Szakdolgozat Tématerv

A Szeged Treebank által alkalmazott XML formátum ISO-24613 szabvány szerinti XML formátumra konvertálása, valamint az ISO-24613 szabványnak megfelelő formátum validálása

Témavezető:	**Dr. Alexin Zoltán**
Készítette:	**Fehér Zoltán**
EHA kód:	**FEZOACT.SZE**

A magyar nyelvben - számos más nyelvhez hasonlóan - a szavak toldalékai (morfémák) és a szószerkezetek (szintagmák) mondatban betöltött szerepe, valamint ezek egymáshoz való viszonya meghatározó szerepű. A természetes nyelvi feldolgozás fontos lépése a szintaktikai elemzés és annotálás, azaz a különböző szintaktikai egységek, pl. főnévi vagy melléknévi csoportok, névutós szerkezetek, igék és vonzatkeretük, stb. bejelölése. Mivel a mondatok többségében, az egész mondat jelentése szempontjából a főnévi csoportok kulcsfontosságú szerepet játszanak. A főnévi csoportok, mint az igék legjellemzőbb bővítményei, hordozzák az információk nagy részét, ezért pl. egy információkinyerési alkalmazásban kiemelkedő szerepük van.

Ezen kívül, ugyancsak a mondatok tartalmának értelmezhetősége szempontjából, fontos szerepe van a tagmondatok elkülönülésének és egymáshoz való viszonyának. További szintaktikai elemzésnél is fontos támpontot jelentenek a tagmondat-bejelölések, hiszen a tagmondathatáron szintaktikai egységek nem nyúlhatnak át.

A fent említetteket a Szeged Treebank 2.0 ki is elégíti. A szakdolgozat célja, hogy a már meglévő Szeged Treebank 2.0-ás XML-ben tárolt információhalmazt az ISO-24613 szabványban definiált formátumúra konvertálja. Az ISO 24613 szabvány egyik fő aspektusa az elektronikus lexikai erőforrások létrehozásának, karbantartásának és kiterjesztésének az optimalizálása tekintettel az általános és a természetes nyelvi feldolgozásra. A szabvány egyfajta absztrakt meta modellt (Lexical Markup Framework, a továbbiakban LMF) definiál, amely az elektronikus lexikonok létrehozására egy általános és standardizált keretrendszert biztosít. Az LMF egy olyan módszert biztosít, amely lehetővé teszi a kódolt lingvisztikai információk újrafelhasználását a különböző alkalmazásokban illetve a különböző feladatokban. Az LMF a lexikális objektumok egy általános reprezentációjára nyújt lehetőséget többek között az alaktani, szintaktikus és szemantikus aspektusokat is figyelembe véve.

Az LMF célja, hogy az elektronikus lexikai erőforrások létrehozására egy általános modellt biztosítson, legyen szó kisebb vagy nagyobb erőforrásokról. Célja, hogy ezen erőforrások közötti adatmozgást lehetővé tegye valamint, hogy elősegítse a már létező és igen nagyszámú egyedileg felépített elektronikus erőforrások beolvasztását így alkotva egy kiterjesztett globális elektronikus erőforrást. A legfőbb célja pedig az, hogy egy moduláris struktúra

létrehozásával elősegítse az elektronikus erőforrások közötti tartalmi együttműködést minden aspektust figyelembe véve.

Megvalósítás

A szakdolgozat célja, hogy a már meglévő Szeged Treebank 2.0-ás rendszerben létrehozott XML-ben tárolt információkat egy új az ISO-24613 szabványnak megfelelő formátumra hozza és azt szintén egy XML fájlban tárolja. Továbbá szükség lesz egy a szabványnak megfelelő formátumú XML-hez tartozó DTD vagy XSD fájl létrehozására is, mivel valahogy ellenőriznünk is kell, hogy egy adott XML formátuma a szabvány által definiált modellnek megfelel-e.

Az említett céloknak megfelelően, két különböző program implementálására fog sor kerülni. Az első program feladata, hogy a két XML formátum között egy XSLT transzformáció segítségével átjárást biztosítson. Ez az átjárás egyirányú, azaz csak a Szeged Treebank 2.0-ás formátumról lehet majd az új ISO-24613 szabványnak megfelelő formátumra konvertálni. A programnak lehetőséget kell biztosítania, hogy mind parancssori mind grafikus módon is használható legyen. Továbbá a programnak gondoskodnia kell a megfelelő memória használatról - lévén, hogy igen nagy XML adatbázisokról van szó - illetve az ebből származó hátrányokról (feldolgozási idő, virtuális memória használat, …) a felhasználót is valamilyen formában értesíteni kell.

A második program feladata, hogy egy adott XML-ről eldöntse, hogy az megfelel-e az ISO-24613 szabványban leírtaknak-e vagy sem melyet egy megfelelően létrehozott DTD vagy XSD fájl alapján fog kiértékelni. A programnak futtathatónak kell lennie, mind parancssori mind grafikus környezetben is egyaránt.

A programok fejlesztése Visual Studio 2008-ban C# nyelven fog történni. A programok által használt XSLT, DTD vagy XSD fájlokat és egyéb XML-el kapcsolatos vizsgálatokat valamilyen XML szerkesztőben fogom végrehajtani ilyen pl.az Oxygene XML Editor is.

Szeged, 2009. szeptember 11.