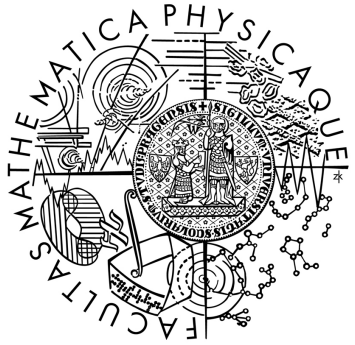
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

**DIPLOMOVÁ PRÁCE**




Marie Konárová

# Školní větné rozbory jako možný zdroj závislostních korpusů (?)

Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: Mgr. Barbora Vidová Hladká, PhD.
Studijní program: Informatika
Studijní obor: Matematická lingvistika

Praha 2011

Děkuji vedoucí své diplomové práce Mgr. Barboře Vidové Hladké, PhD., za cenné připomínky a rady udělované v průběhu práce. Dále děkuji všem, kteří se podíleli na podkladech pro tuto diplomovou práci. Jmenovitě RNDr. Jiřímu Hanovi, PhD., za vytvoření editoru Čapek, jeho úpravy a konzultace k práci s tímto editorem, PhDr. Zdeňce Urešové, PeaDr. Věře Čechákové, PhDr. Jitce Hošnové, Ester Sgallové a Tereze Bártové za kompletní označkování zkoumaných vět, Bc. Tomáši Maiserovi, Ing. Luboši Čechovi a Mgr. Ondřeji Konárovi za technickou podporu. Mé poděkování platí rovněž všem učitelům i žákům, kteří se zúčastnili sběru dat.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 9. prosince 2011 Marie Konárová

Název práce: Školní větné rozbory jako možný zdroj závislostních korpusů (?)

Autor: Marie Konárová

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: Mgr. Barbora Vidová Hladká Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Cílem práce je prozkoumat možnosti využití dat ze školních větných rozborů pro značkování slov v jazykových korpusech. Za účelem ověření této hypotézy byla vybrána množina vět, které byly předloženy žákům základních a středních škol k větnému rozboru. Sběr dat probíhal s využitím funkčního prototypu editoru větných rozborů Čapek. Editor je stále vyvíjen, mimo jiné i na základě zpětné vazby získané při jeho používání žáky i učiteli. Na základě nasbíraných dat byla odvozena transformační pravidla pro konverzi údajů ze školních větných rozborů do datových struktur využívaných Pražským závislostním korpusem. Byla testována jak úspěšnost konverze pomocí navržených pravidel, tak přesnost žáků při provádění větných rozborů.

Klíčová slova: školní větné rozbory, syntaktická analýza, korpusy

Title: A school analysis as a possible source of treebanks (?)

Author: Marie Konárová

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Barbora Vidová Hladká Ph.D., Institute of Formal and Applied Linguistics

Abstract: The aim of this thesis is to explore the possibilities of using data from the school sentence analyses for tagging words in the language corpora. For testing of this hypothesis, a set of sentences has been selected from a common czech language textbook. Students of selected primary and secondary schools were asked to perform the syntactical analysis of these sentences. The data collection was carried out using a prototype sentence analysis editor Capek. The editor is still being developed, also based on feedback gained from the students and teachers who used it during the data collecting process. Several transformation rules for converting data from the school sentence analyses into the data structures used within the Prague Dependency corpus were developed. The accuracy of the conversion using the proposed rules was tested together with the accuracy of students' results.

Keywords: a school analysis, syntactical analysis, treebanks

# Obsah

# Úvod

Jednou z velmi náročných a přitom velmi důležitých úloh v korpusové lingvistice je ruční značkování lingvistických korpusů. Ruční značkování je velice drahou a časově náročnou operací, takže se hledají alternativní metody značkování, které jsou méně časově i finančně náročné. Zatímco hledáme, jakým způsobem co nejlépe a nejlevněji označkovat text, žáci základních a středních škol běžně v rámci výuky zpracovávají jazykové rozbory, což přibližně odpovídá morfologickému a analytickému značkování v korpusech.

Důvodem, proč lingvisté potřebují co nejvíce značkovaných dat, jsou mimo jiné statistické automatické překlady. Na základě značkovaných dat jsou tvořeny jazykové modely, pomocí nichž jsou pak ohodnocovány navržené překlady konkrétní věty a vybírány nejvhodnější z nich.

Cílem této diplomové práce je zjistit, zda by nebylo možné využít větné rozbory zpracovávané žáky během hodin ve škole k alternativnímu způsobu značkování závislostního korpusu. Práce volně navazuje na diplomovou práci Ondřeje Kučery [8], která se zabývala transformací dat z Pražského závislostního korpusu (PDT) do podoby cvičebnice větných rozborů a popularizací PDT.

Pro účely testování byli vybráni žáci šestých až devátých tříd a odpovídajících tříd víceletých gymnázií. Větné rozbory jsou získávány přímo při vyučování na školách a transformovány do formátu používaného v PDT (dále jako formátu PDT), konkrétně do morfologické a analytické roviny. Sběr dat pro experiment probíhal s využitím funkčního prototypu programu Čapek.

V kapitole 2 jsou popsány ideje, nástroje a jazyky, které tvoří podklady k této diplomové práci. Je zmíněn Pražský závislostní korpus ve verzi 2.0, systém STYX a podrobněji popsána aktuální verze editoru Čapek.

V kapitole 3 se mimo jiné nacházejí podrobnosti o výběru testovacích vět z učebnice pro základní školy. Dále jsou v ní podrobně popsány zkušenosti a zážitky z hodin na školách i z individuální práce na rozborech. V neposlední řadě jsou v ní popsány výsledky dotazníků předkládaných zúčastněným učitelům a jejich žákům. Část této kapitoly se věnuje i obsahu motivační prezentace pro žáky, která jim byla přednášena před začátkem práce na větách.

Aby bylo možné testovat využitelnost tohoto způsobu značkování, jsou získané rozbory převáděny pomocí transformací do formátu PDT. Tento převod je popsán v kapitole 4. Zabývám se převodem ve třech víceméně oddělených oblastech. Nejprve získám žádaný tvar stromu. Na školách se na rozdíl od značkování závislostní syntaxe pro lingvistické účely (např. v PDT) sdružují slova do větných členů, které jsou pak mezi sebou propojovány. Pro účely testování je tedy třeba odhadnout závislost mezi slovy ve větném členu, čemuž se podrobně věnuje odstavec 4.1. Dále je třeba převádět morfologické (slovní druhy a jejich kategorie) a analytické funkce (druhy větných členů).

Transformované rozbory vyhodnocuji vzhledem k anotovaným větám ve formátu PDT[1]. Metoda vyhodnocování a vyhodnocení samotné je popsáno v kapitole 5. Zajímavou informací v této kapitole jsou i statistiky odlišnosti zpracování rozborů žáky a učiteli. Tedy jak moc „správně“ žáci rozbory zpracovali.

Implementace je provedena v jazycích Perl a SQL na databázovém serveru

[1] Věty zpracovala dle PDT koncepce Zdeňka Urešová.

MySQL. Podrobnosti k používání napsaných skriptů jsou uvedeny v příloze E. Vybrané detaily implementace jsou uvedeny v příloze F.

# 1. Školní větné rozbory ve světě

Při provádění rešerše dostupných zdrojů nebyla objevena zmínka o používání školních větných rozborů pro anotaci korpusů. To nemusí nutně znamenat, že se nikdo této problematice ve světě nevěnuje. Může se jednat o novou problematiku, k níž zatím nejsou snadno dohledatelné publikace.

Cílem rešerše je proto zmapování provádění školních větných rozborů ve světě. Pak budeme mít představu o tom, pro které jazyky kromě češtiny by se případně dala vyvinutá metodika použít. Vzhledem k tomu, že informace o vzdělávacích programech, případně osnovách, jsou dostupné typicky pouze v úředním jazyce dané země, je rešerše omezena na jazyky, jimž bylo možné alespoň rámcově porozumět.

Ve školách na Slovensku se vyučuje jak morfologie, tak syntax, jak dokládají například osnovy [22], nebo školní vzdělávací program vybrané školy [21]. Vzhledem k dlouhodobé společné historii se dá navíc očekávat, že přístup k větným rozborům na Slovensku bude podobný jako v České republice.

Poněkud pesimističtější jsou poznatky o větných rozborech v angličtině. Podle [7] nebyla v Anglii gramatika prakticky vyučována po celou první polovinu 20. století. Od roku 1960 dochází k postupnému zahrnování výuky gramatiky do školních osnov, nicméně hlavní důraz se klade na užití jazyka. Osnovy [19] jasně ukazují orientaci na tři základní složky – mluvení a poslech, čtení, psaní. Jazykových rozborů se tedy pravděpodobně na anglických školách ještě nějakou dobu nedočkáme.

Ve francouzštině mají větné rozbory rozhodně své místo, jak ukazuje například elektronická učebnice [14]. Otázkou je, zda a v jaké míře se větné rozbory vyučují na základních nebo středních školách. Získání této informace bohužel nebylo snadné. Ve Francii jsou (podobně jako dnes u nás) rámcové vzdělávací plány, které nejsou natolik konkrétní, aby bylo možno vyčíst, zda se vyučují větné rozbory. Jistou nápovědu přináší webová stránka [20] pro studenty základních škol, která uvádí odkaz na elektronickou cvičebnici větných rozborů jako materiál pro šestý ročník. Vzhledem k tomu, že podobných materiálů se vyskytuje mnoho, lze se domnívat, že školní větné rozbory jsou běžnou součástí výuky francouzštiny.

V Německu se větné rozbory vyučují minimálně na gymnáziích, jak je uvedeno v osnovách [18]. Na základních školách se rozbory také řeší, pravděpodobně však v jednodušší formě, jak napovídají osnovy [17].

Ve Švýcarsku se pravděpodobně větné rozbory vyučují až na úrovni středních škol. Zatímco osnovy pro obchodní akademie v kantonu St. Gallen [15] zahrnují větné rozbory, osnovy pro základní školy v kantonu Zürich [16] obsahují orientaci na čtení, psaní, mluvení, podobně jako v anglických školách.

Uvedené zdroje prokázaly, že školní větné rozbory nejsou jen lokální záležitostí. Přinejmenším v rámci Evropy se školní větné rozbory provádějí. V anglických školách nejsou větné rozbory součástí běžné výuky, což může vést ke skeptickému postoji anglicky hovořící vědecké obce. Značkování anglických korpusů se navíc věnuje velká část lingvistické komunity. Problematika užití školních větných rozborů pro značkování korpusů bude tedy spíše zajímavá pro ostatní evropské jazyky.

# 2. Externí datové zdroje a nástroje

Tato kapitola obsahuje informace o projektech, systémech a strukturách, které byly použity pro sběr větných rozborů ve školách. Dále jsou uvedeny podrobnosti o systémech, na které tato práce navazuje, nebo by měla být jejich součástí. Zmíněn je **Pražský závislostní korpus**, do jehož formátu větné rozbory transformujeme, systém **STYX**, který využívá vět z korpusu jako cvičebnice pro školy, a editor **Čapek**, v němž je sběr větných rozborů prováděn.

## 2.1 Pražský závislostní korpus

Pražský závislostní korpus (dále jen PDT) je systémem ručně značkovaných českých textů. Texty v korpusu jsou nezkrácenými články z vybraných ročníků Lidových novin, Mladé fronty Dnes, Českomoravského Profitu a časopisu Vesmír.

Aktuální verze korpusu má označení PDT 2.0 a obsahuje ruční značkování morfologie, povrchové syntaxe, hloubkové syntaxe a sémantiky, aktuálního členění, koreference a lexikální sémantiky (viz [2]). Tato značkování jsou rozdělena do čtyř rovin, které jsou vzájemně provázány. PDT 2.0 obsahuje slovní rovinu, rovinu s morfologickou informací (informací o slovním druhu a jeho kategoriích), analytickou rovinu, obsahující závislosti a vztahy mezi jednotlivými slovy, a tektogramatickou rovinu, zachycující význam věty.

PDT 2.0 obsahuje přibližně 2 milióny slov, přičemž množství označkovaných slov se v různých rovinách liší. Nejvíce dat je anotováno morfologicky a s hlubšími vrstvami množství značkovaných dat klesá, což je vidět z údajů v tabulce 2.1. Značkování v jednotlivých rovinách jsou vzájemně provázána a jsou prováděna více lidmi. Je tedy náročné udržet konzistenci značek. Proto byly vytvořeny anotační manuály [5] pro všechny roviny, z nichž jsem při práci vycházela.

| Název roviny | Počet | | |
|---|---|---|---|
| | anotovaných souborů | vět | slovních jednotek |
| m-rovina | 7 110 | 115 844 | 1 957 247 |
| a-rovina | 5 330 | 87 913 | 1 503 739 |
| t-rovina | 3 165 | 49 431 | 833 195 |

Tabulka 2.1: Počet anotovaných dat v rovinách PDT 2.0. Převzato [2], str. 18.

Rovina slov (w-rovina) vzniká tokenizací vstupního textu, tj. segmentací věty na slova. Obsahuje identifikátory vět a slovních jednotek. Za slovní jednotky jsou považována slova, číslice a interpunkční znaménka. V této rovině se nic neanotuje ani neopravuje. Obsažený text je dál značkován až v dalších rovinách. Data v jednotlivých rovinách jsou zaznamenána pomocí jazyka Prague Markup Language (dále jen PML). PML soubor w-roviny pro větu „To je psí život!“ vypadá následovně:

```
<?xml version="1.0" encoding="utf-8"?>
```

```
<wdata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
 <head>
  <schema href="wdata_schema.xml" />
 </head>
 <meta>
  <original_format>tmt</original_format>
  <lang>cs</lang>
 </meta>
 <doc id="w-100">
  <docmeta>
  </docmeta>
  <para>
   <w id="w-s71-w1">
    <token>To</token>
   </w>
   <w id="w-s71-w2">
    <token>je</token>
   </w>
   <w id="w-s71-w3">
    <token>psí</token>
   </w>
   <w id="w-s71-w4">
    <token>život</token>
    <no_space_after>1</no_space_after>
   </w>
   <w id="w-s71-w5">
    <token>!</token>
   </w>
  </para>
 </doc>
</wdata>
```

Soubor obsahující informace z w roviny se skládá z hlavičky, údajích o slovních jednotkách uzavřených v tagu `<w>` a patičky. Každé slovní jednotce je přiřazen v souboru jednoznačný identifikátor. Uvnitř tagu `<w>` je zaznamenána kromě identifikátoru i povrchová podoba slovní jednotky uzavřená v tagu `<token>`. Pokud mezi slovními jednotkami není mezera, je u slova po němž mezera chybí, přidána informace o chybějící mezeře.

Na w-rovinu navazuje m-rovina, kde jsou k slovním jednotkám w-roviny přiřazeny další atributy. Hlavními atributy jsou `<lemma>` a `<tag>`. Atribut `<lemma>` obsahuje lemma dané slovní jednotky (základní tvar slovní jednotky např. pro sloveso „uměl“ lemma „umět“). Dále pak atribut `<tag>` obsahuje strukturovanou morfologickou značku o patnácti pozicích, která obsahuje informaci o slovním druhu a dalších morfologických kategoriích. K propojení s analytickou rovinou slouží jednoznačný identifikátor `<id>` a pro zpětnou referenci do w-roviny je používán jednoznačný atribut `<w.rf>`. Součástí morfologické roviny je i atribut `<form>`, který slouží k opravám a/nebo normalizacím týkajících se w-roviny. Může se stát, že se ve w-rovině nacházejí například tiskové chyby nebo špatně spojená či

rozdělená slova, případně další technické problémy. Příklad textové podoby:

```
<?xml version="1.0" encoding="utf-8"?>

<mdata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
 <head>
  <schema href="mdata_schema.xml" />
  <references>
   <reffile id="w" name="wdata" href="100.w" />
  </references>
 </head>
 <meta>
  <lang>cs</lang>
 </meta>
  <s id="100-s71">
  <m id="m-s71-w2">
   <w.rf>w#w-s71-w2</w.rf>
   <form>je</form>
   <lemma>být</lemma>
   <tag>VB-S---3P-AA---</tag>
  </m>
  <m id="m-s71-w1">
   <w.rf>w#w-s71-w1</w.rf>
   <form>To</form>
   <lemma>ten</lemma>
   <tag>PDNS1----------</tag>
  </m>
  <m id="m-s71-w4">
   <w.rf>w#w-s71-w4</w.rf>
   <form>život</form>
   <lemma>život</lemma>
   <tag>NNIS1-----A----</tag>
  </m>
  <m id="m-s71-w3">
   <w.rf>w#w-s71-w3</w.rf>
   <form>psí</form>
   <lemma>psí</lemma>
   <tag>AAIS1----1A----</tag>
  </m>
  <m id="m-s71-w5">
   <w.rf>w#w-s71-w5</w.rf>
   <form>!</form>
   <lemma>!</lemma>
   <tag>Z:-------------</tag>
  </m>
 </s>
</mdata>
```

Poslední a nejhlubší rovina používaná pro tuto diplomovou práci, je rovina

analytická, tedy a-rovina. V této rovině je každá věta reprezentována zakořeněným stromem s ohodnocenými hranami a uzly. Každý prvek (slovo nebo interpunkční znaménko) z morfologické roviny odpovídá právě jednomu uzlu stromu v analytické rovině. Závislostní vztah slovních jednotek je vyjádřen hranou mezi příslušnými uzly, kde ohodnocení hrany je dáno typem vztahu (např. závislostním, koordinací, apozicí,...). Každému z uzlů jsou přiřazeny atributy s informací o propojení mezi rovinami, s pořadím slovní jednotky ve větě a s analytickou funkcí. PML soubor a-roviny pro větu „To je psí život!“ vypadá následovně:

```
<?xml version="1.0"?>
<adata xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
  <head>
    <schema href="adata_schema.xml"/>
    <references>
      <reffile id="m" name="mdata" href="100.m"/>
    </references>
  </head>
  <trees>
      <LM id="a-s71-root">
      <s.rf>m#m-</s.rf>
      <ord>0</ord>
      <children>
        <LM id="a-s71-w2">
          <m.rf>m#m-s71-w2</m.rf>
          <afun>Pred</afun>
          <ord>2</ord>
          <children>
            <LM id="a-s71-w1">
              <m.rf>m#m-s71-w1</m.rf>
              <afun>Sb</afun>
              <ord>1</ord>
            </LM>
            <LM id="a-s71-w4">
              <m.rf>m#m-s71-w4</m.rf>
              <afun>Pnom</afun>
              <ord>4</ord>
              <children>
                <LM id="a-s71-w3">
                  <m.rf>m#m-s71-w3</m.rf>
                  <afun>Atr</afun>
                  <ord>3</ord>
                </LM>
              </children>
            </LM>
          </children>
        </LM>
        <LM id="a-s71-w5">
          <m.rf>m#m-s71-w5</m.rf>
          <afun>AuxK</afun>
```

```
          <ord>5</ord>
        </LM>
      </children>
    </LM>
  </trees>
</adata>
```

Větné členy jsou v souboru uzavřeny v tagu `<LM>`, který obsahuje tag `<children>`, uvnitř kterého jsou větné členy závislé na tomto větném členu.

V PDT 2.0 se dále vyskytuje tektogramatická rovina (t-rovina) zachycující hloubkovou strukturu věty. Obdobně jako v analytické rovině je struktura reprezentována zakořeněným stromem. Na rozdíl od analytické roviny jsou však uzly plnovýznamová slova (např. nejsou zachyceny předložky). Na této rovině je vyznačeno aktuální členění a v současné verzi i koreference. Školní jazykový rozbor je někde na pomezí analytické a tektogramatické roviny. Ve školních rozborech totiž uvažujeme o závislosti větných členů (předložka je tedy spojena s plnovýznamovým slovem) a ne o závislosti jednotlivých slov na sobě, jak je tomu v analytické rovině. Na druhou stranu ve školním rozboru nezobrazujeme např. gramatickou ani textovou koreferenci. Tektogramatickou rovinu proto v rámci diplomové práce neuvažujeme.

## 2.2 Systém STYX

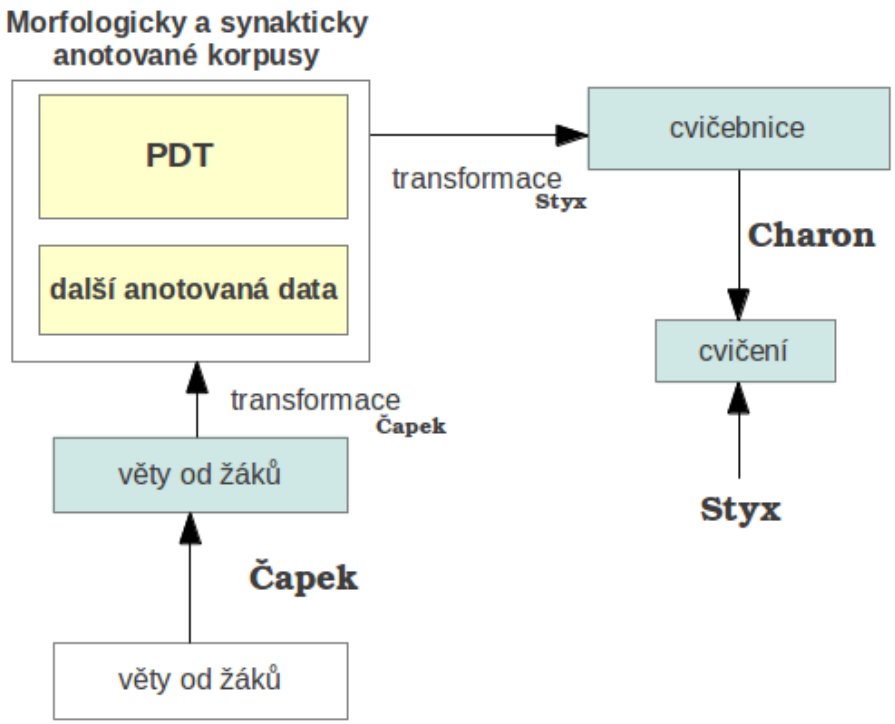
Rozsáhlá databáze ručně anotovaných vět jako je PDT má veliký potenciál pro další využití. Je možné ji použít například k různým statistickým modelům pro překlady nebo kontrolu syntaxe. Dalším využitím může být tvorba a vyhodnocování gramatických pravidel pro generování vět z daného jazyka. Jedním z využití těchto dat je i systém STYX.

Systém STYX je elektronickou cvičebnicí českého jazyka sloužící k procvičování morfologie a syntaxe. Je možné zpracovávat pouze věty z předem připravené množiny vět (viz odstavec 2.2.1), která je výběrem z textů z Pražského závislostního korpusu. Sestavení a propojení jednotlivých modulů systému je znázorněno na obrázku 2.1. Popis struktury systému lze najít v práci [8].

### 2.2.1 Databáze vět

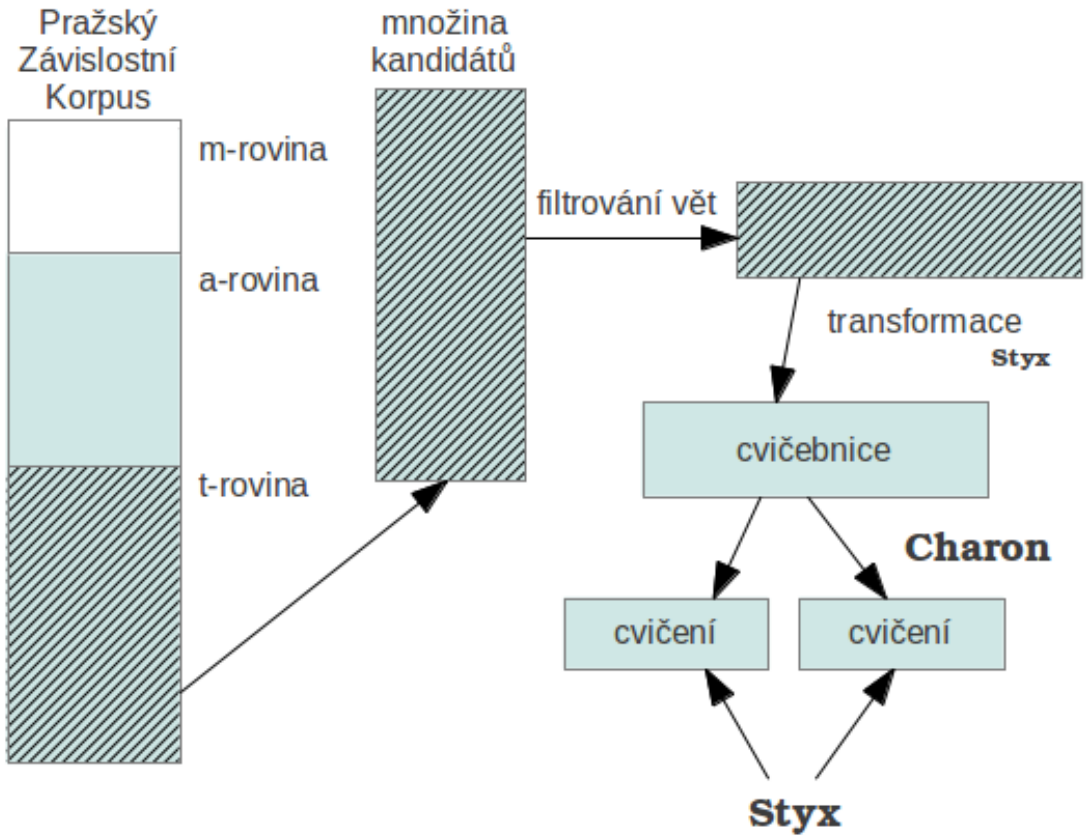
V práci [8] Ondřeje Kučery byla vytvořena databáze pro cvičebnici STYX obsahující jedenáct tisíc vět. Ty byly vybrány přímo z Pražského závislostního korpusu. Zdaleka ne všechny věty nacházející se v PDT jsou vhodné k procvičování pro žáky škol. Nelze například procvičovat jev, o kterém se žáci neučili. Proto byly vhodné věty, tj. ty, na jejichž zpracování se učebnice shodují, filtrovány za použití filtračních kritérií, která vyřazují z PDT věty s danou vlastností (např. elipsy nebo věty se speciálními grafickými symboly).

Vzhledem k odlišnému zpracovávání větných členů ve školních rozborech oproti zpracování v PDT bylo nutno použít transformace vybraných vět pro získání syntaktických stromů, na které jsou žáci a učitelé ve školách zvyklí. Například podmět a přísudek jsou ve škole v jedné rovině, zatímco v PDT je podmět závislý na přísudku. Ve větě „Šel na houby.“ jsou v PDT tři větné členy, zatímco pro




Obrázek 2.1: Schéma sestavení a napojení jednotlivých modulů systému STYX na PDT.

školní rozbor jsou slova „na“ a „houby“ jedním větným členem „na houby“. V diplomové práci Ondřeje Kučery [8] bylo popsáno celkem 23 pravidel ošetřujících transformaci z PDT do tvaru školních větných rozborů. Schéma procesu stavby cvičebnice STYX je znázorněno na obrázku 2.2.




Obrázek 2.2: Schéma procesu stavby cvičebnice STYX.

### 2.2.2 Uživatelské nástroje systému STYX

Pro práci s cvičebnicí byly vytvořeny nástroje *Charon* a *Styx*[1]. Charon umožňuje sestavovat cvičení z databáze vět. Vzhledem k tomu, že množství vět je velké (11 tisíc) a přečíst všechny tyto věty je časově náročné, je možné využít automatického výběru vět obsahující specifické jevy. *Styx* pak umožňuje procvičování morfologie i syntaxe na vybraných větách a okamžitou kontrolu.

Rozšířením systému STYX je i editor větných rozborů *Čapek*, který umožňuje, na rozdíl od původního systému, zadávat vlastní věty určené k rozboru.

[1] *STYX* je systém skládající se z dat i modulů zatímco *Styx* je jedním modulem systému STYX.

## 2.3 Editor Čapek

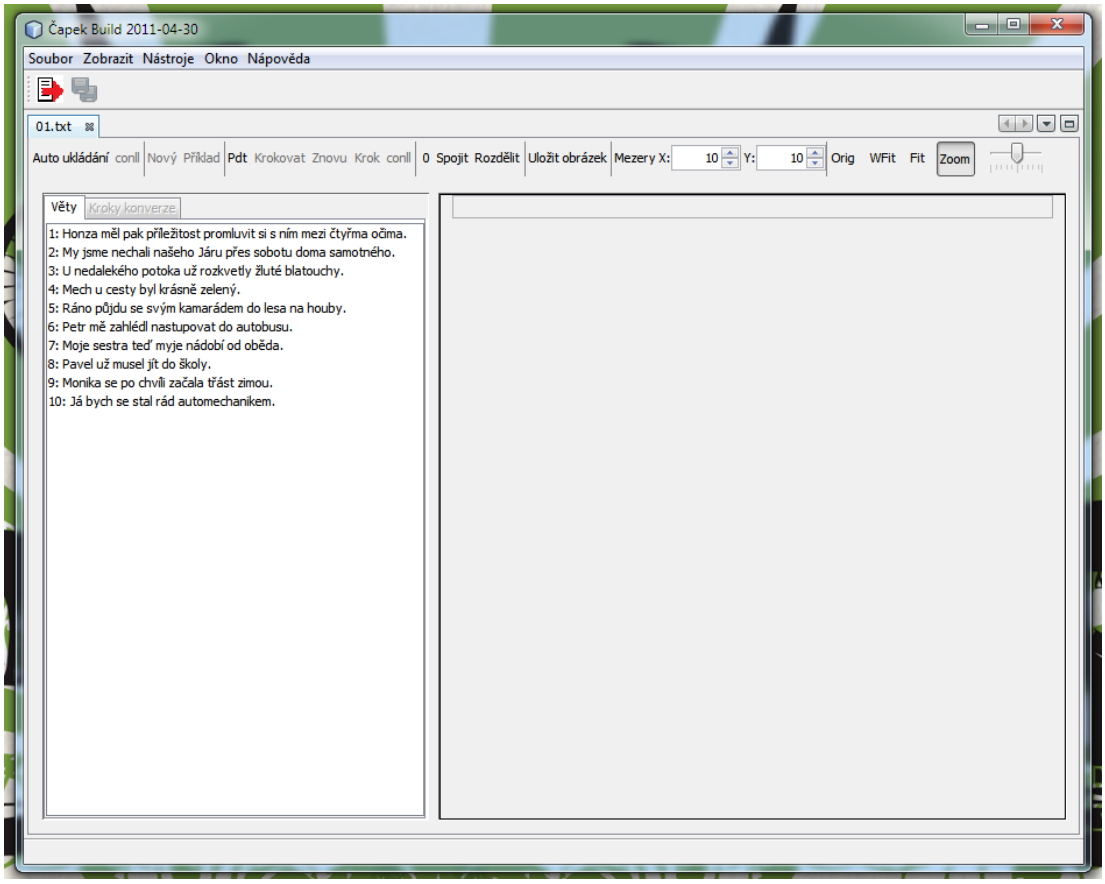
Editor větných rozborů Čapek je nejnovějším nástrojem systému STYX. Dosud se jedná o funkční prototyp nástroje (stále probíhá jeho vývoj na základě situací vzniklých při jeho užívání). Několik úprav vzniklo i na základě zkušeností při sběru dat pro tuto diplomovou práci. Čapek umožňuje učitelům i žákům zpracovávat libovolné věty, které si mohou připravit v textovém formátu tak, že na každém řádku souboru je samostatná věta. Tento soubor je do Čapka importován a je umožněno s ním pracovat. Editor nekontroluje žákům chyby. K výhodám jeho používání patří, že rozbory žáků vykresluje přehledně. Je možné nastavovat velikost mezer mezi jednotlivými uzly analytického stromu a lze zmenšovat velikost uzlů, takže i u složitějších a delších vět je možné udržovat přehled o celé struktuře. Editor umožňuje pracovat s načtenými větami na přeskáčku i otevření více souborů najednou. Ovládání editoru je velice intuitivní.

Při návrhu programu bylo třeba učinit několik rozhodnutí ke struktuře editoru a formátu dat bez podrobné znalosti praxe. Tato rozhodnutí byla při používání editoru přehodnocována, ale dosud program obsahuje několik nepříjemných vlastností, které jsme v průběhu práce s ním objevili. Zadávané morfologické kategorie nejsou vázány na slova, ale na větné členy, takže např. u předložkových spojení „na houby“ nebo u několikanásobných větných členů „kluci a holky hráli hry“ není jasně dané, co bychom vlastně měli určovat. U některých slov pak po transformaci do formátu PDT morfologická informace chybí. Žáci na ni tedy nemohou být testováni. Další nepříjemností je nemožnost navázat doplněk na dva rodičovské uzly, na což jsou žáci na školách zvyklí. Věty zpracovávané v editoru by měly být později automaticky odesílány k dalšímu použití, což zatím není implementováno a věty jsou sbírány ručně.

V následujících odstavcích jsou podrobně popsány úkony prováděné při značkování v editoru Čapek.

### 2.3.1 Tvaroslovný rozbor

Po načtení vstupního souboru s větami má uživatel v editoru Čapek k dispozici seznam vět. Ukázka obrazovky s načteným seznamem vět se nachází na obrázku 2.3. Po kliknutí na větu získá slova dané věty rozdělená do krabiček, jak ukazuje obrázek 2.4. Interpunkce, pokud není z obou stran oddělená mezerou, je přidružena ke slovu, které se nachází před ní. Jsou-li slova zobrazena v oddělených krabičkách, je možné začít s určováním slovních druhů a jejich morfologických kategorií. Určované morfologické kategorie a jejich možné hodnoty byly vybrány z publikace [11]. Vzhledem k tomu, že je v programu Čapek morfologie zatím navázána na větné členy a nikoliv na slova, jak je zvykem při školních rozborech, byla při sběru dat morfologie zpracovávána až po vystavění závislostního stromu. Mohla být určena i před tím, ale pak se po spojení slov do větných členů část zadané informace ztrácí, takže doporučené pořadí nejdříve vytvořit větné členy a až pak určovat morfologii šetří čas při práci. Aby bylo při určování slovních druhů a jejich morfologických kategorií u víceslovných větných členů ve školách zachováno co nejvíce informací potřebných pro transformaci do struktury PDT, byli uživatelé Čapka (žáci ve školách i učitelé) v rámci sběru rozborů pro tuto diplomovou práci požádání o aplikování následujících pravidel při určování morfologie:




Obrázek 2.3: Ukázka úvodní obrazovky editoru Čapek po načtení souboru s větami.


Obrázek 2.4: Ukázka věty na začátku práce v editoru Čapek.

(1) ohebná slova mají přednost před neohebnými,

(2) sloveso v určitém tvaru má přednost před vším,

(3) mnohonásobné větné členy nespojujeme do jednoho členu a určujeme každý zvlášť.

První pravidlo zajišťuje určení maximálního množství informací. Například pokud je dvojslovný větný člen „o houbách“ označen jako podstatné jméno v šestém pádu je velmi pravděpodobné, že druhé slovo větného členu bude předložkou. Pokud bychom však měli určenou pouze předložku, budeme těžko dohledávat morfologické kategorie připojeného podstatného jména.

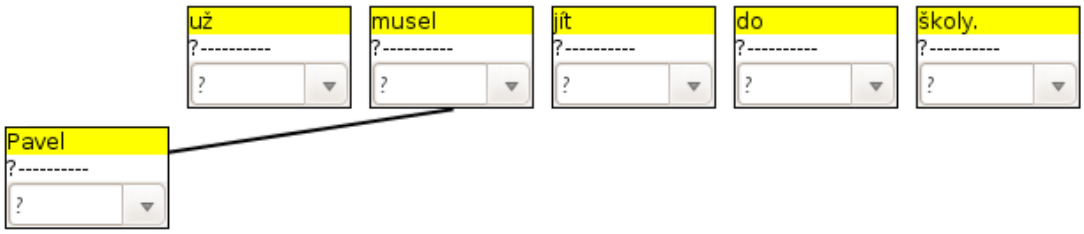
Druhé pravidlo se týká zejména přísudků se sponou, kde je upřednostňováno určení morfologických kategorií spony před jmennými částmi.

Ačkoliv třetí pravidlo neodpovídá zvyklostem při školních větných rozborech, ztratili bychom bez něj velmi mnoho informací a zanášeli bychom do struktury PDT na morfologické rovině velmi mnoho chyb.

### 2.3.2 Větný rozbor

Na začátku tvorby závislostního grafu[2] je vhodné sloučit slova, která společně tvoří jeden větný člen do jednoho uzlu. Uzly lze slučovat i v průběhu práce, ale může dojít ke ztrátě již zapsané informace. Při slučování se totiž přebírá pouze hodnota prvního označeného uzlu. Tuto operaci nelze vzít zpět a týká se též morfologických informací k druhu větného členu.

Graf závislostí je budován pomocí hran mezi větnými členy (dále jako uzly) tak, že hrana vede vždy od závislého uzlu k uzlu, na kterém závisí (dále jako řídící uzel). Záleží nám tedy na směru kterým propojujeme závislé členy. Na začátku práce jsou uzly seřazeny dle pořadí ve větě, ale zobrazované pořadí uzlů lze měnit[3]. Po vytvoření závislosti je závislý uzel posunut o úroveň níže, než se nachází rodičovský uzel. Vzdálenosti mezi uzly jsou konstantní jak v horizontálním, tak ve vertikálním směru. Na obrázcích 2.5 a 2.6 je vidět, co udělá opačný postup zavěšování větných členů s grafickou podobou věty v editoru.

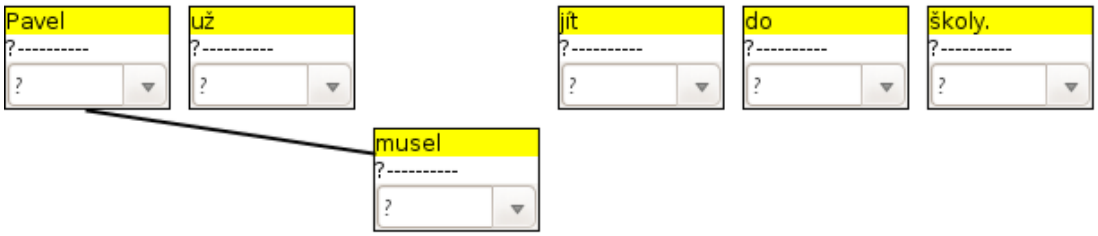



Obrázek 2.5: Ukázka zavěšení uzlů, kdy uzel *musel* je řídícím uzlem pro uzel *Pavel*.

V programu odlišujeme dva typy závislostních hran. Nominálně je použita závislostní hrana, která posune závislý uzel o úroveň níže. Pro případ závislosti mezi podmětem a přísudkem, kterou jsou žáci na školách zvyklí vidět ve stejné

[2]Naším grafem je zakořeněný strom, tedy graf bez kružnic, který má pevně daný kořen (v našem případě je kořenem predikát).

[3]Pořadí slov ve větě však zůstává neměnné.




Obrázek 2.6: Ukázka zavěšení uzlů, kdy uzel *Pavel* je řídícím uzlem pro uzel *musel*.

úrovni, byl zaveden druhý typ závislostní hrany, který zachová uzly ve stejné rovině, ilustrovaný obrázkem 2.7.


Obrázek 2.7: Ukázka zavěšení uzlů, kdy uzly *Pavel* a *musel* leží v jedné rovině.

Chceme-li přidat nevyjádřený větný člen, je možné přidat uzel, který nebude obsahovat žádné slovo. Polohu přidaného uzlu je možné měnit vzhledem k poloze ostatních uzlů. Možnosti přidat uzel je využíváno i v technických pravidlech, která byla nutná zavést pro zachování konzistence značkování:

(1) sloveso je kořenem závislostního stromu,

(2) vedlejší věty jsou připojovány přes spojku nebo přidaný uzel,

(3) jednotlivé členy vícenásobných větných členů jsou navěšeny na spojku nebo přidaný uzel a ten je zavěšen na místo celého větného členu,

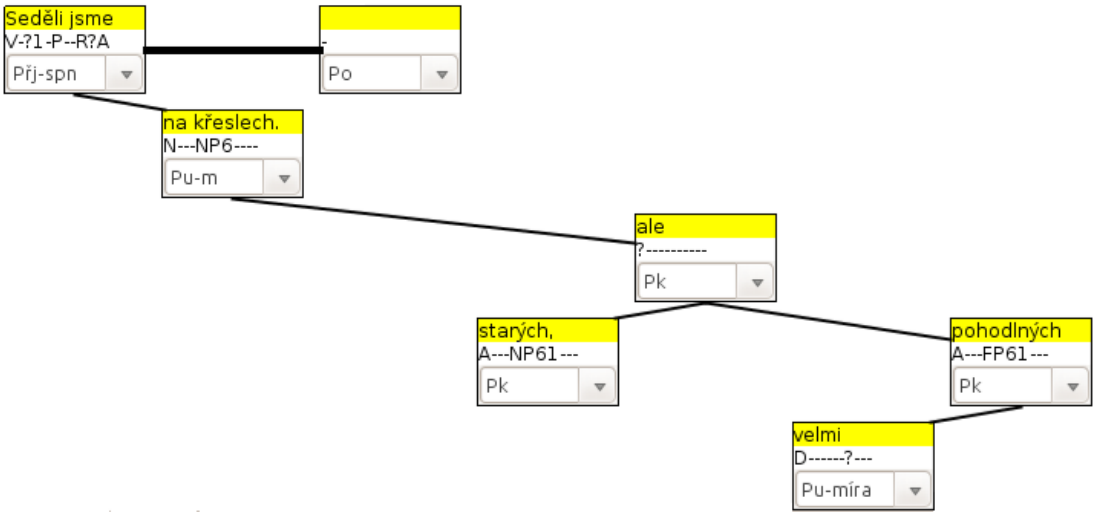
(4) s větami souřadnými zacházíme jako s vícenásobnými větnými členy,

(5) doplněk zavěšujeme na rodiče, o kterém předpokládáme, že se na něj váže více.

První pravidlo může lingvistům připadat nadbytečné, ale ve školách se učí o základní větné skladební dvojici, podmětu s přísudkem. Pokud tedy chceme značení sjednotit, je nutné toto pravidlo zavést. Toto pravidlo lze vypustit a ošetřit při transformacích, protože víme, jaký typ hrany byl zvolen. Je třeba mít tuto odlišnost na zřeteli.

Při sběru dat jsme zjistili, že se na školách málokdy dělají z časových důvodů větné rozbory pokrývající určování jak analytických, tak morfologických funkcí. Pokud se rozebírá souvětí, je nakresleno schéma zapojení vět, a každá z vět je rozebírána zvlášť. Proto bylo třeba zavést pravidlo (2) na stavbu souvětí, které by se co nejvíce podobalo modelu používanému na školách.

Třetí analytické pravidlo pak slouží jako příprava pro uplatnění třetího pravidla morfologického. Ze všech možností (např. postupné závislosti jednotlivých členů na sobě nebo závislosti na prvním ze členů mnohočlenu) byla zvolena reprezentace, která se graficky podobá reprezentaci v PDT. Příklad použití tohoto pravidla je uveden na obrázku 2.8.




Obrázek 2.8: Příklad použití třetího analytického pravidla pro tvorbu závislostního stromu v editoru Čapek.

Čtvrté a páté pravidlo bylo zařazeno z důvodů omezených možností editoru Čapek vzhledem ke zvyklostem žáků při práci s větnými rozbory. Bylo nutno najít takový přístup, který žáci snadno zvládnou, a přitom jej lze použít v rámci možností editoru Čapek.

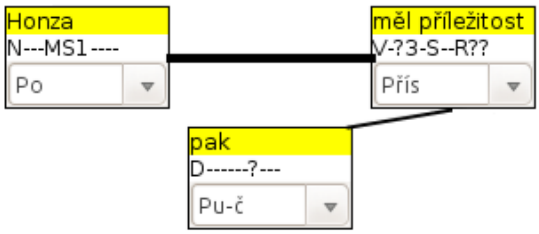
### 2.3.3 Datová reprezentace vět a rozborů

Zpracovávané věty jsou reprezentovány XML strukturou s pevně daným schématem, které bylo připraveno pro program Čapek. Příklad souboru:

```
<?xml version="1.0" encoding="UTF-8"?>
<capek xmlns="http://purl.org/jh/capek/">
  <head>
    <schema href="capek_schema.xml" />
  </head>
  <tagsets>
    <tagset tid="ufal.mff.cuni.cz/czSchoolSyn" version=""
     use="syntax" />
    <tagset tid="ufal.mff.cuni.cz/czMorphSimplified" version=""
     use="morph" />
  </tagsets>
  <sentence id="s1">
    <txt>Honza měl pak příležitost.</txt>
    <token id="s1tok0">
      <txt>Honza</txt>
      <ord>0</ord>
    </token>
    <token id="s1tok1">
      <txt>měl</txt>
      <ord>1</ord>
    </token>
```

```
    <token id="s1tok2">
      <txt>pak</txt>
      <ord>2</ord>
    </token>
    <token id="s1tok3">
      <txt>příležitost</txt>
      <ord>3</ord>
    </token>
    <tree id="s1t1">
      <node id="w">
        <parent.rf>s1t1node1</parent.rf>
        <flat />
        <token.rf>s1tok0</token.rf>
        <fnc>Po</fnc>
        <tag>N-MS1---</tag>
      </node>
      <node id="s1t1node1">
        <token.rf>s1tok1</token.rf>
        <fnc>Přís</fnc>
        <tag>V3-S--RA</tag>
      </node>
      <node id="s1t1node2">
        <parent.rf>s1t1node1</parent.rf>
        <token.rf>s1tok2</token.rf>
        <fnc>Pu-č</fnc>
        <tag>D----?--</tag>
      </node>
      <node id="s1t1node5">
        <parent.rf>s1t1node3</parent.rf>
        <token.rf>s1tok4</token.rf>
        <token.rf>s1tok5</token.rf>
        <fnc>Pk</fnc>
        <tag>P-?S3---</tag>
        <main.rf>s1tok5</main.rf>
      </node>
    </tree>
  </sentence>
</capek>
```

Ve struktuře je po uvozovacím tagu pro větu `<sentence>` uložená mezi tagy `<txt>` rozebíraná věta tak, jak byla zadána na vstupu před jakýmikoliv úpravami. Dále pak následují tagy `<token>`, kde jsou informace o jednotlivých slovech (máme informaci o pořadí slova ve větě, jeho znění a jeho identifikátor, který je jednoznačný pro každý token ze souboru. Následuje záznam závislostní struktury, která je uzavřena v tagu `<tree>` a tag `<node>` reprezentuje větné členy zaznamenávané syntaxe. Uvnitř tohoto tagu máme informaci o typu větného členu, morfologický tag, jednoznačný identifikátor větného členu, na kterém je daný větný člen závislý, a odkazy na tokeny (slova), které se v daném větném členu vyskytují. Příklad grafické podoby věty v editoru Čapek je uveden na obrázku 2.9.




Obrázek 2.9: Příklad grafické podoby věty odpovídající *xml* kódu uvedeného na straně 17 v editoru Čapek.

Informace uvnitř tagu `<tag>` je, stejně jako celý soubor, strukturovaná, tedy každá z pozice tohoto tagu má svůj význam. V editoru byl nejdříve používán osmimístný systém po připomínkách byl tento systém o tři pozice rozšířen. Dvě pozice byly přidány na druhé a třetí místo a poslední pozice byla připojena nakonec. Při čtení souborů je třeba na tuto informaci brát zřetel. V tabulce 2.2 je popis jednotlivých pozic tagů a jsou v ní zmíněny i slovní druhy, kterých se daná pozice bude dotýkat. Pokud se určitá pozice nějakého slovního druhu netýká, je vyplněna pomlčkou, takže víme, že se s touto pozicí nepracuje.

| Pozice | Morfologická kategorie |
|---|---|
| 1 | slovní druh |
| 2 | druh zájmena |
| 3 | způsob |
| 4 | osoba |
| 5 | rod |
| 6 | číslo |
| 7 | pád |
| 8 | stupeň |
| 9 | čas |
| 10 | vid |
| 11 | slovesný rod |

Tabulka 2.2: Poziční systém tagů v editoru Čapek

### 2.3.4 Převod do PML

Soubory z editoru Čapek je nutné převést do struktury PDT. PDT používá jazyk *PML* založený na *XML* základu, který je univerzálně použitelný pro lingvistické značkování textů. Byl vyvinut v Ústavu formální a aplikované lingvistiky na Univerzitě Karlově v Praze. Pro konverzi z jiných datových formátů podporovaných některými korpusy jsou k dispozici konverzní a grafické nástroje. Pro každou z rovin PDT 2.0 bylo vytvořeno speciální *PML* schéma [9]. Stejně jako u editoru Čapek je i v PDT 2.0 využíván poziční systém tagů. Na rozdíl od Čapka, kde jsou tagy osmi až jedenáctimístné, je v *PML* využíván pátnáctimístný systém budování tagů. Jednotlivé pozice tagů používaných v *PML* jsou popsány v tabulce 2.3.

| Pozice | Morfologická kategorie |
|---|---|
| 1 | slovní druh |
| 2 | upřesnění slovního druhu |
| 3 | rod |
| 4 | čísloy |
| 5 | pád |
| 6 | rod vlastníka |
| 7 | číslo vlastníka |
| 8 | osoba |
| 9 | čas |
| 10 | stupeň |
| 11 | negace |
| 12 | slovesný rod |
| 13 | rezerva |
| 14 | rezerva |
| 15 | varianta použití |

Tabulka 2.3: Poziční systém tagů v PDT (převzato z [3])

## 2.4 Nástroje, použité formáty a jazyk

Pro transformaci výstupů z programu Čapek do struktury *PML* používám skripty v jazycích *perl* a *SQL*. Důvodem volby jazyka *SQL* byla jednoduchost filtrování vybraných dat pomocí dotazovacího jazyka a snadná konverze vstupních a výstupních souborů v rámci databázového prostředí. Realizace navržených metod je provedena na databázovém serveru *MySQL*. Jedná se o volně dostupný multiplatformní server s rozšířeným využitím v praxi. Výhodou je i možnost snadné implementace případného webového rozhraní. Jazyk *perl* byl zvolen z důvodu možnosti snadné komunikace s *MySQL* databází v rámci balíku *DBI* a snadnému načítání *xml* souborů pomocí balíku *XPath*.

Vstupem transformačního modulu jsou *XML* soubory vytvořené programem Čapek. Modul umožňuje výstup jak ve formátu čitelném programem Čapek, tak ve formátu *PML* používaném ve verzi 2.0 Pražského závislostního korpusu.

# 3. Práce na školách: získávání školních větných rozborů

Motivací pro tuto práci je možnost využití školních větných rozborů jako alternativní metodu k značkování korpusů. Pro testovaní relevantnosti těchto rozborů pro značkování jsme oslovili některé základní školy a gymnázia. Cílem bylo motivovat češtináře i žáky pro práci na jazykových rozborech v editoru větných rozborů *Čapek* a získat testovací data, na kterých bychom mohli testovat reálnou využitelnost školních rozborů pro značkování. Učitelům byl představen editor a byli požádáni o vyučovací čas ve vybraných třídách pro prezentaci editoru žákům. Žáci byli seznámeni s využitelností větných rozborů a požádáni o zpracování několika vět v editoru.

## 3.1 Výběr vět

Testovací věty jsou vybrány z učebnice [11] ze cvičení zaměřených na jazykové rozbory. Výběr vět z učebnice byl postupný a věty zůstaly ve stejném pořadí, v jakém byly uvedeny v učebnici. Testované jevy se tedy nachází pospolu. Věty neprocházely žádnou systematickou selekcí. Bylo vybráno postupně 101 vět, které byly posléze přepsány do formátu vhodného pro editor Čapek, tj. každá věta je na vlastním řádku.

## 3.2 Prezentace

Žáci byli pro účely zpracování rozborů v editoru Čapek vyučováni v počítačových učebnách. S každou třídou, případně její částí, bylo pracováno 45 minut, tedy jednu vyučovací hodinu. Ihned na začátku hodiny proběhl pro stimulaci komunikace přibližně dvouminutový dotazník zkoumající vztah žáků k jazykovým rozborům. Dotazník obsahoval následující otázky:

1. Kdo má rád češtinu?
2. Kdo má rád větné rozbory?
3. Kdo si myslí, že by větné rozbory mohly být k něčemu užitečné?
4. Kdo má představu, k čemu by mohly být větné rozbory dobré? (odpověď k ničemu se nepočítá)

Otázky byly čteny nahlas a žáci byli vyzváni k zvednutí ruky. Nejprve měli zvednout ruku ti, kdo souhlasili, poté ti, kteří nesouhlasili a nakonec ti, kteří neodpověděli. Po poslední otázce byli ti žáci, kteří měli zvednutou ruku, dotázáni na své představy.

Po dotazníku následovalo přednesení motivační části za podpory krátké prezentace. Žáci byli motivováni užitečností jazykových rozborů pro ně samotné i pro výzkum. Byla jim názorně ukázána víceznačnost tříčlenné věty "Ženu holí stroj"

a využití analytického stromu při překladu z angličtiny do češtiny na jednoduché větě "Marie is going through the forest". Na konci prezentace byly zmíněny výhody používání editoru Čapek při zpracování jazykových rozborů.

Následně byli žáci vyzváni k spuštění editoru a byla jim vysvětlena práce s ním. Bylo jim vysvětleno, jak mají otevírat soubor, jakým způsobem mohou přeskakovat mezi větami, způsob vytváření závislostí, spojování slov, určování morfologie, přidávání krabiček pro nevyjádřené větné členy, doporučený způsob řešení koordinace, doporučený způsob pro řešení vedlejších vět a spojování a rozdělování slov. Následně byli požádáni o dotazy k editoru, k práci s ním, případně jazykovým rozborům jako takovým a poté vyzváni k vypracování kompletních rozborů na připravených souborech. Žáci otevírali připravený *xml* soubor s větami určenými k jazykovému rozboru, který byl uložen na síťovém disku v rámci školní lokální sítě.

Závěrem hodiny vypracovávali žáci připravený dotazník podrobněji prezentovaný v sekci 3.3 a vyzváni k diskuzi ohledně práce v editoru a vypracovávaných vět.

Učitelům byl editor Čapek představen před prezentací žákům. Bylo jim ukázáno prostředí editoru a práce s ním a vysvětlena naše motivace k šíření editoru do škol (snaha o alternativní způsob značkování). Učitelé pak překvapivě na prezentaci editoru žákům nebyli přítomni. Nicméně byli vyzpovídáni a požádáni o vyplnění dotazníku pro učitele (sekce 3.3).

## 3.3 Dotazníky

K vyhodnocení komfortu práce s programem Čapek byly vytvořeny online dotazníky pro žáky i učitele. Jejich účelem bylo zjistit zájem o editor, případně jaká úskalí se při práci s programem vyskytla. Zajímalo nás, jak se žákům i učitelům s programem pracovalo a zda by s ním byli ochotni pracovat i později, případně jaká vylepšení by přivítali. Kompletní znění dotazníků i sesbírané odpovědi na otázky jsou uvedeny v příloze A. Na otázky bylo odpovídáno dobrovolně. Od žáků jsme ke dni 5.11.2011 získali 55 odpovědí.

## 3.4 Návštěvy škol

Sběr rozborů probíhal nejprve při výuce ve třídách na vybraných pražských školách. Zaměřujeme se na nižší třídy víceletých gymnázií a druhý stupeň základních škol, neboť mají výuku školních rozborů ve studijním plánu, a tudíž očekáváme, že v budoucnu bychom dostávali rozbory zpracovávané především v těchto ročnících. Byly navštíveny následující školy a třídy:

- Gymnázium Chodovická v Horních Počernicích (tercie, kvarta, oktáva),
- ZŠ Klánovice (sedmé třídy),
- ZŠ Dolní Počernice (šestá, sedmá, osmá, devátá třída).

Gymnázium Chodovická a základní školu v Klánovicích jsem vybrala, protože jsem je navštěvovala, snáze se mi oslovovali jednotliví učitelé a mám je v blízkosti svého bydliště, takže dojíždění do těchto institucí je jednodušší. Poslední

navštívenou školou byla základní škola Dolní Počernice, protože je opět v blízkosti mého bydliště a znala jsem tam členku učitelského sboru. Pokoušela jsem se získat ke spolupráci gymnázium Jaroslava Seiferta ve Vysočanech. Tam ale ke sběru dat nakonec nedošlo z důvodu obtížné komunikace s vybraným češtinářem a deklarovaného špatného stavu počítačové učebny.

Spolupráce se správci počítačových učeben byla v klánovické a hornopočernické škole snadná a správci mi vyšli vstříc. V případě Horních Počernic jsem byla požádána o software a sdělení svých potřeb. Požádala jsem o přístup na server kvůli nahrávání souborů, na kterých žáci pracovali, a možnost využití dataprojektoru. Ve všem mi bylo vyhověno a software byl připraven. V klánovické škole jsem dostala přístup do učebny a připravovala jsem software na jednotlivých počítačích sama. Na základní škole v Dolních Počernicích nebyl přítomný nikdo, kdo by znal technické detaily o počítačové učebně (kdo má jaká přístupová práva, případně kam lze instalovat software). Vyučující byli nicméně velice vstřícní a software se podařilo zprovoznit. V tabulce 3.1 je zobrazen počet odučených hodin na jednotlivých školách a počet počítačů v učebně, na kterých se pracovalo během hodiny.

| | Počet | |
|---|---|---|
| Název školy | využitých strojů | hod. sběru na škole |
| ZŠ Klánovice | 6 | 4 |
| ZŠ Dolní Počernice | 18 | 4 |
| Gymnázium Horní Počernice | 15 | 3 |

Tabulka 3.1: Počet počítačů a odučených hodin

Očekávala jsem, že žáci budou počítačově gramotní a že jim tudíž nebude dělat problémy editor ovládat. Stejně tak jsem doufala, že každý z žáků zvládne zpracovat během dvaceti pěti minut alespoň pět vět. Dále jsem předpokládala, že učitele bude zajímat, co se žáky při hodině dělám, a budou hodině přítomni. Výše uvedená očekávání se však nenaplnila. Mezi vyplněná očekávání patří nadšení žáků, že mohou trávit hodinu češtiny v počítačové učebně, a snaha velké části žáků řádně splnit zadání práce, i když nebylo součástí jejich klasifikace z češtiny.

Nejprve byl proveden sběr dat u žáků tercií a kvart hornopočernického gymnázia. Vzhledem k neočekávaně malému počtu vypracovaných vět jsem zkusila, kolik vět by byli schopni vypracovat žáci v maturitním ročníku, kteří se k větným rozborům vracejí před maturitní zkouškou. Maturanti měli rozbory více procvičené a zautomatizované, tudíž byli při rozborech o něco rychlejší. Maturanti například určili analytické funkce pro 1492 tokenů, zatímco žáci kvart určili funkce pouze pro 458 tokenů.

Na základě zkušeností z prvních sběrů dat došlo ke změně přístupu k práci s žáky v hodinách. Při prvních sběrech dat žáci dostávali celý soubor vět a mohli si vybrat k vypracování libovolnou větu, což ovšem vedlo k tomu, že žáci nedodělávali věty, a velmi špatně se v hodinách hlídalo, jak kdo pracuje. Nově byl soubor vět rozdělený do oddělených souborů po pěti po sobě jdoucích větách. Každý žák ve skupině pak pracoval na vlastních větách. Soubory byly rozdělovány pokud možno rovnoměrně. Je tedy častým jevem, že první věty ze souborů jsou zpracovány vícekrát než věty z jejich konce.

Následně jsme požádali o spolupráci učitelky češtiny, které s žáky editor viděly, aby vypracovaly referenční jazykové rozbory. Obě učitelky vypracovaly všech 101 vět, každá za přibližně 12 hodin.

Práce na výše zmíněných základních školách byla vedena již s upraveným přístupem k žákům, takže se podařilo získat více ucelených informací. Na obou základních školách měli však někteří žáci problémy s určením byť jen základní stavební dvojice. To značně zpomalovalo práci na větách.

Vzhledem k tomu, že každý z žáků zpracoval maximálně pět vět, požádali jsme o zpracování většího množství vět dvě studentky a jednoho studenta z nižšího stupně gymnázia. Každý ze studentů pracoval na svém souboru vět samostatně. Chlapec navštěvuje kvartu Seifertova gymnázia. Jedna dívka navštěvuje rovněž kvartu, ale na Dopplerově gymnáziu a druhá navštěvuje tercii v pobočce hornopočernického gymnázia na Černém mostě.

### 3.4.1 Poznatky při práci se žáky a jejich reakce

Při práci ve třídách bylo zjištěno několik zajímavých poznatků a objeveno několik drobných chyb v editoru. Před začátkem práce jsme se domnívali, že žáci na školách jsou obeznámeni s prací na počítači. Opak byl nicméně pravdou. Žáci měli například problémy s otvíráním souborů. Při hovoru s žáky jsem dospěla k závěru, že u počítače sice sedí, nicméně typicky jen hrají hry či komunikují na sociálních sítích. Práce s textovými soubory a se souborovými systémy je jim stále cizí, což ovlivňovalo i rychlost zpracovávání vět. K objeveným chybám editoru patřilo například nezavírání rozbalovacích nabídek po vybrání možnosti. Několik žáků se ozvalo, že by uvítali větší možnost využívání klávesnice při provádění rozborů bez nutnosti používat myš. Těchto žáků bylo jen minimum a objevili se pouze na hornopočernickém gymnáziu.

Pokusné věty byly vybrány z učebnice pro osmou třídu, takže jsme předpokládali, že by ji žáci osmé a deváté třídy měli na konci školního roku zvládat. Zjistili jsme však, že se žáci většinou během výuky setkali pouze s jednoduchými a krátkými větami, kde se vyskytoval maximálně jeden určitý sledovaný jev (jako např. najděte doplněk) a komplexnější rozbory dělali ojediněle, pokud vůbec. Na potřebný komplexní rozbor, který byl kombinací větného rozboru a určování morfologie, nebyli vůbec zvyklí, a bylo pro ně náročné se soustředit na tolik jevů po tak dlouhou dobu. Učitelé argumentovali tím, že dlouhé věty během hodiny nestíhají zpracovat a v krátkých větách se vyskytují stejné jevy.

Vzhledem k očekávání počítačové gramotnosti žáků a obeznámenosti s jazykovou tematikou, jsme předpokládali, že přidáme pouze několik pravidel pro práci s editorem *Čapek* a žáci v něm budou schopni bez problémů pracovat. Bohužel však tyto předpoklady nebyly naplněny, a tudíž byli žáci doslova zahlceni informacemi a nevěděli, na co se soustředit dříve. Úkony při práci na počítači jako je práce se soubory (kopírování, ukládání, přejmenovávání, spouštění, zobrazování přípon, ...) nejsou samozřejmostí a je třeba je při prezentaci velice podrobně popisovat. Tyto úkony byly proto ořezány na minimální počet. Ikonu pro spuštění editoru měli žáci na ploše. Otvírali již připravený *xml* soubor, jehož otevírání jim muselo být několikrát a velice pomalu ukázáno. Pokud by měl být program šířen hromadně do škol, bylo by třeba několik přípravných hodin, kde by se postupy pro práci s editorem učily postupně po mnohem menších krůčcích. Toto však nebylo

možné vzhledem k časovým možnostem.

Sběr ve třídách byl problematický i z důvodu, že vypracované věty nebyly součástí klasifikace žáků, tedy někteří z nich volili nejsnadnější cesty pro splnění úkolu a nezamýšleli se nad zadanými větami, případně narušovali hodinu i pro ostatní žáky.

### 3.4.2 Práce s pedagogy a jejich připomínky

Pro získání referenčních „správných“ rozborů jsme požádali o spolupráci dvě učitelky češtiny z hornopočernického gymnázia. Každá z učitelek pracovala samostatně s asistencí v případě nejasností a technických problémů. Vypracování rozborů zabralo každé z nich přibližně 12 hodin a bylo rozděleno do dvou bloků v různé dny.

Při práci s počítačem si učitelky stěžovaly na neobratnost práce s myší a nepřehlednost editoru při zpracovávání dlouhých vět, což vedlo k přidání nové vlastnosti editoru. Bylo umožněno měnit velikost mezer mezi jednotlivými slovy v horizontální i vertikální rovině. Dále jim činilo například problémy sledovat, kde pracují (aktivní krabičky jsou označeny modře), nebo byl prostor, kam se dalo kliknout myší pro označení krabičky, pro ně příliš malý. Aby se s krabičkami lépe pracovalo, byly po tomto sběru upraveny a byl zvětšen prostor, na který je třeba kliknout pro jejich označení.

Některé z připomínek učitelek byly zapracovány do editoru. Například bylo rozlišeno více druhů příslovečných určení. Byly rozlišeny druhy zájmen (např. osobní, přivlastňovací, vztažné, atd.). Ke slovesným morfologickým kategoriím byl přidán způsob (oznamovací, rozkazovací a podmiňovací) a vid (dokonavý a nedokonavý). Dále bylo umožněno přesouvat pořadí uzlů, aby bylo možno vložit například nevyjádřený větný člen kamkoliv do věty a nikoliv jen na začátek. Dalším důvodem pro uvolnění pořadí krabiček v editoru byl následující případ. Při spojování krabiček do jednoho větného členu, pokud mezi nimi bylo slovo, které se s nimi nespojovalo, se krabičky po spojení nemusí řadit tak, jak by to odpovídalo správnému grafickému zobrazení při školních rozborech (místo zobrazení spojené krabičky záleží na pořadí označení krabiček. Jedná se například o větu: Někteří si možná myslíte — spojujeme *si* a *myslíte* — větný člen *si myslíte* se může zařadit na místo před větný člen *možná*, což učitelky nepovažovaly za správně udělaný větný rozbor.

Připomínky učitelek byly ovlivněny i tím, že vyučují na gymnáziu a nikoliv na základní škole. Učitelky na základních školách se při hodinách s žáky vyjadřovaly, že při výuce nezacházejí do takových podrobností (jako jsou např. druhy zájmen), protože není dostatečně procvičen nezbytný základ.

### 3.4.3 Souhrnné výsledky dotazníků

Na obrázku 3.1 vidíme jakým způsobem žáci hodnotili příjemnost ovládání editoru Čapek. Vidíme, že žákům se v editoru dle jejich odpovědí nepracovalo příliš pohodlně. Bohužel u rozšíření otázky, co by chtěli vylepšit, se našlo jen velmi málo odpovědí. Jedinou relevantní odpovědí bylo, že by měl být editor více barevný. Tato odpověď se opakovala.


Obrázek 3.1: Žákovské hodnocení příjemnosti ovládání editoru Čapek.

Na obrázku 3.2 lze vyčíst preference při možnosti výběru mezi papírem a editorem Čapek. Téměř polovina žáků vyplnila, že by upřednostnili papír. Toto by mohlo být způsobené i tím, že počítač považují z velké části jen za hračku a ne za pracovní nástroj.


Obrázek 3.2: Žákovské hodnocení příjemnosti ovládání editoru Čapek.

Zajímali jsme se i s čím konkrétně měli žáci problémy, případně zda je z jejich subjektivního pocitu vůbec měli. Jejich odpovědi znázorňuje obrázek 3.3.


Obrázek 3.3: Problémy žáků při práci s editorem Čapek.

# 4. Transformační pravidla

Jak již bylo zmíněno v odstavci 2.2, převod mezi školním větným rozborem (ŠVR) a syntaktickým stromem v PDT není zcela přímočarý. V práci [8] byl popisován převod z PDT struktury do struktury, na kterou jsou zvyklí žáci a učitelé základních a středních škol. K využití dat nasbíraných na školách pro značkování PDT je zapotřebí transformace opačným směrem, tedy ze struktury, kterou vytvoří školáci, do struktury PDT.

Transformaci poněkud komplikuje skutečnost, že ve školních větných rozborech existují víceslovné větné členy. Takové větné členy je třeba rozdělit na jednotlivá slova. K tomu máme k dispozici, jak bylo zmíněno v kapitole 3, pouze jeden morfologický tag a jednu analytickou funkci ke každému (tedy i víceslovnému) větnému členu. Závislost mezi jednotlivými sloučenými slovy i další morfologické informace se proto velice obtížně zjišťují. Naštěstí alespoň udržujeme informaci o rozdělení větného členu na slova, a známe tedy počet sloučených slov v daném větném členu.

K vytváření pravidel máme k dispozici tzv. *zlatý standard*, který vypracovala Zdeňka Urešová. Jedná se o anotaci vět uvedených v příloze B ve formátu PDT. Dále je k dispozici dvojí kompletní zpracování těchto vět učitelkami gymnázia v editoru Čapek. Obě zpracování rozborů od učitelek považujeme pro tvorbu pravidel za správná. Odlišnosti v těchto zpracováních připisujeme víceznačnosti zvolených vět. Předpoklad správnosti učitelských rozborů je velice silný, nepovažuji se však za osobu povolanou ke kontrole rozborů po středoškolských vyučujících českého jazyka, kteří by měli mít práci se ŠVR dobře zvládnutou. Konkrétní rozdíly ve zpracováních obou učitelek jsou popsány v kapitole 5.

V tabulce 4.1 jsou uvedeny některé číselné údaje o vybraných větách. Tabulka 4.2 ilustruje, že se při převodu funkcí používaných v editoru a ve školních rozborech na funkce používané v PDT nevyhneme nesnázím.

| | |
|---|---|
| Počet vět | 101 |
| Počet tokenů | 1027 |
| Počet interpunkčních znamének | 160 |
| Počet různých analytických funkcí | 19 |
| Počet různých tagů | 207 |

Tabulka 4.1: údaje pro zpracovávané věty ve formátu PDT

| | VC | JH | PDT |
|---|---|---|---|
| Počet různých analytických funkcí | 17 | 15 | 19 |
| Počet různých tagů | 154 | 149 | 207 |
| Počet větných členů | 654 | 654 | 1027 |

Tabulka 4.2: Číselné údaje pro pedagogy (VC a JH) ve srovnání s údaji dle PDT anotace

Samotnou tvorbu pravidel máme rozdělenou do tří částí. V první z nich se zabýváme pouze zapojením jednotlivých slov mezi sebou. Metoda tvorby těchto pravidel a pravidla samotná jsou uvedena v kapitole 4.1. Druhou sadou pravidel

jsou ta, která se týkají převodu analytických funkcí (viz kapitola 4.2). Poslední popisovanou částí jsou pravidla týkající se převodu morfologických tagů (viz kapitola 4.3).

## 4.1 Syntaktická pravidla

Prvním úkolem je zajistit správné závislostní propojení mezi jednotlivými slovními jednotkami (slovo, číslice, interpunkční znaménko), které budeme dále nazývat *tokeny*. V editoru Čapek dále uvažujeme větné členy, které mohou být složeny z více tokenů. Těmto větným členům budeme dále říkat *uzly*. V rámci *xml* struktury jsou značeny tagem `<node>`.

Jako vstup pro transformaci syntaktických pravidel slouží informace z *xml* souboru z editoru Čapek:

- počet tokenů v daném uzlu,
- seznam tokenů patřících do jednoho uzlu,
- pořadí tokenů v daném uzlu,
- závislosti mezi uzly,
- analytické funkce (jedna funkce pro jeden uzel),
- morfologické tagy (jeden tag pro jeden uzel).

Syntaktická pravidla nejsou na rozdíl od morfologických pravidel tvořena automaticky. V editoru Čapek jsou totiž na rozdíl od formátu PDT zaznamenány pouze závislosti větných členů mezi sebou. Čapek nikterak neřeší závislost mezi slovy v rámci jednoho větného členu. Není tedy řešena ani otázka, které ze slov daného větného členu je rozvíjeno slovy obsaženými v závislém větném členu.

Metodou použitou pro tvorbu pravidel byl tzv. odborný pohled, tj. detailní analýza tvaru jazykových rozborů v Čapkovi a jazykových rozborů odpovídajících struktur ve formátu PDT. Pro každý token z editoru Čapek víme, na kterém uzlu je závislý. Nevíme však, k jakému tokenu z příslušného uzlu se daná závislost vztahuje. Školní větné rozbory z editoru Čapek proto potřebujeme modifikovat tak, abychom dostali informaci o vzájemné závislosti mezi všemi tokeny.

Při odvozování pravidel pro transformaci závislostní struktury bylo přihlíženo ke tvaru školních větných rozborů v editoru Čapek, a dále ke tvaru větných rozborů ze zlatého standardu. Závislosti mezi jednotlivými tokeny modelujeme jako zakořeněný strom. Stromy jsou budovány rekurzivně od kořene:

1. V každé větě z editoru Čapek najdeme uzel, který nemá žádného rodiče.
2. Slova v tomto uzlu pospojujeme podle pravidel uvedených v tabulce 4.3, čímž zároveň získáme kořen lokálního stromu pro daný uzel a token, na který se budou vázat závislé uzly.
3. Krok 2 opakujeme pro všechny závislé uzly s tím, že kořen nově vybudovaného stromu (v závislém uzlu) připojujeme na příslušný token z řídícího uzlu.

| Funkce | Počet slov | Pozice slova | Rodič | Otcem |
|---|---|---|---|---|
| Přís-sp | 2 | 1 | 0 | ANO |
| Přís-sp | 2 | 2 | 1 | NE |
| PU | 2 | 1 | 2 | NE |
| PU | 2 | 2 | 0 | ANO |
| Pt | 2 | 1 | 0 | NE |
| Pt | 2 | 2 | 1 | ANO |
| Do | 2 | 1 | 0 | NE |
| Do | 2 | 2 | 1 | ANO |
| Pk | 2 | 1 | 0 | NE |
| Pk | 2 | 2 | 1 | ANO |
| Obecná pravidla | | | | |
| libovolná funkce | libovolný | 1 | 0 | ANO |
| libovolná funkce | libovolný | >1 | 1 | NE |

Tabulka 4.3: Syntaktická pravidla pro převod z Čapka do PML pro tokeny z uzlů neobsahujících sloveso *být* nebo zvratné zájmeno *se*, *si*.

Pravidla zaznamená v tabulce 4.3 interpretujeme následujícím způsobem:

- V levé části tabulky jsou údaje o tokenu, který chceme transformovat:
  - sloupec *Funkce* udává hodnotu analytické funkce přiřazené uzlu, ze kterého token pochází,
  - sloupec *Počet slov* udává počet slov v uzlu, ze kterého token pochází,
  - sloupec *Pozice slova* udává pořadové číslo daného tokenu v rámci uzlu.
- Pravá část tabulky udává transformované hodnoty:
  - sloupec *Rodič* udává pořadí tokenu v rámci daného uzlu, na kterém je vyhodnocovaný token závislý. Pokud je vyhodnocovaný token kořenem lokálního stromu, pak je hodnota ve sloupci *rodič* rovna 0,
  - sloupec *Otcem* udává, zda se na vyhodnocovaný token budou vázat závislé uzly.

Pozice tokenů v daném uzlu se udává jako pořadí slov tak, jak je zadáno v editoru Čapek (v tagu `<ord>`). První slovo má pořadové číslo 1.

K načítané struktuře z editoru Čapek byly přidávány další informace. Například interpunkce je v editoru Čapek připojena k bezprostředně předcházejícímu slovu, zatímco ve formátu PDT je považována za samostatnou slovní jednotku. V průběhu načítání souboru je proto interpunkce vyhledávána a je pro ni vždy vytvořen samostatný token.

## 4.2 Převod analytických funkcí

Dalším úkolem je převedení analytických funkcí zadaných v editoru Čapek na analytické funkce odpovídající koncepci PDT. Jak již bylo zmíněno, koncepce editoru

Čapek a PDT vykazují určité odlišnosti. Prvním příkladem jsou příslovečná určení. V PDT koncepci nerozlišujeme druhy příslovečných určení, máme tedy pouze jednu kategorii, zatímco v ŠVR uvažujeme devět různých druhů příslovečných určení.[1]. Všechny druhy příslovečného určení z editoru Čapek při transformaci sjednocujeme do jednoho druhu ve formátu *PML*.

Obtížnější je převod v situaci, kdy jedné analytické funkci z editoru Čapek můžeme přiřadit více možných analytických funkcí v rámci formátu PDT. Oproti Čapkovi jsou v PDT přidané například funkce pro interpunkci, předložky, zvratná zájmena a modální slovesa.

Pravidla pro převod analytických funkcí vytváříme automaticky s využitím *zlatého standardu* a vět vypracovaných v editoru Čapek. Pracujeme s jednotlivými tokeny. Pro začátek potřebujeme pro každý z tokenů získat analytickou funkci, která byla přiřazena uzlu obsahujícímu daný token. Dále pak při načítání ukládáme informaci o počtu tokenů v uzlu a pořadí daného tokenu v uzlu.

Je-li například již dříve zmíněny větný člen *na houby* označený jako příslovečné určení, pak ke slovu *na* zaznamenáme informaci, že pochází z větného členu o dvou slovech, že se nachází na první pozici a že větný člen, ze kterého pochází, byl označen za příslovečné určení.

Pro tvorbu pravidel přidám ke každému tokenu informaci o analytické funkci převzatou ze *zlatého standardu*. Pravidla získávám pro trojice atributů – počet slov v uzlu, pořadí v uzlu, analytická funkce daného uzlu. Této trojici přiřazuji nejčetněji se vyskytující odpovídající analytickou funkci formátu PDT. Tímto způsobem získám čtveřici obsahující jak původní analytickou funkci, tak funkci dle formátu PDT. Tuto čtveřici nazývám analytickým pravidlem. Výše zmíněným postupem získám sadu pravidel.

Použití vytvořených pravidel je jednoduché. Chci-li převést novou větu vypracovanou v editoru Čapek, pro každý token získám trojici atributů a tu spojím s novou analytickou funkcí pro odpovídající trojici vybranou z pravidel. Tím získám k danému tokenu dosud neznámou analytickou funkci. Příklad vybraných pravidel uvádí tabulka 4.4. Pravidlo 1 například říká, že máme-li jeden token

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Funkce Čapek | Funkce PDT | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | Pk | Atr | 231 |
| 2 | 1 | 1 | Po | Sb | 123 |
| 3 | 1 | 1 | Přís | Pred | 94 |
| 4 | 1 | 2 | Pu-m | AuxP | 94 |
| 5 | 1 | 1 | Pt | Obj | 93 |
| 6 | 2 | 2 | Pu-m | Adv | 74 |

Tabulka 4.4: Příklad pravidel pro určování analytických funkcí.

v uzlu, který je označen jako přívlastek (*Pk*), přiřadíme v PDT tomuto tokenu taktéž přívlastek (*Atr*). Dále vidíme obdobná pravidla pro podmět (*Po*), přísudek (*Přís*) a předmět (*Pt*). Zajímavější je případ víceslovného uzlu. Příkladem je pravidlo 4, které prvnímu slovu ve dvouslovném uzlu označeném jako příslovečné určení místa (*Pu-m*) přiřadí v PDT funkci předložky (*AuxP*). Pravidlo 6

[1]Rozlišení příslovečných určení bylo do editoru přidáno na žádost učitelek. Rozlišení bylo převzato z učebnice [11] Na transformace nemá žádný vliv.

pak druhému slovu v uvažovaném uzlu přiřadí v PDT funkci příslovečného určení (*Adv*).

Výčet všech pravidel pro transformaci analytické funkce je uveden v příloze C. Zatímco počet tokenů v uzlu se počítá bez interpunkce (tak jak je to přirozené v editoru Čapek), pořadí je přiřazováno i interpunkčním znakům (konzistentně se strukturou PDT). Může se proto stát, že pořadové číslo vyhodnocovaného tokenu bude vyšší než celkový počet tokenů v uzlu. Na funkčnost pravidel však tento detail nemá vliv. Pro zajímavost je v tabulce 4.4 i v tabulkách v příloze C uveden také počet tokenů odpovídajících danému pravidlu v datech. Obecně lze pozorovat, že pravidla s vysokou četností tokenů odpovídají přirozené logice, ačkoli bylo sestavování pravidel založené výhradně na zpracování získaných dat. Pravidla s nízkým výskytem mohou na dalších větách fungovat hůře. Na dostupném datovém souboru se bohužel nedají otestovat.

## 4.3 Morfologická pravidla

Jak tagy v editoru Čapek, tak tagy ve formátu PDT, jsou strukturované.[2] Mají však rozdílný počet pozic v tagu a rozdílné údaje na různých pozicích. Tabulka 4.5 popisuje skupiny tagů, které na sebe převádíme. Převodní tabulka vznikla na základě analýzy obou struktur. Při návrhu těchto pravidel tedy nebyla zapotřebí informace z nasbíraných dat.

Výsledkem je sedm skupin morfologických pravidel, která vytváříme zvlášť. Mohli bychom postupovat stejně jako v případě analytických funkcí a pracovat pouze s celými morfologickými tagy. Různých variant morfologických tagů v obou strukturách je však příliš mnoho a tento postup by vedl k velmi vysokému počtu převodních pravidel. Kromě toho by jednotlivé varianty tagů byly nedostatečně zastoupeny. V použitých 101 větách máme v PDT koncepci 207 různých morfologických tagů. Tím, že tagy rozdělíme a mapujeme na sebe skupiny pozic uvedených v tabulce 4.5, získáváme více dat pro tvorbu pravidel.

| | Pozice v tagu v Čapkovi | Pozice v PDT tagu |
|---|---|---|
| 1 | 1 | 1, 2 |
| 2 | 4 | 8 |
| 3 | 5 | 3 |
| 4 | 6 | 4 |
| 5 | 7 | 5 |
| 6 | 8 | 10 |
| 7 | 9 | 9 |

Tabulka 4.5: Mapování pozic tagů editoru Čapek na poziční systém využívaný v PDT.

Při tvorbě transformačních pravidel pro morfologické tagy vycházím z podobné úvahy jako v odstavci 4.2. Rozdílem je, že tvořím sedm skupin pravidel a místo funkce větného členu využívám hodnot z příslušné pozice morfologického tagu. V případě prvního pravidla není nově hledaná hodnota jen jedna, ale tvoří

[2]Například na prvním místě je hodnota odpovídající slovnímu druhu tagovaného slova

ji dvojice atributů. Tyto dvojice získáme obdobným způsobem, jako jsme získávali hodnoty pro tvorbu pravidel pro převod analytických funkcí. První skupina morfologických pravidel bude tedy složena z pětic atributů. Příklad tří pravidel je uveden v tabulce 4.6. Pravidlo číslo 1 například říká, že máme-li jednoslovný uzel, který je v Čapkovi označen jako podstatné jméno, přiřadíme mu ve formátu PDT taktéž podstatné jméno. Totéž platí pro druhé slovo ve dvouslovném větném členu (pravidlo 2). Pravidlo 3 říká, že prvnímu slovu ve dvouslovném větném členu označeném v Čapkovi jako podstané jméno přiřadíme předložku. Tato pravidla odpovídají logickému očekávání, přesto jsou čistě výsledkem automatického zpracování naměřených dat.

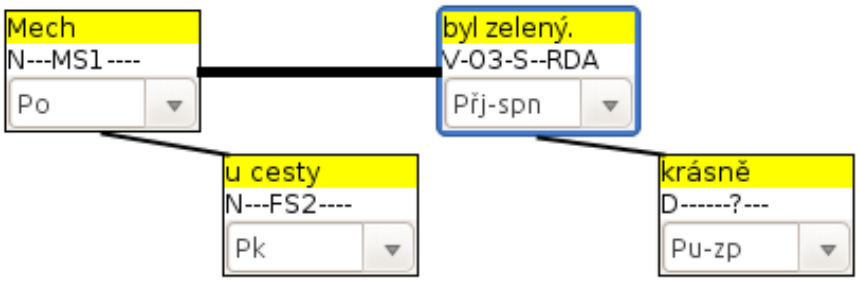
| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag1 | PDT tag1 | PDT tag2 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | N | N | N |
| 2 | 2 | 2 | N | N | N |
| 3 | 1 | 2 | N | R | R |

Tabulka 4.6: Příklad pravidel pro určování slovního druhu.

V dalších skupinách postupuji obdobným způsobem. Seznam všech vytvořených morfologických pravidel je uveden v příloze D.

## 4.4 Příklad aplikace transformačních pravidel

Abychom získali lepší představu o fungování transformace, ukážeme si příklad na jedné vybrané větě. Obrázek 4.1 ukazuje grafickou podobu věty „Mech u cesty byl krásně zelený.“. Na tuto větu aplikujeme transformační pravidla a srovnáme se strukturou této věty v PDT.




Obrázek 4.1: Grafická podoba věty „Mech u cesty byl krásně zelený.“ v editoru Čapek.

Nejprve aplikujeme na jednotlivá slova pravidla pro transformaci analytických funkcí a morfologických kategorií. To ilustrují tabulky 4.7 až 4.12. V tabulkách je pro každé slovo vždy uvedena původní hodnota v Čapkovi (u pozic morfologických tagů je vždy uvedena pozice odpovídající dané pozici v PDT, tj. pozice již převedená dle tabulky 4.5), číslo aplikovaného pravidla (viz přílohy C, D) a transformovaná hodnota.

| „Mech“: pořadí 1 z celkem 1 tokenu v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Po | N | M | S | 1 | - | - | - |
| Pravidlo | 2 | 1 | 3 | 1 | 2 | 1 | 1 | 1 |
| Funkce v PDT | Sb | NN | I | S | 1 | - | - | - |

Tabulka 4.7: Transformace slova *Mech* ve větě „Mech u cesty byl krásně zelený.“

| „u“: pořadí 1 z celkem 2 tokenů v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Pk | N | F | S | 2 | - | - | - |
| Pravidlo | 10 | 2 | 4 | 4 | 10 | 2 | 3 | 2 |
| Funkce v PDT | AuxP | RR | - | - | 2 | - | - | - |

Tabulka 4.8: Transformace slova *u* ve větě „Mech u cesty byl krásně zelený.“

| „cesty“: pořadí 2 z celkem 2 tokenů v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Pk | N | F | S | 2 | - | - | - |
| Pravidlo | 17 | 3 | 6 | 3 | 11 | 3 | 2 | 3 |
| Funkce v PDT | Atr | NN | F | S | 2 | - | - | - |

Tabulka 4.9: Transformace slova *cesty* ve větě „Mech u cesty byl krásně zelený.“

| „byl“: pořadí 1 z celkem 2 tokenů v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Přj-spn | V | - | S | - | 3 | R | - |
| Pravidlo | 8 | 10 | 5 | 4 | 4 | 8 | 8 | 2 |
| Funkce v PDT | Pred | Vp | - | - | - | X | R | - |

Tabulka 4.10: Transformace slova *byl* ve větě „Mech u cesty byl krásně zelený.“

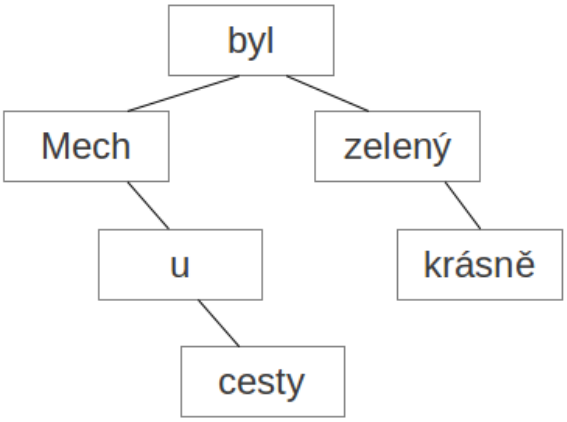
| „krásně“: pořadí 1 z celkem 1 tokenu v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Přj-spn | D | - | - | - | - | - | ? |
| Pravidlo | 9 | 5 | 1 | 2 | 1 | 1 | 1 | 6 |
| Funkce v PDT | Adv | Db | - | - | - | - | - | - |

Tabulka 4.11: Transformace slova *krásně* ve větě „Mech u cesty byl krásně zelený.“

| „zelený“: pořadí 1 z celkem 2 tokenů v uzlu | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An. fun. | tag1-2 | tag3 | tag4 | tag5 | tag8 | tag9 | tag10 |
| Čapek | Přj-spn | V | - | S | - | 3 | R | - |
| Pravidlo | 11 | 12 | 13 | 3 | 5 | 7 | 10 | 3 |
| Funkce v PDT | PNom | Vp | - | S | - | - | - | - |

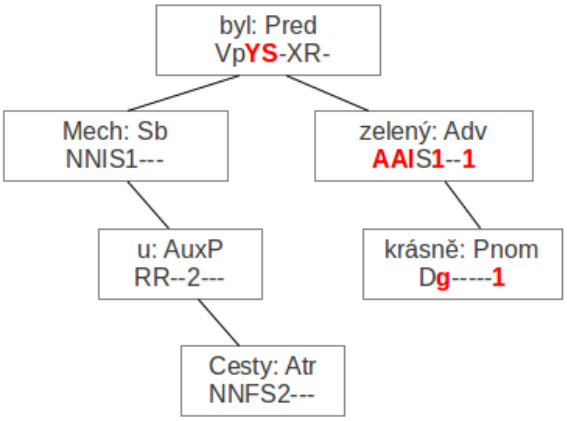
Tabulka 4.12: Transformace slova *zelený* ve větě „Mech u cesty byl krásně zelený.“

Nyní transformujeme stromovou strukturu. Podle tabulky 4.3 určíme stromy ve víceslovných uzlech. V uzlu „byl zelený“ je řídícím tokenem slovo *byl*, na které je zároveň věšen následující uzel. V uzlu „u cesty“ je řídícím tokenem slovo *u*, případný další uzel by se věšel na slovo *cesty*. Výsledná struktura je uvedena na obrázku 4.2




Obrázek 4.2: Transformovaná struktura věty „Mech u cesty byl krásně zelený.“

Zbývá výsledek porovnat s ručně anotovanými daty z PDT. Toto porovnání přináší obrázek 4.3. Vidíme, že struktura stromu je identická, stejně jako analytické funkce. Červeně jsou vyznačené morfologické kategorie, u nichž je transformovaný výsledek jiný.




Obrázek 4.3: Struktura věty „Mech u cesty byl krásně zelený.“ v PDT.

# 5. Vyhodnocení výsledků

V této kapitole jsou vyhodnoceny výsledky práce. Vyhodnocujeme různé aspekty:

1. shoda učitelských rozborů,
2. úspěšnost transformačních pravidel,
3. přesnost žákovských rozborů.

Pro vyhodnocení máme k dispozici následující údaje:

1. dvoje kompletní (všech 101 vět) vypracované rozbory od učitelek ve formátu editoru Čapek,
2. větný rozbor zpracovaný anotátorkou PDT ve formátu PML,
3. rozbory zpracované žáky během školních hodin ve formátu editoru Čapek (maximálně 5 vět zpracovaných jedním žákem, dále jako *žáci ve školách*),
4. dvoje kompletní zpracování vět žáky (dále jako *žáci doma*).

## 5.1 Shoda učitelských rozborů

Vzhledem k tomu, že při tvorbě transformačních pravidel vycházíme z učitelských rozborů, zajímá nás, do jaké míry se rozbory obou učitelek shodují. Tabulka 5.1 uvádí počet (absolutní i relativní) shod v atributech používaných pro transformaci. Relativní hodnoty jsou vztažené k celkovému počtu tokenů zpracovaných oběma učitelkami (1027). Vidíme, že u analytických funkcí se míra shody pohybuje přibližně okolo 80 %. U morfologických pravidel dostáváme přibližně typicky shodu přes 90 %.

| Určovaná kategorie | Počet shod | % shod |
|---|---:|---:|
| Analytická funkce | 807 | 78,58 |
| Slovní druh | 914 | 89,00 |
| Osoba | 953 | 92,79 |
| Rod | 939 | 91,43 |
| Číslo | 916 | 89,19 |
| Pád | 954 | 92,89 |
| Stupeň | 975 | 94,94 |
| Čas | 945 | 92,02 |

Tabulka 5.1: Shoda učitelek při větném rozboru.

Příčinou nízké shody u analytických funkcí by mohlo být například rozlišování druhu příslovečných určení. Když však příslovečná určení nerozlišujeme, dostaneme jen o něco lepší výsledky (823, tj. 80,14 %). Nejčastějším typem neshody je dvojice (přísudek – přísudek jmenný se sponou) – 87 případů, tj. 43 % všech neshod. Z hlediska školního rozboru mohou být z jistého pohledu obě varianty

považovány za správné (přísudek jmenný se sponou je taktéž přísudkem), z pohledu značkování PDT však jde o nezanedbatelný rozdíl. Další četnější neshody byla záměna příslovečného určení místa a přívlastku (11 případů) a situace, kdy jedna učitelka označila přívlastek a druhá pro stejné slovo analytickou funkci neuvedla (11 případů). Ostatní typy neshod byly zastoupeny v počtu menším než 10.

## 5.2 Úspěšnost transformačních pravidel

Při testování úspěšnosti transformačních pravidel považujeme za správná data větné rozbory zpracované anotátorkou PDT. Jako vstupní data pro vyhodnocování používáme rozbory vypracované učitelkami v editoru Čapek.

### 5.2.1 Analytická a morfologická pravidla

Transformační pravidla pro analytické funkce a morfologické kategorie byla konstruována automaticky na základě dat. Pro vyhodnocování bylo proto nutno data rozdělit na *učící* a *vyhodnocovací*. Na *učících* datech jsou pravidla vytvořena, na *vyhodnocovacích datech* je pak testována úspěšnost. Pro přehled o vlivu rozsahu dat bylo použito více rozdělení:

1. 80 učících vět (ozn. Tr80), 21 testovacích vět (ozn. Te80),
2. 70 učících vět (ozn. Tr70), 31 testovacích vět (ozn. Te70),
3. 50 učících vět (ozn. Tr50), 51 testovacích vět (ozn. Te50).

Věty do učících a testovacích souborů byly voleny náhodně (pochopitelně bez opakování), přičemž platí Tr50 $\subset$ Tr70 $\subset$ Tr80. Jako kritérium úspěšnosti slouží podíl správného přiřazení daného atributu v rámci dané množiny dat. Porovnává se tedy počet správně transformovaných tokenů (v tabulkách ve sloupci *Úsp.*) s celkovým počtem tokenů v daném souboru, kterým je přiřazena daná funkce v editoru Čapek (v tabulkách ve sloupci *Celk.*). Pro informaci o rozsahu dané množiny je uveden také absolutní počet správných přiřazení.

Výsledky na učících datech pro všechny uvažované varianty jsou uvedeny v tabulce 5.2. Z hodnot relativních úspěšností je vidět jistá robustnost, úspěšnosti v dané kategorii jsou obdobné pro všechny tři testované množiny dat.

| | Tr50 | | | Tr70 | | | Tr80 | | |
|---|---|---|---|---|---|---|---|---|---|
| Transf. | Úsp. | Cel. | % | Úsp. | Cel. | % | Úsp. | Cel. | % |
| Analytická funkce | 755 | 992 | 76.11 | 1075 | 1433 | 75.02 | 1215 | 1659 | 73.24 |
| Slovní druh | 703 | 1010 | 69.60 | 1029 | 1488 | 69.15 | 1162 | 1693 | 68.64 |
| Rod | 727 | 976 | 74.49 | 1083 | 1427 | 75.89 | 1235 | 1621 | 76.19 |
| Číslo | 822 | 978 | 84.05 | 1218 | 1461 | 83.37 | 1383 | 1659 | 83.36 |
| Pád | 847 | 976 | 86.78 | 1257 | 1427 | 88.09 | 1429 | 1621 | 88.16 |
| Osoba | 854 | 997 | 85.66 | 1242 | 1427 | 87.04 | 1411 | 1621 | 87.05 |
| Čas | 909 | 976 | 93.14 | 1325 | 1427 | 92.85 | 1505 | 1621 | 92.84 |
| Stupeň | 910 | 976 | 93.24 | 1334 | 1427 | 93.48 | 1518 | 1621 | 93.65 |

Tabulka 5.2: Úspěšnost transformace – učící data

Tabulka 5.3 uvádí výsledky úspěšnosti transformace na testovacích datech. Zhoršení oproti trénovacím datům není výrazné (v ojedinělých případech dokonce pozorujeme vyšší úspěšnost, to však může být matoucí, neboť podíly jsou s výjimkou souborů Tr50 a Te50 počítány z různých základů). Vzhledem k vyrovnaným výsledkům lze očekávat, že při použití pravidel naučených na všech dostupných datech budeme dostávat obdobně kvalitní výsledky na nových větách.

| | Te50 | | | Te70 | | | Te80 | | |
|---|---|---|---|---|---|---|---|---|---|
| Transf. | Úsp. | Cel. | % | Úsp. | Cel. | % | Úsp. | Cel. | % |
| Analytická funkce | 703 | 1049 | 67.02 | 400 | 601 | 66.56 | 282 | 431 | 65.43 |
| Slovní druh | 710 | 1098 | 64.66 | 396 | 640 | 61.88 | 263 | 435 | 60.46 |
| Rod | 816 | 1053 | 77.49 | 468 | 610 | 76.72 | 318 | 418 | 76.08 |
| Číslo | 872 | 1040 | 83.85 | 516 | 628 | 82.17 | 351 | 430 | 81.63 |
| Pád | 943 | 1050 | 89.81 | 542 | 608 | 89.14 | 370 | 414 | 89.37 |
| Osoba | 922 | 1089 | 84.66 | 526 | 611 | 86.09 | 357 | 417 | 85.61 |
| Čas | 982 | 1052 | 93.35 | 563 | 604 | 93.21 | 383 | 410 | 93.41 |
| Stupeň | 1004 | 1061 | 94.63 | 582 | 612 | 95.10 | 398 | 418 | 95.22 |

Tabulka 5.3: Úspěšnost transformace – testovací data

### 5.2.2 Syntaktická pravidla

V tabulce 5.4 je vyhodnocena úspěšnost syntaktických vyhodnocovacích pravidel. Kritériem je podíl správně přiřazených řídících tokenů, což odpovídá správnému přiřazení vazby mezi dvojicemi tokenů, ku celkovému počtu tokenů.

| Poč. správných | Poč. celkem | % správných |
|---|---|---|
| 760 | 2054 | 37.00 |

Tabulka 5.4: Úspěšnost transformace – syntaktická pravidla

Je vidět, že ve srovnání s transformací analytických funkcí a morfologických kategorií je úspěšnost transformace syntaktické struktury nízká.

## 5.3 Přesnost žákovských rozborů

Základní myšlenkou této práce je využití žákovských rozborů pro značkování korpusu. Žákovské rozbory však nebyly použity pro tvorbu pravidel. Důvodem byla snaha o minimalizaci chyb v rozborech a předpoklad, že učitelé by měli být v rozborech více zběhlí. Pro případné využití žákovských rozborů pro značkování je tedy důležité mít alespoň představu o chybovosti v jednotlivých transformovaných kategoriích.

Při testování přesnosti žákovských rozborů bereme jako referenční data z rozborů zpracovaných učitelkami. Učitelky zpracovávaly rozbory podle pravidel odpovídajících školním větným rozborům, tj. stejným způsobem jako žáci. Ačkoli se výsledky obou učitelek liší, jak bylo uvedeno v kapitole 4 a jak potvrzují výsledky testů v odstavci 5.1, považujeme při vyhodnocení obě varianty za správné. Důvodem je především absence jakýchkoli podkladů pro rozhodnutí, kterému z řešení

dát v případě neshody přednost. Odlišnosti mohou být například způsobeny víceznačností zvolených vět. Ačkoli nejsme schopni postihnout všechny významy zadaných vět, dvěma nezávislými zpracováními zvyšujeme možnost odhalení více významů.

Za správné vyhodnocení daného větného členu zpracovaného žákem tedy považujeme takové vyhodnocení, které se shoduje aspoň s jedním učitelským vyhodnocením. Testují se jak žákovské rozbory, zpracované ve školách, tak kompletní rozbory dvou žáků získané nad rámec výuky.

### 5.3.1 Přesnost přiřazení analytických funkcí

Tabulka 5.5 zobrazuje přesnost přiřazování analytických funkcí. Hodnotí se počet správně přiřazených tokenů (tj. shoda aspoň s jednou z učitelek) vůči celkovému počtu tokenů, kterým žáci přiřadili danou analytickou funkci. Jinými slovy se jedná o odhad pravděpodobnosti, že pokud žák přiřadí danou analytickou funkci, je tato funkce přiřazena správně. U některých tokenů žáci ponechali analytickou funkci nevyplněnou. Takové tokeny byly z vyhodnocení vyřazeny. Pro představu je však v tabulce (stejně jako ve všech dalších tabulkách) uvedena celková přesnost i se zahrnutím nevyplněných polí. Hodnotu je třeba brát jako orientační, neboť v případě, že jedna z učitelek taktéž nevyplnila analytickou funkci, byla tato prázdná hodnota považována za správnou.

| Analytická funkce | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| | správně | celkem | % správ. | správně | celkem | % správ. |
| podmět | 238 | 289 | 82,35 | 233 | 254 | 91,73 |
| přísudek | 429 | 508 | 84,45 | 737 | 946 | 77,91 |
| přísl. určení | 425 | 500 | 85,00 | 653 | 768 | 85,03 |
| přívlastek | 416 | 473 | 87,95 | 474 | 533 | 88,93 |
| předmět | 170 | 247 | 68,83 | 350 | 595 | 58,82 |
| doplněk | 34 | 54 | 62,96 | 0 | 44 | 0,00 |
| celkem | 1712 | 2071 | 82,67 | 2464 | 3179 | 77,51 |
| vč. nevyplněných | 1763 | 3391 | 51,99 | 2523 | 3682 | 68,52 |
| přísl. urč. typ | 237 | 500 | 47,40 | 596 | 768 | 77,60 |

Tabulka 5.5: Přesnost přiřazování analytických funkcí v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

Povšimněme si, že žáci ve školách měli celkově lepší relativní úspěšnost. Přitom je však třeba přihlédnout k tomu, že celkový počet zpracovaných tokenů byl u žáků ve školách nižší. Důvodem vyšší relativní úspěšnosti může být to, že si žáci ve školách volili přednostně jednodušší věty, které uměli zpracovat. Žáci doma byli naopak požádáni o zpracování všech vět, tudíž museli řešit i obtížnější rozbory, se kterými se v průběhu výuky, dle jejich tvrzení, nesetkali. Žáci doma byli úspěšnější především v určování podmětu, naopak neurčili žádný doplněk. Zajímavým typem je příslovečné určení, kde žáci rozlišovali celkem 7 druhů této analytické funkce. Úspěšnost přiřazení příslovečného určení byla primárně vyhodnocena bez rozlišení těchto druhů (ve formátu PDT toto rozlišení není). Výsledky žáků doma a ve školách jsou při tomto vyhodnocení rovnocenné. Pokud však přihlédneme k rozlišení typů příslovečného určení (poslední řádek tabulky), vidíme, že žáci

doma měli výrazně vyšší úspěšnost. Usuzujeme, že žáci doma věnovali správnosti přiřazení více času a pozornosti.

### 5.3.2 Přesnost přiřazení morfologických tagů

Pravidla pro přiřazování morfologických tagů jsou konstruována odděleně pro jednotlivé pozice v tagu. Z toho důvodu bylo stejným způsobem rozdělení i vyhodnocování přesnosti žáků.

Přesnost přiřazení slovních druhů je vyhodnocena v tabulce 5.6. Z celkového pohledu vykazují žáci pracující doma mírně lepší výsledky. Rozdíly mezi oběma skupinami žáků jsou však patrné pro jednotlivé slovní druhy. Zatímco například podstatná jména určují relativně úspěšně obě skupiny žáků, žáci doma výrazně přesněji určují přídavná jména, číslovky a spojky. Žáci ve školách „lépe“ určují předložky. Určení předložky je však typicky v rozporu s pravidly pro práci s editorem Čapek, uvedenými v odstavci 2.3.1. Žáci doma postupovali podle pokynů a tudíž neurčili žádnou předložku. Obdobný důvod může mít i vyšší úspěšnost v určování infinitivu.

| Slovní | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| druh | správně | celkem | % správ. | správně | celkem | % správ. |
| podst. jm. | 736 | 768 | 95,83 | 706 | 735 | 96,05 |
| příd. jm. | 213 | 260 | 81,92 | 293 | 317 | 92,43 |
| zájmeno | 161 | 176 | 91,48 | 300 | 314 | 95,54 |
| číslovka | 30 | 35 | 85,71 | 68 | 70 | 97,14 |
| sloveso | 387 | 395 | 97,97 | 519 | 545 | 95,23 |
| infinitiv | 46 | 74 | 62,16 | 21 | 48 | 43,75 |
| příslovce | 134 | 154 | 87,01 | 227 | 299 | 75,92 |
| předložka | 10 | 43 | 23,26 | 0 | 12 | 0,00 |
| spojka | 35 | 60 | 58,33 | 86 | 122 | 70,49 |
| částice | 6 | 16 | 37,50 | 0 | 8 | 0,00 |
| celkem | 1758 | 1981 | 88,74 | 2220 | 2470 | 89,88 |
| vč. nevyplněných | 1871 | 3391 | 55,18 | 2309 | 3682 | 62,71 |

Tabulka 5.6: Přesnost přiřazování slovních druhů v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

Přesnost určování rodu zobrazuje tabulka 5.7. Zde pozorujeme výrazně vyšší úspěšnost žáků ve školách. Zajímavé je, že zatímco žáci ve školách rovnoměrně úspěšně určují všechny rody, žáci doma mají výrazně vyšší úspěšnost v určování rodu ženského. Z toho se dá usuzovat, že zaměňují mužský a střední rod. Rozdíl mezi skupinami žáků může být taktéž způsoben skutečností, že žáci na školách nepracovali na celém souboru a mohli tedy určovat rod v pro ně jednodušších případech (jako jsou například podstatná jména).

V tabulce 5.8 je uvedena přesnost určování čísla. Výsledky jsou relativně dobré oproti ostatním morfologickým kategoriím s mírnou převahou žáků pracujících ve školách. Povšimněme si také poměrně vysoké úspěšnosti vyplnění této hodnoty.

Přesnost určování pádu uvádí tabulka 5.9. Za pozornost stojí velmi nízká přesnost určování 3. pádu u žáků ve škole. To však může být způsobeno celkově nízkou četností třetího pádu ve vyhodnocovaných větách.

| Rod | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| | správně | celkem | % správ. | správně | celkem | % správ. |
| mužský | 541 | 582 | 92,96 | 593 | 718 | 82,59 |
| ženský | 333 | 360 | 92,50 | 467 | 491 | 95,11 |
| střední | 159 | 176 | 90,34 | 169 | 210 | 80,48 |
| celkem | 1033 | 1118 | 92,40 | 1229 | 1419 | 86,61 |
| vč. nevyplněných | 1109 | 1239 | 89,51 | 1233 | 1436 | 85,86 |

Tabulka 5.7: Přesnost přiřazování rodu v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

| Číslo | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| | správně | celkem | % správ. | správně | celkem | % správ. |
| jednotné | 1101 | 1138 | 96,75 | 1409 | 1501 | 93,87 |
| množné | 354 | 368 | 96,20 | 435 | 471 | 92,36 |
| celkem | 1455 | 1506 | 96,61 | 1844 | 1972 | 93,51 |
| vč. nevyplněných | 1503 | 1634 | 91,98 | 1844 | 1981 | 93,08 |

Tabulka 5.8: Přesnost přiřazování čísla v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

| Pád | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| | správně | celkem | % správ. | správně | celkem | % správ. |
| první | 313 | 351 | 89,17 | 392 | 434 | 90,32 |
| druhý | 186 | 205 | 90,73 | 229 | 266 | 86,09 |
| třetí | 12 | 20 | 60,00 | 65 | 71 | 91,55 |
| čtvrtý | 206 | 218 | 94,50 | 238 | 303 | 78,55 |
| pátý | 8 | 8 | 100,00 | 0 | 0 | 0,00 |
| šestý | 180 | 184 | 97,83 | 224 | 228 | 98,25 |
| sedmý | 132 | 134 | 98,51 | 115 | 123 | 93,50 |
| celkem | 1037 | 1120 | 92,59 | 1263 | 1425 | 88,63 |
| vč. nevyplněných | 2420 | 3391 | 71,37 | 2883 | 3682 | 78,30 |

Tabulka 5.9: Přesnost přiřazování pádu v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

Přesnost určování osoby uvádí tabulka 5.10. Zajímavé je, že zatímco žáci doma určili správně první osobu ve všech případech, v určování druhé osoby měli pouze šedesátiprocentní úspěšnost. Žáci ve škole naopak dobře určovali třetí osobu, což byla zároveň nejzastoupenější hodnota.

| | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| Osoba | správně | celkem | % správ. | správně | celkem | % správ. |
| první | 96 | 109 | 88,07 | 80 | 80 | 100,00 |
| druhá | 36 | 50 | 72,00 | 26 | 43 | 60,47 |
| třetí | 217 | 223 | 97,31 | 374 | 416 | 89,90 |
| celkem | 349 | 382 | 91,36 | 480 | 539 | 89,05 |
| vč. nevyplněných | 351 | 395 | 88,86 | 480 | 545 | 88,07 |

Tabulka 5.10: Přesnost přiřazování osoby v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

V tabulce 5.11 vidíme přesnost přiřazování stupně. Ta je překvapivě nízká, a to přestože se v této kategorii učitelky nejčastěji neshodly, šance na „úspěšné“ přiřazení se tedy zvyšují.

| | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| Stupeň | správně | celkem | % správ. | správně | celkem | % správ. |
| první | 179 | 244 | 73,36 | 244 | 591 | 41,29 |
| druhý | 2 | 14 | 14,29 | 0 | 0 | 0,00 |
| třetí | 4 | 10 | 40,00 | 9 | 9 | 100,00 |
| celkem | 185 | 268 | 69,03 | 253 | 600 | 42,17 |
| vč. nevyplněných | 290 | 414 | 70,05 | 267 | 616 | 43,34 |

Tabulka 5.11: Přesnost přiřazování stupně v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

V určování časů byli žáci nejúspěšnější v případě času minulého, jak uvádí tabulka 5.12. Zatímco přítomný a budoucí čas lze v některých případech zaměnit. Např. sloveso „jdu“ může referovat jak o činnosti v přítomném čase, např. „já jdu domů“, tak o činnosti v čase budoucím, např. „zítra jdu do kina“. To může být pro žáky matoucí. Vysoká přesnost určování budoucího času v případě žáků ve škole je pravděpodobně opět způsobena výběrem určovaných slov.

| | Žáci ve školách | | | Žáci doma | | |
|---|---|---|---|---|---|---|
| Čas | správně | celkem | % správ. | správně | celkem | % správ. |
| přítomný | 103 | 146 | 70,55 | 150 | 197 | 76,14 |
| minulý | 164 | 168 | 97,62 | 255 | 278 | 91,73 |
| budoucí | 40 | 40 | 100,00 | 42 | 64 | 65,62 |
| celkem | 307 | 354 | 86,72 | 447 | 539 | 82,93 |
| vč. nevyplněných | 335 | 395 | 84,81 | 447 | 545 | 82,02 |

Tabulka 5.12: Přesnost přiřazování času v žákovských rozborech. Položka *celkem** zahrnuje i nevyplněné údaje.

# Závěr

V rámci diplomové práce byla získána řada školních větných rozborů. Část rozborů byla získána od žáků přímo při výuce, úplnější vypracování byla získána od dvou žákyň, které se vypracování věnovaly ve volném čase. Jako referenční údaje byly dále pořízeny rozbory vypracované dvěma středoškolskými učitelkami českého jazyka.

Sběr dat proběhl pomocí funkčního prototypu editoru větných rozborů Čapek vyvíjeného v Ústavu Formální a Aplikované Lingvistiky MFF UK. V rámci získávání dat byla zjišťována i spokojenost s editorem. Byly sledovány reakce na uživatelské rozhraní i na vlastní problematiku větného rozboru. Z reakcí žáků vyplývá, že práce s editorem pro ně není tak lákavá, jak jsme očekávali. Důvodem může být i skutečnost, že žáci obecně velmi málo pracují s počítačem a počítač typicky vnímají jako zdroj zábavy, nikoli jako pracovní nástroj. S tím souvisely i technické problémy žáků při práci s editorem. Pokud by měl být editor rutinně používán ve školách, bylo by zapotřebí žáky důkladněji proškolit v práci s editorem. Přinejmenším by pomohlo, kdyby žáci běžně ovládali základní pracovní úkony na počítači, jako je například práce s textovým editorem. Na základě reakcí učitelů byly provedeny některé funkční změny v editoru.

Na základě sesbíraných dat byla vytvořena pravidla pro transformaci školních větných rozborů do formátu *PML* používaného v Pražském závislostním korpusu. Pravidla byla vytvořena kombinací expertního přístupu a automatického zpracování větných rozborů. Pro zlepšení úspěšnosti transformace v morfologické rovině by bylo vhodné upravit editor Čapek tak, aby byly morfologické kategorie vázány na slovo, nikoli na větný člen, jak je tomu doposud. Z časových důvodů bohužel nebylo možné tuto úpravu provést před zahájením sběru dat. Pravidla byla tvořena na základě učitelských rozborů, u nichž byla očekávána nižší chybovost.

Navržená transformační pravidla byla podrobena testování. Dostupná data byla rozdělena na učící a testovací v různých poměrech pro otestování vlivu velikosti učícího vzorku. Výsledky byly pro všechny testované poměry srovnatelné, a to jak na učících datech, tak na testovacích datech. U analytických funkcí dosahujeme úspěšnosti kolem 70 %. Pravidla pro transformaci morfologických kategorií byla vesměs úspěšnější, výjimkou je transformace slovního druhu, která byla nejméně úspěšná (přesnost cca 60 % na testovacích datech). V některých morfologických kategoriích dosahujeme úspěšnosti přes 90 %.

Základní myšlenkou práce je využití žákovských rozborů pro značkování. Proto byla také vyhodnocována chybovost žáků při větných rozborech. Chybovost byla posuzována vůči učitelským rozborům. Analytické funkce přiřazují žáci s přibližně 80% přesností. Přesnost přiřazení morfologických kategorií je různorodá. Žáci dosahují relativně dobrých výsledků (přibližně 90% úspěšnost) při určování slovního druhu, rodu, čísla, pádu a osoby. Čas určují o něco hůře (úspěšnost kolem 85 %), což je dáno zejména problémy s rozlišováním přítomného a budoucího času. Zdaleka nejhorší výsledky mají žáci při přiřazování stupně. Výsledky však mohou být zkreslené nízkým zastoupením slov, u nichž se tato kategorie určuje.

Školní větné rozbory jsou možným zdrojem anotace závislostních korpusů. Před započetím jejich rozsáhlejšího využití je však třeba vyřešit několik úkolů. Především by bylo zapotřebí získat velkou podporu vyučujících českého jazyka.

Vyučující by měli být motivováni používat pro výuku jazykových rozborů počítač. Zajímavou motivací by například bylo, kdyby používaný program umožňoval snadnou a rychlou kontrolu žákovských rozborů proti zadanému vzoru. Zároveň by bylo vhodné přizpůsobit vzdělávací plány v oblasti informatiky tak, aby žáci získali nezbytné základy práce s počítačem. Úspěšnost značkování pomocí školních větných rozborů není v současné době oslnivá, nicméně by bylo například zajímavé otestovat kombinaci využití žákovských rozborů a automatického parseru.

# Seznam použité literatury


[1] Bunce, Tim. *DBI* [online]. URL: <http://search.cpan.org/∼timb/DBI-1.616/DBI.pm> [cit. 2011-12-04].

[2] Hajič, Jan a kol. *Průvodce PDT 2.0* [online], 2006. URL:<http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/pdf/pdt-guide.pdf> [cit. 2011-11-28].

[3] Hajič, Jan. *Positional Tags:* Quick Reference (Czech „HM“ Morphology) [online]. URL:<http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html> [cit. 2011-04-12.

[4] Hajič, Jan a kol. *PDT 2.0 Annotation Markup Reference* [online]. URL:<http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml-markup/index.html> [cit. 2011-10-22].

[5] Hajič, Jan; Panevová, Jarmila; Buráňová Eva; Urešová Zdeňka; Bémová, Alla. *Anotace na analytické rovině*: Návod pro anotátory. [online] ÚFAL MFF UK Praha, 1999. URL: <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/pdf/a-man-cz.pdf> [ cit. 2011-11-28].

[6] Hana, Jiří; Zeman, Daniel a kol. *Manual for Morphological Annotation.* Revision for the Prague Dependency Treebank 2.0, ÚFAL Technical Report No. 2005-27 [online]. URL:<http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html> [cit. 2011-10-22].

[7] Hudson, R. ; Walmsley, J. *The English Patient: English grammar and teaching in the twentieth century.* Journal of Linguistics 43.3, 2005, pp. 593–622.

[8] Kučera, Ondřej. *Pražský závislostní korpus jako cvičebnice jazyka českého.* Diplomová práce, MFF UK, 2006.

[9] Pajas, Petr; Štěpánek, Jan. *XML-Based Representation of Multi-Layered Annotation in the PDT 2.0.* In Proceedings of the LREC Workshop on Merging and Layering Linguistic Information LREC, Genova, 2006, pp. 40–47. ISBN 2-9517408-2-4.

[10] Sergeant, Matt. *XML-XPath-1.13* [online]. URL: <http://search.cpan.org/∼msergeant/XML-XPath-1.13/XPath.pm> [cit. 2011-12-04].

[11] Styblík, Vlastimil a kol. *Český jazyk: Pro 8. ročník základní školy.* 2. vydání. SPN, Praha, 2005. ISBN 80-7235-126-5.

[12] Vidová Hladká, Barbora; Hajič, Jan; Hana, Jirka; Hlaváčová, Jaroslava; Mírovský, Jiří; Raab, Jan. *The Czech Academic Corpus 2.0 Guide.* The Prague Bulletin of Mathematical Linguistics, MFF UK, 2008.

[13] *Extensible Markup Language* [online]. URL: <http://en.wikipedia.org/wiki/XML> [cit. 2011-10-30].

[14] *Fin de l'exercice de français Analyse de la phrase: Exercices pratiques (1)* [online]. URL: <http://www.francaisfacile.com/exercices/exercice-francais-2/exercice-francais-34348.php> [cit. 2011-12-06].

[15] *Kanton St.Gallen Wirtschaftsmittelschule mit Schwerpunkt Sprachen (WMS) mit Schwerpunkt Informatik (WMI) Lehrplan.* Erziehungsrat des Kantons St.Gallen [online], 2011. URL: <http://www.schule.sg.ch/home/mittelschule/ausbildungsgaenge/wirtschaftsmittelschule/_jcr_content/Par/downloadlist/DownloadListPar/download_2.ocFile/WMS%20Lehrplan%202011.pdf> [cit. 2011-12-06], p. 21.

[16] *Lehrplan für die Volksschule des Kantons Zürich* [online]. URL: <http://www.vsa.zh.ch/internet/bildungsdirektion/vsa/de/schulbetrieb_und_unterricht/faecher_lehrplaene_lehrmittel0/_jcr_content/contentPar/downloadlist_1/downloaditems/lehrplan.spooler.download.1291724804933.pdf> [cit. 2011-12-06].

[17] *Lehrplannavigator Grundschule: Deutsch – Kompetenzen.* Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2011 [online], rev. 2009-08-13. URL: <http://www.standardsicherung.schulministerium.nrw.de/lehrplaene/lehrplaene-gs/deutsch/lehrplan-deutsch/kompetenzen/kompetenzen.html> [cit. 2011-12-06].

[18] *Lehrplannavigator Gymnasium: Deutsch – Kompetenzen.* Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2011 [online], rev. 2011-03-09. URL: <http://www.standardsicherung.schulministerium.nrw.de/lehrplaene/kernlehrplaene-sek-i/gymnasium-g8/deutsch-g8/kernlehrplan-deutsch/kompetenzen/kompetenzen.html> [cit. 2011-12-06].

[19] *The National Curriculum for English.* Department for Education, 2011 [online]. URL: <http://www.education.gov.uk/schools/teachingandlearning/curriculum/primary/b00198874/english> [cit. 2011-12-06].

[20] *Signets par classes* [online], rev. 2011-08-28. URL: <http://www.pomverte.com/Signets.htm> [cit. 2011-12-06].

[21] *Školský vzdelávací program: ISCED 3A – gymnázium.* Gymnázium sv. Uršule, Bratislava, 2010 [online]. URL: <http://www.gsurba.sk/pdf/dokumenty/svp.pdf> [cit. 2011-12-06]., str. 27-28

[22] *Učebné osnovy gymnázia: osemročné štúdium.* Štátny pedagogický ústav, Bratislava, 1997 [online]. URL: <http://www.statpedu.sk/files/documents/nereformne_rocniky/uo_8r_gym_sj_a_literatura.pdf> [cit. 2011-12-06].

# Seznam tabulek

# Seznam použitých zkratek

**PDT** Prague Dependency Treebank

**PML** Prague Markup Language

**ŠVR** Školní větný rozbor

**XML** Extensible Markup Language [13]

# Přílohy

# A. Dotazníky

V této příloze jsou uvedeny dotazníky pro získání zpětné vazby o pohodlnosti a smysluplnosti práce s editorem Čapek. Dotazníky byly vyplněny učiteli a žáky po dokončení práce s editorem. Získané výsledky jsou shrnuty v kapitole 3.

## A.1 Dotazník pro žáky

1. Napiš typ školy
   (a) základní škola
   (b) osmileté gymnázium
   (c) šestileté gymnázium
   (d) jiné
2. V kolikátém ročníku jsi? (Pokud jsi na víceletém gymnáziu, vyplň prosím odpovídající ročník základní školy.)
   (a) 6
   (b) 7
   (c) 8
   (d) 9
3. Jak se ti editor ovládal? (1–5)
4. Jak se ti líbil vzhled editoru? (1–5)
5. S jakou částí práce na procvičované látce jsi měl(a) potíže?
   (a) Vše zvládám levou zadní.
   (b) Mám problémy s určováním slovních druhů a jejich kategorií.
   (c) Mám problémy s určováním větných členů.
   (d) Nevím, kam zavěšovat jednotlivé větné členy.
   (e) Pořádně jsem se zapotil(a).
6. Pokud by sis měl(a) vybrat, jak bys zpracovával(a) domácí úkol z větných rozborů, pracovala bys:
   (a) rozhodně v editoru,
   (b) asi v editoru,
   (c) neumím se rozhodnout,
   (d) asi na papíře,
   (e) rozhodně na papíře.
7. Zdůvodni prosím jednou větou svůj výběr v předchozí otázce.
8. Pokud bys nám chtěl(a) sdělit své nápady a připomínky, prosím piš.

## A.2 Dotazník pro učitele

1. Jak se vám v programu pracovalo?
2. Vidíte nějaké přednosti oproti větnému rozboru na papír? Jaké?
3. Zadával(a) byste žákům vypracovávat úkoly v tomto programu?
   (a) Ano
   (b) Ne
4. Co vám na programu nejvíce vadilo, případně co byste udělal(a) jinak?
5. Myslíte si, že je to pro školy využitelný nástroj?
   (a) Ano
   (b) Ne
6. Řeknete o programu svým kolegům češtinářům?
   (a) Ano
   (b) Ne
7. Na tomto místě bude vítána libovolná poznámka k programu, na kterou nebylo jinde místo.

# B. Analyzované věty

Honza měl pak příležitost promluvit si s ním mezi čtyřma očima.
My jsme nechali našeho Járu přes sobotu doma samotného.
U nedalekého potoka už rozkvetly žluté blatouchy.
Mech u cesty byl krásně zelený.
Ráno půjdu se svým kamarádem do lesa na houby.
Petr mě zahlédl nastupovat do autobusu.
Moje sestra teď myje nádobí od oběda.
Pavel už musel jít do školy.
Monika se po chvíli začala třást zimou.
Já bych se stal rád automechanikem.
Maso a zeleninu doměkka podusíme.
Nebyla to nijak složitá, ale velice nudná práce.
Jsi milý, ale hrozně upovídaný kluk.
Hlavou zeď neprorazíš.
Před svítáním začalo drobně pršet.
Včera jsem potkal Vaška, našeho bývalého spolužáka.
Seděli jsme na starých, ale velmi pohodlných křeslech.
Z malé mýtinky před námi byl půvabný pohled po zalesněných úbočích až do údolí.
Malá dřevěná chata stála pod několika vysokými borovicemi.
Tělo kosatky, velkého delfína, je přizpůsobené rychlému pohybu na vodě.
Co dávají večer na prvním televizním programu?
To se v žádném případě nemělo stát.
Voda v hrnci už začíná být horká.
Bylo by nádherné prožít několik měsíců na ostrově v Tichomoří.
Naši se určitě budou na nás hněvat.
Jako poslední do cíle dorazil Ondřej.
Co se stalo?
Babiččin pes býval vděčný za každé pohlazení.
Kvůli rozorávání mezí a práškování polí byly koroptve u nás málem vyhubeny.
Je zbytečné plakat nad rozlitým mlékem.
Kdopak by se na něho zlobil!
Vašku, přestaň si už konečně vymýšlet nesmysly!
Co bych měl udělat?
Po všech těch špatných zkušenostech se stal vůči němu nedůvěřivým.
Mladí ležáci - staří žebráci.
Staré domy na kraji města budou muset být již v tomto roce zbourány.
Brtník je v Evropě nejmohutnější šelma.
Dříve žil na mnoha místech i u nás, nyní se s ním můžeme setkat ve slovenských horách.
Medvěd je oblíbené zvíře slovanských pohádek a v cirkusech bývá miláčkem návštěvníků.
Jeho pohyby vypadají jako nemotorné a má dobrácký pohled.
Ale povaha tohoto bručouna není tak pěkná jako jeho pohled.
Brtník je potměšilý, často se přetvařuje a dokáže se úplně bez příčiny vrhnout na

svého krotitele.
Na svobodě se medvěd člověku zdaleka vyhne, jakmile slyší jeho kroky.
Krems je velmi malebné městečko.
Jeho název si okamžitě vybavíme z názvu kremžská hořčice, ale tu tady neznají.
Měšťanské domy, prozrazující zámožnost původních majitelů, se tu zachovaly dodnes neporušené.
Ty nejstarší pocházejí ze 16. století.
Upravené ulice jsou plné obchůdků.
Z rozvalin hradů si zaslouží zvýšenou pozornost Dürstein.
Toto sídlo bylo kdysi vězením.
V době křížových výprav tu byl internován Richard Lví srdce.
Stal se romantickým hrdinou, ale historici ho nehodnotí příznivě.
Rád vyhledával dobrodružství a pohrdal vší rozvahou.
Po otcově smrti vyprázdnil pokladnu a vydal se na křížovou výpravu do Svaté země.
Jeruzalém nezískal, ale dobyl Kypr.
Rok krutě vládl křesťanské Palestině, pak se chtěl vrátit domů.
Jeho loď však ztroskotala a on byl zajat.
Dva roky byl vězněn v Rakousku.
Když zaplatil velké výkupné, byl z vězení propuštěn.
Vrátil se do Anglie, jeho stoupenci ho vítali, ale se svým bratrem musel bojovat o trůn.
Pak válčil s Francií a v jedné bitvě zahynul.
Během své vlády strávil v Anglii jen asi čtyři měsíce.
Před mnoha lety směřovala jedna plachetnice úžinou mezi dvěma ostrovy do Mexického zálivu.
Kapitán lodi chtěl přistát v nějakém přístavu na pobřeží Mexika.
Osud však její plavbě nepřál, loď se dostala do větrné smršti a v prudké bouři ztroskotala na širém moři.
Zachránila se jen část posádky, kterou vylovila jiná loď.
Z hrdého trojstěžníku zbyl pouze vrak, vyčnívající z části nad hladinu.
Mrtvá loď bez jediného muže na palubě se stala nebezpečím pro ostatní plavidla.
V noci se tajemně vynořovala před jejich příděmi a kormidelníci, pozorně sledující směr plavby, museli dělat pravé divy, aby zabránili srážce s vrakem.
Kožené boty jsou dost drahé.
To je psí život!
Šumavská příroda je nádherná.
V samoobsluze prodávají i francouzské sýry.
Zdáli se mi, že mě tajemná noční řeka snad pohltí.
Nemám nejmenší představu, jak to udělám.
U savců, kteří žijí trvale buď pod zemí nebo ve vodě, srst degradovala.
Až za tmy vylézáme z kánoe na břeh, celí prokřehlí a unavení.
Když se mladší sestra vrátila z letního tábora, celé dny nadšeně vyprávěla, co všechno zažila.
To červené auto, zahýbající právě za roh, muselo prudce přibrzdit, protože do vozovky vběhly dvě malé děti.
Ani jsem nepostřehl, odkud ses vynořil.
Abychom natankovali benzin, zastavili jsme u benzinové pumpy.

Po letech byl stále takový, jakého jsme ho znali.
Zdálo se, že ho nikdo nepřemůže.
Kdo jinému jámu kopá, sám do ní padá.
Domníval jsem se, že ten film hrají dneska.
Byl jsem velmi pyšný na to, že jsi to dokázal.
Je vždycky tam, kde se něco děje.
Vrať se, odkud jsi přišel.
Šel, kam ho oči vedly.
Jakmile přijedu v sobotu k babičce, půjdeme spolu do lesa.
Až napíšeš úkoly, přijď za mnou.
Tvářil se, jako by mu ulítly včely.
Kolikrát se vsadil, tolikrát prohrál.
Protože půjdu zítra na prohlídku k lékaři, budu ve škole první dvě hodiny chybět.
Musím se s ním sejít, abych mu všechno vysvětlila.
Zavři dveře, ať netáhne.
Kdybys té matematice nerozuměl, vysvětlím ti to.
Ačkoliv je teprve šest hodin, musíme už svítit.
Jestli přijdu pozdě, už mám připravenou omluvu.
Dům, v němž bydlíme, stojí hned u hlavní silnice.
Není už takový, jakého jsem si ho pamatoval.

# C. Analytická pravidla

V tabulkách C.1 až C.3 jsou uvedena pravidla pro přiřazování analytických funkcí. Jedná se o transformační pravidla vygenerovaná automaticky podle metody popsané v odstavci 4.2. Sloupec „Tokenů v uzlu“ udává celkový počet tokenů v daném uzlu nezahrnující interpunkci. Z pohledu pořadí se však interpunkce vyhodnocuje jako samostatný token.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Funkce Čapek | Funkce PDT | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | Pk | Atr | 231 |
| 2 | 1 | 1 | Po | Sb | 123 |
| 3 | 1 | 1 | Přís | Pred | 94 |
| 4 | 1 | 2 | Pu-m | AuxP | 94 |
| 5 | 1 | 1 | Pt | Obj | 93 |
| 6 | 2 | 2 | Pu-m | Adv | 74 |
| 7 | 1 | 1 | Pu-č | Adv | 56 |
| 8 | 1 | 2 | Přj-spn | Pred | 44 |
| 9 | 1 | 1 | Pu-zp | Adv | 38 |
| 10 | 1 | 2 | Pk | AuxP | 35 |
| 11 | 2 | 2 | Přj-spn | Pnom | 34 |
| 12 | 1 | 1 | ? | Coord | 33 |
| 13 | 1 | 2 | Pt | AuxP | 32 |
| 14 | 3 | 2 | Pu-m | AuxK | 31 |
| 15 | 2 | 2 | Přís | Pred | 30 |
| 16 | 1 | 2 | Pu-č | AuxP | 30 |
| 17 | 2 | 2 | Pk | Atr | 28 |
| 18 | 1 | 2 | Přís | Pred | 28 |
| 19 | 2 | 2 | Pu-č | Adv | 28 |
| 20 | 2 | 1 | Přís | AuxK | 26 |
| 21 | 3 | 2 | Přj-spn | AuxK | 25 |
| 22 | 2 | 1 | Pt | AuxK | 21 |
| 23 | 1 | 1 | Pu-m | Adv | 21 |
| 24 | 3 | 2 | Přís | AuxK | 18 |
| 25 | 2 | 1 | Pk | AuxX | 15 |
| 26 | 2 | 1 | Po | AuxK | 14 |
| 27 | 3 | 3 | Přís | Obj | 14 |
| 28 | 2 | 2 | Pt | Adv | 14 |
| 29 | 2 | 2 | Pt | Obj | 14 |
| 30 | 1 | 1 | Pu-míra | Adv | 13 |

Tabulka C.1: Seznam automaticky generovaných pravidel pro převod analytické funkce – 1. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Funkce Čapek | Funkce PDT | Četnost v datech |
|---|---|---|---|---|---|
| 31 | 3 | 2 | Pk | AuxK | 11 |
| 32 | 4 | 3 | Přís | AuxK | 11 |
| 33 | 1 | 3 | Přís | AuxT | 10 |
| 34 | 2 | 3 | Přís | Pred | 8 |
| 35 | 1 | 3 | Přj-spn | Pred | 8 |
| 36 | 1 | 1 | Do | Atr | 7 |
| 37 | 2 | 1 | Do | AuxK | 7 |
| 38 | 3 | 2 | Pt | AuxK | 7 |
| 39 | 2 | 3 | Přj-spn | AuxV | 5 |
| 40 | 2 | 1 | ? | AuxX | 4 |
| 41 | 3 | 3 | Přj-spn | Pnom | 4 |
| 42 | 4 | 3 | Přj-spn | AuxX | 4 |
| 43 | 1 | 1 | Pu-příč | AuxC | 4 |
| 44 | 1 | 1 | Pu-účel | AuxC | 4 |
| 45 | 3 | 2 | Pu-zp | AuxK | 4 |
| 46 | 1 | 2 | Pu-příč | AuxP | 3 |
| 47 | 2 | 1 | Pu-zp | AuxK | 3 |
| 48 | 1 | 2 | Pu-zp | AuxC | 3 |
| 49 | 1 | 2 | Pu-zp | AuxP | 3 |
| 50 | 2 | 2 | Pu-zp | Adv | 3 |
| 51 | 1 | 1 | Přj | Atr | 2 |
| 52 | 1 | 4 | Přj-spn | AuxV | 2 |
| 53 | 5 | 4 | Přj-spn | AuxK | 2 |
| 54 | 2 | 1 | Pu-č | AuxK | 2 |
| 55 | 3 | 2 | Pu-č | AuxK | 2 |
| 56 | 2 | 1 | Pu-m | AuxX | 2 |
| 57 | 2 | 1 | Pu-m | AuxK | 2 |
| 58 | 1 | 3 | Pu-m | AuxZ | 2 |
| 59 | 2 | 3 | Pu-m | AuxP | 2 |
| 60 | 3 | 3 | Pu-m | Adv | 2 |
| 61 | 1 | 2 | Pu-míra | AuxP | 2 |
| 62 | 2 | 2 | Pu-míra | Atr | 2 |
| 63 | 1 | 1 | Pu-podm | AuxC | 2 |
| 64 | 2 | 2 | Pu-příč | Adv | 2 |
| 65 | 1 | 1 | Pu-příp | AuxC | 2 |
| 66 | 1 | 2 | Do | AuxY | 1 |
| 67 | 1 | 2 | Do | AuxC | 1 |
| 68 | 2 | 2 | Do | AtvV | 1 |
| 69 | 2 | 2 | Do | Atr | 1 |
| 70 | 1 | 3 | Pk | AuxZ | 1 |

Tabulka C.2: Seznam automaticky generovaných pravidel pro převod analytické funkce – 2. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Funkce Čapek | Funkce PDT | Četnost v datech |
|---|---|---|---|---|---|
| 71 | 2 | 3 | Pk | AuxP | 1 |
| 72 | 3 | 3 | Pk | Adv | 1 |
| 73 | 4 | 3 | Pk | AuxX | 1 |
| 74 | 1 | 3 | Po | Sb | 1 |
| 75 | 2 | 3 | Po | Atr | 1 |
| 76 | 3 | 3 | Po | Obj | 1 |
| 77 | 4 | 3 | Po | AuxK | 1 |
| 78 | 1 | 4 | Přís | AuxV | 1 |
| 79 | 2 | 4 | Přís | Pred | 1 |
| 80 | 3 | 4 | Přís | AuxV | 1 |
| 81 | 4 | 4 | Přís | Obj | 1 |
| 82 | 5 | 4 | Přís | AuxK | 1 |
| 83 | 1 | 2 | Přj | Coord | 1 |
| 84 | 2 | 2 | Přj | Sb | 1 |
| 85 | 3 | 2 | Přj | AuxK | 1 |
| 86 | 1 | 1 | Přj-spn | Sb | 1 |
| 87 | 1 | 1 | Přj-spn | Adv | 1 |
| 88 | 2 | 4 | Přj-spn | AuxT | 1 |
| 89 | 2 | 4 | Přj-spn | Pred | 1 |
| 90 | 3 | 4 | Přj-spn | Pred | 1 |
| 91 | 3 | 4 | Přj-spn | AuxV | 1 |
| 92 | 4 | 4 | Přj-spn | Adv | 1 |
| 93 | 4 | 4 | Přj-spn | Obj | 1 |
| 94 | 4 | 3 | Pu-m | AuxX | 1 |
| 95 | 4 | 3 | Pu-m | AuxK | 1 |
| 96 | 1 | 2 | Pu-pros | AuxP | 1 |
| 97 | 2 | 2 | Pu-pros | Adv | 1 |
| 98 | 2 | 1 | Pu-příč | AuxK | 1 |
| 99 | 3 | 2 | Pu-příč | AuxK | 1 |
| 100 | 1 | 2 | Pu-účel | AuxP | 1 |
| 101 | 2 | 2 | Pu-účel | Atr | 1 |
| 102 | 3 | 2 | Pu-účel | AuxK | 1 |

Tabulka C.3: Seznam automaticky generovaných pravidel pro převod analytické funkce – 3. část.

# D. Morfologická pravidla

V následujících tabulkách jsou uvedena pravidla pro přiřazování morfologických kategorií, konkrétně slovního druhu (tabulky D.1 a D.2), rodu (tabulka D.3), čísla (tabulka D.4), pádu (tabulky D.5 a D.6), osoby (tabulka D.7), času (tabulka D.8) a stupně (tabulka D.9). Jedná se o transformační pravidla vygenerovaná automaticky podle metody popsané v odstavci 4.3.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag1 | PDT tag1 | PDT tag2 | Četnost v datech |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | N | N | N | 225 |
| 2 | 2 | 2 | N | N | N | 178 |
| 3 | 1 | 2 | N | R | R | 162 |
| 4 | 1 | 1 | A | A | A | 137 |
| 5 | 1 | 1 | D | D | b | 86 |
| 6 | 2 | 1 | N | Z | : | 80 |
| 7 | 3 | 2 | N | Z | : | 73 |
| 8 | 1 | 1 | V | V | B | 66 |
| 9 | 3 | 2 | V | Z | : | 57 |
| 10 | 1 | 2 | V | V | p | 46 |
| 11 | 2 | 1 | V | Z | : | 39 |
| 12 | 2 | 2 | V | V | p | 32 |
| 13 | 1 | 1 | J | J | , | 29 |
| 14 | 1 | 1 | P | P | D | 27 |
| 15 | 1 | 2 | P | R | R | 24 |
| 16 | 4 | 3 | V | Z | : | 19 |
| 17 | 1 | 1 | C | C | l | 16 |
| 18 | 3 | 3 | V | V | f | 15 |
| 19 | 1 | 3 | V | P | 7 | 13 |
| 20 | 2 | 1 | A | Z | : | 12 |
| 21 | 2 | 2 | P | P | 5 | 12 |
| 22 | 2 | 3 | V | P | 7 | 11 |
| 23 | 2 | 1 | D | Z | : | 10 |
| 24 | 1 | 1 | T | J | ^ | 7 |
| 25 | 1 | 1 | f | V | f | 6 |
| 26 | 1 | 2 | A | J | , | 4 |
| 27 | 2 | 2 | A | A | A | 4 |
| 28 | 2 | 1 | P | Z | : | 4 |
| 29 | 3 | 2 | P | Z | : | 4 |
| 30 | 5 | 4 | V | Z | : | 3 |

Tabulka D.1: Seznam automaticky generovaných pravidel pro převod slovního druhu – 1. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag1 | PDT tag1 | PDT tag2 | Četnost v datech |
|---|---|---|---|---|---|---|
| 31 | 3 | 3 | N | N | N | 2 |
| 32 | 4 | 3 | N | Z | : | 2 |
| 33 | 1 | 3 | P | J | ^ | 2 |
| 34 | 2 | 3 | P | R | R | 2 |
| 35 | 3 | 3 | P | P | P | 2 |
| 36 | 4 | 3 | P | Z | : | 2 |
| 37 | 1 | 4 | V | V | B | 2 |
| 38 | 2 | 4 | V | V | f | 2 |
| 39 | 3 | 4 | V | V | f | 2 |
| 40 | 4 | 4 | V | V | s | 2 |
| 41 | 2 | 1 | C | Z | : | 1 |
| 42 | 1 | 2 | D | R | R | 1 |
| 43 | 2 | 2 | D | N | N | 1 |
| 44 | 1 | 2 | f | P | 7 | 1 |
| 45 | 2 | 2 | f | V | f | 1 |
| 46 | 1 | 3 | N | N | N | 1 |
| 47 | 1 | 3 | N | T | T | 1 |
| 48 | 2 | 3 | N | A | A | 1 |
| 49 | 2 | 3 | N | R | R | 1 |

Tabulka D.2: Seznam automaticky generovaných pravidel pro převod slovního druhu – 2. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag5 | PDT tag3 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | - | - | 307 |
| 2 | 1 | 1 | F | F | 166 |
| 3 | 1 | 1 | M | I | 93 |
| 4 | 1 | 2 | F | - | 85 |
| 5 | 1 | 2 | - | - | 83 |
| 6 | 2 | 2 | F | F | 82 |
| 7 | 1 | 2 | M | - | 80 |
| 8 | 3 | 2 | - | - | 64 |
| 9 | 1 | 1 | N | N | 57 |
| 10 | 2 | 1 | - | - | 55 |
| 11 | 2 | 1 | M | - | 51 |
| 12 | 2 | 2 | M | I | 47 |
| 13 | 2 | 2 | - | - | 46 |
| 14 | 2 | 1 | F | - | 33 |
| 15 | 1 | 2 | N | - | 32 |
| 16 | 2 | 2 | N | N | 31 |
| 17 | 3 | 2 | M | - | 30 |
| 18 | 1 | 3 | - | - | 26 |
| 19 | 3 | 2 | F | - | 26 |
| 20 | 2 | 3 | - | - | 23 |
| 21 | 4 | 3 | - | - | 20 |
| 22 | 3 | 2 | N | - | 19 |
| 23 | 3 | 3 | - | - | 18 |
| 24 | 2 | 1 | N | - | 11 |
| 25 | 1 | 4 | - | - | 3 |
| 26 | 2 | 4 | - | - | 3 |
| 27 | 5 | 4 | - | - | 3 |
| 28 | 3 | 4 | - | - | 2 |
| 29 | 4 | 4 | - | T | 2 |
| 30 | 1 | 3 | M | M | 1 |
| 31 | 2 | 3 | M | N | 1 |
| 32 | 3 | 3 | M | N | 1 |
| 33 | 4 | 3 | M | - | 1 |
| 34 | 1 | 3 | N | - | 1 |
| 35 | 2 | 3 | N | - | 1 |
| 36 | 3 | 3 | N | N | 1 |
| 37 | 4 | 3 | N | - | 1 |

Tabulka D.3: Seznam automaticky generovaných pravidel pro převod rodu.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag6 | PDT tag4 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | S | S | 422 |
| 2 | 1 | 1 | - | - | 233 |
| 3 | 2 | 2 | S | S | 221 |
| 4 | 1 | 2 | S | - | 166 |
| 5 | 1 | 1 | P | P | 163 |
| 6 | 3 | 2 | S | - | 110 |
| 7 | 2 | 1 | S | - | 90 |
| 8 | 2 | 2 | P | P | 58 |
| 9 | 2 | 1 | P | - | 45 |
| 10 | 1 | 2 | P | - | 42 |
| 11 | 3 | 2 | P | - | 22 |
| 12 | 4 | 3 | S | - | 19 |
| 13 | 1 | 3 | S | S | 18 |
| 14 | 3 | 3 | S | S | 17 |
| 15 | 2 | 1 | - | - | 16 |
| 16 | 2 | 3 | S | S | 11 |
| 17 | 2 | 3 | S | X | 11 |
| 18 | 1 | 2 | - | S | 8 |
| 19 | 2 | 2 | - | S | 7 |
| 20 | 3 | 2 | - | - | 7 |
| 21 | 1 | 3 | P | X | 4 |
| 22 | 2 | 3 | P | P | 4 |
| 23 | 3 | 3 | P | - | 4 |
| 24 | 4 | 3 | P | - | 4 |
| 25 | 1 | 3 | - | S | 2 |
| 26 | 2 | 3 | - | S | 2 |
| 27 | 1 | 4 | P | P | 2 |
| 28 | 2 | 4 | P | - | 2 |
| 29 | 3 | 4 | P | - | 2 |
| 30 | 4 | 4 | P | P | 2 |
| 31 | 5 | 4 | P | - | 2 |
| 32 | 3 | 3 | - | - | 1 |
| 33 | 3 | 3 | - | S | 1 |
| 34 | 4 | 3 | - | - | 1 |
| 35 | 1 | 4 | S | S | 1 |
| 36 | 2 | 4 | S | X | 1 |
| 37 | 3 | 4 | S | S | 1 |
| 38 | 4 | 4 | S | S | 1 |
| 39 | 5 | 4 | S | - | 1 |

Tabulka D.4: Seznam automaticky generovaných pravidel pro převod čísla.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag7 | PDT tag5 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | - | - | 368 |
| 2 | 1 | 1 | 1 | 1 | 209 |
| 3 | 1 | 2 | - | - | 104 |
| 4 | 1 | 1 | 4 | 4 | 104 |
| 5 | 2 | 2 | - | - | 76 |
| 6 | 3 | 2 | - | - | 64 |
| 7 | 1 | 2 | 6 | 6 | 63 |
| 8 | 2 | 2 | 6 | 6 | 63 |
| 9 | 1 | 1 | 2 | 2 | 61 |
| 10 | 1 | 2 | 2 | 2 | 57 |
| 11 | 2 | 2 | 2 | 2 | 57 |
| 12 | 2 | 1 | - | - | 55 |
| 13 | 2 | 1 | 1 | - | 34 |
| 14 | 1 | 1 | 6 | 6 | 34 |
| 15 | 2 | 1 | 4 | - | 33 |
| 16 | 1 | 2 | 7 | 7 | 32 |
| 17 | 2 | 2 | 7 | 7 | 32 |
| 18 | 1 | 1 | 7 | 7 | 29 |
| 19 | 1 | 1 | 3 | 3 | 28 |
| 20 | 1 | 2 | 4 | 4 | 28 |
| 21 | 2 | 2 | 4 | 4 | 28 |
| 22 | 3 | 2 | 2 | - | 24 |
| 23 | 2 | 3 | - | - | 23 |
| 24 | 1 | 3 | - | - | 21 |
| 25 | 4 | 3 | - | - | 20 |
| 26 | 3 | 3 | - | - | 19 |
| 27 | 2 | 1 | 2 | - | 18 |
| 28 | 3 | 2 | 4 | - | 18 |
| 29 | 3 | 2 | 6 | - | 15 |
| 30 | 3 | 2 | 7 | - | 10 |
| 31 | 1 | 2 | 1 | - | 8 |
| 32 | 2 | 2 | 1 | 1 | 8 |
| 33 | 2 | 2 | 3 | 3 | 8 |
| 34 | 1 | 2 | 3 | 3 | 7 |
| 35 | 3 | 2 | 1 | - | 6 |
| 36 | 2 | 1 | 7 | - | 5 |
| 37 | 2 | 1 | 3 | - | 4 |
| 38 | 3 | 2 | 3 | - | 4 |
| 39 | 1 | 4 | - | - | 3 |
| 40 | 3 | 4 | - | - | 3 |

Tabulka D.5: Seznam automaticky generovaných pravidel pro převod pádu – 1. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag7 | PDT tag5 | Četnost v datech |
|---|---|---|---|---|---|
| 41 | 5 | 4 | - | - | 3 |
| 42 | 1 | 3 | 2 | - | 3 |
| 43 | 2 | 3 | 2 | 2 | 3 |
| 44 | 3 | 3 | 2 | 2 | 3 |
| 45 | 4 | 3 | 2 | - | 3 |
| 46 | 2 | 4 | - | - | 2 |
| 47 | 4 | 4 | - | - | 2 |
| 48 | 2 | 1 | 6 | - | 2 |
| 49 | 1 | 3 | 1 | 1 | 1 |
| 50 | 2 | 3 | 1 | 4 | 1 |
| 51 | 3 | 3 | 1 | 4 | 1 |
| 52 | 4 | 3 | 1 | - | 1 |
| 53 | 1 | 1 | 5 | 3 | 1 |
| 54 | 2 | 1 | 5 | - | 1 |

Tabulka D.6: Seznam automaticky generovaných pravidel pro převod pádu – 2. část.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag4 | PDT tag8 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | - | - | 736 |
| 2 | 1 | 2 | - | - | 212 |
| 3 | 2 | 2 | - | - | 198 |
| 4 | 2 | 1 | - | - | 113 |
| 5 | 3 | 2 | - | - | 84 |
| 6 | 1 | 1 | 3 | X | 62 |
| 7 | 2 | 2 | 3 | - | 53 |
| 8 | 1 | 2 | 3 | X | 40 |
| 9 | 3 | 2 | 3 | - | 37 |
| 10 | 2 | 1 | 3 | - | 31 |
| 11 | 1 | 1 | 1 | 1 | 22 |
| 12 | 3 | 3 | 3 | - | 18 |
| 13 | 1 | 2 | 1 | 1 | 12 |
| 14 | 3 | 2 | 2 | - | 12 |
| 15 | 1 | 3 | 3 | - | 12 |
| 16 | 4 | 3 | 3 | - | 12 |
| 17 | 1 | 2 | 2 | 2 | 11 |
| 18 | 2 | 2 | 1 | X | 9 |
| 19 | 1 | 1 | 2 | 2 | 9 |
| 20 | 2 | 2 | 2 | X | 9 |
| 21 | 3 | 2 | 1 | - | 8 |
| 22 | 3 | 3 | 1 | - | 8 |
| 23 | 2 | 3 | 3 | X | 8 |
| 24 | 2 | 3 | 3 | - | 8 |
| 25 | 4 | 3 | 1 | - | 7 |
| 26 | 1 | 3 | 1 | 1 | 6 |
| 27 | 4 | 3 | - | - | 5 |
| 28 | 2 | 1 | 1 | - | 5 |
| 29 | 2 | 3 | 1 | - | 5 |
| 30 | 1 | 3 | - | - | 4 |
| 31 | 2 | 3 | - | - | 4 |
| 32 | 3 | 3 | - | - | 4 |
| 33 | 2 | 1 | 2 | - | 3 |
| 34 | 1 | 3 | 2 | 3 | 2 |
| 35 | 2 | 3 | 2 | - | 2 |
| 36 | 3 | 3 | 2 | - | 2 |
| 37 | 1 | 4 | 3 | 3 | 2 |
| 38 | 2 | 4 | 3 | - | 2 |
| 39 | 3 | 4 | 3 | - | 2 |
| 40 | 4 | 4 | 3 | X | 2 |
| 41 | 5 | 4 | 3 | - | 2 |
| 42 | 1 | 4 | 1 | 1 | 1 |
| 43 | 2 | 4 | 1 | - | 1 |
| 44 | 3 | 4 | 1 | X | 1 |
| 45 | 4 | 4 | 1 | - | 1 |
| 46 | 5 | 4 | 1 | - | 1 |

Tabulka D.7: Seznam automaticky generovaných pravidel pro převod osoby.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag9 | PDT tag9 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | - | - | 790 |
| 2 | 2 | 2 | - | - | 218 |
| 3 | 1 | 2 | - | - | 213 |
| 4 | 2 | 1 | - | - | 113 |
| 5 | 3 | 2 | - | - | 84 |
| 6 | 1 | 1 | R | R | 62 |
| 7 | 1 | 1 | P | P | 42 |
| 8 | 1 | 2 | R | R | 42 |
| 9 | 2 | 2 | P | - | 31 |
| 10 | 2 | 2 | R | - | 30 |
| 11 | 3 | 2 | R | - | 30 |
| 12 | 1 | 2 | P | P | 29 |
| 13 | 2 | 1 | R | - | 21 |
| 14 | 3 | 2 | P | - | 19 |
| 15 | 1 | 1 | F | P | 15 |
| 16 | 3 | 3 | R | - | 13 |
| 17 | 2 | 1 | P | - | 10 |
| 18 | 4 | 3 | R | - | 9 |
| 19 | 2 | 1 | F | - | 8 |
| 20 | 1 | 3 | R | - | 8 |
| 21 | 2 | 3 | R | R | 8 |
| 22 | 3 | 3 | P | - | 7 |
| 23 | 3 | 3 | - | - | 6 |
| 24 | 1 | 3 | - | - | 5 |
| 25 | 4 | 3 | - | - | 5 |
| 26 | 1 | 3 | P | P | 5 |
| 27 | 2 | 3 | P | - | 5 |
| 28 | 4 | 3 | P | - | 5 |
| 29 | 2 | 3 | - | - | 4 |
| 30 | 1 | 2 | F | - | 3 |
| 31 | 2 | 2 | F | P | 3 |
| 32 | 3 | 2 | F | - | 3 |
| 33 | 1 | 3 | F | - | 2 |
| 34 | 2 | 3 | F | F | 2 |
| 35 | 3 | 3 | F | - | 2 |
| 36 | 4 | 3 | F | - | 2 |
| 37 | 1 | 4 | F | F | 2 |
| 38 | 2 | 4 | F | - | 2 |
| 39 | 3 | 4 | F | - | 2 |
| 40 | 4 | 4 | F | X | 2 |
| 41 | 5 | 4 | F | - | 2 |

Tabulka D.8: Seznam automaticky generovaných pravidel pro převod času.

| Číslo pravidla | Pořadí tokenu | Tokenů v uzlu | Čapek tag8 | PDT tag10 | Četnost v datech |
|---|---|---|---|---|---|
| 1 | 1 | 1 | - | - | 632 |
| 2 | 1 | 2 | - | - | 336 |
| 3 | 2 | 2 | - | - | 317 |
| 4 | 3 | 2 | - | - | 141 |
| 5 | 2 | 1 | - | - | 130 |
| 6 | 1 | 1 | ? | - | 110 |
| 7 | 1 | 1 | 1 | 1 | 82 |
| 8 | 1 | 3 | - | - | 38 |
| 9 | 2 | 3 | - | - | 37 |
| 10 | 3 | 3 | - | - | 30 |
| 11 | 4 | 3 | - | - | 24 |
| 12 | 2 | 1 | 1 | - | 12 |
| 13 | 2 | 1 | ? | - | 10 |
| 14 | 1 | 1 | 3 | 3 | 6 |
| 15 | 1 | 4 | - | - | 3 |
| 16 | 2 | 4 | - | - | 3 |
| 17 | 3 | 4 | - | - | 3 |
| 18 | 4 | 4 | - | - | 3 |
| 19 | 5 | 4 | - | - | 3 |
| 20 | 1 | 2 | 1 | - | 3 |
| 21 | 2 | 2 | 1 | 1 | 3 |
| 22 | 1 | 2 | ? | - | 2 |
| 23 | 1 | 1 | 2 | 2 | 2 |
| 24 | 2 | 2 | ? | - | 1 |
| 25 | 2 | 2 | ? | - | 1 |

Tabulka D.9: Seznam automaticky generovaných pravidel pro převod stupně.

# E. Uživatelská dokumentace k transformačnímu balíku

## E.1 Požadavky na systém

Aby mohl být převodní skript správně používán, je třeba mít k dispozici:

- *mysql server* verze 5 a vyšší – je k dispozici k volnému stažení na stránce http://dev.mysql.com/downloads/mysql/5.0.html (navštíveno 4.12.2011),
- *perl* verze 5 a vyšší (v linuxu zjistíme verzi perlu zadáním příkazu `perl -v` v terminálu.)
- kromě standardních modulů perlu verze 5 jsou zapotřebí moduly:
  - *XPath* (modul pro práci s XML soubory) [10]
  - *DBI* (modul pro práci s databází) [1]

## E.2 První použití (instalace)

Instalaci provedeme spuštěním skriptu `create_structure.pl` . Tento skript vytvoří strukturu databáze a načte pravidla pro transformace. Po spuštění instalačního skriptu je uživatel vyzván k zadání uživatelského jména k databázi a následně o zadání hesla k tomuto jménu. V posledním kroku je třeba zadat jméno databáze, do které chceme strukturu pro převod vytvořit. Tímto je vytvořena připravená databáze zadaného jména.

## E.3 Transformace souborů

Při transformaci souborů z editoru Čapek do formátu PML postupujeme v následujících krocích:

- soubory k transformaci připravíme do samostatného adresáře,
- spustíme skript `transform.pl`,
- zadáme uživatelské jméno k *mysql* databázi, které jsme zadali při instalaci (viz odst. E.2),
- zadáme příslušné heslo (taktéž zadáno při instalaci),
- zadáme název adresáře, ve kterém se nachází soubory určené k transformaci,
- zadáme název adresáře, do kterého se mají transformované soubory vypisovat.

Pro každý z transformovaných souborů budou vytvořeny tři nové soubory se shodným jménem a s příponami `.w`, odpovídající slovní rovině, `.a`, odpovídající analytické rovině a `.m`, odpovídající morfologické rovině v PDT.

# F. Programátorská dokumentace k transformačnímu balíku

## F.1 Model databáze

Na obrázku F.1 je vidět návrh schématu relační databáze. Pro větší přehlednost byly vybrány jen některé tabulky (schéma pro všechny tabulky pravidel i jejich zapojení jsou si velice podobná.

Struktura tabulek je volena tak, aby bylo možné snadno načítat jak struktury dle PDT, tak struktury vytvořené editorem Čapek. V průběhu práce byl výstupní xml formát editoru změněn. Morfologické tagy začaly být součástí informace tokenu, oproti původní verzi, kde byly tagy součástí uzlu. Atribut *token_id* je proto použit duplicitně jak v tabulce *node*, tak v tabulce *token*. Pravidla jsou obsažená v tabulkách označených *rules_(název pravidla)*. Tyto tabulky jsou při spuštění instalačního skriptu načteny do nově vytvořené databáze z připravených csv souborů. Načítaná data jsou ukládána do tabulek:

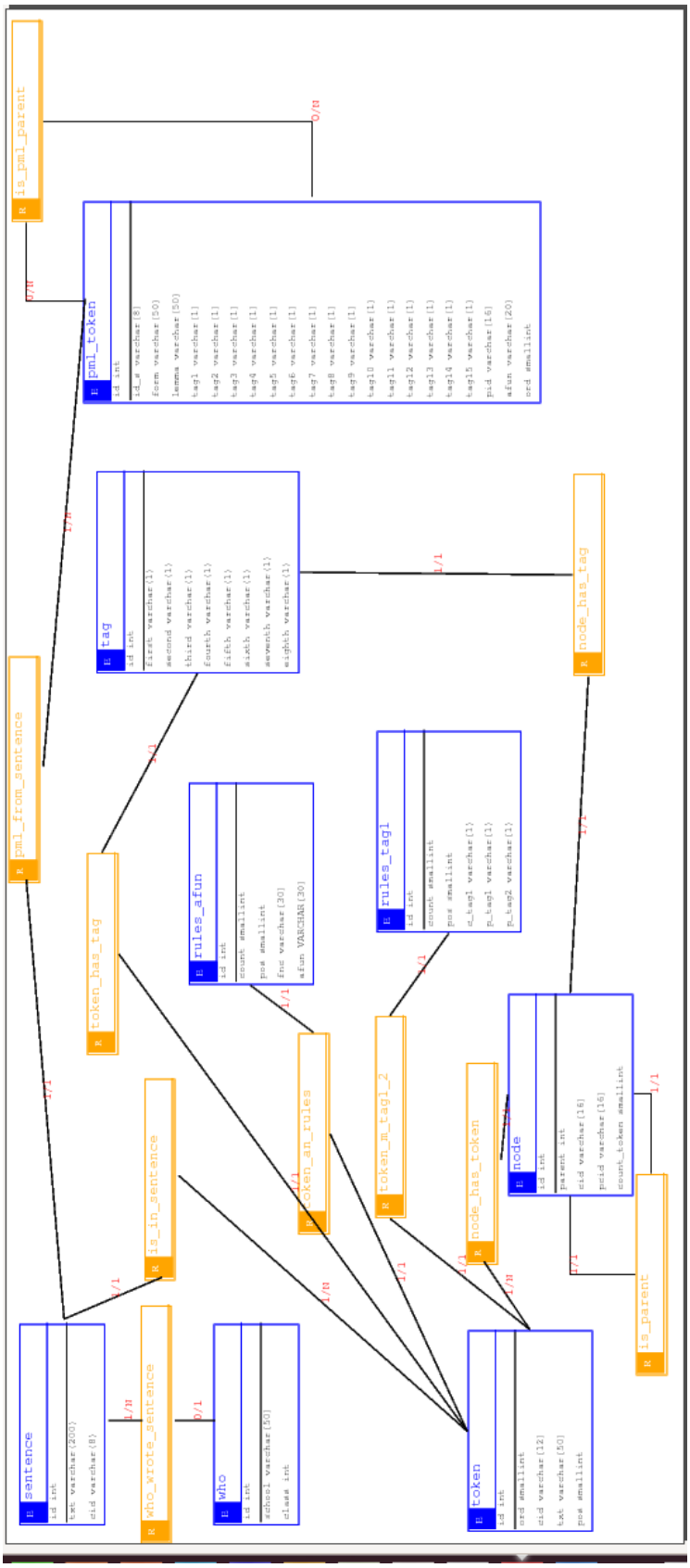
- *token*: zde jsou informace o slovech, jako například pořadí slova ve větě, z jak velkého uzlu pochází, jaká je jeho pozice v rámci uzlu. Zároveň zde máme i atribut *token_id*, který je i v tabulce *node*. Oba odkazují na strukturu morfologického tagu, ale v každém z načítaných souborů je plněn jen jeden z nich podle toho, ve které verzi Čapka byl načítaný soubor vytvořen.

- *node*: v této tabulce jsou uchovávány informace o stromové struktuře vytvořené v editoru Čapek a informace o analytických funkcích,

- *who*: do této tabulky jsou data vkládána uživatelem transformačních skriptů. Není povinné při načítání xml souborů z Čapka tyto údaje zadávat. Tato tabulka byla vytvořena pro vyhodnocování nasbíraných souborů, aby bylo možné načtená data více třídit.

- *tag*: v této tabulce jsou uloženy všechny načtené morfologické údaje zadané v editoru.

Během načítání do těchto struktur je třeba ošetřit délku tagů (v nové verzi editoru byly přidané pozice do morfologického tagu), počítat slova v uzlech a oddělovat interpunkci. Ta je v editoru Čapek přilepena ke slovu, které jí předchází, zatímco ve formátu PDT je samostatným tokenem.




Obrázek F.1: ER diagram databáze sloužící k transformacím.