MSD-KR harmonizáció a Szeged Treebank 2.5-ben

Farkas Richárd[1], Szeredi Dániel[2], Varga Dániel[2], Vincze Veronika[3]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
rfarkas@inf.u-szeged.hu
[2] BME Média Oktató és Kutató Központ
daniel@bme.mokk.hu, daniel@szeredi.hu
[3] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu

Kivonat: A magyar morfológiai erőforrások közül az egyik legelterjedtebben használt a morphdb.hu, amelynek morfológiai annotációs formalizmusa az úgynevezett KR-kódolás. A legnagyobb, kézzel egyértelműsített magyar nyelvi korpusz, a Szeged Treebank kódrendszere ezzel szemben az MSD-kódolást követi. A két kódolás nem kompatibilis egymással. Ez azt jelenti, hogy ha egy statisztikus módszerekkel tanított nyelvi elemző komponensben (POS-tagger, konstituenselemző, dependenciaelemző stb.) mindkét erőforrást ki kívánjuk aknázni, akkor nehézkes, információvesztéssel járó konverziós műveleteket kell végeznünk. Ebben a munkában beszámolunk a két kódrendszer (MSD és KR) közös nevezőre hozásáról, harmonizációjáról, amely megoldja a fenti problémát. A munka mindkét erőforrásban alapvető átalakításokkal járt. A konfliktusok nagyobb részében a harmonizációt közös finomítással igyekeztünk elvégezni, melynek hozadékaként jelentős mennyiségű manuális munka befektetésével a Szeged Treebank 2.5 által hordozott morfológiai információ részletgazdagabbá vált az előző verziókhoz képest.

1 Bevezetés

A magyar vonatkozású nyelvtechnológiai kutatásoknak és fejlesztéseknek alapfeltétele, hogy rendelkezésre álljon egy (lehetőleg egységes) nyelvi előfeldolgozó alapeszköztár. A rendelkezésre álló nyelvi elemzők egységesítésének legnagyobb akadálya a különböző morfológiai kódrendszerek használata. Cikkünkben beszámolunk két magyarra alkalmazott kódrendszer (MSD és KR) közös nevezőre hozásáról, harmonizációjáról. Ehhez tételesen ismertetjük a kódolások közötti elméleti különbségeket, majd az összehangolás során meghozott kompromisszumos döntésekről is beszámolunk. Az átalakított kódrendszernek megfelelően a morphdb.hu-ban [4] is változásokat eszközöltünk és a Szeged Treebank [2] szövegállományát is újrakódoltuk (a létrejött új verziót Szeged Treebank 2.5-nek kereszteltük). Célunk, hogy az egységes morfológiának köszönhetően létrejöhessen egy olyan morfológiai elemző, amely a Szeged Korpusszal is kompatibilis, annak érdekében, hogy a morfológiai elemzőre egy olyan POS-tagger legyen építhető, amely a magasabb szintű elemzé-

sekhez, illetve alkalmazásokhoz (dependenciaelemzés, információkinyerés) hasznos bemenetet szolgáltat.

2 Morfológiai kódrendszerek a magyar nyelvre

Az MSD morfológiai kódrendszer [3] több nyelvre, többek közt a magyarra lett kifejlesztve. A kódokon belül az első pozíció adja meg a fő szófaji kategóriát, míg a további pozíciók egyéb nyelvtani információkat tartalmaznak (pl. ige esetében az ige típusát, módját, idejét, számát, személyét, ragozását: a **Vmis2s---y** kód például egy kijelentő módú, múlt idejű, egyes szám második személyű tárgyas ragozású főigét jelöl).

A KR kódrendszer a magyar nyelv morfológiáját szem előtt tartva lett kidolgozva, bár alapvető szintaxisa nyelvfüggetlen, és a későbbiekben több más nyelvhez is készült a szintaxisra és a kódrendszer alapelveire épülő morfológiai erőforrás [4]. Magyar nyelvre történő implementációja, a morphdb.hu morfológiai elemző erőforrás létrehozásakor a legfontosabb célkitűzések a teljesség és az elméleti nyelvészeti szempontból való megalapozottság voltak, valamint hangsúlyos szempont volt a nyílt forráskódú szabad hozzáférhetőség. A kódrendszer hierarchikus jegy-érték struktúrában kódolja a nyelvészeti információkat: vannak alapértelmezett (default) jegyek (például egyes szám, harmadik személy), és csak az ettől eltérők jelennek meg a kódban. A fenti példa KR-kódolása a következő: **VERB<PAST><PERS<2>><DEF>**. A kódok inflexiós és derivációs információt is tartalmaznak.

A HUMor morfológiai kódrendszer az unifikációs nyelvleíráson alapul, azaz a tövek és morfémák más morfémákkal való együttes előfordulásra való képességük alapján jegyekkel vannak ellátva. E jegyek lehetnek egymást megengedők vagy egymásnak ellentmondók: egy szóalak csak olyan morfémákból épülhet fel, amelyek jegyei nem zárják ki egymást [5]. Az elemzés eredményeképpen a szó morfémákra bontott változatát kapjuk, minden morféma mögött szerepel a szófaji megjelölése, és ha eltér a szótári alakja, az is (*megy~me*), például: **mehetsz --megy[IGE]=me+het[HAT]+sz[e2]**.

Mivel a Szeged Korpusz építéséhez a szófaji előelemzést a HUMor morfológiai elemzőprogram végezte, melynek végeredményét automatikusan konvertálni kellett MSD-kódokra [1], az MSD és a HUMor kódrendszer harmonizációja már korábban megtörtént: a végeredmény a Szeged Treebank szófaji kódjaiban is tükröződik. Jelen cikkben a KR és MSD kódrendszerek összehangolására teszünk kísérletet.

3 A KR és MSD kódrendszerek harmonizációja

A kódrendszerek összehangolásában azt az alapelvet követtük, hogy a morfológiai kódoknak olyan (és csak olyan) információkat kell tartalmazniuk, amelyek a későbbi feldolgozás (szintaxis, különféle alkalmazások) szempontjából hasznosak. Ennek fényében mérlegeltük az egyes esetekben, hogy az MSD vagy pedig a KR rendszer megközelítését építsük-e be a harmonizált morfológiába.

Az egyik lényegi különbség a képzések kezelésében nyilvánul meg: míg a KR abszolút, addig az MSD relatív szótövekkel dolgozik. Ennek megfelelően a képzők nincsenek is kódolva MSD-ben, míg KR-ben igen, így adott esetben a szóalakok lemmája is eltér egymástól. A képzés hiányából adódóan az MSD kódrendszer nem tudja megkülönböztetni például ugyanannak az igének a műveltető vagy ható képzős alakjait a kód szintjén (természetesen a lemma eltérő) – ezzel szemben a KR-ben a lemma ugyanaz, de a kód különbözik.

Megoldásunk ebben az esetben az lett, hogy mindkét rendszerből átvesszük az indokolható megkülönböztetéseket. A relatív lemmák általában elég információt szolgáltatnak az alkalmazásoknak (pl. információ-visszakeresés), és a képzők annotálása a Szeged Korpuszban irreálisan nagy feladat lett volna, így a harmonizált kódrendszer is relatív lemmákkal dolgozik. Néhány esetben azonban indokolt volt kivételt tenni. A műveltető, gyakorító és ható[1] igék esetében fontos, hogy a képző csak aspektuális, illetve modális változást jelent, melyeket más nyelvek más – nem morfológiai, hanem például szintaktikai – eszközökkel fejeznek ki, aminek például a gépi fordításban lehet jelentősége. Ha pl. egy műveltető igealakot tartalmazó mondatot akarunk gépi úton angolra fordítani, akkor az MSD-kódolást használva abba a problémába ütközünk, hogy nagy valószínűséggel nem találunk a lemmának megfelelő szóalakot a szótárban. A KR-elemzést tekintve azonban a szótárban is megtalálható lemmából indulunk ki, és ha megfelelő fordítási szabályokat rendelünk a műveltetés (például használd a have + tárgy + ige 3 alakja szerkezetet) megfelelő kezeléséhez, akkor eljuthatunk a helyes fordításhoz.

Ezek alapján fontosnak tartottuk, hogy ezek az információk kódolva legyenek az MSD kódrendszerben is. Az igetípus pozíciójában azt is megjelöljük, hogy az ige műveltető (kódja: s), ható (kódja: o) vagy gyakorító (kódja: f) alakban szerepel-e.

Egy másik nagy elvi különbség a kódrendszerek között a névmások kezelése. Míg az MSD-ben külön szófaji kategóriának számítanak, addig a KR a helyettesített szófaj szerint kódolja őket. Az egységesítés eredményeképpen a KR rendszerbe is bevezettük a névmásokat PRONOUN jelöléssel.

A határozószavak kezelésében is mutatkoznak eltérések: az MSD-ben alosztályokba vannak sorolva, a KR-ben pedig egységesen <ADV> kóddal rendelkeznek. Az egységesítés folyamán az alosztályok megkülönböztetését választottuk, ugyanis ennek például a fokozásban van jelentősége. Az MSD kódrendszer képes jelölni a határozószavak fokozását, míg a KR-ből ez hiányzik: a *lejjebb*, *közelebb* alakok lemmája *lejjebb*, *közelebb*, kódolása pedig ADV. Az MSD-n belül mindez Rxc kódú (a c jelöli a középfokot), a lemmák pedig *lent* és *közel*. Viszont nem minden határozószó fokozható (a kérdő vagy általános határozószók például nem), ezért úgy szükséges módosítani a KR-kódolást, hogy csak bizonyos altípusok esetén legyen megengedve a fokozás lehetősége.

Az ún. személyes névmási határozószavak kérdése jelentette az egyik legjelentősebb elvi különbséget a két kódrendszer között. Míg MSD-ben a határozószavak egy altípusaként voltak kódolva (pusztán számot és személyt kódolva), addig a KR-ben

[1] Megjegyezzük, hogy az eredeti KR rendszerben a *-hAt* toldalék inflexióként jelenik meg, a harmonizált kódrendszerben azonban hasonlóképpen kezeljük a műveltető és gyakorító igeképzőkhöz, ezért itt tárgyaljuk.

főnévként: a határozórag alapúaknál (pl. *nekem*, *veled*) a személyes névmás szerepelt lemmaként, és a főnévi paradigmához hasonlóan kaptak esetet, a névutóból képzettek (*mögötted*, *szerintünk*) kódja pedig tartalmazta az eredeti névutót. Néhány példa: a *nekem* KR-elemzése én/NOUN<CAS<DAT>>, az MSD-elemzése Rl--s1 (*neki* lemmával), a *szerintem* szó esetében pedig én/NOUN<POSTP<SZERINT>>, illetve Rl--s1 (*szerinte*). A példákból ismét csak megmutatkozik az az eltérés a kódrendszerek között, hogy míg MSD-ben a kódolások megegyeznek, de a lemmák eltérnek, a KR rendszerén belül a lemmák megegyeznek, de a kódok különböznek.

Ennél a problémakörnél teljes egészében egyik rendszer megoldását sem vettük át. Mivel személyes névmásokból származtatjuk az alakokat, ezért a személyes névmási rendszerbe illesztjük be őket.

Szavak és szóalakok szófaji besorolását tekintve is találhatunk különbségeket a két kódrendszer között: jellemzően a kötőszavak és a határozószavak csoportjában fordul elő, hogy az egyik kódrendszerben kötőszó, a másikban határozószó az adott szóalak (pl. *majd*, *persze*). Ezek státuszáról egyenként hoztunk döntést, nyelvi disztribúciójukat mérlegelve.

Néhány kisebb horderejű különbség is megfigyelhető a két kódrendszer között. A főnevek kategóriáján belül ilyen például a köznév-tulajdonnév megkülönböztetés, mely az MSD sajátja. Mivel úgy gondoljuk, hogy nem a morfológiai elemző feladata eldönteni egy adott főnévről, hogy az tulajdonnév-e vagy sem (hanem egy NE-felismerőé), úgy döntöttünk, hogy az MSD-n belül sem érdemes ezt az elkülönítést alkalmazni. A familiáris többes számot a KR külön kódolja <FAM> jeggyel, az MSD-ben azonban ez nem szerepel. Mivel alkalmazási szempontból nem tűnt szignifikánsnak a többes szám kétféle jelölése, az egységes morfológiában csak egy "általános" többes számot használunk.

A Szeged Treebank 2.5 munkálatai nem csak elvi morfológiai átalakításokban öltöttek testet: 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.

4 Konklúzió

Az előző fejezetben bemutatott harmonizációs lépéseket a morphdb.hu és a Szeged Korpusz manuális átalakításával valósítottuk meg. A két nyelvi erőforrás átalakításának statisztikai mutatóinak bemutatására hely hiányában nincs lehetőségünk, de részleteiben is elérhetőek a www.inf.u-szeged.hu/rgai/krmsd honlapon.

A cikkben bemutatott egységes morfológiának köszönhetően lehetővé vált olyan morfológiai elemző építése, amelynek kimenete a Szeged Treebankkel teljes összhangban van, és ezért a rá épülő, magasabb szintű nyelvi elemzést végző szövegfeldolgozó rendszerek (mint a magyarlanc[2] és hun* eszközláncok) a Szeged Treebank által hordozott minden morfológiai információt ki tudják használni statisztikus modelljeik tanításakor.

[2] www.inf.u-szeged.hu/rgai/magyarlanc

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Alexin, Z., Csirik, J., Gyimóthy, T., Bibok, K., Hatvani, Cs., Prószéky, G., Tihanyi, L.: Manually Annotated Hungarian Corpus. In: Proceedings 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) 53-56
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. Erjavec, T. (ed.): MULTEXT-East morphosyntactic specifications. Version 3 (2004) http://nl.ijs.si/ME/V3/msd/msd.pdf
4. 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
5. Prószéky, G., Tihanyi, L.: Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. La tribune des industries de la langue 10, OFIL, Paris, France (1993) 28–29