

## Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz

Csendes Dóra<sup>1</sup>, Hatvani Csaba<sup>1</sup>,  
Alexin Zoltán<sup>1</sup>, Csirik János<sup>1</sup>, Gyimóthy Tibor<sup>1</sup>, Prószéky Gábor<sup>2</sup>, Váradi Tamás<sup>3</sup>

<sup>1</sup> Szegedi Tudományegyetem Informatikai Tanszékcsoport  
6720 Szeged, Árpád tér 2.  
{dcsendes, hacso, alexin, csirik, gyimi}@inf.u-szeged.hu  
<http://www.inf.u-szeged.hu>

<sup>2</sup> MorphoLogic Kft Budapest  
1118 Budapest, Késmárki u. 8.  
[proszeky@morphologic.hu](mailto:proszeky@morphologic.hu)  
<http://www.morphologic.hu>

<sup>3</sup> MTA Nyelvtudományi Intézet  
1068 Budapest, Benczúr u. 33.  
[varadi@nytud.hu](mailto:varadi@nytud.hu)  
<http://www.nytud.hu>

**Absztrakt:** A Szeged Korpusz jelenlegi állapotában egy 1.2 millió szövegszóból álló szófajlag egyértelműsített, szintaktikai szempontból laposan elemzett adatbázis. Az elemzések szabályok alapján történő automatikus előelemzést követően kézi ellenőrzéssel és javítással történtek. A folyó munkálatok keretében egy bővebb szintaktikai elemzés, azaz egy magyar nyelvű treebank építése a cél, amelyben már szemantikai információk is szerepelni fognak. A korpusz regisztráció után hozzáférhető<sup>1</sup>, oktatási és kutatási célokra ingyenesen letölthető.

### 1. Bevezetés

A Szeged Korpusz jelenlegi állapota három projekt eredményeként született. Az első projekt (IKTA 2000)<sup>2</sup> keretében a Szegedi Tudományegyetem Informatikai Tanszékcsoportja a MorphoLogic Kft.-vel közösen egy 1 millió szövegszavas korpusz morfológiai elemzését és szófaji egyértelműsítését vállalta. A második projektben (NKFP 2001)<sup>3</sup> a konzorcium a Magyar Tudományos Akadémia Nyelvtudományi Intézetével kibővülve egy 200 ezer szavas szövegállomány feldolgozására, és az ez alapján történő automatikus információkinyerésre vállalkozott. Mivel a feldolgozás célja ezúttal

<sup>1</sup> A Szeged Korpusz a <http://www.inf.u-szeged.hu/III> oldalról tölthető le.

<sup>2</sup> A projekt az Oktatási Minisztérium által támogatott, *Magyar nyelvi szófaji egyértelműsítő módszer fejlesztése gépi tanuló algoritmusok felhasználásával* című IKTA 27/2000. pályázat keretében valósult meg.

<sup>3</sup> A projekt az Oktatási Minisztérium által támogatott, *Automatikus információszerezés rövid (politikai, üzleti, piaci) hírekből* című NKFP 2/017/2001 pályázat keretében valósult meg.

irányított információszerzés volt, ezért nem csupán morfológiai elemzést foglalt magába, hanem lapos szintaktikai elemzést (főnévi csoportok, ill. tagmondatok bejelölését), és bizonyos szintű szemantikai elemzést (a főnevek és melléknevek szemantikai attribútumainak megadását, igék és vonzatkereteik közötti szemantikai összefüggések azonosítását) is tartalmazott. A harmadik, jelenleg is folyó projekt (IKTA 2002)<sup>4</sup> keretében a MorphoLogic Kft.-vel és a Nyelvtudományi Intézettel karöltve az előző projektekben összegyűjtött, 1,2 millió szövegszavas korpusz bővebb szintaktikai feldolgozása, azaz egy treebank építése a cél.

A Szegedi Tudományegyetem Informatikai Tanszékcsoportjának egyik mesterséges intelligenciával foglalkozó kutatói csoportja 1998-ban kezdte meg munkáját a természetes nyelvi feldolgozás területén a fent említett projektek konzorciumi partnereként. A csoport a korpusz szövegeinek összegyűjtésében és az elemzett szövegek alapján történő gépi tanulási módszerek kialakításában vállalt aktív szerepet. Mivel a szövegek kézi annotálása teljes egészében e csoport irányítása alatt zajlott, ezért kapta az adatbázis a Szeged Korpusz nevet. A nyelvi feldolgozással kapcsolatos feladatokat a csoport menedzsmentjén kívül egy 4-10 fős programozói csapat és kb. 20 nyelvész-hallgató végzi. A korpuszon folyó tevékenységek két fő irányba mutatnak: egyrészt a nyelvfeldolgozás egymásra épülő szintjeinek kialakítására és annotálására, másrészt a már elkészült szintek minőségének javítására.

## 2. A Szeged Korpusz szöveganyaga

A korpusz szöveganyagának összeválogatásánál a fő szempont az volt, hogy a korpusz tematikailag minél sokszínűbb legyen, azaz a mai magyar nyelvnek több, egymástól különböző rétegét képviselje. Ennek megfelelően a Szeged Korpuszt hat, egymástól eltérő témájú szövegcsoporthoz alkotja. Az első öt csoport az említett IKTA 2000 projekt keretében került összeválogatásra a szövegszavak morfológiai elemzése és szófaji egyértelműsítése céljából. A hatodik csoport a későbbi NKFP 2001 projektben, kiegészítésként került a meglévő szövegekhez. Ez utóbbi szövegcsoporthoz az automatikus információszerzés alapjául. A vállalt projektmunkák, és a szövegeken ezen felül elvégzett elemzési munka eredményeként ma egy megközelítőleg 1,2 millió szövegszó + 225 ezer írásjel nagyságú, morfológiailag annotált, szintaktikai szempontból laposan elemzett korpusz áll rendelkezésre.

Természetesen az említett terjedelem még így sem teszi lehetővé egy átfogó, a teljes írott nyelvet felölelő szöveganyag összegyűjtését, de a konzorcium törekedett arra, hogy a korpusz a lehető legrepresentatívabb legyen, illeszkedjen az Interneten található nagy mennyiségű szöveg témáihoz. A fentieket figyelembe véve az alább részletezett témakörökből kerültek ki szövegek, egyenként kb. 200-200 ezer szavas terjedelemben. [1]

<sup>4</sup> A projekt az Oktatási Minisztérium által támogatott, *Mondatszintaxis gépi tanulása (gépi tanulási módszerek a magyar nyelv szintaktikai szabályainak létrehozására)* című IKTA 37/2002. pályázat keretében valósul meg.

Szövegtípus	Méret				Többértelmű szavak aránya szövegtípusonként	Szavak száma témakörönként
	100 mondatos fájlok száma	Szavak	Írásjelek	Többértelmű szavak		
Iskolás 1.	87	104818	22329	62705		
Iskolás 2.	73	97786	20705	52837		
Iskolás 3.	71	20454	4174	12279	57.30%	223058
Szépirod. Utás	52	60202	15946	33719		
Szépirod. Pizskos Fred	63	46578	14391	24284		
Szépirod 1984	65	80411	17631	42965	53.94%	187191
Szám.tech. Comp.World	62	124043	21018	57159		
Szám.tech. Win2000	31	57937	10888	25539	45.44%	181980
Újság (nsz)	41	63966	11980	31463		
Újság (nv)	13	22630	3900	11244		
Újság (hvg)	23	57647	9810	27990		
Újság (mh)	24	43091	7258	20678	48.78%	187334
Jogi (gazd.)	57	134872	22908	64165		
Jogi (szerz.)	35	87314	15807	42416	47.97%	222186
Rövid MTI	96	188345	25817	82813	43,7%	188345
<b>Összesen:</b>	<b>793</b>	<b>1190094</b>	<b>224562</b>	<b>592256</b>	<b>49.52%</b>	<b>1190094</b>

## 1. A korpusz összefoglaló adatai

*1. Általános iskolások fogalmazásai*

Ebben a témakörben 8. és 10. osztályos tanulók fogalmazásai kerültek feldolgozásra. A fogalmazások az ország különböző részein (Magyarországot egyenletesen lefedve), két különböző témában („Egy szép napom”, ill. „Miért (nem) szeretek iskolába járni?” címmel) íródtak. Az előbbiből 347 elbeszélő, az utóbbiból 492 érvelő dolgozatot tartalmaz a korpusz.

*2. Szépirodalmi szövegek*

Irodalmi alkotások közül három regény került feldolgozásra. Szerb Antal: *Utás és holdvilág*, George Orwell: *1984*, és Rejtő Jenő: *Pizskos Fred, a kapitány* című művek teljes szövege bekerült a korpusz szöveganyagába. (Az Orwell-korpuszban<sup>5</sup> korábban

<sup>5</sup> Az összesen 98.042 szövegszót, ill. írásjelet tartalmazó, ún. Orwell-korpuszt az 1995-97 között folyó MULTTEXT-EAST nemzetközi projekt keretében dolgozták fel.

már feldolgozott 1984 c. regényt a projekt keretében saját módszerekkel újra elemeztük.)

### 3. Újságnyelv

Magyar napilapok, ill. heti újságok közül a Magyar Hírlap 1999. január 4-i, a Népszabadság 1999. április 3-i, a Népszava 1999. november 25-i, és a HVG 1999. szeptember 4-i teljes száma került a korpuszba. A napi- és hetilapok számai a Népszava kivételével az újságok archív CD-ROM-jairól származnak, a Népszava cikkei pedig az újság internetes honlapjáról valók<sup>6</sup>.

### 4. Számítástechnikai szövegek

Számítástechnikai szövegek képviseltetése céljából az IDG kiadóvállalat *ComputerWorld Számítástechnika* c. újságjának összegyűjtött számai, valamint Kis Balázs: *Windows 2000 – haladó könyv haladó szoftverhez* c. könyvének 4., 5., 6. fejezete szerepelnek a korpuszban. A korpusznak ez a része az Interneten gyakran előforduló számítástechnikai, technológiai jellegű szövegeket reprezentálja.

### 5. Törvénytörvények

Jogi témájú szövegek közül az 1997. évi 144. törvény *A gazdasági társaságokról*, illetve az 1999. évi 76. törvény *A szerzői és szomszédos jogokról* teljes szövege került be a korpusz anyagába. A szöveg a *CD Jogtár: Hatályos magyar jogszabályok* CD-ROM-ról származik.

### 6. Rövid pénzügyi és gazdasági hírek

A különböző üzleti, tőzsdei, piaci és egyéb gazdasági témájú szövegek reprezentálása céljából az MTI-ECO gazdasági hírszolgálat *Business Plus* nevű adatbázisából származó hírek is bekerültek a korpuszba. Az anyag a 2001. márciusától 2002. januárjáig terjedő időszak híreit öleli fel. A hírek összeválogatásánál fontos szempont volt, hogy egyetlen bekezdésből álljanak. A korpusz összefoglaló adatai az 1. táblázatban láthatók.

Törekedve arra, hogy a korpusz nemzetközileg elfogadott szabványoknak megfeleljen, a szövegek reprezentációja XML-formátumú<sup>7</sup>, a leíró séma pedig a TEI DTD-nek<sup>8</sup> felel meg.

## 3. Morfológiai elemzés a Szeged Korpuszon

A korpusz nyelvi feldolgozásához, ill. a morfológiai elemzés megvalósításához a szövegeknek bizonyos előkészítő munkálatokon kellett átesniük. A feldolgozás egyik előfeltétele a szövegek mondatokra, majd szavakra bontása, vagyis a szöveg szavainak, nyelvi jeleinek szegmentálása, a tokenek kialakítása volt (pl. az írásjelek leválasztása az előtte álló szóról, több szóból álló tulajdonnevek megjelölése stb.). Ezután a kapott tokenekből alakult ki a szótár, amelynek címszavai tehát nem a szótövek, hanem az egyes előforduló szóalakok (jeles, ragos formák). A szótár elkészülte után a korpusz tokenjei mellé morfo-szintaktikai kódok kerültek.

<sup>6</sup> A Népszava honlapjának címe: <http://www.nepszava.hu/>.

<sup>7</sup> Az XML-honlap címe: <http://www.xml.org>

<sup>8</sup> A TEI DTD leírása az alábbi honlapon található meg: <http://www.tei-c.org>

Az így előállt szótárban a szófaji előelemzést a MorphoLogic Kft. által kifejlesztett HuMor magyar szófaji elemző programmal [6] végezte a csoport. A korpusz hivatalos szófaji kódolására azonban nem a HuMor által megadott kódokat, hanem az MSD-(Morpho-syntactic Description)-kódrendszer [3] magyar változatát használták, de ezzel párhuzamosan tárolásra kerültek a generált HuMor-kódok is. A HuMor-MSD közti megfeleltetést egy konvertáló programmal biztosították. Ennek a programnak a felhasználásával lehetett előállítani a még nem egyértelműsített, de a szófaji elemzés eredményét, vagyis minden lehetséges szófaji kódot tartalmazó korpuszfájlokat. Ha egy szóalak többjelentésű volt, és a jelentéskülönbségek morfológiailag megfogalmazhatók voltak, akkor egy-egy szócikkhez ennek megfelelően több MSD-kódot is tartozhatott. Az előkészítő munkálatok nyelvészeti vezetője Dr. Bibok Károly, a Szegedi Tudományegyetem Orosz Filológiai Tanszékének docense volt.

A fenti előkészületek után az egyetem 20 nyelvészhallgatója kézi ellenőrzés és javítás formájában végezte az egyértelműsítési munkát az előelemzett korpuszon. Az egyértelműsítés során a több kóddal rendelkező szavak esetében eldöntötték, hogy a lehetséges kódok közül melyik érvényes az adott szöveggörnyezetben, és ezt helyes kódként jelölték meg. A szófaji elemzésben és egyértelműsítésben felmerülő problémák megoldásakor a Magyar Értelmező Kéziszótár (1972)<sup>9</sup> volt az irányadó.

Az egyértelműsítési munka eredményeként elkészült az 1 millió szavas (első öt szövegcsoport) kézzel annotált magyar nyelvű tanuló adatbázis, amely egy automatikus szófaji egyértelműsítő módszer kialakításának alapja, és amely a további természetes nyelvi feldolgozáshoz kiindulásként szolgál. [2], [4], [5]

#### 4. Szintaktikai elemzés a Szeged Korpuszon

A szintaktikai elemzést több lépcsőben végzi a csoport. Első lépéseként a főnévi csoportok bejelölését vállalta. A választás nyelvészetiileg is megindokolható, hiszen a mondatot jellemzően egy ige (vagy más predikatív csoport), és annak (kötelező és szabad) bővítményei alkotják, a bővítmények pedig jellemzően főnévi csoportok. A magyar nyelvben ezek a bővítmények nemcsak nyelvtani szempontból játszanak központi szerepet, de a legtöbb esetben tartalmilag is ezek hordozzák a legfontosabb információt. Ezt a kezdeti elemzési lépést a nemzetközi szakirodalom lapos szintaktikai elemzésként (shallow parsing) azonosítja.

Definíció szerint a főnévi csoport (<NP> és </NP> címkék között) egy adott mondaton belül előforduló, egy vagy több szót, ill. írásjelet magába foglaló szerkezet, amelyre a következő feltételek teljesülnek:

- A csoport feje főnév (vagyis a csoportban szereplő utolsó szó főnévi szófajú);
- A csoport folyamatos (vagyis a szerkezetben szereplő szó/szavak, ill. írásjel/ek egymást követik);
- A csoport mérete maximális (vagyis nem lehet további elemmel bővíteni).

A főnévi csoportok lehetnek egymásba ágyazottak, de nem keresztezhetik egymást.

9 Juhász J., Szőke I., Nagy G. O., Kovalovszky M. (ed.): Magyar Értelmező Kéziszótár Akadémiai Kiadó, Budapest, (1972)

A morfológiai előfeldolgozáshoz hasonlóan a főnévi csoportok bejelöléséhez is egy automatikus előelemző programot használt a csoport. Az előelemzés előre meghatározott reguláris szabályok alapján a CLARK rendszer<sup>10</sup> segítségével történt. Ezt kézi szakértői ellenőrzés és javítás követte. Ott, ahol az előre meghatározott reguláris szabályok nem fedték le a főnévi szerkezeteket, a kézi annotálók speciális, a szerkezetre vonatkozó információval bővítették az annotálást. A munkálatok eredményeként előállt a korpusz elemzésének második szintje, amely a későbbiekben egyrészt gépi tanuló algoritmusok tanulási adatbázisaként szolgált, másrészt a természetes nyelvi mondatok szemantikai elemzésének kiindulópontja. [2], [4]

A következő lépésben a tagmondathatárokat keresték meg és jelölték be az annotálók. A (teljes) mondatban található tagmondatokat egy nyitó <CP> és egy záró </CP> címkével jelölték. A tagmondatok bejelölésénél egyrészt meg kellett határozni magukat a tagmondatokat, másrészt el kellett helyezni azokat az egész mondat szerkezetben. A tagmondatok alárendeltek vagy mellérendeltek lehetnek. Az alárendelt tagmondatok teljes egészében benne találhatóak egy másik (tag)mondatban, a mellérendelt tagmondatok együtt (egyenrangúan) alkotnak egy bővebb tagmondatot. Csak azok a kötőelemek tartoznak a tagmondatba, melyek abban mondatrészi szerepet töltenek be. Így pl. a hogy kötőszó nem része az alárendelt tagmondatnak, a vonatkozó névmások viszont igen.

## 5. Folyó és jövőbeli munkák

A természetes nyelvi feldolgozó csoport jelenleg folyó munkálatai egy bővebb szintaxis (*treebank*) kialakítására irányulnak. Ez egyrészt az igéknek, az igeneveknek és a főnévi csoportként nem azonosított mondat szintű bővítményeknek (határozószói csoportok, névutós csoportok, elváló igekötők) az annotálását, másrészt a főnévi csoportok belső szerkezetének finomítását jelenti.

A főnévi csoportként nem azonosított mondat szintű bővítmények bejelölése a következő irányelvek alapján történik. Névutós csoportot (PP: postpositional phrase) akkor kapunk, ha főnévi csoporthoz névutót kapcsolunk. A névutó általában az NP után áll, de ritkán meg is előzheti (Pistával szemben, szemben Pistával). Segítség: a névutók MSD-kódja: St.

A határozószói csoportok (ADVP: adverbial phrase) közé tartoznak a határozószók (*most, itt, tavaly, mindig* stb.; MSD-kódjuk első karaktere jellemzően: R), a „határozóragos” mellénevek (*betegesen, angolul, jogilag* stb.; MSD-kódjuk jellemzően: Afs-sw) és számnevek (*ötten, sokan* stb.; MSD-kódjuk: M?-sw?), és a „névutós” személyes névmások (*melleltem, miattad, utánunk* stb.; MSD-kódjuk: RI--??). Nem tartoznak ide a határozói igenevek (-va/-ve; -ván/-vén; MSD-kódjuk: Rv). Jellemző mondati szerepük: igei jellegű rész határozói bővítménye (*tegnap jött*), vagy fok- mértékhatózó (*nagyon finom*).

<sup>10</sup> A CLARK rendszert Kiril Simov, a Bolgár Tudományos Akadémia dolgozója fejlesztette ki a BulTreeBank projekt keretében (<http://www.bultreebank.org>).

Az elváló igekötők (pl: meg, el, be, stb.) bejelölése a <PREVERB> és </PREVERB> címkék segítségével történik. Meghatározásukban az igekötők MSD-kódja (Rp) lehet iránymutató.

A treebank kialakításának másik fontos eleme a főnévi csoportok belső szerkezetének finomítása, azaz a melléknévi jelzők bejelölése. A melléznevek (ADJP: adjective phrase) minőségjelzői szerepben találhatók, rendszerint egy NP fejének (főnévnek) jelzőjeként, a *milyen?* vagy a *melyik?* kérdésre válaszolva. A predikatív melléznevek NP-ként már be vannak jelölve, így ezeket nem jelöljük melléknévi szerkezetként, ill. ha a melléknév egy NP feje, akkor sem kap <ADJP>-címkét (pl. *az első* jöjjön be). Az ADJP-knek lehet bővítményük (leginkább ADVP), ill. mellérendelés is előfordulhat a melléznevek között, ha a két melléknév egyenrangú. Ez azt jelenti, hogy akár fel is cserélhető a sorrendjük, és tartalmilag így is ugyanaz a jelentésük marad (pl: szép és okos gyerek, okos és szép gyerek). Két egymás mellé rendelt ADJP újabb ADJP-szerkezetet alkot.

A fenti szintaktikai annotálás eredményét az alábbi példamondat szemlélteti:

```

<CP>
  <ADVP>
    Ekkor
  </ADVP>
  <NP>
    egy
  <ADJP>
    <ADVP>
      olyan
    </ADVP>
    hangos
  </ADJP>
  jármű
</NP>
robogott
<PREVERB>
  be
</PREVERB>
,
  hogy
  <CP>
    megfájdult
    <NP>
      tőle
    </NP>
    <NP>
      a
      fülünk
    </NP>
  </CP>
</CP>

```

Az igei csoportok bejelölése ugyanazt a stratégiát követi, mint a morfológiai elemzés és a főnévi csoportok bejelölése. Az annotátorok feladata ezúttal is egy automatikus előelemzés eredményének ellenőrzése, ill. javítása lesz. Az annotálás eredményeként a szövegben előforduló igék és igevek kerülnek bejelölésre, majd a szövegkörnyezetnek megfelelő hozzájuk tartozó vonzatkeretek egy táblázatból kerülnek kiválasztásra. A táblázatot egy domain-ontológia egyszerűsített változataként kell elképzelni, amely az egyes igék kötelező vonzatainak szemantikai attribútumait tartalmazza.

A csoport távolabbi terveiben szerepel egy bővebb ontológiai adatbázis (hierarchikus fogalomháló) kialakítása is, amely elsősorban a szemantikai információk annotálásában, a szemantikai szerepek azonosításában, és az ezek alapján történő információkinyerésben játszik fontos szerepet. A szemantikai annotálás során a különböző tartalmi elemek azonosítása és kódolása történik, ún. szemantikai keretek segítségével. [4] Ezt az említett NKFP projekt keretében a konzorcium tagjai meg is valósították a 200 ezer szavas rövid híreket tartalmazó szövegállományon. A továbbiakban a fennmaradó 1 millió szón végzi el az egyetem természetes nyelvi csoportja ugyanezt a munkát. A szemantikai keretek olyan kötelező, ill. szabad alegységek sorozatából állnak, amelyek ráilleszthetők a morfológiailag és szintaktikailag elemzett szöveg elemeire. Az alegységek ezáltal olyan információkkal telnek meg, amely alapján megbízhatóbban lehet válaszolni egy információszerzés céljából feltett kérdésre. Ha például valaki egy konkrét cég felvásárlása után érdeklődik, akkor a CÉG ADÁSVÉTELE szemantikai keretet felépítő VEVŐ, ELADÓ, CÉG, ÖSSZEG alegységekben szereplő információ alapján választ kaphat kérdésére.

## Irodalomjegyzék

1. Alexin Z., Csirik J., Gyimóthy T., Bibok K., Hatvani Cs., Prószéky G., Tihanyi L.: Manually Annotated Hungarian Corpus in Proc. of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03, Budapest, Hungary 15-17 April, (2003) pp. 53-56
2. Alexin Z., Váradi T., Pravec Cs., Prószéky G., Csirik J., Gyimóthy T.: FGT – A Framework for Generating Rule-based Taggers. ILP-99 Late-Breaking papers, Bled, Slovenia 24-27 June, (1999) pp. 1-7
3. Erjavec, T., Monachini, M., (ed.): Specification and Notation for Lexicon Encoding. Copernicus Project 106 „MULTEX-EAST”, Work Package 1 – Task 1.1, Deliverable D1.1F, (1997)
4. Hócza A., Alexin Z., Csendes D., Csirik J., Gyimóthy T.: Application of ILP methods in different natural language processing phases for information extraction from Hungarian texts in Proc. of the Kalmár Workshop on Logic and Computer Science, Szeged, Hungary, 1-2 October (2003), pp. 107-116
5. Horváth T., Alexin Z., Gyimóthy T., Wrobel S.: Application of Different Learning Methods to Hungarian Part-of-Speech Tagging in Proc. of the 9th International Workshop on Inductive Logic Programming (ILP99) Bled, Slovenia, in the LNAI series vol 1634, pp. 128-139, Springer Verlag (1999)
6. Prószéky G., Kis B.: A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages in Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA (1999), pp. 261-268