi, před rekurzivním zpracováním sentencí, klauzí a větných členů. V závěrečné části generování se vyhodnocují zbylé dvě skupiny pravidel, a to před a po realizaci uzlů pro konektory. Toto rozdělení nemá žádné teoretické důvody.

Současný stav vznikl na základě našich nesprávných předpokladů. Věděli jsem, že aktuální větné členění, tedy faktor, který s pořadím uzlů zamíchal v náš neprospěch, byl anotován až dodatečně. Bez hlubšího zkoumání jsme předpokládali, že tyto změny budeme schopni zvrátit v přípravné fázi.

Pro rozhodnutí, kdy uspořádání uzlů ve stromě provést, potřebujeme znát kritéria, která tvoří kontextové podmínky, na jejichž základě k přeuspořádání dochází. Z analýzy následujícího seznamu pravidel vyplývá, že při rozhodování o pozici hraje roli syntaktický slovní druh uzlu i rodiče, případná změna klauze na infinitiv, rozvrstvení výpovědní dynamičnosti, funktor, obligatornost doplnění, větná modalita.

Z uvedeného plyne, že pokoušet se o přeuspořádání v přípravné fázi je nerozumné. V kódu tak část pravidel vyhodnocujeme předčasně jen z toho důvodu, že už na nich máme závislé další dva moduly. Implementace obrácené dominace číslovek tyto změny neoprávněně předpokládá a přidávání interpunkce máme také zařazeno příliš brzy.

Všechny přesuny zmíněné v této sekci jsme zavedli na základě pozorování, ve sledovaných větách některé obraty nevypadaly přirozeně. Nejedná se tedy o úplný výčet potřebných změn, který by byl podložen soustavným výzkumem.

Přesuny první fáze

i. RSTR
Shodné přívlastky jsou umístěny pod svým rodičem z našeho pohledu většinou nepříznivě. Uzly stojí za svým rodičem, a to v opačném pořadí, než je jejich přirozený povrchový slovosled. Překlopíme je tedy před rodiče. Při této operaci kontrolujeme, že se přesouvané uzly neocitnou vlevo od případných ukazovacích nebo zájmenných přívlastků.

Kandidátem na přesun není každý přívlastek. Více strukturované přívlastky naopak musíme ponechat v jejich pozici napravo. Nevíme přesně, co takový přívlastek charakterizuje. V implementaci prozatím testujeme, zda není rozvitý sémantickým substantivem.

Uzly s funktorem RSTR mohou být vyjádřeny i vedlejší větou. V tom případě přesouváme uzel vpravo za rodiče, což je potřeba, když se jedná o kontextově zapojený uzel.

ii. MANN
S adverbiálními doplnění způsobu je to podobné jako s přívlastky. Doplnění způsobu zpravidla přináší novou informaci. Tím je v anotaci určena pozice vpravo od rodičovského uzlu. Zdá se nám, že pak ale není z čeho usoudit, zda větší výpovědní dynamičnost má rodič, nebo jeho doplnění s funktorem MANN.

Bezpříznaková pozice pro taková doplnění, pokud mluvčí nechce tuto okolnost zdůraznit, je nalevo od rodiče. Tam také všechny uzly s funktorem MANN přesouváme. Ve větách, kde způsob je hlavní sdělovaná informace, tak generujeme chybné pořadí.

iii. časová a místní a další doplnění u substantiv
Funktory PAT, ORIG, ADDR, APP, DIR1, DIR2, DIR3 a LOC přesouváme za rozvíjené substantivum.

iv. EXT
U všech sémantických slovních druhů, které jsou rozvity uzlem s funktorem EXT, volíme přerovnání vlevo. U sloves, kde je doplnění obligatorní, by ale mělo stát vpravo. V takových případech děláme chybu.

v. REG, DIFF
V levé pozici chceme mít funktory REG a DIFF, které netvoří vedlejší větu a rozvíjí adjektivum nebo adverbium.

vi. ID
Do pozice nalevo od rodiče rovnáme uzly s funktorem identity.

vii. indeftype = inter
Ve větách s uspořádáním typickým pro tázací klauze přesouváme tázací zájmena, adverbia i číslovky na počátek. Nejedná se pouze o věty s anotovanou tázací větnou modalitou. Přesouváme všechny výskyty vyjmenovaných uzlů.

viii. CPR
Volné doplnění způsobu vyjádřené pomocí srovnání je anotované tak, že pod uzlem vyjadřujícím míru podobnosti je zopakovaný uzel pro děj/entitu, se kterou se srovnává. V případě srovnání dvou dějů se tak z druhého srovnávaného děje standardním způsobem stane vedlejší věta. Je však vložená mezi větné členy prvního děje, což nevede k přijatelnému povrchovému slovosledu. Srovnávací vedlejší větu zařadíme až za všechny členy prvního děje. Se srovnávanými entitami nakládáme stejně jako s ději.

V první fázi přerovnání posuneme uzel vyjadřující míru podobnosti tak, aby byl poslední mezi svými bratry. Po rekurzivním průchodu, tj. ve druhé fázi přerovnání, vyrobíme neprojektivní konstrukci, kdy se uzel míry vrátí zpět na své místo, ale jemu podřízená srovnávací vedlejší věta zůstane za srovnávaným dějem.

Přesuny druhé fáze

i. posesiva
Derivovaná přivlastňovací m-lemata budou stát před rodičem a před případnými bratry s funktorem RSTR. Pokoušeli jsme se vyjádřit kontextovou nezapojenost vynucením genitivního tvaru, při kterém uzel stojí za rodičem a nese tak větší důraz. Výsledky jsou dobré, až na možné odchylky u jmen ulic, ústavů a podobně. Nevíme, podle čeho o formě v takových případech rozhodnout.

ii. slova v genitivu
Slova v genitivu musí přijít jako první pravé děti za svým rodičem. Tam, kde uzlům přiřazujeme genitivní povrchovou formu, pro ně zajistíme i pozdější přeuspořádání. Vyloučíme tak případy, kdy uzlu byl druhý pád přidělen na základě substantivní shody.[1] Uzel může dostat genitivní formu při nezvolení posesiva v derivaci, při užití rámce z vallexu a nebo při odvození genitivu jako vhodného vyjádření pro daný kontext, když záznam ve vallexu není k dispozici.

iii. první pozice mezi levými
Požadavek na přerovnání vzniká v průběhu rekurzivního průchodu stromem. Pokud je vedlejší věta uvozená konektorem a jedná se o autosémantikum, má takový konektor ve stromě svůj uzel. Jeho pozice je dána vztahem závislosti, takže často není na první pozici mezi uzly, které patří do klauze, jak to povrchový slovosled vyžaduje. Přerovnání na první pozici se uplatní i v případě souřadného spojení vět pomocí konektoru *což*.

Přesuny třetí fáze

i. klitiky
Stálé příklonky patří na druhé místo v klauzi. V případě, že je zájemců o Wackernagelovu pozici více, řadíme klitiky za sebe podle klesající priority. Největší prioritu přiznáváme tvarům kondicionálního *být*. Následuje pomocné *být* od préterita. S nejmenší prioritou přesouváme uzly zvratného zájmena *se/si*, které má původ ve vybrané diatezi, pochází z t-lematu slovesa nebo bylo derivováno z *#PersPron*. Do výčtu patří i krátké tvary zájmen, ty ale zatím neimplementujeme.

Přesuny druhé fáze provádíme před tím, než do stromu doplníme uzly pro interpunkci a pro kontektory. Při přerovnání na první pozici si tak situaci nekomplikujeme dalšími typy uzlů, které bychom museli brát v úvahu. Pro klitiky je ale přítomnost uzlů konektorů důležitá pro správné zařazení.

2.9 Konektory

Téměř před koncem procesu generování dodáme uzly pro interpunkci a pro spojovací výrazy, které ve stromě zatím chybí. Budeme je souhrně označovat jako konektory. Uzly vytváříme na

[1] Například *bez Miloše Zemana*.

základě informací, které jsme si poznamenali při rekurzivní průchodu stromem. K jednomu uzlu vždy může náležet sada levých konektorů, které nakonec budou stát jako jeho děti nejvíce vlevo z celého podstromu, a podobně i sada pravých konektorů. Mimo konektorů přidáváme i bezprostřední levé a pravé bratry za účelem ozávorkování částí stromů, kde mají všechny uzly nastavený atribut is_parenthesis. Popíšeme kroky z rekurzivního průchodu pro jednotlivé typy větných částí, které v této fázi vedou k vytvoření nových uzlů.

Sentence

Každý uzel přímo pod technickým kořenem věty je pro nás hlava sentence. Jiné uzly jsou hlavou sentence, pokud mají nastavený atribut větné modality sentmod. Se sentencí se pojí jen interpunkční znaménka. Pro přímou řeč[1] přidáme oboje uvozovky, pokud se nejedná o první člen, tak i uvozovací dvojtečku. Mezi levé konektory zařadíme také speciální uzel, který se ve fázi postprocesingu postará o počáteční velké písmeno. Původně jsme ho zařazovali pro všechny sentence, ale je vhodnější užití velkého písmena omezit jen na sentenci bezprostředně pod kořenem a pro přímou řeč. Přiřazené koncové znaménko je určeno větnou modalitou. Ne vždy je ale správné koncovou interpunkci opravdu přidat. Pokud je v podstromě přítomná subsentence a nenásleduje za ní už žádný člen nadřazené sentence, tak finální koncovou značku určuje atribut sentmod vnitřní sentence.

Klauze

Uzly sentencí mají zároveň i status klauze. Navíc je klauzí i každý uzel, který je syntaktickým slovesem a neprošel infinitivizací. Klauzi je potřeba oddělit od okolních klauzí pomocí čárek a spojovacích výrazů. Po neúspěšném pokusu naimplementovat přidávání potřebných levých i pravých konektorů jsme dodefinovali, že klauze je odpovědná za svoji separaci od okolí pouze na levé straně. Za konektory na pravé straně se postará buď následující větný člen, klauze nebo sentence. Při zpracování podřadící věty zapíšeme do kontextu rodiče příznak, signalizující takovou nutnost oddělení. Dokud se tento příznak nezpracuje, nebo nenastanou podmínky pro jeho vynulování,[2] předává se i do nadřazených kontextů.

Pro podřadící věty je mimo levé čárky třeba zajistit i přítomnost odpovídajícího spojovacího výrazu. Nejprve hledáme vhodný spojovací výraz mezi větnými členy klauze, tj. v podstromu uzlu klauze, ze kterého vyloučíme všechny subklauze. Pokud vhodný spojovací výraz nenajdeme, vytvoříme ho jako nový levý konektor. Lema určíme podle funktoru a subfunktoru.

V případě, že klauze je členem koordnice nebo apozice, oddělujeme ji od předchozích členů čárkou. Pokud se jedná o poslední člen ze skupiny, užijeme t-lema z uzlu koordinace. Víceslovné spojovací výrazy zatím nerealizujeme.

[1] Pokud je označená atributem is_dsp_root.

[2] Nulujeme na konci zpracované sentence.

Větné členy

Pro větné členy vyplývá povinnost zajistit oddělení od případné předcházející podřadící věty. Dalším jevem je koordinace a apozice větných členů, tu řešíme stejně jako v případě klauzí. Společně s konektory zpracováváme i atribut is_parenthesis. Takové parentezi říkáme členská parenteze, narozdíl od běžné vsuvky, která je anotována funktorem PAR. Technicky rozdíl spočívá v tom, že v případě funktoru PAR se v závorkách octne celý podstrom, u členské parenteze tomu tak není. Pro každý větný člen testujeme, jestli je hlavou členské parenteze. Tj. zda má narozdíl od svého rodiče nastavený sledovaný příznak. Pro všechny hlavy členských parentezí poté v této fázi hledáme jejich děti s příznakem, které stojí nejvíce vlevo a vpravo v odpovídajících podstromech. Těm vyrobíme bezprostřední levé a pravé bratry – závorky.

Na oddělení volných přívlastků jsme zatím nepomýšleli, protože funktor DES byl z anotace vypuštěn.

Poznámky

Popsaný přístup se ukázal jako schůdný, chybou je, že jsme k výpočtu interpunkce přistoupili příliš brzy. Myslíme si, že vhodnější bude znovu rekurzivně projít strom, a to jen za účelem přidání interpunkčních znamínek a chybějících synsémantických konektorů. Umožní nám to zrušit spornou první fázi přeuspořádání a o případné infinitivizaci bychom pak mohli rozhodovat až v prvním rekurzivním průchodu, kdy už známe subjekt rodičovské klauze. To, že současná implementace má také obstojné výsledky, je dáno tím, že ze stromů je velice dobře poznat, jestli v původní větě byl slovesný uzel realizovaný infinitivem nebo nějakým finitním tvarem. To nám umožnilo zvolit typ klauze velmi brzy a vyspravit tím předchozí chybná rozhodnutí o posloupnosti jednotlivých fází. Podle nás by se ale o infinitivizaci mělo uvažovat až nepoměrně později, alespoň pokud se jedná o infinitivy, které jsou důsledkem jazykové ekonomie. O takové infinitivizaci podle nás rozhoduje hlavně fakt, zda kontrolující člen z rámce nadřazeného slovesa je shodný s takříkajíc implicitním subjektem infinitivního vyjádření. V opačném případě úsporné užití infinitivu není možné.

2.10 Postprocesing

Velké písmeno

Změnu velikosti počátečního písmena u toplevel sentence a vybraných subsentencí provedeme ještě na stromové struktuře. Z podstromů, které mají za kořen uzly se speciálním levým konektorem AuxBig, vybíráme vždy uzel postavený nejvíce vlevo, který není elidovaný a má spočítanou formu.

Posloupnost slov

Protože na uzlech stromu udržujeme úplné uspořání, je odvození posloupnosti tvarů jednoduché. Drobnou komplikací je distribuce mezer mezi tokeny. Obecně se každý další token odděluje od přechozího jednou mezerou vlevo. Pro výjimky jsme definovali seznamy interpunkčních znamínek, které nemohou mít mezeru před sebou nebo za sebou. Zvláštní zacházení vyžadují uvozovky, protože na nich není na první pohled poznat, jestli se jedná o uvozovku otevírací nebo ukončovací. Poznáme to podle binárního příznaku, který si překlopíme s každou další uvozovkou na výstupu.

Vokalizace

Předložky převádíme na jejich vokalizovanou podobu až ve výsledné posloupnosti tvarů, protože Encyklopedický slovník [ES–2002, 349] uvádí, že rozhodující je fonetická podoba začátku následujícího slova. Zpráva [Petkevič–1998] ale uvádí i případy, kdy vokalizace slouží k odlišení homonymních tvarů.[1] Než zpracujeme pravidla doc. Petkeviče, používáme regulární výraz pro základní vokalizaci předložek *s, z*.

[1] Například *beze vší úcty* oproti *vlasy bez vší*.

Kapitola 3

Implementace a vyhodnocení

Náš skript pro btred je zároveň i makrem pro TrEd. Instalujeme ho proto do adresáře tredlib/contrib/generate. Výhodou zvoleného řešení je možnost v TrEdu krokovat jednotlivé fáze generování. V kontextu generate funguje klávesová zkratka Ctrl-g pro spuštění generování na aktuálním stromě. Zaplatíme za to tím, že při každém startu takového TrEdu se při inicializaci našeho prostředí do paměti načte valenční slovník.

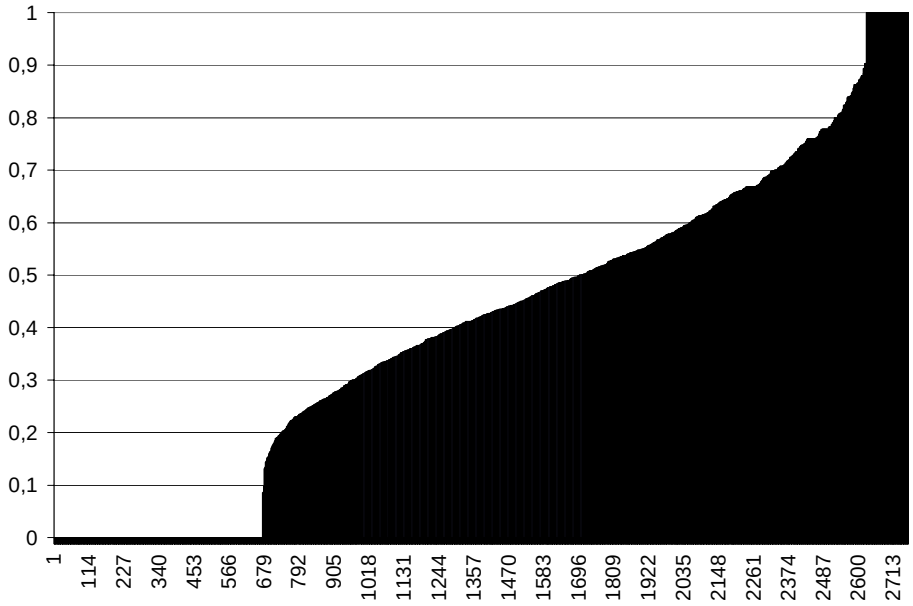
V celém textu jsme se vyjadřovali, jako kdyby neexistovalo zmnožení pozice díky koordinaci nebo apozici. Ve většině případů opravdu nechceme rozlišovat, zda pracujeme s uzlem typu complex nebo coap. Proto jsme si nadefinovali řadu metod pro třídu FsNode, které interně testují, zda je nevoláme na koordinaci nebo apozici, a pokud ano, provedou operaci na prvním členu.

3.1 Datové soubory

V průběhu generování využíváme následující zdroje informací:

- zápis vstupního stromu v tektogramatickém stromě,
- morfologický generátor a analyzátor,
- definiční soubor pro inverzní derivace doplněný o kontextové podmínky pro derivace,
- seznam příslovcí a údajů o jejich schopnosti tvořit negaci a možnosti stupňování,
- modul `Lexicon.pm`, kde definujeme některé slovníkové seznamy – měsíce, dny, vazbu rámec → m-lema.

Pokusem jsme ověřili, že například omylem nevyužíváme linkované informace z nižších rovin. Pokud smažeme soubory, které související data z analytické a morfologické roviny obsahují, generátor vygeneruje stejný výstup. Chybět bude pouze rekapitulace předlohy, sestavovaná na základě smazaných dat.



Obrázek 3.1: BLEU skóre pro vzorek 2761 vět ze sady DTEST

3.2 Vyhodnocení pomocí BLEU skóre

Na posloupnost anotace – generování se můžeme dívat jako na překlad z češtiny do češtiny. To nám umožní hodnotit úspěšnost pomocí BLEU skóre, jak je to v oblasti strojového překladu běžné.[1] Pomocí skriptu mtEval1.1 jsme napočítali skóre pro vzorek 2761 vět ze sady DTEST a pro vzorek 780 vět ze sady ETEST. Ohodnocení jednotlivých vět pro obě sady jsme seřadily vzestupně a vynesli do grafu. Skóre 0.0000 mají věty, ve kterých nesouhlasí ani jeden 4-gram. V souborech je hodně segmentů, které mají například jen 3 slova.

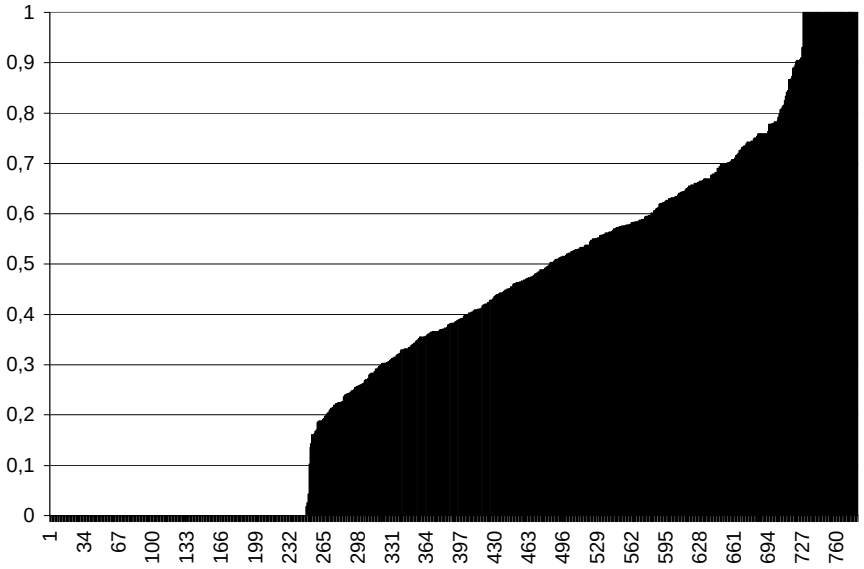
dtest	0.4773
etest	0.4801

3.3 Rychlost

Vygenerování celkem 378 vět z šesti souborů, které jsme užívali pro ladění, trvá na naší pracovní stanici přibližně deset minut.[2] Časově náročné je hledání záznamů ve valenčním

[1] Podrobnosti v [Papineni-2001].

[2] Pentium M 1.6GHz, 768MB takže nedochází ke swapování, prostředí cygwin, morfologické servery lokálně



Obrázek 3.2: BLEU skóre pro vzorek 780 vět ze sady ETEST

slovníku pro jiné slovní druhy než sémantická slovesa. Ta mají v atributu val_frame.rf vyplněný identifikátor valenčního rámce. Pro ostatní slovní druhy to neplatí a jejich sadu rámců hledáme podle lematu. Rozdíl mezi těmito dvěma způsoby dotazování do vallexu jsme odhadli pomocí testu, ve kterém jsme se u ostatních slovních druhů pro simulaci náročnosti ptali na rostoucí posloupnost identifikátorů rámců. Čas zpracování se pak vrátil na hodnoty před využitím valenčního slovníku i pro neslovesa, tj. na přibližně 3 minuty.

Podstatné zrychlení je tedy možné pomocí vyplnění identifikátoru rámce i pro ostatní slovní druhy. Myslíme si, že identifikátor je součástí tektogramatické roviny. Samotné t-lema není jednoznačné, jak jsme poznali při problémech s přiřazením indexovaného m-lematu. Teprve rámec definitivně určí sémantický význam uzlu.

Další pomalým místem je interface na morfologický generátor a současné řešení fuzzy tagů. Nejen, že je potřeba využít vlastnosti generátoru a seznam všech možných fuzzy tagů posílat v jednom dotazu a omezit tak zbytečnou síťovou komunikaci. Myslíme si, že kód generující sadu fuzzy tagů logicky patří celý dovnitř morfologického generátoru.

Kapitola 4

Závěr

Zdá se nám, že na každé stránce naší práce je alespoň jedna poznámka pod čarou o tom, že zvolené řešení není ideální. I přesto si subjektivně myslíme, že výsledné věty jsou dobré. Utvrzuje nás v tom i objektivní kritérium: nárůst skóre oproti baseline.

Závěrem bych chtěl požádat generátor, aby se několika slovy pokusil zhodnotil budoucnost hráčů na poli strojového překladu.

ln94204_141.t.pls

19: V podmínkách rodícího se kapitalismu na mnohé z nich čeká krach.
G: V podmínkách rodícího kapitalizmu mnohé z nich čeká krach.

Literatura

[Tman–2005] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z. (2005): Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka, Prague.

[Tman–2000] Panevová, J., Böhmová, A., Hajičová, E., Sgall, P., Ceplová, M., Řezníčková, V. (2000): A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank, Prague.

[Mluvnice I–1986] Mluvnice češtiny, sv.1 (1986). Academia. Praha

[Mluvnice II–1986] Mluvnice češtiny, sv.2 (1986). Academia. Praha

[Mluvnice III–1987] Mluvnice češtiny, sv.3 (1987). Academia. Praha

[Čas a modalita–1971] Panevová, J., Benešová, E., Sgall, P. (1971): Čas a modalita v češtině. Universita Karlova. Praha.

[Panevová–1980] Panevová, J. (1980): Formy a funkce ve stavbě české věty. Academia. Praha.

[VALLEX–2003] Lopatková, M., Žabokrtský, Z., Skwarska, K., Benešová, V. (2003): VALLEX 1.0 Valency Lexicon of Czech Verbs, Prague.

[Razímová–2005] Razímová, Magda (2005): Meanings of Morphological Categories on the Tectogrammatical Level, In WDS'05 Proceedings of Contributed Papers, pp. 72-77 (eds. Jana Šafránková), MFF UK, Trója, Prague, Czech Rep., June 7-10.

[Sgall–1986] Sgall, P., Hajičová, E., Panevová, J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Academia. Praha.

[Hana, Hanová–2002] Hana, J., Hanová, H., Hajič, J., Vidová-Hladká, B., Jeřábek, E. (2002): Manual for Morphological Annotation, Prague.

[Kuryłowicz–1936] Dérivation lexicale et dérivation syntaxique. Bulletin de la Société de liguistique de Paris, 37, s. 79-92. Český překlad v: Principy strukturní syntaxe I. Praha, Univerzita Karlova, s. 87-94.

[ES–2002] Karlík, P., Nekula, M., Pleskalová, J. (editoři) (2002): Encyklopedický slovník češtiny. Lidové noviny, Praha.

[Petkevič–1998] Petkevič, V. (1995): Vocalization of Prepositions. In V. Petkevič (ed.): Linguistic Problems of Czech. Final Research Report for the JRP PECO 2824 project. Prague, 1995, pp. 147-157 (11 stran)

[Papineni-2001] Papineni, K., Roukos, S., Ward, T., Zhu, Wei-Jing (2001): BLEU – A Method for Automatic Evaluation of Machine Translation. IBM Research Report. New York.

Příloha A

Ukázky vygenerovaných vět

ln94207_84.t.gz:
soubor použitý pro vývoj: BLEU score = 0.5134, bez vallexu = 0.4788, baseline = 0.0326

1: Růžová bublina
B: bublina růžový
G: Růžová bublina.

2: Vybrané kapitoly z dějin žvýkačky
B: kapitola dějiny žvýkačka vybraný
G: Vybrané kapitoly z dějin žvýkačky.

3: Daniel Kumermann
B: Kumermann Daniel
G: Daniel Kumermann.

4: Který čuně zase?!
B: čuně který zase
G: Které čuně zase!

5: Řev rázu spíše symbolického:
B: řev ráz spíše symbolický
G: Řev spíše symbolického rázu.

6: Ten, kdo křičí, se totiž neptá.
B: totiž ten kdo křičet ptát se
G: Totiž se ten, kdo křičí, neptá.

7: Dobře ví, o koho jde.
B: vědět dobrý jít kdo
G: Dobře ví, o koho jde.

8: Stejně tak si je i adresát výtky podle ostrosti a výšky tónu okamžitě jist nejen tím, že jde o něj, ale i tím, co skandál vyvolalo.
B: stejně tak být i adresát výtka ostrost a výška tón okamžitý jistý si nejen jít ale i skandál vyvolat co
G: Stejně tak je i adresát výtky podle ostrosti a podle výšky tónu okamžitě jistý, nejen že jde o něj, ale i co skandál vyvolalo.

9: Jeho milovaný kousek žvýkací gumy, který si tak pečlivě odložil na spodek desky stolu, se stal kořistí nepřítele a asi jej čeká potupný konec v odpadním koši.
B: kousek guma žvýkací milovaný který odložit tak pečlivý spodek deska stůl stát se kořist nepřítel a čekat asi konec potupný koš odpadní
G: Jeho milovaný kousek žvýkací gumy, který si tak pečlivě odložil do spodku desky stolu, se stal kořistí nepřítele a jeho čeká asi potupný konec v odpadním koši.

10: Že by šel ještě vyndat?
B: jít ještě vyndat
G: , Že by šlo ještě jeho vyndat?

11: Ale ne, to už není ono...
B: ale ne, ten už být
G: Ale, ne, to už není ono.

12: Scéna číslo dvě: Kráva patří do chlíva.
B: scéna číslo dva: kráva patřit chlív
G: Scénu dvou čísla: , že kráva patří do chlív.

13: Lidi nežvýkají, to jenom krávy.
B: člověk žvýkat, ten žvýkat jenom kráva
G: Lidi nežvýkají, ten žvýkají jenom krávy.

14: Tak si miláčku laskavě rozmysli, jestli chceš patřit do téhle třídy, nebo se radši postěhuješ do chlíva?
B: tak miláček rozmyslet si laskavý patřit třída tenhle nebo raději stěhovat se chlív
G: Tak si, miláček, laskavě rozmysli, že chceš patřit do téhle třídy nebo se raději budeš stěhovat do chlív!

15: Opravdu volba!
B: opravdu volba
G: Opravdu volba!

16: Autorka nekompromisních slov se nad miláčkem obludně tyčí a on si může být zcela jist, že kdyby odpověděl po pravdě, skončil by s poznámkou v žákovské knížce.
B: kompromisní slovo autorka miláček tyčit se obludný a být jistý si zcela odpovědět pravda skončit poznámka knížka žákovský
G: Autorka nekompromisních slov se nad miláčkem obludně tyčí a může být zcela jistý, že když odpověděl pravdou, by skončil s poznámkou v žákovské knížce.

17: Nezbývá tedy než se opět pokořit a až na dno duše ukrýt vzdor spolu s vírou, že pravda jednou zvítězit musí.
B: tedy zbývat pokořit se opět a až dno duše ukrýt vzdor víra pravda jeden zvítězit
G: Tedy nezbývá pokořit se opět a až do dna duše ukrýt vzdor s vírou, že pravda jednou musí zvítězit.

18: Již staří Řekové...
B: Řek již starý
G: Již staří Řeci.

19: Ano, pravda.
B: ano, pravda
G: Ano, pravda.

20: Pravda o tom, že žvýkání pro žvýkání bylo odjakživa činností veskrze lidskou - kam paměť lidského rodu sahá.
B: pravda žvýkání žvýkání být odjakživa činnost lidský veskrze paměť rod lidský sahat kde
G: Pravda, že žvýkání pro žvýkání bylo odjakživa veskrze lidská činnost (kam paměť lidského rodu sahá).

21: Začněme co nejtradičněji: Již staří Řekové...
B: začít tradiční co: Řek již starý
G: Začněme co tradiční: již staří Řeci.

22: Filozofovali, řešili geometrické úlohy, bojovali s bohy i mezi sebou navzájem.
B: filozofovat řešit úloha geometrický, bojovat bůh bojovat i navzájem
G: Filozofovali, řešili geometrické úlohy, bojovali proti bohům, proti bojovali i navzájem.

23: To vše a mnohé jiné nás škola s radostí naučí.
B: radost který ten a mnohý jiný škola naučit
G: S radostí všechno to a jiné mnohé škola nás naučí.

24: Se stejnou radostí však zamlčí, že Řekové často a s oblibou žvýkali kousky ztuhlé mízy mastikového keře (pistacia lenticus), který se pěstuje především na ostrově Chios.
B: však radost stejný zamlčet Řek často a obliba žvýkat kousek míza ztuhlý keř mastikový pistacia lenticus který pěstovat především ostrov Chios
G: Však s radostí zamlčí, že Řeci často a s oblibou žvýkali kousky ztuhlé mízy mastikový keře (pistacia lenticus), který se pěstuje především v ostrově Chios.

25: Známý antický lékař a botanik Dioscorides psal v prvním století našeho letopočtu obsáhle o léčebném a hygienickém účinku žvýkání.
B: Dioscorides lékař a botanik antický známý psát obsáhlý století letopočet jeden účinek žvýkání léčebný a hygienický
G: Známý antický lékař a botanik Dioscorides obsáhle psal při prvním století našeho letopočtu o léčebném a hygienickém účinku žvýkání.

26: Pro historii žvýkací gumy, jak ji známe dnes, se však musíme přenést na jiný kontinent.
B: však historie guma žvýkací jak znát dnes přenést se kontinent jiný
G: Však se pro historii žvýkací gumy, jak ji známe dnes, musíme přenést do jiného kontinentu.

27: Keř, kterého si Kolumbus na ostrově Santo Domingo povšiml, je příbuzným řecké mastiky a jeho mízu místní Indiáni používali stejně jako Řekové.
B: keř který Kolumbus ostrov Domingo Santo povšimnout si být příbuzný mastika řecký a míza Indián místní používat stejný používat Řek
G: Keř, kterého si Kolumbus v ostrově Santo Domingo povšimnul, je příbuzný řecké mastika a jeho mízy místní Indiánové používali stejně, jak jí používali Řeci.

28: Zatímco karibští Indiáni strčili do úst kousek surové gumy v té podobě, jak jej utrhli od kůry, Mayové na poloostrově Yucatán přivedli žvýkání na vyšší úroveň.
B: Indián karibský strčit ústa kousek guma surový ten podoba jak utrhnout kůra Mayo poloostrov Yucatán přivést žvýkání úroveň vysoký
G: Zatímco karibští Indiánové strčili do úst kousek surové gumy tou podobou, jak jeho utrhnuli z kůry, Mayo v poloostrově Yucatán přivedli žvýkání vyšší úroveň.

29: Mízu stromu sapodilla (achras sapota) sklízeli a upravovali systémem, který se používá dodnes.
B: míza strom sapodilla sapota achras sklízet a upravovat systém který používat dodnes
G: Mízu stromu sapodilla sapota achras sklízeli a upravovali pomocí systému, který se používá dodnes.

30: Kůru stromu nařízli do tvaru písmene v a do špičky řezu umístili nádobu, do níž šťáva ukapávala.
B: strom kůra naříznout tvar písmeno v a řez špička umístit nádoba který šťáva ukapávat
G: Kůru stromu naříznuli tvarem písmene v a do špičky řezu umístili nádobu, do které šťáva ukapávat.

31: Získanou mléčnou gumovitou látku pak čistili, vařili.
B: látka gumovitý mléčný získaný potom čistit, vařit
G: Získanou mléčnou gumovitý látku potom čistili, vařili.

32: Teprve výsledný substrát byl hoden žvýkání.
B: substrát teprve výsledný být hodný žvýkání
G: Teprve výsledný substrát byl hodný žvýkání.

33: Zrodila se žvýkačka
B: zrodit se žvýkačka
G: Zrodila se žvýkačka.

34: Významnou roli v dějinách žvýkačky sehrál mexický diktátor Antonio Lopez de Santa Anna.
B: role žvýkačka dějiny významný sehrát Lopez de Santa Anna Antonio diktátor mexický
G: Role v dějinách žvýkačky sehrál mexický diktátor Antonio Anna Santa de Lopez.

35: Poté, co byl v roce 1845 jako prezident svržen a na deset let vypovězen na Kubu, vydal se do New Yorku s jedinou myšlenkou - získat zpět vládu nad Mexikem.
B: rok 1845 prezident svrhnout a rok deset vypovědět Kuba vydat se York New myšlenka jediný získat zpět vláda Mexiko
G: Co se v roce 1845 svrhnul a deset let se vypověděl do Kuby vydal se do New Yorku s jedinou myšlenkou získat zpět vládu Mexika.

36: K tomu jsou ovšem zapotřebí peníze, spousty peněz.
B: ovšem ten být zapotřebí peníze, spousta peníze
G: Ovšem pro to jsou zapotřebí peníze, spousty peněz.

37: Jednou z nejžádanějších komodit na světě byl v té době kaučuk, kterého nebylo dost a dovážel se z daleka.
B: jeden komodita žádaný svět ten doba být kaučuk který být dost a dovážet daleko
G: Jeden z komodit nejžádanějších ve světě v té době byl kaučuk, který dost nebyl a dovážel se z daleka.

38: Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle (prý z mayského slova tsictle), a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku.

B: tak Mexičan který i Anna Santa znát a občas žvýkat míza sapodilla zvaný chicle prý slovo tsictle mayský a tak zrodit se nápad pokusit se chicle udělat náhražka kaučuk
G: Tak ono každý Mexičan i Santa Anna znala a občas žvýkala mízu sapodilla zvanou chicle (prý z mayského slova tsictle) a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku.

39: Právě v té době přihrála náhoda Santa Annovi do cesty Thomase Adamse, fotografa a především vynálezce všeho druhu.
B: právě ten doba Santa Anna přihrát náhoda cesta Adams Thomas, fotograf a především vynálezce druh který
G: Právě v té době Santa Anna přihrála náhoda do cesty Thomase Adamse, fotografa a především vynálezce všeho druhu.

40: Oba si okamžitě padli do oka a dohoda byla jistá:
B: oba okamžitý padnout oko a dohoda být jistý
G: Oba si okamžitě padnuli do oka a dohoda byla jistá.

41: Santa Anna má chicle (zanedlouho jí nechal do New Yorku přivézt celou tunu) a Adams technické schopnosti.
B: Santa Anna mít chicle zanedlouho nechat New York přivézt tuna celý a Adams mít schopnost technický
G: Santa Anna má chicle (zanedlouho nechal do New Yorku přivézt celou tunu jí) a Adams má technické schopnosti.

42: Asi rok se Adams a jeho nejstarší syn snažili - chicle vařili, čistili, přidávali množství různých látek a míchali s pravým kaučukem.
B: Adams a syn starý rok asi snažit se - chicle vařit, čistit přidávat množství látka různý a míchat kaučuk pravý
G: Adams a jeho nejstarší se syn v asi roce snažili - chicle vařili, čistili, přidávali množství různých látek a je míchali s pravým kaučukem.

43: Vše s nulovým výsledkem.
B: který výsledek nulový
G: Všechno do nulového výsledku.

44: Když asi po roce své úsilí vzdali, rozhodl se Adams, že vše, co mu z chicle ještě zbylo, hodí do řeky.
B: rok asi úsilí vzdát Adams rozhodnout se co co chicle ještě zbýt hodit řeka
G: Když po asi roce svoje úsilí vzdali, Adams se rozhodnul, že se všechno, co z chicle jemu ještě zbylo, hodí do řeky.

45: Psal se rok 1869 a do hry vstoupila další náhoda:
B: psát rok 1869 a hra vstoupit náhoda další
G: Psal se rok 1869 a do hry vstoupila další náhoda.

46: Thomas Adams vstoupil do drogistického obchodu na rohu Broadwaye a Chambersovy ulice a spolu s ním tam byla i malá holčička, která požádala o žvýkací gumu za jednu penci.
B: Thomas Adams vstoupit obchod drogistický roh Broadway a ulice Chambers a tam být i holčička malý který požádat guma žvýkací pence jeden
G: Thomas Adams vstoupil do drogistického obchodu v rohu Broadway a Chambersovy ulice a tam s ním byla i malá holčička, která požádala o žvýkací gumu.

47: V tom okamžiku se Adamsovi zúročily roky vynalézání.
B: ten okamžik Adams zúročit se rok vynalézání
G: V tom okamžiku se pro Adamse zúročily roky vynalézání.

48: Vzpomněl si totiž, jak Santa Anna čas od času uloupl kus gumy, strčil do pusy a žvýkal.
B: totiž vzpomenout si Santa Anna čas čas uloupnout kus guma strčit pusa a žvýkat
G: Totiž si vzpomněl, že Santa Anna čas času uloupnout kus gumy, jeho strčila do pusy a jeho žvýkala.

49: Vyptal se proto prodavače, co to vlastně děvče chtělo.

B: proto vyptat se prodavač vlastně ten děvče chtít co
G: Proto se vyptat prodavače, co vlastně ten děvče chtělo.

50: Byl to parafin značky White Mountain, jehož žvýkání prý žádné veliké potěšení nepůsobí.
B: ten být parafin značka White Mountain který žvýkání který potěšení veliký působit prý
G: Ten byl parafin značky White Mountain, jehož žvýkání žádné veliké potěšení nepůsobí prý.

51: Tehdy se zrodila skutečná žvýkací guma.
B: tehdy zrodit se guma žvýkací skutečný
G: Tehdy se zrodila skutečná žvýkací guma.

52: Dvě stovky kuliček
B: stovka kulička dva
G: Dva stovky kuliček.

53: Adams se vrátil domů a ještě tentýž večer vyrobil se svými čtyřmi syny dvě stovky gumových kuliček bělošedavé barvy.
B: Adams vrátit se domů a ještě tentýž večer vyrobit syn čtyři stovka kulička gumový barva bělošedavý dva
G: Adams se vrátil domů a ještě při tomtéž večeru vyrobil se svými čtyřmi syny dvě stovky gumových kuliček bělošedavý barvy.

54: Druhý den je vzal ke svému příteli drogistovi z Jersey City s tím, že by mu mohly vydržet na pár měsíců.
B: den dva vzít přítel drogista City Jersey vydržet pár měsíc
G: Druhého dne je vzal ke svému příteli drogista z Cit Jersey, že by jemu mohly vydržet pár měsíců.

55: Nevydržely, celá dávka se při ceně dvě kuličky za cent vyprodala během dopoledne.
B: vydržet, dávka celý cena kulička dva cent vyprodat dopoledne
G: Nevydržely, celá dávka se vyprodala během dopoledne.

56: Adamsova rodina dala dohromady své veškeré jmění - podle některých zdrojů 35, podle jiných 55 dolarů - a zrodil se žvýkačkový průmysl.
B: Adams rodina dát dohromady jmění který rodina dohromady zdroj který dát dolar 35, jiný zdroj dát dolar 55 a zrodit se průmysl žvýkačkový
G: Adamsova rodina dala dohromady svoje všechno jmění (rodina dohromady podle některých zdrojů mohla dát 35 dolarů, podle jiných zdrojů mohla dát 55 dolarů.) a zrodil se žvýkačkový průmysl.

57: Když o deset let později obrátil ke gumě pozornost louisvilleský lékárník John Colgan, existovala již řada žvýkačkových milionářů (mezi nimi Adams).
B: pozdě rok deset guma obrátit pozornost Colgan John lékárník louisvilleský existovat již řada milionář žvýkačkový Adams
G: Když o deset let později ke gumě obrátil pozornost louisvilleský lékárník John Colgan, existovala již řada žvýkačkový milionářů (mezi nimi Adams).

58: Přesto však byly dveře pro zlepšovatele otevřeny dokořán, většina gumy byla stále ještě jen povrchově oslazený či ochucený kousek chicle.
B: přesto však dveře zlepšovatel otevřít dokořán, guma většina být stálý ještě jen kousek chicle oslazený či ochucený povrchový
G: Přesto se však dvéře pro zlepšovatel otevřely dokořán, většina gumy byla stále ještě jen povrchově oslazený či ochucený kousek chicle.

59: Colgan použil aromatický balzám ze stromu tolu, který sloužil jako základ sirupů proti kašli, a smíchal jej s gumou.
B: Colgan použít balzám aromatický strom tolu který sloužit základ sirup kašel a smíchat guma
G: Colgan použil aromatického balzámu ze stromu tolu, který sloužil a jeho smíchal s gumou.

60: Výsledek, který dostal obchodní jméno Taffy Tolu , se ujal okamžitě, Colgan zavřel lékárnu a během krátké doby se přidal k milionářskému klubu.
B: výsledek který dostat jméno Taffy Tolu obchodní ujmout se okamžitý Colgan zavřít lékárna a doba krátký přidat se klub milionářský

G: Výsledek, který dostal obchodní jméno Taffy Tolu, se ujmul okamžitě, Colgan zavřel lékárnu a během krátké doby se přidal k milionářskému klubu.

61: Vousáč místo vepříka
B: vousáč vepřík
G: Vousáč.

62: Podobnou cestu nastoupil v roce 1881 jiný lékárník Edward E. Beeman z Clevelandu.
B: podobný cesta nastoupit rok 1881 Beeman Edward E lékárník jiný Cleveland
G: Podobnou cestu nastoupil v roce 1881 jiný lékárník E Edward Beeman z Clevelandu.

63: Nejen to.
B: nejen ten
G: Nejen to.

64: Jako první pozdvihl dosud jen předmět zábavy na vyšší rovinu.
B: jeden pozdvihnout dosud předmět jen zábava rovina vysoký
G: Pozdvihl dosud předmět jen zábavy do vyšší roviny.

65: Svou předchozí úspěšnou praxi postavil na pepsinovém prášku, který zlepšoval zažívání a který sám objevil.
B: praxe úspěšný předchozí postavit prášek pepsinový který zlepšovat zažívání a který sám objevit
G: Svoji předchozí úspěšnou praxi postavil na pepsinový prášku, který zlepšoval zažívání a který objevil.

66: Ještě úspěšnější kariéru výrobce žvýkaček založil na přidávání pepsinového prášku do gumy.
B: kariéra výrobce žvýkačka ještě úspěšný založit přidávání prášek pepsinový guma
G: Ještě úspěšnější karieru vyrábitele žvýkaček založil na přidávání pepsinový prášku do gumy.

67: Jeho výrobek se však skutečně rozšířil až v okamžiku, kdy nahradil obrázek vepříka na obalu vlastní důstojnou vousatou tváří.
B: však skutečně výrobek rozšířit se až okamžik kdy nahradit obrázek vepřík obal vlastní tvář vousatý důstojný
G: Však se skutečně jeho výrobek rozšířil až v okamžiku, kdy nahradil obrázek vepřík v obale pomocí důstojné vousaté vlastní tváře.

68: Díky tomu se stal jedním z nejznámějších lidí Ameriky.
B: ten stát se jeden člověk známý Amerika
G: Kvůli tomu se stal jeden z nejznámějších lidí Ameriky.

69: Tehdy došlo ve stejném městě k podstatnému technickému průlomu:
B: tehdy dojít město stejný průlom technický podstatný
G: Tehdy došlo v stejném městě k podstatnému technickému průlomu.

70: William J. White, obchodník s popcornem, vyřešil problém, jak do chicle dostat libovolnou příchuť a především jak ji v ní udržet.
B: White William J, obchodník popcorn vyřešit problém chicle dostat jak příchuť libovolný a především udržet jak
G: J William White, obchodník popcorn vyřešil problém jak do chicle dostat libovolnou příchuť a především jak ji v ní udržet.

71: Zjistil, že se guma dobře váže s kukuřičným sirupem, který je schopen absorbovat prakticky jakoukoliv chuťovou esenci.
B: zjistit guma vázat se dobrý sirup kukuřičný který být schopný absorbovat esence chuťový prakticky jaký
G: Zjistil, že se guma dobře víže s kukuřičným sirupem, který je schopný absorbovat jakoukoliv chuťovou esenci.

72: První, kterou on sám přivedl na trh, byla peprmintová.
B: jeden esence který sám přivést trh být peprmintový

G: První esence, kterou sám on přivedl do trhu, byla peprmintový.

73: V rámci reklamní kampaně předal osobně každému členu washingtonského Kongresu jednu krabici své gumy značky Yucatan.
B: kampaň reklamní předat osobní člen Kongres washingtonský který krabice guma značka Yucatan jeden
G: Na reklamní kampani osobně předal každému členovi washingtonského Kongresu jednu krabici svojí gumy značky Yucatan.

74: Tak se mu tam zalíbilo, že o něco později do Kongresu sám úspěšně kandidoval.
B: tam zalíbit se tak co pozdě Kongres sám kandidovat úspěšný
G: Tam se jemu zalíbilo tak, že o něco později do Kongresu úspěšně kandidoval.

75: Jeho postavení mu při pozdější návštěvě Velké Británie umožnilo audienci u krále Edwarda VII.
B: postavení návštěva Británie Velký pozdní umožnit audience Edward VII král
G: Jeho postavení jemu při pozdnější návštěvě Velké Británie umožnilo audienci u krále VII Edwarda.

76: I král dostal svou krabici gumy a nádavkem i doslova trhoveckou prezentaci.
B: i král dostat krabice guma a dostat nádavek i prezentace doslova trhovecký
G: I král dostal svoji krabici gumy a dostal nádavkem i trhovecký prezentaci.

77: Samozřejmě, že novinové zprávy o králi s krabicí gumy byly reklamou k nezaplacení.
B: samozřejmě zpráva král krabice guma novinový být reklama zaplacení
G: Samozřejmě novinové zprávy o králi s krabicí gumou byly reklama s ohledem na zaplacení.

78: I když konzervativní Anglie jeho čin odsoudila, guma se zde chytila a Británie se pro žvýkačku stala bránou do Evropy.
B: Anglie konzervativní čin odsoudit guma tady chytit se a Británie žvýkačka stát se brána Evropa
G: Ač konservativní Anglie jeho čin odsoudila, guma se tady chytila a Británie se pro žvýkačku stala bránou do Evropy.

79: Pokusy a omyly
B: pokus a omyl
G: Pokusy a omyly.

80: Ještě jeden milník si zaslouží zmínku - zrod bublinové žvýkačky.
B: zasloužit si zmínka milník ještě jeden - zrod žvýkačka bublinový
G: Zaslouží si zmínku ještě jeden milník - zrod bublinový žvýkačky.

81: Objevila se v roce 1906, kdy Frank Henry Fleer vyrobil první syntetickou gumovou bázi, která umožnila vyfouknout bublinu.
B: objevit se rok 1906 Fleer Frank Henry kdy vyrobit báze gumový syntetický jeden který umožnit vyfouknout bublina
G: Objevila se v roce 1906, kdy Henry Frank Fleer vyrobil první syntetickou gumovou bázi, která umožnila vyfouknout bublinu.

82: Vyfouknout ano.
B: vyfouknout ano
G: Vyfouknout ano.

83: Co však následovalo, bylo neštěstí.
B: však co následovat být neštěstí
G: Však, co následovalo, bylo neštěstí.

84: Bublina totiž snadno praskla a guma byla velice lepkavá - k dětskému obličeji lnula takřka dokonale.
B: totiž bublina prasknout snadný a guma být lepkavý velice - obličej dětský lnout dokonalý takřka
G: Totiž bublina snadno prasknula a guma byla velice lepkavá - k dětskému obličeji takřka dokonale lnula.

85: Trvalo to až do roku 1928, než se tento problém podařilo překonat.
B: trvat až rok 1928 podařit se tento problém překonat

G: Trvalo až do roku 1928, že se podařilo tento problém překonat.

86: Mladý účetní z Fleerovy firmy Walter Diemer, inspirován svým šéfem, míchal ve volném čase gumu s rozmanitými ingrediencemi metodou pokusu a omylu (jeho znalosti chemie byly nulové).
B: účetní mladý Fleer firma Walter Diemer inspirovat šéf míchat metoda pokus a omyl čas volný guma ingredience rozmanitý znalost chemie být nulový
G: Mladý účetní z Fleerovy firmy Walter Diemer inspirovat svůj šéf míchal methodou pokusu a omylu při volném čase gumu s rozmanitými ingrediencemi (jeho znalosti o chemii byly nulové)

87: Trvalo mu čtyři měsíce, než svůj objev dokázal stabilizovat tak, že byl průmyslově použitelný.
B: trvat měsíc čtyři dokázat objev stabilizovat tak být použitelný průmyslový
G: Jemu trvalo čtyři měsíce, že dokázal svůj objev stabilizovat tak, že byl průmyslově použitelný.

88: V den, kdy svůj výrobek představil vedení podniku, měl k dispozici pouze růžovou potravinářskou barvu.
B: den kdy výrobek představit podnik vedení mít dispozice pouze barva potravinářský růžový
G: Dne, kdy svůj výrobek představil vedení podniku, měl k dizpozici pouze růžovou potravinářskou barvu.

89: Takže dnes nejcharakterističtější žvýkačková barva je ryzí náhodou.
B: takže barva žvýkačkový dnes charakteristický být náhoda ryzí
G: Takže dnes nejcharakterističtější žvýkačkový barva je ryzí náhoda.

90: Kráva nebo drak?
B: kráva nebo drak
G: Kráva nebo drak?

91: Koncem roku 1928 byly tedy položeny všechny základy současné světové žvýkačkové produkce - vlastní gumu dnes nevyrábějí jen některé z nejmenších státečků světa;
B: tedy rok 1928 položit základ produkce žvýkačkový světový současný který - guma vlastní dnes vyrábět jen který státeček malý svět
G: Tedy se konec roku 1928 položily všechny základy současné světové žvýkačkový produkce - vlastní gumu dnes nevyrábějí jen některé z nejmenších státečků světa.

92: případně ty, v nichž je žvýkačka chápána jako ideologická diverze.
B: případně vyrábět ten který žvýkačka chápat diverze ideologický
G: Případně nevyrábějí ty, v kterých se žvýkačka chápe ideologickou diverze.

93: Všechny technologické změny od té doby jsou již jen menší úpravy.
B: změna technologický ten doba který být již jen úprava malý
G: Všechny technologické změny od té doby jsou již jen malejší úpravy.

94: Nutno zdůraznit slovo technologické.
B: nutný zdůraznit slovo technologický
G: Nutné zdůraznit slovo technologické.

95: Změny struktury trhu, které přicházejí jako odpověď na konkrétní společenskou objednávku, bývají totiž často významné.
B: totiž změna struktura trh který přicházet odpověď objednávka společenský konkrétní být často významný
G: Totiž změny struktury trhu, které přicházejí, jsou často významné.

96: Jde především o zdravotní (či pseudozdravotní) námitky proti žvýkání.
B: jít především námitka zdravotní či námitka pseudozdravotní žvýkání
G: Jde především o zdravotní námitky či o pseudozdravotní námitky proti žvýkání.

97: Tento boj začal již počátkem století značkou Dentyne.
B: tento boj začít již století značka Dentyne
G: Tento boj začal již počátek století pomocí značky Dentyne.

98: Tehdy ještě obsahovala malé množství cukru, v současnosti tatáž guma nemá cukr žádný.
B: tehdy ještě obsahovat množství cukr malý, současnost tentýž guma mít cukr který

G: Tehdy ještě obsahovala malé množství cukru, při současnosti tatáž guma nemá žádný cukr.

99: Mnohé dnešní žvýkačky obsahující cukr mívají i přídavek uhličitanu vápenatého, který neutralizuje ústní kyseliny.
B: žvýkačka dnešní obsahující cukr mnohý mít i přídavek uhličitan vápenatý který neutralizovat kyselina ústní
G: Mnohé dnešní žvýkačky obsahující cukr mají i přídavek vápenatého uhličitanu, který neutralizuje ústní kyseliny.

100: Máte problémy s plombami, žádný problém, řada firem již vyrábí nepřilnavou gumu.
B: mít problém plomba problém který, řada firma již vyrábět guma přilnavý
G: Máte problémy s plombami, žádný problém, řada firem již vyrábí přilnavý gumu.

101: O kladném vlivu žvýkaček na psychiku dnes již lékaři vesměs nepochybují.
B: žvýkačka vliv psychika kladný lékař dnes již pochybovat vesměs
G: O kladném vlivu žvýkaček na psychiku lékaři dnes již vesměs nepochybují.

102: Podle průzkumů uskutečněných v řadě zemí se většina dentistů domnívá, že při vhodné volbě zaručují žvýkačky ústní hygienu, a proto je jejich vliv spíše pozitivní.
B: průzkum uskutečněný řada země většina dentista domnívat se volba vhodný žvýkačka zaručovat hygiena ústní a proto vliv být spíše pozitivní
G: Podle průzkumů uskutečněných v řadě zemích se většina dentista domnívá, že při vhodné volbě žvýkačky zaručují ústní hygienu a proto jejich vliv je spíše pozitivní.

103: Nemluvě o tom, kdy přímo slouží k rovnoměrnému vstřebávání léku, či dokonce k odnaučování kouření.
B: nemluvě přímo kdy sloužit vstřebávání lék rovnoměrný či dokonce odnaučování kouření
G: .

104: A tu se dostáváme zpět k počátku tohoto textu.
B: a tady dostávat se zpět počátek tento text
G: A tady se dostáváme zpět k počátku tohoto textu.

105: Všimli jste si někdy, že velká většina skvělých učitelů, kteří ve spojení se žvýkačkou tak rádi mluví o dobytku, vyznává estetiku kouře vycházejícího z úst?
B: všimnout si kdy většina velký učitel skvělý který žvýkačka mluvit tak rád dobytek vyznávat estetika kouř vycházející ústa
G: Všimnuli jste si někdy, že velká většina skvělých učitelů, kteří se žvýkačkou mluví tak rádi o dobytkovi, vyznává estetiku kouře vycházejícího z úst?

106: Krávou tedy člověk být nesmí, drakem však ano...
B: tedy člověk kráva být však drak být ano
G: Tedy člověk kráva nesmí být, však drak smí být ano.

107: Foto archív
B: foto archív
G: Foto archív.

ln95049_013.t.gz:
soubor zvolený nepřítelem, BLEU score = 0.4948, bez vallexu = 0.4169, baseline = 0.0000

1: Rasisté útočí ve skupinách
B: rasista útočit skupina
G: Rasisté útočí ve skupinách.

2: Praha, Plzeň (toh, vet) -
B: Praha, Plzeň toh, vet
G: Praha, Plzeň.

3: Pod vlivem rasové nesnášenlivosti páchají v ČR trestnou činnost hlavně organizované skupiny.
B: vliv snášenlivost rasový páchat ČR činnost trestný hlavně skupina organizovaný
G: Vlivem rasové snášenlivosti pášou v ČR trestnou činnost hlavně organizované skupiny.

4: Nejčastěji takto útočí deseti až třicetičlenné party pachatelů.
B: často takto útočit pachatel parta deset až třicetičlenný
G: Nejčastěji takto útočí deset part až třicetičlenný part.

5: LN to řekla poradkyně náměstka ministra vnitra pro oblast bezpečnosti Jitka Gjuričová, autorka zprávy "Interetnické konflikty".
B: LN ten říci Gjuričová Jitka poradkyně náměstek ministr vnitro oblast bezpečnost, autorka zpráva konflikt interetnický
G: LN ten řekla Jitka Gjuričová poradkyně náměstka ministra vnitra pro oblast bezpečnosti, autorka zprávy interetnický konflikty.

6: Právě tento čtvrtek napadla čtyřčlenná skupina mladíků s obušky a sečnou zbraní podobnou mačetě na Centrálním autobusovém nádraží v Plzni dva Romy.
B: právě tento čtvrtek napadnout skupina mladík obušek a zbraň sečný podobný mačeta čtyřčlenný nádraží autobusový centrální Plzeň Rom dva
G: Právě v tento čtvrtek napadnula čtyřčlenná skupina mladíků s obušky a se sečný zbraní podobnou mačetě v centrálním autobusovém nádraží v Plzni dva Romy.

7: Při útoku utržil dvaadvacetiletý Rom sečnou ránu do krku.
B: útok utržit Rom dvaadvacetiletý rána sečný krk
G: Při útoku utržil dvaadvacetiletý Rom sečný ránu do krku.

8: Přihlížející přivolali lékaře a ihned poté policii.
B: přihlížející přivolat lékař a potom ihned přivolat policie
G: Přihlížející přivolali lékaře a potom ihned přivolali policii.

9: Násilníci však stačili z místa činu uprchnout.
B: násilník však stačit čin místo uprchnout
G: Násilníci však stačili z místa činu uprchnout.

10: Podle svědků se jednalo o šestnácti až devatenáctileté příznivce hnutí skinheads.
B: svědek jednat se příznivec hnutí skinheads šestnáct až devatenáctiletý
G: Podle svědků se jednalo o šestnáct až devatenáctileté příznivce hnutí skinheads.

11: "Soudy musejí v takových případech prokázat, že se dav pachatelů skutečně dopustil rasistického útoku.
B: soud takový případ prokázat pachatel dav dopustit se skutečně útok rasistický
G: "Soudy musejí prokázat, že se dav pachatelů dopustil skutečně rasistického útoku."

12: To je - vzhledem k množství pachatelů - velmi obtížné, někdy skoro nemožné," tvrdí Gjuričová.
B: Gjuričová tvrdit ten být pachatel množství obtížný velmi, kdy být možný skoro
G: Gjuričová tvrdí: "To je (s ohledem na množství pachatelů) velmi obtížné, někdy je skorem nemožné."

13: Upozorňuje přitom na známý "písecký případ", kdy skupina osmnácti skinů zahnala do řeky Otavy čtveřici mladých Romů a nedovolila jim z vody vylézt.

B: přitom upozorňovat případ písecký známý skupina skin osmnáct kdy zahnat řeka Otava čtveřice Rom mladý a dovolit voda vylézt
G: Přitom upozorňuje na známý písecký případ, kdy skupina osmnácti skinů zahnala do řeky Otava čtveřici mladých Romů a jim nedovolila z vody vylézt.

14: Osmnáctiletý Rom Tibor Danihel se utopil.
B: Danihel Tibor Rom osmnáctiletý utopit se
G: Osmnáctiletý Rom Tibor Danihel se utopil.

15: Dva ze skinheadů byli odsouzeni k ročnímu podmíněnému trestu odnětí svobody.
B: skinhead dva odsoudit trest podmíněný roční odnětí svoboda
G: Dva ze skinheadů se odsoudili k ročnímu podmíněnému trestu odnětí svobody.

16: Ostatní útočníci byli osvobozeni.
B: útočník ostatní osvobodit
G: Ostatní útočníci se osvobodili.

17: Obhájci při procesu argumentovali tím, že příslušnost k hnutí skinheads nelze určit podle vnějších znaků - účesu či oblečení.
B: obhájce proces argumentovat lze skinheads hnutí příslušnost určit - znak vnější účes či oblečení
G: Obhájci při procese argumentovali, že nelze příslušnost k hnutí skinheads určit podle vnějších znaků - podle účesu či podle oblečení.

18: Gjuričová má za to, že i když je podíl rasově motivovaných trestných činů na celkové kriminalitě nevýznamný, "svým dopadem a obecnými následky jde o významný jev se společenskými a politickými dopady i mimo území ČR".
B: Gjuričová mít ten rasový motivovaný trestný čin podíl kriminalita celkový být významný dopad a následek obecný jít jev významný dopad i společenský a politický území ČR
G: Gjuričová má za ten, že ač podíl rasově motivovaných trestných činů na celkové kriminalitě je nevýznamný, s ohledem na svůj dopad a s ohledem na obecné následky jde o významný jev s i společenskými a politickými dopady mimo území ČR.

19: Podle autorky zprávy "Interetnické konflikty" tvoří převážnou část pachatelů takových trestných činů členové hnutí skinheads nebo osoby s nimi sympatizující.
B: interetnický konflikt zpráva autorka tvořit takový trestný čin pachatel část převážný člen hnutí skinheads nebo osoba sympatizující
G: Podle autorky zprávy interetnický konflikty tvoří převážnou část pachatelů takových trestných činů členové hnutí skinheads nebo s nimi sympatizující osoby.

20: Poškozenými jsou téměř vždy Romové.
B: poškozený být kdy téměř Rom
G: Poškození jsou téměř vždy Romové.

Příloha B

Ukázky výstupu generátoru číslovek

Zlomkové

plurál, instrumentál, neutrum

1 celek
8 osminami
15 patnáctinami
22 dvaadvacetinami
29 devětadvacetinami
36 šestatřicetinami
50 padesátinami
71 jednasedmdesátinami
106 stošestinami
120 stodvacetinami

Řadové

singulár, dativ, feminimum

1 první
124 sto čtyřiadvacáté
247 dvě stě sedmačtyřicáté
493 čtyři sta třiadevadesáté
616 šest set šestnácté
985 devět set pětaosmdesáté
1108 tisíc sto osmé
1231 tisíc dvě stě jednatřicáté
1477 tisíc čtyři sta sedmasedmdesáté
1600 tisíc šesti stému
1846 tisíc osm set šestačtyřicáté
1969 tisíc devět set devětašedesáté
2092 dva tisíce dvaadevadesáté
2338 dva tisíce tři sta osmatřicáté