Univerzális dependencia és morfológia magyar nyelvre

Vincze Veronika[1,2], Farkas Richárd[1], Simkó Katalin Ilona[1],
Szántó Zsolt[1], Varga Viktor[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{kata.simko, viktor.varga.1991}@gmail.com, {szantozs,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat Ebben a cikkben beszámolunk az univerzális dependencia és morfológia elveinek magyarra történő alkalmazásáról, és bemutatjuk a kihívást jelentő nyelvi jelenségeket és az azokra nyújtott megoldásainkat. A kidolgozott elvek alapján részben automatikus, részben kézi átalakítás segítségével létrehozzuk a Szeged Treebank egy újabb változatát.

Kulcsszavak: szintaxis, dependencia, morfológia

1. Bevezetés

A szófaji egyértelműsítés és a szintaktikai elemzés napjainkban is a számítógépes nyelvészet leginkább kutatott területei közé sorolható. A téma népszerűségét mutatja, hogy az utóbbi években több versenyt is hirdettek, ahol számos nyelv szövegeinek morfológiai, illetve szintaktikai elemzése volt a feladat [1,2]. A különféle nyelvű szövegeken elért eredmények összevetése azonban nehézségekbe ütközik, hiszen a különböző nyelvű adatbázisok eltérő címkekészleteket használnak, illetve más annotációs elvek alapján készültek. Ezen problémák áthidalását célozza az Univerzális Dependencia és Morfológia (UD) című, nemzetközi együttműködésben megvalósuló projekt [3].

Az UD projekt fő célja, hogy egy „univerzális", azaz nyelvfüggetlen szintaktikai és morfológiai reprezentációt dolgozzon ki, mely számítógépes nyelvészeti oldalról elősegíti a többnyelvű morfológiai és szintaktikai elemzők fejlesztését, továbbá elméleti nyelvészeti oldalról megkönnyíti a nyelvtipológiai és kontrasztív nyelvészeti vizsgálatok elvégzését. E cikkben az UD elveinek magyarra való alkalmazását mutatjuk be, különös figyelmet fordítva a speciálisan magyar nyelvi jelenségekre. Ehhez kiindulópontként a Szeged Korpusz és Treebank 2.5-ös verzióját [4] használtuk.

2. Az UD projekt

Az Univerzális Dependencia projekt célja, hogy számos nyelven ugyanazokra az annotációs elvek alapján hozzanak létre morfológiai és szintaktikai korpuszokat,

ugyanazokat az annotációs kódkészleteket használva. Ehhez hasonló egységesítési törekvések már korábban is megfigyelhetők voltak a számítógépes nyelvészetben. Például a Stanfordban kialakított függőségi címkekészletet [5] több nyelv reprezentációjában is hasznosítják. A morfológia terén az MSD kódrendszert közép- és kelet-európai nyelvekre alakították ki, többek között magyarra is [6]. Az Interset kódkészlet egyfajta közvetítő nyelvként szolgál különféle kódkészletek között, a rá épülő konverziós eljárások lehetővé teszik a kódkészletek azonos morfológiai reprezentációra történő átalakítását [7]. Rambow és munkatársai [8] a szófaji egyértelműsítést és szintaktikai elemzést szem előtt tartva megalkottak egy több nyelvre is alkalmazható morfológiai kódkészletet, míg a CoNLL-2007 verseny [9] adatai alapján McDonald és munkatársai [10] 8 fő univerzális szófajt azonosítottak. A későbbiekben Petrov és munkatársai [11] 12 fő univerzális szófajt alkalmaztak 22 nyelvre.

Az univerzális és többnyelvű morfológiai és szintaktikai kódkészletekre való törekvés legújabban az Univerzális Dependencia projektben jelenik meg. A 2015 novemberében publikált 1.2-es verzióban összesen 33 nyelv annotált adatbázisait találhatjuk meg, melyen között az angol, német, francia ugyanúgy szerepel, mint a magyar vagy a koreai.

3. Univerzális morfológia a magyarban

Az alábbiakban röviden bemutatjuk az univerzális morfológiai kódkészlet legfontosabb jellemzőit. A morfológai információt szófaji kód és jegy–érték párok formájában tároljuk. A szófajok és a jegyek halmaza kötött, azaz nincs lehetőség újabbak felvételére, ezzel szemben az értékek között szerepelhet nyelvfüggő érték, amennyiben szükséges. A jegyek között lexikai és inflexiós jegyeket egyaránt találunk: a lexikai jegyek magukra a lemmákra jellemző tulajdonságokat kódolnak, míg az inflexiós jegyek a szóalakot írják le. A jegyek lehetnek hierarchikusak is: a magyarban például a szám jegy többszörösen is megjelenhet a főnéven, így a főnév számát és a főnév birtokosának a számát két külön hierarchikus jeggyel írjuk le.

Az univerzális morfológia magyarosításakor a Szeged Korpusz 2.5-ben is használt morfológiai jellemzők nagy részét automatikusan át tudtuk konvertálni UD formátumra, ugyanakkor néhány megoldandó problémával is szembesültünk. A legnagyobb nehézséget a birtokjelölés jelentette, ugyanis a projekt addigi nyelveiben a birtokos jelölése elsődlegesen determináns segítségével valósult meg, így a magyar birtokos és birtokjel szám- és személyjelölésére külön morfológiai jellemzőket kellett felvennünk. Így a *házaiménak* szó morfológiai elemzése az alábbi lesz, ahol a Number a főnév számát, a Number[psor] és Person[psor] jegyek a birtokos számát és személyét, a Number[psed] pedig a birtok számát jelöli:

NOUN
Case=Dat|Number=Plur|Number[psed]=Sing|Number[psor]=Sing|Person[psor]=1

További sajátosságot jelentett például a determinánsok és a sorszámnevek kezelése. A sorszámnevek a Szeged Korpusz hagyományai szerint számnévként kódolandók, az univerzális morfológiában azonban melléknévként kell jelölni őket.

Ezek átkonvertálása automatikus eszközökkel történt. A mutató névmások kezelése szintén eltérő a Szeged Korpusz eredeti annotációjában és az univerzális morfológiában. Míg az *ez/az* névmások pozíciótól függetlenül névmásként kódolandók az eredeti korpuszban, addig az univerzális morfológiában a névelő előtti használat (*Olvastam azt a könyvet*) determináns, míg az önálló NP-értékű használat (*Olvastam azt*) névmás címkét követel meg. Ezek átkódolása szintén automatikus úton valósult meg.

Szintén magyar sajátosságnak számított az ige tárgyi egyeztetése, azaz a külön igei paradigma használata határozott és határozatlan tárgy mellett. Így a Definiteness jegy a magyarban az igére is alkalmazandó lett, továbbá a második személyű tárgy által megkövetelt *-lAk* morféma miatt új értéket is fel kellett venni a meglévő *határozott* és *határozatlan* érték mellé.

Az univerzális morfológia nem tartalmaz külön igekötő szófajt, az igei partikulák és igekötők szófaji besorolását az eredeti szófaj szerint végzi. Ennek megfelelően a magyarban is összeállítottunk egy táblázatot, melyben feltüntettük az igekötőként kezelt nyelvi elemek eredeti szófaját (pl. az *el* igekötőt határozószóként vettük fel, az *agyon*-t pedig főnévként). Ezt követően a táblázat alapján automatikusan átcímkéztük az igekötőket.

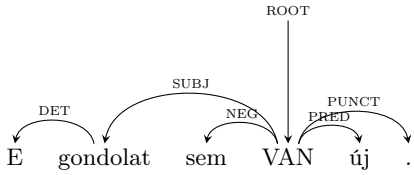
4. Univerzális dependencia a magyarban

Az univerzális dependencia címkekészletének magyarosításakor szintén automatikusan tudtuk konvertálni a függőségi viszonyok többségét, időnként a szintén rendelkezésre álló konstituens-, illetve koreferenciaannotációt is figyelembe véve. Néhány problémás jelenséget azonban részletesebben is bemutatunk.

Klasszikusan a függőségi elemzésekben a mondatok feje a főmondat ragozott igéje, viszont a kopulát és névszói predikátumot tartalmazó mondatok, és főként a fonológiailag üres kopulát tartalmazók esetén problémába ütközik ez a leírás. Mel'čuk [12] szerint azokban az esetekben, amikor a kopula csak bizonyos szám, személy, idő esetén üres (mint a magyarban), feltételeznünk kell egy zéró igealakot. A szerkezet feje ez a zéró igealak, ehhez kapcsolódik a névszói predikátum. Ezt az úgynevezett funkciószó fej elemzést követi a Szeged Dependencia Treebank is, melyet az 1. ábrán láthatunk.

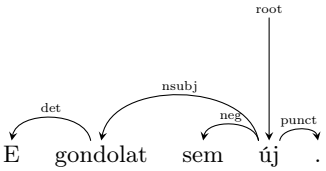
A tartalmas fej elemzés abban tér el a funkciószó fej változattól, hogy a funkciószavak helyett a tartalmas szavakat preferálja fejként. Az elmélet szerint a mondat vázát a benne szereplő tartalmas szavak és közöttük lévő kapcsolatok fejezik ki. A funkciószavak ehhez a vázhoz kapcsolódnak. A funkciószó fej elemzéstől a kopulás és adpozíciós szerkezetek kezelésében tér el. Mindkét esetben a tartalmas szó (névszói predikátum vagy az adpozíció névszói vonzata) lesz a fej a funkciószó (kopula vagy adpozíció) helyett.

Az univerzális elvek szerint a tartalmas fej elemzést kell követni. Ezen okokból automatikusan átalakítottuk a névszói és névszói-igei predikátumot tartalmazó mondatokat: mindegyik esetben a predikatív címkét viselő szót tüntettük fel fejként (vö. 2. ábra), és amennyiben szerepelt mellette kopula, az `cop` (kopula) címkével kapcsolódott a fejhez. Hasonlóképpen a névutós szerkezetekben



1. ábra: Virtuális kopula a Szeged Dependencia Treebankben.

a főnevet szerepeltettük fejként, a névutó pedig ehhez kapcsolódik `case` (eset) címkével.

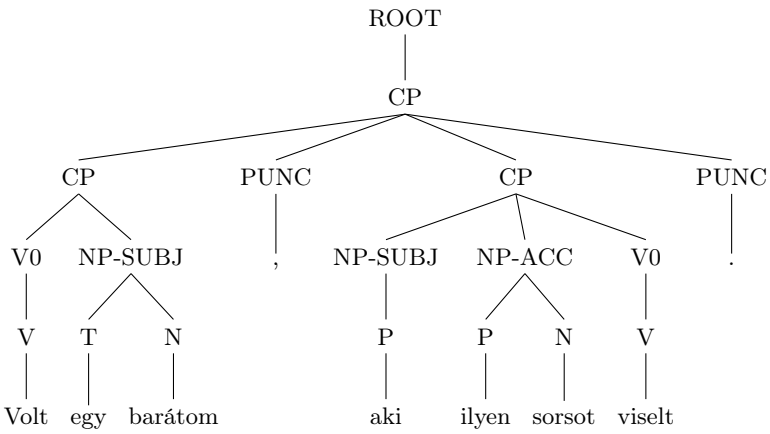


2. ábra: Tartalmas szó mint fej.

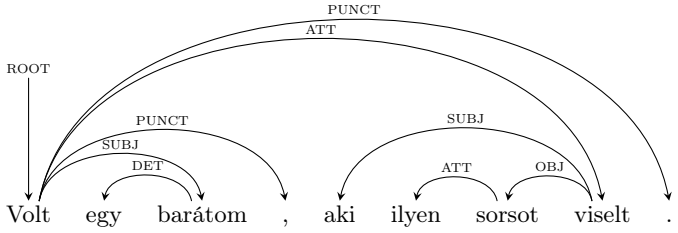
Sajátos problémát jelentett az alárendelő mellékmondatok kezelése, ahol is az univerzális elvek – a Szeged Treebanktől eltérően – megkülönböztetik az alárendelő mondatok több fajtáját is, így például az alanyi, tárgyi és határozói alárendelést, illetve a vonatkozó mellékmondatokat. Ezzel szemben a Szeged Dependencia Treebank egységesen alárendelő mellékmondatként jelölte ezeket, az altípusokat meg nem különböztetve. A Szeged Treebank konstituens változatában ezek egy része megkülönböztető jelöléssel rendelkezett, így azokat át tudtuk onnan emelni, más esetekben pedig nyelvészeti szabályok segítségével valósult meg az átalakítás, azonban az esetek egy részében az automatikus konverziót kézzel kellett javítanunk. Az alábbi példán látjuk, hogy a vonatkozó mellékmondat ATT címkét visel az eredeti dependencia treebankben (4. ábra), és a főmondat igéjéhez van kötve, a konstituens treebankben pedig nincs jelölve a CP szerepe (3. ábra).

A koreferenciaviszonyokból azonban láthatjuk (5. ábra), hogy az *aki* névmás voltaképpen az *egy barátom* szókapcsolatra vonatkozik, így vonatkozó mellékmondatról van szó, melyet a *barátom* szóhoz kell csatlakoztatni. Ezek után a szükséges átalakítások segítségével átkonvertálhatjuk a mondatot az univerzális dependencia elveinek megfelelően (6. ábra).

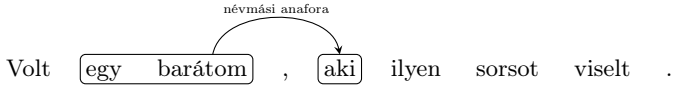
A többtagú tulajdonnevek esetében az univerzális elvek szerint az első tagot kellene fejnek jelölni. A magyarban azonban morfoszintaktikai okok miatt az utolsó tagot tekintettük fejnek, hiszen az ragozódik (*Kovács Jánosnak* vs. **Kovácsnak János*). Így ebben az esetben eltértünk az univerzális elvektől, és az utolsó tagot jelöltük fejként, míg a többi treebank esetén az első elem szerepel a szerkezet fejeként.



3. ábra: Konstituenselemzés a Szeged Treebankben.



4. ábra: Függőségi elemzés a Szeged Dependencia Treebankben.

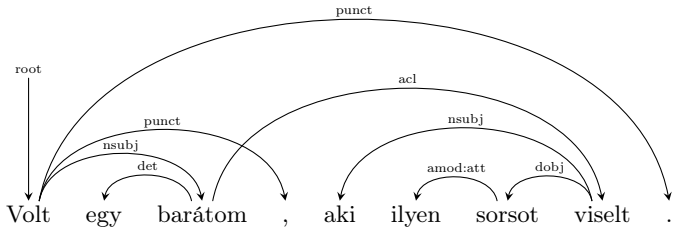


5. ábra: Koreferenciaviszonyok a SzegedKorefben.

A dativus kezelése szintén problematikusnak bizonyult. A magyarban a *-nAk* ragos főnevek számos különböző szerepet tölthetnek be a mondatban, például:

- részeshatározó: *Laci adott a padtársának egy almát.*
- birtokos: *Laci elvette a padtársának a könyvét.*
- dativus ethicus: *Nekem nehogy eladd az autódat!*
- experiens: *Nekem nagyon tetszett az előadás.*
- szemantikai alany: *Lacinak bocsánatot kellett kérnie a padtársától.*

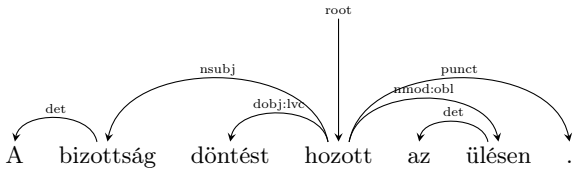
Míg morfológiai szinten a fenti alakok teljesen egybeesnek, addig szintaktikai és szemantikai szinten különféle szerepeket jelölnek. Így amellett döntöttünk, hogy míg a fenti példák morfológiai annotációját egyformán jelöljük, addig



6. ábra: Függőségi elemzés az Univerzális Dependencia Treebankben.

szintaktikai szinten elkülönítjük őket. A részeshatározót `iobj` (indirect object) címkével látjuk el, a birtokost `nmod:att` címkével (főnévi módosító), az egyéb előfordulásokat pedig `nmod:obl` címkével (főnévi vonzat) láttuk el. Természetesen ezen átalakítások kézi annotációt igényeltek, hiszen pusztán morfológiára és szintaxisra hagyatkozva ezeknek az eseteknek a nagy részében nem tudtuk volna egyszerűen és egyértelműen megvalósítani az automatikus konverziót (vö. *Nekem nehogy eladd az autódat!* és *Nehogy eladd nekem az autódat!*, ahol az első mondatban dativus ethicust találunk, a másodikban pedig részeshatározót, a két esetet egyedül a szórend különbözteti meg).

Az úgynevezett félig kompozicionális szerkezetek kezelése az 1.2-es verzióban egyelőre nem egységes a különböző nyelvek között: az UD treebankek vagy nem jelölik a szerkezeteket, vagy ha pedig jelölik, akkor ezt vagy a szokványos vonzatjelöléstől eltérő szerkezettel és/vagy speciális címkézéssel teszik [13]. A magyar az utóbbi csoportba tartozik, azaz speciális címkékkel látja el a félig kompozicionális szerkezetek tagjait. Például a 7. ábrán a `dobj:lvc` címke köti össze a szerkezet főnévi és igei tagját, azaz a *döntést* és a *hoz* szavakat, jelölve ezáltal, hogy szintaktikai értelemben ige–tárgy kapcsolatról van szó, azonban szemantikai értelemben sajátos a két összetevő viszonya. Az UD projekt legújabb egységesítési törekvései szerint ezt az elemzést terjesztjük ki a későbbiekben a többi UD treebankre is.



7. ábra: Félig kompozicionális szerkezet az Univerzális Dependencia Treebankben.

5. Szeged Univerzális Treebank

A fenti elveknek megfelelően elkészítettük a Szeged Treebank univerzális morfológiára konvertált változatát. Emellett elkészült a Népszava-alkorpusz univerzális dependenciára konvertált változata is, és folyamatosan dolgozunk a további alkorpuszok dependenciakonverzióján is: az automatikus konverzió már lezajlott, a kézi ellenőrzést igénylő annotációs lépések pedig folyamatban vannak. Az elkészült adatbázisok elérhetők az UD projekt honlapján[1].

Köszönetnyilvánítás

Szeretnénk megköszönni az Univerzális Dependencia projekt tagjainak, különösen Joakim Nivrének, Chris Manningnek és Daniel Zemannak a magyar UD treebank annotálási elveinek kialakításában nyújtott önzetlen segítségüket.

Hivatkozások

1. Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J.D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Marton, Y., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A.: Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA, ACL (2013) 146–182
2. Seddah, D., Kübler, S., Tsarfaty, R.: Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, Dublin, Ireland, Dublin City University (2014) 103–109
3. Nivre, J.: Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer (2015) 3–16
4. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC 2014, Reykjavik, Iceland, ELRA (2014) 1074–1078 ACL Anthology Identifier: L14-1241.
5. de Marneffe, M.C., Manning, C.D.: Stanford dependencies manual. Technical report, Stanford University (2008)
6. Erjavec, T.: MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation **46**(1) (2012) 131–142
7. Zeman, D.: Reusable tagset conversion using tagset drivers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) http://www.lrec-conf.org/proceedings/lrec2008/.

[1] http://universaldependencies.github.io/docs/

8. Rambow, O., Dorr, B., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K.J., Mitamura, T., Reeder, Florence, Siddharthan, A.: Parallel syntactic annotation of multiple languages. In: Proceedings of LREC. (2006)
9. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. (2007) 915–932
10. McDonald, R., Nivre, J.: Characterizing the errors of data-driven dependency parsing models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 122–131
11. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of LREC. (2012)
12. Polguère, A., Mel'čuk, I.A., eds.: Dependency in Linguistic Description. Studies in language companion series. Amsterdam Philadelphia, Pa. J. Benjamins (2009)
13. Nivre, J., Vincze, V.: Light verb constructions in universal dependencies (2015) Poszter. PARSEME 5th General Meeting.