Morfológiai újítások a Szeged Korpusz 2.5-ben

Vincze Veronika[1,2], Varga Viktor[2], Simkó Katalin Ilona[2],
Zsibrita János[2], Nagy Ágoston[2], Farkas Richárd[2]

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

Kivonat: A Szeged Korpusz a legnagyobb, kézzel annotált adatbázis, amely a szóalakok lehetséges morfológiai kódjait és lemmáit is tartalmazza. Ebben a munkában bemutatjuk a korpusz újabb változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot.

1 Bevezetés

A Szeged Korpusz a legnagyobb, kézzel annotált magyar adatbázis, melyben a szavak lehetséges és a szövegkörnyezetnek megfelelő morfológiai kódjai, illetve a szavak lemmái kézzel be vannak jelölve [1]. A korpusz 2.0 verziójában található morfológiai kódok az MSD kódrendszernek felelnek meg [2]. Ebben a munkában bemutatjuk a korpusz újabb változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is kézzel hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot.

2 Harmonizált morfológiai kódok

Egy korábbi munkánkban már lefektettük a KR [3] és MSD [2] kódrendszerek harmonizálásának alapelveit [4]: a harmonizálás során arra törekedtünk, hogy az új morfológiai kódoknak olyan (és csak olyan) információkat kell tartalmazniuk, amelyek a későbbi feldolgozás (szintaxis, különféle alkalmazások) szempontjából hasznosak.

A 2.5 verzióban így a korábbi 2.0-s verzióhoz képest az alábbi morfológiai újítások találhatók:

- a gyakorító, ható és műveltető igék lemmája a képző nélküli igető lett, és a kódban jelöljük azt, hogy az ige milyen alakban áll;
- a melléknévi igenevek önálló kódot kaptak (korábban a melléknevek és az igenevek nem voltak elkülöníthetők MSD-kóduk alapján);
- tulajdonnév és köznév elkülönítésének megszüntetése;
- a személyes névmási határozószóknak a névmási rendszerbe való beillesztése.

A fenti esetekben az egyes szóalakok mellé felvettük az új morfológiai kódokat, valamint szófajilag is egyértelműsítettük a szövegeket, azaz manuálisan kiválasztottuk, hogy melyik lehetséges kód illik az adott szövegkörnyezetbe. Az alábbiakban részleteiben is ismertetjük az egyes morfológiai újításokat.

2.1 Gyakorító, ható és műveltető igék

A KR kódrendszer a gyakorító és műveltető igéket (pl. *olvasgat*, *futtat*) az alapalakból képzett igének tekinti, tehát a gyakorító és műveltető szuffixumokat képzőként kezeli. A ható igék (*mehet*) toldaléka ezzel szemben inflexiós toldaléknak számít a KR rendszerében. Az MSD kódrendszer eredetileg mindezen toldalékokat a lemma részeként kezelte, azaz míg például az *olvastak* és *olvashattak* morfológiai kódja azonos volt (Vmis3p---n), addig lemmájuk eltért: *olvas* és *olvashat*. A harmonizációnak köszönhetően a Szeged Korpuszban is jelöljük azt, hogy az ige gyakorító, műveltető vagy pedig ható-e. Az igei MSD-kód második pozíciójában jelenítjük meg ezeket az információkat, lemmának pedig az ige toldalékolatlan alakját tüntetjük fel. Arra is figyelmet fordítottunk, hogy ezen toldalékok nem zárják ki egymást, tehát egy adott igealak lehet egyszerre például műveltető és ható is. Így a toldalékok lehetséges kombinációját is meg tudjuk jeleníteni a harmonizált kódrendszerben. Az alábbi táblázat mutatja be a harmonizált kódokat:

1. táblázat: Igei harmonizált kódok.

Leírás	**Kód**	**Toldalék**	**Példa**
fő (main)	m	-	*megy*
segéd (auxiliary)	a	-	*fogok (menni)*
ható (modal)	o	-hAt	*mehetek*
gyakorító (frequentative)	f	-gAt	*pofozgat*
műveltető (causative)	s	-(t)At	*etet*
gyakorító+ható	1	-gAthAt	*boncolgathat*
műveltető+ható	2	-(t)AthAt	*fektethet*
műveltető+gyakorító	3	-(t)AtgAt	*etetget*
műveltető+gyakorító+ható	4	-(t)AtgAthAt	*futtatgathat*

Az igék újrakódolásakor különös figyelmet fordítottunk a kétértelmű esetekre, amikor ugyanaz az igealak jeleníti meg a műveltető és nem műveltető alakot. Ez el-

sődlegesen a múlt idejű igealakoknál fordult elő, amikor például a *festetted* alak jelölheti a *fest* és a *festet* múlt idejű E/2. tárgyas ragozású alakját is, kontextustól függően.

2.2 Melléknévi igenevek

Míg a KR kódrendszer a melléknevektől elkülönítve kezelte a melléknévi igeneveket, addig az MSD-ben az A szófaji kód vonatkozott a melléknevekre és a melléknévi igenevekre egyaránt. Azonban a melléknevek és a melléknévi igenevek morfológiai és szintaktikai viselkedése eltérő vonásokat mutat: a melléknevek fokozhatók, míg a melléknévi igenevek nem, vö. *az okos fiú – az okosabb fiú* és *az éneklő fiú - *az énekelőbb fiú*, továbbá a melléknévi igenév igen gyakran megőrzi az eredeti ige vonzatszerkezetét: *a slágert jó hangosan éneklő fiú*. Mivel úgy gondoljuk, hogy e különbségek kihatással vannak a mondatok szintaktikai elemzésére is, a harmonizált kódrendszerben is bevezettük e megkülönböztetést. A melléknévi MSD-kód második pozíciójában jelenítjük meg azt az információt, hogy melléknévről vagy melléknévi igenévről van-e szó, illetve utóbbi esetben megadjuk a melléknévi igenév típusát is (folyamatos, befejezett vagy beálló). A kódokat az alábbi táblázat részletezi:

2. táblázat: Melléknévi (igenévi) harmonizált kódok.

Leírás	**Kód**	**Képző**	**Példa**
melléknév	f	-	*friss*
folyamatos melléknévi igenév	p	-Ó	*sétáló*
befejezett melléknévi igenév	s	-t/-tt	*megvásárolt*
beálló melléknévi igenév	u	-AndÓ	*felveendő*

Bizonyos szóalakok mind melléknévként, mind melléknévi igenévként használatosak, vö. *égető kérdések – a kertben tüzet égető gondnok*. Az egyértelműsítés során is a fenti különbségeket (fokozás, vonzatok) használtuk nyelvi tesztként.

2.3 Köznevek és tulajdonnevek

Az MSD kódrendszer korábbi verziójában a köznevek és tulajdonnevek külön kóddal rendelkeztek. Azonban úgy gondoljuk, hogy a köznév-tulajdonnév elkülönítés nem bír jelentőséggel a morfológia szintjén, így egy morfológiai elemzőnek nem is lehet feladata a tulajdonnevek felismerése, meghagyva az a névelem-felismerő alkalmazásoknak. Mindezekből kifolyólag a Szeged Korpusz 2.5-ös változatában eltöröltük a köznév-tulajdonnév megkülönböztetést, így minden főnévi kód egységesen Nn- kezdettel rendelkezik.

2.4 Személyes névmási határozószók

A magyar nyelvben a hagyományos terminológiával személyes névmási határozószóknak hívott szóalakok két csoportra bonthatók. Az első csoportot azok alkotják, amelyek etimológiájukat tekintve határozóragra vezethetők vissza (*bennem*, *neki*). A második csoportba azok tartoznak, amelyek névutóból eredeztethetők (*szerinted*, *mögöttünk*). Az eredeti MSD-rendszerben e szóalakok egységesen a határozószavak egy alosztályát képezték, míg a KR rendszerében mindkét csoport főnévként szerepeltek (bár a morfológiai kód felépítése eltért a két esetben).

A harmonizált kódrendszerben egyik megoldást sem vettük át, hanem névmásként kezeljük ezeket az alakokat, a személyes névmási rendszerbe illesztve. A névutóból eredeztethető alakok esetében lemmaként a névutót tüntetjük fel, a határozóragból eredeztethető alakoknál pedig a személyes névmást. Néhány példát mutatunk az alábbiakban:

3. táblázat: Névmási harmonizált kódok.

Szóalak	Lemma	Morfológiai kód
szerintem	szerint	Pp1-sn
nálunk	mi	Pp1-p3

Ezek az alakok automatikusan lettek átcímkézve, esetükben nem volt szükség kézi egyértelműsítésre.

2.5 Írásjelek

Az írásjelek morfológiai kódolásán szintén változtattunk. Az alábbi 8 írásjelet tekintjük relevánsaknak (az írásjelek mögött az ASCII kódjuk szerepel): !(33) ,(44) -(45) .(46) :(58) ;(59) ?(63) –(8211).

A releváns írásjelek lemmája maga az írásjel lesz, morfológiai kódja szintén. Egyéb nem releváns írásjelek (olyan karaktersorozatok, melyek nem tartalmaznak sem betűt, sem számot) lemmája szintén maga az írásjel lesz, de kódja K (központozás) lesz.

2.6 Elváló igekötők

Az elváló igekötőt tartalmazó igei elemek (igék, főnévi, melléknévi és határozói igenevek) lemmájában megjelöltük az igekötő-igei elem közti morfémahatárt. Mivel bizonyos szintaktikai műveletek hatására az ige és igekötő elválhat egymástól, úgy döntöttünk, hogy ezekben az esetekben jelöljük a morfémahatárt a lemmában.

3 Helyesírási hibák javítása

A morfológiai javítások mellett figyelmet fordítottunk a helyesírási hibák javítására is. A korpusz 2.0 változatában külön MSD-kóddal rendelkeztek a rossz helyesírású

(elírt, elgépelt) szavak (pl. *kiráj*), illetve azok, melyek értelmes magyar szavak, azonban a szövegkörnyezetbe nem illettek bele (*mer úgy gondolom* vs. *mert úgy gondolom*). Amennyiben a helyes és az elírt alak azonos tokenszámú egységet tartalmazott, úgy a helyesírási hibát vagy elírást tartalmazó szóalakok mellé felvettük azok helyes alakját is annak lehetséges MSD-kódjaival együtt, majd a szövegkörnyezetnek megfelelően kiválasztottuk az aktuális kódot. Azokban az esetekben pedig, ahol a helyes és helytelen alakok tokenszáma között eltérés mutatkozott (pl. *areggel* vs. *a reggel*), a fő szóalak morfológiai kódját vettük fel (pl. egy egybeírt névelő és főnév esetén a főnévi címkét).

4 Statisztikai adatok

A Szeged Korpusz 2.0 verziója 1,2 millió tokent tartalmazott (egy tokennek számítva a többtagú tulajdonneveket). Ezek közül 11 461 token minősült ismeretlen vagy rossz helyesírású szónak. A 2.5-ös verzióban e szavak száma mindösszesen 1563 lett, azaz a morfológiai elemzés számára problematikus szavak aránya 1%-ról 0,13%-ra csökkent, ami jelentős – egy nagyságrendnyi – változást jelent: a problémás szavak 86,4%-át sikerült kijavítani.

A korpusz jelen változatában az ismeretlen szavak legnagyobb része angol számítástechnikai terminus. Ez arra vezethető vissza, hogy a számítógépes szövegek alkorpuszban gyakran szerepelnek az eredeti angol megnevezések is a felhasználói kézikönyvek szövegeiben.

A korpusz 2.5 változatában összesen 1315 morfológiai kód szerepel. Az alábbi táblázat mutatja be az újonnan bevezetett kódok előfordulásait:

4. táblázat: Új kódok gyakorisága

Leírás	**Kód**	**Előfordulás**
Folyamatos melléknévi igenév	Ap*	23483
Befejezett melléknévi igenév	As*	12588
Beálló melléknévi igenév	Au*	520
Melléknévi igenév összesen	Ap*, As*, Au*	36591
Műveltető ige	Vs*	1698
Ható ige	Vo*	8415
Gyakorító ige	Vf*	327
Műveltető/ható/gyakorító kombinációja	V1*, V2*, V3*, V4*	67
Műveltető/ható/gyakorító igék összesen	Vs*, Vo*, Vf*, V1*, V2*, V3*, V4*	10057

A személyes névmási határozószók újrakódolása további 8232 tokent érintett. Ha összegezzük tehát a megváltozott kódú szavakat (melléknévi igenevek, műveltető/ható/gyakorító igék, személyes névmási határozószók, javított helyesírási hibák), akkor összesen 64 788 szóalak kódja változott meg, ami a korpusz szavainak 4,36%-a.

5 Morfológiai elemző

A Szeged Korpusz 2.5 változata lehetővé tette, hogy a *magyarlanc* nevű adatvezérelt nyelvi elemző [5] morfológiai és szófaji egyértelműsítő moduljait az új adatbázison tanítsuk be, létrehozva ezzel az elemző újabb változatát, mely a morfológiai elemzés és szófaji egyértelműsítés végeredményeként az új harmonizált morfológiának megfelelő kódokat ad vissza.

A korpusz teljes állományát véletlenszerűen osztottuk fel tanító és kiértékelő adatbázisra 80:20 arányban, majd a tanítást követően értékeltük a szófaji egyértelműsítő teljesítményét. Akkor fogadtuk el helyesnek a *magyarlanc* által adott elemzést, ha mind a lemma, mind pedig a morfológiai kód egyezett az etalon korpuszban lévővel. Eredményeink szerint a *magyarlanc* szófaji egyértelműsítő modulja az új kódrendszer használatával 96,32%-os pontosságot ér el, ami megegyezik a korábban publikált, Szeged Korpusz 2.0 verzión tanított rendszer eredményességével [5], vagyis az elemzés minőségét nem befolyásolja érdemben a megnövekedett kódhalmaz.

6 Összegzés

Ebben a munkában bemutattuk a Szeged Korpusz 2.5 változatát, amelyben az új, harmonizált KR-MSD kódrendszernek megfelelő morfológiai kódok találhatók, illetve a rossz helyesírású szavak nagy részéhez is hozzárendeltük a szándékolt szónak megfelelő morfológiai kódot. A korpusz lehetővé tette azt is, hogy a magyarlanc morfológiai elemző és szófaji egyértelműsítő modulját az új szófaji kódokra tanítsuk be. Eredményeink alapján a szófaji egyértelműsítés minősége változatlanul magas a megnövekedett kódhalmaz ellenére is.

A korpusz kutatási és oktatási célokra szabadon hozzáférhető a http://www.inf.u-szeged.hu/rgai/SzegedTreebank oldalon.

Köszönetnyilvánítás

A kutatás – részben – 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. 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. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf

3. Kornai, A., Rebrus, P., Vajda, P., Halácsy, P., Rung, A., Trón, V.: Általános célú morfológiai elemző kimeneti formalizmusa. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2004) 172–176
4. Farkas, R., Szeredi, D., Varga, D., Vincze, V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (2010) 349–353
5. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013)