morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis

Trón Viktor[1], Halácsy Péter[2], Rebrus Péter[3], Rung András[3],
Simon Eszter[4], és Vajda Péter[3]

[1] International Graduate College Language Technology and Cognitive Systems
University of Edinburgh és Saarland University
v.tron@ed.ac.uk
[2] BME – Média Oktató és Kutató Központ
hp@mokk.bme.hu
[3] MTA Nyelvtudományi Intézet – MTA-ELTE Elméleti nyelvészet program
{rebrus,runga,vajda}@nytud.hu
[4] BME – Kognitív Tudományi Tanszék
esimon@cogsci.bme.hu

Kivonat: Cikkünkben a `morphdb.hu` adatbázist mutatjuk be, amely a magyar nyelv egy minden eddiginél teljesebb és elméleti alapokon álló morfológiai leírása. A leírás a `hunlex` keretrendszerben van megfogalmazva, így a `hunmorph` szóelemző eszközkészlet segítségével az adatbázis helyesírásellenőrzéshez, tövezéshez, morfológiai elemzéshez és számos egyéb annotációs feladathoz használható elsődleges nyelvi erőforrásként.

Bevezetés

A szabály alapú szóelemzők működése – akár a végesállapotú, akár affixumlevágásos architektúrára gondolunk – a nyelv szókészletét és morfológiáját leíró erőforrásokat feltételez. Cikkünkben a `morphdb.hu` adatbázist mutatjuk be, amely a magyar nyelv minden eddiginél teljesebb és elméleti alapokon álló morfológiai leírása.

A `morphdb.hu` szóanyaga a helyesírásellenőrzésre használt Magyar Ispell szótár [6], az Elekfi László jegyezte Magyar Ragozási Szótár [2], valamint az ún. FKP-szótár [3] szóanyagának kritikus összefésülésével készült.

A szótár és a hozzá tartozó morfológiai nyelvtan leírását a `hunlex` keretrendszerben [8] végeztük el. A `hunlex` a morfológiai leírásból olyan kimeneti állományokat állít elő automatikusan, amelyeket a `hunmorph` szóelemző-algoritmusai (és hasonló `ispell` típusú erőforrást használó affixumlevágásos szóelemzők [9]) igényelnek. A `hunlex` és a `hunmorph` segítségével a `morphdb.hu` adatbázis így helyesírásellenőrzéshez, tövezéshez, morfológiai elemzéshez és számos egyéb annotációs feladathoz használható elsődleges nyelvi erőforrásként.

A hunlex rendszer morfológiai jelenségek formalizálására kidolgozott leíró nyelve alkalmasnak bizonyult a magyar nyelv komplex morfológiai jelenségeinek leírására. Az alábbiakban először a morfológiai nyelvtan néhány vonását mutatjuk be (§1), majd rátérünk a szótári anyag ismertetésére (§2). Összegzésképpen megmutatjuk, hogy miként mérhető az adatbázis lefedettsége és pontossága, valamint kitérünk jövőbeni terveinkre (§3).

1. A morfológiai leírás alapelvei és szerkezete

Tőtár és morfológiai folyamatok A hunlex rendszer leíró nyelve Item-and-Process [4] típusú morfológiai leírások formális keretéül szolgál. Egy hunlex morfológiai leírás lényegében egy tőtárból (lexicon állomány) és egy morfológiai operációkat formalizáló nyelvtani leírásból (grammar) áll. Az operációk konceptuálisan két részre oszthatók: (i) egyik részük konkrét morfoszintaktikailag additív szabályok, vagyis olyanok, amelyeket morf hozzáadásként is értelmezhetünk; (ii) másik részük pedig absztrakt morfofonológiai folyamatokat ír le, vagy diakritikumok (idioszinkratikus jegyek) és fonológiai-ortográfiai mintázatok közötti megfeleléseket fogalmaz meg. Az utóbbi, (ii)-es típusú szabályokat *szűrőszabályoknak* is nevezhetjük. Ilyen például a hangkivetés vagy a rövidülés mint absztrakt morfofonológiai folyamatok, vagy az ikességnek és a múlt idejű ragozásnak az összefüggését kimondó szabályok. Bizonyos szabályok (például a rövidülést leíró szabályok) egyazon absztrakt folyamat (rövidülés) megvalósulási formáinak tekinthetők, amelyeket közösen *szűrőnek* hívtunk.

Toldalékszabályok és szűrők A nyelvtan implementálásakor konkrét döntéseket kellett hoznunk arról, hogy a nyelvészeti elemzések által feltételezett morfofonológiai folyamatok közül melyeket általánosítjuk és melyeket nem. A leíráskor szem előtt tartottuk azt, hogy a szabályok leírása a legközvetlenebbül tükrözze a hagyományosan allomorfiának nevezett váltakozásokat, így a nem-konkatenatív morfofonológiai folyamatokat (pl. a magánhangzó-harmóniát kezelő fonológiai szabályt) nem absztrakt ((ii)-es típusú) szabályként írtuk le, hanem az allomorfikus szabályokban adtuk meg. Ezt illusztrálja az 1. ábra. (A hunlex leíró nyelvének technikai részleteire itt nem áll módunkban kitérni.)

```
CAS_INE
        IF: analytic lengthened cas_ine
        TAG: <CAS<INE>>

, +ban IF: back
, +ben IF: front
;
```

1. ábra. A vesszővel elválasztott szabályok a toldalékmorféma allomorfjainak (*-ban*, ill. *-ben* morfok) feleltethetők meg, a teljes szabály (CAS_INE) pedig így magának az affixummorfémának (az *inessivus* esetragnak).

A nyelvtan morfoszintaktikailag additív ((i)-es típusú) szabályait aszerint csoportosítjuk, hogy milyen morfoszintaktikai tulajdonságok (az 1. ábrán például az

inessivus eset) kifejezéséért felelnek. És megfordítva, egy ilyen szabálycsoport hagyományosan egyetlen morfémának tekintett toldalékot ír le.[52] Így ezeket a szabályegyütteseket (az 1. ábrán `CAS_INE`) némiképp lazán toldalékmorfémának nevezzük, az egyes szabályokat pedig toldalékallomorfoknak. A fenti elvek eredményeképpen lényegesen egyszerűsödik és átláthatóbb lesz a nyelvtan, és ezáltal új toldalékokkal való bővítése az Item-and-Arrangement szemléletet ismerő nyelvész vagy a fonológiai folyamatokhoz nem értő laikus számára is könnyebb.

Toldalékváltozatok kondicionálása A folyamatoknak a bemenetre való alkalmazása számos feltételhez lehet kötve. A `hunlex` leírásban a feltételek között szerepelhet mind a mintaillesztés, mind a jegyekre való hivatkozás (az 1. ábrán például a magánhangzó-harmónia jegyeire hivatkoznak a szabályok). A toldalékallomorf kiválasztása gyakran idioszinkratikus tulajdonsága a tőnek (pl. *hal<u>at</u>*, de *dal<u>t</u>*), ezeket tehát absztrakt jegyek ellenőrzésével kell kezelnünk. Ráadásul a fonológiailag megjósolhatóan kondicionált váltakozásokat a helyesírási konvenciók miatt egyes esetekben szintén önkényesen kellett kezelnünk (pl. *Voltaire-rel*). A legtöbb allomorfikus szabály alkalmazási feltételeit tehát jegyek segítségével fejeztük ki.

Lexéma-alapú tőtár és alulspecifikáció A lexikon bővítése akkor a legegyszerűbb, ha a (i) szótári bejegyzések lexémáknak[53] felelnek meg, valamint (ii) a szótári bejegyzéshez a lehető legminimálisabb morfológiai információt kell megadni. A szótárak a lexikográfiai hagyomány szerint lexéma-alapúak és a bejegyzések az ún. 'szótári alak' mellett csak a kivételességek (és megjósolhatatlan információ) feltüntetésével kerülnek az állományba. Ezt szem előtt tartva a nyelvtan részeként olyan folyamatokat (szűrőszabályokat) is feltételeznünk kellett, amelyek (i) a szótári alakokból előállítják a tőváltozatokat, valamint amelyek (ii) a toldalékváltakozások feltételeként használt megjósolható, de a lexikonban alulspecifikált jegyeket a tőváltozatokhoz asszociálják. A nyelvtan érdekes része tehát az a szűrőlánc, amely a morfofonológiai jegyeket kiosztja és a tőalternánsokat előállítja (például hangbetoldási, rövidülési és nyúlási folyamatok alkalmazásával).

Unáris licenciálás és opcionalitás A morfológiai adatbázisban túlnyomó részben unáris (egyértékű) jegyeket használtunk. Unáris jegynek nevezünk egy jegyet, ha a szabályok alkalmazási feltételei csak a jegy meglétére hivatkoznak. Például azoknál a toldalékolási folyamatoknál, ahol a kötőhangzó középső, illetve alsó nyelvállású (nyílt) is lehet (pl. többesszám *-ak*, *-ok*) a bemeneti relatív tőtől függően, ott a nyílt, illetve középső nyelvállást egy-egy jegy engedélyezi. A nyitótöveket (pl. *hal*, ill. nyitó relatív töveket, pl. *fák-*) a `low`, a nemnyitókat (pl. *dal*) pedig a `non_low` jegygyel kell ellátni ahhoz, hogy a megfelelő toldalékokat megkaphassák. Azok a tövek, amelyek opcionálisan nyitók (pl. *öröm*) egyszerűen mindkét jeggyel rendelkeznek. Ezzel a megközelítéssel elérhető, hogy (i) a szabályokat átlátható módon pozitív feltételhez kössük, ugyanakkor (ii) már a jelölés szintjén megmutatkozzon az opcionalitás jelöltsége, hiszen egy opcionális tőnél az adott jegydimenzió (nyitás)

[52] Ettől csak néhány praktikus esetben tértünk el, amikor két morféma szétválasztása bonyolítaná a nyelvtant, például a múlt idő jelét és az azt követő személyjeleket nem választottuk külön.

[53] Pontosabban alaplexémáknak (azaz nem megjósolhatóan képzett lexémáknak).

mindegyik unáris jegyét specifikálni kell. Azt, hogy egy kategória (pl. névszó) az adott dimenzióban (pl. nyitás) mindenképp felvesz egy értéket, itt azt jelenti, hogy a dimenziót kifejező unáris jegyek közül a jelöletlen vagy default jegyet (a nyitás esetében `non_low`, vagyis nemnyitó) az adott kategória minden eleme meg kell, hogy kapja. Ezt a nyelvtanban egy külön szűrő fejezi ki, amely a nyitást szabályozza: ez a szűrő a 2. ábrán látható.

```
NOM_LOWERING_FILTER

    FREE: false
    FILTER: low non_low
    OUT: NOM_KEEP_ALL_FEATURES
    OUT: NOM_ACC_FILTER


,OUT: non_low
;
```

2. ábra. Példa egy szűrőre: a nyitás szűrője minden nyitásra nem specifikált névszóhoz a (`non_low`) jegyet rendeli.

A nyitáshoz hasonlóan, azoknál a morfofonológiai tulajdonságoknál, amelyeknél felmerül az opcionalitás, következetesen unáris jegyeket használtunk. Ilyen tulajdonságra példa még az előlségi harmónia, az ikesség, a tárgyasság és a legtöbb tőváltakozás. Továbbá minden olyan tulajdonságra, amelynél az idioszintkatikus jegyek közül az egyik jelöltnek tekinthető, ott a nyitóhoz hasonló szűrőszabályt vezettünk be. Ilyenre példa az igéknél a tárgyasság.

Analitikusság és tőváltakozások Hasonlóan kezeltük a tőváltozatok választását befolyásoló ún. szintetikus illetve analitikus toldalékolást is [7]. Tőalternációt mutató lexéma esetén az egyes tőváltozatok kapják a jegyek valamelyikét. Itt a tőváltakozást nem mutató, egyalakú tövek opcionális töveknek felelnek meg, hiszen az egyalakú tő mind analitikus mind szintetikus toldalékolási folyamatokban részt vehet. Az analitikusság jegyei a nyelvtan belső jegyei, közvetlenül a lexikonban csak a tőváltakozás típusát (rövidülő, hangkivető, *v*-vel bővülő, *sz-d* tő, stb.) kódoló jegy jelenik meg. A szintetikus és analitikus tőváltozatok előállítása során a megfelelő jegyek a változatokhoz rendelődnek, az egyalakú szótári tövek pedig automatikusan mindkét jegyet megkapják.[54]

Magánhangzónyúlás Hasonlóan kezeltük a névszói magánhangzónyúlást is. Azok a toldalékok, amelyek kiváltják a magánhangzó-nyúlást (többesszám, birtokjel, a legtöbb esetrag, stb.) az affixumszabály feltételeként ún. „hosszú magánhangzós tő" jegyet (lengthened) követelnek meg, míg a nem nyújtó toldalékok (pl. *-ként*, *-kor*, *-ság/ség*) a „nem hosszú magánhangzós tő" jegyet (`non_lengthened`). A magánhangzónyúlás, mint folyamat így egy virtuális morf segítségével kezelhető, amely a

[54] Bizonyos tövek egyalakúak, de mégsem vehetnek fel minden szintetikus és analitikus toldalékokat: ezek az ún. defektív tövek (pl. *rejlik* (vö. **rej(e)ljen*), amelyek már a lexikonban szintetikusként szerepelnek.

megnyújtott véghangzós tövet (*fá-*) előállítja és ellátja a hosszú jeggyel, a szótári tövet (*fa*) pedig a nem-hosszú jeggyel látja el. Azok a tövek, amelyek végződésük szerint soha nem nyúlnak (pl. *mozi*, *bot*) megkapják mind a hosszú mind a nem-hosszú jegyet, így természetes módon az egyalakú tövekhez hasonlóan viselkednek.[55]

Idegen helyesírású szavak kiejtés szerinti toldalékolása A fonológiailag és ortográfiailag kondicionált allomorfia jegyekkel történő kezelése egyéb pontokon is hasznosnak bizonyult. Egyrészt az affixumszabályok leírásakor az absztrakt feltételek megadása tömörebbé válik, és ezek koordinálhatók lesznek (több konjunktív feltétel megadható egyszerre, pl. hosszú és mély hangrendű), ami az elemző algoritmus korlátai miatt nem lehetséges közvetlen illesztési kifejezés esetén. A másik nagyon fontos előny pontosan ahhoz kapcsolódik, hogy a nyelv hangtanilag kondicionált szabályszerűségeit az ortográfia speciális szabályai szerint írjuk le. Az idegen helyesírású, de kiejtés szerint ragozott szavak (pl. *Voltaire*) kezeléséhez a toldalékválasztást adó jegyek elengedhetetlenek. Ez a fajta ortográfiai önkényesség egyszerű módon kezelhető, ha bizonyos idegen szavak a kiejtésükkel együtt vannak felvéve a lexikonba (pl. *Voltaire/volter*). A `hunlex` keretrendszer lehetővé teszi, hogy az ilyen szavaknál, az egyes toldalékok az írásképhez járuljanak, de úgy, hogy a toldalékallomorfok kiválasztása mégis a kiejtésük szerint történjen. Ehhez csupán annyi kell, hogy a kivételes kiejtésű szótári tételeknél fel legyen tüntetve a filterekben található minták szempontjából releváns kiejtés. Ekkor a tő a jegyeket a kiejtés szerint kapja meg, majd a szűrőlánc végén egy szabály egyszerűen törli a kiejtésre vonatkozó részt a tőből. A `hunlex` azt is lehetővé teszi, hogy bizonyos kiejtési szabályokat is a nyelvtanban adjunk meg. Ezt egyelőre csak a szavak egy speciális csoportjánál, a betűszavaknál (pl. *http*) vezettük be, ahol a kiejtés egyértelműen rekonstruálható (*hátétépé*).

Morfoszintaktikai jegyek és hiányos paradigmák Mivel az igék és a névszók releváns morfofonológiai jegyei nagyrészt diszjunktak, ez a két nagy kategória külön szűrőláncot kívánt. Hasonlóan a névszókon belül számos (főként képzési) folyamat érzékeny arra, hogy az illető névszó főnév, melléknév vagy számnév. Az ilyen alkategória-érzékeny toldalékolás miatt az egyes morfémák (*–s* képző, *-szoroz* képző, stb.) és morfémacsoportok (esetrag, birtokos személyjel, stb.) kapcsolódását szintén jegyek meglétéhez kötöttük. Az ilyen (unáris) morfoszintaktikai jegyek alkotják a `morphdb.hu` jegyeinek másik részét. Hasonlóképpen ilyen jegyekkel értük el, hogy az igéknél a tárgyasságra, vagy a névszóknál a köznév–tulajdonnév különbségére érzékeny szabályok ne generáljanak túl. Egyes morfoszintaktikai jegyek a lexéma alaptulajdonságai (főnév, számnév, stb.), így a tőtárban minden bejegyzésnél szerepelnek. A lexikonállományban más jegyek csak jelölt esetben (tulajdonnév, tárgyas ige) szerepelnek, ilyenkor az adott dimenzión belüli default tulajdonságokat egyszerű

[55] A nyúlás jegyei szintén a nyelvtan belső jegyei, azonban a kiosztásukat szabályozó filter nem lexikai jegyekre, hanem fonológiai mintázatokra hivatkozik. A nyúlást leíró folyamat default szűrőként való implementálását az motiválja, hogy egyes alakok (*la*, *Che*, *Mandrake*) a releváns mássalhangzó ellenére nem váltakoznak, vagyis ezeket a lexikonban az opcionális alakokhoz hasonlóan a dimenzió minden jegyével el kell látni.

szűrőszabályok segítségével rendeltük az alakokhoz.[56] Ezek a morfoszintaktikai jegyek szabályozzák a hiányos paradigmájú alakokat is (pl. *léptek*, *két*, *ismerszik*).

Összefoglalásképpen a `morphdb.hu` igei szűrőláncának összetevőit soroljuk fel.

- az ikesség felismerése a szótári alak alapján automatikus, tő előállítása
- a default tőtípus hozzárendelése a nemkivételes egyalakú lexémákhoz
- az elölségi harmóniát szabályozó unáris jegyek hozzárendelése mintaillesztéssel
- tőváltozatok előállítása: hangkivetés és -betoldás, *sz-d*(-*v*)-tövek, *v*-tövek, rövidülés
- kerekségi harmónia unáris jegyeinek hozzárendelése mintaillesztéssel
- szótagszámot kódoló jegyek hozzárendelése (a szótagszámérzékeny szabályok miatt) mintaillesztéssel
- alanyi ill. tárgyas ragozást engedélyező jegyek hozzárendelése: a default csak alanyi ragozás
- kvázianalitikus toldalékolást szabályozó jegyek hozzárendelése mintaillesztéssel (*hoznak* vs. *vonzanak*)
- a múlt idő változatait szabályozó jegyek hozzárendelése mintaillesztéssel (*hoztak* vs. *vonzottak*)
- összetételi határokat törlő szabályok
- kategóriaérzékeny képzést engedélyező morfoszintaktikai jegyek kiosztása

Alulspecifikáció és kivételesség A mintaillesztéssel történő allomorfkiválasztás (jegy-hozzárendelés) csak akkor lehetne lehetséges, ha a szóban forgó jelenség tökéletesen megjósolható lenne. Bár számos ilyen jelenség van (pl. az ikesség a szótári alak alapján, vagy a kerekségi harmónia a kiejtés alapján tökéletesen megjósolható), sok esetben a toldalékolás önkényes, azaz megköveteli, hogy egyes jegyeket a lexikonban adjunk meg. Ilyen esetekben azt az elvet követtük, hogy minél tágabb körű általánosítást fogalmazunk meg a nyelvtan szűrőiben úgy, hogy a kivételesnek tekintett (tehát a lexikonban jeggyel ellátott) alakok lehetőség szerint véges zárt osztályt alkossanak. Egy adott dimenzióra nézve ennek a zárt osztálynak a tagjai tekinthetők kivételesnek. Ezzel a szótárbővítési munkát remélhetőleg minimálisra csökkentjük, hiszen ez a módszer csak olyan jegyek specifikációját kényeszeríti a szótárfejlesztőkre, amelyek egy nyílt osztályra is megjósolhatatlanok. A morfofonológiai jegyek közül csak a névszói birtokos toldalékolás típusát (*-a/e* vagy *-ja/-je*) szabályozó jegy ilyen.

[56] Mivel vannak csak tárgyas részparadigmát megengedő igék (*megemberel(i magát)*), ezért a tárgyasság dimenziója is két unáris jeggyel ábrázolódik: `verb_indef` (engedélyezi az alanyi részparadigma toldalékait), és `verb_def` (engedélyezi az tárgyas részparadigma toldalékait). Hasonlóan a hangrendileg ingadozó opcionalitáshoz, a tárgyas igéket (amelyek mindkét részparadigmát engedélyezik) a két jegy együttes jelenlétével adjuk meg.

2. A magyar morfológiai adatbázis szóanyaga

A szótári anyag A `morphdb.hu` szótári anyagának elkészítéséhez három önmagában is nagy lefedettségű elektronikus szótárat használtunk fel. A Magyar Ispell szótár [6] a szabad forráskódú szoftverek világában domináns `ispell` alapú helyesírásellenőrző magyar nyelvű erőforrása. A Németh László vezetése alatt közös munkaként elkészült Magyar Ispell szótár a mai magyar nyelv egyik legteljesebb és legnaprakészebb szóanyagát tartalmazó anyag. A `magyarispell` átalakításakor a `hunmorph` morfológiai elemzőhöz készült szótár nyers változatából indultunk ki. A több mint 100 ezer szavas szóállomány témakörök, stílusminősítések alapján van csoportosítva, az ún. alapszókincs mintegy 37 ezer szóból áll. A tematikus szótárak anyagát is átvettük, megtartva a témakört, mint a szótári bejegyzéshez adott használati információt.

Másik forrásunk a Magyar Ragozási Szótárnak [2] a Nyelvtudományi Intézet Korpusznyelvészeti osztálya által digitalizált változata. Az Elekfi-szótár az Értelmező Kéziszótár közel 70 ezer lemmáját tartalmazza paradigmaosztályokba sorolva. A szótár jelöli a komplex szavak belső szerkezetét is, és számos összetettségi típust elkülönít. Ezek közül némelyeket egybevontunk, de az összetett szavakat valamint igekötős igéket felbontásukkal együtt átvettük.

Harmadik forrásunk, az FKP-szótár, Papp Ferenc Szóvégmutató Szótárából és a Füredi-Kelemen-féle Gyakorisági szótárból állt elő [3] és kb. 70 ezer tételt tartalmaz.

A források átalakítása A `morphdb.hu` szótár előállításának első részeként ezen létező szótárak anyagát kellett a `hunlex` lexikon formátumára hozni, úgy, hogy a morfológiai információikat az első részben vázolt elvek figyelembevételével a `hunlex` nyelvtanban használt jegyekké átalakítsuk. A morfofonológiai jegyeket egy-egy szónál csak akkor kívántuk felvenni, ha a nyelvtanban meghatározott szabálytól eltérően viselkednek, azaz a szó ebből a szempontból kivételes. A források viszont minden információt tartalmaznak, így a nyelvtanunk szempontjából szabályosnak tekinthető folyamatokat szabályozó default jegyeket is. Az átalakításnál ennek a redundanciának a kiküszöbölése volt az egyik probléma.

Második lépésként az átalakított és redundanciamentesített forrásszótárakat kellett összefésülni. Ez a több forrásban is előforduló bejegyzések esetében annak eldöntését is megkívánta, hogy melyik „autoritás” által adott morfológiai leírást fogadjuk el a legpontosabbnak.

Célunk nem csupán egy minden eddiginél nagyobb, de elméletileg is jobban megalapozott morfológiai leírást követő szótár előállítása volt. A morfológiai információ átvételében csak a nyílt szóosztályok (névszók és igék) esetén követtük az eredeti források morfológiai leírását. A határozószavak, névutók, névmások, kötőszók és egyéb szófajok feldolgozásánál csupán a szóanyagot vettük át, de megkíséreltünk egy adekvátabb leírást, csoportosítást kialakítani.

A jegyek hozzárendelése: nyílt szóosztályok A magyarispell szóállomány már bizonyos morfofonológiai tulajdonságok alapján (hangrend, birtokos *-j* megléte, tőallomorfia, tárgyasság, stb.) csoportosítva tartalmazza a szavakat és a kivételeket, ezért ezek átalakítása a nyelvtanunkban használt jegyekké nem ütközött nehézségekbe.

A Magyar Ragozási Szótár a bejegyzéseket lehetséges toldalékaik alapján diszjunkt csoportokba, paradigmákba sorolja. Egy-egy paradigmába csak a teljesen

egyformán viselkedő (ugyanazon affixumokat felvevő) szavak kerültek, és a paradigmaosztályok száma összesen 1700. Az osztályok számozása nyelvészeti szempontból nem rendszerezett, ami azt jelenti, hogy a kódok közti különbség a legtöbb esetben (kivéve az előlségi és a kerekségi harmóniát, illetve az ikes és nem-ikes igéket) nem tükrözi az adott kódú csoportokba sorolt szavak közti ragozásbeli különbséget. Ezért az átalakításához egy olyan táblázatot kellett elkészíteni, mely minden egyes paradigmakódhoz megadja, hogy a `morphdb.hu` szótárban az adott csoport milyen jegyeket kap. Bár a paradigmák nem veszik figyelembe a tárgyas és nem tárgyas igék különbségét, maga a szótár tartalmazza ezt az információt, amelyet ily módon egyszerűen át tudtunk venni.

Az FKP-szótár bejegyzései a lemmákhoz tartozó morfológiai információt nem paradigmabesorolással, hanem a `morphdb.hu` szelleméhez közelebb álló módon tartalmazza. Például egyes mezők egy toldalék vagy toldalékcsoport allomorfjai közül való választást adják meg (pl. tárgyeset, *-at*, vagy *t*), mások közvetetten absztrakt tulajdonságokat (pl. tőtípus) specifikálnak. Az FKP szótár átalakítását úgy végeztük, hogy egy táblázatban minden mezőértékhez megadtunk egy `morphdb.hu`-beli jeghalmazt, amelyet a mezőérték jelenléte esetén a szótári bejegyzéshez asszociáltunk.

Mivel a lemmákhoz csak a minimális, nem megjósolható információt kívántuk tárolni, az átalakítás részeként a nyelvtan által a szavakhoz rendelt, és így a szótárban redundáns jegyeket töröltük. Ezt a `morphdb.hu` szűrőinek alkalmazásával tettük meg: ha egy szűrő egy bizonyos dimenzió jegyeit helyesen rendeli a szóhoz, akkor a szótárból annál a szónál töröljük az illető jegyeket.

A szótárak összefésülése A három forrásszótár szóanyagának átalakítása és redundanciamentesítése után a közös szókincs kiszűrését kellett megoldanunk. A mindhárom szótárban előforduló szavak száma 28 ezer, és tízezres nagyságrendű a páronként közös, illetve a csak az egyik szótárban előforduló szavak száma. A három szótárat összevonva jelenleg a lemmák száma 150 ezer körüli. Megfigyelhető, hogy az egyes szótárak egyedi hozadéka szintén 10 ezres nagyságrendű (az `Ispell` 70 ezer egyedi tétele a hatalmas tulajdonnévtárnak köszönhető, amit a többi szótár nem tartalmaz), vagyis bármelyik nélkül lényegesen szegényebb lenne a szóállományunk. Az egyes forrásállományok bejegyzéseinek, valamint az átfedő tételek számát mutatja az 1. táblázat.

1. táblázat. A forrásszótárak számokban

SZÓTÁR	BEJEGYZÉSEK SZÁMA
Ispell	105580
Elekfi	67047
FKP	68316
Ispell $\cap$Elekfi	32898
Ispell $\cap$ FKP	30754
Elekfi $\cap$ FKP	54607
Ispell \ (Elekfi $\cup$ FKP)	70591
Elekfi \ (Ispell $\cup$ FKP)	8155
FKP \ (Ispell $\cup$ Elekfi)	11568
Ispell $\cap$ Elekfi $\cap$ FKP	28663

Az összefésülésnél ellenőrizni is tudtuk az egyes átalakításokat, mivel a több szótárban szereplő szavak esetében ugyanolyan bejegyzéseknek kell kijönniük. Ez azonban az esetek kis részében fordult csak elő. A többi esetben a különbségek egyik oka az volt, hogy az egyes erőforrások másként kezelnek egy-egy szót, más-más alakjait tartják helyesnek. Ez különösen azokban az esetekben fordult elő, ha az egyik szótár egy szót (bármilyen szempontból) ingadozónak tüntet fel.

A különbségeket automatikusan csoportosítottuk – egy-egy csoportba azok a szavak kerültek, amelyekhez azonos módon rendelődött többféle jegyhalmaz. Például azok a szavak, amelyhez az egyik forrás csak a *preverb* jegyet egy másik pedig a *preverb*, *trans* jegyeket rendelte egy csoportba kerültek. Ezeket a csoportokat kézi feldolgozással átnézve alakult ki a nyílt tokenosztályok szótárának végső állapota.

Egyéb kategóriák A névmások és névutók nem kerültek közvetlen módon átvételre, ezeket főként a nyelvtanon belül külön jegyek és toldalékolási folyamatok segítségével kezeljük. A forrásszótárak összes egyéb szófajú bejegyzését kézzel átnéztük és újra szófajokba soroltuk őket. Ezek a szófajok nem feltétlenül követik a források besorolását, és gyakran a forrásokban is ellentmondásos besorolással szerepelnek. A 2. táblázat a `morphdb.hu` főkategóriáit és a `hunlex` által kompilált kimeneti szótárban a hozzájuk tartozó bejegyzések számát adja meg.

Morfológiai annotáció A `morphdb.hu` természetesen tartalmazza azt az információt, amellyel a kimenetet használó morfológiai elemző a szavakat címkézi. A szótári adatbázist úgy késztettük el, hogy az egyes morfémákhoz tartozó morfoszintaktikai tagek változtathatóak legyenek, vagyis az elemző kimeneti annotációja rugalmas. Jelenleg a `morphdb.hu` az ún. KR-kódolást [5] támogatja. A KR jelölés a morfoszintaktikai tulajdonságoknak a jelöletlenséget jól kifejező, hierarchikus gráfreprezentációja.

2. táblázat. A morphdb.hu főkategóriái

FŐKATEGÓRIA-CÍMKE	FELOLDÁS	BEJEGYZÉSEK SZÁMA
NOUN	főnév	88026
ADJ	melléknév	17514
VERB	ige	12549
ADV	határozószó	1932
UTT-INT	mondatszó/interjekció	498
CONJ	kötőszó	258
NUM	számnév	209
DET	determináns	164
POSTP	névutó	146
PREV	igekötő/igevivő	132
ONO	hangutánzó szó	96
PUNCT	központozás	28
PREP	prepozíció	14
ART	névelő	2

A szótár egyéb információi Mindhárom forrás a toldalékolásra vonatkozó gazdag morfológiai leírás mellett számos egyéb hasznos információt tartalmaz, pl. témakör (Magyar Ispell), belső szerkezet (Elekfi), stiláris és használati információ (FKP), amelyet a `morphdb.hu` adatbázisában is megőriztünk.

Ezen kívül a névmások és a határozószók külön alkategóriákba lettek sorolva. Az alkategóriára vonatkozó információk a `hunlex` beállításával a kimeneti állományokba kerülhetnek, s ezáltal az elemző tetszőleges annotációs eszközként is használható.

3. Konklúzió

A `morphdb.hu` szóanyagának és nyelvtanának pontosságát és fedettségét a kézzel taggelt Szeged Korpusszal [1] való összehasonlítással ellenőriztük. A korpuszban magadott MSD-kódokat a `morphdb.hu` kimeneti annotációjára [5] alakítottuk át egy konverziós táblázat segítségével, majd a korpusz szavait leelemeztük a `morphdb.hu` erőforrást használva is. Így a kézzel és az elemzővel taggelt változat összehasonlíthatóvá vált. A két elemzés közötti eltérések egy része a készítők eltérő nyelvészeti felfogásából ered. Az összehasonlítás érdekében ezeket a konkrét elemzéseket átsoroltuk, azaz az MSD-kódokat nem a neki megfelelő `morphdb.hu` kódra (ha volt ilyen[57]), hanem az általunk helyesnek tartott elemzésre cseréltük. Például az *is* szó tagjének (`Cccp`) „fordítása” `CONJ` (kötőszó), a `morphdb.hu` azonban `ADV`-ként tartja számon. A névutók, névmások kapcsán sok ilyen különbséget szisztematikusan kezeltünk. E folyamat közben a Szeged Korpusz számos hibásan taggelt szavát is sikerült azonosítanunk.

Az átalakítások után az elemző jellemző fedettsége 90%-os, a látszólag magas 10% elemzetlen alak közül a legtöbb tulajdonnév, rövidítés, illetve olyan névszók, amely az Elekfi és FKP szótárokban egyszerűen a koruknál fogva nem szerepelnek (pl. *rendszergazda*, *internetező, limit, fájl*). Már az első teszteléstől fogva ezekkel a szavakkal bővítjük a `morphdb.hu` anyagát, végső célunk a közel 100%-os fedettség elérése.

A `morphdb.hu` pontosságának mérése nehezebb feladat, hiszen még a korpuszban előforduló szóalakoknak is számos a korpuszban nem szereplő alternatív elemzése lehet. Emiatt a pontosságot a korpusz segítségével csak azzal tudjuk tesztelni, hogy a `morphdb.hu` erőforrással kiadott elemzések között szerepel-e a korpusz szerint helyes elemzés.[58] Az ilyen hiányzó elemzés jelenleg a szóalakok kevesebb, mint 1%-ánál fordul elő, és ezek is ritka szavak, hiszen tokenszázalékban ez a korpusznak hozzávetőleg 0.1%-a. Célunk az ilyen hibák lehetőleg teljes kiküszöbölése. Ez ezután is szakértői munkát igényel, hiszen szem előtt kell tartani a `morphdb.hu` kategorizálásának a Szeged Korpusztól való szándékolt eltéréseit.

A jövőbeni terveink között szerepel a `morphdb.hu` további kézzel egyértelműsített szövegekkel (pl. Nemzeti Szövegtár) történő tesztelése és ezzel párhuzamosan a szótár bővítése. Számos tulajdonnévlista (magyar vezetéknevek, cégne-

[57] Pl. a *-hAt* morfémát az MSD-kód nem jelöli, a `morphdb.hu`-ban pedig inflexióként szerepel.

[58] Ezzel persze a túlelemzések nem kezelhetők, azok szűrésére más módszert kell találnunk.

vek, földrajzi nevek) is rendelkezésünkre áll, amelyeknek `hunlex` formátumra történő alakítása nem ütközhet nagy nehézségbe. Fontos tervünk a szóanyag normatív ellenőrzése is, vagyis a szubsztandard (helyesírású) alakok pontosabb megjelölése, hogy az adatbázisból egy a Magyar Ispell szótár minőségét elérő helyesírás-ellenőrző erőforrást lehessen generálni. Tervünk a szóösszetételi modul finomítása is, amely jelenlegi formájában igen megengedő.

Köszönetnyílvánítás

A `morphdb.hu` létrejöttében sokan segítségünkre voltak. Külön köszönet illeti Kornai András, Németh László és Varga Dánielt. Köszönet a Magyar Telecomnak a projekt anyagi és infrastrukturális támogatásáért.

Bibliográfia

1. Csendes Dóra, Hatvani Csaba, Alexin Zoltán, Csirik János, Gyimóthy Tibor, Prószéky Gábor, Váradi Tamás. Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In II. Magyar Számítógépes Nyelvészeti Konferencia, 238–245. Szegedi Tudományegyetem, 2003.
2. László Elekfi. Magyar ragozási szótár. MTA Nyelvtudományi Intézet, Budapest, 1994.
3. Mihály Füredi, András Kornai, and Gábor Prószéky. A szo1ta1r adatbázis. Kézirat, 2004.
4. Charles F. Hockett. Two models of grammatical description. Word, 10:210–234, 1954.
5. András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, and Viktor Trón. általános célú morfológiai elemző kimeneti formalizmusa. In II. Magyar Számítógépes Nyelvészeti Konferencia, 172–176. Szegedi Tudományegyetem, 2004.
6. László Németh. Magyar Ispell – Válasz a Helyes-e?-re. In IV. GNU/Linux szakmai konferencia, pages 99–107. Linux-felhasználók Magyarországi Egyesülete, 2002.
7. Péter Rebrus. Morfofonológiai jelenségek [morphophonological phenomena]. In Ferenc Kiefer, editor, Strukturális magyar nyelvtan. 3. Morfológia. [Hungarian structural grammar. 3. Morphology.], 763–948. Akadémiai Kiadó, Budapest, 2000.
8. Viktor Trón. Hunlex - morfológiai szótárkezelő rendszer. In II Magyar Számítógépes Nyelvészeti Konferencia, 177–182. Szegedi Tudományegyetem, 2004.
9. Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In Proceedings of the ACL05 Software Workshop. Ann Arbour, 2005.