Pražský závislostní korpus jako elektronická cvičebnice češtiny

Ondřej Kučera *
ondrej.kucera@centrum.cz

Abstrakt Pražský závislostní korpus patří mezi nejvýznamnější jazykové korpusy na světě. Cílem naší práce je představit softwarový systém, který z vět Pražského závislostního korpusu vytvoří elektronickou cvičebnici českého jazyka. Procvičování probíhá ve dvou oblastech: tvarosloví (určování slovních druhů a jejich morfologických kategorií) a větný rozbor (určování větných členů a závislostí mezi nimi). Vzhledem k odlišnostem mezi akademickými rozbory vět a rozbory tak, jak jsou vyučovány ve školách, však nelze věty z Pražského závislostního korpusu použít zcela přímočaře. Mnoho z nich je potřeba z dat úplně vyřadit, na ostatních je nutné provést množství transformací, které převedou původní reprezentaci do tvaru, na nějž jsou žáci zvyklí ze školy. Představovaná elektronická cvičebnice nabízí okamžitou kontrolu správnosti prováděných rozborů.

Klíčová slova: elektronická cvičebnice češtiny, korpusová lingvistika, Pražský závislostní korpus

1 Úvod

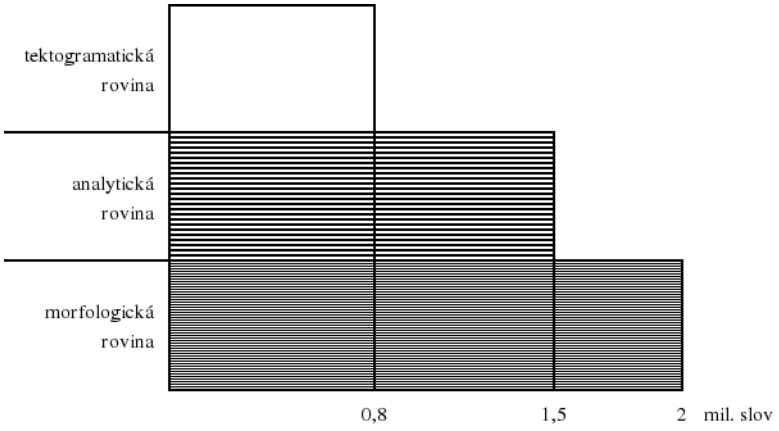
Jako každá jiná vědní disciplína i lingvistika se dělí na řadu podoborů. Pro nás zde nejdůležitějším z nich bude lingvistika korpusová. Samotná myšlenka jazykového korpusu, tedy nashromážděného množství textů toho kterého jazyka, není nikterak nová. Teprve příchod počítačů však posunul korpusovou lingvistiku na dnešní úroveň. Díky nim je možné nashromáždit, ale především rychle zpracovávat řádově mnohem větší množství dat, než by kdy bylo v lidských silách. Vzniklo tak další nové odvětví, počítačová (či komputační) lingvistika.

Na běžný elektronický korpus můžeme pohlížet jako na řetězec po sobě jdoucích slov (tvořících jednotlivé věty). Můžeme v něm velice snadno například hledat výskyty určitých slovních tvarů, porovnávat jejich počet nebo třeba zjišťovat průměrný počet slov ve větách. To vše můžeme provádět s korpusem prakticky libovolného jazyka, dokonce aniž bychom měli jeho zásadnější znalosti. Pro náročnější úlohy, jakými může být třeba strojový překlad mezi dvěma jazyky nebo systém zodpovídání dotazů, však potřebujeme hlubší analýzu korpusových dat. Tak například chceme-li v korpusu nalézt všechny výskyty všech tvarů slova *kočka*, musíme znát morfologické údaje o každém slovu v korpusu. Lze sice namítnout, že můžeme vyjmenovat jednotlivé tvary sami a pak hledat libovolný z nich, avšak tato strategie již selže v případě, kdy budeme hledat tvary podstatného jména *hnát*, protože bez morfologických informací obsažených přímo v datech nedokážeme posoudit, zda nalezený výskyt neodpovídá *slovesu* hnát. Můžeme však jít ještě dál a v korpusu uchovávat také třeba údaje o syntaktickém rozboru věty. Korpus obohacený o takováto metadata, tzv. anotace, se nazývá korpusem *anotovaným*.

Pro český jazyk takový anotovaný korpus existuje – je jím *Pražský závislostní korpus* (Prague Dependency Treebank, dále často jen PDT, viz [8])[1]. Proč pražský a závislostní?

* Univerzita Karlova, Matematicko-fyzikální fakulta, Malostranské nám. 25, 118 00 Praha 1
[1] Druhá verze PDT vyšla v červenci 2006.

Pražský pochopitelně proto, že vznikl v Praze na Ústavu formální a aplikované lingvistiky MFF UK a navazuje na tradici pražské lingvistické školy. Závislostní pak znamená, že z hlediska syntaktického je zvolen závislostní přístup, kdy za hlavní člen věty je považován predikát (nejčastěji sloveso), který je rozvíjen dalšími, závislými členy (které mohou být rovněž rozvíjeny). Pražský závislostní korpus je anotován na třech úrovních: morfologické (určení lemmat, slovních druhů a gramatických kategorií, jako jsou rod, číslo, pád, ...), syntaktické (syntaktické informace – analytická funkce, závislosti jednotlivých uzlů) a tektogramatické (rozbor sémantiky, významu). Anotace probíhaly od nejjednodušího ke složitějšímu, tedy od morfologie k sémantice, čemuž odpovídá i množství označkovaných slov na jednotlivých rovinách. Slov s anotacemi na všech třech úrovních je 0,8 mil. (to odpovídá přibližně 50 000 větám), slov anotovaných morfologicky a analyticky 1,5 mil. a konečně na morfologické rovině je označkováno celkem 2,0 mil. slov (viz Obrázek 1). Svými vlastnostmi se PDT řadí k předním světovým korpusům.



Obrázek 1: Rozložení počtu anotovaných slov v PDT na jednotlivých rovinách

Pražský závislostní korpus zpřístupňuje nejenom nové možnosti ověřování dřívějších lingvistických teorií, ale především umožňuje vytváření (a rovněž ověřování) teorií nových, zvláště v oblasti statistických metod a metod strojového učení. Našim cílem bylo vytvořit počítačový systém, jenž bude využívat dat z PDT k sestavování úloh k procvičování tvarosloví a větných rozborů. Pokud je nám známo, tak u nás ani ve světě neexistuje podobný projekt, který by zpřístupňoval myšlenku jazykového korpusu školním dětem. Klademe si za cíl rovněž popularizovat PDT jako akademický produkt mezi širokou veřejností.

2 Motivace

Dnešní žáci základních a středních škol běžně počítače používají. Hrají hry, surfují po Internetu, chatují s kamarády, píšou si deník nebo malují. Snadno se nabízí otázka, proč by se prostřednictvím počítačů nemohli i vzdělávat, konkrétně v našem případě proč by nemohli určovat morfologické kategorie slov či rozebírat větu a označovat větné členy. Pochopitelně nelze očekávat, že by snad děti byly z této možnosti nadšené tolik jako z těch předchozích jmenovaných. Na druhou stranu však gramatiku procvičovat tak jako

tak potřebují a jak doufáme, pro mnohé z nich to bude takto snazší a třeba i zábavnější než při použití tradičních tištěných cvičebnic.

Řada elektronických učebnic a výukových programů již existuje, v tomto směru nejsme první. Dostupné produkty jsme testovali a nechali jsme se inspirovat jejich přednostmi, nebo jsme se naopak poučili z jejich chyb. Ačkoliv všechny testované programy ([5], [7], [4], [6]) nějakým způsobem umožňují procvičovat tvarosloví a syntax, ani jeden z nich se nepřibližuje našemu cíli – vzít větu a komplexně ji z těchto dvou hledisek rozebrat. Navíc žádný neobsahuje na výběr k procvičení více než několik desítek, možná stovek vět. Především však žádný z nich na úrovni syntaktické nejde dál než k určení některých (popřípadě všech) větných členů ve větě, uživateli nejsou závislosti mezi jednotlivými větnými členy ani zobrazeny (například se správným řešením), natož aby mu bylo umožněno si označování těchto závislostí procvičit. V tomto směru jsou výsledky naší práce jedinečné.

3 Sestavení cvičebnice

Chceme-li vytvořit cvičebnici češtiny (nebo obecně i jiného jazyka), můžeme postupovat dvěma způsoby. Za prvé si můžeme všechnu práci udělat sami ručně. Věty si buď vymyslíme nebo je odněkud opíšeme (případně zkombinujeme obojí) a jednu po druhé je zpracujeme, určíme všechny slovní druhy, jejich mluvnické kategorie, větné členy a závislosti a cvičebnice je hotová. Tento přístup má hned několik nevýhod. Je pro autora nesmírně pracný, rovněž je dost náchylný k chybám. Nejspíš se nepodaří dát dohromady bohatší výběr vět než několik desítek (možná stovek), ale především je vysoce pravděpodobné, že zvolené věty nebudou příliš dobře reflektovat skutečné používání jazyka – patrně budou v průměru jednodušší a kratší. Výhodou tohoto řešení je, že je lze použít prakticky vždy.

Alternativně se můžeme pokusit sestavit cvičebnici automaticky (nebo možná přesněji poloautomaticky), ovšem za předpokladu, že máme k dispozici anotovaný korpus. Tento postup odstraňuje nevýhody předešlého – nejtěžší práce je již hotová, korpus existuje a je označkován. Chyby se v něm sice zajisté rovněž vyskytují, ale pravděpodobně v podstatně nižším rozsahu. Anotování korpusových vět totiž obvykle provádí více lidí najednou a šance, že se anotátoři shodnou na chybném řešení, je relativně malá. Hlavně však je-li korpus sestaven tak, aby odrážel současný stav jazyka, bude tak činit i výsledná cvičebnice, stejně jako její velikost bude přímo úměrná velikosti celého korpusu.

V naší práci jsme se vydali právě touto druhou cestou. Jako anotovaný korpus jsme použili *Pražský závislostní korpus*. Jeho využití však nemohlo být úplně přímočaré, bylo potřeba provést řadu úprav, o kterých budeme pojednávat v následujících odstavcích.

4 Filtrování vět

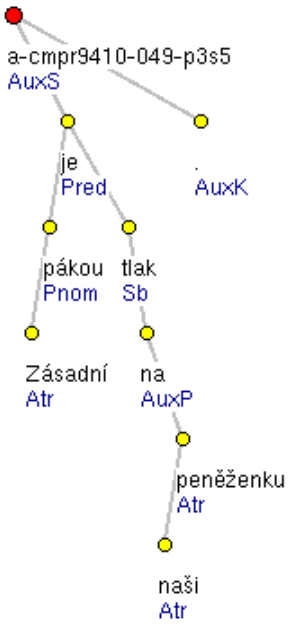
Na samém začátku jsme do cvičebnice zahrnuli ty věty z PDT, které jsou anotované na všech třech rovinách, tedy celkem 49 442 vět. Bohužel mezi těmito větami se vyskytuje celá řada vět nevhodných k procvičování žáků – vět obsahujících takové jevy, na jejichž klasifikacích se různé školské učebnice neshodují či je dokonce nezmiňují vůbec (ať už pro jejich komplexnost, nebo pro jejich okrajovost). Takovéto věty tedy bylo nutno z cvičebnice odstranit (protože není možné studenty procvičovat v látce, o níž se neučili), pochopitelně automatickou cestou (vzhledem k vysokému počtu vět).

Vzniklo tak několik filtrů, které byly na data použity a které z nich postupně odstraňovaly nevhodné věty. Nejdůležitějším z nich byl filtr odstraňující všechna souvětí, neboť ani na základních, ani na středních školách se neučí, jak zpracovávat souvětí. Stejně tak jsme se rozhodli mezi zbylými odstranit příliš dlouhé věty (více než dvacet slov), protože i když se jedná technicky o věty jednoduché, mohou svojí stavbou a složitostí žáky zbytečně mást, navíc největší část takovýchto vět tvoří dlouhé řetězce shodných či neshodných přívlastků, které jsou z hlediska procvičování „nezajímavé“. Obdobně jsme vyřadili i věty příliš krátké (méně než pět slov), obvykle pouze krátké nadpisy, které opět nemá velký smysl rozebírat. Celkem jsme filtrů sestrojili devět, další odstraňovaly například věty obsahující vsuvky, výpustky, přístavky nebo věty zcela bez přísudku. Podrobnosti lze nalézt v [2], kapitola 5. Z původních 49 442 vět, které do filtračního procesu vstupovaly, jich po všech filtracích zbylo **11 705**, tedy asi 24 %.

5 Transformace syntaktických stromů

Dalším krokem byla transformace anotací těch vět, které žádný z filtrů nevyloučil, do podoby, se kterou by uživatelé byli schopni snadno pracovat. Zatímco na úrovni morfologické toto nevyžadovalo většího úsilí, bylo již předem jasné, že na úrovni syntaktické bude nutné množství úprav.

Syntaktická úroveň PDT se totiž v nemálo ohledech značně liší od syntaxe vyučované ve školách. Předně obsahuje mnoho větných členů, ke kterým neexistují školské protějšky. Za druhé na ní každému slovu věty (včetně interpunkce) odpovídá právě jeden uzel, zatímco ve školské reprezentaci může obsahovat (a velice často také obsahuje) jediný uzel slov několik. Celému uzlu (nebo chceme-li všem těmto slovům) pak přísluší pouze jeden větný člen. Situace je ilustrována na Obrázku 2, který zachycuje větu *„Zásadní pákou je tlak na naši peněženku.“*.



Obrázek 2: Akademický rozbor věty *„Zásadní pákou je tlak na naši peněženku.“*

Transformacemi jsou pak myšleny operace nad takovýmto stromem, které jej převedou do podoby větného rozboru tak, jak je ve školách vyučován. Základními operacemi, které tvoří jádro všech transformací, jsou následující tři:

- *Připojení k rodiči.* Při této operaci daný uzel zaniká a všechna jeho slova jsou přesunuta do rodičovského uzlu. Provádí se například s jmennou částí přísudku nebo s pomocným slovesem *být*. V naší ukázce by operace byla provedena na uzlu *pákou*.
- *Pohlcení dětí.* Tato operace je k předchozí do jisté míry inverzní – je ekvivalentní provedení operace připojení k rodiči na všech dětských uzlech daného uzlu. Provádí se nejčastěji s předložkami či spojkami, v našem příkladu by byla provedena na uzlu *na*.
- *Přiřazení větného členu.* Tato operace pak pouze výsledným uzlům stromu přiřadí větné členy odpovídající patřičným analytickým funkcím PDT. Mapování mezi analytickými funkcemi a větnými členy shrnuje Tabulka 1.

Výsledek transformace ukázkové věty demonstruje Obrázek 3, více o transformacích viz [2], kapitola 6.

Větný člen PDT	**Školský větný člen**	**Popis**
Pred	Přs	slovesný přísudek
Pnom	Přj	jmenný přísudek
Sb	Po	podmět
Obj	Pt	předmět
Atr, AtrAdv, AdvAtr, AtrAtr, AtrObj, ObjAtr	Pk	přívlastek
Adv	Pu	příslovečné určení
Atv, AtvV, Obj	D	doplněk
Coord	–	koordinace
AuxC, AuxP, AuxZ, AuxO, AuxV, AuxR, AuxY, AuxK, AuxX, AuxG	–	pomocné větné členy

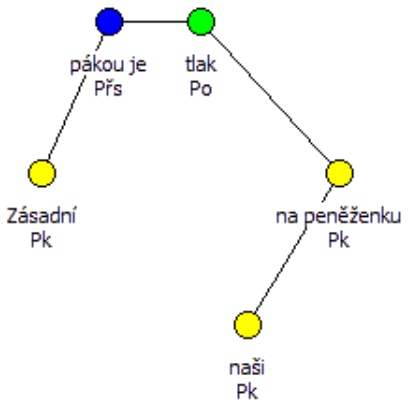
Tabulka 1: Analytické funkce PDT vs. školské větné členy

6 Implementace

Implementaci cvičebnice (a pomocných utilit) jsme se rozhodli provést v jazyce *Java*. K tomuto rozhodnutí jsme měli několik důvodů, mimo jiné jsme si tím zajistili možnost snadné přenositelnosti i na jiné operační systémy než MS Windows (přestože ty byly platformou, na kterou jsme se zaměřovali, a to z důvodu její rozšířenosti ve školách).

Pro práci s grafickým rozhraním byla použita knihovna SWT z projektu Eclipse[2]. Ta má oproti standardnímu modulu Swing obsaženému přímo v Javě dvě přednosti. Na každé platformě, pro kterou její implementace existuje, jsou použity standardní prvky této platformy, proto vzhled aplikace odpovídá tomu, na co je uživatel zvyklý. Navíc odezva grafického rozhraní je i rychlejší.

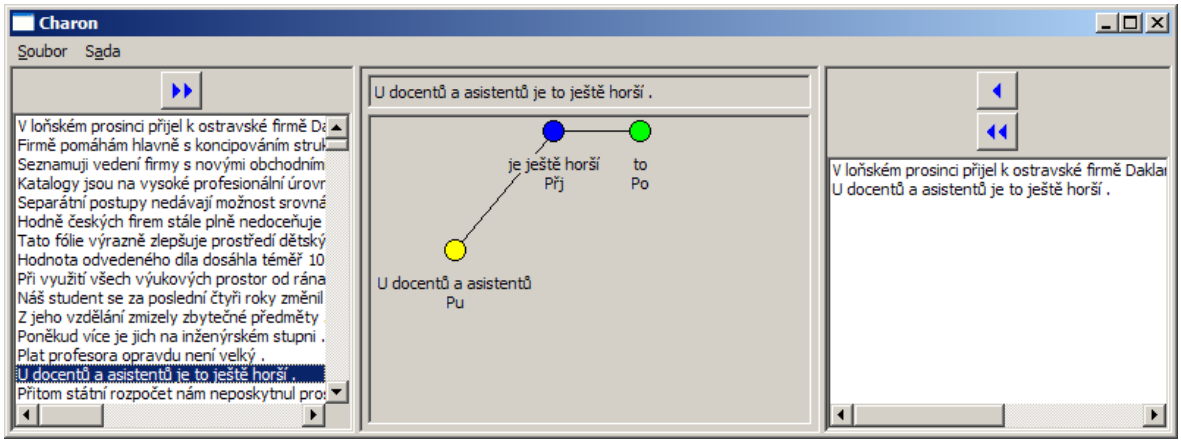
[2] Více viz http://www.eclipse.org/



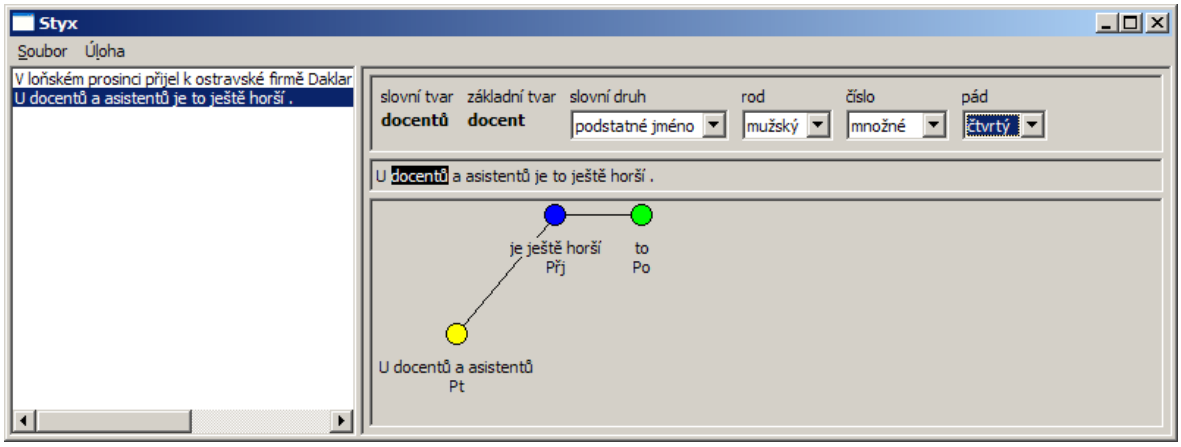
Obrázek 3: Školský rozbor věty *„Zásadní pákou je tlak na naši peněženku.“*

Samotným aplikačním výsledkem naší práce jsou krom tohoto textu především následující tři programy tvořící celý systém cvičebnice.

- ***FilterSentences***. Slouží k přípravě dat vhodných k použití ve cvičebnici (k filtrování vět), koncový uživatel s ním nepřijde do styku.
- ***Charon***. Administrační nástroj, slouží k zobrazování a prohlížení všech dostupných vět a vytváření cvičení (viz Obrázek 4). Předpokládá se, že jej bude používat vyučující.
- ***Styx***. Samotná cvičebnice, na které si budou žáci prověřovat své znalosti u cvičení vytvořených v programu Charon. Procvičování probíhá interaktivně na obrazovce počítače, jednotlivé hodnoty morfologických kategorií a větných členů jsou vybírány pomocí rozbalovacích seznamů či kontextových nabídek, provádění větných rozborů probíhá přímým posouváním uzlů po pracovní ploše metodou *drag & drop* (viz Obrázek 5).



Obrázek 4: *Charon* – sestavování cvičení



Obrázek 5: Procvičování v programu *Styx*

7 Závěr

Celý projekt cvičebnice češtiny postavené na Pražském závislostním korpusu úspěšně prošel první fází – podrobným prozkoumáním kroků, které je při zpracovávání PDT potřeba učinit, a vytvořením funkční implementace cvičebnice. Má-li však být skutečným přínosem, je potřeba v práci na něm dále pokračovat. Dalším krokem musí být přenesení projektu z čistě akademické sféry mezi uživatele, posbírat jejich připomínky a nápady a tyto do programu zapracovat.

Na domovské stránce projektu ([9]) je volně ke stažení aplikace Styx a cvičení. Zveme čtenáře, aby si aplikaci spolu s větami stáhli a vyzkoušeli si provést tvaroslovný a větný rozbor. Doufáme, že se nám podaří u čtenářů-uživatelů zahnat neveselé vzpomínky na nudné hodiny češtiny, během kterých se větné rozbory prováděly.

Literatura

1. HLADKÁ, Barbora, KUČERA, Ondřej. *Prague Dependency Treebank as an exercise book of Czech.* Proceedings of HTL/EMNLP 2005 Interactive Demonstrations, Vancouver, BC, Canada, 2005
2. KUČERA, Ondřej. *Pražský závislostní korpus jako cvičebnice jazyka českého.* Diplomová práce. Univerzita Karlova, 2005
3. KUČERA, Ondřej, HLADKÁ, Barbora. *Cvičebnice češtiny netradičním způsobem.* Tradičně netradiční metody a formy práce ve výuce českého jazyka na základní škole, Katedra českého jazyka a literatury PdF UP v Olomouci, v tisku, 2006
4. Silcom Multimedia. *Didakta Český jazyk 1* [software].
5. SRP, Pavel, SKLENÁŘOVÁ, Ivana, FRITSCHOVÁ, Martina. *Český jazyk, přijímací zkoušky na SŠ* [software]. 2004
6. ŠÁRA, Lubomír, ŠÁRA, David. *PON Škola – Český jazyk* [software]. 2003
7. Terasoft, a. s. *TS Český jazyk 2 – jazykové rozbory* [software]. 2003
8. HAJIČ, Jan, HAJIČOVÁ, Eva, PANEVOVÁ, Jarmila, SGALL, Petr, PAJAS, Petr, ŠTĚPÁNEK, Jan, HAVELKA, Jiří, MIKULOVÁ, Marie *Prague Dependency Treebank 2.0* [CD-ROM]. ISBM: 1-58563-370-4. Linguistic Data Consortium. 2006
9. *Styx* [online]. `http://ufal.mff.cuni.cz/styx/`