Különírás-egybeírás – automatikusan

Ludányi Zsófia[1, 2], Miháltz Márton[2], Hussami Péter[3]

[1] ELTE BTK Nyelvtudományi Doktori Iskola
[2] MTA Nyelvtudományi Intézet, Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
[3]Alkalmazott Logikai Laboratórium

{ludanyi.zsofia, mihaltz.marton}@nytud.mta.hu, hussami@all.hu

Kivonat: Jelen tanulmány a helyesiras.hu automatikus helyesírási tanácsadó rendszer külön- vagy egybeírással foglalkozó webes alkalmazását mutatja be. A modul attribútum-érték struktúrás környezetfüggetlen nyelvtani elemzésen alapszik. Az elemzés morfológiai és szemantikai tulajdonságokra támaszkodik. A rendszer általános működésének, illetve a nyelvtani elemző felépítésének és működésének ismertetése után a rendszer egyik fontos alappillérét képező formális nyelvtan részletes bemutatása következik. Végezetül néhány bonyolultabb helyesírási probléma nyelvtechnológiai megoldását ismertetjük példákkal illusztrálva (mozgószabályok, szótagszámlálási szabály stb.).

1 Bevezetés

Az MTA Nyelvtudományi Intézete 2009 óta dolgozik egy olyan szakértői rendszeren, amely nyelvtechnológiai eszközök segítségével kísérel meg a felhasználók helyesírási kérdéseire automatikus választ adni (Miháltz et al 2012, Pintér et al 2009). A helyesiras.hu névre keresztelt készülő tanácsadó portál hét különböző helyesírási területen próbál interaktív segítséget nyújtani: külön- és egybeírás, helyesírás-ajánló, elválasztás, tulajdonnevek írása, számnevek helyesírása, keltezés, betűrendbe sorolás. A felhasználónak a nyitóoldalon felkínált menüből kell kiválasztania, hogy milyen típusú helyesírási kérdésre szeretne választ kapni (1. ábra).

Jelen tanulmány célja a helyesiras.hu projekt külön- és egybeírással foglalkozó moduljának bemutatása.

1. ábra. A helyesiras.hu nyitóoldala

1.1 A külön- és egybeírás problémája

A magyar helyesírás egyik legproblematikusabb kérdésköre a különírás és az egybeírás. A szabályok megfelelő alkalmazása némi grammatikai alaptudást igényel, mivel a helyesírás rendszerszerűségét a magyar nyelvtan szabályai alakították ki. Különbséget kell tudni tehát tenni a szószerkezetek, illetve szóösszetételek között. Ez sok esetben problémás, mivel a két nyelvtani kategória között sokszor nem éles a határ, gyakran fordulnak elő nem egyértelmű, többféleképpen megítélhető esetek (Laczkó–Mártonfi 2004).

Az összetétellé válás oka sok esetben például a jelentésváltozás. Olyan esetekben, amikor az adott kifejezés konkrét és elvont jelentésben is szerepel (s a megfelelő írásmódot éppen ez dönti el), emberi beavatkozás nélkül nem adható egyértelmű megoldási javaslat. A különírás-egybeírás sok egyéb területére is jellemző ez. Kimondható, hogy jelenleg nem lehetséges olyan helyesírás-ellenőrző, helyesírási tanácsadó alkalmazás kifejlesztése, amely teljesen önállóan, az ember anyanyelvi kompetenciáját segítségül hívó beavatkozás nélkül képes a külön- és egybeírás minden területét hatékonyan kezelni (Pintér et al. 2009).

1.2 A helyesiras.hu újszerűsége

A létező online helyesírási tanácsadók egyszerű szójegyzéken alapulnak, és csak akkor adnak kielégítő eredményt, ha a beírt szó eleve helyesen van leírva, és megtalálható a rendszer mögött álló szótárban (Pintér et al. 2009). Léteznek olyan szójegyzék alapú online tanácsadók, amelyek bizonyos hiányos bemeneteket is elfogadnak (pl. a www.magyarhelyesiras.hu online szótár elfogadja a *j-ly* cserével keletkezett hibás bemeneteket, ékezet nélküli alakokat), de a pusztán szótár alapú megközelítés nem elég hatékony.

A helyesiras.hu külön- és egybeíró modulja mögött ezzel szemben egy formális nyelvtan áll, amelyet felhasználva a kifejlesztett nyelvtani elemző létrehozza a megadott bemenetből generálható lehetséges jó megoldásokat.

2 A különírás-egybeírás webalkalmazás felépítése

2.1 Általános felépítés

A rendszer felhasználókkal történő interakcióját szemléltei a 2. ábra, a 3. ábra pedig moduljainak kapcsolatát.

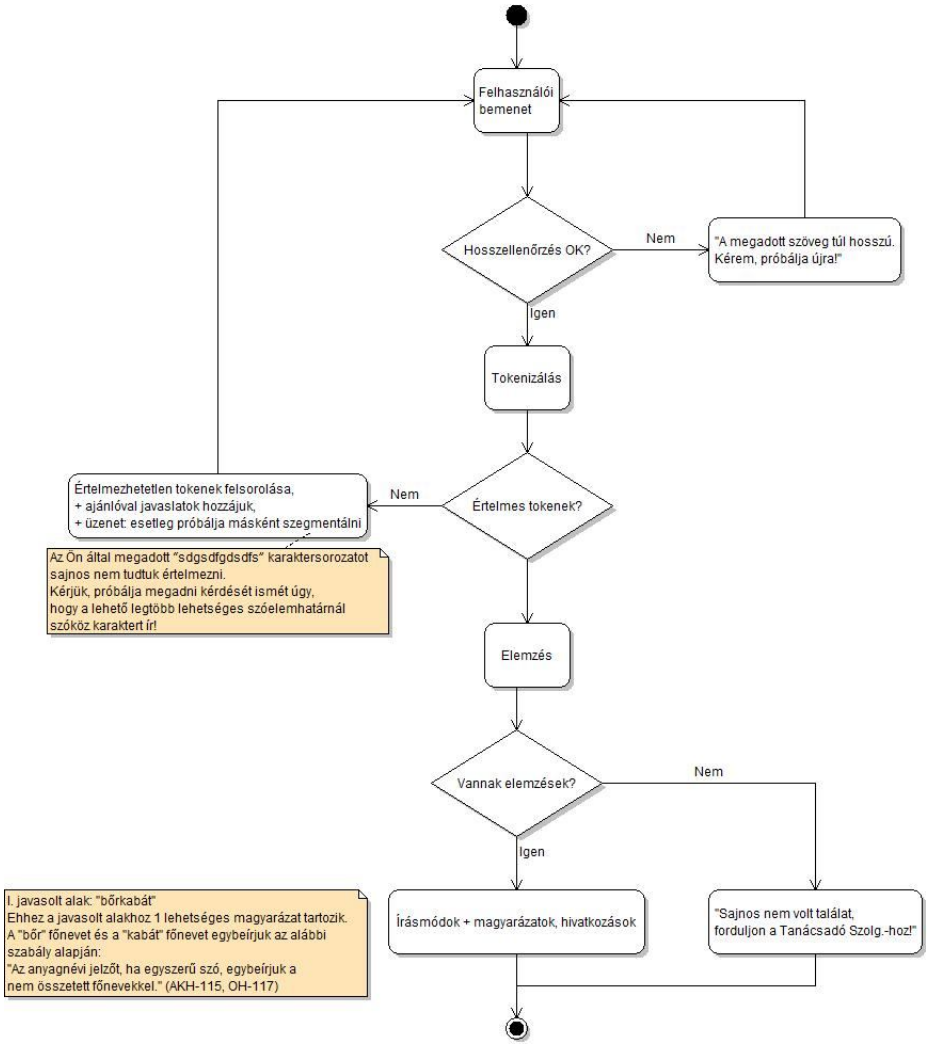
A felhasználói bemenetet néhány egyszerűbb ellenőrzésnek vetjük alá. A beírt szöveg hossza legfeljebb 70 karakter lehet; amennyiben ennél hosszabb bemenetet kapunk, a rendszer hibaüzenettel válaszol. Ha a karakterhossz megfelelő, következik a tokenizálás: a rendszer megkísérli tokenekre bontani a bemenetet.

A tokenizáló modul eltávolítja a felhasználó által megadott kötőjeleket (ha vannak), azokat szóközre cseréli, és az így kapott elemeket próbálja atomi szintű tokenekre bontani, illetve a HuMor morfológiai elemzővel (Novák–Pintér 2009) értelmezni (szófaji, morfológiai információkkal ellátni). Ha a tokenizálás sikertelen (nem sikerült minden tokent a morfológiai elemzővel azonosítani, illetve tovább-bontani), a rendszer megkéri a felhasználót, hogy újból adja meg a kívánt bemenetet, az összes lehetséges helyen szóközzel elválasztva. Sikeresen értelmezés esetén következik az elemzés, ellenkező esetben újabb hibaüzenetet kapunk, illetve az oldal átirányítja a felhasználót a *Helyes-e így*? és a *Névkereső* modulokhoz.

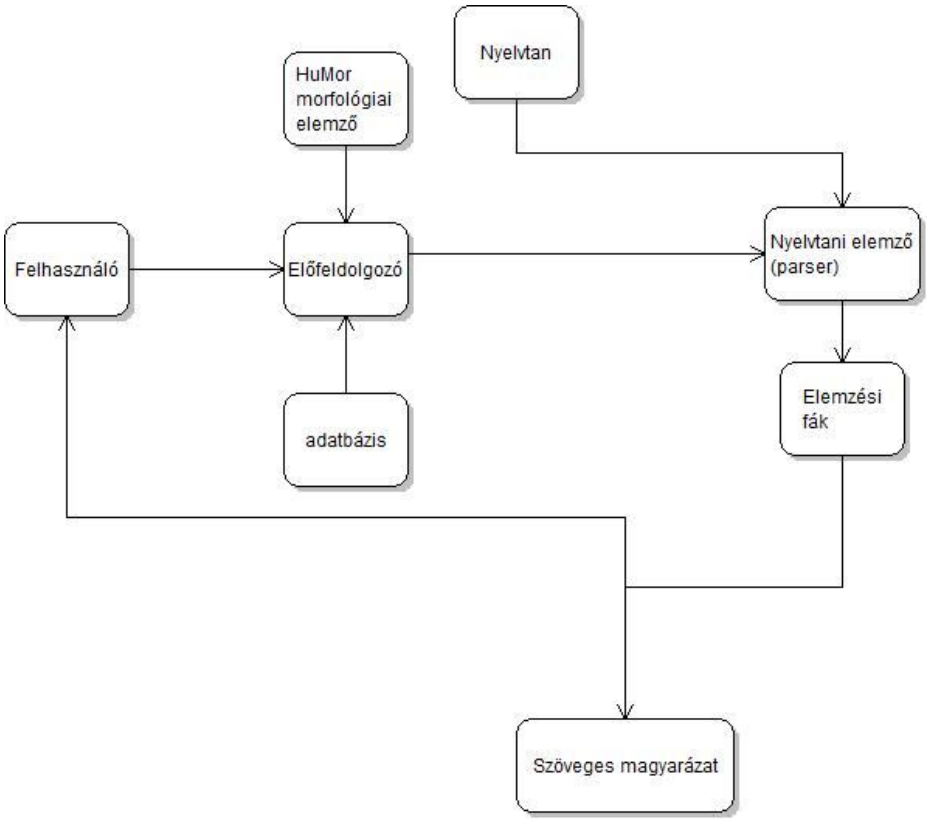
Amennyiben a tokenizálás sikeres volt, és megtörtént a bemenet morfológiai elemzése (szófaji, morfológiai, szótagszámra és összetételi tagok számára vonatkozó információkkal történő ellátása), az adatbázisban eltárolt szemantikai kategóriákat is hozzárendeljük a tokenekhez (ha vannak). Ilyen szemantikai kategóriák pl. a színnevek, foglalkozások és rangok, számnevek, földrajzi jellegű jelzők és köznevek, közterületek nevei, keresztnevek, népek és nyelvek nevei, rövidítések, közszói betűszók, önálló szóként nem használatos előtagok, a helyesírási szabályzatban az egyes szabályokban hivatkozott további kategóriák és különösen az egyes kivételek listája, melyeknek száma jelenleg 2100 körülire tehető.

A morfológiai, szemantikai tulajdonságokkal felruházott tokenek képezik az elemző modul bemenetét. Az elemző a mögöttes nyelvtanban található, speciális formális nyelven megfogalmazott helyesírási szabályokat próbálja alkalmazni a bemeneti tokenekre. Sikeres elemzés esetén megkapjuk a lehetséges megoldásokat a hozzájuk tartozó magyarázatokkal és az érvényben lévő akadémiai helyesírási szabályzat (AkH. (esetleg OH.)-beli hivatkozásokkal együtt. Ha az elemzés sikertelen,

azaz a megadott helyesírási szabályok egyike sem alkalmazható a bemenetre, a rendszer felajánlja a humán szakértői segítséget: a Nyelvtudományi Intézet Közönségszolgálatának telefonos vagy e-mailes igénybevételét (Mihaltz et al. 2012).



2. ábra. A rendszer általános felépítése



3. ábra. Az elemző felépítése

2.2 Az elemző felépítése és működése

Az elemző modul bemenetét az előző részben említett szegmentált, morfológiai és szemantikai jegyekkel ellátott tokenek képezik.

A szabályok illesztésére használt algoritmus lényegében egy bottom-up modell szerint működő elemző.

A célja az, hogy előállítsa az összes olyan szintaktikai fát, amely a bizonytalanságokat is tartalmazó bemeneti adatokból a rendelkezésre álló szabályok sorozatos alkalmazásával létrejöhet. A bemeneti tokenek fölé exponenciálisan sok fát lehet építeni, az algoritmus gyakorlati megvalósítása ezt optimalizálja. Az optimalizáláshoz felhasználunk egy ún. „kiértékelési lánc” konstrukciót. Ez ilyen lánc nem más, mint egy hipotézis a szabályalkalmazási sorozatra, nevezetesen, hogy melyik szabályt hol alkalmaztuk. Egy *n* argumentumú szabály alkalmazása helyettesíti a lánc *n* csomópontját egy eredménycsomóponttal. Egy láncon általánosságban több szabály is alkalmazható, így a lánc utódjai többen is lehetnek. Ha egy láncon nem alkalmazható több szabály, de a kilépési kritériumnak még nem felel meg, akkor az a lánc lekerül a jelöltek listájáról.

Kezdetben a lánc maga a tokenlánc. Kilépési kritériumnak azt választottuk, hogy a lánc egyelemű legyen, azaz egy teljes fát reprezentáljon. A láncok evolúciójának alapművelete a helyi szabályalkalmazás: az algoritmus a lánc összes csomópontján begyűjti, milyen szabályokat tudna ott alkalmazni. Ahhoz, hogy egy lánc *k*-adik elemén alkalmazhassuk az *n* argumentumú *X* szabályt, arra van szükség, hogy a lánc *k*-adik, *k+1*-edik, …, *k+n–1*-ik eleme megfeleljen *X* első, második, … *n*-edik elemének. A megfelelés szükséges és elégséges feltétele, hogy az adott csúcsok morfológiai címkéi egybeessenek, és a bemenet attribútumai megfeleljenek a szabályban megkövetelt feltételeknek.

2.3 A nyelvtan

A modul alapját képező környezetfüggetlen, jegystruktúrás nyelvtan formális leírása független a modul programkódjától, így könnyen karbantartható, fejleszthető.

A nyelvtan jelenleg mintegy 230 darab szabályt tartalmaz. Egy szabály felépítését az alábbi példán keresztül mutatjuk be:

```
id: M_EK_ANYAGNEV_1_2_1

rule: N(sem="Material", ncomparts>=2, type!="Qualificative") +
N(ncomparts=1) == N(sep=' ', hasnesep="1")

comment: Ha az anyagnévi jelzős kapcsolatnak valamelyik vagy mindkét
tagja össze tett szó, az anyagnevet különírjuk jelzett szavától.

refs: AKH-115, OH-117

ex: műbőr + kabát = műbőr kabát, nyersselyem + ing = nyersselyem ing

kill: M_EK_JELOLETLEN_BIRTOKOS
```

A kettőspontra végződő mezők jelentése a következő:

– Az **id** mező a szabály egyedi azonosítóját tartalmazza.

– A **rule** mezőben található maga az újraíró szabály (ennek kifejtését lásd alább).

– A **comment** mező tartalmazza a szabálynak a szöveggel történő megfogalmazását.

– A **refs** mezőben találjuk az AkH. megfelelő szabálypontjaira és/vagy az Osiris-helyesírás releváns témaköreire történő hivatkozásokat (az AkH. esetében szabálypontot, az OH. esetében oldalszámot).

– Az **ex** mezőben példákat találunk. Ezekre a szabályok automatizált tesztelésénél van szükség.

– A **kill** mezőbe (opcionális) azoknak a szabályoknak az egyedi azonosítóit tüntetjük fel, amelyek a a szabályalkalmazó algoritmus futtatásakor konkurensek lehetnek az adott szabályra nézve. Ilyenkor a kill mezőben megadott címkéjű szabályt letiltjuk, így szűkíthető a lehetséges jó megoldások halmaza (illetve az esetlegesen illeszkedő, de valójában nem jó megoldások is eltávolíthatók a kimenetről).

A **rule** mezőben található újraíró szabályok felépítése a következő:

(1) X(a=v, …) + … == Y(a=v, …),

ahol *X* a bal oldali, *Y* a jobb oldali szimbólumot jelenti; *a* az attribútum nevét, *v* pedig annak értékét jelöli. A bal oldali szimbólumokban az attribútumok és az értékek közötti operátorok értékvizsgálatot, a jobb oldalon értékadást jelentenek.

A leíró nyelvtan szimbólumai az angol szófaji kategóriák kezdőbetűi vagy -betűcsoportjai: N (főnév), A (melléknév), V (ige), Adv (határozószó), Num (számnév).

A szabályok bal oldalán a következő attribútumok állhatnak:

– **sem**: A token szemantikai tulajdonságait tartalmazza (egyszerre több értéke is lehet). Szemantikai tulajdonság például: színnév, anyagnév, foglalkozásnév stb. A példában szereplő *sem="Material"* attribútum-érték páros jelentése: a bemenetnek olyan tokennek kell rendelkeznie, amely rendelkezik a „Material", azaz anyagnév szemantikai jeggyel.

– **match**: Értéke egy reguláris kifejezés, amely illeszkedik a morfológiai elemző által előállított címkesorozatra. Ha például azt szeretnénk, hogy a folyamatos melléknévi igenevekre illeszkedjen a szabály, úgy tudjuk beállítani, hogy a *match* attribútumnak megadjuk a következő értéket:

[1] `match~"IGE,_OKEP",`

amelynek jelentése: egy ige és egy *-ó*/*-ő* képző. A match után szereplő ~ egy speciális operátor, amely reguláris kifejezések illeszkedését vizsgálja (az operátorokról lásd később).

– **wordform**: A bemenet felszíni alakja.

– **stem**: A felszíni alak töve.

– **ncomparts**: Értéke egy egész szám. Megadja, hogy hány összetételi tagból áll az adott szimbólumnak megfelelő token(rész)sorozat. Igazodva a szótagszámlálási szabály (AkH. 138.) előírásához, a két vagy több szótagból álló igekötők összetételi tagnak számítanak (pl. *ellen-*, *elő-*), míg az egy szótagos igekötők nem. Így például az *előadás* token *ncomparts* értéke 1, míg az azonos felépítésű *beadás*-é csupán 1 (azaz egyszerű – nem összetett – szó).

– **nsylls**: Értéke egy egész szám. Megadja, hogy hány szótagból áll az adott token. Az *ncomparts* jegyhez hasonlóan a szótagszámlálási szabály előírásainak megfelelően számolódik ki az értéke: az összetett szó jel és rag nélküli alakjának szótagszámát értjük alatta. Így pl. mind a *kerékpárjavítás*, mind a *kerékpárjavításnak* tokenek *nsylls* értéke 6. A képzők viszont már beleszámítanak a szótagszámba: a *kerékpár-javítási* alak *nyslls* értéke 7 az *-i* képző miatt.

– **ntoks**: A bemenetben megadott tokenek száma. Ha egy vagy több token összetett szó, a tokenek és az összetételi tagok száma eltérő lehet (*ncomparts* és *ntoks* értéke nem mindig egyenlő).

– **join1**, **join2**, **join3**: Ezen attribútumok a kivételes (nem formalizálható) írásmódú összetételek kezelésére szolgálnak. Az előfeldolgozás során, ha a tokenek felszíni alakjai valamilyen kombinációban szerepeltek a

kivételszótárban, megkapják értékül a kivétel kategóriáját (pl. *Jelentessurito*), így az adott kivételeket kezelő szabályok érvényesek lesznek rájuk.

– **type**: Bizonyos speciális típusú főnévi csoportok megjelölésére szolgáló attribútum.

– **sep**: Ez az attribútum kötelezően szerepel egy értékadásban minden szabály jobb oldalán, ahol a bemeneti tokenek közé kerülő szeparátort (üres sztring, szóköz, kötőjel stb.) jelöli.

– **ortho**: az adott elemzési lépésben, a sep attribútum segítségével kiszámított helyesen írt alak.

– **hasnesep**: Értéke egész szám. Azt jelzi, hogy a generált alak hány darab nem egybeírást jelző szeparátort tartalmaz (azaz hány szóközt, kis- vagy nagykötőjelet). Ha nem tartalmaz egyiket sem a felsoroltak közül, értéke 0. A szótagszámlálási szabályoknál van rá szükség: ezek a szabályok csak akkor hajtódnak végre, ha a generált alak egyáltalán nem tartalmaz kötőjelet vagy szóközt, vagyis egybeírt alakok esetén.

– **63exception**: Értéke ’YES’ lehet. A 6:3-as szabály alól kivételt jelentő írásmódú szavak megjelölésére szolgál. A 6:3-as szabályok csak akkor hajthatók végre, ha a 63exception attribútum értéke nem ’YES’ (azaz nem kivételes írásmódú szavakról van szó).

– **3idcons**: Ha új összetétel keletkezésekor három azonos mássalhangzó kerül egymás mellé, speciális szabályt kell alkalmazunk, ezen esetek megjelölésére szolgál ez a jegy. (Bővebben l. 3.5, 3.6 pontokban.)

A szabályok jobb oldalán az alábbi attribútumok állhatnak:

– **sep**: Az attribútum értéke a bemeneti szóelemek közé kerülő elválasztó karaktert kódolja. Értékei lehetnek:

- ’’ (üres sztring), amely egybeírást jelöl;
- ’ ’ (szóköz), amely különírást jelöl;
- ’-’, kötőjellel írást jelöl;
- ’--’ (kötőjel-kötőjel), nagykötőjellel írást jelöl;
- ’@ ’ (kukac-szóköz), amely az anyagnévi mozgószabály speciális szeparátora; a különírt anyagnévi jelzős szókapcsolatot alkalmilag egybeírja („összerántja”), és a jelzett szót különírva kapcsolja hozzá;
- ’@-’ (kukac-kötőjel), amely a második mozgószabály speciális szeparátora; a különírt minőségjelzős szókapcsolatot alkalmilag egybeírja („összerántja”), és az új összetételi utótagot kötőjellel kapcsolja hozzá;
- ’-@’ (kötőjel-kukac), amely a második mozgószabály speciális szeparátora; a különírt minőségjelzős szókapcsolatot alkalmilag egybeírja („összerántja”), és az új összetételi előtagot kötőjellel kapcsolja hozzá;
- ’-1’, ’-2’, …, -6, -n, amely egy többszörös összetétel 1., 2., …, 6. szava mögé kötőjelet szúr be (-n esetén az utolsó tag elé) – a 6:3-as szabályok használják;

- ’-#’, amelyet azon speciális szabályok használnak, amelyek az (állandó vagy alkalmi) összetételek keletkezésekor egymás mellé kerülő három azonos mássalhangzó esetét hivatottak kezelni: a ’-#’-jel kódolt szeparátor egy eredetileg már egybeírt összetételt „igazít ki” kötőjellel;
- ’$’, amely egybeírást jelöl, de csak azon speciális esetekben használandó, amikor egy *-sz*-re végződő szóhoz járul a *-szerű* ún. képzőszerű utótag, ilyenkor ugyanis a jelenlegi szabályozás szerint egyszerűsítést alkalmazunk (AkH. 94.): szsz helyett ssz-t írunk (*ész + szerű* > *ésszerű*), ezt az *szsz* > *ssz* cserét hajtja végre tulajdonképpen a ’$’ szeparátor.

Az attribútumok és értékeik között a szabályok bal oldalán többféle operátor szerepelhet. Ezek a következők:

= : Karakterláncok (stringek) közti egyenlőséget vizsgál.

!= : Karakterláncok közti nemegyenlőséget vizsgál.

~ : Az operátor jobb oldalán reguláris kifejezésnek kell állnia. Az *a~v* jelentése: az *a* attribútumnak illeszkednie kell a *v* reguláris kifejezésre. A ~ operátort legtöbbször a match attribútum használatakor alkalmazzuk (a match csak ily módon kaphat értéket, l. a match attribútum leírását); ritkábban máshol is előfordulhat. Például azoknál a szabályoknál, ahol a szóösszetételi határon előforduló három azonos mássalhangzó meglétét vizsgáljuk, szintén a ~ operátort alkalmazzuk (az ortho jegyre):

[2] `ortho~.*([bcdfghjklmnpqrstvwxy])\1\1.*"`.

<= : A bal oldali egész szám kisebb vagy egyenlő, mint a jobb oldali.

>= : Egész számok közti nagyobb vagy egyenlő, mint a jobb oldali.

< : A bal oldali egész szám kisebb, mint a jobb oldali.

> : A bal oldali egész szám nagyobb, mint a jobb oldali.

3 Néhány bonyolultabb szabály formális leírása

3.1 Anyagnévi mozgószabály

Az AkH. 1984. nem szól azoknak a szerkezeteknek az írásmódjáról, amikor egy eredetileg különírt szószerkezet (pl. *valódi bőr*) tölti be az anyagnévi jelző szerepét. Az ilyen esetekre vezette be az Osiris-helyesírás az ún. anyagnévi mozgószabályt, amely a következőt javasolja: az eredetileg különírt szerkezetet az anyagnévi mozgószabálynak megfelelően egybeírjuk, az alkalmi összetétellé váló jelző és a jelzett szó azonban különírandó (Laczkó–Mártonfi 2004: 134). Pl.: *valódi bőr + kabát > valódibőr kabát*.

A *valódi + bőr + kabát* bemenetek megadásakor első lépésben az első két token összevonódik egy különírt minőségjelzős szerkezetté. A szabály jobb oldalán beállítjuk, hogy a *type* attribútum értéke „Qualificative” legyen; ezzel azt jelezzük, hogy a további szabályok bemenetét képező *valódi bőr* főnévi csoport egy különírt (szóközt tartalmazó) szerkezet.

[3] `A() + N(sem="Material") == N(sep=' ', hasnesep="1", type="Qualificative")`

A bal oldalon szükséges egy olyan megszorítást tennünk a második bemenetre, hogy az anyagnév legyen; ezt a *sem* attribútum megfelelő értékadásával állítjuk be.

Mivel nyelvtanunkban – megállapodás szerint – a fej (a jobb oldalon az utolsó szimbólum) bizonyos jegyeit automatikusan megörökli a szabály jobb oldalán álló szimbólum, ha értékük specifikálva van (Miháltz et al. 2012), ezért jelen esetben a szabály által generált *valódi bőr* szimbólum is megőrzi a „Material" (anyagnév) szemantikai jegyet. Éppen ezért az anyagnévi mozgószabály bal oldalán az egyik feltétel – hogy az első bemenet anyagnév legyen – teljesülni fog. A szabály a következőképpen néz ki:

[4] `N(sem="Material1", type="Qualificative") + N() == N(sep='@ ', hasnesep="1")`

A szabály jobb oldalán a *sep* attribútum értéke a '@ ' szeparátor lesz, amely az eredetileg különírt szókapcsolatot összerántja (egybeírja), és az utótagot szóközzel elválasztva kapcsolja hozzá. A *hasnesep* attribútum értékét 1-re állítjuk, ami azt jelzi, hogy a generált szimbólum nem egybeírt szó, hanem tartalmaz legalább egy szóközt vagy kötőjelet.

Megemlítendő, hogy az ilyen típusú bemenetekre (*valódi* + *bőr* + *kabát*) az említett anyagnévi mozgószabályos alakulatok mellett egyéb szabályok legenerálják a *valódi bőrkabát* alakokat is (ahol a *valódi* jelző nem a *bőr*re, hanem a *kabát* szóra vonatkozik). Ez néha értelmes, néha kevésbé értelmes szókapcsolatot eredményez (pl. *edzett* + *acél* + *szerszám* > *edzett acélszerszám* [?]); ilyenkor a felhasználónak kell kiválasztania a kapott megoldások közül – a részletes magyarázatok segítségével –, hogy melyik forma áll közelebb az ő elgondolásához.

3.2 Anyagnévi jelzős szerkezetek, ahol az anyagnév jelzője melléknévképzővel ellátott földrajzi név

Az előbb ismertetett szabálynak előfordulnak olyan speciális esetei, amikor a különírt anyagnévi jelző nem vonható össze az anyagnévi mozgószabály végrehajtásával. Ez az eset akkor áll fenn, ha az anyagnév jelzője melléknévképzővel ellátott földrajzi név: *carrarai márvány*, *herendi porcelán*. Ilyenkor nem hajtható végre az anyagnévi mozgószabály, tehát nem írhatjuk: **carraraimárvány szobor*, **herendiporcelán étkészlet*, hanem teljes különírást alkalmazunk: *carrarai márvány szobor*, *herendi porcelán étkészlet* (Laczkó–Mártonfi 2004: 134).

Ebből következően szükség van egy olyan szabályra, amely az *-i* képzős földrajzi névből és az anyagnévből legenerálja azt a jelzős szerkezetet, amely együtt alkotja majd az anyagnévi szabály első bemenetét. A *match* attribútum értékét megszorítjuk, hogy *-i* képzős szóalakot fogadjon el, a sem értéke pedig „ProperGeo", azaz földrajzi név lesz. A második bemenetet az eddigiekhez hasonlóan anyagnév. A szabály jobb oldalán újdonság a *type* attribútum „ProperGeo" értéke. Ez azt jelzi, hogy olyan főnévi csoportról van szó, amelynek alaptagja (feje) anyagnév, bővítménye *-i* képzős földrajzi név (pl. *carrarai márvány*).

[5] `A(match~".*_IKEP,NOM", sem="ProperGeo") + N(sem="Material") == N(sep=' ', hasnesep="1", type="ProperGeo")`

Ezután már alkalmazhatjuk a következő szabályt a *carrarai márvány* + *szobor* bemenetekre:

[6] `N(type="ProperGeo") + N() == N(sep=' ', hasnesep="1")`

Vagyis az első bemenetet megszorítjuk, hogy csakis olyan főnévi csoportokat fogadjon el, amelyek *type*-értéke „ProperGeo", azaz különírt földrajzi névi jelzős

szerkezet. A jobb oldalon a *sep* attribútumnak ’ ’-t (szóközt) adunk értékül, ezzel jelezvén a különírást.

Hasonlóan az előzekhez, itt is előállnak az említett írásmódú formák mellett a *carrarai márványszobor* stb. típusú alakulatok. Ugyanígy, ha például a *velencei* + *arany* + *pénz* szavakat kapja bemenetül az elemző, legenerálja mind a *velencei arany pénz*, mind a *velencei aranypénz* alakulatokat. Noha valószínűsíthető, hogy a felhasználó a második esetre gondolt (a *pénz* jelzője az *arany*), az elemző felkínálja azt az eshetőséget is, hogy a *velencei arany* lehet egyfajta speciális anyagnév (mint a *carrarai márvány*), holott ilyen anyag nem létezik, de ha létezne, az említett módon – mindent külön szóba – kellene írni az anyagnévi jelzős alakulatot.

3.3 Második mozgószabály

Az AkH. 139. b) szabálya a következőképpen szól: ha egy különírt szókapcsolat (pl. hajlított bútor) olyan utótagot kap (pl. *gyár*), amely az egészhez járul, az egyébként különírandó előrészt az új alakulatban egybeírjuk, és ehhez az utótagot kötőjellel kapcsoljuk: *hajlítottbútor-gyár*.

A *hajlított* + *bútor* + *gyár* bemenetek megadásakor a következő történik: a minőségjelzős szerkezetekért felelős szabály összefűzi az első két tokent egy különírt szószerkezetté: *hajlított* + *bútor* > *hajlított bútor*. A *hajlított bútor* + *gyár* tokenek lesznek a második mozgószabály bemenetei.

Azt, hogy az első bemenet egy különírt szószerkezet, valamilyen módon meg kell jelölnünk. Ehhez a *type* attribútumot használtuk fel. A minőségjelzős szerkezetek helyesírásáért felelő szabályok jobb oldalára felvettük a type=”Qualificative” attribútum-érték párt.

[7] `A(match~".*,NOM", type!="Acronym") + N() == N(sep=' ', hasnesep="1", type="Qualificative")`

A minőségjelzős szerkezetekért felelős szabályok beállítják a *type* attribútumok értékét a megfelelő módon, innentől kezdve a további szabályok tudni fogják, hogy különírt szószerkezetről van szó.

A második mozgószabály formális leírása a következőképpen szól:

[8] `N(type="Qualificative", sem!="Material") + N() == N(sep='@-', hasnesep="1")`

Vagyis a jobb oldalon álló főnévi csoport egy különírt (minőségjelzős) szerkezet. A *sem* attribútum értékét azért kell figyelni, hogy ne legyen anyagnév, mert ha nem tennénk ezt a korlátozást, azokra a bemenetekre is végrehajtódna (párhuzamosan) a második mozgószabály, amelyekre az anyagnévi mozgószabályt kell alkalmazni. Például a *nyers* + *selyem* + *ing* bemenetre az egyébként helyes *nyersselyem ing* megoldás mellett megkapnánk a második mozgószabály alkalmazásával a *nyersselyem-ing* írásmódot, amely helytelen.

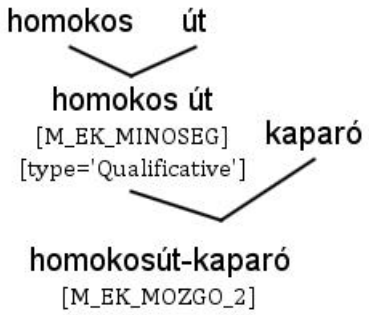
A szabály jobb oldalán a *sep* attribútum értéke a ’@-’ szeparátor lesz, amely az eredetileg különírt szókapcsolatot összerántja (egybeírja), és az utótagot kötőjellel kapcsolja hozzá. A jobb oldalon a *hasnesep* attribútum értékét egyúttal 1-re állítjuk jelzendő, hogy a szabály által generált szóalak egy darab kötőjelet tartalmaz.

A második mozgószabály „fordítva” is működik: a különírt szókapcsolat egészéhez nem utó-, hanem előtag járul. Ilyenkor az előtagot kapcsoljuk kötőjellel az alkalmilag egybeírt szerkezethez: *történelem* + *házi feladat* > *történelem-házifeladat*. Az ilyen típusú bemenetek feldolgozása az előzőekhez hasonlóan történik, a mozgószabályt

leképező újraíró szabály jobb oldalán ’@-’ operátor helyett az ellenkező irányban működő ’-@’ attribútumot alkalmazva, amely a különírt utótagot „rántja össze”.

Hasonlóan az anyagnévi mozgószabályhoz, az itt említett eseteknek is létezik másféle értelmezése s ezáltal írásmódja, pl. *érettségi + követelmény + rendszer > érettségi követelményrendszer* (a mozgószabályos *érettségikövetelmény-rendszer* mellett). Ez esetben a mozgószabály végrehajtása nélküli alakulat jelentése alig különbözik a mozgószabályosétól; az Osiris-helyesírás ilyen esetekben ezt az írásmódot is elfogadja (Laczkó–Mártonfi 2004: 132). Van, hogy a mozgószabály nélküli írásmód egészen mást jelent, olykor pedig egészen komikus-bizarr formákat eredményez: *csuklós + autóbusz + vezető > csuklós autóbuszvezető*, *homokos + út + kaparó > homokos útkaparó*, *ortopéd + cipő + készítő > ortopéd cipőkészítő*. Ismét hangsúlyozzuk, hogy a rendszer az összes lehetséges írásmódot felkínálja megoldásnak (nem törődvén az értelmezéssel); a megfelelő alak kiválasztása a felhasználó feladata.

A kétféle kimenetet szemléltetendő láthatjuk alább a kétféle értelmezés elemzési fáját (a végrehajtott szabályokat). A 4. ábra, illetve az alatta található egyszerűsített attribútum-érték páros leírás a kötőjeles, mozgószabályos megoldást, a 5. ábra és az azt követő leírás egy másik értelmezést demonstrál. A leírásokban egy-egy sor egy facsomópontnak felel meg, az első sor a fa gyökere. A csomópont szimbólum és attribútum-érték párjainak felsorolása után kettősponttal elválasztva a elemzési létrehozó szabály azonosítója.



4. ábra. A *homokosút-kaparó* megoldás elemzési fája

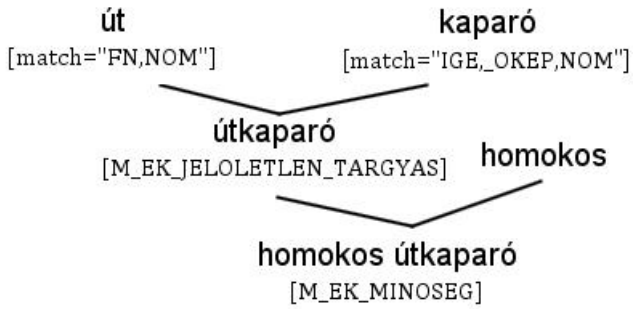
```
N(ortho="homokosút-kaparó", sep=['@-'], ntoks="3", hasnesep="1") :
M_EK_MOZGO_2_1

  N(ortho="homokos út", sep=[' '], ntoks="2", type="Qualificative",
hasnesep="1") : M_EK_MINOSEG_1_1_1

    A(A(wordform="homokos", match="MN,NOM", ntoks="1") : 0.

    N(wordform="út", match="FN,NOM") : 1.

  N(wordform="kaparó", stem="kaparó", match="FN,NOM”, ntoks="1"): 2.
```



5. ábra. A *homokos útkaparó* megoldás elemzési fája

```
N(ortho="homokos útkaparó", sep=[' '], ntoks="3",
type="Qualificative", hasnesep="1") : M_EK_MINOSEG_1_1_1

  A(wordform="homokos", match="MN,NOM", ntoks="1") : 0.

  N(ortho="útkaparó", sep=[''], ntoks="2") :
M_EK_JELOLETLEN_TARGYAS_1_2_1

    N(wordform="út", match="FN,NOM", ntoks="1") : 1.

    A(wordform="kaparó", match="IGE,_OKEP,NOM", ntoks="1") : 2.
```

3.4 Szótagszámlálási szabály (6:3-as szabály)

Az AkH. 138. pontja szerint a legalább három tagból álló többszörös összetételeket kötőjellel tagoljuk a fő összetételi határon, amennyiben a szótagszám meghaladja a hatot. A szótagszámba nem számítanak bele a ragok és a jelek, de a képzők igen. Külön összetételi tagnak számítanak a legalább két szótagból álló igekötők (pl. *ellen-*, *elő-*).

Ezen helyesírási szabály formalizálásához az *ncomparts*, illetve *nsylls* attribútumok értékeit kell vizsgálnunk. A 6:3-as szabálynak megfelelő újraíró szabály legtöbbször utolsóként hajtódik végre, miután egyéb (pl. birtokos jelzői alárendeléseket egybeíró) szabályok legenerálták a többszörös összetételt. Jelen esetben tehát egyargumentumú szabályról van szó, amelynek bal oldalán az említett attribútumok értékeinek kell megfelelniük a 6:3-as szabály feltételeinek. A szabály jobb oldalán a -1, -2, …, -n operátorokat használjuk az eredetileg már egybeírt szó kötőjellel való tagolására, illetve a *hasnesep* értékét 1-re állítjuk.

Léteznek bizonyos kivételes alakok, amelyek egybeírandók annak ellenére, hogy megfelelnek a 6:3-as kritériumoknak. Azt, hogy egy adott szó kivétel-e, a *63exceptions* jegy értékének vizsgálatával ellenőrizzük.

3.5 Három azonos mássalhangzó egymás mellett szóösszetételi határon

Előfordulnak olyan esetek, amikor a szóösszetételi határon három azonos mássalhangzó szerepelne egymás mellett az egybeírás eredményeként (**spicccipő*), ilyenkor kötőjellel tagoljuk az összetételt (AkH. 262. b). Ilyen esetekben a szabályok reguláris kifejezéssel írják le az egymás melletti három azonos mássalhangzót:

[9] `A(ortho~".*([bcdfghjklmnpqrstvwxy])\1\1.*", sep="", 3idcons!='YES') == A(sep='-#', hasnesep="1", 3idcons='YES')`

Ugyanígy, amikor a duplázott kétjegyű mássalhangzó az első tag végén, illetve ugyanaz a mássalhangzó a második tag kezdetén szerepel, külön-külön reguláris kifejezéssel vizsgáljuk az összes lehetséges kétjegyűbetű-kombinációt (a példában az *ssz* + *sz* egymás mellé kerülését).

[10] `N(ortho~".*sszsz.*", sep="", 3idcons!='YES') == N(sep='-#', hasnesep="1", 3idcons='YES')`

Ezek az egyargumentumú szabályok egy – egyéb szabályok által létrehozott – összetételt korrigálnak. Például a *hossz* + *számítás* esetében a birtokos jelzős alárendeléseket kezelő szabály létrehozza az egybeírt alakot (*hosszszámítás*), majd a 10-es példa szabálya ellenőrzi, van-e a reguláris kifejezésnek megfelelő illeszkedés az előző szabály által létrehozott írásmódban (*ortho* jegy értékében). Amennyiben igen, végrehajtja a szabályt. A jobb oldalon szereplő '-#' operátor szúrja be a kötőjelet az eredetileg egybeírt alakba, egyúttal beállítjuk a *3idcons* értékét 'YES'-re, amellyel azt jelezzük, hogy jelen esetben három azonos mássalhangzó került egymás mellé. A bal oldalon a *3idcons* értékvizsgálatával a szabály többszörös végrehajtását tiltjuk le.

3.6 Három azonos mássalhangzó egymás mellé kerül a második mozgószabály alkalmazása során

Mi történik olyankor, amikor egy különírt szószerkezet első tagjának dupla mássalhangzója, illetve a szószerkezet egészéhez kapcsolandó utótag első betűje megegyezik? Például: *barokk kép* + *restaurálás*, *festett tapéta* + *bolt*. A második mozgószabályt kellene alkalmazni, de annak végrehajtása (az eredetileg különírt szószerkezet „összerántása") három egymás melletti azonos mássalhangzót eredményezne (**barokkkép-restaurálás*, **festetttapéta-bolt*). Ilyen esetekben – az előzőekhez hasonlóan – a mozgószabály eredményezte alkalmi összetételben kötőjellel oldjuk fel a három azonos mássalhangzó egymás mellé kerülését, azaz összesen két kötőjelet írunk (Laczkó–Mártonfi 2004: 133). Helyesen: *barokk-kép-restaurálás*, *festett-tapéta-bolt*.

Hogyan modellezi le ezt a rendszer? A korábban ismertetett módon ilyenkor is végrehajtja a második mozgószabályt, és az így keletkezett, három egymás melletti azonos magánhangzót tartalmazó alkalmi összetételeket „igazítja ki" a következőképpen:

[11] `N(ortho~".*([bcdfghjklmnpqrstvwxy])\1\1.*", sep="@-", 3idcons!='YES') == N(sep='-#', hasnesep="1", 3idcons='YES')`

Ez a szabály majdnem megegyezik az előző pont 9-es példájával, a különbség csupán a bal oldali *sep* attribútum értékében van. Míg ott a *sep* értéke üres sztring, ami azt jelzi, hogy korábban egybeírás történt, itt a '@-' szeparátor szerepel a bal oldalon, jelezvén, hogy mozgószabály megkövetelte alkalmi egybeírás történt (nem pedig „sima" egybeírás).

3.7 *Sz*-re végződő szavak és képzőszerű utótagok alkotta összetételek

Az AkH. 11. kiadásának 94. pontja kimondja, hogy a *-szerű* ún. képzőszerű utótagot – helyesírásunk egyszerűsítő elvét érvényesítve – csonkítva (toldalék módjára) kapcsoljuk az sz végű szavakhoz: *ésszerű*, *mésszerű* stb. Az AkH. készülő 12. kiadásában ez változni fog: a képzőszerű utótag mint kategória meg fog szűnni, és a *-szerű* is szabályos utótagnak számít ezentúl: „Nem egyszerűsítjük […] az összetett szavak tagjainak határán találkozó azonos kétjegyű betűket: *kulcscsomó*, *jegygyűrű*, *nagygyűlés*, *fénynyaláb*, *díszszázad*, ***mészszerű*** stb." (MTA Magyar Nyelvi Bizottság 2010–2011: 26). Mindezek ellenére jelenleg még az egyszerűsítő írásmód van érvényben, ezért tanácsadó rendszerünkben is ezt kell követnünk.

Szükség van tehát egy olyan operátorra, amely *-sz* végű szavak és a *-szerű* utótag kapcsolásakor elvégzi az egyszerűsítést. Ez az operátor a már említett '$', amely az összetételi határon keletkező *szsz* > *ssz* egyszerűsítést végzi. A szabály első operandusa wordform attribútumának értékét kell figyelni, hogy (pontosan egy) sz-re végződik-e. Ezt reguláris kifejezéssel tehetjük meg.

```
[12]    N(match~".*,NOM",                    type!='Qualificative',
wordform~".*[^sz]sz$")      +      A(match~"MN",      stem="szerű",
type!='SuffixComPart') == A(sep='$', type='SuffixComPart')
```

Természetesen a dupla *sz*-re végződő szavakra nem vonatkozik ez az előírás, ott kötőjellel kapcsoljuk az utótagot: *dzsessz-szerű*. Ezek írásmódjáról külön szabály gondoskodik.

4 Összegzés

Tanulmányunkban bemutattuk a helyesiras.hu automatikus helyesírási tanácsadó rendszer külön- vagy egybeírással foglalkozó webes alkalmazását. A modul attribútum-érték struktúrás környezetfüggetlen nyelvtani elemzést alkalmaz, morfológiai és szemantikai tulajdonságokra támaszkodva. Mivel a magyar helyesírási szabályzat vonatkozó rendelkezései igen összetettek, bizonyos esetekben nehezen algoritmizálhatóak, a rendszer megalkotása során számos kihívással szembesültünk. Igyekeztünk több példán keresztül bemutatni megoldásainkat az egyes jelenségek kezelésére. Noha egy ilyen automatizált rendszer sosem lehet teljes és helyes, javításokra, kiegészítésekre, új jellemzők implementációjára szükség lesz a jövőben is, rendszerünk mostani formájában is jóval többet nyújt, mint a jelenleg létező hasonló megoldásuk, így reményeink szerint hatékony segítséget nyújthat a magyar nyelvet használók közösségének.

Irodalom

Laczkó, K., Mártonfi, A. 2004. *Helyesírás.* Budapest: Osiris Kiadó.

Miháltz, M., Hussami P., Ludányi Zs., Mittelholcz I., Nagy Á., Oravecz Cs., Pintér T., Takács D. 2012. Helyesírás.hu – Nyelvtechnológiai megoldások automatikus helyesírási tanácsadó rendszerben. In: Tanács, A., Vincze, V. (szerk.) 2012. *MSZNY 2013. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress, 135–147.

MTA Magyar Nyelvi Bizottság 2010–2011. *A magyar helyesírás szabályai.* 12. kiadás. Kézirat.

Novák, A., M. Pintér T. 2006. Milyen a még jobb Humor? In: Alexin, Z., Csendes D. (szerk.) 2006. *MSZNY 2006. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem, 60–69.

Pintér, T., Oravecz Cs., Mártonfi A. 2009. Online helyesírási szótár és megvalósítási nehézségei. In: Tanács, A., Szauter D., Vince V. (szerk.) 2009. *MSZNY 2009. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress, 172–182.

Pomázi Gyöngyi (szerk.) 2000. *A magyar helyesírás szabályai.* 11. kiadás, 12. (példaanyagában átdolgozott) lenyomat. Budapest: Akadémiai Kiadó.