Váradi Tamás: Gépesített helyesírási tanácsadás
Argumentum 9 (2013), 321-331
Debreceni Egyetemi Kiadó

Váradi Tamás

Gépesített helyesírási tanácsadás

Nyelvtechnológiával a helyesírásért

Abstract

The present paper introduces the online spelling advisor portal helyesírás.mta.hu. It describes its principles of operation, the set of language resources it utilizes, illustrates the algorithms it employs and suggests lines of future enhancements.
Keywords: spelling advisor, spelling rules, language use, language technology, corpus linguistics, norms of language use

1 A portál létesítésének háttere

Az MTA Nyelvtudományi Intézete alapítása óta társadalmi küldetésének tekinti a nyelvhelyességi tanácsadást. Egykor külön osztály, a Nyelvművelő osztály végezte ezt a tevékenységet, élükön olyan országos ismertséget szerzett nyelvészekkel, mint Lőrincze Lajos vagy Grétsy László. A nyelvhelyességi tanácsadás iránt változatlanul igen széles az igény, melynek az utóbbi időben egyre kevésbé tudott az Intézet a meglevő eszközökkel és csökkenő személyi állománnyal eleget tenni. A telefonszolgálat csak szűkös lehetőséget biztosít, és a hívás díja is korlátozó tényező. Az elektronikus levelezés, bár minőségi előrelépést jelent a hagyományos levelekhez képest, nem ad azonnali választ.

Ugyanakkor alapvetően megváltozott a környezet, amelyben az Intézet ezt a közszolgáltatást végzi. Az Internet megjelenése két szempontból is gyökeres változást hozott. Lényegesen kitárult az írásbeli nyilvánosság: a közösségi média, a blogok, fórumok, hozzászólások tömegesen adtak nyilvános megszólalást olyan embereknek, akiknek addig erre nem volt lehetőségük. Másrészt az Internet teremtette meg annak a lehetőségét is, hogy e technológia segítségével egyszerre tömegeknek lehessen nyelvi tanácsot adni.

A fent leírt szükség és a lehetőség vezetett ahhoz a felismeréshez, hogy az Intézet online helyesírási tanácsadó portállal egészítse ki nyelvi tanácsadó szolgáltatását. Nyilvánvaló volt azonban, hogy az Internet pusztán a hordozó közeget jelenti, a tanácsadó portálnak interaktívnak kell lennie, ami feltételezi magának a tanácsadásnak a számítógépesítését. Ehhez jelentős nyelvtechnológiai fejlesztés kell.

Váradi Tamás: Gépesített helyesírási tanácsadás
Argumentum 9 (2013), 321-331
Debreceni Egyetemi Kiadó

2 A szakmai kihívások

A helyesírásban eligazodni gyakran az embereknek is nehéz feladat. Ebben a részben először megvizsgáljuk, hogy milyen elvekre épül a magyar helyesírás, és milyen tanulási problémát jelent az embereknek, majd azt vizsgáljuk, hogy a számítógép számára hogyan jelentkeznek ezek a kérdések.

2.1 Mitől nehéz a helyesírás az embereknek?

A helyesírási rendszerünk (Magyar Tudományos Akadémia, 1984) „betűíró, latin betűs, hangjelölő és értelemtükröző" írásrendszerként definiálja magát (AKH. 2. paragrafus). A hangjelölő betűírás azt jelenti, hogy az írás elemi egységei hangokat rögzítenek, nem szótagokat vagy szavakat. Az értelem tükrözése pedig abban mutatkozik meg, hogy a szavak belső összetételét tiszteletben tartjuk, például az elválasztásban, valamint abban, hogy a különírás-egybeírás tükrözi a szemantikai hatóköröket (l. két emeletes ház vagy kétemeletes ház).

A helyesírási szabályzat két alappillére a hangzás szerinti ejtés és a szóelemző írásmód elve. A két elv egymással ellentétes hatású azokban az esetekben, amikor morfémahatáron mássalhangzó-hasonulás történik. Ez esetben a szóelemző írásmód elve érvényesül, amely a hasonulás ellenére is megőrzi az eredeti mássalhangzó alakját (l. *ad-ja,* ejtés szerint *aggya*). Az ejtés szerinti írásmód alapja a „helyesírás által is őrzött mai köznyelvi kiejtés (AKH. 17.). Ez a deklarált elv szándékosan figyelmen kívül hagyja a köznyelven kívül eső tájnyelvi vagy egyéb nem sztenderd változatokat, a beszélt nyelvi alakokat. Még az ily módon idealizált köznyelvi beszélt norma sem teljesen egységes. Az *i, u* és *ü* (valamint kisebb mértékben az *o és ö*) magánhangzók hossza ingadozást mutat sok, egyébként a köznyelvi normát követő személy beszédében (l. *színes, ágyú, ígér, körút, posta, turista* szavak ejtése).

A különírás és egybeírás témája helyesírásunk legbonyolultabb és legingoványosabb területe. Gyakran több tényezőt együttesen kell figyelembe venni a helyes alak megtalálásához, és ezek között kisebb szerepet kapnak az alaki jellemzők, mint például a szótagszám vagy a szóösszetétel. Döntő mértékben a kifejezések értelmezése, az elemek szintaktikai és szemantikai szerkezete, szemantikai osztályokhoz (pl. anyagnév, foglalkozásnév) való tartozásuk, valamint lexikalizált jelentésük szolgál alapul annak meghatározásához, hogy egy kifejezést külön vagy egybe kell írni. Ezek a szempontok a laikus nyelvhasználók számára nehezen értelmezhetőek, alkalmazásuk már csak ezért is bizonytalanságot szül.

2.2 A számítógépesítés nehézségei

Miután számba vettük a helyesírási nehézségek forrását az emberek számára, vizsgáljuk meg, hogy a számítógép szempontjából milyen kihívást jelent a helyesírás? Természetesen a számítógép „gyárilag" semmilyen nyelvi tudással nem rendelkezik, nem hagyatkozhatunk tehát arra a nyelvi kompetenciára, amelyet feltételezünk az anyanyelvi beszélőkről a helyesírási problémák megoldásában. Meg kell tehát határozni azt a tudást, amelyet kiindulásként be kell táplálni a rendszerbe ahhoz, hogy a számítógépes eljárás egyáltalán működésbe tudjon lépni. Nyilvánvalóan szükség van egy szótárra, amelyben szerepel az összes tőszó, valamint az összes szóösszetétel vagy több szavas kifejezés, amelynek a jelentése nem kompozicionális, azaz jelentésük nem állítható elő produktív szabályok segítségével az elemeik jelentéséből. Ezen felül szükség van egy morfológiai elemző programra, amely megállapítja a szóalak tövét, szófaját, a toldalékok morfológiai jellemzőit. Ideális esetben előállítja a felszini alak morfológiai derivációs történetét is, hogy az összetett és képzett alakokat vissza lehessen vezetni összete-

vőikre. Gyakorlatilag a szótár és a morfológiai elemző nem két független komponens: a morfológiai elemző saját szótárral működik, és annak mérete szabja meg, mire képes a számítógép. Ugyanis hiába van meg egy szó a szótárban, ha a morfológiai elemző nem ismeri fel, nem tud vele mit kezdeni.

Felmerülhet még az a lehetőség is, hogy szótár és elemző helyett pusztán egy óriási méretű korpusz, egy gigantikus méretű szóalakhalmaz álljon rendelkezésre. Ebben az esetben a helyesírás-ellenőrzés abból állhatna, hogy a számítógép megnézi, hogy a keresett kifejezés megtalálható-e ebben a korpuszban. Csábító az elképzelés, de rögtön belátható, hogy vannak korlátai is. Először is, a magyar esetében már az összetétellel és képzéssel előállítható szóalakok sokasága is milliárdos nagyságrendű. Ha ehhez hozzátesszük a külön írt többszavas kifejezéseket, akkor nagyságrendekkel nagyobb halmazt kapunk. Minél hosszabb kifejezést keresünk, annál nagyobb az esélye annak, hogy bármilyen óriási korpuszt veszünk is, nem tartalmazza majd az illető kifejezést. Két további korlát is felmerül. Az egyik az, hogy honnan vehetnénk ilyen méretű és száz százalékosan pontos adathalmazt? Természetes, azaz emberek által spontán írt szövegekből nem származhat, hiszen a referenciakorpusz célja éppen az emberek ingadozó helyesírásának javítása lenne. Az viszont egészen bizonyos, hogy ilyen tömegű *szerkesztett*, azaz szerkesztőségi ellenőrzésen átesett szöveg nem áll rendelkezésre. A másik korlát azonban az, hogy még ha lenne is egy ekkora méretű és minőségű korpuszunk, az nem tartalmazna semmi járulékos nyelvi információt az egyes szóalakokon kívül. Azonban már a morfológiai elemző kimenete is csak egy részét tartalmazza a jellemzők azon halmazának, amelyekre szükség van ahhoz, hogy meghatározhassuk a helyes alakot. Ráadásul önmagában a szóalak jólformáltsága nem elégséges ahhoz, hogy helyes is legyen. Gyakori ugyanis, hogy az *adott mondatban* vagy a szándékolt értelemben helytelen alak egyébként *más kontextusban vagy más értelemben* helyes megoldás. Gondoljunk az *igyekezet* főnév – *igyekezett* ige, a *helység – helyiség,* a *szép asszony – szépasszony* stb. különbségére. Általánosságban is megállapítható tehát, hogy a helyesírás az esetek nagyobb részében az adott *alak és jelentés* adott kontextusban érvényes helyességéről rendelkezik. Ennélfogva pusztán az alakokra hagyatkozni kilátástalan vállalkozás.

2.3 Milyen tudással oldható meg a feladat?

Nézzük tehát, hogy milyen tudással rendelkeznek az emberek? Kicsit idealizálva a képet, mondhatjuk, hogy a laikus beszélők egy kommunikációs igényeiket kielégítő méretű mentális szótárral és egy teljeskörűnek mondható morfológiai kompetenciával rendelkeznek. A szavak saját nyelvváltozatuk szerinti természetes kiejtésével, beszédük fonológiai és prozódiai jellemzőikkel tisztában vannak, és képesek azokat finom szemantikai különbségek megtételére is alkalmazni. Ugyanakkor alkalmanként tanácstalanok beszédük írásos képe felől. Milyen segítségre számíthatnak, ha helyesírási tanácsra szorulnak? A legfontosabb tudásforrás a helyesírási szabályzat és az azt kiegészítő szótár. A magyar helyesírás szabályai 14 fejezetben 299 pontban foglalja össze a szabályokat. A szabályok a legjobb igyekezet ellenére sem könnyen olvashatóak és főleg értelmezhetőek. Bár az egyes paragrafusok sok példát is tartalmaznak, azok végén szinte kivétel nélkül ott szerepel a *stb.* szócska. Legjobb esetben is csak sejtetésre alkalmasak. Ráadásul a szöveg maga is tele van hezitáló, bizonytalan megfogalmazással. Gyakoriak az olyan homályos jelentésű szavak mint *általában (68), gyakran (29), rendszerint (7).* A külön- és egybeírás témakör preambulumául szolgáló 95. paragrafus vége bevallottan teljes bizonytalanságot áraszt:

Váradi Tamás: Gépesített helyesírási tanácsadás
Argumentum 9 (2013), 321-331
Debreceni Egyetemi Kiadó

> A különírás és az egybeírás szabályai a szavak összekapcsolásának, illetőleg az összetett szavak alkotásának törvényszerűségein alapulnak. Helyesírásunk e területén mégis meglehetősen nagy számban vannak ingadozások, többféleképpen is megítélhető esetek. Ennek legfőbb oka az, hogy a szókapcsolatok és az összetételek között nincs éles határ: sok olyan szókapcsolat van, amely még nem igazi összetétel ugyan, de tagjai már alig tekinthetők egymáshoz képest önálló szavaknak. […] Ezekből a körülményekből következik az, hogy a különírás és egybeírás szabályainak megfogalmazása más szabálypontokhoz képest olykor határozatlannak látszik, bár valójában csak a hangos nyelv és az írás természetéhez alkalmazkodik.

A szabályzat mellett a másik segédeszköz a szótár. Ez olyan köznyelvi szavak és szókapcsolatok válogatását tartalmazza, amelyek „helyesírási szempontból eligazítást kívánnak" (i.m. 129.) A szótár szerkesztői feltehetően saját tapasztalatból előre látták a szótárhasználók bizonytalanságait, és egy-egy helyes alak idézésével igyekeztek nemcsak az adott alak helyesírását megadni, de egyben mintát adni hasonló típusú vagy hasonló összetételű kifejezések írásképére is. A szótár rengeteg összetett és többszavas kifejezést tartalmaz. Ha egy kifejezés különírva éppúgy helyes, mint egybeírva, csak éppen más jelentéssel, akkor a kifejezéspár elemei mellett egyértelműsítő jelentésmagyarázatok találhatók (l. *tanárfeleség 'tanár felesége'* és *tanár feleség 'olyan feleség, aki tanár'*).

Mind a szabályzat, mind a hozzá tartozó szótár alapvetően csak példákat szolgáltat, a listák sehol sem kimerítőek, még a szótár is a használók analógiás képességére bízza azt, hogy a szótárban szereplő alakok nyomán az ott nem szereplő hasonló alakok helyességét el tudják dönteni.[1]

A tudás, amelyet a laikus olvasótól várnak, vegyes: a szabályzat hivatkozik elvont, nehezen értelmezhető terminusokra, mint például *mellérendelő szókapcsolatok, minőségjelző*, *nyomósító szerepű melléknévi jelző,* valamint világos szemantikai osztályokra (népnevek, foglalkozásnevek). Fontos kiemelni, hogy mindkét típusú kifejezés a szavak kisebb-nagyobb (adott esetben akár végtelen halmazát) jelöli. Ezen kifejezésekkel arra apellálnak, hogy a használó mentális szótára tartalmazza például az összes foglalkozásnevet, felesleges tehát felsorolni őket, a felhasználó legalább felismeri őket, ha találkozik velük.

3 A helyesírás.mta.hu portál

A helyesírási tanácsadó portál olyan számítógépes rendszer, amelynek célja, hogy segítséget nyújtson a mindenkori helyesírási szabályzat szerinti helyes alak megtalálásához. Fontos hangsúlyoznunk, hogy nem csak a megoldást adja meg, hanem magyarázatul is szolgál, eligazítást ad a többértelmű azonos alakok használatában. Ez utóbbi jellemző megkülönbözteti a helyesírás-ellenőrzőktől, amelyek a szövegszerkesztő programokba beépítve kínálnak alakváltozatokat akkor, ha egy szót helytelen alakban találnak. Megjegyezzük, hogy a portál egyik szolgáltatása, a *Helyes-e így?* modul éppen ezt a helyesírás-ellenőrző funkciót nyújtja.

3.1 Funkciók

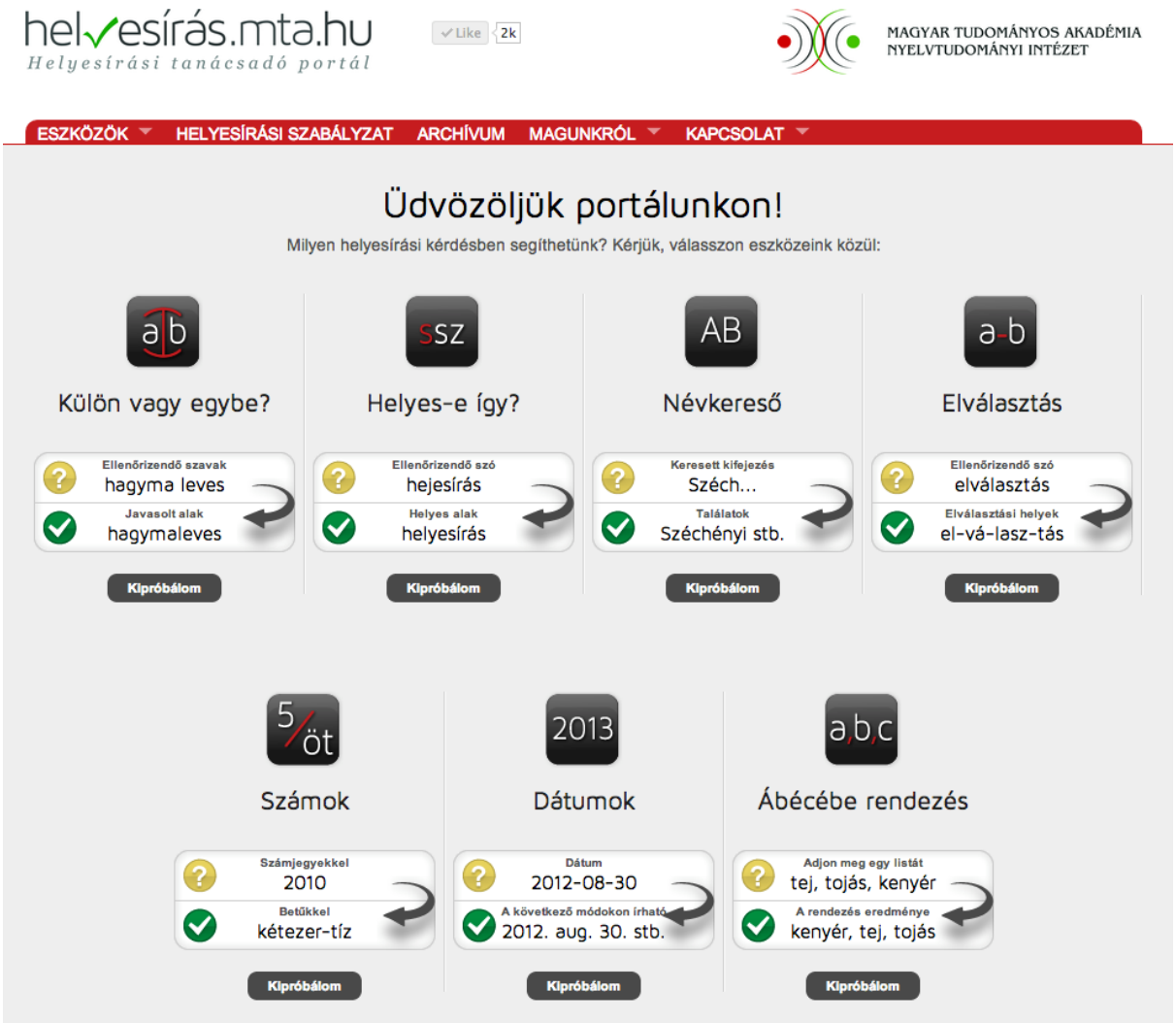
A portál egyelőre az alábbi hét témakörben szolgál segítséggel:

1. Külön vagy egybe?
2. Helyes-e így?
3. Névkereső
4. Elválasztás

[1] Igaz ez még arra a kétségkívül monumentálisnak mondható szótárra, amely az Osiris Kiadó által megjelentetett helyesírási tanácsadóban található. (Laczkó & Mártonfi 2005)

5. Számok
6. Dátumok
7. Ábécébe rendezés

A fenti témák felölelik a helyesírási szabályzat majdnem egészét: csupán az írásjelek használata, valamint a rövidítések és a mozaikszók kérdésköre maradt ki. Az írásjelek témája jelentős nyelvtechnológiai fejlesztést igényel, mert ehhez olyan részletes, mély szintaktikai elemző megléte szükséges, amely egyelőre nem áll rendelkezésre. A témák sorrendje tükrözi az Intézet nyelvhelyességi tanácsadást végző munkatársainak a tapasztalatát, melyet az első napok forgalma is megerősít. A portál összes oldalának látogatottságában a *Külön vagy egybe?* modul 34%-ot, a *Helyes-e így?* 24%-ot, míg a *Névkereső* 3,3%-ot tett ki.



1. ábra

3.2 Működési elvek

Az eddigi próbálkozások a helyesírási tanácsadás számítógépesítésére a szabályok+listák módszerének elektronikus megvalósítását jelentették.[2] Ez azt jelenti, hogy működésük többnyire kimerül abban, hogy a keresett kifejezés helyességét szótári egybevetéssel állapítják meg. Ha megtalálták a keresett kifejezést a szótárban, gyakran utalást is adnak a helyesírási szabályzat megfelelő paragrafusára. Ha viszont nincs találat, a felhasználó nem tudhatja, vajon a szótár hiányos, vagy tényleg helytelen a kérdéses alak.

A *helyesírás.mta.hu* alapjaiban más elveket követ. A leglényegesebb különbség, hogy a rendszer elemzi a keresett kifejezést, megkísérli azonosítani minden olyan jellemzőjét, amely a helyes alak eldöntéséhez szükséges, majd egy szabályrendszer segítségével azonosítja a helyesírási szabályzat vonatkozó pontját, és alkalmazza azt a helyes alak generálásában. Röviden tehát itt egy „bottom-up" megközelítés működik: a keresett kifejezés formai és helyenként szemantikai jellemzői alapján a rendszer egy szabályalapú algoritmus segítségével megtalálja azt a szabályt, amely leírja a megfelelő kifejezés helyes alakját. A rendszer részletesebb ismertetését l. Miháltz és mtsai (2013).

3.3 A felhasznált erőforrások

Egy ilyen összetett feladat számítógépesítéséhez számos nyelvi erőforrásra és eszközre van szükség. A rendszer motorja itt is a morfológiai elemző: a HUMor (Prószéky & Tihanyi 1993) és a Hunspell[3] eszközöket alkalmazzuk, és a kimeneteik unióját használjuk fel. A helyesírási szabályzat gyakran hivatkozik szemantikai osztályokra, mint például anyagot, színt, foglalkozást vagy népet jelentő szavakra. Ezt a tudást az ezekkel a jegyekkel ellátott szótárak segítségével tápláltuk be a rendszerbe. A szótáraknak egyenként fel kell sorolniuk a kérdéses szemantikai osztályba tartozó összes szót, mivel a számítógépes rendszer a szemantikai jegyekre nem tudja értelmezni a 'stb.' kifejezést.

Külön több százezer tételt tartalmaz a tulajdonnevek listája. A keresett kifejezés beírása során azonnal megjelenik a lista azon része, amely az addig beírt betűsorral kezdődik. Ez az inkrementális keresés nemcsak kényelmes, hanem hatékony is, mert előre láttatni engedi a betűsor lehetséges folytatását a listában.

Háttéradatforrásként természetesen rendelkezésre áll a Magyar Nemzeti Szövegtár (Váradi 2002), a Webkorpusz (Halácsy et al. 2004) és az Internet is. Közvetlen felhasználásuk azonban meggondolást igényel. Itt is felmerülnek ugyanis azok az adathelyességgel kapcsolatos aggályok, amelyeket a 2.2. pont alatt érintettünk.

Különös dilemmával találtuk magunkat szemben még olyan viszonylag korlátozott területen is, mint az intézmények és vállalkozások nevei, ahol a hiteles adatokat be lehet szerezni. Azt tapasztaltuk ugyanis, hogy a bejegyzett nevek nagy számban olyan írott alakban szerepelnek a cégbíróság nyilvántartásában, amely nem felel meg a helyesírás hatályos szabályainak. Egyelőre még nem találtuk meg az optimális megoldást, ezért az intézmény- és cégnevek nem szerepelnek a kereshető elemek listáján.

A hét modul közül egyedül a Névkereső esetében használtunk fel úgy listát, hogy a megoldást közvetlen szótári egybevetés szolgáltatja. A keltezés, a dátumok betűzése, az elválasztás és az ábécébe rendezés nyilvánvalóan procedurális feldolgozást kíván, és, amint hangsúlyoz-

[2] L. www.magyarhelyesiras.hu. Tudomásom szerint ezen az elven működik az Akadémiai Kiadó által üzemeltetett helyesírási weboldal.

[3] http://hunspell.sourceforge.net/

tuk, a portál a különírás-egybeírás kérdését, valamint a szavak helyes alakjának megtalálását is produktív, szabályalapú eljárással végzi.

3.3 A különírás-egybeírás megoldása

A különírás-egybeírás kérdéskörével még e rövid írás keretében is kicsit részletesebben kell foglalkoznunk. Ebben a témában mutatkozik meg ugyanis legvilágosabban a portálon alkalmazott nyelvtechnológiai megközelítés a listás megoldással szemben. A produktív, szabályokra épülő nyelvtechnológiai megközelítés már csak azért is indokolt, mert a nyelvhasználat olyan területéről van szó, ahol az elemek kombinációja olyan tömegű halmazt jelent, amelyet listával reménytelen követni. Másrészt viszont a nyelvhasználat is egyre változik, a szókapcsolatok és összetételek egyre újabb lexikai elemek között jönnek létre. Összekapcsolódásuk szabályai viszont változatlanok a konkrét lexikai elemektől függetlenül. Ezért is teljesebb megoldásnak tartható a szabályalapú rendszer. Az azonban fokozatos átmenet kérdése, hogy a szabályok milyen általános kategóriákkal operálnak, és milyen a hatókörük. A pusztán szófaji kategóriákkal definiált szabályok képezik az egyik végletet, az egyedi szavakat tartalmazók a másikat. A köztes területen találunk szemantikai osztályokra (pl. anyagnevekre) hivatkozó szabályokat, amelyek változó méretű halmazokat ölelnek fel. A helyesírási szabályzatban a különírás és az egybeírás kérdésére vonatkozó pontok (AKH. 95–142. pontok) egyaránt hivatkoznak mindháromfajta szabályra.

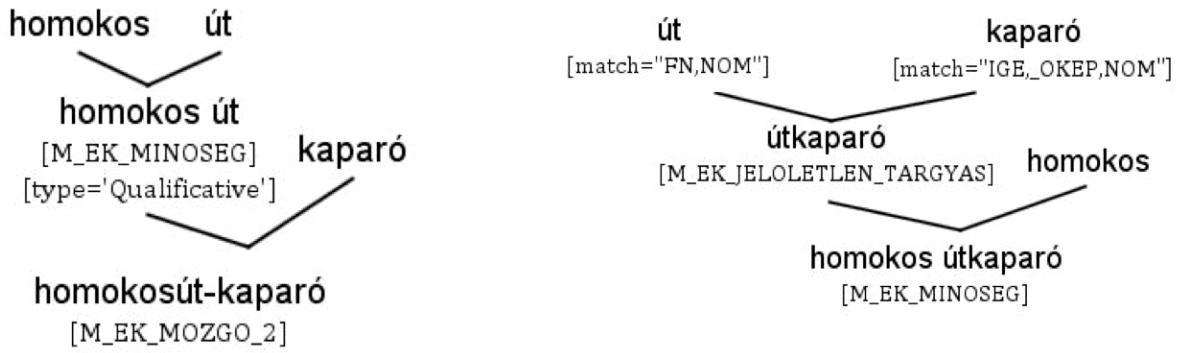
• A *Külön vagy egybe?* modul jelenleg nem tud kezelni minden jelenséget, csak azokat, amelyek algoritmizálhatóak (l. Kis Ádám áttekintését (Kis 1999), különösen az *Összefoglalás* részt, amelyet teljes egészében az AKH. 95. pontjának szentel[4]). A portál az alábbi jelenségekre igyekszik megoldást adni:

- jelölt és jelöletlen alárendelői összetételek/szintagmák,
- az ún. 6:3-as szabály,
- a mozgószabályok,
- a rövidítéseket és mozaikszókat tartalmazó összetételek,
- a szemantikai osztályokra vonatkozó szabályok (pl. színnévi összetételek, anyagnévi összetételek).

A modul a kérdéses kifejezést szóközökkel elválasztva kéri megadni, és a kimenet a helyesen (egybe-, külön- vagy kötőjellel) írt alakot adja meg, magyarázattal és a helyesírási szabályzatra utalással.

Az ily módon megadott kifejezést egy „bottom-up” környezetfüggetlen, jegystruktúrás nyelvtan dolgozza fel, amelyben morfológiai jegyek (szófaj, a morfológiai elemzés jegyei), alaki sajátosságok (szótagok és összetételi elemek száma), valamint lexikális szemantikai jegyek (anyagnév, színnév, népek és nyelvek nevei, rangok, keresztnevek stb.) szerepelnek. A rekurzív helyesírási szabályrendszer alkalmazása során a különböző derivációk elvezetnek a különböző ajánlott írásmódhoz, az elemzési fák bejárásából generálhatók a magyarázó szövegek.

[4] Meg kell jegyeznünk, hogy a cikk elsősorban a helyesírás-ellenőrző programok működése szempontjából ad hasznos áttekintést, és az 1999-ben keletkezett cikk érthető módon nem tárgyalja kellő mélységben az annotált nyelvi erőforrásokban, valamint a gépi tanulásban rejlő lehetőségeket, amelyek kitágították a számítógépes eljárások eredményességét a jelentés megragadásában.



2. ábra: A homokosút-kaparó és homokos útkaparó alakok elemzési fái (Miháltz és mtsai 2013)

A rendszer a 3. ábrán látható módon adja meg a helyes alakokat a magyarázatokkal és a helyesírási szabályzat megfelelő pontjaira történő utalással.

4 A társadalmi küldetés

A szakmai kérdések mellett szólnunk kell a portál társadalmi küldetéséről. A portál mottója: „Helyesírás mindenkinek!” Olyan eszközt szándékozunk kifejleszteni, amely mindenki számára könnyen elérhető, használható és főleg érthető. Az okostelefonok tömeges elterjedése reményeink szerint mindenki tenyerébe adja az univerzális helyesírási segédletet, és ezáltal jelentősen enyhítheti a stresszt, amit sokan éreznek a helyesírással kapcsolatban.

Az, hogy a portál remélhetőleg mindenkihez elviszi a helyesírást, nemcsak úgy értendő, hogy az okostelefonok és az Internet jóvoltából mindenkinek alkalma lesz azonnali választ kapnia helyesírási problémáira. Ezenfelül a mottó azt is feltételezi, hogy mindenki a saját természetes nyelvváltozatára hagyatkozva legyen képes használni a rendszert és megérteni annak tanácsait. A jelenlegi változat egyelőre ezt még csak töredékesen képes nyújtani. Ez jelenleg stratégiai cél, amely közvetlen összefüggésben van azzal, hogy a rendszer milyen bemenetet fogad el, és milyen információt ad a javasolt megoldásról.

4.1 A bemenet problémái

Ugyan a helyesírási szabályzat és különösen a kapcsolódó szótár tekintetbe veszi a feltételezett hibás alakokat, amelyeket a használók gyakran elkövetnek, mivel nem interaktív eszközök, nem feladatuk az, hogy módszeresen foglalkozzon a felhasználók által vélhetően beadott kifejezésekkel. Ez a kérdés azonban igen élesen vetődik fel a helyesírási tanácsadó portál számára. Itt ugyanis fel kell készülni mindenféle alakú kifejezésre, beleértve a köznyelvben ingadozó alakok (pl. *levő ~ lévő*) vagy esetleg a hibás alakokra is. Ha például valaki az „utban levő” szavak külön- vagy egybeírására kíváncsi, akkor a rendszer nem állhat le azért, mert a kifejezés egyik tagja nem helyes alakban szerepel.

Ennél átfogóbb és stratégiai fontosságú kérdés az, hogy a helyesírási szabályzat csak a köznyelvi beszélt norma szerinti ejtés helyesírásával foglalkozik. Azaz nem foglalkozik a köznyelvi beszélt sztenderdben is teljesen elfogadott laza artikulációjú változatokkal, amelyek gyakran hangkiesés vagy pótlónyújtás révén állnak elő, pl. *hónap* (= holnap), *mér* (=miért). Innen csak egy lépés a kötetlen stílusra jellemző alakok (pl. *mé* (=miért), *ne má (=ne már), tom* (=tudom)) kezelése.

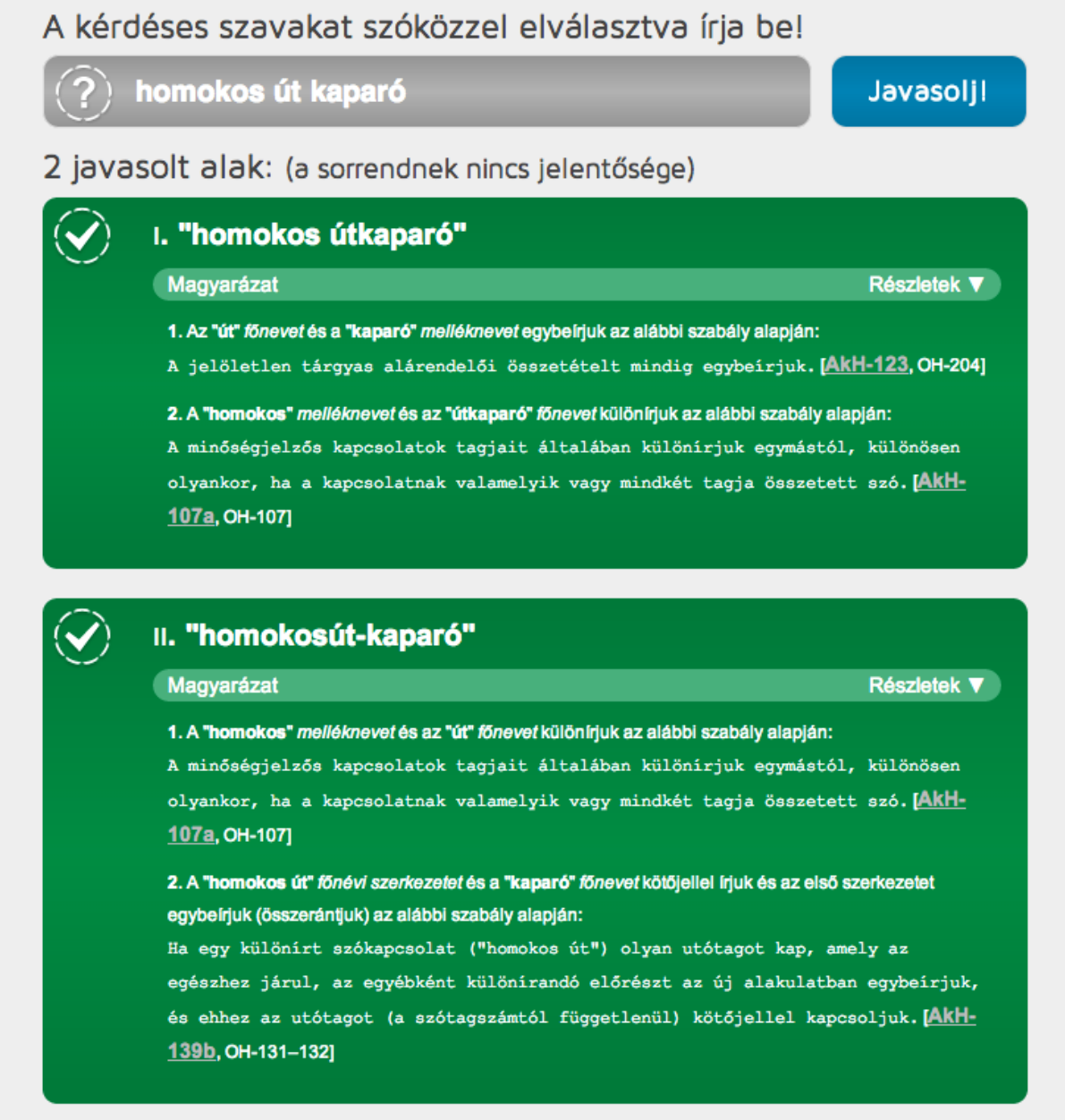
Még feltűnőbb a dialektális alakok teljes negligálása. Itt nem a dialektális alakok helyesírási norma szerinti leírására gondolunk, hanem arra, hogy keresési kifejezésként (azaz szótári címszóként vagy változatként) szerepeljen tájnyelvi alak is. Ha úgy tetszik, persze szigorúan véve itt kétlépcsős konverzióról van szó: úgy is lehet tekinteni a megfelelést, mint ami először a dialektális alak köznyelvi megfeleltetéséről szól, és aztán annak szabályos helyesírásáról. Az első mozzanat tulajdonképpen stiláris vagy szociolingvisztikai kérdés. Ha elfogadjuk a helyesírási szabályzat kiindulópontját, amely már eleve feltételezi a felhasználótól, hogy a köznyelvi beszélt norma szerint ejti a szavakat, akkor akár azt is elfogadhatjuk, hogy ez nem helyesírási kérdés. Ha viszont a helyesírási szabályzat (és különösen a tanácsadó) szerepét tágabb értelemben úgy definiáljuk, mint ami a hangzó beszéd és a helyes íráskép közti megfelelést definiálja, akkor gondoskodnunk kell arról, hogy ezen alakváltozatok is szerepelhessenek az input oldalon.

4.2 A kimenet kérdései

A rendszer által adott megoldásoknak és legfőképpen az azokhoz fűzött üzeneteknek stílusukban és nyelvezetükben igazodniuk kell a mottóban rejlő társadalmi küldetéshez, azaz mindenki számára érthetőnek és világosnak kell lenniük. Ezzel sem lehetünk még egyelőre elégedettek, mert jelenleg a rendszer magyarázatai és üzenetei még túlságosan támaszkodnak a helyesírási szabályzatra, gyakran szó szerint idézik azt. A szabályzat sem stílusában, sem terminológiájában nem olyan, amelyet a felhasználók széles köre minden gond nélkül megért és alkalmazni tud. Törekednünk kell tehát arra, hogy közérthetőbbé tegyük a szabályzat előírásait. Ennek egyik jó eszköze a jelenségek illusztrálása jól megválasztott példákkal.

A másik eszköz, amivel a rendszer hasznosságát növelhetjük a laikus használók számára, az az interaktivitás fokozása. A mostani verzióban gyakran előfordul, hogy több változatot egyaránt elfogad a rendszer, és ad ugyan magyarázatot mindegyikre, de a magyarázatok leginkább azt mutatják be, hogyan jutott a rendszer a deriváció eredményeként az eredményhez. Mindezt tömör, szakszavakkal terhelt szövegben közli a felhasználóval.

Ami leginkább hiányzik, az annak a világos szavakkal történő bemutatása, hogy az egyes változatok milyen jelentésbeli különbségeket hordoznak. Különösen fontos ez akkor, amikor a felkínált megoldások között sok olyan van, amely elvileg létezhet, de gyakorlatilag nagyon ritkán fordul elő, és csak hosszas belegondolás után értelmezhető.



3. ábra: A két megoldás megjelenítése magyarázatokkal és utalásokkal.

5 Tervek

A portál többéves fejlesztés eredményeként jött létre, de még korántsem mondható tökéletesnek. A legnagyobb feladat a rendszer túlgenerálásának korlátozása: jelenleg a szabályok sok olyan összetételt előállítanak, amelyek elvileg szabályosan képzettek ugyan, de amelyeknek az értelmezése nehéz. Ez óhatatlan ugyan, és bizonyos értelemben a helyesírási szabályzat tesztje is: a szabályok következetes implementálása ugyanis a kifejezések ezen halmazát hozzák létre – valószínűleg a szabályzat alkotóinak tudta és szándéka ellenére. A puding próbája az evés, a rendszert a hasznossága minősíti, nem a szabályok korrekt számítógépesítése.

A fejlesztés másik iránya a beszélők természetes nyelvhasználatához való nagyobb mérvű igazodás, amelyet a 4. pontban érintettünk. Ez a rendszer által elfogadott alakok körének kiszélesítését jelenti a spontán beszélt nyelvi változatokkal, amelyek adott esetben tájnyelvi elemeket is tartalmazhatnak. Távlati célként az lebeg szemünk előtt, hogy a felhasználó a saját természetes beszédmódja szerint kiejti az okostelefonján futó rendszernek a kért kifejezést, és a kijelzőn megjelenik annak a szabályzat szerinti köznyelvi írott alakja.

Irodalom

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I. & Trón, V. (2004): Creating open language resources for Hungarian. In: *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*, 1201-1204.

Laczkó, K. & Mártonfi, A. (2005): *Helyesírás*. Budapest: Osiris Kiadó.

Kis, Á. (1999): Az akadémiai helyesírási szabályzat és a számítógép. *Magyar Nyelvőr* 123 (2), 149-168.

Magyar Tudományos Akadémia (1984): *A magyar helyesírás szabályai*. Tizenegyedik kiadás. Pomázi, G. (szerk.): Budapest: Akadémiai Kiadó.

Miháltz, M., Husami, P., Ludányi, Zs., Mittelholcz, I., Nagy, Á., Oravecz, Cs., Pintér, T. & Takács, D. (2013): Helyesírás.hu – Nyelvtechnológiai megoldások automatikus helyesírási tanácsadó rendszerben– az intelligens helyesíróportál. In: Tanács, A. & Vincze, V. (szerk.): *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: SZTE, Informatikai Tanszékcsoport, 135-147.

Prószéky, G. & Tihanyi, L. (1993): Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. *TRIBUNE DES INDUSTRIES DE LA LANGUE* 5 (10): 28-29.

Váradi, T. (2002): The Hungarian National Corpus. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*. Paris: European Language Resources Association, 385-389.

VáradiTamás
MTA Nyelvtudományi Intézet
1068 Budapest
Benczúr u 33
varadi.tamas@nytud.mta.hu