AUTOMATICKY VYTVOŘENÁ CVIČEBNICE ČEŠTINY

Kučera O., Hladká B., Hrstková K.
Ústav formální a aplikované lingvistiky, UK MFF, Praha

Abstrakt
Cílem naší práce je představit elektronickou cvičebnici českého jazyka o celkovém objemu 12 tisíc vět k procvičování ve dvou oblastech: tvarosloví (určování slovních druhů a jejich morfologických kategorií) a větný rozbor (určování větných členů a vztahů mezi nimi). Uživatelům je k dispozici pro všechny věty i klíč k řešení. Cvičebnice je sestavena z výběru vět, které obsahuje akademický Pražský závislostní korpus.

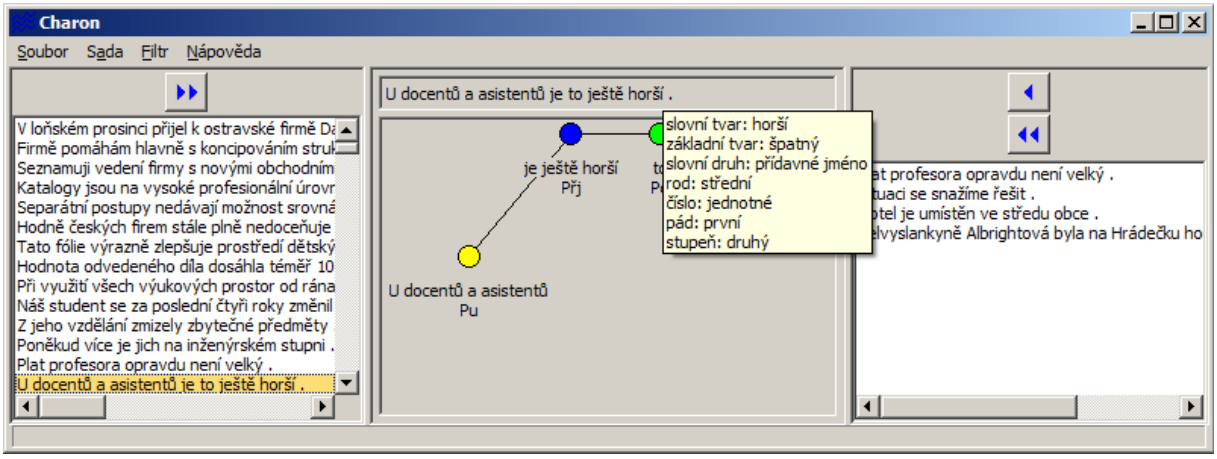
Motivace
V dnešní době již samozřejmě celá řada elektronických učebnic a výukových programů existuje, nabízí se tedy otázka, proč vytvářet nějaký další. Při práci na projektu jsme se seznámili s již existujícími produkty (např. [5], [6], [7], [8]) a snažili se prozkoumat jejich přednosti i nedostatky. Hlavním problémem, který u všech vidíme, je, že z pohledu tvarosloví a větného rozboru umožňují procvičovat vždy jenom některé vybrané jevy a žádný z nich neobsahuje cvičení na sestavení kompletního větného rozboru. Naším cílem je naopak umožnit uživatelům provádět komplexní rozbor (z hlediska morfologie i syntaxe) všech vět ve cvičebnici.

Budování cvičebnice
Cvičebnici jazyka chápeme jako sbírku vět, pro které jsou k dispozici řešení jazykovědných úkonů, na které je cvičebnice zaměřena. My se soustřeďujeme na cvičebnici pro procvičení tvarosloví a syntaxe.
Chceme-li vytvořit cvičebnici češtiny (nebo obecně i jiného jazyka), můžeme postupovat dvěma způsoby. Jedna možnost je udělat všechnu práci 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), to můžeme udělat ovšem pouze za předpokladu, že máme k dispozici tzv. *anotovaný korpus* – banku textů (řádově desítky tisíc vět) s doplněnými jazykovědnými informacemi u slov a vět. 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 slov a vět totiž obvykle provádí více lidí najednou a šance, že se anotátoři shodnou na chybném řešení, je relativně malá. Hlavně však je-li korpus dobře sestaven, 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* (dále PDT, viz [1]).

Implementace cvičebnice
Využití PDT však nemohlo být úplně přímočaré, obsahuje totiž množství vět, které jsou svou stavbou příliš složité, než aby mohly být ve cvičebnici použity. Takovéto věty bylo potřeba nejdříve ze vstupních dat odfiltrovat. Na ostatních větách jsme museli provést ještě řadu transformací odstraňujících rozdíly mezi akademickým přístupem k větnému rozboru a přístupem vyučovaným na základních a středních školách. Více o filtrování a transformacích viz [4].
Celý systém je naprogramován v jazyce Java s použitím grafické knihovny SWT z projektu Eclipse. Díky tomu je cvičebnice použitelná v prostředí MS Windows, GNU/Linuxu, Mac OS X (a pravděpodobně několika dalších platforem, které jsme nezkoušeli), navíc v každém z nich používá ovládací prvky pro danou platformu přirozené.
Cvičebnice ([2]) je rozdělena do tří samostatných aplikací.

FilterSentences
Konzolová aplikace *FilterSentences* slouží k aplikaci výše zmíněných filtrů na vstupní data PDT. Koncový uživatel s ní nikdy nepřijde do styku, bude pouze používat její výstup – celkový soubor vět cvičebnice.

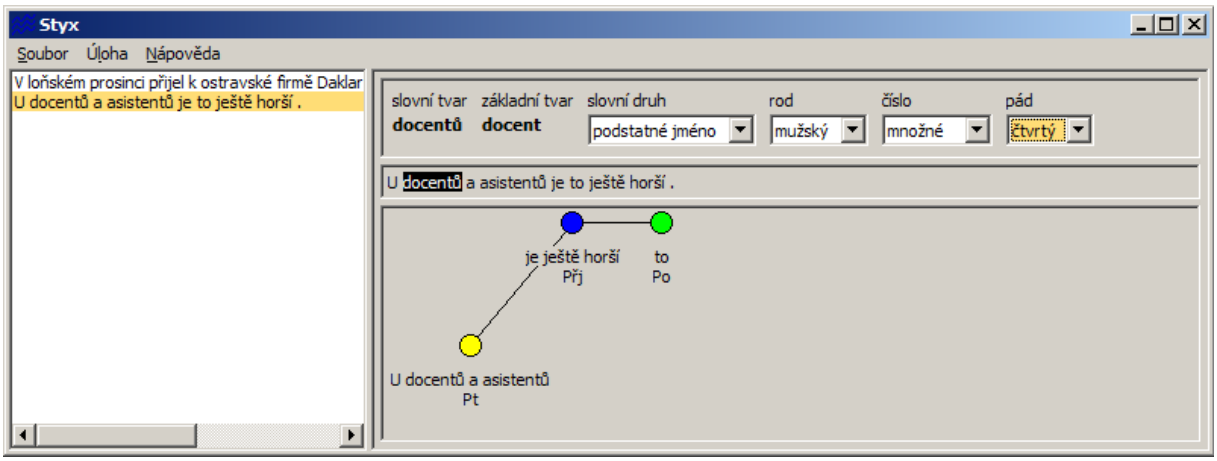
Charon
Program *Charon* (viz Obrázek 1) je administračním nástrojem určeným především vyučujícím. V jeho levé části jsou zobrazeny věty, které cvičebnice obsahuje, z nichž si uživatel může vytvářet vlastní cvičení. Pro vybranou větu je v prostřední části zobrazen její rozbor z hlediska tvarosloví a syntaxe. V pravé části potom uživatel vidí věty, které dosud do cvičení vybral.
Celkově cvičebnice obsahuje k použití přibližně 12 tisíc vět, pro větší přehlednost je toto množství ovšem rozděleno do deseti menších sad, mezi kterými lze snadno přepínat. Navíc je možno nastavit zobrazování pouze takových vět, které obsahují (nebo naopak neobsahují) určitý gramatický jev, například lze zobrazit pouze věty neobsahující nevyjádřený podmět.



Obrázek 1: Vytváření cvičení v programu *Charon*

Styx
Program *Styx* (viz Obrázek 2) je cvičebním programem samotným, na kterém si žáci mohou 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, tvorba syntaktických stromů probíhá přímým posouváním uzlů po pracovní ploše metodou

drag & drop.
Na závěr si studenti mohou zobrazit souhrnnou tabulku informující je o tom, kolik a jakých chyb udělali. Krom toho uvidí i vzorová řešení rozboru jednotlivých vět zobrazená vedle svých vlastních řešení, takže snadno uvidí, které jevy jim dělaly potíže (chybná řešení jsou navíc barevně odlišena).



Obrázek 2: Procvičování v programu *Styx*

Závěr
V loňském roce jsme cvičebnici při několika příležitostech představili na veřejnosti. Důležitá pro nás byla především prezentace na Gymnáziu Nad Alejí, v souvislosti s níž vznikl i metodický list ([3]), jenž může sloužit pedagogům pro připravení vyučovací hodiny se cvičebnicí Styx. Studentům, kteří si měli možnost na místě program přímo vyzkoušet, se samotný nápad i jeho realizace líbila, nicméně zároveň jsme od nich získali i několik užitečných připomínek.
V současné době se snažíme v systému pracovat především na vylepšování uživatelského rozhraní obou hlavních aplikací, Charonu a Styxu. V nejbližší době bude na webu zveřejněna nová verze obsahující právě několik novinek v této oblasti.

Reference

[1] *Pražský závislostní korpus*. http://ufal.mff.cuni.cz/pdt2.0/.
[2] *STYX*. http://ufal.mff.cuni.cz/styx.
[3] Klára Hrstková. *Metodický list pro první hodinu práce se cvičebnicí větného rozboru STYX*. 2007.
[4] Ondřej Kučera. *Pražský závislostní korpus jako cvičebnice jazyka českého*. Diplomová práce, Univerzita Karlova, 2005.
[5] Silcom Multimedia. *Didakta Český jazyk 1*. Software.
[6] Pavel Srp and Ivana Sklenářová and Martina Fritschová. *Český jazyk, přijímací zkoušky na SŠ*. 2004. Software.
[7] Lubomír Šára and David Šára. *PON Škola – Český jazyk*. 2003. Software.
[8] Terasoft, a. s. *TS Český jazyk 2 – jazykové rozbory*. 2003. Software.