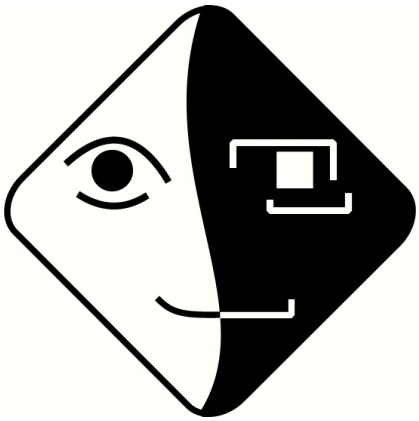
VI. Magyar Számítógépes Nyelvészeti Konferencia



MSZNY 2009

Szeged, 2009. december 3-4.
http://www.inf.u-szeged.hu/mszny2009

Online helyesírási szótár és megvalósítási nehézségei

Pintér Tibor[1], Mártonfi Attila[1], Oravecz Csaba[1]

[1] MTA Nyelvtudományi Intézet, Benczúr utca 33.,
1068 Budapest, Magyarország
{tpinter, martonfi.attila, oravecz}@nytud.hu

Kivonat: A magyar társadalom helyesírás és nyelvhelyesség iránti igénye mármár szakmai közhelynek számít. A helyesírás számítógépes modellezésének eddigi gyakorlata azt mutatja, hogy egy online helyesírási szótár, nyelvi tanácsadó szolgáltatás triviálisan nem oldható meg csupán gépi erőforrással, például egy nyelvtan mögött álló szótárral. A helyes alak felismeréséhez mindenképpen szükség van morfológiai elemzőre, illetve az elemzés kimeneteként keletkező homonimák egyértelműsítésekor bizonyos mértékben a kérdező interaktivitására is. A morfológiai elemzést segíti a főként szemantikai szempontok alapján szerkesztett szótár, amelyben az egyes lexikai tételek több szempontból annotálva vannak (ehhez a szótárat különféle szemantikai kategóriák alapján egyértelműsítettük, valamint az interakciót elősegítendő, egyszerű mondatokkal rávezetjük a kérdezőt az adódó lehetőségek közti választásra). Sok esetben a morfológiai elemző és a szótár önmagában nem elegendő a helyes alak kiválasztásához, így némely esetben a lokális szintaktikai környezet elemzését is fel kell vállalnunk. Az online helyesírási tanácsadó rendszer erősen formális felépítésű. Hatékony működése érdekében teljesen új – formális rendszert követő – alapokon kell leírnunk a helyesírás számos részrendszerét.

1 Bevezetés

A magyar nyelvre alkalmazott nyelvtechnológiai kutatások mostohán kezelik a helyesírási relevanciájú internetes segédeszközöket. Bár a hibátlan, „helyes" írás megmozgatja a művelt magyar társadalmat, ezekben a kérdésekben leginkább az e-mailes és telefonos segítség, illetve a különféle fórumok által közvetített ember-ember interakció az, amit a nyelvhasználók leginkább igénybe vesznek. Ennek oka nem elsősorban a megfelelő nyelvtechnológiai eszköz hiánya (általában morfológiai elemzővel kiegészített, szótári keresésen alapuló eszközök vannak forgalomban; MorphoLogic: Helyes-e?; Németh László: Hunspell, Szabad magyar szótár), hanem a magyar helyesírásnak az a tulajdonsága, hogy bizonyos pontokon a szabályalkalmazók anyanyelvi kompetenciájára és szövegértelmezésére hivatkozik, illetve számos, a szabályrendszernek ellentmondó íráshagyományt is továbbörökít. E miatt az összetett függés miatt valószínűtlennek tartjuk egy olyan program kifejlesztését, amely emberi segítség (felhasználói interaktivitás) nélkül képes lenne hatékonyan kezelni a magyar helyesírás minden pontját (vö. [1, 2]).

Az MTA Nyelvtudományi Intézete éppen ezért olyan portál elkészítésén dolgozik, amely megszüntetné a fent említett űrt: egy pontos és gyors, mindenki által elérhető, azonnal segítséget nyújtó internetes nyelvi tanácsadó portál, a **helyesiras.hu** megalkotásán. A rendszer működőképessége három alappilléren, 1. egy robusztus, többrétegű, annotált szótáron, 2. pontos, formális nyelvtanon és 3. a kérdező interaktivitásán alapszik (ez utóbbira a helyesírás egyes részeinek erőteljes szemantikai beágyazottsága, az ún. értelemtükröztetés miatt van szükség). A már működő internetes helyesírási segédletekhez képest a most készülő rendszer nagyobb fedésű és remélhetőleg jóval megbízhatóbb és pontosabb lesz, nem pusztán egy helyesírási szótár szolgai számítógépes másolata. A pontosság mellett egyéb olyan tulajdonságai is lesznek, amelyek reményeink szerint nem csak a helyesírási alapismeretekkel rendelkezőket és nem csak a magyarországi nyelvhasználókat ösztönzik majd a portál használatára. A **helyesiras.hu** számos újítása miatt új felhasználói irányban is nyit.

2 A nyelvtan

2.1 Milyen nyelvtanra van szükség?

Az előmunkálatok folyamán nyilvánvalóvá vált, hogy a helyesírási problémák nagy része lefedhető szótárral, vagy megoldható egyszerű grammatikával. A valódi kihívást ezért csupán a magyar helyesírás bizonyos pontjai jelentik (ám önmagukban ezek megoldása jelentős munkával jár). A magyar helyesírás létező számítógépes modelljei azt mutatják, hogy hatékony helyesírási tanácsadás nem valósítható meg csupán gépi erőforrással és a nyelvtan mögött álló szótárral (még több százezer szavas háttérkorpusz esetén sem). Az egyszerű szójegyzéken alapuló tanácsadás (ezt csinálják az interneten jelenleg elérhető helyesírási tanácsadók) csak akkor ad kielégítő eredményt, ha a beírt (lekérdezett) szó eleve helyesen van írva, valamint megtalálható a rendszer mögött álló szótárban (illetve jobb esetben a mögöttes nyelvtan össze tudja rakni). A helyesen írt, ugyanakkor nem ismert szavakat az ilyen elemzők hibás írásmódúként adják vissza, vagyis nem nyújtanak többet egy átlagos, szabályzattal nem rendelkező papírszótárnál. Pontosabban lényegesen kevesebbet nyújtanak, ugyanis egy papírszótár készítője az anyag elrendezésével (tehát a keresett elem betűrendi és szócikkbeli környezetével) tekintélyes mértékű információt tud adni a szótárt lapozgató felhasználónak, hiszen ezen a módon interakcióba tud lépni a szótárhasználó anyanyelvi intuíciójával, egyéb ismereteivel és kognitív működésével.

Az általunk fejlesztett rendszerben a kérdező által beírt szót vagy többtagú kifejezést a webfelület mögött működtetett elemző értelmezi, megpróbálja azonosítani a lehetséges helyesírási problémakört, majd megválaszolni, illetve jóváhagyni a helyes alakot. A keresett alak felismeréséhez mindenképpen szükség van morfológiai elemzésre (pl. a különféle, különösen az *-ó/-ő* képzős igenevek felismerése, az alkotó tagokban szereplő tőmorfémák számlálása). A nyelvtan és a szótár együttes használata sem jelent azonban minden esetben megoldást, hiszen például a keresés kimenetén megjelenő homonimák egyértelműsítése bizonyos mértékben már a kérdező interak-

tivitását igényli. A helyesírásukban eltérő, kiejtésükben (vagy legalábbis a szegmentális hangszerkezetben) azonos, tehát homofón alakpárok, -többesek esetében számos alakváltozat helyes lehet (pl. *klónozottkukorica-termesztő* 'klónozott kukoricát termesztő személy' – *klónozott kukoricatermesztő* 'olyan kukoricatermesztő, akit klónoztak', *adalékanyag* 'az adalék anyaga' – *adalék anyag* 'adalékul használt anyag', *csuklósbusz-vezető* 'csuklós busz vezetésére alkalmazott gépkocsivezető' – *csuklós buszvezető* 'csuklásra hajlamos autóbusz-vezető'), mivel azonban az éppen keresett alak azonosítása magas szintű, tág szövegkörnyezetre támaszkodó nyelvi elemzést igényelne, és a tanácsot kérő csak egy szót vagy szókapcsolatot ad meg, a tanácsadó a megfelelő alak kiválasztása érdekében ilyen esetekben kénytelen az elemzési folyamatba bevonni a kérdezőt is.

Milyen morfológiai elemzésekre is van a helyesírás szempontjából szükség? A bemeneti karaktersorozaton végrehajtandó elsődleges elemzés a tőmorfémákra bontás (mivel a helyesírásban használt ÖSSZETÉTELI TAG fogalom valójában ennek a nyelvtani kategóriának felel meg) – nem mindegy például, hogy az elemző hogyan szegmentálja például a következő szavakat: *rendszer (= rend+szer), valószínűség (= va[ló]+szín[űség]); szemöldök* (képzett alak, nem összetétel), hiszen a helyes szegmentálás képezi a magyar helyesírás különírás-egybeírási részrendszerében a szótagszámlálás szabályának egyik bemenetét (*valószínűség-számítás* és nem **valószínűségszámítás*, mivel a *valószínűség* összetett szóalak, így megvan a 3 tőmorféma és a 7 szótag). Ugyancsak a különírás és egybeírás kategóriájához tartozik a toldalékmorfémák pontos szegmentálása és típusok szerinti elkülönítése (a fenti szótagszámba beleszámítanak a képzők, de a jelek, ragok nem), ez azonban teljes mértékben gépesíthető.

A program kezeli továbbá többek között a különféle, hagyományokon alapuló külön- és egybeírást. Rendszerszerű hagyomány szerinti írásúnak tekinthetők például az anyagnevek, a színnévi jelzős összetételek vagy a számnévi jelzős, *-s, -i, -ú/-ű/ -jú/-jű, -nyi, -nként, -nta* toldalékra végződő alakulatok. Ha a jelzői szerepű szó és az alaptag egyszerű szó, akkor egybe kell őket írni (1+1=1), s ezt a program követi is. Ha valamelyik tag önmagában is összetett szó, akkor már különírandók (2+1|1+2=2): *selyemköntös ~ nyersselyem köntös*, *ötéves ~ öt hónapos*, *kétévnyi ~ tizenkét évnyi*, *kéthavonta ~ tizenkét havonta*. Hasonló algoritmus mozgatja az anyagnévi mozgószabályt is, ahol a különírt szószerkezet anyagnévi jelzőként szerepel: *valódi bőr,* de: *valódibőr kabát*; *fehér márvány,* de: *fehérmárvány vízcsap; tömör arany,* de: *tömörarany nyaklánc*. A fenti helyes írásmódok kialakításához arra is szükség van, hogy a program meghatározza az egyes alkotótagok közötti szintaktikai függéseket, valamint felismerje az ANYAGNÉV szemantikai kategóriát. Ez utóbbiban kapnak szerepet az annotált szótárak.

A magyar szavak külön- és egybeírása a felhasználó számára is meglehetősen bonyolult, egy helyesírási tanácsadó számára is szinte megoldhatatlan, bár részlegesen nyelvtannal és szótárral jól kezelhető. (A gépi választ nem eredményező esetekben, illetve azokban, amelyek során a kérdező nem elégedett a válasszal, a rendszer felkínálja a humán tanácsadói segítség igénybevételének lehetőségét.) A morfológiai elemzőknek általában alapvető problémájuk, hogy az elemzést két szóköz között hajtják végre, így csak a hibás egybeírást képesek észrevenni, a különírást viszont nem, vagy csak korlátozott mértékben (l. pl. a Helyesek „zöld aláhúzása"). A **helyesiras.hu** a részletesen annotált szótárak segítségével hatékonyan (bár nem teljes körű-

en) kezeli a magyar külön és egybeírás szemantikai jellegű komponenseit is. A szótárakkal és visszakérdező modullal kiegészített rendszer képes szemantikailag is különbséget tenni (és így a kérdezett alakot helyesen visszaadni) például az *-ó/-ő* képzős melléknévi igeneves szerkezetek vagy az összetett főnevek külön- és egybeírásának kérdésében (*csomagoló papír* 'olyan papír, amely éppen csomagol' – *csomagolópapír* 'csomagolásra készített papír', *napra forgó* 'a nap hatására meg-megforduló' – *napraforgó* 'magjáért, olajáért tartott haszonnövény', *járólapos* 'járólappal rendelkező, azzal felszerelt' – *járó lapos* 'gyalogló kismellű', *vendégfogadó* 'vendégül látó személy, ill. panzió' – *vendég fogadó* 'vendégségbe jött bukméker', *tanulószoba* 'tanulás tevékenységére rendszeresített helyiség' ~ *tanuló szoba* 'olyan szoba, amely tanul').

A tőmorfémák számának megállapítására irányuló szegmentálás mellett a morfoszintaktikai komponensnek kezelnie kell a szófajokat is. Erre is elsősorban a külön- és egybeírás miatt van szükség, hiszen például a színnévi jelzős összetételek, bizonyos fokozó szerkezetek vagy akár az anyagnévi mozgószabály helyes kezeléséhez ez elengedhetetlen. Lássunk erre is pár példát: a fokozó szerepű melléknévi vagy főnévi etimonú szó (azaz fokozópartikula) mindig külön áll a rákövetkező melléknévtől, például: *borzasztó rossz, böszme nagy, csoda jó, jó nagy, kutya hideg, marha erős, szép kövér, tök hangos.* Ettől eltér a hasonlítást kifejező jelentéssűrítő összetételek írásmódja, például: *csodaszép* 'a csodához hasonlatosan szép', *hófehér* 'a hó színéhez hasonlóan fehér', *hollófekete* 'a holló színéhez hasonlóan fekete'.

A magyar helyesírás, illetve a mögötte álló grammatikai modell összetett volta miatt a nyelvtani modulnak ki kell egészülnie kivételszótárral. Ez az MTA Nyelvtudományi Intézetében évtizedek óta működő helyesírási tanácsadói munkatapasztalat, az ezeket rögzítő jegyzőkönyvek, illetve a helyesírási szabályzatok szerkesztésekor felhalmozott tudás alapján készült.

2.2 Morfológia mellett lokális szintaxis

Mint erre korábban utaltunk, sok esetben a morfológiai elemző és a szótár önmagában nem elegendő a helyes alak kiválasztásához, így némely esetben a lokális szintaktikai környezet elemzését is fel kell vállalnunk (pl. bizonyos bővítmények megléte kulcsként szolgálhat annak eldöntésében, hogy egy alakulat szókapcsolat vagy összetétel-e, pl. *takarítónő* 'foglalkozásszerűen helyiségeket tisztává tevő nő' – *takarító nő* 'olyan nő, aki helyiségeket éppen most tesz tisztává' – *sokat takarító nő* 'olyan nő, aki sokat takarít'). Elsősorban a homofon alakok egyértelműsítése érdekében ennek a kérdező segítségét igénybe kell vennie – rávezető kérdéseken keresztül.

3 A szótár

A legtöbb helyesírás-segítő szolgáltatás szótár alapján működik: ez elkerülhetetlen alap, önmagában azonban nem megoldás, mivel a végeredmény így számos hiányt, kívánnivalót hagy maga után. A pusztán szótáron alapuló megoldás hátránya, hogy a keresés kimenete csak azt adja meg, hogy a beírt szó (karaktersorozat) megvan-e az

adatbázisban: akkor sem fogunk pozitív eredményt kapni, ha olyan szót keresünk, amely helyesen van ugyan írva, de az adatbázis nem tartalmazza. A fentiek ismeretében a morfológiai elemző sem elég hatékony megoldás önmagában, gazdag és részletesen annotált szótárak nélkül nem képzelhető el jól működő helyesírás-elemző és tanácsadó rendszer. A **helyesiras.hu** morfológiai elemzőjét főként szemantikai szempontok alapján annotált részszótárak gyűjteménye segíti, amelyben az egyes lexikai tételek több szempontból is kódolva vannak (ehhez a szótárat különféle szemantikai kategóriák alapján egyértelműsítettük). A kiejtésben az írásképtől jelentősen eltérő szavak, nevek, mozaikszók esetében szükség van a szótárban kiejtésjelölésre is az elválasztás, a toldalékolás, illetve a névelőzés helyes meghatározásához.

3.1 Szótári erőforrások

A portál alapvető lexikális erőforrásait egyrészt a Magyar Nemzeti Szövegtár 187 millió szavas, kontextuális stílusok szerint tagolt korpusza, másrészt egy külön erre a célra összeállított több mint 400 millió szavas, címkézett gyűjtemény adja. Ez utóbbi több mint 4 millió elemzett szóalakot, közel 2 millió szótövet tartalmazó, műfaji kategóriákba sorolt gyakorisági adatbázis. Az adatbázishoz kapcsolódó lekérdező felület már működik, ezzel a szótárnak a kritikus helyesírási problémákat tartalmazó, jellemző szóalakok feletti fedése vizsgálható közvetlenül (1. ábra). Ezek mellett az alapvető források mellett a rendszert a felhasználói kérdésre adott pontos válasz megtalálásában egy több tízezer többtagú kifejezést tartalmazó szótár, valamint több, specifikus szemantikai jegyek alapján összeállított szólista támogatja (pl. csak kis- és nagybetűben vagy különírás-egybeírásban eltérő stb. minimális párok, anyagnevek, számnevek, jelzők, állatnevek, növénynevek, településnevek, magyar családnevek és kiejtésük, különböző szókapcsolatok listája [*-ó/-ő* képzős melléknévi igeneves szerkezetek, fn+fn, mn+fn], *a* végű szavak listája). Az aktuális problémának a számítógép számára érthető formális meghatározásában további segítséget nyújt egy mintegy 6000 rekordos adatbázis, amely a közönségszolgálati jegyzőkönyvekben található kérdés-válaszokat rendszerezi és osztályozza.

Az annotált részszótárak közül külön érdemes foglalkozni a minimális párokat, anyagneveket, melléknévi igeneves szerkezeteket stb. feldolgozó szótárakkal. A minimális párok szótára 1040 olyan párt tartalmaz, amelyek között egykarakternyi eltérés található (ez lehet akár kis- és nagybetű, illetve szóköz is).

1. táblázat: Mutatvány a minimális párok szótárából.

abba *(nm.)*	**abba-** *(ik.)*
Ábrahámhegy *(település)*	**Ábrahám-hegy** *(hegy)*
adalékanyag 'az adalék anyaga'	**adalék anyag** 'adalékul használt anyag'
adóvevő *(fn.)*	**adó-vevő**
afelé *(hsz.)*	**a felé** *(nm.)*
afelett *(hsz.)* ~ afölött	**a felett** *(nm.)* ~ a fölött
afelől *(hsz.)*	**a felől** *(nm.)*
Ag <ezüst>	**AG**

ági	**Ági**
ágrólszakadt 'nyomorult'	**ágról szakadt** 'olyan, ami leszakadt egy ágról'
ahelyett *(hsz.)*	**a helyett** *(nm.)*
akadémia 'főiskola'	**Akadémia** 'Magyar Tudományos Akadémia'
akár	**akár-**
akárcsak 'mint' *(ksz.)*	**akár csak** 'akár csupán'
akárhogy 'bármilyen módon'	**akár hogy** *(kihagyásos szerkezetben)*

A minimális párok megfelelő kezelése elsősorban a visszakérdezés során oldható meg, mivel a két elem közti eltérések főként szemantikaiak, így a pontos alak kiválasztásában legfőként a kérdező tud segíteni interaktív kérdéseken keresztül (hiszen a kérdező szándékát közvetlenül nem ismerhetjük). A kérdező a helyesírás fogalmi rendszerében gyakran nem tudja artikulálni teljes pontossággal a kérdését (ha tudná, nem kérdezne), így a rávezető kérdéseknek olyan releváns és főképpen egyszerűen közölt információkat kell tartalmaznia, amelyek nyelvtani-helyesírási ismeretekre nem építenek, csupán a kérdező anyanyelvi kompetenciájára, és amelyekből a kérdező számára kiderül, pontosan melyik alakváltozatra is van szüksége (pl. *tanítónő – tanító nő*).

tanítónő	» éppen a cselekvést, tevékenységet végzi, esetleg folyamatot átéli, elszenvedi (nő, aki éppen most tanít)	» *tanító nő*
	» valamire rendeltetett, valamit általában, foglalkozásszerűen űz, nem vagy nem pusztán pillanatnyi cselekvést, tevékenységet végez, illetve folyamatot átél, elszenved (tanításra való nő)	*» tanítónő*
kávéscsésze	» valamit tartalmazó, valamivel szennyezett edény (kávét tartalmazó, kávéval szennyezett csésze)	» *kávés csésze*
	» valaminek a felszolgálására, fogyasztására használt, szokásosan meghatározott méretű és formájú edény (kávé felszolgálására, fogyasztására szolgáló csésze)	» *kávéscsésze*

Bár tudjuk, hogy a szemantikai információ megfelelő minőségű kezelésétől még távol vagyunk, nem kerülhetjük meg a szavak bizonyos jelentésjegyeinek beépítését. Erre alakítottuk ki az annotált szótárakat, amelyek a megfelelő nyelvtani szabályokkal kiegészítve hatékonyan kezelik a helyesírás azon pontjait, ahol a morfológiai-szintaktikai elveket kiegészítik a szemantikai kategóriák.

3.2 Feldolgozó modulok

A rendszer működését a helyesírás részrendszerei köré szervezett modulok vezérlik, amelyeket az alábbi attribútumok jellemeznek:

1. a modul feladata: a modul által kezelt jelenség leírása;

2. a modul működéséhez szükséges erőforrások és jellemzőik specifikációja (pl. milyen speciális szólista szükséges a kérdéses jelenség kezeléséhez);

3. a modulhoz rendelhető felhasználói kérdés géppel azonosítható jegyei, illetve ezek hiányában a felhasználótól bekérendő további információ meghatározása;

4. a modul működésének forgatókönyve: a modulok működését forgatókönyvek írják elő, amelyek megadják, hogy amennyiben az adott felhasználói lekérdezés a modulhoz rendelődik, milyen processzáló lépések szükségesek a válasz megadásához (pl. a lekérdezett alak szerepel-e a modulhoz rendelt lexikális erőforrásokban → igen → rendben; → nem → felhasználótól további információ, ennek alapján válasz generálása).

4 A további, speciálisabb részrendszerek kezelése

A szavak, egyszerűbb szókapcsolatok szótár és nyelvtan egységén alapuló kezelésének vázlatát mutattuk be az eddigiekben. Szükséges azonban szólni azokról a részrendszerekről, amelyeknek a működtetéséhez ezek a műveleti elemek nem nyújtanak elégendő támpontot. Ezek többnyire diffúzabb problematikát mutatnak, így a számítógépes kezelésük is nehezebben körülhatárolható, ugyanakkor alapvető jelentőséggel bír, hogy az MTA Nyelvtudományi Intézet közönségszolgálati jegyzőkönyveinek tanúsága szerint a felvetett kérdések túlnyomó többsége a különírás és egybeírás kérdéskörét érinti elsősorban. Mindazonáltal nem maradhatnak megválaszolatlanul az alábbi részrendszereket érintő kérdések sem.

4.1 Tulajdonnevek

A legnagyobb összetartozó problémakört a különféle tulajdonnevek jelentik. Noha ezt a kategóriát szófaji megnevezésként is szokás használni, számítógépes nyelvészeti értelemben nem érdemes szófajnak tekinteni – túlnyomó többségük ugyanis többszónyi terjedelmű (azaz a tulajdonnévi egységet adó karakterláncok rendszerint tartalmaznak szóközt). Ezen a ponton természetesen érintkezik a tulajdonnevek írásának kérdésköre a különírás és egybeírás területével, ez kiegészül azonban a kis- és nagybetűk használatának problematikájával is. Itt talán még fokozottabb szerepe van a szemantikának, hiszen a denotátum tulajdonnévi osztályai is tükröződhetnek az írásképben, például: *Magyar Nyelv* (folyóiratcím) – *Magyar nyelv* (könyvcím), *Tátrai-vonósnégyes* 'Tátrai Vilmos által alapított, általa vezetett kvartett, illetve őáltala komponált, ilyen összeállítású hangszeregyüttesre írt ciklikus mű' – *Tátrai vonósnégyes* 'Tátrai Vilmos emlékére, tiszteletére elnevezett kvartett' – *Tátrai Vonósnégyes* 'ez utóbbi mint jogilag is intézménnyé alakult társaság', *Gellért-hegy* 'domb Budán a Duna jobb partján az Erzsébet hídnál' – *Gellérthegy* 'ez mint városrész', *Tisza híd* 'Tisza Kálmánról elnevezett híd' – *Tisza-híd* 'a Tiszán átívelő híd', *magyar állam* (közszói megnevezés) – *Ohio állam* (országrésznév, vö. *Csongrád megye*) – *Izrael(i) Állam* (államnév, vö. *Magyar Köztársaság*), *Szent István* 'a magyar államot megalapító király' – *Szentistván* (település), *Madách Színház – Madách mozi, Béke Szálló –*

Béke étterem; Békás patak (a patak neve önmagában a *Békás*) – *Gombás-patak* (a patak nevének része a *patak* földrajzi köznévi utótag is).

A kategoriális különbségek megjelennek az *-i, -s, -beli* képzős alakokban is. Itt külön szerepe van az egyes alkotótagok tulajdonnévi vagy közszói voltának is: *kossuthi – shakespeare-i – rippl-rónais – Csokonai Vitéz-i, nemzeti színházi – Madách színházi, Békás pataki – Békás-szorosi* (mert az előtag a *Békás* patak tulajdonneve) – *gombás-pataki, országos Széchényi könyvtári – holt-Tisza-bereki, móriczi – Móricz-féle; kosztolányis – Népszabadság-os – nyugatos* (egyszeri kivétel) stb.

További problémát jelent bizonyos tulajdonnévi kategóriák esetében a kodifikáció és az úzus között feszülő oly mértékű diszkrepancia, amelyről valamilyen formában már a tanácsadásnak is tudomást kell vennie (pl. események, rendezvények elnevezésének, illetve intézmények alegységeinek szabálytalan, de általánosan elterjedt nagybetűs írása), valamint azok a tulajdonnévtípusok, amelyeket nem vagy csak nagyvonalakban kodifikált az 1984-ben megjelent, ma is hatályos helyesírási szabályzat (pl. a címadási szokások megváltozása; a címmel ellátható műfajok sokaságának megjelenése; a programok, akciók, pályázatok korábban elképzelhetetlen változatosságban való használata; a márkanevek jogi kérdéseket is felvető írásproblémái; a legkülönbözőbb fajtájú alapítványnevek; a díjak, kitüntetések elnevezésének alapjaiban új típusai).

A földrajzi nevek bonyolult szaknyelvi szabályozásáról vagy a kémiai elnevezések helyesírásáról, az állat- és növényneveknek a taxonómiát tükröző írásmódjaival csak a távolabbi jövőben lesz mód foglalkozni.

4.2 A magyar nyelvbe bekerülő idegen elemek

Az idegen szavak, nevek, illetve kifejezések részrendszere alapvetően két lényegi kérdést vet fel.

Az első és általánosabb annak problémája, hogy egy újonnan a magyar nyelvbe kerülő szó, kifejezés idegenes vagy magyaros írásmóddal írattassék-e. Az ennek meghatározásához szükséges, formális és kategoriális szempontokon alapuló döntési fa a szükséges kommentárokkal együtt megtalálható az Osiris Helyesírásban [3]. Ezt egészíti ki az egyszavas köznevek kezelésére vonatkozó eljárás. Ennek lényege, hogy azon idegen eredetű szavak esetében, amelyek korábban nem szerepeltek normatívnak tekinthető szótárban, 40%-os vagy a feletti magyar írásmódú korpusz-előfordulás esetén (ha egyéb, releváns szempont nem merül fel), a magyaros írásmód támogatandó. Korlátozottan, de ugyanez követendő, ha szerepel az adott szó normatív szótárban, de idegenes írásmóddal (ekkor ugyanis nyelvhasználati változás tehető fel).

A második és speciálisabb probléma az idegen írásrendszerből való átírás kérdéskörre. Mivel az átírási szabályzatok jól formalizálhatók (akár az eredetiből, akár más átírásból indulunk ki), ennek számítógépes támogatása igen sikeres lehet.

4.3 Írásjelhasználat

Az írásjelhasználat szabályozása sok tekintetben fakultatív, alapjául azonban mégis csak a szintaktikai szerkezet elemzése szolgál. Ebben a tekintetben – igaz korlátozot-

tan – használhatók parciális szintaktikai szabályok (pl. két azonos esetű főnév általában nem követheti közvetlenül, írásjel nélkül egymást, de: *a városban decemberben;* két véges igealak között általában kell lennie egy írásjelnek, de problémát jelentenek a befejezett melléknévi, illetve az igei igenevek mint a véges igealakokkal homonim formák: *ettem az anyám sütötte kenyérből, ettem az anyám által sütött kenyérből*). A felvethető kérdéseknek ezek azonban csak szűkebb körére adnak választ. Szükséges tehát a mélyebb szintaktikai elemzés kialakításán túlmenően bizonyos szövegtani, stilisztikai, pragmatikai szempontok figyelembevétele. Hogy ezekből mennyi formalizálható, illetve milyen módon lehet ezeknek az esetében az interaktív felületet felhasználni, további megfontolásokat igényel. Ezek kifejlesztése csak a távolabbi időben lehetséges.

4.4 Rövidítések, mozaikszók

A rövidítésekre és mozaikszókra különféle helyesírási szabályok sokasága vonatkozik, a tény azonban mégiscsak az, hogy a szabályos írásmódú formák kisebbségben vannak a különféle hagyományos esetekkel szemben. Így ebben a körben a szabályismertetésen és a szótári keresésen túlmutató megoldást tervezni jelen ismereteink szerint nem lehetséges.

4.5 Keltezés, a számok írása

A keltezéssel, illetve a számok írásával kapcsolatos helyesírási tudnivalók igen egyszerűek és eleve formálisak, tehát számítógépes támogatásuk nem okoz komolyabb nehézséget.

5 További feladatok – kiejtéskövető írás vs. helyesírás, hibás szavak gyűjteménye, illetve a magyar nyelv határon túli változataiban használatos szavak gyűjteménye

A helyesírási segédletek (legyen az könyv vagy számítógép) elsősorban azok számára jelentenek támogatást, akik tisztában vannak a helyesírás alapvető kategóriáival (pl. a hangjelölés alapelveivel [kiejtés szerinti, szóelemző, hagyományos, egyszerűsítő írásmód], a helyesírás alapfogalmaival [pl. értelemtükröztetés, tulajdonnévosztályok], illetve a helyesírási kodifikáció mögött álló nyelvtani modell felépítésével és fogalomhasználatával). A szélesebb felhasználói kör kiszolgálásának érdekében a tervek közt szerepel egy olyan modul beiktatása is, amely hatékonyan kezeli a kiejtéskövető írásmódot is. A magyarországi helyesírási segédeszközök között újítás lenne, hogy a szoftver nemcsak a helyesírási vétség(ek), illetve az elütés(ek) miatt hibásan leírt szavakat ismerné fel és tudná javítani, hanem a köz- vagy tájnyelvi kiejtést tükrözve leírtat is. A hibásan beírt szavak esetében egyrészt a szokásos eljárás szerint felkínálja a lehetséges jó változatokat (ez elsősorban elgépelésnél lehet hasznos), másrészt egy

speciális elemző modul segítségével felismeri a kiejtés alapján a mögöttes morfémaszerkezetet, s végül felkínálja a helyesírás szerinti alakot.

Ez azért meghatározó újdonság, mivel azok, akik nincsenek tisztában a helyesírás alapvető szabályaival sem, a kiejtést tükröző alakot hallás után leírva eleve nem férnek hozzá a helyesírási szótárakban elérhető ismeretanyaghoz. A magyar nyelv szavainak, kifejezésének írott és beszélt formája között feszülő eltérés alapvető szabályait felhasználva lehetőség nyílik a kiejtést tükrözve leírt szavak írott alakúra történő változtatására (illetve a kétszintű morfológiához hasonlatos módon az ellenirányú átalakítás is megoldható szükség esetén). Hangtani szabályok ismerete alapján a rendszer felismeri a kérdező szándékát, és ez alapján generálja a szóelemzés elvét is figyelembe vevő alakot, például:

szimpad » szinpad [mp↔np], színpad [i↔í]
teccik » tetszik [cc↔tsz]
aggyá » adjá [ggy↔dj], adjál [szó végi l↔Ø]
kiszcica » kiscica [szc↔sc]
egésség » egészség [ss↔szs]
pallament » parlament [ll↔rl]
tejjes » teljes [jj↔lj]
báttya » bátyja [tty↔tyj]

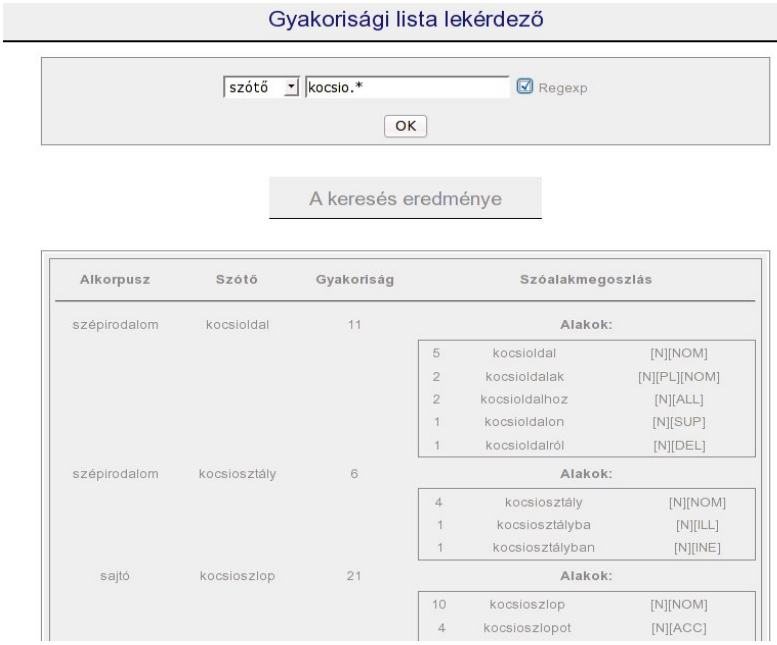
A hibásan írt szavak kezelésének további erőforrása a leggyakrabban hibásan írt szavak gyűjteménye (mintegy 120 ezer tétel), amely javarészt az MTA Nyelvtudományi Intézetében zajló helyesírási tanácsadás gyakorlatából származik, a gyakran előforduló, tipikus hibák gyűjteményén alapszik.

Amint látható, a hibás alakban keresett szót több szűrőn keresztül ellenőrizve jutunk el a helyesen leírt alakig, amely még korántsem a végső alak, mivel több lehetséges megoldás esetén itt is szükség lehet még a kérdező általi egyértelműsítésre.

A **helyesiras.hu** célközönségeként nemcsak a magyarországi nyelvhasználókra, hanem a legtágabb értelemben vett magyar nyelvközösségre is gondolunk. Éppen ezért a szótár nemcsak a magyarországi magyar nyelvváltozatok szókészletét tartalmazza majd. (Természetesen a magyarországi magyar nyelvváltozatok közül a kizárólag beszélt nyelvi formában élő területi, illetve csoport- és rétegnyelvi változatok problémáival, tehát azon lexikai tételekkel, amelyeknek nincs és esetleg nem is lehet kodifikált helyesírásuk, nem foglalkozunk.) Már az alapvető erőforrásnak számító MNSz. is tartalmaz mintegy 23 millió szövegszónyi határon túli korpuszt, amely mellé bekerül egy közvetlen kölcsönszavakból álló, annotált, ún. ht-szólista (**http://ht.nytud.hu**). Ez még kiegészül az MTA határon túli kutatóállomásai által gyűjtött magyar etimonú földrajzi nevek, intézménynevek, díjak és címek megnevezéseit tartalmazó szóanyaggal. (A földrajzi neveknek a Földrajzinév-bizottsággal való egyeztetése ehhez elkerülhetetlen.) Ez utóbbiak országra utaló megkülönböztető jelzéssel lesznek ellátva, így lehet ugyanis kezelni a nyelvváltozatok helyesírási vetületének esetleges ütközéseit is, bár az ilyen esetek számát a minimálisra kell szorítani a helyesírás egysége érdekében.

Hasonló módon kezelhetők a jövőben egyes szaknyelvi részszótárak is. Ezek, illetve általában a szaknyelvi helyesírás kérdései további bővítési-fejlesztési lehetőséget

kínálnak a **helyesiras.hu** portál számára. Ezek megoldásához az egyes szakmák művelőivel is ki kell építeni a megfelelően szoros munkakapcsolatot.



1. ábra. Az adatbázis már működő lekérdezőfelülete.

Hivatkozások

1. Kis Ádám: Gépszerű helyesírás. Az akadémiai helyesírási szabályzat és a számítógép. http://mek.iif.hu/porta/szint/tarsad/nyelvtud/gepscikk/ (1997)
2. Kis Ádám: Az akadémiai helyesírási szabályzat és a számítógép. Magyar Nyelvőr 123 (1999) 149–168.
3. Laczkó Krisztina, Mártonfi Attila: Helyesírás. Osiris Kiadó, Budapest. (2004)