KORPUS ŠKOLNÍCH ROZBORŮ

DOKUMENTACE K ROČNÍKOVÉMU PROJEKTU

KAROLÍNA KUCHYŇOVÁ

ZADÁNÍ

Tématem ročníkového projektu je vytvoření korpusu školních rozborů. Věty k rozboru jsou v aplikaci STYX (http://ufal.mff.cuni.cz/styx). Jde o cvičebnici obsahující 11 tisíc vět s tvaroslovným a větným rozborem. Věty jsou vybrány z Pražského závislostního korpusu (PDT, https://ufal.mff.cuni.cz/pdt3.0) tak, aby byly srozumitelné školákům. Jsou v aplikaci uloženy, ale jejich transformace z PDT anotací do školních rozborů se dělá za běhu.

Cílem projektu tedy je tyto věty transformovat do podoby školních rozborů a jejich korpus začlenit do repozitáře LINDAT (https://lindat.mff.cuni.cz/repository/xmlui/#) se všemi náležitostmi, včetně vytvoření aplikace, která umožní větné rozbory z korpusu prohlížet.

UŽIVATELSKÁ DOKUMENTACE

A) TRANSFORMACE NA ŠKOLNÍ ROZBORY

FORMÁT VÝSTUPNÍHO SOUBORU TRANSFORMACE

Věty z PDT v aplikaci STYX jsou anotovány na několika rovinách – morfologické, analytické a tektogramatické, přičemž každý typ anotace je v jiném souboru.

Tento formát převádíme na pozměněnou verzi formátu CoNLL-U. Výstupem je textový soubor (s kódováním UTF-8). Rozbory jednotlivých vět jsou uvozeny řádkem začínajícím křížem # s id dané věty a vzájemně odděleny prázdným řádkem.

V rozboru je každé slovo/token na jednom řádku ve stejném pořadí jako ve větě. Informace o slově jsou rozděleny do 8 polí oddělených jedním tabulátorem s následujícím významem:

1. ID: Index slova/tokenu, jeho pořadí ve větě. První token má číslo 1.
2. FORM: Tvar slova nebo interpunkční znaménko.
3. PDTLEMMA: Lemma (základní tvar) dle PDT.
4. PDTPOST: Tag slovního druhu a mluvnických kategorií slova dle PDT.
5. PDTDEPRAL: Analytická funkce dle PDT.
6. PDTHEAD: ID tokenu, na kterém dle PDT anotace dané slovo závisí, nebo 0, pokud na žádném dalším větném členu nezávisí.
7. SCHOOLDEPREL: Česká školní větná funkce slova (- pro koncovou interpunkci)
8. SCHOOLHEAD: ID větného členu, na kterém slovo závisí dle pravidel školních větných rozborů, nebo 0 pro podmět a přísudek (- pro koncovou interpunkci)

PRÁCE S PROGRAMEM

Program pro transformaci na školní rozbory bude uživatel spouštět z příkazové řádky příkazem (*doplnit*). Prvním parametrem spuštění je název adresáře, kde jsou uloženy soubory s PDT rozbory. V adresáři by měly být pro každou sadu vět tři soubory s rozbory - na rovině morfologické (soubory s příponou .m), analytické (soubory s příponou .a) a tektogramatické (soubory s příponou .t). Rozbory

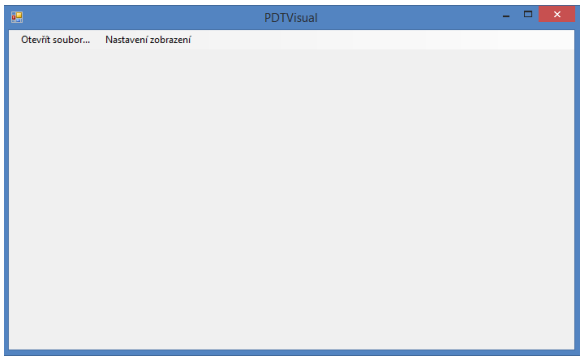
na úrovni slovní (přípona .w) program nepotřebuje, protože veškeré informace z nich se nachází i v morfologických rozborech.

Program prochází všechny soubory s příponou .m v adresáři a očekává, že ke každému budou existovat dva další stejně pojmenované soubory s koncovkami .a a .t. Pokud tyto soubory nenajde vypíše na příkazovém řádku zprávu, že soubor se nepodařilo najít.

Načtené věty jsou porovnány se seznamem vět ze STYXu a ponechají se jen ty, které jsou vhodné pro školní rozbory (dle kritérií uvedených v diplomové práci o STYXu). Dále probíhá transformace jednotlivých vět. Pokud se některou z vět nepodaří transformovat, program vypíše její identifikátor v chybové zprávě. Po dokončení transformace jsou věty ze všech souborů v adresáři uloženy do jediného souboru s příponou .o, jehož jméno dostává program jako druhý parametr svého spuštění.

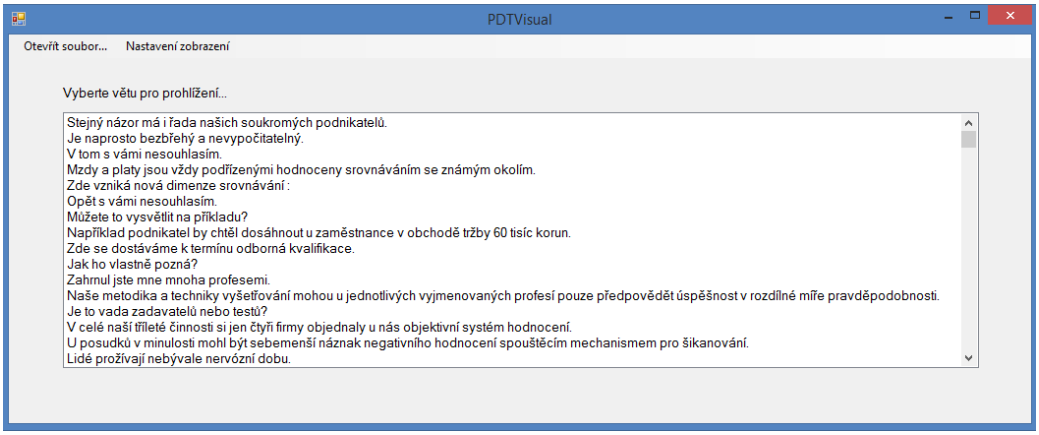
B) VISUALIZACE

Okenní aplikace umožňující prohlížet věty korpusu se spustí ze souboru PDTVisual.exe. Po zapnutí aplikace se objeví úvodní obrazovka.



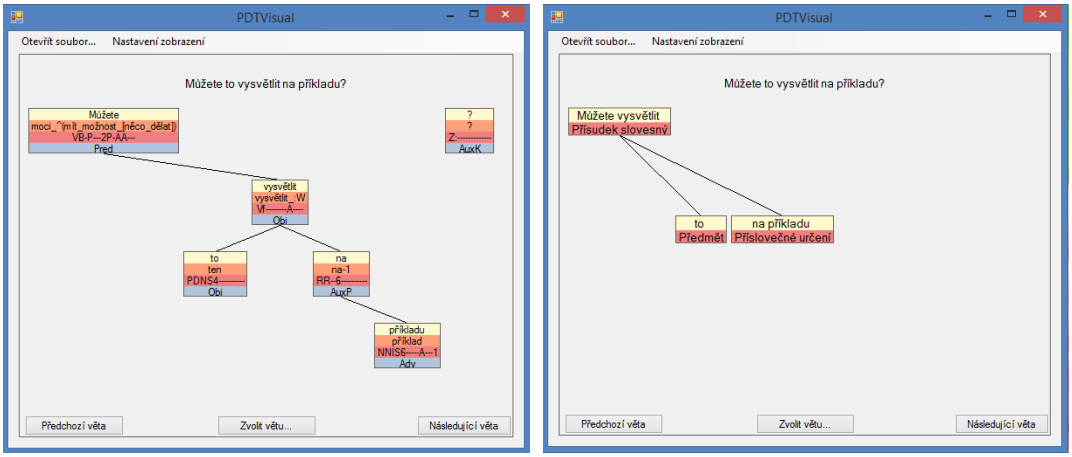
Korpusový soubor je možné vybrat pomocí menu v záhlaví a zvolení možnosti *Otevřít soubor*.... Je potřeba zvolit korpusový soubor ve formátu přesně popsaném výše. Vhodné soubory by měly mít příponu .o.

Po vybrání korpusu se objeví seznam vět v korpusu. V seznamu se dá pohybovat posuvníkem nebo pomocí šipek klávesnice. Dvojím kliknutím si uživatel může zvolit větu pro zobrazení.

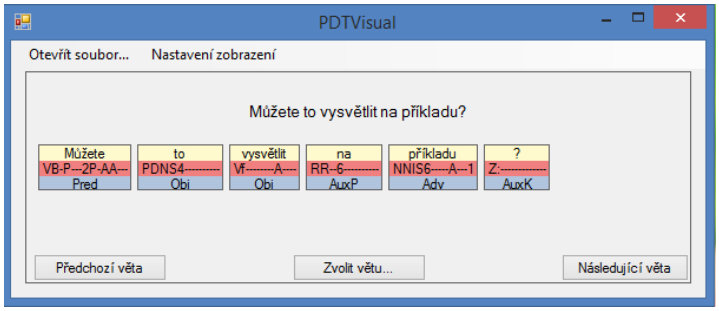


Po výběru věty se uživatel přesune do panelu zobrazení větného rozboru. Nahoře uprostřed je aktuálně prohlížená věta. Mezi větnými rozbory může uživatel přecházet použitím tlačítek *Předchozí věta* a

Následující věta, eventuálně se může vrátit na výchozí seznam všech vět zmáčknutím tlačítka *Zvolit větu*…. Velikost okna s rozborem si může uživatel nastavit libovolně.



Existuje více možností zobrazení větného rozboru. Základní je volba mezi rozborem podle PDT a školním rozborem. U PDT představuje každé slovo věty a každé interpunkční znaménko samostatný token, ke kterému se vztahují další informace – dle nastavení se může zobrazovat lemma, tag a funkce. Naopak u školních rozborů jsou tokeny spojeny do větných členů a doplňující informace se vztahují k těmto celkům. Uživatel se může rozhodnout, jestli chce zobrazovat popisek, o jaký větný člen se jedná. U obou formátů rozborů je možno zobrazit závislosti, pak jsou jednotlivé tokeny (respektive větné členy) v různé výšce a čáry představují závislostní vztahy (dolní člen závisí na horním). U školních rozborů je navíc dvojitou čarou spojena základní skladební dvojice. Pokud je zobrazení závislostí vypnuto, všechny tokeny nebo větné členy se zobrazují v řadě za sebou ve stejné výšce.



Uživatel si variantu zobrazení rozborů může navolit výběrem v *Nastavení zobrazení* v záhlavním menu. Tam si nejprve vybere, jestli má zájem o školní rozbor, nebo rozbor z PDT a dále pomocí zaškrtávání navolí další informace, které chce do rozboru zahrnout.