Helyesiras.mta.hu – az intelligens helyesíró portál

Ludányi Zsófia – Miháltz Márton – M. Pintér Tibor – Takács Dávid

Kulcsszavak: nyelvtechnológia, helyesírás, formális grammatika, lexikonok, web

1. Bevezetés

A magyar helyesírás gépesítésére, számítógépes feldolgozására már számos kísérlet történt (lásd pl. Naszódi 1997, Kis 1999, Varasdi–Rebrus 2003). Az MTA Nyelvtudományi Intézete 2009 óta fejleszti a *helyesiras.mta.hu*-t, egy olyan internetes portált, amely a magyar nyelvre elérhető legfejlettebb nyelvtechnológiai eszközök segítségével próbál a magyar helyesírásra fogékony közönségnek tanácsadással szolgálni (a munkálatokról bővebben lásd Pintér et al. 2009, Miháltz et al. 2012). A portál 2013. április 30-án mutatkozott be. Jelen tanulmányban összefoglaljuk a portál egyes elemeinek (moduljainak) működését.[1]

Az eredeti elképzelés szerint (Pintér et al. 2009) a felhasználók egyetlen beviteli mezőbe írhatták volna be a keresendő kifejezéseket. A feladat összetettsége miatt azonban nyilvánvalóvá vált, hogy a beírt szöveg pontos elemzése nem vállalható. Az automatikusan generált válasz pontossága érdekében már a kérdés feltevésekor aktívan be kell vonni a kérdés feltevőjét is (ennek fő oka a magyar helyesírás értelemtükröztető jellegében rejlik). A pontosabb válasz érdekében ugyanis a portál által felkínált hétféle helyesírási terület közül kell a felhasználónak választania. Kérdéseire az alábbi területeken kaphat választ: külön- és egybeírás, helyesírás-ajánló (a szó szintjén), elválasztás, tulajdonnevek írása, számnevek helyesírása, keltezés, betűrendbe sorolás. A felhasználónak a nyitóoldalon (http://helyesiras.mta.hu) felkínált menüből kell kiválasztania, hogy milyen típusú helyesírási kérdésre szeretne választ kapni. A portál célja, hogy hasznos és hatékony segédeszköze legyen mindazoknak, akik helyesen szeretnének írni – az érvényben levő akadémiai szabályzat, *A magyar helyesírás szabályai* 11. kiadása alapján.

2. A *helyesiras.mta.hu* moduljainak bemutatása

2.1. Számok

A számjegyek betűvé alakítása aránylag jól automatizálható terület, mivel világosan megfogalmazott, egyértelmű helyesírási szabályai vannak. Létezik ugyan egy nemzetközi oldal (http://numbertext.org), amely kifejezetten e célból készült, és – számos egyéb nyelv mellett – a magyart is ismeri. Ez az eszköz azonban csupán a tőszámneveket, valamint – bizonyos határokig – a tizedes törteket kezeli, a sorszámneveket és a törtszámokat (pl. ⅔) már nem ismeri.

A *helyesiras.mta.hu Számok* modulja olyan eszköz, amely számokat alakít át betűvel leírt magyar szavakká, a hatályos akadémiai helyesírási szabályzat előírásainak megfelelően. Az eszköz ismeri a tő- és sorszámneveket, a hagyományos és a tizedes törteket. A keresőmezőbe a számjegyeken kívül előjelet (−), tizedesvesszőt (,) és törtvonalat (/) lehet írni.

Előfordulhat, hogy egy számjegyet többféleképpen is át lehet alakítani szóvá. Tipikusan ilyen a *2*-nek a *két* és *kettő* alakváltozata. A rendszer ilyenkor igyekszik a létező összes jó megoldást megadni, mégpedig a megfelelő magyarázatokkal ellátva. Ha például a *2/3* számot gépeljük a keresőmezőbe, négy lehetséges átírást is fogunk kapni: 1. *kétharmad csésze liszt*, 2. *két harmad nagyobb, mint egy harmad*, 3. *kettőharmad csésze liszt*, 4. *kettő harmad több, mint egy harmad*. A rendszer az 1–2. és a 3–4. közötti stílusbeli különbségre is felhívja a felhasználó figyelmét, miszerint a *kettő* alakváltozat nem része a sztenderd nyelvváltozatnak.

2.2. Dátumok

A dátumok helyesírása gyakran szokott problémát okozni – például mikor kell az évszám vagy a nap után pontot tenni. Viszonylag jól szabályozott területről lévén szó, a szabályok itt is könnyedén automatizálhatók.

A *Dátumok* eszköz a felhasználó által *évszám-hónap-nap* formában beírt vagy egy naptárból kiválasztott dátumot fogad el bemenetként. Válaszul visszaadja a leggyakrabban használt dátumformátumokat. Például a *2013-05-21* bemenetre a következő válaszokat kapjuk: *2013. május 21., 2013. máj. 21., 2013. V. 21.* Nemcsak az alapalakokat, hanem a leggyakrabban használt toldalékos alakokat is megkapjuk: *2013. május 21-én, 2013. május 10-e óta, 2013 októberében, 2013. évi, 2013 óta*.

2.3. Ábécébe rendezés

Az *Ábécébe rendezés* elnevezésű alkalmazás célja a felhasználó által megadott latin betűs tételek betűrendbe sorolása *A magyar helyesírás szabályai* 14–15. pontjainak megfelelően. Az eszköz néhány előfeldolgozási lépést követően – amilyen például az összetételi tagokra bontás, ennek alapján a betűhatárok megállapítása; a kettőzött többjegyű mássalhangzóbetűk feloldása [*ccs* > *cscs*] és a kivételes esetek kezelése – a klasszikus rendezési módszert alkalmazza. Ennek során mindig két tételt hasonlít össze balról kezdve, betűnként. Az első különböző betűpár összehasonlítása adja a két tétel egymáshoz képesti rendezését.

Az alkalmazás az úgynevezett általános magyar betűrend szerint tetszőleges tételeket képes betűrendbe sorolni. A szabályok értelmében az alkalmazás csak akkor tesz különbséget az egybeírt, kötőjellel írt vagy különírt alakok között, továbbá a kis- és nagybetűk, a (magyar) magánhangzók hosszú és rövid változatai, illetve az idegen mellékjeles betűk között, ha a tételek között ezeken kívül nincs más különbség *(Eger, egér, éger; Jáger, Jäger)*. A hagyományos írásmód szerint írt neveket – a szabályzatnak megfelelően – az írásképük (nem pedig hangalakjuk) alapján rendezi (így például a *Dessewffy* nem a *Dezső* mellé kerül).

*A magyar helyesírás szabályai*nak 16. pontja által említett kivételes betűrendbe sorolási eseteket (a bibliográfiai tételek betűrendje), tekintve, hogy ezek önálló algoritmusok lennének, továbbá a számokat is tartalmazó tételeket nem kezeli az eszköz. A magyar ábécén kívül más ábécék szerinti rendezést nem végez az alkalmazás.

2.4. Elválasztás

Az *Elválasztás* elnevezésű alkalmazás segítségével a szavak elválasztását ellenőrizhetjük. A keresőmezőbe írt szóalakot az alkalmazás az összes lehetséges helyen elválasztja.

Az alkalmazás a magyar nyelven elérhető nyílt forrású elválasztó program, a huhyphn, illetve a Humor morfológiai elemző alapján működik (vö. Novák és M. Pintér 2006). A Humor elemző segítségével azonosítjuk a szóösszetételi határokat, így, ha szükséges, a huhyphn által adott (szótagolási mintákra épülő) megoldást ki tudjuk egészíteni, pl. *megint* → *me-gint* (huhyphn), *meg*|*-int* (Humor) (az elválasztási és egyben szóösszetételi határokat az eszköz „|-” karakterekkel jelöli).

A huhyphn-t sok esetben módosítani kellett, mivel főleg tipográfiai célokra készült, így nem engedélyez például olyan elválasztásokat, mint *a-pa-i*, mivel egy karakter leválasztása nyomdai szövegben nem esztétikus. Az adatbázist egy egymillió szavas, a Magyar Nemzeti Szövegtár (Váradi 2000) szavainak gyakorisági listájából készült listán teszteltük: ha a tesztprogram olyan szótagot talál, amelyben a magánhangzók száma nem egy, akkor jelzi, és ha szükséges, azt kézzel javítjuk (pl. a *Mar-seille* esetében nem kell).

Az *Elválasztás* alkalmazás nem létező szóalakokra is ad elválasztási javaslatokat, ha részben megfelelnek a szótagolási szabályoknak (például: *ezdegbe* = *ez-deg-be*). A nem létező szavak helyes elválasztására azonban nem nyújt garanciát az eszköz.

A tulajdonnevek közül – főként a régies írásmódú magyar családnevek esetén – csak a leggyakrabban előfordulóakat képes az eszköz helyesen elválasztani, mivel mind a huhyphn, mind a Humor adatbázisa főként csak közneveket tartalmaz.

2.5. Névkereső

Mivel a nyelvtechnológiai alkalmazások nem tudják teljes mértékben kezelni a szemantikát, a hatékonyabb segítség, a helyesírási kérdésekben történő pontosabb válaszadás érdekében a tulajdonnevekkel, illetve azok bizonyos csoportjaival külön alkalmazás keretében foglalkozunk. A *Névkereső* alkalmazás a földrajzi tulajdonnevek és személynevek helyesírásához szolgál tanácsokkal. Az áttekinthetőbb válaszadás érdekében a tulajdonnevek közti keresés a többi alkalmazástól eltérően működik. A rendszer által tárolt több ezer felcímkézett tulajdonnév között nem lehetséges a szabad keresés. Ellenben begépeléskor a rendszer megjelenít egy, az újabb karakterek beírásával folyamatosan szűkülő találati listát, amely tartalmazza az eddig begépelt karakterekkel kezdődő összes tulajdonnevet.

A találati ablakban minden egyes tulajdonnév mellett feltüntetjük annak tulajdonnévi kategóriáit is, így az azonos alakú vagy hasonló tulajdonnevek esetében kiolvasható, hogy milyen lehetséges kategóriákba tartoznak (a kategóriák a tulajdonnévre kattintva érhetők el), pl. *Kaba = tulajdonnév – földrajzi név – településnév – személynév – vezetéknév – keresztnév – férfinév*.

Fontos hangsúlyozni, hogy a rendszer csak a szótáraiban tárolt földrajzi és személyneveket tartalmazza. Nem tartalmaz más tulajdonnévi kategóriákat, mint például magyarországi és nemzetközi intézményneveket, valamint cégneveket (utóbbiakat fel szerettük volna venni a rendszer mögötti adatbázisba, azonban a szabályostól eltérő, ugyanakkor már bejegyzett alakok ellentmondásos kezelhetősége miatt inkább lemondtunk róla).

2.6. Helyes-e így?

A *Helyes-e így?* elnevezésű alkalmazás célja a szóközöket nem tartalmazó jelsorozatok létezésének, illetve helyességének vizsgálata. Az alkalmazás újdonsága, hogy nem pusztán szótáralapú keresésre van beállítva (azaz nem csak arról tud döntést hozni, ami a mögöttes szótárban benne van), hanem kiegészül a helyesírást támogató formális nyelvtannal is (amelynek része például, hogy a találati ablakban toldalékolt és a rendszer által összerakott alakokat is felajánl). Válaszadáskor, illetve ajánláskor így nemcsak a szótárában található szavakról tud döntést hozni, hanem bizonyos mértékben a rendszer által nem ismert szavakat is tudja kezelni. Nem létező szavak vagy a rendszer által nem ismert szavak esetén a rendszer a keresett szóhoz karakterben legközelebb álló szóalakokat javasolja – tekintet nélkül a keresett szó jelentésére. Így lehet, hogy az egyszerűnek tűnő tévesztések javításakor is több, jelentésben oda nem illő szót ajánl (a keresés az alkalmazás mögött álló Hunspell[2] program szótáraiban található vagy a szótári tételekből szabályok alapján előállítható szavak között történik, miközben a keresés nem kezeli a beírt szó jelentését). A hibásan írt *papagály* szó esetén – a helyes alak mellett – számos olyan, a szótól csupán egy karakterben eltérő alakot is javasol a program, amelyek elképzelhetőek vagy létező jelentéssel bírnak, tehát potenciális, ugyanakkor nem lexikalizálódott vagy aktuálisan nem használatos alakok: *papagáj, papagála, papaggály, papragály, papdagály, papagálya, papapály*. Végül a kérdezőnek kell eldöntenie, hogy a felkínált lehetőségek közül melyik szót szerette volna – helyesen – leírni.

A *Helyes-e így?* helyesírás-ajánló nem csupán helyesírási, hanem a leggyakoribb nyelvhelyességi kérdésekben is segít. Például ismeri a nákolás (*én csinálnák*), a suksükölés (*mi lássuk a hibákat; ők elosszák a pénzt; az orromat is tisztítsa*), valamint a szukszükölés jelenségét is (*mi ébresszük őt fel; mi kelesszük a tésztát*). Ilyen esetekben egy-egy, a sztenderd nyelvváltozatnak megfelelő példamondattal hívja fel a figyelmet a megfelelő használatra.

Az alkalmazás kezeli az ún. alaki hasonlóság vagy paronímia jelenségét is (gondoljunk csak a nyelvművelők és a helyesírás által legalább írásban megkülönböztetett, a szubstandard, mindennapi beszédben azonban egyre inkább egybemosódó *egyelőre ~ egyenlőre, helység ~ helyiség* vagy *szabad tér ~ szabadtér* párokra). Ezekben az esetekben példamondattal illusztrálja a keresett szót, illetve

felajánlja a szó hasonló alakú párját, amelyre rákattintva annak normatív jelentését is megnézhetjük. Az alaki hasonlósággal bíró szópárokat folyamatosan bővülő szótárban kezeljük.

A többféle módon annotált szótárakkal és a szavakat, szótöveket toldalékoló minimális szabályrendszerrel ellátott alkalmazás sajnos nem használható minden beírt szó esetén. Mivel az alkalmazás a két szóköz közötti karaktersorozatokat vizsgálja, előfordulhat, hogy nagyon hosszú, kötőjeles szót is kezelnie kell. A többszörösen összetett szavak esetében az alkalmazás azonban nem tud helyes választ adni, viszont az összetételi tagok nagyobb száma miatt több lehetséges, de nem feltétlenül jó alakot ajánl. A többszörösen összetett szavak kezelésével a rendszer egy másik modulban foglalkozik. Ha ilyen szó helyesírását szeretnénk ellenőrizni, a *Külön vagy egybe?* nevű alkalmazást kell használni – ilyen input esetén a rendszer erre felhívja a felhasználó figyelmét.

2.7. Különírás-egybeírás

A *Külön-egybe* eszköz bemenetül szóközzel elválasztott szavakat vár. Válaszul megadja a rendelkezésre álló szabályok által létrehozható összes lehetséges (egybe-, külön-, kis- vagy nagykötőjellel írt) megoldást. A helyes szóalakokon kívül részletes magyarázattal is szolgál, és hivatkozik az akadémiai helyesírási szabályzat (vagy az Osiris Kiadó *Helyesírás* c. kézikönyvének: Laczkó Mártonfi 2004) megfelelő szabálypontjaira (az Osiris-helyesírás esetén szabálypontok helyett az oldalszámokra hivatkozik). Az Akadémiai Kiadó által gondozott *A magyar helyesírás szabályai* 11. kiadásának szabálypontjai hiperhivatkozások, amelyekre rákattintva a releváns szabályokat is elolvashatjuk.

A különírás-egybeírás a magyar helyesírás egyik legkomplexebb területe. A nehézséget főként az okozza, hogy a helyesírási szabályok helyes alkalmazásához szükség van bizonyos grammatikai alapfogalmak ismeretére, például különbséget kell tudni tenni a szószerkezetek és -összetételek között (Laczkó–Mártonfi 2004).

Az eddigi különírást-egybeírást ellenőrző eszközök szótáralapon működnek: ellenőrzéskor azt vizsgálják, hogy a bemeneti szó valamilyen formában (külön- vagy egybeírva) szerepel-e a szótárban. Ebből az következik, hogy csak véges számú szóalak ellenőrzését képesek elvégezni. Ezzel szemben a *helyesiras.mta.hu Külön vagy egybe* eszköze mögött egy formális nyelvtan áll, amelynek alapján a lehetséges jó megoldások generálódnak: ez tulajdonképpen azt jelenti, hogy a *Külön vagy egybe* eszköz végtelen mennyiségű összetett szót és többszavas kifejezést tud előállítani. Szabályok alapján elvben tetszőleges szó vagy kifejezés helyes alakját képes megadni.

Az eszköz működése nagy vonalakban a következő (erről bővebben lásd Miháltz et al. 2012, Ludányi et al. 2013): a felhasználó által szóközzel elválasztott bemeneteket az előfeldolgozó modul morfológiai (pl. szófaj, eset, szótagszám) és szemantikai információkkal látja el (pl. a *bőr* szóról megmondja, hogy anyagnévről van szó). A morfológiai és szemantikai jegyekkel felruházott tokeneket az ún. nyelvtani elemző kapja meg bemenetül: az elemző a formális nyelven leírt helyesírási szabályokat próbálja alkalmazni a felhasználó által beírt szavakra. A nyelvtani elemző elemzési szerkezeteket (elemzési fákat) épít a megadott szavakból: egy megadott bemenet esetén tehát többféle értelmezés, ezáltal többféle helyes írásmód is lehetséges. Például a *klónozott + kukorica + termesztő* bemenetre a következő két elemzés jön létre: az egyik elemzési fa először felépíti a *klónozott* és a *kukorica* szavakból a *klónozott kukorica* minőségjelzős szerkezetet, majd a különírt minőségjelzős szerkezet egészéhez kapcsolja a *termesztő* utótagot. Ilyenkor az ún. második mozgószabályt alkalmazza, és az eredetileg különírt szókapcsolatot alkalmilag egybeírja: *klónozottkukorica-termesztő*. A másik elemzési fa a szintén helyes, bár más jelentésű *klónozott kukoricatermesztő,* ahol az elemzési fa felépítése a másik irányból indul el. Az elemző először létrehozza a *kukoricatermesztő* jelöletlen tárgyas alárendelő összetételt, majd ehhez kapcsolja minőségjelzőként a *klónozott* szót.

A külön- és egybeírás területén sok olyan szabály létezik, amelyeket formálisan is meg tudunk fogalmazni, így automatikusan is végrehajthatók: az összes jelölt és jelöletlen alárendelői összetételre és szintagmára vonatkozó szabály (AkH. 104–128.), a szótagszámlálási (6 : 3-as) szabály (AkH. 138.), a mozgószabályok közül az első és a második (AkH. 139. a, b), az anyagnevek (AkH. 115.) és színnevek speciális szabályai stb. Más területek viszont egyáltalán nem algoritmizálhatók, például a

mellérendelő szószerkezetek (pl. *ment-ment, sürög-forog*), bár egy részük kivételszótárakkal megfelelően kezelhető.

3. Összefoglalás

Az MTA Nyelvtudományi Intézete által létrehozott *helyesiras.mta.hu* intelligens helyesíró portál lényegében válasz és egyben bizonyosság az eddig felmerült nyelvtechnológiai helyesírás-feldolgozás kérdéses pontjaira, a feldolgozásban rejlő lehetőségekre. A már létező és elérhető online helyesírási segédletek mellett a *helyesiras.mta.hu* valódi alternatívát jelent az automatikus és egyben intelligens helyesírási tanácsadásra. A kérdező interaktív bevonásával, precízen annotált szótárakkal, illetve formális nyelvtannal megvalósítható az a feladat, amelyet eddig nem sikerült megvalósítani: a magyar nyelvet lefedő helyesírási szolgáltató, amely azonnal pontos választ ad. A rendszer természetesen nem tökéletes – vannak olyan területei, amelyeken még nem képes az embert helyettesíteni (de valljuk be, a magyar helyesírással sokszor még az egyébként helyesen író emberek sem birkóznak meg egykönnyen). A *helyesiras.mta.hu* célközönségét elsősorban a helyesírási kérdésekre fogékony, néminemű alaptudással rendelkező érdeklődők jelentik, mint ahogy a helyesírási szabályzatot lapozgató, illetve helyesírásban eligazításra vágyó kérdezőnek is kell minimális helyesírási tudással rendelkezniük – másként hogyan merülne fel kérdésként például az, hogy a *muszály* szó helyesen van-e írva. A portálon található statikus *(A magyar helyesírás szabályai)* és dinamikus tartalmak (a helyesírást támogató alkalmazások) ízléses köntösben tálalva próbálnak meg minél szélesebb közönséget megszólítani.

A *helyesiras.mta.hu* portál az első bizonyítéka annak, hogy a magyar helyesírás nagy százalékban kezelhető nyelvtechnológiai alkalmazásokkal. A teljes körűen működő portál fejlesztését azonban még nem fejeztük be. A folyamatos karbantartás és a javítások mellett számos egyéb (helyesírási és nyelvhelyességi) tartalom bővítésén dolgozunk.

Jegyzetek

1. A *helyesiras.mta.hu* használatáról, valamint az egyes modulok működési elveiről a http://helyesiras.mta.hu/helyesiras/default/howitworks oldal szolgál bővebb információkkal.
2. http://hunspell.sourceforge.net

Irodalom

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

Kis Á. 1999. Az akadémiai helyesírási szabályzat és a számítógép. *Magyar Nyelvőr* 123. évf. 2. szám.149–168.

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

Ludányi Zs., Miháltz M., Hussami P. 2013. Különírás-egybeírás – automatikusan. Megjelenés alatt: *VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia.*

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.

Naszódi M. 1997. Nyelvhelyesség-ellenőrzés számítógéppel (Parciális szintaxis). In: *VII. Országos Alkalmazott Nyelvészeti Konferencia*. I. kötet. Budapest: Külkereskedelmi Főiskola. 256–260.

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., Vincze V. (szerk.) 2009. *MSZNY 2009. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress. 172–182.

Váradi T. 2000. Szótár, korpusz – Magyar nemzeti szövegtár. In: Gecső T. (szerk.) 2000. *Lexikális jelentés – aktuális jelentés.* Budapest: Tinta Könyvkiadó. 263–270.
Varasdi K., Rebrus P. 2003. *A helyesírás mint default öröklődési hálózat.* Elhangzott: *A mai magyar nyelv leírásának újabb módszerei VI.* Szegedi Tudományegyetem, Szeged, 2003. október 16–17.