

## Morfoszintaktikailag annotált néprajzi korpusz<sup>1</sup>

Szauter Dóra<sup>1</sup>, Vincze Veronika<sup>1</sup>, Almási Attila<sup>1</sup>, Alexin Zoltán<sup>2</sup>, Kiss Márton<sup>1</sup>

<sup>1</sup> Szegedi Tudományegyetem, Informatikai Tanszékcsoport  
H-6720 Szeged, Árpád tér 2.  
{szauter, vinczev, mkiss}@inf.u-szeged.hu, vizipal@gmail.com

<sup>2</sup> 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 egyrészt a korpusz tematikája, hiszen néprajzi témájú szöveges adatbázisok eddig nem vagy alig bizonyultak elérhetőnek elektronikus formátumban (a néprajzi adatbázis-építés nehézségeiről l. [2]), másrészt – tudásunk 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-

---

<sup>1</sup> 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 találhatóak, és gyűjteményes formában, könyv alakban is hozzáférhetőek [4].

A hiedelemszövegek a hétköznapi élet szinte valamennyi területéről tartalmaznak közléseket: az emberi élet fő állomásai (születés, keresztelés, 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 találhatóak. 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 ismeretőjegyeiről és képességeiről, leginkább róluk szóló rövid elbeszélések formájában, a tájegységnek megfelelő nyelvváltozatban.

## 3 Morfoszintaktikai annotáció

A korpusz szövegállományának digitalizálását követően Darányi Sándor, a Stockholmi Egyetem kutatója kezdett foglalkozni az anyaggal. Egy közös kutatás-fejlesztési 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 Treebankból, 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 Kézipiszótár kiadásaira támaszkodunk, annak a kategóriarendszerét vesszük át.

Sorszám	Szavak	Előfordu...	Szótövek	MSD kódok	Helyes íráské...
7253	kötény	1	köténykötény	Nc-sa-s3Nc-sa-s	kötényét
7254	kötőszókködebe	1	kötőszókködekköde...	Nc-ek-s3Nc-ek-s	
7255	kötözve	2	kötözve	Rv	
7256	kötőféket	5	kötőfék	Nc-sa	
7257	kötőfékel	1	kötőfék	Nc-ef	
7258	kötőfékaszat	1	kötőfékaszár	Nc-sa	
7259	kötőt	1	kötőkötés	Nc-sa/Np-sa	
7260	kötő	1	kell	Vmp.3a-n	kell
7261	következ	1	köt	Nc-pt	
7262	következő	1	következőkvetkező	Nc-ani/Np-en	következő
7263	következőkvetkező	1	következő	Rv.3a	
7264	következőkvetkező	2	következőkvetkező	Nc-ep/Np-pe	
7265	következőkép	1	következőkép	Nc-en	
7266	következőképen	2	következőkép	Nc-sp	
7267	következőketnek	2	következőket	Vmp.3p-n	
7268	következőknek	1	következő	Rf-on	
7269	következő	1	következők	Vmn	
7270	következő	2	következő	Afp-en	
7271	következő	1	következő	Nc-en-s3	
7272	következő	2	következő	Nc-sb	
7273	közbenközben	1	közbenközben	Rf/Np	közben
7274	közbevetése	1	közbevetés	Nc-en-s3	
7275	közdenek	1	küzd	Vmp.3p-n	küzdönek
7276	közébe	3	közékközé	Afp-en	
7277	középen	3	közép	Nc-ep-s3	középen
7278	középre	2	közép	Nc-s-s3	középre
7279	középre	1	közép	Nc-s2-s3	középreben
7280	középteni	1	középtész	Vmn	
7281	közéjük	1	közéje	Rf-p3	közéjük
7282	közönség	1	közönséges	Afp-en	közönséges
7283	közönség	3	közönség	Nc-sx	
7284	közönségben	1	közönség	Nc-s2-s1	
7285	közület	1	közüle	Rf-p2	

  

Sorszám	Szavak	Előfordulások	Szótövek	MSD kódok
81301	következőkvetkező	4	következőkvetke...	Nc-ef/Ccaw/Ccaw
81302	következőkvetkező	13	következőkvetkeppenk...	Ccaw/Nc-ef/Ccaw
81303	következőkvetkező	3	következőkvetke	Vmp.3a-n
81304	következőkvetkező	1	következőkvetke...	Vmia.3a-n/Np-en
81305	következőkvetkező	2	következőkvetke	X
81306	következőkvetkezők	119	következőkvetke	Vmp.3a-n
81307	következőkvetkezők	2	következőkvetke	Nc-e
81308	következőkvetkezők	2	következőkvetke	Nc-en
81309	következőkvetkezők	25	következőkvetke	Nc-en-s3
81310	következőkvetkezők	10	következőkvetke	Nc-en-s3
81311	következőkvetkezők	1	következőkvetke	Nc-ep-s3
81312	következőkvetkezők	1	következőkvetke	Nc-ep-s3
81313	következőkvetkezők	1	következőkvetke	Nc-pd-s3Nc-pg-s3
81314	következőkvetkezők	2	következőkvetke	Nc-ep-s3
81315	következőkvetkezők	1	következőkvetke	Nc-ph-s3
81316	következőkvetkezők	9	következőkvetke	Nc-pa-s3
81317	következőkvetkezők	1	következőkvetke	Nc-pb-s3
81318	következőkvetkezők	1	következőkvetke	Nc-pj-s3
81319	következőkvetkezők	4	következőkvetke	Nc-pk
81320	következőkvetkezők	6	következőkvetke	Nc-pa
81321	következőkvetkezők	4	következőkvetke	Nc-ef-s3
81322	következőkvetkezők	10	következőkvetke	Nc-pi
81323	következőkvetkezők	1	következőkvetke	Nc-pb
81324	következőkvetkezők	1	következőkvetke	Nc-sg-s3Nc-sd-s3
81325	következőkvetkezők	1	következőkvetke	Nc-sf-s3Nc-si-s3
81326	következőkvetkezők	1	következőkvetke	Nc-sa
81327	következőkvetkezők	1	következőkvetke	Afp-en
81328	következőkvetkezők	2	következőkvetke	Vmp.3a-n
81329	következőkvetkezők	7	következőkvetke	Vmp.3a-n
81330	következőkvetkezők	4	következőkvetke	Vmn
81331	következőkvetkezők	3	következőkvetke	Vmp.3a
81332	következőkvetkezők	447	következőkvetke	Nc-ani/Np-en
81333	következőkvetkezők	1	következőkvetke	X

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 (*ígízis, 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-kódrendszer [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: Proceedings 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.): MULTTEXT-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)