A magyar helyesírás-ellenőrzők mai állása

Naszódi Mátyás, e-mail: naszodim@morphologic.hu

MorphoLogic, 1122 Ráth György utca 36.

Kivonat A helyesírás-ellenőrzők jósága függ az előállítás módjától, karbantartásától, de az adatbázis méretének növekedésével objektív korlátokba ütközik a minőség. Jelen cikk kitér az objektív minősítés módszertanára, elvi korlátaira. Összeveti az elérhető helyesírás-ellenőrzőket. Megkísérli pártatlan módon összevetni az elérhető programokat, és megmutatni, hogy a nyelvi adatbázis építésénél alkalmazott módszereknek milyen előnyük, hátrányuk van. A cikk végén keresi a további hatékony fejlesztés irányát.

Kulcsszavak: szóellenőrzés, statisztika, nyelvminőség

1. Bevezető

A helyesírás-ellenőrzők a személyi számítógépek megjelenésével terjedtek el. Angol, majd francia, spanyol, olasz nyelveken íróknak könnyítette meg a dolgát. Magyarra készített szpellerek a 90-es évek elején jelentek meg. A késést nyelvünk összetettsége okozta. Míg az indoeurópai nyelveknél elegendő pár százezer szóalakot azonosítani egy gépi lektornak, addig magyar, finn, török nyelveknél az eszköznek több milliárd alakot kell felismernie.
Mostanában jelent meg a palettán a Microsoft és a Google ellenőrzője. Nyelvtanunk hivatalosan is megújult, melyet az eszközöknek is követnie kell.
Jelentek meg helyesírás-ellenőrzők tesztjéről szóló cikkek[1][2], de ha tesztanyag az eszköz előállításánál szerepet játszott, akkor arra az eszközre aránytalanul jó eredményhez vezet.

2. Technikai áttekintés

A 80-as években olvastam egy írást azzal a címmel: Hogyan készítsünk helyesírás-ellenőrzőt?. A recept a következő: egy szótárba gyűjtsük az ismeretlen szavakat. Ha a program találkozik egy új szóval a szövegben, a felhasználó döntsön, kell-e. A szavak gyakorisági statisztikája miatt a szöveg felét a szóalakok kis hányada lefedi – akár ezer szó – a módszer az angolra be is válik. A magyarra az ilyen próbálkozás teljes kudarcba fulladt.
Ragozó, agglutináló nyelvekben túl sok szóalak létezik. Nem lehet összegyűjteni annyit, hogy ezekkel elfogadható lefedettséget érjünk el. 2016-os cikkemben[3] említem, hogy exponenciálisan csökkenő valószínűséggel előforduló egyedek gyűjtése megfelelő hibaszázalékkal csak korlátos mennyiségben lehetséges. Nyelvi adatbázisok építésénél ez maximum 200 000 körüli érték. Hasonló adatokat említ

Kornai András *Frequency in morphology*[4] című írásában.
Szavakat kell gyűjteni, és generatív modell alapján kell előállítani a szóalakokat – vagy a nyelvi leírás alapján kell visszavezetni a szóalakot morfémák sorozatára. Olyan megoldásokat, melyekkel a szóellenőrzők nyelvi adatbázisainak mérete a kritikus alá csökkenhetett, csak a 80-as évek végétől készítettek.
A generatív modell miatt a szóalakok nem feltétlen gyakoriságuk miatt kerülnek be a készletbe. Ha egy szót regisztrálunk, akkor annak minden szabályosan toldalékolt alakját is, még ha nem is használatosak. Ezek olyan közel lehetnek egy gyakori szóalakhoz, hogy nagy az esélye, hogy a helyes szó elütése következtében került a papírra. A *tan* főnév *-i* képzős alakja tárgyesetben *tanit*, amit gyakran írnak le a *tanít* helyett. A magyar nyelv nagyon sűrű, különböző szavak nagyon közel vannak egymáshoz. Emiatt magyar nyelvnél az előbb említett probléma gyakrabban merül fel, mint angolban, németben, olaszban...

3. A választék

Jelenleg a következő általánosan használható helyesírás-ellenőrzők léteznek:

- Helyes-e?: A MorphoLogic terméke. MS Office-ok része volt. Sok más alkalmazásba került bele. Megjelenése: 1992. Alkotói: Prószéky Gábor, Pál Miklós, Tihanyi László. Jelen fejlesztők közül kiemelném Novák Attilát.
- Lektor: Seregy Lajos nyelvész és a MicroSec programozóinak terméke. Elsőnek, még a 80-as évek végén jelentették be, de végül 1992-ben lett belőle eszköz. Sajnos azóta nem fejlődött.
- Helyeske: Elekfi László ragozási paradigmaszótárára épülő véges automata elven működő ellenőrzőt Farkas Ernővel készítettem. 1993-ban lett a MorphoLogic terméke, de azóta nem fejlődött tovább.
- ISPELL, MYSPELL és HUNSPELL: a szabad szoftverek világában fejlődő vonal. Két szempontból is jelentős. Egyrészt a HUNSPELL, a legfejlettebb változat magyar gyártmány. Szabad szoftver lévén sok helyen használják böngészőnél, levelezőnél. Legmarkánsabb javulását a Szószablya[5] keretében végezték rajta. Alkotója Németh László. A nyelvi leírásnak számtalan „bedolgozója" volt.
- Kimmo-féle kétszintű morfológia: a XEROX-nál, IBM-nél használják. Ezek magyar nyelvi kiindulási anyagát a MorphoLogic állította elő, de nem helyesírás-ellenőrző céljából, és azóta sokat változott.
- A Microsoft ellenőrzője: Egyetlen program kezeli a különböző nyelveken írt szövegek javítását. A Microsoft ellenőrzője 2015 óta működik. Az új MS Office-ok szerves része, emiatt sokaknak lesz hozzá szerencséje.
- Hozzáférhetők webes felületű helyesírási tanácsadók[6][7][8]. Ezeket három okból nem vettem górcső alá.
  1. Tömegfelhasználásban kevésbé játszanak szerepet.
  2. Nehéz a 2-es típusú hibát detektálni (lásd később)
  3. Nem lehet vele nagy tömegű anyagot tesztelni.

A Helyes-e, Helyeske, HUNSPELL forrásai számomra hozzáférhetőek, ezért minősítéseimet megalapozottak, míg a Microsoft forrásanyagára csak a viselkedés alapján következtethetek.

4. Mennyiségi teszt lektorálatlan szövegen

Kétfajta tévedés lehetséges.

1. Helyes szót nem ismer fel, tehát hibásnak tart.
2. Helytelen szót helyesnek minősít, ezért elfogadja

Ha a szövegszerkesztőben az 1-es típusú tévedés fordul elő, a program jelez. A második esetben a felhasználónak nem jut tudomására a szöveghiba, emiatt a szöveg javítatlan marad. Kiss G Gábor cikkében[9] 10-szeres súllyal bünteti a 2-es hibát. A fent vázolt gondok miatt ennél jóval nagyobb a jelentősége.

4.1. Elvi megfontolások

Az, hogy egy karakterlánc magyar szó-e, valószínűségi kérdés. Hibásnak ítélt szó is lehet helyes: *nemecsek*, *frisssssss*, de a *böszmeség* is csak azóta ismert, mióta kiszivárgott az öszödi beszéd. A szövegekben előforduló sztringek többségéről minden magyar anyanyelvű határozottan tud dönteni. Ennek az oka, hogy a valószínűségek elég karakterisztikusak. A többség vagy megüt egy szükséges szintet, vagy egy nagyon alacsony szint alatta marad. A kettő közötti hányad, mely esetekben esetleg még nyelvészek sem értenek egyet, elenyésző.
A nagy valószínűségű szavaknál a statisztikai becslés megbízhatósága elfogadható, de a szavak többségénél, még ha megütik az elfogadható szintet, a statisztikai becslés megbízhatósága alacsony.

1. A szóalakok előfordulási valószínűsége szövegkörnyezettől függ.
2. A szóalakok előfordulási valószínűsége írótól függ.
3. A szóalakok előfordulási valószínűségét csak a gyakoribb esetekben lehet megbízhatóan becsülni.
4. Ha lenne is megbízható becslés, ennek felhasználása a mai számítástechnika mellett túl nagy erőforrást igényelne.
5. A felhasználót irritálná, ha a szavakról a program nem jó-rossz választ adna. Még a „talán” válasszal sem tudna mit kezdeni.

Mindezek miatt a nyelvi adatbázisok és az erre épülő programok igen-nem döntést hoznak a szóalakokról, melynek egy küszöbszint elérése lehet az alapja.

4.2. Technikai meggondolások

Hogy egy szót elfogad-e vagy sem a program, a futtatás választ ad. Arra a kérdésre viszont, hogy helyes-e a szó, nincs objektív mérce. Vagy nagy kompetenciával rendelkező emberi erőforrást kell igénybe vennünk, vagy le kell mondanunk a szavak egyedi minősítéséről. Mivel a teszt során feldolgozandó anyag mérete tetemes, az emberi minősítés nem jöhet szóba.
Az eszközök összevetéséhez nem kell vizsgálni azokat a szavakat, melyekről mindegyik azonosan dönt. A relatív minősítésben csak az eltérések játszanak szerepet.

A jelen vizsgálatnál az eltérően bírált szavak száma 1000-es, 10 000-es nagyságú. Az egyszerű előfordulási statisztika nem segít, mert számos, mindenki által elfogadott szóalak létezik, mely egyszer sem volt leírva. (Valószínűleg az a szó, hogy *testetlenítettséggel* most lett először leírva, de helyes szó.) Egyes hibás alak előfordulási gyakorisága ezt jóval meghaladja. (*Hüje, írts, szervíz...*)
Ha csupán két ellenőrzőt vetünk össze, akkor kikeressük azokat a szavakat, melyeknél ellentétes döntés született. A legegyszerűbb kiértékelés, ha a hibás döntések számát vetjük össze. Amelyiknél kisebb ez az érték, az lehet a jobb ellenőrző. Ennél egy fokkal jobb, ha a 2-es típusnak nagyobb súlyt adunk.
Ha ismernénk a szavak valószínűségét, súlyozhatnánk vele. Egy gyakori szó elhibázása nagyobb baj, mint egy ritkáé. Hát még egy gyakori hiba megengedése. A legpontosabb minősítés az lenne, ha azt is felismernénk, mekkora kárt jelent egy ilyen téves szó. Vagyis a globális képlet:

$$\sum W(alak) = \sum_{alak} p(alak) * e(alak), \tag{1}$$

ahol W a hiba súlya, p az alak valószínűsége, e pedig a hiba által okozott kár, tehát a helyes szóalakoknál $e(alak) = 0$.
Ha a kárt abban mérjük, hogy milyen szóalakok elírásából adódhatnak, a következő becslést adhatjuk.

$$\sum_{alak} p(alak) * e(alak) = \sum_{alak} p(alak) * \sum_{alak2 \in Helyes} \frac{p(alak2)}{e^{m(alak,alak2)}} \tag{2}$$

ahol m a két szóalak közti távolság, amit már mérni, számolni lehet[3].
Ha nincs a valószínűségre sem jó becslés, akkor egyszerűbb képletet kell alkalmazni. A korábbiak szerint a valószínűség becslése csak a gyakran előforduló szavaknál lehetséges.

4.3. A tesztkorpusz

A Népszabadság 1993-as szerkesztősége rendelkezésünkre bocsátott egy nagyobb mennyiségű anyagot. Egyéb forrásunk nagyobb hányada a magyar szpellerek megszületését megelőző időkből származik.
A tesztkorpusz mérete 5 585 000 karakter, 745 900 szó 131 000 különböző szóalak. A 30 leggyakoribb szóalak lefedi a szöveg 25 %-át. Az első 15 alak:

a	54394		*hogy*	11215		*volt*	2356		*vagy*	2141		*kell*	1579		*el*	1427
az	20280		*A*	9789		*de*	2277		*s*	2059		*szerint*	1533		*ki*	1356
és	13520		*nem*	8658		*már*	2167		*még*	2054		*van*	1494		*mert*	1265

A ritkán előforduló szóalakok számából látszik, hogy a többség csak egyszer fordul elő:

1-szer fordul elő	77820 szóalak	2-szer fordul elő	21351 szóalak
3-szor fordul elő	8604 szóalak	4-szer fordul elő	5085 szóalak
5-ször fordul elő	3299 szóalak	6-szor fordul elő	2261 szóalak
7-szer fordul elő	1699 szóalak	8-szor fordul elő	1272 szóalak

A 131 000 szóalakból az ellenőrzők más-más szavakat tartottak hibásnak:

Office 6	Office XP	Office 2002	Office 2016	HUMOR 97	HUMOR 2000
15500	12900	12000	15500	11000	16000
ISPELL	MYSPELL	HUNSPELL	Libre Office	Lektor	Helyeske
17500	17900	13300	13100	17000	20300

A táblázat a 2-es típusú hiba becslésére nem ad lehetőséget. Vizsgáljuk meg, melyek azok a szavak, melyeket az egyik ellenőrző elfogad, a másik elutasít.

	Off 6	OXP	2002	2016	O 97	H 97	2000	ISP	MYS	HUN	Lekt	Heke
Office 6		4166	3516	6258	3129	4897	1402	5113	2615	3449	2291	1527
Office XP	1552		706	3980	861	3027	1996	3828	1565	2721	2129	1968
Office2002	926	730		3664	206	2485	1390	3940	1716	2165	1920	1364
Office2016	2794	3130	2790		2872	3569	2925	3341	2997	3467	2436	2929
Office 97	750	1096	416	3958		2435	1253	4181	2033	2126	2027	1295
HUMOR97	399	1143	577	2535	316		126	2960	991	730	833	481
HUM2000	1918	5126	4496	6905	4148	5139		5541	3140	4008	3103	2414
ISPELL	4758	6086	6144	6449	6204	7103	4672		1344	5309	3666	4672
MYSPELL	4628	6192	6312	8474	4625	7501	4637	3743		5274	4314	3847
HUNSPELL	1252	3138	2558	4733	2307	3030	1294	3468	1063		1599	1062
Lektor	3837	6988	6055	7744	5951	6876	4132	5567	3846	5342		4062
Helyeske	6352	9407	8779	11218	8500	9804	6722	9184	6659	8085	7342	

Ha a szóalakok előfordulási gyakoriságát is figyelembe venném, a fenti teszt nem mutatna ki még ilyen kis különbséget sem. A lefedettség mindegyiknél 97 % körüli érték. Szubjektív módon érzi a felhasználó, hogy melyik a jobb, de ezt nehéz így számszerű adattal igazolni. Emiatt finomabb különbségtételre van szükség.

5. Teszt mesterséges tesztanyaggal

A magyar ABC kisbetűiből álló legfeljebb 6 karakteres sztringeket ellenőriztem egy ponttal lezárva. Majd 2 200 000 000 szóalak keletkezik. A szó végi pontot mindegyik program aszerint kezelte, hogy kötelező vagy nem a szó után.

2 176 782 336	Office XP	Office 2002	Office 2016	HUNSPELL	Helyeske
futási idő	6 óra	3 nap	10 nap	30 perc	1 perc
helyes szavak	600 037	594 409	3 910 312	776 515	281 511
ebből ponttal a végén	80	101	1 298 036	290	68

Ezek az adatok még markánsabban mutatják a különbségeket. A HUNSPELL-nél azért magasabb a ponttal végződők száma, mert a római számokat csak ponttal lezárva fogadja el. Az Office 2016-nál az a hiba állt elő, hogy rövidítéseket is megenged szóösszetételben. Ez okozza a mérhetetlen nagy számot.
Az eggyel felette levő sorban az Office 2016 imponáló adata onnan ered, hogy rengeteg hibás szóalakot fogad el: sok kötőjellel toldalékolandó szót kötőjel nélkül. Ráadásul ezeket szóösszetételben is használja. Ilyen mellélövések mellett az egyéb hibák száma eltörpül.
Az Office 2016 hétkarakteres szavaknál 1 évig futott volna! A Helyeske imponáló ideje lenyűgöző akkor is, ha a tesztágy különböző volt. A sebesség egy szövegszerkesztőnél nem lényeges. A szöveg beírása jóval lassabb ennél. Azt is figyelembe lehet venni, hogy az algoritmusok helyes szavaknál sokkal hatékonyabbak, mint hibás szó esetén, és ez utóbbi tesztnél szinte mindegyik szó hibás volt.
Érdemes a kereszteszt adatait is megtekinteni, hisz ebből már olyan adathalmazok keletkeznek, melyeket közvetlen emberi erővel nem, de mintavételezés után érdemes lenne vizsgálni.

	Office XP	Office 2002	Office 2016	HUNSPELL	Helyeske
Office XP		48 936	3 440 268	333 014	98 563
Office 2002	54 564		3 446 664	327 906	95 098
Office 2016	129 992	130 761		168 872	105 112
HUNSPELL	156 536	145 800	3 320 668		109 458
Helyeske	417 089	407 996	3 733 913	604 462	

Az adatok most is az Office 2016 oszlopában a legnagyobbak. Ha belenéz valaki az állományokba, kiderül az oka. Több mint 3 000 000 hibásan elfogadott szó. A becslés onnan ered, hogy véletlenszerűen kiválasztva az elfogadott szavakból egy részhalmazt, annak legalább három negyede helytelen forma.
Utólag még ráengedtem ezt az iromány és a Tinta kiadó helyesírási szótárát is az ellenőrzőkre. A tanulság kettős. Egyrészt a kiadott szótárban is találtam hibákat. A másik, hogy – mivel itt többnyire helyes szavak vannak felsorolva – a MS új ellenőrzője gyakori jó szavakat sem mindig ismer fel.

6. Szubjektív kiértékelés

A szubjektív kiértékelés a kereszteszt alapján objektív módon kinyert szóalakok vizsgálatából származik.

– Helyeske: Toldalékolása a ragok és jelek esetén a legpontosabb. A képzőknél kicsit túlgenerál. Korlátlan számú képzőt elfogad, (*legeslegellovasíthatatlanítottabbak*), és olyan toldalékokat is kezel, melyeket mások egyáltalán nem (*zsákosdi*). Az igeneves összetételekkel (*macskafogó, padlófeltörlés...*) nincs baj, a számnevek is pontosak, de egyéb összetételt ritkán enged meg. Kötőjeles összetétele szabad. Tiltó szabályok nincsenek. A betűn, számjegyen kívüli karaktereket nem kezeli. (*§-ának, °C, %-ot...*) A tulajdonnevek kisbetűsítését (pl. *-i* képző) algoritmikusan elvégzi. A forrásleírása a legtömörebb.
– HUNSPELL: Akad pontatlanul osztályozott szó. Szóösszetétele engedékeny, de legalább nem mond ellent az általános nyelvi szabályoknak. Szókészlete elég jó. Ez kezeli egyedül megkülönböztetően a rövid és a hosszú kötőjeleket. Van lehetőség tiltó szabályok alkalmazására, ezért elvileg még sokat javulhatna – ha lenne egy metaszintje a leírásoknak. A 6-3-as szabály ugyan nincs benne, de ritkán téved. A tulajdonnevek kisbetűsítését (pl. *-i* képző) algoritmikusan elvégzi. Jelenleg csak ez engedi meg felhasználói szótárában a ragozható tételek felvételét. Adatbázisa súrolja a kezelhető méret határát – mintegy 150 000 tétel.
– Office XP: Szókészlete elég jó. Van pár hiba a toldalékolásban – lelke még a régi 16 bites, ahol korlátok voltak a leírás összetettségére. Szóösszetétele elfogadható – talán a betűvel írt számok körül lehetnek nagyobb gubancok. Sok betűn és számon kívüli szót is jól kezel. Már nem fejlődik. Nem is érdemes, mert van jobb helyette. Adatbázisa súrolja a kezelhető méret határát – mintegy 150 000 tétel.
– Office 2002: A toldalékolása elég pontos, és szóösszetételben a legpontosabb. A tiltó szabályok hatékonyak. Sok betűn és számjegyen kívüli szót is jól kezel. Létezik metaleírás. A legjobban karbantartható. A kereszttesztek alapján legtöbbször ennek volt igaza a vitatott szóalakoknál. A 6-3-as szabályt már algoritmikusan kezeli. A tulajdonnevek kisbetűsítését szótári bejegyzésekkel oldja meg. A számok kezelése majdnem tökéletes. A felhasználói szótárban nincs lehetőség ragozható alakok felvételére. [10]Adatbázisa kezelhető méretű – mintegy 60 000 tétel.
– Office 2016: Egyedül a lefedettségi paraméterei jobbak a többinél, de ennek nagy az ára. A módszert nem ismerem, hogyan készült, de zsákutcának tűnik. Több sebből vérzik, és tulajdonképpen mindenben lemarad a többitől.
– Lektor: Látszik, hogy rég nem fejlődött, nem bővült. Én az 1993-as adatokkal dolgoztam. Főként tulajdonnevekből van hiánya, de egy-két gyakori köznév is hiányzik. Szóösszetételben nem erős. Toldalékolása precíz, kicsit konzervatív.

7. Összefoglaló

Az ellenőrzők mind hasznosak, de ez ma már nem elég. A minőség három összetevője: alapszókészlet, toldalékolási pontosság, szóösszetételek kezelése. Az ellenőrzők mindegyike valamiben erősebb a többinél, kivéve a legújabb MS szpellere. Kezdetekben a toldalékolásokon volt a fő hangsúly. A ragok, jelek használatára pontos leírások léteznek, de magyarban nem lehet csupán felszíni szabályok alapján osztályozni a szavakat. Ezzel kapcsolatos, hogy forrásleírása tömör legyen,

és lehetőleg ne lépje túl a 100 000-es tételszámot. Míg a HUNSPELL szóosztályozásának algoritmusa statisztikai eszközökre is támaszkodik[11], a Helyes-e szóbővítésénél mintaalapú az automatikus osztályozás módszere. Egyik sem kerülheti el az utólagos emberi felülvizsgálatot. Valószínűleg a neuronhálózatos megoldások sem eredményeznek jó megoldást, de ezt tudtommal még senki nem próbálta ki a magyarra, hacsak a Microsoft vagy a Google nem tette.
Ma a sarkalatos probléma a szóösszetételek kezelése. A kifinomult összetételkezelés érdekében szükség lenne pontosabb szabályrendszerre, amit az elemzők használnának. Addig is statisztikák segíthetnek, de a lehetséges szóösszetételek száma meghaladja azt a mértéket, amivel a statisztikai módszer elbír.
A lefedettség növelése nem kritikus. Persze szakszövegeknél fontos lenne kiegészítő szótárakra, amire volt is példa (orvosi, katonai Helyes-e?). Bővíteni lehet a szótárat, de inkább a toldaléktárakat kéne javítani, pontosítani. Minden bővítésnél figyelembe kell venni a 4.2 képletet a 2-es típusú hiba elkerülése érdekében. Ahol van kifinomult tiltó szabály, ott nagyobb esély van a javulásra.

8. Utóirat

- Nem teszteltem a Google tisztán valószínűségre alapozó, esetleg neuronhálózatos megoldását, annyira gyengének mutatkozik.
- Megszülettek az újított nyelvtant figyelembe vevő ellenőrzők.
- Fél év alatt a Microsoft másodlagos hibáinak száma harmadára csökkent. A becslésem szerint tíz éven belül elérik az elfogadható szintet.

Hivatkozások

1. Dömötör Andrea: HELYESÍRÁS-ELLENŐRZŐ PROGRAMOK VERSENYE
   `http://anyanyelvapolo.hu/helyesiras-ellenorzo-programok-versenye/`
2. ORIGO: Szövegszerkesztők helyesírásversenye
   `http://www.origo.hu/techbazis/szamitogep/20080923-megvizsgaltuk-a-helyesirasellenorzoket-microsoft-office-vs-openoffice.html`
3. Naszódi Mátyás: Statisztika megbízhatósága a nyelvészetben
   Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015) Szeged, 2015
4. András Kornai: Frequency in morphology
   In I. Kenesei (ed): *Approaches to Hungarian* Vol 4 (1992) 246-268
5. Németh László: A Szószablya fejlesztés
   Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003) Szeged, 2003
6. MTA, Nyelvtudományi Intézet: Helyesírási tanácsadó
   `http://xn-helyesrs-fza2j.mta.hu/helyesiras`
7. WEB: helyesírás http://www.magyarhelyesiras.hu/
8. webforditas.hu: Fordítási és helyesírási szolgáltatás
   `http://www.webforditas.hu/helyesiras`
9. KISS G. Gábor: Magyar helyesírás-ellenőrző programok ellenőrzése és összehasonlítása *Könyv Papp Ferencnek* Debrecen KLTE (1991) 325–333.
10. Novák Attola: emMorph http://e-magyar.hu/hu/textmodules/emmorph
11. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I. és Trón V.: A szógyakoriság és helyesírás-ellenőrzés
    In: I. Kenesei (ed): *Approaches to Hungarian* Vol 4 (1992) 246-268

State of the Hungarian Spell Checkers

Mátyás Naszódi, e-mail: naszodim@morphologic.hu

MorphoLogic, 1122 Ráth György utca 36. Hungary

Kivonat The quality of spell checkers depends on the applied method of constructing and maintaining their databases. The size of the database may limit the achievable quality. The present article discusses the methodology of the objective evaluation of spell checkers and the theoretical limits of testing. It attempts to compare the available programs impartially, and to show the advantages of the applied methods used for the construction of linguistic databases. Finally, it reviews the directions of possible improvement.

Keywords: spell checker, statistics, linguistic tool, quality of spell checker