Morfoszintaktikailag annotált néprajzi korpusz[1]

Szauter Dóra[1], Vincze Veronika[1], Almási Attila[1], Alexin Zoltán[2], Kiss Márton[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{szauter, vinczev, mkiss}@inf.u-szeged.hu, vizipal@gmail.com
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
alexin@inf.u-szeged.hu

Kivonat: Az első, néprajzi tematikájú, nyelvileg elemzett magyar nyelvű korpusz szövegállománya a Néprajzi Múzeum Ethnológiai Adattárából származik. A szövegek két téma köré csoportosulnak: népi hiedelemvilág és táltosszövegek. A korpusz tartalmazza a szövegszavak lehetséges és az adott kontextusban helytálló morfoszintaktikai MSD-kódjait. A korpusz bővíthető más jellegű néprajzi szövegekkel, illetve a későbbiekben lehetséges lesz az állomány szintaktikai annotációjának elvégzése is.

1 Bevezetés

Cikkünkben bemutatjuk az első, néprajzi tematikájú, nyelvileg elemzett magyar nyelvű korpuszt. Újdonságot jelent egyrészről a korpusz tematikája, hiszen néprajzi témájú szöveges adatbázisok eddig nem vagy alig bizonyultak elérhetőnek elektronikus formátumban (a néprajzi adatbázis-építés nehézségeiről l. [2]), másrészről – tudomásunk szerint – magyar nyelvű néprajzi szövegeket még nem vetettek még alá számítógépes nyelvészeti elemzésnek.

A néprajzi korpusz feldolgozása követi a Szeged Treebankben [1] megszokott jelölésrendszert. A korpusz ebben az esetben is TEI XML formában készül, amelyben a szöveget szakaszokra, bekezdésekre, mondatok és szavakra bontják fel. Minden egyes szó mellett szerepelnek majd a lehetséges morfoszintaktikai elemzései, illetve a kontextusnak megfelelően kiválasztott morfoszintaktikai kód. A munka elvégzéséhez a kutatók azokat a szoftvereket fogják használni, amelyeket korábban a Szeged Treebank elkészítéséhez is igénybe vettek. Szükség esetén kisebb javításokat és korrekciókat végeznek a programokon.

A néprajzzal foglalkozó kutatók számára ez a fajta munka újdonságot jelent, mivel korábban a feldolgozásokat többnyire kézzel végezték. Sok esetben az összegyűjtött szövegek számítógépes formára hozása – begépelése, rögzítése sem történt még meg. Vélhetően ez a kisebb, mintegy 110 ezer szövegszó méretű korpusz elegendő vonzerőt gyakorol majd a néprajzos szakma képviselőire, hogy további anyagokat gyűjt-

[1] Az itt ismertetett kutatást az NKTH Jedlik Ányos program 2008, MASZEKER (Modell Alapú Szemantikus Kereső Rendszer) kódnevű kutatás-fejlesztési projektje támogatta.

senek össze, adjanak át feldolgozásra, s a tőlünk visszakapott anyag pedig újabb eredményeket hozhat a kutatásban.

A következőkben részletesen bemutatjuk a korpuszt, ismertetjük a nyelvi annotáció folyamatát, végül statisztikai adatokat közlünk az adatbázisról.

2 A korpusz tematikája

A néprajzi korpusz két témából tartalmaz szövegeket: népi hiedelemvilág (2704 szöveg) és táltosszövegek (432 szöveg). A szövegek lejegyzése a XX. század elején történt, a történelmi Magyarország csaknem minden tájegységéről származnak adatok. Az eredeti kéziratok a Néprajzi Múzeum Ethnológiai Adattárában találhatók, és gyűjteményes formában, könyv alakban is hozzáférhetőek [4].

A hiedelemszövegek a hétköznapi élet szinte valamennyi területéről tartalmaznak közléseket: az emberi élet fő állomásai (születés, keresztelő, férjszerzés, betegség, halál, túlvilág), időjárás, jeles napok, háziállatok. A rövid, egymondatos hiedelmeket hol magyarázat kíséri, hol rövid elbeszélések illusztrálják. A gyűjteményben egyszerű leírásokon kívül versformába szedett ráolvasások is találhatók. Bizonyos hiedelmek több változatban is előfordulnak. A szövegekből gyakran népszokáselemekre is következtethetünk:

Ha a menyasszony cipőjét ellopják a lakodalom éjjelén s lekaparva a talpáról a földet felteszik a füstre – ez a házas társak nyugodt életét megrontja.

A hiedelemközlésekhez fűzött megjegyzések az adott közösség életéről is hordoznak információt:

Ha a fiatal asszony közvetlen esküvő után 3-szor egymásután belenéz a kutba: meghal minden gyereke. Ez a szokás általános lett nálunk!

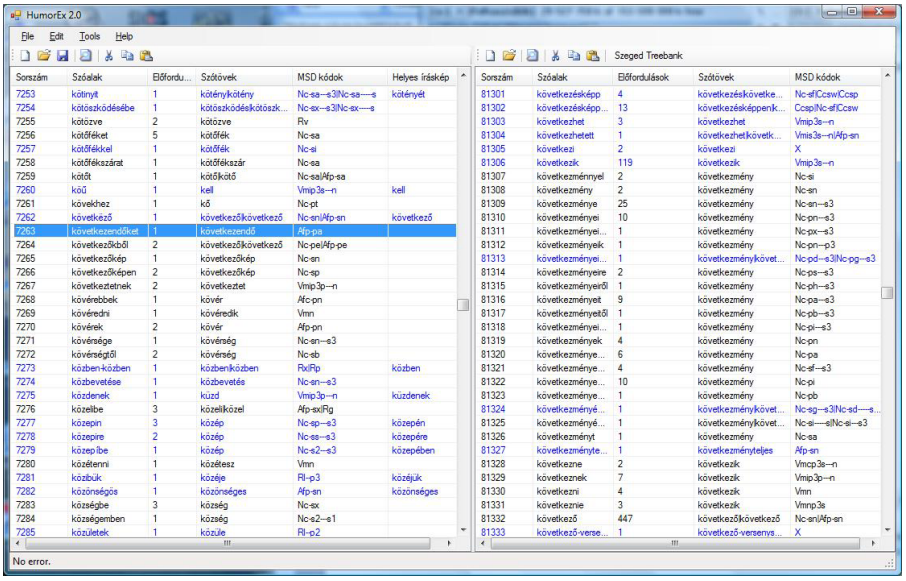
A táltosszövegekben a Kárpát-medence több tájegységéről található információ garabonciásokról, tudósemberekről, tudósasszonyokról, táltosokról, illetve azok ismertetőjegyeiről és képességeiről, leginkább róluk szóló rövid elbeszélések formájában, a tájegységnek megfelelő nyelvváltozatban.

3 Morfoszintaktikai annotáció

A korpusz szövegállományának digitalizálását követően Dalányi Sándor, a Stockholmi Egyetem kutatója kezdett foglalkozni az anyaggal. Egy közös kutatásfejlesztési projekt keretében jutottunk hozzá a szövegekhez, melyeken számítógép segítségével végzünk további nyelvi elemzéseket.

A feldolgozás első lépése a korpuszban található szavak összegyűjtése és morfoszintaktikai elemzése volt. A kapott 25 034 szóból álló listát a kutatók két részre bontották aszerint, hogy az adott szó megtalálható-e a Szeged Treebankben. Az ismert és korábban már elemzett szavakat ebben a munkafázisban félretettük, kódolá-

sukat egy az egyben átemeltük a Szeged Treebankből, és csak a korábban elő nem forduló, ismeretlen szavakkal foglalkoztunk. 14347 ilyen szó fordult elő a néprajzi szövegekben. Az annotálási munkálatokhoz az 1. ábrán látható programot használtuk. Először a szavakhoz számítógépes elemzéssel morfoszintaktikai kódokat rendeltünk, amelyeket azután át kellett nézni és jóvá kellett hagyni. Továbbra is tartottuk magunkat ahhoz, hogy a nyelvi elemzésben az Értelmező Kéziszótár kiadásaira támaszkodunk, annak a kategóriarendszerét vesszük át.



1. ábra. A szövegszavak morfológiai annotálásához készített szoftver.

A program két panelből áll, amelyekbe egyrészt az eddig feldolgozatlan szavak listáját (bal oldal), illetve a Szeged Treebank szótárát lehet betölteni (jobb oldal). Amennyiben az új szó csak kis mértékben, pl. esetragban tért el egy korábban már elemzett szótól, akkor a korábbi szóhoz rendelt morfológiai kódokat korrekcióval át lehet emelni. A programnak van is egy ilyen másolási funkciója. Új elemként jelent meg a program baloldali paneljében egy oszlop, amelyben a szavak ma szokásos írásmódját lehet megadni. Ha ez a modern alak előfordult a Szeged Treebank szótárában, akkor a program át tudta emelni a kódot a meglévő adatbázisból. A múlt századi vagy annál is régebbi népies vagy tájnyelvi szövegekben található szavaknál gyakori jelenség, hogy a helyesírásuk megváltozott.

A tájnyelvi szavak (*goroboncás*, *slájer*) mellett sajátos problémát jelentettek a következő esetek:

- népies helyesírású szavak (*ígízís*, *abbú*): ezek mellett feltüntettük a sztenderd magyar helyesírású alakot (*igézés*, *abból*), és ezek MSD-kódja a legtöbbször már átemelhető volt a Szeged Korpuszból. Amennyiben a szóalak nem szerepelt benne, akkor természetesen megadtuk a megfelelő kódo(ka)t.

- ha a népies helyesírású szó egybevág egy másik, létező szóalakkal (*mellül*, *aggyá*): ezek különös figyelmet igényeltek az egyértelműsítésnél, hiszen már volt egy – sztenderd helyesírás szerinti – MSD-kódjuk, azonban a szövegekben többnyire a népies változat fordult elő, így külön meg kellett adni annak sztenderd alakját (*mellől*, *adjál*) és MSD-kódját/kódjait.

A korpusz jelenleg morfoszintaktikai annotációt tartalmaz az MSD-kódrendszert [3] követve: minden szövegszó mellett szerepel annak összes lehetséges morfoszintaktikai kódja, és ezek közül az adott kontextusba illő is jelölve lesz (ez a munkafázis jelenleg zajlik).

4 Statisztika

A korpusz 109760 szövegszót tartalmaz összesen (a hiedelemszövegekben 65715, a táltosszövegekben 44045 szövegszó szerepel). Mivel a szövegszavak egyértelműsítése még folyamatban van, további statisztikákat például a morfoszintaktikailag egy-, illetve többértelmű szavak arányáról a későbbiekben közlünk.

5 További tervek

A morfoszintaktikai elemzésen kívül szintaktikailag is elemezni kívánjuk a teljes szövegállományt (dependenciaelemzés). A korpusz később esetleg más jellegű szövegekkel (például népmesék) is bővíthető.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceddings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123-131
2. Pávai I.: A néprajzi adatbázis-építés akadályai. Néprajzi Hírek 1-4 (1996) 86-89
3. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf
4. Verebélyi K. (szerk.): Néphit szövegek. Magyar Néprajzi Társaság, Budapest (1998)