A szógyakoriság és helyesírás-ellenőrzés

Halácsy Péter[1], Kornai András[2], Németh László[1], Rung András[3], Szakadát István[1]és Trón Viktor[4]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Média Oktatási és Kutató Központ
{halacsy,szakadat,rung}@mokk.bme.hu
[2] MetaCarta Inc.
andras@kornai.com
[3] Budapesti Műszaki és Gazdaságtudományi Egyetem, Kognitív Tudományi Központ
rung@itm.bme.hu
[4] International Graduate College of Language Technology and Cognitive Systems
Saarland University – University of Edinburgh
v.tron@ed.ac.uk

Kulcsszavak magyar webkorpusz, gyakorisági szótár, helyesírás-ellenőrzés, lexikográfia

Absztrakt A Szószablya projekt márciusában indult az Informatikai és Hírközlési Minisztérium támogatásával, a Budapesti Műszaki és Gazdaságtudományi Egyetemen működő Média Oktatási és Kutató Központ vezetésével. A projekt egyik kiemelt célja, hogy a magyar weboldalak szövegtartalma alapján egyedülálló teljességű magyar szógyakorisági szótárat, illetve ehhez kapcsolódó nyitott forráskódú alkalmazásokat készítsen. A cikkben röviden bemutatjuk a gyakorisági szótár készítésének menetét (§1), valamint a beépített szótárat használó alkalmazások, mint a helyesírás-ellenőrző szóanyagának hatékony bővítésére kidolgozott gyakorisági információn alapuló technológiát. A cikkben megmutatjuk, hogy a helyesírás-ellenőrzők pontosságát elsősorban a tőtáruk mérete határozza meg (§2), azonban a tőtár bővítése a Zipf törvény miatt egyre kisebb pontosságnövekedéssel jár (§3).

1. A magyar web és feldolgozása és a Szószablya gyakorisági szótár

A letölthető magyar web durva becslés alapján 20–25 millió oldalból áll, amelynek jelentős része szövegtartalom nélküli, idegen nyelvű, vagy duplikált oldal; egyes oldalak folyamatosan változnak, így a letöltését egy szűkebb időintervallumra kell korlátoznunk. A SzóSzablya első korpusza 2,4 millió weboldalt tartalmazó `lar0` 2002 decemberében készült el, míg a második, jóval teljesebb, 18 millió oldalt tartalmazó `gul0` 2003 őszét tükrözi. A letöltött nyers korpuszok szűrését és feldolgozását itt csak röviden ismertetjük (részletesen ld Németh 2003). A lapformátumok, a szövegtartalmak és a karakterkódolás normalizálását a Szószablya keretében kifejlesztett Hunnorm alkalmazás segítségével végeztük el. Ezután a szövegtartalom alapján történő duplikátumszűrésre került sor, amelyet a magyar ékezetes karaktereket nem tartalmazó szövegek kizárása követett. A szógyakorisági szótár elkészítéséhez a korpusz szövegeit tokenizáltuk a Huntoken program segítségével.[5]

A gyakorisági számítások jelentős torzulását okozhatja, hogy az azonos szolgáltatóhoz tartozó weboldalak egységes automatikus generált címszavai a tényleges gyakoriságuknál jóval nagyobb számban jelennek meg a gyűjtésben. Emiatt az oldalakat csak az első mondatzáró pont utántól vettük figyelembe, és az így kapott halmazon újabb tartalomalapú duplikátumszűrást végeztünk. A szűrések után a korpuszok az oldalak számát tekintve rendre 46, illetve 75%-kal csökkentek.

Az így kapott szöveganyag a nyers eredetihez képes jelentősen javult minőségileg, a magyar webköznyelvhez jobban közelítő gyakorisági adatok kinyeréséhez azonban egy további kritériumot alkalmaztunk: a Hunspell helyesírás-ellenőrző által elutasított szavak arányát. Ha ugyanis feltételezzük, hogy egy szöveg helyesírása egységes, akkor a helyesírás-ellenörző által helyesebbnek ítélt oldalakból nyert szóalakok is nagyobb valószinűséggel lesznek helyes szavak, vagyis a szóanyag minőségi előszűrése elvégezhető a helyesírásellenőrző lefedettségétől függetlenül. A Szeged Korpusz állományain elvégzett előzetes helyességvizsgálat alapján a tisztított gyakorisági szótárunk számára a 4% százalékos felső Hunspell-hibaküszöböt tartottuk megfelelőnek. Ez a kritérium a szűrt weboldalak további 60%-át szűrte ki.

Az így nyert anyag, a `web0` 433238 weboldalból áll, $N = 113385165$ szövegszót (token) és $V(N) = 4527456$ szóalakot (típus) tartalmaz. Becslésünk szerint a `web0` típusainak 80%-a helyes magyar szóalak, ami lényeges javulás a teljes webkorpuszhoz képest, amelynek szótípusain ez az arány 50% alatt van. Még szembetűnőbb a hibás szövegszavak előfordulásának csökkenése: ez becslésünk szerint 15%-ról 2,5%-ra csökkent. A teljesen automatikus módszerekkel nyert `web0` alapján elkészített első Szószablya szógyakorisági szótár elérhető a projekt honlapján.

A gyakorisági szótárak jelentősége felbecsülhetetlen mind a nyelvtechnológiában mind nyelvészeti kutatásokban, felhasználási területei számosak a fordítástámogatástól a pszicholingvisztikai kísérletekig. Az alábbiakban egy speciális területre koncentrálunk: azt elemezzük, hogy a gyakorisági információ milyen módon segítheti a a nyelvtechnológiai alkalmazások beépített szótárának hatékony bővítését, vagyis az alkalmazás lefedettségének a növekedését. A technológiát a helyesírás-ellenőrző alkalmazáson szemléltetjük.

2. Helyesírásellenőrzés és pontosság

A helyesírás-ellenőrzés alapvető célja, hogy segítségével a szövegben található hibák számát csökkentsük. Egy helyesírás-ellenőrző program tényleges hibáját az határozza meg, hogy (i) a szövegben lévő helyes szavak közül hányat utasít

[5] Elsősorban tulajdonnevek és a rövidítések későbbi hatékonyabb feldolgozása érdekében a szövegek mondatra bontását is elvégeztük.

el, és (ii) a helytelen szavak közül hányat fogad el. Ha tehát egy szöveg ellenőrzésekor minden 100 szó közül az ellenőrző csak egyszer hibázik (akár (i), akár (ii) típusú hibát vét), akkor azt mondjuk, hogy az ellenőrző hibája h=1%, pontossága $1-h$= 99%. Méréseinket tokenekre, nem pedig típusokra alapozzuk. Ennek a módszertani döntésnek a hátterében az a megfigyelés áll, hogy a gyakori alakokon vett pontosság sokkal fontosabb a ritka, egzotikus szóalakok hibátlan kezelésénél. A hibaszázalék (h) meghatározásakor tehát akkor nyerünk a felhasználó gyakorlati tapasztalaival egybevágó eredményt, ha az egyes szóalakokat gyakorisági súlyuknak megfelelően vesszük számításba[6]

A helyesírásellenőrző működését a továbbiakban automatikus javítási módban képzelhetjük el, vagyis amikor az ismeretlen szavakra szerkesztési távolság és gyakoriság alapján tesz helyettesítést. Azokra az alakokra, amelyekre nincs megbízható cserejavaslat, elfogadásra kerülnek. A pontosságot így három tényező befolyásolja döntően: a helyesírás-ellenőrző lefedettsége l, belső hibája b, és a maradványhiba m. A lefedettség egyszerűen azon tokenek aránya a szövegben, amelyekre nézve a helyesírás-ellenőrző képes javaslatot tud tenni (maga a szóalak vagy egy ahhoz közeli karakterfüzér szerepel a szótárában). A belső hiba magából a javaslattevésből adódó hiba, a maradványhiba pedig a javaslattevésből kimaradt és így elfogadott szavak között szereplő helytelen alakok aránya. Az összevont h hiba tehát a belső hiba és a maradványhiba lefedettséggel súlyozott átlaga:

$$h = lb + (1-l)m \tag{1}$$

A lefedettséghez csak a javaslattevések és az elfogadott szavak számát kell tudnunk, l mérése tehát triviális. m mérése azonban jóval összetettebb feladat, amely nem automatizálható. Az egyszerűség kedvéért feltesszük, hogy m felülről becsülhető az egész szövegben levő hibától, vagyis m l-lel nem növekszik.[7] Akkor érdemes egyáltalán helyesírásellenőrzőt használni, amikor az aktív ellenőrzés belső pontatlansága kisebb, mint a szöveg hibája ($b < m$), ekkor azonban h optimalizálása l növelésével, vagyis a tőtár növelésével érhető el. Felvethető a kérdés: mi a leghatékonyabb módszere korpuszok alapján történő szótárbővítésnek,[8] vagyis milyen módszerrel garantálható, hogy a legkisebb befektetéssel a hibaszázalék legnagyobb növekedését érjük el?

3. Szógyakoriság és a lefedettség növelése

A naiv helyesírás-ellenőrző rendszerek alapját a helyes alakok gyakoriság szerint rendezett rögzített listája adja. A magyarban $l > 0.5$ eléréséhez már az első

[6] A tokenszintű hibaszámítás a kontextuális információkat is felhasználó helyesírás-ellenőrző esetén még fontosabb, így ez az eljárás lehetővé teszi az eredmények összevetését.

[7] Tudjuk, hogy a szövegre jellemző hibaérték nagyban függ a szöveg típusától: *ekezettelen irasmodban* akár a 30-40%-ot is elérheti, átlagos szövegben 5-6% körüli, gondozott szövegben tipikusan 0.1% alatt marad.

[8] A helyesírás-ellenőrző szótárbővítése természetesen részben kivetelezhető új tövekben gazdag anyagok (szótárak, helységnévtárak, cégjegyzékek, stb.) átvételével is, de a kézi átnézésre még szótáraknál is szükség van a sajtóhibák miatt.

néhány ezer alak figyelembevétele is elégséges: ha ezeket kézzel átnézzük akkor $b \approx 0$ garantálható,[9] így (1) alapján $h < m/2$ minden nehézség nélkül elérhető.

Az alábbiakban a kvantitatív állításokat a Magyar Webkorpusz két változatán is illusztráljuk, az adatokat `web0` gyűjtés mellett összehasonlításképpen a $N = 670076633$ szövegszót és $V(N) = 15057395$ típust tartalmazó szűretlen `lar1` gyűjtésre is megadjuk.

Az 50%-os lefedettség garantálásához a 2913 (`lar1`), illetve 6486 (`web0`) szóalak listába vétele szükséges. Még $l > 0.666$ (tehát $h < m/3$) is csupán 15 ezer (`lar1`), illetve 24 ezer alak (`web0`) kézi átnézését igényli. A naiv módszer korlátját az adja, hogy az alacsony tokengyakoriságú szavakból típus szerint nagyon sok van. Legyen a mintában pontosan a k-szor előforduló típusok száma $V(k, N)$, tehát $N = \sum_{k=1}^{\infty} kV(k, N)$ és $V(N) = \sum_{k=1}^{\infty} V(k, N)$. Általános tapasztalat (ld. pl. Baayen 1996), hogy ha $N > 10^6$, akkor $V(N)$ domináns tagja $V(1, N)$, a típusok több mint fele olyan, hogy az egész szövegben csak egyszer fordul elő, vagyis ún. hapax legomenon. A `lar1` korpuszban $V(1, N) = 8133805$, a `web0`-ban $V(1, N) = 2567665$, vagyis a szóalakoknak rendre 54.0, illetve 56.7 százaléka hapax, bár gyakorisággal súlyozott arányuk csak rendre 1.21%, illetve 2.26%. A hapaxok kézi átnézése ezért lehetetlen, vagyis a lefedettség ezzel a módszerrel nem vihető 98% fölé, vagyis a h alsó korlátja $m/50$. Gyakorlatilag még a kétszer és háromszor előforduló szavak is komoly akadályt jelentenek: példáinkban ezek együttes gyakorisága mindössze 7.2% (`lar1`), illetve 4.2% (`web0`), a lista bővítéséhez azonban így is alakok millióit (8.1, illetve 4.7 millió) kellene kézzel átnézni! Gyakorlatban tehát a naiv módszer alsó hibahatára inkább $h \approx m/20$; ez egybevág a gyakorlati tapasztalattal, miszerint az átlagos (5-6% hibaarányú) nyers szövegből a naiv elven működő helyesírás-ellenőrzők lényegesen jobb (0.3% hibájú), azonban a gondozott szöveg minőségi követelményeit (0.1% alatti hiba) el nem érő javított változatot állítanak elő.

A tisztán izoláló nyelveknél, mint pl. a vietnami, a naiv módszer teljesen kielégítő, hiszen ezekben a nyelvekben a helyesírás ismerete nem jelent többet, mint az egyes szavak helyes ismerete[10] A komplexebb morfológiájú nyelveknél, mint amilyen a magyar, a helyes szóalakok ismerete nem merül ki a szótövek és a ragok ismeretében, hiszen ezek konkatenációja csak a morfológia szabályainak figyelembevételével eredményez helyes szóalakot. A helyesírás-elemzőbe tehát be kell építeni nemcsak a töveket és a ragokat, hanem a morfológiát is.[11] A hiba az ilyen rendszer esetén is a lefedettség lineáris függvénye, utóbbi viszont három komponensre bontható: a tő , a ragok, és a hasonulási szabályok lefedettsége. A nem-rekurzív, tehát tételesen felsorolandó kombinációk száma nem túl nagy (pl. Veenker (1968) 3020 ragkombinációt vesz lajstromba, a Hunspell jelenleg

[9] Szigorú értelemben b nem nulla, hiszen egy biztosan jó szótípus adott környezetben való előfordulása lehet, hogy hibásan kerül elfogadásra. Az ilyen hibákat elhanyagolhatónak tekintjük

[10] A megfontolás lényegében változtatás nélkül átvihető az olyan nyelvekre is, mint az angol, ahol az azonos tőhöz tartozó alakok száma kicsi.

[11] Ezt látjuk a legtöbb magyar helyesírás-ellenőrzőnél, pl. Hunspell, Helyes-e, Lektor.

4936 főnévi, 4041 melléknévi, és 59 igei alakkal dolgozik).[12] Miután ezek kézzel könnyen ellenőrizhetők, feltehető, hogy a rendszer lefedettsége ebben a ragkombinációk tekintetében 100%. Érettebb rendszer esetén még a morfofonológia szabályainak teljes ismerete is elvárható, így a belső hiba legfontosabb forrása az lehet, hogy a tőtárban egyes elemek hibás morfológiai információval szerepelnek. A morfológiai elemzést használó helyesírás-ellenőrzőknél a tőtár bővítése nem egyszerűen az új tövek felvételét jelenti, hanem előfeltételezi a tövek morfológiai osztályozását is, így a fentiekkel ellentétben $b = 0$ nem garantálható. A hiba csökkentésének alapvető módszere itt is a mohó algoritmus: először bevesszük a leggyakoribb alakokat akár elemzetlenül is, különösen ha elemzésük túlságosan komplikálná a morfonológiai szabályrendszert (hiszen ennek hibáját 0-n akarjuk tartani), és csak akkor lépünk tovább a $T + 1$-edik tőhöz, ha az első T tő már minden gyakoribb alakot lefed. Ezzel az eljárással olyan alakokat is lefedünk, amelyek önmagukban nem kerülnének be (vagy mert hapaxok, vagy akár elő sem fordulnak az adott mintában), de mint gyakoribb tövek ritkább ragokkal való kombinációi most mégis elérhetővé válnak. Ilyen pl. a *decembereinknek* alak, amely kétségkívül jólformált magyar szó, még a 670m szövegszavas mintánkban sem fordul elő (és a naiv alak-gyakorisági megfontolások alapján soha nem is kerülne be a listába), így viszont a *december* gyakoribb tőszármazékai, valamint a produktív toldalékolási minták jóvoltából elfogadásra kerül.

Az általunk használt szótárbővítési eljárás tehát három lépésre bomlik: először gyakoriság szerint sorba rendezzük a szóalakokat, másodszor vesszük a még hiányzó leggyakoribb alak tövét. Ezt automatikusan állítjuk elő, a projekt keretében kifejlesztett Hunstem tövezővel, amely Hunspell-lel azonos morfológiai szabályrendszerrel dolgozik. Végül megállapítjuk a tő helyes besorolását. A mohó algoritmus egyre kisebbeket tud csak kivenni a maradékból, és a csökkenés mértéke is jól megbecsülhető. A szóalakok gyakoriságát első közelítésben Zipf törvénye adja meg: az r-edik alak valószínűsége $1/r^B$-vel arányos:[13]

$$p_r \sim 1/r^B \tag{2}$$

Miután az alakok valószínűsége a tő valószínűségének konstansszorosa (Kornai 1992) az összefüggés a tövekre is érvényes. Ennek alapján tehát a mintában pontosan k-szor előforduló tövek arányára a következő adódik (a levezetés részleteit ld. Kornai 1999):

$$V(k, N)/V(N) = 1/k^{1+1/B} \tag{3}$$

Éppen a helyesírási hibák miatt, ez még további korrekciót igényel, de $k = 1$-re áll, hogy $V(1, N)/V(N) > 0.5$. A magasabb tagok igen jól illeszkednek a Zipf törvény által jósolt (3) értékekhez: az alábbi táblázat $k \leq 10$-re mutatja $V(k, N)$ mért, illetve a $V(N)/2k^{1.8}$ formulával becsült értékeit, valamint a becslés relatív pontosságát a `lar1` és a `web0` korpuszokra:

[12] Bár a ragok és rag-kombinációk száma elvileg végtelen, a rekurzív esetek (túlzófok *legeslegesleges...*, birtokos *éééé...*) viszonylag egyszerű szabályokkal kezelhetők.

[13] A B Zipf konstans a magyarban 5/4 körül van (ld. Füredi et al 2003).

	lar1			web0		
k	$V(k,N)$	$V(N)/2k^{1.8}$	b/m	$V(k,N)$	$V(N)/2k^{1.8}$	b/m
1	8133805	7528697	0.925606	2567665	2263728	0.881629
2	2393113	2162050	0.903447	632426	650085	1.02792
3	1006086	1042081	1.03578	280327	313333	1.11774
4	628829	620886	0.987369	167517	186688	1.11444
5	385044	415503	1.0791	112747	124933	1.10809
6	297081	299259	1.00733	82215	89981.2	1.09446
7	211134	226748	1.07395	63211	68178.6	1.07859
8	175753	178303	1.01451	50017	53612	1.07188
9	137100	144239	1.05207	41053	43369.8	1.05644
10	115697	119322	1.03133	34623	35877.7	1.03624

Zipf törvénye szerint, ha lista bővítését a rangsor r-dik tagjánál abbahagyjuk, a lefedetlen rész aránya, $1 - l = \sum_{i=r}^{\infty} 1/r^B$. Ebből, az összeget integrállal helyettesítve, azt kapjuk, hogy a lefedetlen rész r függvényében csak mint r^{1-B} (tehát $B = 5/4$ mellett a negyedik gyök reciprokával) csökken. A gyakorlatban ez annyit jelent, hogy míg a 100 ezer leggyakoribb tő felvétele 5.6% lefedetlenséget hagy, 1 millió tőnél ez 3.2%, 10 millió tőnél 1.8%, és 100 millió tőnél (ennyit korpuszunk nem is tartalmaz) 1%, és általában a lefedetlenség egy nagyságrendnyi csökkenéséhez a kezdeti korpuszt négy nagyságrenddel kell növelni.

4. Konklúzió

A fentiekben megmutattuk, hogy a helyesírás-ellenőrzők pontosságát elsősorban lefedettségük (tehát a tőtár mérete) határozza meg. A tokenek és típusok összefüggésének jólismert törvényeire támaszkodva levezettük a bővítés várható hatását a lefedettségre, amit a méréseink alá is támasztottak. Az emberi erőforrások minimalizálása miatt tehát a lefedettség növelését egy határon túl a feldolgozandó szövegek heurisztikákkal operáló előszelekciójával (a mondatközi nagy kezdőbetűs szavak nagy valószínűséggel tulajdonnevek, amelyeket fel kell venni, a csupa számjegyből álló szövegszavak viszont telefonszámok vagy dátumok, amiket viszont automatikusan el lehet hagyni) érdemes csak megkísérelni.

Hivatkozások

1. Baayen, R. H. 1996: The effect of lexical specialisation on the growth curve of the vocabulary. *Computational Linguistics* **22**, 455–480.
2. Füredi M., Kornai A., Prószéky G. 2003: A SZÓTÁR adatbázis. Kézirat.
3. Kornai A. 1992: Frequency in morphology. In: I. Kenesei (ed): Approaches to Hungarian IV 246–268.
4. Kornai A. 1999: Zipf's law outside the middle range. Proc. Sixth Meeting on Mathematics of Language, University of Central Florida, 347–356.
5. Németh L. 2003: A szószablya fejlesztés. Az V. Linux konferencián elhangzott előadás cikk változata. URL http://konf2003.linux.hu/.
6. Veenker, W. 1968: Verzeichnis der Ungarische Suffixe und Suffixkombinationen. Mitteilungen der Societas Uralo-Altaica 3, Hamburg.