Javaslatok a tárgyilagos minősítéshez

Mikor jó egy helyesírás-ellenőrző?

A tudománytörténészek számos esetről tudnak,
amikor ugyanazt a technikai újítást
vagy tudományos felismerést egymástól függetlenül,
egy időben többen is bejelentették.
Hasonló esett meg most is: egyszerre két professzionális,
a magyar nyelvű helyesírást ellenőrző program bukkant fel
a hazai szoftverpiacon.
Rögtön felvetődik a kérdés: hogyan lehet
tárgyilagosan összehasonlítani, egyértelműen kimutatni
két (vagy több), azonos célú program milyenségét,
minőségét?
Hogyan lehet a használhatóságot, a végeredményt
vizsgálni, függetlenül a belső működés módjától,
jelen esetben attól, hogy a helyesírás-ellenőrző
a szóalakok analizálásán vagy generálásán alapszik-e.

Tudjuk, hogy az angol nyelvre már nagyon korán számtalan helyesírás-ellenőrző (spelling checker) készült. Ebben több más tényező mellett bizonyára szerepet játszott az angol nyelv tipológiája is.

De a magyarhoz tipológiailag közelebb álló nyelvekre (német, cseh) szintén jóval előbb kidolgozták a PC-n működő helyesírás-korrektorokat.

A magyar helyesírás-ellenőrző talán a következő négy „nyelvi ok" miatt késett sokáig (a társadalmi, gazdasági, technikai okokat nem említve):

— A magyar toldalékolási rendszer igen változatos (egy főnévnek képzett alakok nélkül is több mint 1000 alakja lehet, például asztalt, asztalomat, asztalaimat stb.), és ezt megsokszorozhatják a továbbképzett alakok: asztalos, asztali...

— Némely toldaléknak több (kettő, három, sőt öt) alakváltozata is él (például asztal-t, madar-at, szék-et, kalapot, ördög-öt), és ezek sem viselkednek mindig „logikusan" (például viz-e-t, de: nyil-a-t).

— A magyarban (is) könnyű új szavakat összetétellel képezni (például asztalláb, gépkocsi). A kb. 70 igekötővel ugyancsak más-más jelentésű igéket hozhatunk létre.

Ez a jelenségcsoport a szavak elválasztását nehezíti, mert az összetett szavakat az alkotó tagok határán kell elválasztani (szem-üveg, vas-út, megőriz), és befolyásolja a szavak egybevagy különírását is.

— Egyes magyar szótövek nyelvtörténeti okok miatt változnak a toldalékolásnál (bokor, de: bokrok; víz, de vizet; ló, de: lovat).

Objektív vizsgálati módszer

Mindenekelőtt rendszerezni kell azokat a tulajdonságokat, amelyeket elvárunk egy jó helyesírás-ellenőrzőtől. Talán — mint minden szoftver esetében — az alábbi fő összetevőktől függ a helyesírás-ellenőrző milyensége is.

— Technikai paraméterek: memóriaigény, géptípus, az operációs rendszer fajtája, a feldolgozás sebessége.

— Komfortosság: a felhasználónak mennyire kényelmes, mennyire „kézreálló" a program, milyen mértékben működik együtt a helyesírás-ellenőrző a szövegszerkesztőkkel, lehetséges-e a hibás alakok cseréje.

— Nyelvi, szakmai helyesség: a program a vállalt feladatot hogyan végzi el, mennyire szakszerű, milyen mértékben felel meg a helyesírási, esetleg nyelvhelyességi szabályoknak.

— Kereskedelem, szolgáltatások: a szoftvertermék ára, a program továbbfejlesztett verzióival való felújítása, a program hozzáférhetősége és védelme.

Első „kályha"

Milyen aspektusokból lehet a helyesírás-ellenőrző nyelvi, szakmai helyességét vizsgálni. E szempontok alapján nemcsak egy adott helyesírás-ellenőrző szakmai oldala ítélhető meg, hanem több ilyen jellegű szoftvertermék is érdemben hasonlítható össze. De előbb nézzük meg, milyen kiindulópontok felől közelíthető meg ez a kérdéscsoport.

— Feltételezhetjük, hogy a helyesírás-ellenőrző a szavak szintjén működik. Pragmatikai, szemantikai, szintaktikai, stilisztikai hiányosságokra nem hívja fel a figyelmet. Így nem követeljük meg tőle, hogy hibát jelezzen a következő mondatokban:

Az Eiffel-torony Budapesten van. Minden reggelek kávé iszom. Adidas cipőbe futok reggelente.

— Vesszőhibák, a mondatvégi írásjelek használatának ellenőrzésére nem alkalmas.

Ennek figyelembevételével a helyesírás-ellenőrző szószintű működése során a vizsgált szöveg szavankénti ellenőrzésekor 4 alapeset lehetséges:

	Input (A vizsgált szó)	Output (A jelzés)	A program működése
1	Helyes szó	Jó	Jó
2	Helyes szó	Rossz	Rossz
3	Hibás szó	Rossz	Jó
4	Hibás szó	Jó	Rossz

Eldöntendő, hogy az 1., 2., 3., 4. esetek egyenrangúak-e, ugyanazon súlyozással veendők-e figyelembe a megmérettetésnél. (Szerintem semmiképpen sem. Nyilvánvaló, hogy egészen súlyos vétség az ellenőrző program részéről, ha egy hibás szóalakra nem hívja fel a figyelmet, „elengedi" azt. Ennél kisebb hibának tartom, ha egy jó szóalakot kérdésesnek — netán rossznak — ítél, más szóval nem ismer fel.)

Ajánlás az esetek minősítésére, a „súlyokra":

1 = 1 piros pont (+ 1 pont)
2 = 50 fekete pont (– 50 pont)
3 = 25 piros pont (+ 25 pont)
4 = 250 fekete pont (– 250 pont)

Eldöntendő, hogy a nyelvészszempontból korrekt vizsgálatok eredményei milyen faktorral szerepeljenek a végső osztályzat képletében, és ez az osztályzat milyen súllyal essen azután

latba a technikai paraméterek és a komfortosság mellett a „végbizonyítvány" kiállításakor.

Második „kályha"

A helyesírás-ellenőrző vizsgálata összetett feladat, mivel sokféle helyesírási hibát lehet elkövetni. Minden helyesírás-ellenőrzőnek vannak erős, de gyenge oldalai is. A sokoldalú tesztelés vezethet csak tárgyilagos eredményre. (Ahogy a személygépkocsik összehasonlításánál is több jellemzőt szoktunk figyelembe venni, nem egy-két kiragadott paramétert.)

Az egyetlen szempont szerinti vizsgálat félrevezető lehet. Fontoljuk meg, hogy a gazdag szóösszetételi lehetőségből adódóan — amennyiben ezt tágan értelmezi a helyesírás-ellenőrző — furcsa dolgokkal találhatjuk szemben magunkat, ha erre élezzük ki a próbát. Egyetlen példa: feltesszük, hogy ismeri a program a „rá" és a „lyuk" szót, a tág értelmezés miatt pedig elírásként értelmezheti az rájuk szót (rályuk) — mint a rá és a lyuk összetételét.

Harmadik „kályha"

A helyesírás-ellenőrző tesztelésének legelején tisztázni kell, hogy eleve milyen speciális szócsoportok vizsgálatát nem vállalja, milyen nyelvi jelenségek felismerésére, javítására nem is kívántak az alkotók megoldást találni. Többek között ilyesmire:

a) Tulajdonnevek írásmódjának vizsgálata (ezek között szerepelhetnek a mozaikszavak is). Hiszen a mondatok elején álló — és ezért nagybetűvel írt — szavakról nehéz eldönteni, hogy tulajdonnév-e vagy tévesen írt köznév.

b) Az idézetekben lévő idegen nyelvű szavak felderítése, kezelése.

Negyedik „kályha"

Az elválasztás problémakörénél a következőkre kell felhívnunk a figyelmet, hiszen tudnivaló, hogy a helyesírás-ellenőrzőknek gyakran éppen az automatikus elválasztást kell majd ellenőrizniük.

Magyar szavakat helyesen elválasztó programot nem nehéz írni, ha megelégszünk kb. 90 %-os helyességgel. Minden újabb százalékért azonban már komolyan meg kell küzdenünk. Az elválasztás alapvető szabályait ugyanis egyéb szempontok és körülmények gyakran felülbírálják. (Nem me-gír, hanem meg-ír.) Úgy vélem, hogy a helyesírás-ellenőrző elválasztással foglalkozó részének vizsgálatakor irányt adhat az is, hogy a következő 5 elválasztási szint közül melyeken működik jó hatásfokkal a program, és mely szinteket nem ismeri. Szerintem 5 minőségi szint különböztethető meg az elválasztás automatizálásában:

— Az alapszabály ismerete: le-het, vá-ros, di-ó-fa, las-san.

— A hosszú kétjegyű mássalhangzók helyes kezelése: hosz-szú, asz-szony, kong-resz-szus, eny-nyi-re. (Közbevetőleg: ezeket az eseteket nem lehet rejtett, előre elhelyezett kötőjellel megoldani.)

— Az olyan összetételek megfelelő elválasztása, ahol az összetétel egyik tagja igekötő vagy a leg- melléknévfokozó szócska, vagy a névelők valamely származéka: fel-ír, meg-ad, ösz-sze-olvas, leg-ú-jabb, ez-e-lőtt.

— A két vagy több szó összetételéből keletkezett szavak helyes szétválasztása a szóhatáron: szak-em-ber, hír-adó, bör-tön-ab-lak, ki-lo-gramm.

— Ha a szövegkörnyezettől függően kell elválasztani a kérdéses szót: meg-int/me-gint, gép-e-lem/gé-pe-lem, kikért/ki-kért.

A vizsgarend

Most nézzük meg, hogy a „kályháktól" milyen lépésekben célszerű elindulni. Ezek között két általános (az 1. és a 2.), valamint három speciális van (a 3., a 4. és az 5.).

1. Ismert és nem ismert tőszavak

Felderítendő, hogy milyen szavakat ismer (és milyen szavakat biztos, hogy nem ismer) a helyesírás-ellenőrző. (Természetesen felhasználótól függően lehetnek különleges igények — például bizonyos szakmai szavak —, és ezért izgalmas a bővíthetőség kérdése.)

Fontos, hogy a fel nem ismertek a magyar szavak általános gyakorisági listáján hol helyezkednek el. A gyakorisági értékek (csökkenő sorrendben) erősen esnek. Néhány száz szó adja a szövegszavak nagyobb felét. Ezért a fel nem ismert szavakat súlyozni kell gyakoriságuk alapján. Szerencsére már vannak szógyakorisági listáink:

• Füredi Mihály: A mai magyar nyelv széppróza gyakorisági szótára (Akadémiai Kiadó, 1989). Félmillió szövegszót tartalmazó korpuszból kiindulva felsorolja a leggyakoribb 3500 szót és minden előfordult toldalékos alakot.

• Csirikné — Csirik János: Újságnyelvi gyakorisági szótár I-II. (Szeged — Bp., 1986). Több újság teljes anyaga alapján, 201 000 szövegszót összeszámolva közli a legalább kétszer előfordult szavak listáját.

Javaslatom: a Füredi-féle lista minden elemérőI tudni kell, hogy ismeri-e a vizsgált helyesírás-ellenőrző. Természetesen nem csupán tőszó formájában, hanem leggyakrabban előforduló toldalékos alakjában is.

2. A pontozás számszerűsítése

A vizsgálat magja lenne próbaszöveg segítségével a helyesírás-ellenőrző működésének tesztelése. Legcélszerűbbnek látszik a szavankénti ellenőrzés kapcsán vázoltakat egy minimum 100 000 szövegszót tartalmazó, különböző (szépirodalmi, újságnyelvi, hivatalos iratokból idézett stb.) szövegfajtákból álló, és gépelési hibákat (kb. 10 000, véletlenszerűen szétszórt elütést) magában rejtő szövegen megmérni.

3. Iskolai dolgozatok ellenőrzése

Javaslom, hogy 100 gimnazista, egyenként kb. 300 szavas magyar dolgozatát vizsgálja meg a helyesírás-ellenőrző program. A dolgozatok a diákok által leírt formában kerüljenek a helyesírás-ellenőrző elé. Az értékelésnél számításba kell venni, hogy a különböző típusú hibáknál és az egy, két, három pontos hibák esetében (a tanár által pirossal egyszer, kétszer, háromszor aláhúzott tévesztések felismerésekor, értékelésekor) miképpen viselkedik a helyesírás-ellenőrző.

4. Célirányosan összeállított lista

A fenti vizsgálatot ki kellene egészíteni rosszul írt szóalakok kb. 2500 elemű (szakemberek által összeállított) listájának kijavíttatásával. A listának az elemei (toldalékolt alakjukban) mintegy reprezentálnák egy elképzelt helyesírás-tanító kurzus témaköreit. Csak példaképpen: a kétfajta „j", a szövégi „ő", a múlt idő „t"-je, magánhangzók/mássalhangzók ejtési időtartamának eltévesztése, egybeírás/különírás, ige felszólító módja.

5. Az elválasztás ellenőrzése

A speciális vizsgálatok közé kell iktatnunk a problémák sorát magában rejtő elválasztás helyességének tesztelését. Az elválasztási képesség vizsgálatára egy kb. 25 000 szóalakot tartalmazó szöveg alkalmas. Ennek a szövegnek minden szavát, minden lehetséges helyen el kellene választania a programnak. A program által adott eredményt kell utána összehasonlítani a kézzel helyesen elválasztott anyaggal.

Kiss G. Gábor