Diplomamunka tématerv

XML technológiák alkalmazása a számítógépes nyelvészetben

Készítette: Ferencsik István (???)

Témavezető: Dr. Alexin Zoltán

A számítógépes nyelvészetben a wordnet, elterjedt magyar nevén a fogalomháló vagy nyelvi ontológia a fogalmakra vonatkozó strukturált emberi tudás egyfajta gépi reprezentációja. A fogalomháló formálisan definiált fogalmak és relációk adatszerkezete, amelynek segítségével szemantikai következtetések végezhetők. A fogalomhálók alapegységei az egyes fogalmak szinonimáit egybegyűjtő szinonimahalmazok (angolul synsets), amelyeket különböző szemantikai relációk kötnek össze. A világban létrehozott fogalomhálók közül kiemelkedik a Princeton WordNet és az EuroWordnet kezdeményezés. A Princeton WordNet kizárólag angol nyelvű, míg EuroWordNet jelenleg nyolc nemzeti nyelven elérhető. Az EuroWordNet kifejezett célja a többnyelvűség támogatása. Jelenleg körülbelül még húsz további nyelven folyik Euro-WordNet-fejlesztés, köztük magyarul is.

A Lexical Markup Framework (LMF) egy ISO **(ISO 24613)** szabvány a természtes nyelvfeldolgozás és a számítógép álltal értelmezhető szótárak terülén amelynek célja hogy egy közös modellt adjon a nyelvészeti erőforrások készítéséhez és használatához, segítve ezen erőforrások közötti adatcserét, illetve lehetővé téve az egyes munkák összevonását, egyetlen globális, elektronikus erőforrásá olvasztását. Célszerű és hasznos feladat az eddig elkészült magyar fogalomhálót, ennek a szabványnak megfelelő formába átalakítani, ehhez a szabványhoz csatlakozni.

TrEd egy teljeskörűen programozható és testre szabható, grafikus felülettel rendelkező szoftver, amivel fa alapú struktúrákat jeleníthetünk meg és szerkeszthetünk. A program GNU licensz alatt hozzáférhető, széleskörűen elterjedt és használt. A VisDic egy a világon általánosan elfogadott program elsősorban a WordNet-ekkel kapcsolatos munkára tervezték, de bármilyen elektronikus szótárat is szerkesztetünk vele, többek között a magyar WordNet is ezzel készült.

Feladat:

A TrEd egy nagy hiányossága hogy saját bináris fájltípust használ. A cél az, hogy ezt a fájltípust átalakítsuk egy általánosabb formátumra ez pedig az XML ami széleskörűen használt a számítógépes nyelvészetben.

A feladat két program létrehozása a következő funkcionalitásokkal.

Mindkét program rendelkezzen egyszerű grafikus felülettel és paracsorból is futtatható legyen. A programok Windows platformra készülnek, C# programozási nyelven. A programok legyenek képesek egyszerre több fájlt is kezelni. A fejlesztési környezet a Microsoft Visual Studio 2008.

Az **első** program célja, egyrészt hogy a TrEd saját formátumából *.FS, XML formátumba konvertáljon illetve ugyanebből a formátumból LMF szabványnak megfelelő formájú XML fájlba konvertáljon.

A **második** program célja hogy a VisDic program által használt fogalomháló leíró XML file okat az LMF szabványnak megfelelő XML fájllá konvertáljon.

Szeged, 2009. 09. 12.