

## magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés

Zsibrita János<sup>1</sup>, Vincze Veronika<sup>2</sup>, Farkas Richárd<sup>1</sup>

<sup>1</sup> Szegedi Tudományegyetem, Informatikai Tanszékcsoport  
{zsibrita, rfarkas}@inf.u-szeged.hu

<sup>2</sup> MTA-SZTE Mesterséges Intelligencia Kutatócsoport  
vinczev@inf.u-szeged.hu

**Kivonat:** Ebben a cikkben bemutatjuk a magyarlanc nyelvi elemző újabb változatát, amely a hatékonyabb implementációnak köszönhetően a korábban publikált verzióhoz képest jóval gyorsabban képes magyar szövegek mondatra és szövegszavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére a pontosság javulása mellett. A magyarlanc 2.0 továbbá tartalmaz a mondatok függőségi nyelvtan szerinti szintaktikai elemzéséért felelős modult is. A rendszer teljes egésze JAVA-ban implementált, így platformfüggetlenül használható. Az elemző kutatási célokra bárki számára szabadon hozzáférhető.

### 1 Bevezetés

A természetesnyelv-feldolgozás magasabb szintű alkalmazásainak elengedhetetlen alapfeltétele a szövegek mondatokra és szavakra szegmentálása, a szövegszavak morfológiai elemzése és szófaji egyértelműsítése, illetve a mondatok szintaktikai elemzése. Cikkünkben bemutatjuk a magyarlanc elemző újabb változatát, amely a hatékonyabb implementációnak köszönhetően az eddiginél jóval gyorsabban képes magyar szövegek mondatra és szövegszavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, mindemellett újdonságot jelent a mondatok függőségi nyelvtan szerinti szintaktikai elemzése. A rendszer teljes egészében JAVA-ban implementált, így platformfüggetlenül használható.

### 2 Az elemző modulok

A korábban bemutatott magyarlanc 1.0 [11] magyar szövegek mondatra és szövegszavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére volt képes. Az újabb munkálatoknak köszönhetően az elemzés felgyorsult, illetőleg pontosabbá vált, továbbá a lánc kibővült a mondatok függőségi nyelvtan szerinti elemzésével, így tudomásunk szerint a magyarlanc 2.0 az első olyan eszköz, amely a szegmentálástól kezdve egészen a szintaktikai elemzésig képes végrehajtani a magyar nyelvű szövegek nyelvi előfeldolgozását.

## 2.1 Morfológiai elemzés

A szófaji elemző és egyértelműsítő (lemmatizáló és POS-tagger) a Stanford POS-tagger egy módosított változata, amely az ismeretlen szavakra a morphdb.hu-alapú [9] morfológiai elemző által adott lehetséges elemzéseket használja fel. Az elemzőt a kézi morfoszintaktikai annotációval rendelkező Szeged Korpuszon [2] tanítottuk, az eredeti MSD-kódok egy redukált kódhalmazán, azonban az elemzés eredményeképpen teljes értékű MSD-kódokat kapunk vissza. A kódhalmaz redukálásánál azt az irányelvet követtük, hogy a csökkentett kódkészletet használó szófaji egyértelműsítő modul kimenete egyértelműen megfeleltethető legyen az eredeti MSD-kódoknak. Tehát például az Nc-sd és Nc-sg kódok redukált alakja különbözik, míg a Nc-sd és Nc-sd---s3 ugyanarra a kódra redukálódik, mert soha nem fordulhat elő, hogy egy szóalaknak ez a két kód lehetséges elemzése (és a szófaji egyértelműsítőnek döntenie kell köztük).

A névszók (főnév, melléknév, számnév, névmás) és a nyílt tokenosztályba tartozó elemek esetében alapesetben az MSD-kódok a fő szófajra (az MSD-kód első pozíciójában álló elemre) redukálódnak. Birtokos és részes határozós esetük azonban egybeesik, ezért birtokos és részes esetben a redukált kódok megtartják az eset értékét (pl. Nd, Ng). Az essivusi (-an/-en, -ul/ül) és superessivusi (-n/-on/-en/-ön) esetragok szintén egybeeshetnek, pl.: *szépen*. Ezért superessivusi esetben a redukált kódok megtartják az eset attribútum értékét (pl. Ap). A névszók E/3. birtokos alakja egybeeshet a névszó birtokos nélküli alakjával, pl.: *Ajkán*. Ezért ezekben az esetekben szintén különbözőek lesznek a redukált kódok. Egy magas hangrendű névszó E/3. birtokos alakjának ragozott változata egybeeshet a névszó -é birtokjeles ragozott változatával, pl.: *énekét* (Nz és Ns). A névmások esetében a fenti megkötések mellett a három legfontosabb névmáscsoport (személyes, kérdő és vonatkozó) megtartja típusát is (Pe/Pq/Pr). Törtszámok esetén a redukált kódok megtartják a típust (Mf).

Alapesetben az igei MSD-kódok egyszerűen V-re redukálódnak. A segédigék kódja Va-ra redukálódik. A feltételes módú, T/1. és T/2. igék alanyi és tárgyias ragozású alakja egybeesik, pl.: *olvasnánk*. Ezért az ilyen igék tárgyias ragozású alakjainak MSD-kódja Vcp-re redukálódik. Az ikes igék E/1. alakjainak alanyi és tárgyias ragozása egybeesik, pl.: *iszom*, ezért a tárgyias ragozású MSD-kódok Vip-re redukálódnak. Feltételes módban, jelen időben, a magas hangrendű igék E/1. alanyi ragozású és T/3. tárgyias ragozású alakja egybeesik pl.: *ennék*. Ezért a T/3. tárgyias MSD-kódok alapesetben V3p-re redukálódnak. A kijelentő módú, múlt idejű, E/1. igék alanyi és tárgyias alakja egybeesik pl.: *osztottam*, ezért a tárgyias alakok MSD-kódja Vy-ra redukálódik. A felszólító módú igék kódja Vm-re redukálódik. Bizonyos esetekben egy adott ige múlt ideje és egy másik ige jelen idejű alakja egybeeshet (pl.: *ért*), ezért a korábbi szabályokra nem illeszkedő jelen idejű igék kódja Vp-re redukálódik.

Alapesetben a határozószói MSD-kódok egyszerűen R-re redukálódnak. A négy legfontosabb határozószó csoport (igekötő, kérdő, vonatkozó és személyes névmási) megtartják típusukat (Rp/Rq/Rr/Rl). A névelők MSD-kódja T-re redukálódik. A kötőszavak, névutók, indulatszavak, helyesírási hibát tartalmazó szavak, ismeretlen szavak és rövidítések esetében az eredeti MSD-kód megegyezik a redukált kóddal.

Az MSD-kódrendszer valamennyi attribútumához hozzárendeltünk egy-egy morfológiai jegyet, a magyar nyelv sajátosságainak megfelelő, a *CoNLL-2009 Shared Task*<sup>1</sup> kiírásnak eleget tevő módon. A szintaktikai elemző ezeket a jegyeket használja az elemzés során. Részletesen: típus – SubPOS, szám – Num, eset – Cas, birtokos száma – NumP, birtokos személye – PerP, birtok(olt) száma – NumPd, mód/forma – Mood, Idő – Tense, személy – Per, határozottság – Def, fok – Deg, klitikum – Clitic, forma – Form, mellérendelés típusa – Coord, altípus – Type. Az 1. táblázat mutatja, mely szófaj esetén mik a releváns jegyek.

1. táblázat: A szófajok és a morfológiai jegyek kapcsolata.

jegy	N	V	V	A	P	T	R	R	S	C	M	I	I	X	Y	Z	O	O
Sub-POS	•	•	•	•	•	•	•	l	•	•	•		o				•	eldn
Num	•	•	•	•	•			•			•						•	•
Cas	•			•	•						•						•	•
NumP	•			•	•						•						•	•
PerP	•			•	•						•						•	•
NumPd	•			•	•						•						•	•
Mood		•	n															
Tense		•																
Per		•	•		•			•										
Def		•																
Deg				•			•	•										
Clitic																		
Form										•	•							
Coord										•								
Type																		•

## 2.2 Szintaktikai elemzés

A szintaktikai elemzésnek két a leggyakoribb reprezentációs módja a konstituensfa és a függőségi fa. A függőségi fákkal dolgozó elemzők különösen jól használhatóak szabad szórendű nyelvek elemzésére, így a magyarra is, ezek ugyanis könnyebben teszik lehetővé az egymással nem szomszédos, de összetartozó szavak összekapcsolását is.

A függőségi elemzők két fő megközelítésre épülnek. A gráfalapú modellek a mondat szavait mint csúcspontokat tartalmazó teljes gráfon keresik a legvalószínűbb feszítőt [3,7]. A tranzakcióalapú modellek balról jobbra haladva szavanként elemzik a mondatot [6,8]. A magyarlanca beépítendő függőségi elemző kiválasztása előtt három függőségi elemző-implementációval is kísérleteket végeztünk: egy átmenetalapú modellt (Malt [8]) és két gráfalapú modellt (MST [7] és Bohnet-parser [1]) vizsgáltunk [4].

A mérések alapjául a Szeged Dependencia Treebank [10] szolgált. A treebank eredeti változatában a többtagú tulajdonnevek össze voltak vonva, azaz egy tokenként

<sup>1</sup> <http://ufal.mff.cuni.cz/conll2009-st/task-description.html>

voltak kezelve. A valóságban azonban nem létezik olyan algoritmus, amely hiba nélkül vonja össze a többtagú tokeneket, így méréseinkhez mi is több részre bontottuk ezeket. Az új tokenek tulajdonnévi kódot kaptak, alapértelmezett morfológiai jegyekkel, kivéve az utolsó tokent, amely megtartotta az eredeti elemzést. Így például a *Kovács és társa kft.* frázis az új annotációban N N N N szófaji kódokat kapott (megjegyezzük, hogy a Penn Treebank annotációs elveit követve N C N N kódokat kellett volna kapnia, azaz a tulajdonnévben előforduló kötőszó kötőszói kódot kap). A tulajdonnevek belső szintaktikai szerkezetét nem jelöltük be: egy láncot alkotnak az első tagtól az utolsóig. E döntések háttérében az áll, hogy legjobb tudomásunk szerint nem léteznek olyan alkalmazások, amelyek hasznosítani tudják a tulajdonnevek belső szerkezetét.

Az ige nélküli tagmondatok (túlnyomórészt névszói állítmány) esetében a Szeged Dependencia Treebank virtuális csomópontokat alkalmaz (16 000 előfordulás). A megoldás előnyei közé tartozik, hogy így hasonló faszerkezetet tulajdonítunk a mondatnak kijelentő mód jelen idő egyes/többes szám harmadik személyben, mint más módban, időben és számban/személyben.. A jelenleg elérhető szintaktikai elemzők azonban nem képesek a virtuális csomópontok megfelelő kezelésére. Éppen ezért, összhangban a Prague Dependency Treebankben alkalmazott megoldással [5], a virtuális csomópontokat töröltük a fából, és gyermekeiket a virtuális csomópont szülő csomópontjához kötöttük, illetve az Exd címkével láttuk el. Amennyiben a mondat gyökéreleme virtuális csomópont volt, ennek törlése azt eredményezte, hogy a mondatban nem maradt gyökérelem, aminek következtében az ilyen mondatokat kiszűrtük a korpuszból, és kísérleteinkben nem használtuk fel őket.

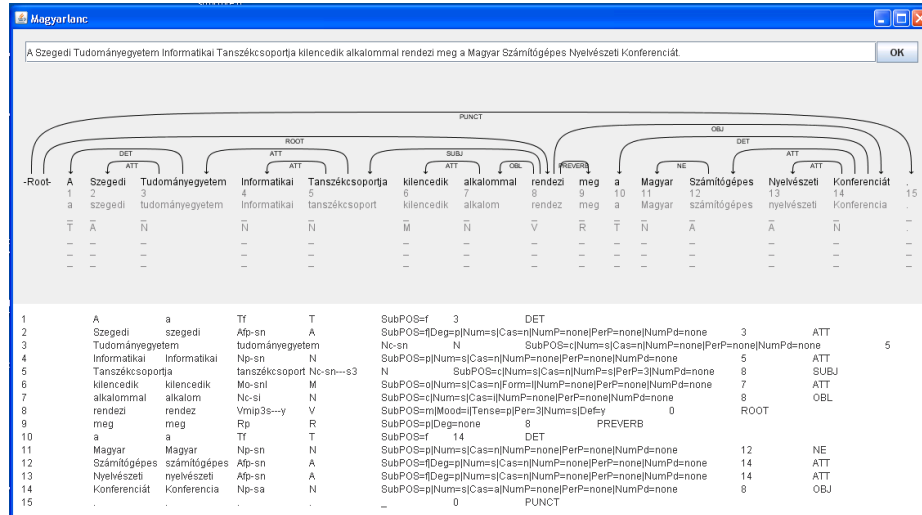
Mivel a kísérletek eredményei alapján a Bohnet-parser bizonyult a legpontosabbnak és leghatékonyabbnak, így ezt a függőségi elemzőt integráltuk az elemző láncba. A magyarul 2.0-ba integráltuk a *whatswrong*<sup>2</sup> megjelenítőt is, így a szintaktikailag elemzett mondatok ágrajzának vizuális megjelenítésére is van lehetőség.

### 2.3 Az elemző lánc kimenete

Az 1. ábra bemutatja az elemző felhasználói felületét egy mintaelemzéssel. A képernyő közepén látható a mondat függőségi elemzése, majd a képernyő alján található a részletes morfológiai elemzés.

---

<sup>2</sup> <https://code.google.com/p/whatswrong/>



1. ábra. A magyarlanc 2.0 felhasználói felülete.

Az elemzett kimeneti fájlok felépítése a következő. Egy sor egy tokennek felel meg, a mondatokat üres sorok választják el egymástól. Az első oszlopban a token mondatbeli sorszáma, a másodikban a szóalak, a harmadikban a lemma, a negyedikben az MSD-kód, az ötödikben a szófaj, a hatodikban a morfológiai jegyek, a hetedikben a szülő csomópont sorszáma, a nyolcadikban pedig a függőségi élcímke látható. Az alábbiakban közlünk egy példát a kimeneti fájlformátumra.

```

1  Az      az      Tf      T      SubPOS=f      2      DET
2  elnök  elnök  Nn-sn  N      SubPOS=n | Num=s | Cas=n | NumP=none | PerP=none | NumPd=none 3      SUBJ
3  megígérte  megígér  Vmip3s---y  V      SubPOS=m | Mood=i | Tense=s | Per=3 | Num=s | Def=y 0      ROOT
4  '      '      '      '      3      PUNCT
5  az      az      Tf      T      SubPOS=f      7      DET
6  észlelt  észlelt  Afp-sn  A      SubPOS=f | Deg=p | Num=s | Cas=n | NumP=none | PerP=none | NumPd=none 7      ATT
7  hibákat  hiba  Nn-pa  N      SubPOS=n | Num=p | Cas=a | NumP=none | PerP=none | NumPd=none 14      OBJ
8  a      a      Tf      T      SubPOS=f      9      DET
9  szövetség  szövetség  Nn-sn  N      SubPOS=n | Num=s | Cas=n | NumP=none | PerP=none | NumPd=none 10      ATT
10  vezetése  vezetés  Nn-sn---s3  N      SubPOS=n | Num=s | Cas=n | NumP=s | PerP=3 | NumPd=none 14      SUBJ
11  45      45      Mc-snd  M      SubPOS=c | Num=s | Cas=n | Form=d | NumP=none | PerP=none | NumPd=none 12      ATT
12  napon  nap  Nn-sp  N      SubPOS=n | Num=s | Cas=p | NumP=none | PerP=none | NumPd=none 13      OBL
13  belül  belül  St      S      SubPOS=t      14      TLOCY
14  kijavítja  kijavít  Vmip3s---y  V      SubPOS=m | Mood=i | Tense=p | Per=3 | Num=s | Def=y 3      ATT
15  .      .      .      .      -      0      PUNCT

```

### 3 Eredmények

A magyarlanc 2.0 elemzési pontosságát megállapítandó kísérleteket végeztünk mind a szófaji egyértelműsítés, mind a szintaktikai elemzés terén. Méréseinkhez a Szeged Dependencia Treebanket [10] használtuk fel. A treebank mondatait véletlenszerűen osztottuk fel tanító (80%) és kiértékelési (20%) adatbázisra. Az alábbiakban ezen mérések eredményeit ismertetjük.

#### 3.1 A szófaji egyértelműsítés eredményei

A szófaji egyértelműsítés a tesztadatbázison 96,33%-os pontosságot ért el. Az átalakításoknak köszönhetően a korábbi magyarlanc 1.0 verzióhoz képest pontosabbá vált a számok és a nyílt tokenosztályba tartozó tokenek elemzése.

#### 3.2 A szintaktikai elemzés eredményei

A szintaktikai elemzés kiértékeléséhez a Labeled Attachment Score (LAS) és az Unlabeled Attachment Score (ULA) metrikákat használjuk. A LAS esetében a teljes egyezéshez szükséges mind a szülő, mind az élcímke egyezése az etalonhoz képest, míg az ULA esetében elégséges a szülő csomópont egyezése (itt nem számít hibának a rossz élcímke). A függőségi elemzés a tesztkorpuszon 91,42%-os (LAS) és 93,22%-os (ULA) eredményt ért el.

#### 3.3 Az elemzés sebessége

A magyarlanc jelen verziójának működési sebességét az Egri csillagok című regényen teszteltük. A teljes elemzési láncot futtatva 1 GB RAM felhasználásával percenként 1000 mondat elemzése történik meg. Amennyiben csak szófaji egyértelműsítést szeretnénk végezni, az 3000 mondat/perc sebességgel zajlik, ami a korábban publikált 1.0 verzióhoz képest harmincszoros gyorsulást jelent.

## 4 Összegzés

Cikkünkben bemutattuk a magyarlanc 2.0 elemző láncot, amely magyar nyelvű szövegek nyelvi előfeldolgozására – szegmentálás, morfológiai elemzés, szófaji egyértelműsítés és szintaktikai (függőségi nyelvtan szerinti) elemzésre – hivatott, és a korábban publikált verzióhoz képest jóval gyorsabban képes mindezt megnevelkedett pontosság mellett. A rendszer teljes egészében JAVA-ban implementált, így platformfüggetlenül használható. Az elemző lánc kutatási célokra szabadon hozzáférhető a <http://www.inf.u-szeged.hu/rgai/magyarlanc> oldalon.

## Köszönetnyilvánítás

A kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

## Hivatkozások

1. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (2010) 89–97
2. 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
3. Eisner, J. M.: Three new probabilistic models for dependency parsing: an exploration. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96 (1996) 340–345
4. Farkas, R., Vincze, V., Schmid, H.: Dependency Parsing of Hungarian: Baseline Results and Challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (2012) 55-65
5. Hajič, J., Böhmová, A., Hajičová, E., Vidová-Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Abeillé, A. (ed.): Treebanks: Building and Using Parsed Corpora. Amsterdam, Kluwer (2000) 103–127
6. Kudo, T., Matsumoto, Y.: Japanese dependency analysis using cascaded chunking. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20, COLING-02 (2002) 1–7
7. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: Non-Projective Dependency Parsing using Spanning Tree Algorithms. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (2005) 523–530
8. Nivre, J., Hall, J., Nilsson, J.: Memory-Based Dependency Parsing. In: HLTNAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004) (2004) 49–56
9. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation (2006)
10. Vincze, V. Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010)
11. Zsibrita J., Vincze V., Farkas R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem (2010) 275-283

## Szerzői index, névmutató

- Ács Zsombor, 289  
Alberti Gábor, 236
- Berend Gábor, 251  
Biró Tamás, 21
- Csernyi Gábor, 85  
Csirik János, 213
- Dobó András, 35, 213  
Durst Péter, 97
- Ehmann Bea, 121  
Endrédy István, 297
- Farkas Richárd, 193, 251, 263, 289,  
368  
Fegyő Tibor, 13
- Grósz Tamás, 3
- Gyarmathy Zsófia, 275
- Héja Enikő, 59  
Hussami Péter, 135, 302
- Indig Balázs, 305, 310
- Jelasity Márk, 251
- Károly Márton, 236, 318  
Kilián Imre, 225, 236  
Kiss Gábor, 324  
Kiss Márton, 324  
Kleiber Judit, 236  
Kornai András, 62
- Lackó Tibor, 85  
Laki László János, 71, 331  
László János, 121  
Lendvai Piroska, 121  
Ludányi Zsófia, 135
- Makrai Márton, 62  
Mátyus Kinga, 338  
Mihajlik Péter, 13  
Miháltz Márton, 121, 135, 343  
Mittelholcz Iván, 135
- Nagy Ágoston, 135  
Nagy T. István, 47  
Nagy Tímea, 13  
Nemeskey Dávid Márk, 106, 346  
Novák Attila, 71, 148, 159, 170, 297
- Oravecz Csaba, 135  
Orosz György, 159, 331
- Pintér Tibor, 135  
Pólya Tibor, 124  
Prószéky Gábor, 148, 159, 310  
Pulman, Stephen G., 35
- Rákosi György, 85  
Recski Gábor, 346
- Sass Bálint, 348  
Siklósi Borbála, 71, 148  
Simon Eszter, 106  
Simonyi András, 275  
Subecz Zoltán, 263



Szabó Martina Katalin, 97  
Szász Levente, 124  
Szécsényi Tibor, 205  
Szekeres Péter, 351  
Szóts Miklós, 275

Takács Dávid, 59, 135  
Tarján Balázs, 13  
Tóth Ágoston, 85, 354  
Tóth László, 3

Vadász Noémi, 236  
Vincze Orsolya, 121  
Vincze Veronika, 47, 97, 182, 251,  
361, 368

Wenszky Nóra, 170

Zséder Attila, 346  
Zsibrita János, 47, 97, 251, 361, 368