Egy általános célú morfológiai annotáció kiterjesztése

Recski Gábor

MTA SZTAKI, Nyelvtechnológiai Kutatócsoport
recski@sztaki.hu

Kivonat: Egy szó nyelvtani jegyeinek kódolására számos különböző annotációs formalizmus létezik. Az ezek közötti veszteségmentes konverzió jelentősen megkönnyíti a különböző nyelvfeldolgozó eszközök és más nyelvi erőforrások közötti együttműködést. Az angol nyelvre elterjedt BNC kódolást, valamint a magyarra kidolgozott MSD-kódrendszert a KR-formalizmus címkekészletére képezzük le úgy, hogy szükség esetén utóbbit bővítjük az információveszteség elkerülése érdekében. Mivel a morfológiai elemzés véges állapotú transzdúcerekkel elvégezhető, a konverziók pedig maguk is transzdukciók, ezért a leírt átalakítások segítségével könnyedén előállítható több típusú annotációt kibocsátani képes elemző is.

1 Bevezetés

Jelen cikkben bemutatjuk, miként valósítható meg több, széles körben elterjedt morfológiai annotációs formalizmus konverziók általi egységesítése, és miként válik lehetővé ezáltal számos nyelvi erőforrás együttes használata. Ehhez a magyar nyelv feldolgozásában kulcsszerepet játszó KR-formalizmus (Rebrus et al. 2012) címkekészletét bővítjük ki úgy, hogy a más konvenciók szerint készült elemzések veszteségmentesen konvertálhatóak legyenek KR-kódokra. A 2. fejezetben bemutatjuk a konverzióban érintett kódolási rendszereket (KR, MSD, BNC). A 3. fejezetben bemutatjuk a konverzió során egyedi megfontolást igénylő esetek kezelését. Végül a 4. fejezetben felvázoljuk egy mindhárom kódrendszer szerint elemezni és generálni képes véges állapotú morfológiai eszköz működését. Jelen cikk kizárólag a KR-kód lehetséges változtatásaival foglalkozik, azonban az MSD és KR rendszerek közötti – mindkét formalizmust érintő – harmonizáció is folyamatban lévő munka (Farkas et al. 2010).

2 Annotációs rendszerek

2.1 A KR-kód

Egy szó morfológiai jegyeinek reprezentációjára számos különböző konvenció létezik. A `hun*` nyelvtechnológiai eszközök az ún. KR-formalizmust követik, így ezt a reprezentációt támogatja többek között a `hunpos` morfológiai címkéző (Halácsy et al. 2007), a `hunmorph` morfológiai elemző (Trón et al. 2005) és az egyebek közt

Alknyelvdok7
Szerk.: Váradi Tamás
MTA Nyelvtudományi Intézet, Budapest, 2013
ISBN 978-963-9074-59-0

tulajdonnévfelismerést és mondatrész-azonosítást végző `huntag` szekvenciális címkéző (Recski–Varga 2012) is. A KR-konvenciót alkalmazó `morphdb.hu` morfológiai adatbázist (Trón et al. 2006) használja továbbá a `magyarlanc` nyelvtechnológiai eszközkészlet is (Zsibrita et al. 2013).

A KR-formalizmus előnye, hogy egy szó valamennyi morfológiai jegyét külön karaktersorozatnak felelteti meg, így a KR-kódok jelentése kompozicionális (például az *asztalát* szó a `NOUN<POSS><CAS<ACC>>` kódot kapja). Ugyancsak a más rendszerekkel való együttműködést könnyíti meg, hogy a KR-reprezentáció könnyedén feleltethető meg attribútum-érték mátrixoknak (AVM), így a fenti reprezentációval ekvivalens a `[CAT=NOUN, POSS=1, CAS=ACC]` struktúra. Ez teszi lehetővé például, hogy a KR-formalizmus közvetlen bemenetként szolgáljon a népszerű NLTK nyelvtechnológiai eszközkészlet (Bird et al. 2009) számos mondatelemző (parszer) moduljának (ld. Recski 2010).

A KR-kód általában – az annotáció tömörségének és jobb olvashatóságának érdekében – nem tünteti fel egy szó azon morfológiai jegyeit, melyek alapértelmezett (a szótári alakkal egybeeső) értéket vesznek fel; így például nem jelzi, ha egy ige harmadik személyű, ha egy főnév alanyesetű, stb. Olyan alkalmazások számára azonban, melyek megkívánják, hogy egy szó nyelvtani jegyei mindig kitöltöttek legyenek, könnyedén elkészíthető a KR kód azon változata, ahol a szám, személy, eset, stb. jegyek a `SING, 3, NOM`, stb. értékeket is felvehetik.

Végül megemlítjük a KR-kód egy, a címkekészlet későbbi kiterjesztésénél szerephez jutó tulajdonságát is: a KR a legtöbb kódolással ellentétben képes egy szóalak képzésére vonatkozó információk megjelenítésére is. Így például a *felértékelődése* szó a `hunmorph` elemzőtől a `felértékel/VERB[MEDIAL]/VERB[GERUND]/NOUN<POSS>` címkét kapja, melyről rendre leolvasható a *felértékel*, *felértékelődik*, *felértékelődés* szavak kategóriája és az őket előállító képzés módja is. Ezt a szintaxist használja a KR az alkategóriák megjelenítésére is. Így egy sorszámnév például az `[ORDINAL]/NUM` címkét kapja. Ez a párhuzam azt tükrözi, hogy egy morfoszintatkikai kategóriát az őhozzá vezető képzési utak is alkategóriákra bontják.

2.2 Az MSD-kód

Az `MULTEX-EAST` projekt keretében a magyar nyelv morfológiájának kódolására kidolgozott MSD-kód (Erjavec 2004) a KR-kódtól alapvetően eltérő felépítésű. Egy címke a főkategóriát jelző betűből (N, V, A, stb.) és különböző nyelvtani jegyeknek megfelelő további betűk sorozatából áll, melyek egyenkénti értelmezéséhez a főkategóriát és az adott betű pozícióját is figyelembe kell venni. Az *asztalát* szó elemzése az MSD-kódrendszer szerint például `Nc-sa–s3`, melyben az N a főnévi kategóriát jelöli, a további jelekből pedig rendre az derül ki, hogy a szó köznév, egyes számú, tárgyesetű, a birtokos pedig egyes számú és harmadik személyű.

A magyar nyelvre kidolgozott MSD-kódot használja többek között a Szeged Treebank korpusz (Csendes et al. 2005) és az ennek részét képező HunNER korpusz (Simon et al. 2006), melyek számos felügyelt tanuláson alapuló nyelvtechnológiai eszköznek szolgáltatnak tanítóadatot, köztük a `hun*` eszközkészlet részét képező mondatrész-azonosítónak és tulajdonnév-felismerőnek is. Mivel ez utóbbi eszközök elsősorban KR-kódolás szerint annotált szövegek elemzésére vannak felkészítve, ezért a tanulóadat létrehozásához is biztosítani kell az MSD-kódról KR-re való konverziót.

2.3 A BNC-kód

A `hun*` eszközkészlet valamennyi komponense – megfelelő modell birtokában – alkalmas angol nyelvű bemenet kezelésére is. Ahogy a magyar modellekhez a Szeged Treebank, úgy az angol nyelvű modellekhez leggyakrabban a Penn Treebank (Marcus et al. 1994) nyújt megfelelő tanulóadatot, mely a morfológiai annotációhoz a British National Corpus (Burnard 1995) által is használt, és emiatt gyakran *BNC*-nek nevezett címkekészletet használja, mely angol nyelvű szövegek szófaji címkézésében gyakorlatilag egyeduralkodóvá vált. Ugyan a `hun*` eszközök képesek bármilyen formátumú adatból modellt építeni, a `hunmorph` elemző azonban angol nyelvű szövegre is KR-formátumú elemzést bocsát ki, így a megfelelő modellek előállításához a BNC és KR készletek közötti leképezést is meg kívánjuk valósítani. Elterjedtsége miatt a BNC címkekészletnek számtalan változata van használatban, a jelen munka során mi a Penn Treebank-ben használatos dialektust tekintjük irányadónak.

A BNC címkekészlet a másik két bemutatott rendszertől jelentősen eltér abból a szempontból, hogy nem az általános morfológiai leírás a célja, hanem csupán az angol nyelv szavainak osztályozása. Így külön címkét tart fenn olyan, az angolban gyakran használt nyelvi elemeknek, mint az ún. *existential there* (pl. ***There** is no asbestos in our products now*) vagy a *'s* birtokos klitikum (pl. *A Fed spokesman denied Mr LaFalce**'s** statement.*), ugyanakkor csak az angol nyelvben létező distinkciók megtételére képes.

3 Konverzió

Konverzió alatt két címkekészlet elemei közti leképezést értjük. Amennyiben két annotációs séma pontosan ugyanazokat a különbségeket tenné meg, tehát a nyelv bármely két szavát vagy mindkét séma egyforma címkével látná el, vagy mindkettő megkülönböztetné, akkor ez a leképezés *kölcsönösen egyértelmű* lenne, azaz mindkét irányba *veszteségmentesen* lenne elvégezhető. Jelen munka célja, hogy a konverzió a KR-kód irányába veszteségmentessé váljon, tehát ahol az MSD vagy BNC rendszerek finomabb felosztással élnek, ott a KR-kódot kibővítsük. Ebben a fejezetben az ilyen bővítéseket, kiterjesztéseket mutatjuk be.

3.1 Névmások

A KR és MSD kódok közti egyik legalapvetőbb különbség, hogy utóbbi külön főkategóriába sorolja a névmásokat, míg a KR-kód ilyen kategóriát nem tart fenn, az egyes névmáscsoportokat disztribúciójuk alapján sorolja valamely szófajhoz. Így tehát az MSD kódolás a névmás főkategória alá sorol olyan szavakat, melyek a KR-reprezentáció szerint főnevek (pl. *aki, valaki, bárki*), melléknevek (*milyen, effajta, afféle*), határozók (*mennyire, ahányszor, valamennyien*), számnevek (*ennyi, néhány, valamennyi*), vagy névelők (*amazon, mindezen, ugyanazon*). Az MSD a névmások kódjában feltünteti azok típusát – megkülönböztet mutató, kérdő, személyes, birtokos, vonatkozó, kölcsönös, visszaható, határozatlan és általános névmásokat.

Az MSD által megjelenített különbségeket úgy őrizzük meg, hogy a KR-ben érintett főkategóriáknak új alkategóriájaként (vö. 2.1. alfejezet) vesszük fel a névmás (`PRON`) címkét, mely pedig jegyként veheti fel az MSD-ben definiált névmástípusokat.

Egyes esetekben az MSD-kódból még nem egyértelműen következik a KR-főkategória – ld. például a mutatónévmásokat, ahol azonos MSD-kód KR-ben négyféle szófajnak is megfelelhet) – ilyenkor a konverzió a szóalakot is figyelembeveszi, a konverziós táblák tehát nem csupán (`MSD, KR`) címkepárokat, hanem (`szóalak, MSD, KR`) hármasokat is fognak tartalmazni. Az 1. táblázat a konverziót valamennyi típus esetében példával is illusztrálja.

	MSD	**KR**
ő	Pp3-sn	[PRON<PER>]/NOUN
ez	Pd3-sn	[PRON<DEM>]/NOUN
ezen	Pd3-sn	[PRON<DEM>]/ART
effajta	Pd3-sn	[PRON<DEM>]/ADJ
ennyi	Pd3-sn	[PRON<DEM>]/NUM
egyik	Pi3-sn	[PRON<INDEF>]/NOUN
valamilyen	Pi3-sn	[PRON<INDEF>]/ADJ
valamennyi	Pi3-sn	[PRON<INDEF>]/NUM
saját	Ps3-sn	[PRON<POSS>]/NOUN
mi	Pq3-sn	[PRON<INTER>]/NOUN
milyen	Pq3-sn	[PRON<INTER>]/ADJ
mennyi	Pq3-sn	[PRON<INTER>]/NUM
ami	Pr3-sn	[PRON<REL>]/NOUN
amilyen	Pr3-sn	[PRON<REL>]/ADJ
ahány	Pr3-sn	[PRON<REL>]/NUM
magához	Px3-st	[PRON<REFL>]/NOUN
egymást	Py3-sa	[PRON<RECIP>]/NOUN
semmi	Pg3-sn	[PRON<GEN>]/NOUN
valamennyi	Pg3-sn	[PRON<GEN>]/NUM
mindegyik	Pg3-sn	[PRON<GEN>]/ADJ
mindannyian	Pg3-sn	[PRON<GEN>]/ADV

1. táblázat. Névmások kezelése az MSD-KR konverzió során

A Penn Treebank címkekészlete nem tesz különbséget a személyes névmások között sem nem, sem szám, sem személy, sem eset szerint, így egyaránt a `PRP` címkét kapja a *he*, a *she* és a *they*, a *you* és az *I*, valamint a *we* és az *us*. A birtokos esetű névmások külön címkét kapnak ugyan (`PRPS`), köztük azonban a predikatív használatúak (*mine, yours, ours, theirs*) sincsenek megkülönböztetve *my, your, our, their* párjaiktól. Mivel a KR-kód ezeket a különbségeket megjeleníti, ezért azokban az esetekben, ahol a szavak felszíni alakjából a megfelelő jegyek értéke kikövetkeztethető, az előző alfejezetben bevezetett, a szóalakra is hivatkozó szabályokkal ezek a címkék a konverzió során egyértelműsíthetők.

3.2 Többértelműség

Egyes esetekben a felszíni alakból nem minden jegy egyértelmű. Például a *his* szóról nem tudhatjuk, hogy predikatív használatú-e (*This cup is his*) vagy nem (*This is his cup*), a *you* névmás pedig lehet egyes- és többesszámú is. Ezekben az esetekben valódi kétértelműségről van szó, amit a felszíni alak figyelembevételével sem lehet felol-

dani – gazdagabb eszközökkel, például a környező szavak címkéinek figyelembevételével igen, ez azonban nem képezheti egy címkekonverzió részét. Így az ilyen szavaknál a KR-kódban a kétértelműséget jelöljük (az érintett jegy elé tett kérdőjellel), ezzel meghagyva a lehetőségét, hogy a szövegfeldolgozás magasabb szintű eszközei azt egyedileg kezeljék. Így például ha a morfológiai annotáció – a 2. fejezetben említett módon – jegy-érték struktúraként lesz értelmezve, akkor ezek a jegyek kitöltetlenek maradhatnak. Az angol személyes névmások kezelését a 2. táblázat foglalja összse.

	BNC	**KR**
it	PRP	NOUN
she	PRP	NOUN<FEM><NOM>
her	PRP	NOUN<FEM><ACC>
her	PRPS	NOUN<FEM><POSS>
hers	PRPS	NOUN<FEM><POSS<PRED>>
him	PRP	NOUN<MASC><ACC>
he	PRP	NOUN<MASC><NOM>
his	PRPS	NOUN<MASC><POSS<? PRED>>
me	PRP	NOUN<PERS<1>><ACC>
I	PRP	NOUN<PERS<1>><NOM>
us	PRP	NOUN<PERS<1>><PLUR><ACC>
we	PRP	NOUN<PERS<1>><PLUR><NOM>
our	PRPS	NOUN<PERS<1>><PLUR><POSS>
ours	PRPS	NOUN<PERS<1>><PLUR><POSS<PRED>>
my	PRPS	NOUN<PERS<1>><POSS>
mine	PRPS	NOUN<PERS<1>><POSS<PRED>>
you	PRP	NOUN<PERS<2>><? PLUR>
your	PRPS	NOUN<PERS<2>><? PLUR><POSS>
they	PRP	NOUN<PLUR><NOM>
their	PRPS	NOUN<PLUR><POSS>
theirs	PRPS	NOUN<PLUR><POSS<PRED>>
its	PRPS	NOUN<POSS>

2. táblázat. Névmások kezelése a BNC-KR konverzió során

3.2.1 Tulajdonnevek

Mind az MSD, mind a BNC kódolás a főneveken belül külön címkéket tart fenn a tulajdonneveknek. A `hun*` eszközkészlet tartalmaz tulajdonnév-felismerőt (`hunner`, Varga–Simon 2006), az azonban támaszkodik a morfológiai elemzés kimenetére. A `hunmorph` elemző a tulajdonneveket tehát a köznevekhez hasonlóan kezeli, ez azonban nem akadálya annak, hogy a KR-kódot alkalmassá tegyük a más elemzők (és kódrendszerek) révén már rendelkezésre álló információ tárolására. Az alkategóriák rendszere itt is könnyen használható, legegyszerűbben úgy, hogy a tulajdonnevek számára létrehozzuk a `[PROPER]/NOUN` alkategóriát. Ha azonban ennél gazdagabb, a tulajdonnevek kategóriájára vonatkozó információt is tárolni szeretnénk – mint például amilyet a `hunner` állít elő –, indokolt lehet több alkategória bevezetése: a `hunner` által is követett CoNLL szabvány szerint megkülönböztetjük a személy-, helység- és

cégneveket, valamint az egyéb kategóriába tartozó tulajdonneveket, és ezeket rendre a `[PER]/NOUN`, `[LOC]/NOUN`, `[ORG]/NOUN`, `[MISC]/NOUN` címkékkel látjuk el.

3.3 További tervek

Egy morfológiai elemzést végző eszköz megalkotásakor rendkívül fontos szempont a sebesség, ennek érdekében pedig kívánatos az elemző által használt nyelvtan bonyolultságának csökkentése. A morfológiai elemzés mint szimbólummanipulációs feladat általában megoldható *véges állapotú transzdúcerekkel* [FST, l. pl. Beesley–Karttunen (2003)]. A véges állapotú transzdúcereknek a sebesség mellett további nagy előnye, hogy az elemzéssel párhuzamosan a generálást is lehetővé teszik – hiszen egy transzdúcer birtokában mindkét irányba futtatható keresés. A magyar nyelv morfológiájára véges állapotú nyelvtant adott már Kornai (1994), a mai napig pedig több FST-alapú morfológiai elemző is készült magyar nyelvre, ilyen a 2. fejezetben már említett `magyarlanc` egy komponense, valamint a `foma`[i] nyílt forráskódú FST-implementációra épülő `hunmorph-foma`[ii]. Magyar nyelvű elemző készült még az ugyancsak véges állapotú technológiákon alapuló NooJ rendszerben[iii], melynek forrása azonban jelen cikk születésekor nem volt elérhető.

Egy elemző lehetséges kimenetei alkotják a címkekészletet, melyről már bármilyen konverzió – akár a jelen cikkben bemutatottak valamelyike is – egy véges táblázattal megadható, tehát az elemzési folyamat nem válhat bonyolultabbá attól, hogy a kimenetet egy más annotációs séma elemeire fordítjuk le. Mindebből következik, hogy egy konverziós tábla birtokában a fent említett véges állapotú morfológiai elemzők bármelyike könnyedén képessé tehető MSD vagy BNC kódok generálására is.

Irodalom

Beesley, K. R., Karttunen, L. 2003. *Finite-state morphology: Xerox tools and techniques*. Cambridge: Cambridge University Press.

Bird, S., Klein E., Loper E. 2009. *Natural language processing with Python*. Sebastopol: O'Reilly Media.

Burnard, L. 1995. *Users reference guide. British National Corpus version 1.0.* Oxford: Oxford University Computing Services.

Csendes, D., Csirik J., Gyimóthy R., Kocsor A. 2005. The Szeged Treebank. In: Matousek, V. et al. (szerk.) *Text, speech and dialogue.* Karlovy Vary: Springer. 123–131.

Erjavec, T. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In: Lino, M. T. et al. (szerk). *Fourth international conference on language resources and evaluation, LREC.* Lisbon; ELRA. 1535–1538.

Farkas, R., Szeredi D., Varga D., Vincze V. 2010. MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: Tanács, A., Vincze, V. (szerk.) *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, 354–357.

Halácsy, P., Kornai A, Oravecz Cs. 2007. Hunpos: An open source trigram tagger. In: Ananiadou, S. et al. (szerk.) *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Prague: Association for Computational Linguistics. 209–212.

Kornai, A. 1994. *On Hungarian morphology*. Budapest: Magyar Tudományos Akadémia Nyelvtudományi Intézete.

i https://code.google.com/p/foma/

ii http://freecode.com/projects/hunmorph-foma

iii http://www.nooj4nlp.net/pages/download.html

Marcus, M. P., Santorini B., Marcinkiewicz M. A. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19: 313–330.

Rebrus, P., Kornai A., Varga D. 2012. Egy általános célú morfológiai annotáció. In: Prószéky, G., Váradi, T. (szerk.) *Általános Nyelvészeti Tanulmányok XXIV*. Budapest: Akadémiai Kiadó. 47–80.

Recski, G., Varga, D. 2012. Magyar főnévi csoportok azonosítása. In: Prószéky, G., Váradi, T. (szerk.) *Általános Nyelvészeti Tanulmányok XXIV*. Budapest: Akadémiai Kiadó. 81–95.

Recski, G. 2010. Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In: Tanács, A, Vincze, V. (szerk.) *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 333–341.

Simon, E., Farkas R., Halácsy P., Sass B., Szarvas Gy., Varga D. 2006. A HunNER korpusz. In: Alexin, Z., Csendes, D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged: Szegedi Tudományegyetem. 373–377.

Trón, V., Kornai A., Gyepesi Gy., Németh L., Halácsy P., Varga D. 2005. Hunmorph: Open source word analysis. In: *Proceedings of the workshop on software*. Stroudsburg: Association for Computational Linguistics. 77–85.

Trón, V., Halácsy P, Rebrus P., Rung A., Simon E., Vajda P. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In: Calzolari, N. et al. (szerk.) *Fifth international conference on language resources and evaluation, LREC06*. Genoa: ELRA.

Varga, D., Simon, E. 2006. Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 16: 293–301.

Zsibrita, J., Vincze V., Farkas R. 2013. magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács, A., Vincze, V. (szerk.) *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 368–374.