Szóhasonlóság mérése analógiás megközelítésben

Rung András[1]

[1] MTA Nyelvtudományi Intézet, Elméleti Nyelvészet Tanszék, Benczúr utca 33., 1068 Budapest, Magyarország
rungandras@gmail.com

Kivonat: A magyar szavak, elsősorban a főnevek hasonlóságát meghatározó tényezők leírására törekszem. Ebben a szabályalapú nyelvtanok helyett a mentálisan reálisabbnak és rugalmasabbnak tűnő analógiás keretrendszert tekintem kiindulási alapnak. Munkámban a számítógépes nyelvészet eredményeire és módszereire támaszkodom. Számításaim, megállapításaim kizárólagosan korpuszból vett adatokon alapszanak. Kutatásomnak közvetlen hozadéka is lehet a nyelvtechnológia számára a szótárbővítés és –karbantartás területén, mivel a hasonlóságot mérő algoritmusom 95%-os pontossággal ismeri fel főnévi tövek hangkivető voltát, amely már lehetővé teszi az ilyen szavaknak akár automatikus besorolását is.

1 Bevezetés

A szabályalapú nyelvtanok sokszor jó közelítő leírást adnak az alaktani viselkedésről, azonban képtelenek olyan nyelvi jelenségeket magyarázni, mint például a fokozatosság, nyelvi ingadozás, a gyakoriság hatása a nyelvi változásra [8]. Ezekben az esetekben az analógiás nyelvtan jobban közelíti a pszichológiai realitást, azaz a valós nyelvi működést. Ha bármely nyelv analógiás nyelvtanát kívánjuk megírni, annak egyik alapfeltétele az, hogy tudjuk mely fonémák-hangok, alakok [1], konstrukciók [2] hasonlóak az adott nyelvben, és ezek hasonlósága milyen mértékű, min alapszik.

Feltételezésem szerint az egyes nyelvi elemek nem elszigetelten léteznek, hanem szoros és állandó interakcióban vannak egymással, amelynek egyik legfontosabb mozgatórugója az analógia. Ha a nyelvben valahol változás következik be, akkor az az erőviszonyokra azonnal hatással van, és a rendszer egészének változásához vezet. Ezt a legtöbb 20. századi nyelvelmélet el is ismeri. Ennek megfelelően az egyes állapotok leírásával foglalkozik a szinkrón nyelvészet, míg az ezek közti átmenetek vizsgálatával a diakrónia.

Ez a megközelítés azonban kimondva vagy kimondatlanul azt közvetíti, hogy vannak mindig stabil, önmagukban megfigyelhető állapotok. A nyelv változik, de maga a változás nem alapvető minősége. Az analógiás nyelvtan által megfigyelt tények, jelenségek azonban cáfolják ezt a szigorú és merev szétválasztást. A változás és az állapot nehéz szétválaszthatóságából következik, hogy a rendszer a maga statikusságában nem létezik, vagy legalábbis olyan absztrakt fogalom, amely a nyelvvel való valós munkára alkalmatlanná teszi.

Ha a nyelvi jelenségeket szorosan összefüggőnek vesszük és változásukat lényegi elemüknek tekintjük, akkor felvetődik a kérdés, hogy egyáltalán lehet-e és értelmes-e a nyelvnek bármely részjelenségét leírni anélkül, hogy más részleteit ne vennénk figyelembe, hisz az összefüggések feltárása nélkül, a jelenség értelmezhetetlen vagy csak részlegesen értelmezhető lesz.

A leírás ebben az esetben valóban nem lesz tökéletes, de mivel a nyelvi változás teljességét az arról való meglehetősen korlátozott és szerény tudásunk miatt semmiképp sem tudjuk megragadni, így mégis kénytelenek vagyunk csak egyes darabjait vizsgálni. A hagyományos megközelítésekkel ellentétben azonban nem állítom, hogy ezeket a részleteket önállóan és pontosan le tudjuk írni, hanem úgy vélem, hogy újabb leírások fényében majd kiegészítésre szorulnak a későbbiekben. Adatainkat folyamatos változásukból kifolyólag sosem tudjuk megragadni, de ez nem is gond, hisz a nyelvészet célja nem feltétlenül a leírás, hanem a leírást meghatározó nyelvi folyamatok megértése és pszichológiailag reális feltárása.

2 Célkitűzések

A szavak összehasonlításában azok felszíni szerkezetét és alakjaik gyakoriságát veszem alapul, mivel az analógiás keretrendszer a mögöttes szerkezeteket inkább gátló, semmint hasznos elméleti konstrukcióknak tartja [1]. Ezekre támaszkodva az ingadozással és fokozatossággal szorosan összefüggő analógiás kiegyenlítődés is jól megragadható jelenséggé válik [5]. Feltételezésem szerint az analógiás alapú változásokat több további szempont is meghatározza (használati mód, jelentés, stb.), de ezekből a legfontosabb a hangtani/fonológiai hasonlóság.

A szavak fonológiai hasonlóságának meghatározására keresek egy egyértelmű algoritmust, amellyel modellálhatom ezt az analógiában egyik legfontosabb szerepet játszó komponenst. Amennyiben kitűzött célomat elérem, az a magyar nyelvtechnológia számára is értékes hozadékkal jár, hisz egy pontos, a valós folyamatokat jól megragadó algoritmus segítségével lehetővé válik a szavak hatékony szótárba sorolása (megfelelő jegyeikkel), illetve a már meglévő szótári anyag frissítése, karbantartása is, amely komoly kihívást jelent, ha csak emberi erőre hagyatkozunk. A szótárak automatikus bővítésével lehetővé válik akár rétegnyelvi (szleng, szaknyelv, stb.) szövegek hatékonyabb elemzése is.

Leggyakrabban az edit distance algoritmust [6] használják két szó, karakterlánc hasonlóságának eldöntésére. Ez a megközelítés azonban számos problémát vet fel. Egyrészt az algoritmus az összehasonlítást betűk és nem fonémák alapján végzi. Két betűt azonosnak vagy teljesen különbözőnek vesz, így az *o* és az *ó* ugyanannyira különböző az algoritmus számára, mint az *o* és a *k*. Továbbá az algoritmus feltételezi, hogy a törlés, beillesztés és megfordítás pont ugyanakkora változást okoz a szón belül, és ezeknek a beavatkozásoknak a helye is lényegtelen. Korábbi [8] és jelenlegi vizsgálatom is megmutatja, hogy ez az algoritmus emberi nyelvek szavainak (legalábbis a magyar nyelv szavainak) hatékony és megbízható összehasonlítására nem alkalmazható, így legfeljebb csak kiindulási pont lehet olyan kifinomultabb megközelítések számára, amelyek jobban megragadják az emberi nyelvi rendszer sajátságait és működését.

3 A hasonlóság mérésére használt algoritmus

A szavak közti hasonlóság mérésére egy python programnyelvben megírt algoritmust használok, amely egy hasonlósági mátrix alapján végzi számításait. Ez a mátrix adja meg, hogy két fonéma mennyire hasonlít egymáshoz. A hasonlóság értéke a 0 és 1 közti skálán helyezkedik el. Így két fonéma nem csak azonos vagy eltérő lehet, hanem az analógiás nyelvi megközelítéssel összhangban több, bár diszkrét fokozatban adható meg hasonlóságuk.

A fonémákat alapul véve korrigálom az edit distance algoritmus azon hiányosságát, hogy az összehasonlításban nem a szavakat, hanem azok grafémikus leképzését veszi alapul. A fonémák kiválasztásában és jegyeik meghatározásában a Magyar Strukturális Nyelvtant [4] és az Ipa (International Phonetic Alphabet) leírásait tartottam irányadónak.

Mivel az analógiás megközelítés egyes irányzatai [1] szerint a szavakat jelentésük mellett konkrét hangalakjukkal is tároljuk, érdemes lett volna a hasonlóságot fonetikus alapon (is) számolni. Ettől a lehetőségtől azonban kénytelen voltam eltekinteni, mivel jelenleg a fonetika nem tudja egyértelműen meghatározni, hogy két hang mikor és mennyire hasonlít, valamint egy ilyen vizsgálathoz szükséges beszéd korpusz vagy az ez alapján készített beszélt nyelvi gyakorisági szólista sem áll rendelkezésre. Így ezek hiányában maradtam a fonémák összehasonlításánál, amelyek még ha meglehetősen durva összehasonlítást is tesznek lehetővé, mégis legalább közelítik a nyelvi realitásokat.

A fonémák hasonlóságának mértékét a fonémák megkülönböztető jegyei alapján számolom. A magánhangzók esetében a nyíltságot, ajakkerekítést, hosszúságot, elölképzettséget, a mássalhangzók esetében pedig a zöngésséget, a képzés helyét és módját veszem figyelembe. Minden eltérő jegy esetén az összehasonlított két fonéma hasonlóságát felére csökkentem. A mássalhangzók és a magánhangzók egymáshoz számított hasonlósága rendszeremben 0, nincsenek közös jegyeik. A későbbiekben tervezem köztes kategóriák bevezetését bizonyos hangokra (*j* és *v*), illetve a hangok egységes kezelésének érdekében azonos mássalhangzók kapcsolatait nem geminátának venném, hanem e megoldás helyett a mássalhangzóknak is lenne hosszúság jegye.

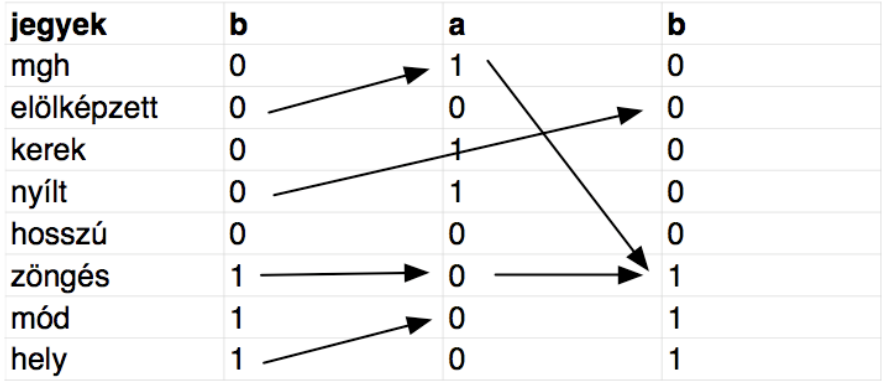
Ezek alapján az *o* fonéma hasonlóságának mértéke egy másik *o* fonémához 1, az *ö*-höz és az *ó*-hoz 0,5 ($1:2^1$, mivel egy jegyben az elölképzettségben, illetve a hosszúságban különböznek), míg az *ő*-höz 0,25 ($1:2^2$ hiszen két jegyben az elölképzettségben és a hosszúságba különböznek). A magyar nyelvleírás hagyományát követve az edit distance algoritmussal ellentétben nagyobb súlyt adok a szóvégek hasonlóságának. A fonémahasonlóság súlya a szóvégétől a szó eleje felé logaritmikusan csökken. Az algoritmusomban 1,8-es alapú logaritmust használok, mivel korábbi vizsgálataimban ez bizonyult a leghatékonyabbnak[7]. A szavak önmagukkal vett hasonlósági értéke 1. Programom számítása alapján a *bab* és a *púp* szavak hasonlósága a következőképp alakul:

b:p = 0,5 (eltérő jegy: zöngésség)
a:ú = 0,25 (eltérő jegy: nyíltság, hosszúság)
b:p = 0,5 (eltérő jegy: zöngésség)

Hasonlóság kiszámítása a logaritmikusságot is figyelembe véve (a könnyebb átláthatóság kedvéért 2-es alapú logaritmussal):

((0,5*1)+(0,25*2)+(0,5*4))/7=0,5

A fentebb leírt algoritmus mellett egy másik algoritmus tesztelését is megkezdtem, amely a hasonlóság számításában az edit distance-hez hasonlóan nem ad súlyt annak, hogy a szavak mely részei hasonlóak. A hasonlóságot az alapján határozza meg, hogy a két szó fonémáinak jegyeiből épített mátrixokban hány közös részgráf van. Így a *bab* és a *púp* tartalmazza a CVC a CV, VC, CC, C és V illetve a zárhang-hátulképzett-zárhang, zárhang-hátulképzett, stb. láncokat. A CC példából látható, hogy a gráfokban megszakításokat is megengedtem, hisz például a magánhangzó harmónia esetén a magyar nyelvben is a releváns összetevők nem közvetlenül követik egymást.



1. ábra. Néhány lehetséges gráf, amely az összehasonlítás alapját képezheti.

Ez az algoritmus az edit distance algoritmussal ellentétben azonban hangsúlyt ad a fonémák hasonlósági jegyeinek. Elvárásaim szerint jó teljesítményt hozhatott volna, mivel algoritmusom fő gyengéjének korábban a magánhangzó harmónia és a szótag-szerkezet iránt mutatott kisebb érzékenység mutatkozott. Azonban a sorrendiséggel szemben való semlegessége olyan hátránynak bizonyult, amelynek köszönhetően még az edit distance algoritmusnál is rosszabbul teljesített.

4 Adataim

Korpuszalapú vizsgálataimban a Szószablya webkorpusz [3] gyakorisági adatait használtam fel, amelyet azért választottam, mert jelenleg ez a legnagyobb, mintegy 19,1 millió szóalakot tartalmazó korpusz, amely 3,493 millió weboldal és 1,486 milliárd szó alapján készült. Tövek helyett szóalakokat, elnagyolt gyakorisági kategóriák helyett pedig pontos gyakoriságú számokat tartalmaz, amely egyedülálló módon

alkalmassá teszi nyelvészeti és nyelvtechnológiai kutatásokra. A Szószablya webkorpusz további nagy erénye, hogy válogatatlan, sokszor a beszélt nyelvhez közelibb nyelvhasználatot és írásmódot rögzítő anyagokat (fórumok, blogok) is nagy arányban tartalmaz, így az ez alapján tett megállapításaink is jobban közelíthetik a magyar nyelvi valóságot.

Vizsgálatomhoz a hangkivető főneveket választottam. Választásom több okból kifolyólag esett rájuk. A magyar hangkivető tövek habár zárt osztályt alkotnak, meglehetősen nagyszámúak, így a viselkedésükből levonható következtetések nem szórványos, egyedi és ritka adatokon alapszanak. A hangkivető főnevek viselkedésére jellemző a fokozatos ingadozás, amelyről még a szabályalapú megközelítések engedékenyebb változatai sem tudnak teljesen számot adni.

A hangkivető töveket kiindulási pontnak az is kiválóan alkalmassá teszi, hogy korábban alapos leírást készített róluk Rebrus Péter[7] a kormányzásfonológia eszköztárát használva, amely jó viszonyítási alapot képez vizsgálódásaimhoz.

A Szószablya webkorpuszon túl vizsgálatomban még a BME MOKK morphdb.hu szótárára támaszkodok, amely jelenleg a legnagyobb, ingyenesen is hozzáférhető nyelvi adatbázis (130 ezer tő, [10]), ahonnan összesen 1205 hangkivető főnévi tövet választottam ki. Ezekből összesen 1097 volt a szótárban hangkivetőként megjelölve, amelyekből kivettem a *kelet, sportberkek, sodor, terem* szavakat. A *kelet* szó egyértelműen a *kelte* szóalak miatt került be hangkivetőként rögzítve, ahelyett, hogy már ragozott főnévként vették volna fel a szótárba. A *sportberkek* már szóalak, és nem tő, helyesen a szótárban *sportberek*-ként kellene szerepelnie, amelynek hiányos a paradigmája. A *sodor* és a *terem* szavak valóban hangkivető főnevek, de mivel alanyesetük és számos további ragozott alakjuk egybeesik a náluk sokkal gyakoribb *sodor* és *terem* igei tövek alakjaival, ezért célszerűbbnek tartottam ezek kihagyását a vizsgálatból. Úgy véltem, hogy elegendő nyelvi adat birtokában ezek elhagyása nem vezet az eredmények jelentős módosulásához.

A szótárban hangkivetőként megadott töveken túl további 102 tövet választottam ki, amelyek valóban hangkivetők és szerepelnek is a szótárban, de nincsenek hangkivetőként megjelölve. Ezek a szavak a hangkivetőként megjelölt 1093 szótőből létrehozott összetett szavak, amelyekben a hangkivető tő az összetétel jobb oldali záró tagját adja. Az adatbázis mindösszesen 4 tövet ad meg ingadozónak (hangkivetés a megfelelő ragok előtt a Szószablya webkorpusz alapján *bajusz*: 35%, *fókabajusz*: nincs adat, *harcsabajusz*: 57%, *macskabajusz*: 25%), amelyeket jól azonosít.

A korpuszra támaszkodva a morphdb.hu szótára több ponton is javítható lenne, amelynek szóanyaga több korábbi szótár automatikus módszerekkel végrehajtott egyesítésével jött létre. A morphdb.hu más szótári adatbázisokhoz hasonlóan nem kezeli a nyelvre jellemző ingadozást. Szótáraink, mint láthattuk legjobb esetben is csak megjelölik az ingadozást, de annak mértékéről nem tudnak számot adni. Ezzel tulajdonképpen azon szabályalapú nyelvelméletek gyakorlatát követik, amelyek készítésükre a legnagyobb hatást gyakorolták. Elsősorban a hibás besorolások és az ingadozás helyes megadásával lehetne javítani az adatbázisok minőségét.

Az ingadozás mértékének jelölése a morfológiai elemzésnek csak bizonyos, nem túl gyakori eseteinél lehetne szükséges, ahol egy kiegyenlített alak egybeeshet egy másik szó alakjával (pl. *sodort* = *sodor* ige + múlt idő vagy *sodor* főnév + tárgyeset, *kéreget* = *kér* + *gAt* vagy *kéreg* + tárgyeset). Az ingadozás mértékének megadása elsősorban egy olyan adatbázisban nyerne létjogosultságot, amelyet már szóalakok

produkciójára is jól lehet használni, hisz nem mindegy, hogy adott esetben a ritkább hangkivetéses alakot használjuk vagy a már analógiásan kiegyenlített gyakoribb alakot. Vélhetőleg az ingadozás mértéke nyelvi regiszterenként is eltérő lehet, de ennek megállapítására a Szószablya webkorpusz alkalmatlan forrás.

Az egyes tövek ingadozását az összes rag, jel jellegű toldalék előtt megvizsgáltam, amelyek hangkivetést válthatnak ki: tárgyeset, szuperesszívusz, többes szám, birtokos személyragok. A Szószablya webkorpusz alapján megállapítható, hogy a *pityer*, *szlalom*, *vicikvacak* szavak már nem hangkivetők, míg további 114 tő tekinthető ingadozónak, mert ezeknél az esetek legalább 1%-ban a hangkivetést kiváltó toldalékok előtt nem történik meg a hangkivetés. 43 tőnél ez az arány meghaladja a 10%-ot, 15-nél pedig több, mint 50%. Habár célszerű lenne ezeknek az adatoknak az alapján szótárunkat frissíteni, 2009-es Google lekérdezések alapján látható (.hu domain alatt), hogy az analógiás kiegyenlítődés tovább folytatódik (pl. *fátyolt* aránya 2003-ban 67%-ban hangkivető, 2009-ben már 42% a *fátylat* helyett). Természetesen ezen folyamatok pontos leírása és értékelése külön vizsgálatot érdemel az összes alak figyelembevételével.

Adataimat vizsgálataimhoz átkonvertáltam egy olyan írásrendszerbe, ahol egy fonémának egy betű felel meg. Az egyes alakokban itt a szóbelseji zöngésedési folyamatoknak megfelelő fonémát tüntetem fel, amelyeket eredetileg az íráskép nem rögzít, így lesz a *virágcsokor*ból *virákčokor*.

5 Kísérlet és eredmények

A szóhasonlóságot megállapító algoritmusom pontosságát olyan tesztekkel ellenőriztem, amelyek során valós magyar szavakat sorolok be már meglévő szócsoportokba. A besorolások helyessége alapján látható, hogy egy algoritmus mennyire jól ragadja meg azt a feltételezett nyelvi képességet, amely alapján analógiás hasonlítások elvégzésére képesek vagyunk.

Első számításaim a magyar helységnevek lokatívuszaival végeztem, amelynek során azt tapasztaltam, hogy az analógiás keretrendszer jól meg tudja ragadni ezek viselkedését, ellentétben a szabályalapú megközelítésekkel, amelyek általánosításainak a valós adatok nem egyszer ellent mondanak.

A leggyakoribb 100-100 harmónia szempontjából megfelelő alak alapján meghatároztam (100-100 *–bAn*, 100-100 *–(V)n* végű), hogy a következő 40 leggyakoribb alak szuperesszívuszt vagy inesszívuszt vár-e el. A szavak szótári és ragozott alakjai alapján 87,5%-os pontossággal választotta ki algoritmusom a megfelelő szócsoportot. Ritkább alakok esetében már az anyanyelvi beszélők ítéletei is ingadoznak, így ez a 87,5%-os teljesítmény megközelíti az ő eredményeiket, az edit distance teljesítményét pedig messze meghaladja.

Eredményeim megerősítése végett a hangkivető tövekkel is elvégeztem egy a korábbival megegyező vizsgálatot, amelyben az edit distance algoritmus, és a gráfi hasonlóságot figyelembe vevő algoritmus eredményességét hasonlítottam össze saját algoritmusommal.

Az 1205 hangkivető tőből sorrendben az 501. leggyakoribb tőtől a 600. tőig megvizsgáltam, hogy ha egy már meglévő szólistához hasonlítom ezeket a töveket, akkor

milyen pontossággal találnak az egyes algoritmusok a szólistában szereplő hangkivető tövet hasonlósági alapon. A 100 hangkivető tőből összesen 7-nél a hangkivetés elmaradása a releváns toldalékok előtt meghaladta a 10%-ot (*hatökör, ködfátyol, lombsátor, sulyok, tündérfátyol, szalmakazal, zsákvászon*), míg a listán szereplő *pityer*nél a korpusz alapú vizsgálatok alapján láttuk, hogy az analógiás kiegyenlítődés befejeződött vagy befejeződés közeli állapotban van.

A hangkivető tövekhez kontroll csoportként véletlenszerűen kiválasztott, velük azonos gyakoriságú 100 nem hangkivető tövet vettem. Ezek ragozatlan alakjainak a korpusz 93 és 57 közti előfordulást adott meg, azaz ritka, de még használt és valamelyest ismert szavakról van szó. A szavak gyakoriságát besorolásaimhoz minden esetben ragozatlan alakjaik alapján vettem. Ez némileg eltérhet a szó összes alakjai alapján számított gyakoriságától, mégis alkalmazhatjuk ezeket a számokat besorolásukhoz. A hangkivető szavaknak mind ragozatlan alakjairól, mind összes előforduló alakjukról pontos adataim vannak a korpusz alapján, és a kétféle módon számított gyakoriság közt igen magas, 0,758-as korreláció figyelhető meg.

A hangkivető és nem hangkivető tesztszavakat összesen 4 eltérő méretű szólistához hasonlítottam. Ezekben az 50, 100, 200 illetve 500 leggyakoribb hangkivető tő, illetve az ezekkel egyenlő vagy nagyobb gyakoriságú nem hangkivető főnévi tövek szerepeltek, amely listák pontos méretét az 1. táblázat adja meg. Mint látható a hangkivetők aránya a tőszámmal együtt egyenletesen nő, de nem változik olyan radikális mértékben, hogy az egy vizsgálat eredményére jelentős kihatással lehessen.

1. táblázat: Szólisták száma és a hangkivető tövek aránya ezekben.

Hangkivetők száma	Tőszám	hangkivető tövek aránya
50	2828	1,7%
100	5468	1,8%
200	10315	1,9%
500	15333	3,2%
1205 (összes tő)	55762 (összes tő)	1,8%

A teszt során a két 100-100 darabos szócsoportot a nagyobb szólistákhoz hasonlítottam, amelynek eredményét a 2. táblázat mutatja. A százalékok arra utalnak, hogy a 100 tőből hány százalékban választott az adott algoritmus az adott listából azonos típusú tövet. Amennyiben hangkivető tőt kellett választanunk, úgy a találgatás küszöbe 1,7 és 3,2% közt lett volna. Ezt láthatjuk, hogy minden esetben sikerült a vizsgált algoritmusoknak meghaladnia. A nem hangkivető tövek esetében ez a szám jóval magasabb 96,8-98,3%, hisz ezek a tövek jóval nagyobb arányban voltak képviselve a szólistákban, így véletlenszerű kiválasztásukra is nagy esély lett volna. Ezt a szintet egyedül saját algoritmusom haladta meg, azonban csak a legkisebb 50-es szólista esetén, amikor azonban hibátlanul teljesített. Mivel a leggyakoribb tövekhez hasonlítanak algoritmusaim, a gyakorisági szempontok is szerepet kapnak, de csak mérsékelten, hisz a gyakoribb tövek közt a nagyon gyakori és a kevésbé gyakori tő már egyforma súllyal bír.

2. táblázat: Az egyes algoritmusok eredményessége a szavak összehasonlításában.

Szólisták	Edit distance, hangkivető	edit distance, nem hangkivető	saját algoritmus, hangkivető	saját algoritmus, nem hangkivető	gráf alapú, hangkivető
50 hangkivető	39%	98%	51%	100%	7%
100 hangkivető	75%	93%	73%	97%	14%
200 hangkivető	64%	98%	84%	97%	
500 hangkivető	63%	100%	95%	98%	

A gráf alapú algoritmussal a 200 és az 500 hangkivető alakot tartalmazó szólista esetében nem végeztem el az összehasonlításokat, mert az algoritmus jelenlegi implementációja nem teszi lehetővé belátható időn belül ekkora adattömeg összehasonlítását. Kihagytam a táblázatból a nem hangkivető tövekkel való összehasonlítást is ennek az algoritmusnak az alapján, mivel a hangkivetőkkel való összehasonlítás során már megmutatkozott, hogy az algoritmus jelen formájában nem tud megfelelő eredményt hozni.

A táblázat alapján látszik, hogy saját algoritmusom nagy mennyiségű adattal összesen 95%-os, illetve 98%-os eredményt hozott. Eredményeim azt mutatják, hogy algoritmusom megfelelő hatékonysággal tud emberi beavatkozás nélkül is szavakat megfelelően besorolni, amely egybecseng korábbi tapasztalataimmal. A nagy számok természetesen relatívak, hisz a hasonlításhoz felhasznált szavak mennyisége még így is csak a negyede annak, amivel szótárunkban rendelkezünk.

Algoritmusom 5 esetben sorolta be rosszul a következő hangkivető töveket: *pityer, bugyor, orrnyereg, lombsátor, csöbör*. A *pityer* esetében nem beszélhetünk hibázásról, hisz ezt a besorolást a korpusz adatai is támogatják. Ha a *bugyor* (leghasonlóbb *hunyor*) , orrnyereg (leghasonlóbb *hadsereg*) és csöbör (leghasonlóbb *csömör*) esetében a hozzájuk 10 leghasonlóbb szót vesszük, akkor azt figyelhetjük meg, hogy ezek közt már van 3, 2 illetve 4 hangkivető szó. Azaz az algoritmus felfedezi a hangkivető tövekhez a hasonlóságot, csak nem ad ezeknek megfelelő súlyt. Egyedül a *lombsátor*hoz nem talált megfelelő hangkivető szót még az első 10 közt sem, ami jól tükrözi, hogy a *lombsátor* szó ingadozik, de az algoritmus ítélete túlzó. Az algoritmus következetesen, de tévesen az *-átor* végű latin eredetű szavakhoz hasonlítja: *pankrátor, diktátor, organizátor* stb. Még a rontott példákból is látszik, hogy az algoritmus ilyenkor is jól közelíti a hasonlóságot, de teljesítményét célszerűbb lenne nem csak egy választott alak alapján kiértékelni. Az algoritmusnak két tulajdonsága, miszerint hátulról számol, illetve meglehetősen engedékeny egy szekvencián belül kisebb eltérésekre, alkalmassá teszi, hogy hatékonyan hasonlítson.

A 100 nem hangkivető tőhöz való hasonlítás során az algoritmus két hibát követett el: *bikacsök:bütyök* illetve *csucsor:csupor*. Az első esetben a korábbi tesztelések során is tapasztalt hibát figyelhetjük meg, miszerint az algoritmus nem elég érzékeny a hangrendi harmóniára, hisz a második magánhangzó már elég távol van a szó végétől, hogy kis súlyt kapjon. Ezért nem zavarja az algoritmust az *aö* szekvencia hasonlítása az *öö*-höz. A sokkal megfelelőbb jelölt, a *lopótök* csak a 10. leghasonlóbb szónak kerül elő.

Az edit distance algoritmus gyengébb teljesítménye egyértelműen a már leírt hiányosságaira vezethető vissza, a gráf alapú algoritmus pedig jelenlegi implementáció-

jában leginkább az azonos hosszúságú szavakat választja, amely szintén többnyire rossz választáshoz vezet.

Természetesen felmerülhet a kérdés, hogy az ilyen jellegű besorolás mennyire használható a szótárbővítésben, hisz a hangkivető tövek zárt csoportot alkotnak, amely nem bővíthető tovább. Látnunk kell azonban, hogy a szótárbővítés valójában nem új szavak besorolása egy szótári csoportba, hisz ezek a szavak szótárunktól is függetlenül már hangkivetők vagy sem. Esetünkben csak az történik, hogy ezek hangkivető voltát „felismerjük". Igen sok szó van, amelyek besorolása a digitális szótárakba még nem történt meg. Ezek esetében is hasznos lehet az automatikus, de a valós folyamatokhoz közeli besorolás, amely nem alapulhat csak azon, hogy egy új szó esetleg valamely a szótárban már meglévő szóból létrehozott összetett szó-e (lásd *lé:levet*, de *baracklé:baracklét/baracklevet*).

Másrészt ha egy szócsoportot zártnak is veszünk, nem kizárt, hogy a valóságban, ha elég nagy analógiás erővel bír, be tud vonzani új szavakat, mint például a *motrok, bútrok* alakok esetében, amely adatokat gyakran félresöprik, de mégsem hagyhatjuk ezeket figyelmen kívül, mert a nyelvi változás lényegéről beszélnek nekünk. Egy szó besorolása alapvetően hasonló feladat, mint amikor egy szónak egy alakját hozzuk létre beszéd közben, ha már hallottuk ezt az alakot vagy egyenesen nagy gyakoriságú is, akkor jó esélyünk van arra, hogy az „elvárt", hangkivető alakot ejtjük ki. Azaz a gyakoribb hangkivető töveknél nagyobb az esély hangkivető változatok létrehozására, hisz ott több minta mutatja ezt.

6 További kutatási lehetőségek

Habár algoritmusom immár két vizsgálatban is sikeresen bizonyította, hogy elegendő nyelvi minta birtokában hatékonyan tud analógiás párokat találni, számos lehetőség van továbbfejlesztésére úgy, mint nyelvtechnológiai eszköz, és úgy is mint a nyelvi folyamatokat reálisan modelláló algoritmus. Elsősorban a magánhangzó harmónia és a szótagszerkezet iránti érzékenységét lenne érdemes növelni. Erre a célra lehet alkalmas az egyébként nem olyan jól teljesítő gráf alapú algoritmussal való ötvözése. A két nyelvi jelenséggel való vizsgálat már sokat elárult természetéről, de célszerű lenne még további nyelvi jelenségeken is megvizsgálni hatékonyságát (tőbelseji magánhangzó rövidülés, *v*-vel való bővülés).

A rendszer látszólagos hibázásainak felderítése közben korábbi és jelenlegi kutatásaimban is az körvonalazódott, hogy jobb eredményt kaphatnánk, ha az analógiás hasonlításnál nem feltétlenül egy szóhoz, hanem egy valamilyen szempontból konzisztens csoporthoz hasonlítjuk szavainkat, amelyet klaszterezéssel lehetne felderíteni. Ezzel párhuzamosan fel kellene térképezni az egyes szavakra ható analógiás nyomást, amely mentén egy adott szó részt vesz az analógiás folyamatokban. Egy ilyen vizsgálatban a gyakoriságnak már kiemelkedő szerepe lenne, amelynek azonban a jelenlegi irányvonal továbbfejlesztésében is nagyobb szerepet kellene kapnia.

Hivatkozások

1. Bybee, J. L. : Phonology and Language Use, CUP, Cambridge (2001)
2. Goldberg, A.: Constructions. A Construction Grammar approach to argument structure, University of Chicago Press, Chicago (1995)
3. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: Creating Open Language Resources for Hungarian, In: Proceedings of LREC. (2004) 1201–1204
4. Kiefer F.(szerk.): Strukturális Magyar Nyelvtan 2. Fonológia. Akadémiai Kiadó, Budapest (1994)
5. Kraska-Szlenk, I.: Analogy. The relation between Lexicon and Grammar. Lincom, Muenchen (2007)
6. Levenshtein, V.: I. Binary codes capable of correcting deletions, insertions, and reversals, Doklady Akademii Nauk SSSR, 163(4): 845–848 (Russian). English translation in Soviet Physics Doklady, 10(8): (1965-1966) 707–710
7. Rebrus, P.: Morfofonológiai jelenségek. In: Kiefer F. (szerk.) Strukturális magyar nyelvtan 3. Morfológia. Akadémiai Kiadó, Budapest (2000) 763–947
8. Rung, A.: Determining word similarity in the Hungarian language. Papers from the Mókus Conference. Tinta Kiadó, Budapest (2008) 112–118
9. Skousen R.: Analogical Modeling of Language Kluwer Academic Publishers, Dordrecht Boston London (1989)
10. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA (2006) 1670–1673