A HG-1 treebank: a nyelvtanírástól az online konkordanciáig

Tóth Ágoston[1]

[1] Debreceni Egyetem, Angol-Amerikai Intézet
4010 Debrecen, Egyetem tér 1.
tagoston@delfin.unideb.hu

Kivonat: Kutatócsoportunk a magyar nyelv LFG nyelvtanának kifejlesztését és egy treebank elkészítését tűzte ki célul. Az implementált nyelvtani jelenségek vonatkozásában a korpuszt mondattani annotációval látjuk el, majd egy kiválasztott alkorpuszt manuálisan egyértelműsítünk. Ehhez kapcsolódóan olyan eszközt is fejlesztünk, amellyel az alternatív elemzéseket kapó mondatok elemzését grafikus felületen szerkeszthetjük. A korpusz hozzáférhetőségének és kereshetőségének biztosítására online konkordanciaprogram-alkalmazást fogunk megvalósítani, így lehetővé válik a weben keresztüli keresés tetszőleges szóra, szófajra, alaktani jegyre, mindezt igény esetén meghatározott mondattani címkével ellátott összetevőre szűrve, a kimeneten megfelelően vizualizálva. Jelenleg a korpuszfejlesztési munka tervezési fázisának lezárásáról számolunk be, és áttekintjük a folyamatban lévő fejlesztéseket és az előttünk álló feladatokat.

1 Bevezetés

A Debreceni Egyetem Angol-Amerikai Intézetének Laczkó Tibor által vezetett Lexikai-Funkcionális Grammatikai Kutatócsoportja megkezdte egy 1,5 millió szavas, magyar írott nyelvi szövegeket tartalmazó treebank összeállítását és annotálását a saját készítésű LFG nyelvtanának felhasználásával. A készülő korpuszunkat, melynek a HunGram-1 (HG-1) nevet adtuk, mondattani annotációval látjuk el automatizált módon, majd egy kiválasztott alkorpuszt manuálisan ellenőrzünk és egyértelműsítünk. A kézi annotálás tapasztalatait a nyelvtanírásban felhasználjuk. A korpusz hozzáférhetőségének és kereshetőségének biztosítására online konkordanciaprogram-alkalmazást fejlesztünk.

2 Az elméleti keret és az implementációs környezet

A treebank projekt lényegi hátterét adó nyelvtanírási munka a **lexikai-funkcionális grammatikát** (LFG) használja, amely egy erős lexikalizmusra építő nem transzformációs generatív keret.

Kutatócsoportunk munkája közvetlenül kapcsolódik a nemzetközi **ParGram** együttműködéshez, melyben több nyelvhez (angol, német, norvég, francia, japán,

urdu, arab, török stb.) készül LFG nyelvtan, folyamatos egyeztetések mellett. A grammatika minden elemét – a többi ParGram projekthez hasonlóan – a Xerox Linguistic Environmentben (XLE) implementáljuk.

3 A treebank projekt feladatai

A tervezési fázisban meghatároztuk az alkalmazandó adatbázis-szerkezetet és a kifejlesztendő adatbázis-kezelő rendszer funkcióit. A mondattani fák tárolását a TigerXML [1] segítségével oldjuk meg, amely kiváló eszköz fák reprezentálására, és felhasználták – többek között – a Penn Discourse Treebank XML-re konvertálására is [2]. A projekt szoftver-infrastruktúráját – a parser kivételével – házon belül hozzuk létre.

Az induláshoz két kész korpuszt használunk nyersanyagként: a *Hunglish* korpuszt, mely egy nem annotált magyar-angol párhuzamos korpusz, és a *Szeged Treebank 2.0* korpuszt, ami egy 1,2 millió szavas magyar treebank (melynek annotációját projektünkben nem használjuk fel). Mindezt kiegészítjük egy saját gyűjtésű „nyers" korpusszal, ami főleg szépirodalmat, technikai dokumentációkat és híreket tartalmaz.

A programozási feladataink:

1) Mondatok elemeztetése a készülő nyelvtannal feltöltött XLE elemzővel, és a kimenet rögzítése (alternatív elemzésekkel). A korpuszt ettől a ponttól XML dokumentumban tároljuk.
2) Az összes lehetséges elemzés c-struktúrájának kibontása és tárolása.
3) Alkorpuszok kezelése:
   - korpuszfájlok darabolása és egyesítése,
   - indexelés, statisztikák készítése (faszélesség, -mélység, szavak és mondatok száma).
4) Kiválasztott mondat kézi egyértelműsítése, illetve annotációja saját fejlesztésű, grafikus felületű szerkesztőprogrammal, melynek a tervezett főbb funkciói a következők: ábrázolás, ágrajz kézi szerkesztése (melyhez bármelyik automatikusan generált elemzés kiindulópontként választható; a többszavas kifejezések lexikai egységként megjelölhetők, a morfológiai címkék megváltoztathatók; az ágrajzon élek és csomópontok létrehozhatók és törölhetők), a felhasználó által helyesnek vagy rossznak ítélt elemzések megfelelő feljelölése, megjegyzések elhelyezésének lehetősége.
5) Online lekérdezési felület a következő főbb funkciókkal:
   - keresés szóra vagy lemmára reguláris kifejezések használatával,
   - keresés szűrése morfológiai jegyekre és a keresett szót tartalmazó összetevőre (szűrés beállítása űrlap segítségével),
   - a találatok KWIC konkordanciaként való megjelenítése,
   - a konkordanciából kiválasztott mondat ágrajzainak megjelenítése.

A korpuszt tartalmazó XML dokumentumot több lépésben hozzuk létre, a fent említett eszközök segítségével. A nyelvtanunkkal feltöltött XLE parser PROLOG kódot ad vissza elemzésként, mely tartalmazza a „csomagolt" (a többértelműségeket részfákra lokalizáltan tároló) LFG c-struktúrát és f-struktúrát. Az XML fájl első változata

az eredeti mondaton (és az ahhoz kapcsolódó alapadatokon, valamint az esetlegesen meglévő angol fordításon kívül) ezt, az XLE-ből közvetlenül kapott elemzést tárolja. A következő lépésben a PROLOG kódból automatikusan létrehozzuk az összes lehetséges elemzést, majd ezzel egységes szerkezetben tároljuk a kézi annotáció eredményét.

A mondattani fák reprezentálását egy Tiger-XML alapú leírónyelv segítségével oldjuk meg. Egy ágrajz kódolása a gyökérelem kijelölésével indul, utána a terminális szimbólumok felsorolása következik, melynek során a lexikai egységekhez kapcsolódóan a szófajt, a lemmatizált alakot és a morfológia által visszaadott összes jegyet tároljuk. Ezt követi az összes többi csomópont leírása legalább 1-1 kapcsolódó él meghatározásával.

4 Felhasználási lehetőségek

A készülő korpusz (a tervezett lekérdezési lehetőséggel) felhasználható a nyelvoktatás, nyelvtanulás területén, a konkordanciaalapú megoldások összes előnyével: autentikus élőnyelvi szövegekkel dolgozhatunk olyan módon, hogy a tanulás nyelvi felfedezéssé válik. Ugyancsak fontosak számunkra a lehetséges lexikográfiai alkalmazások, valamint a korpusz felhasználása elméleti nyelvészeti kutatásokban: ez utóbbira példa a kutatócsoportunk nyelvtanírási projektje is, amelyhez a korpuszfejlesztési alprogram folyamatos tesztelési lehetőséget és visszajelzést biztosít.

Köszönetnyilvánítás

A munkát részben az OTKA (K 72983), részben a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatja. A projekt az Új Magyarország Fejlesztési Terven keresztül az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósul meg.

Bibliográfia

1. Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G.: The TIGER Treebank. In: Proceedings of the Workshop on Treebanks and Linguistic Theories (2002) 24–41
2. Yao, X., Borisova, I., Alam, M.: PDTB XML: the XMLization of the Penn Discourse TreeBank 2.0. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) (2010) 2022–2027