Klaszterek helyett prototípusok

Kálmán László[1], Rung András[1,2]

[1] MTA Nyelvtudományi Intézet, Elméleti Nyelvészet Tanszék, Benczúr utca 33.,
1068 Budapest, Magyarország
[2] BME Fizikai Intézet, Budafoki út 8.,
1111 Budapest, Magyarország
kalman@nytud.hu
rungandras@gmail.com

Kivonat: Írásunkban bemutatjuk, hogy nyelvi elemek viselkedésének jellemzése és modellezése lehetséges klaszterekre való hivatkozás nélkül prototípusok segítségével is. Vizsgálatunkban gépileg kiválasztott prototípusok segítségével a hangkivető főnevek ingadozását modelleztük eredményesen. 282 hangkivető főnévből választottunk ki 8 prototípusnak tekinthető szót. Az egyes szavak és a hozzájuk alakjában leghasonlóbb prototípus közt mérhető távolság szignifikáns pozitív együttjárásban ($r(280) = 0{,}419$, $p < 0{,}001$) van a viszonyított szavak hangkivetési mértékével a Szószablya Gyakorisági Szótár [3] adatai alapján. Ebből láthatjuk, hogy azok a szavak, amelyek a prototípusokra jobban hasonlítanak hangalakjukban, azokhoz közelítő módon is viselkednek, azaz az egyes szavak viselkedését klaszterekre és szabályokra való hivatkozás nélkül is modellezni tudtuk.

1 Bevezetés

A statisztikai alapú számítógépes nyelvészetben (így pl. a korpusznyelvészetben és az automatikus nyelvtanindukcióban) fontos szerepe van az egymáshoz hasonlóan viselkedő egységek felfedezésének és csoportosításának, vagyis a klaszterezésnek. Vannak azonban olyan nyelvészeti feladatok, amelyeknél nem annyira magukra a klaszterekre, hanem az őket legjobban képviselő elemre van szükségünk. Ilyen például az esetalapú, példányalapú vagy általában analógiás okoskodás használata a számítógépes nyelvészetben. Ennél a fajta okoskodásnál olyan elemet keresünk, amely bizonyos szempontokból a lehető leghasonlóbb egy adatbázisban lehetőleg minél nagyobb gyakorisággal szereplő elemekhez. (Az adatbázis a beszélő korábbi nyelvi tapasztalatait kívánja ábrázolni.)

Az ilyen elven működő algoritmusokban nem annyira a korábbi elemek hasonlósági osztályai (klaszterei) játszanak szerepet, mint maguk azok az elemek, amelyek ezeket az osztályokat mind gyakoriságuk, mind tulajdonságaik alapján a legjobban képviselik, középponti szerepet játszanak bennük, vagyis prototipikusak.

A statisztikai megközelítésekben a klaszterek prototípusának fogalmát úgy szokták értelmezni, hogy az a klaszter ún. centroidjához (súlypontjához) legközelebb eső elem. Ennek a prototípus-fogalomnak azonban több hátránya is van. A legfontosabb

az, hogy a klaszter prototípusának meghatározásához először is magát a klasztert kell meghatározni, ennek a folyamatnak minden nehézségét le kell küzdeni, hiszen — ebben az esetben teljesen szükségtelenül — döntést kell hozni a klaszter határának kérdésében. A másik probléma, hogy a súlyponthoz legközelebb eső elem nem feltétlenül a klaszter legsűrűbb részére esik.

Írásunkban egy teljesen más megközelítést javasolunk: azokban a feladatokban, amelyekben a prototípusokra szükség van, de magukra a klaszterekre nem, olyan algoritmusokat is alkalmazhatunk, amelyek közvetlenül a prototípusok megtalálására irányulnak, maguknak a klasztereknek a határait pedig nem próbálják meghúzni [10]. Az általunk javasolt algoritmus a következő egyszerű alapfeltevéseken alapul:

- A klaszter prototipikus eleme legyen minél gyakoribb [1, 6].
- A klaszter prototipikus elemének közelében minél több minél gyakoribb elem legyen, vagyis sok elem hasonlítson rá.
- A klaszter prototípusa minél távolabb legyen, minél kevésbé hasonlítson más klaszterek prototípusaira.

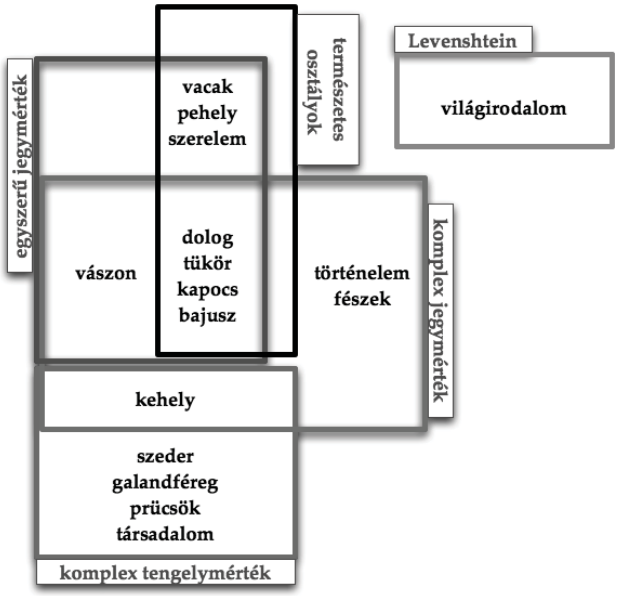
Megmutatjuk, hogy ezeknek a kritériumoknak az alapján viszonylag egyszerű algoritmusokkal hatékonyan megtalálhatóak a klaszterek prototípusai, amelyek nagyjából egybeesnek a hagyományos meghatározás szerinti prototípusokkal, de a módszer egyes nyelvészeti feladatokban talán még kedvezőbb eredményekhez is vezet, mint a hagyományos megközelítés.

2 Prototípusok gépi meghatározásának módja

A prototípusok kiválasztása során saját fejlesztésű algoritmusokkal számítjuk a hasonlóságot, amelyek a kurrens, hasonlóság mérésére használt algoritmusoknál (pl. [9]) finomabb összehasonlításokat is lehetővé tesznek. A komplex jegymérték és a komplex tengelymérték nevű algoritmusok a szavak hasonlóságát azok jobb szélétől véve számítják ki úgy, hogy a megfeleléseknek, hasonlóságoknak egyre kisebb súlyt adnak a szavak bal széle felé haladva. Így mind a két számítógépes algoritmus a *vas* és *sas* szavakat hasonlóbbnak tekinti, mint a *vas* és a *vaj* szavakat. Az algoritmusok a hasonlítást az egyes fonémák jegyei alapján végzik el, de a komplex jegymérték [7, 8] fonémákat hasonlít össze, míg a komplex tengelymérték az egyes jegyek tengelyeinek hasonlósága alapján számítja ki két szó hasonlósági értékét. Ezeket az értékeket egy 0-1 terjedő skálán adtuk meg.

További összehasonlítási eljárásunk a fonológiai természetes osztályokon alapszik, amellyel a komplex jegymértékhez hasonlóan vetettük össze a szavakat, csak két fonéma hasonlóságát annak révén határoztuk meg, hogy hány közös és hány eltérő természetes fonológiai osztályban szerepelnek ezek [2]. Összehasonlításainkban összesen 13 fonológiai jegyet vettünk figyelembe. Egyedül a komplex jegymérték egyszerűsített variánsa (egyszerű jegymérték) esetében alkalmaztunk 8 jegyet. Saját algoritmusainkat a közismert és általánosan szó-összehasonlításra is használt Levenshtein-algoritmus [5] teljesítményéhez mértük.

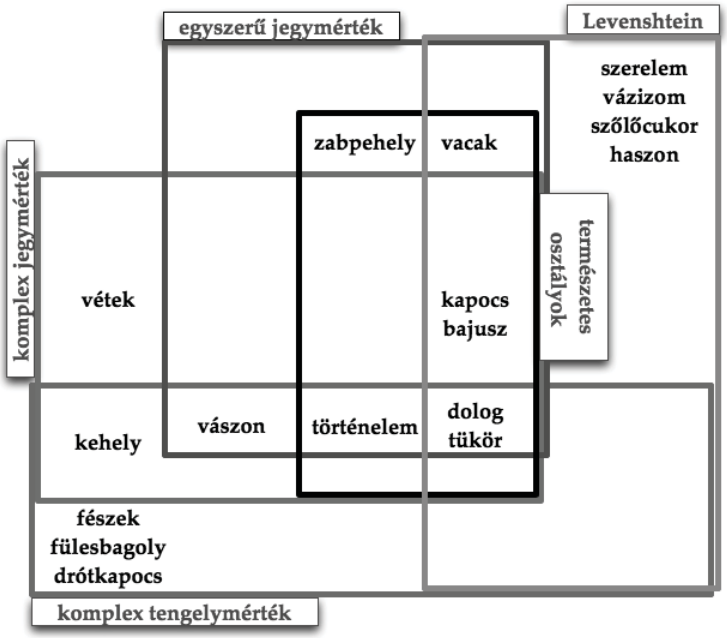
Kísérletünkben magyar hangkivető főnevek ingadozását (pl. *sátrat* : *sátort,* de *szerelmet* : **szerelemt*) modelleztük úgy, hogy ingadozásuk mértékét a legközelebbi (leghasonlóbb) prototípushoz való hasonlóság alapján határoztuk meg. Korábbi elemzések tapasztalatai alapján [4, 7, 8] a prototípusok kiválasztására egy olyan algoritmust hoztunk létre, amely a bevezetőben megadott kritériumok alapján működik. A prototípuskiválasztásban többféle küszöbértéket is megadhatunk, amelynek növelésével algoritmusunk egyre szigorúbban alkalmazza a hasonlósági szempontokat, miszerint a prototípushoz sokan hasonlítanak, de az más prototípusokra nem hasonlít. Két eltérő küszöbérték mellett kiválasztott prototipikus szavainkat a 1-2. ábrák mutatják meg.



1. ábra. 0,9-es értékkel növelt küszöbérték[1] mellett kiválasztott prototípusok[2].

[1] A teszt ismertetése szempontjából nem fontos, hogy a küszöbérték-növelések értéke pontosan hogyan járul hozzá az algoritmus működéséhez. Az ábrák értelmezéséhez csak annyit kell tudnunk, hogy minél magasabb ez az érték, az algoritmus annál szigorúbban alkalmazza a kiválasztásban a hasonlósági kritériumainkat.

[2] Az ábrákon Venn-diagramokat láthatunk az Edwards-féle módosításban, ami lehetővé teszi öt halmaz elemeinek is az összehasonlítását. A halmazok megjelenítését tartalmuk függvényében átalakítottuk a könnyebb áttekinthetőség érdekében.



2. ábra. 0,5-ös értékkel növelt küszöbérték mellett kiválasztott prototípusok.

Mielőtt áttekintenénk, hogy az egyes prototípusok mennyire jól modellezték a hangkivető szavak hangkivetésének mértékét, érdemes őket szemügyre venni. Minden mérték esetében jellemző a gyakori alakok preferenciája. Ez legszembetűnőbben a *dolog* mint prototípus választásában jelenik meg, mivel az összes hangkivető előfordulás mintegy 16,1%-át teszi ki (348 ezer egyes szám alanyesetű előfordulás a *Szószablya Gyakorisági Szótár*ban), és 2,42-szer gyakoribb, mint az őt közvetlenül követő *társadalom*. A választásokban további nagyon gyakori szavak is szerepelnek még: *szerelem* (68 ezer)*, társadalom* (144 ezer)*, történelem* (68 ezer). A *dolog*-gal együtt ezek már az összes hangkivető főnév alanyesetű előfordulásainak 29,1%-át fedik le. E kiugróan gyakori elemeken túl azonban a prototípusválasztó algoritmus inkább a hasonlósági szempontokat veszi figyelembe, hisz a következő leggyakoribb szó, a *tükör* (29., 21 ezer) már jóval elmarad ezek mögött. Az összes mértéken alapuló választásnál megfigyelhető, hogy habár a gyakori *-alom, -elem* végűek alkotják a legszámosabb alcsoportját a hangkivető szavaknak, mégis ezek vannak leginkább alulreprezentálva a prototípusok tekintetében. Általában az egyes prototípuscsoportokban csak *-elem* végű prototípus jelenik meg, ami egyaránt jól lefedi az *-alom* végűeket és a többi *-e-e* végű szót is.

A prototípusválasztó algoritmus azonban kevésbé gyakori szavakat is választ, ha azok a hasonlósági kritériumoknak jobban megfelelnek. Ezek gyakran összetett szavak, hisz hosszúságuk alapján jobban reprezentálják a zömükben összetett hangkivető szavakat, mint a példánygyakoriságban gyakoribb, de típusgyakoriságban ritkább alapszavak. Ilyen szavak a *zabpehely* (egyszerű jegymérték, természetes osztályok)*, szőlőcukor, vázizom* (Levenshtein-algoritmus)*, drótkapocs, galandféreg, fülesbagoly,*

(komplex tengelymérték). Kisebb, de jól elkülönülő szócsoportok is több esetben kapnak önálló prototípust: *vacak, bajusz* (utolsó magánhangzó nem középső nyelvállású), *vászon* (*-(á/ó)CVC* végűek), *zabpehely, kehely* (hangátvetés), *vázizom, pityer* (*-iCVC* végűek).

Az egyes hasonlósági mértékek és a két eltérő küszöbérték mentén kiválasztott prototípusokat az olyan hangkivető főnévekkel hasonlítottuk össze, amelyek legfeljebb 99,99%-ban mutattak hangkivető viselkedést (282 szó) az olyan toldalékokkal, amelyek esetében hangkivető alakokat várnánk el. Vizsgálatunkból azért zártuk ki a 100%-ban hangkivető főneveket, mert ezek esetében legfeljebb csak a kiugróan gyakoriaknál tudhatjuk, hogy az ingadozás hiányának oka következetes viselkedésük, és a 100%-ban hangkivető viselkedés nem adathiánynak tudható be. A prototípusokhoz az egyes szavakat mindig olyan mérték alapján hasonlítottuk, amilyen mértéket a prototípus kiválasztásában is alkalmaztunk. Miután minden, a vizsgálatra kiválasztott hangkivető főnevet minden prototípuscsoporttal (2 x 5 db) összehasonlítottuk, megvizsgáltuk, hogy az egyes szavak hangkivetési mértéke[3] mennyire korrelál a hozzá legközelebbi prototípushoz való hasonlóságával.

Feltételezésünk szerint egy szó minél jobban hasonlít a hozzá leghasonlóbb prototípushoz, annál nagyobb a hangkivetési mértéke is. Az együttjárások számítása során a prototípushoz való hasonlósági értéket súlyoztuk a hasonlítandó hangkivető főnév releváns toldalékos alakjai alapján meghatározott gyakoriságának 8. gyökével (pl. *dolog* esetében 5,23, a *sátor*-nál 3,34), mivel a gyakoribb főneveknél magasabb hangkivetési mértéket vártunk, de nem kívántunk ennek az értéknek túlzott súlyt sem adni. Az 1-2. ábrákon bemutatott prototípusokon túl a szavakat hasonlítottuk a *Szószablya Gyakorisági Szótár*ban az egyes szám alanyesete alapján 50 leggyakoribb hangkivető főnévhez is, mint olyan prototípusokhoz, amelyeket kizárólag gyakoriságuk alapján választottunk ki a hasonlósági szempontok figyelmen kívül hagyásával. Gyakorisági prototípusnak azért választottunk ki viszonylag több szót, mert a 10 leggyakoribb hangkivető főnévből 8 *-alom/-elem* végű volt, így ennél több szóra volt szükségünk ahhoz, hogy ne csak az *-alom/-elem* csoporthoz való hasonlóságot mérjük. A prototípusok számának növelése nem jár szükségszerűen együtt a korreláció mértékének növelésével, hisz ha az összes hasonlítandó szót felvennénk prototípusnak, akkor az önmagukhoz való hasonlóságuk 1 lenne, aminek következtében egyáltalán nem tudnánk érdemleges együttjárásokat megfigyelni a változó hangkivetési mértékek és a konstans 1-es értékek közt.

3 Kísérletünk eredményei

Az 1. táblázat alapján láthatjuk, hogy – a Levenshtein-algoritmust leszámítva – már az összes legközelebbi prototípushoz való hasonlóság közepesen korrelál a szavak hangkivetési mértékével. A komplex tengelymérték a legmagasabb együttjárást mu-

[3] Hangkivetési mérték alatt az értjük, hogy a hangkivetéssel együttjáró toldalékok (pl. tárgy, szuperesszívusz, birtokos ragok stb.) esetében mennyire stabilan jelentkezik a hangkivetés. Így ez az érték az *–alom* végű szavaknál többnyire 100%, a *sátor* esetében 81%, míg a *bajusz*-nál csak 36%.

tatja. A hangkivetés mértékét legjobban megragadó prototípusaink: a *dolog, történelem, tükör, vászon, fészek, kehely, fülesbagoly, drótkapocs*. Ez a néhány szó viszonylag jól fedi a lehetséges végmintázatokat és záró magánhangzó-szekvenciákat is, amelyek a viselkedés szempontjából a legfontosabbak lehetnek. A *fülesbagoly* és a *drótkapocs* a nagyszámú összetett szót, a *kehely* egy speciális mintát, a *vászon* pedig a mérsékelten hangkivető szavak csoportját képviseli.

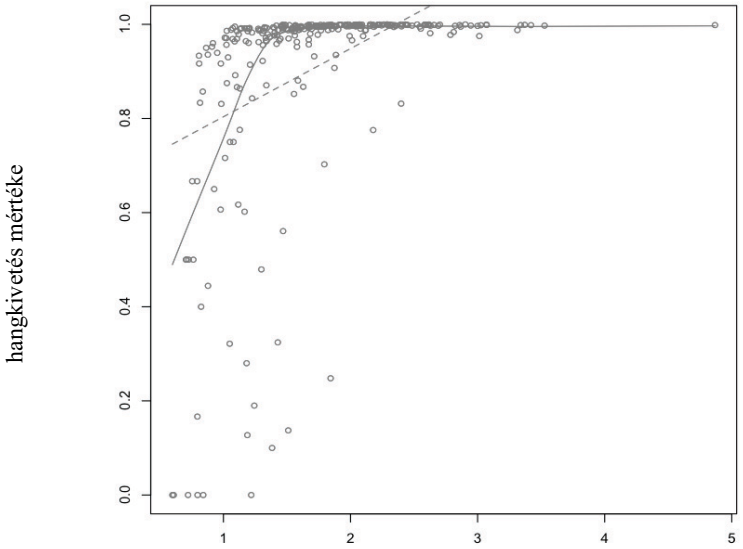
1. táblázat: A hangkivetési mérték és a prototípushoz való hasonlóság együttjárásának mértéke a felhasznált prototípusok függvényében.

	Levenshtein	**Egyszerű jegyek**	**Komplex jegyek**	**Természetes osztályok**	**Komplex tengelymérték**
0,5 küszöbérték	0,241***	0,352***	0,364***	0,371***	0,419***
0,9 küszöbérték	0,248***	0,352***	0,362***	0,370***	0,409***
gyakori szavak	0,346***	0,458***	0,461***	0,455***	0,423***

** = $p < 0,01$
*** = $p < 0,001$

Mindösszesen ennek a 8 szónak az alapján szabályoknál hatékonyabb és könnyebb módon 274 másik szó viselkedését tudjuk viszonylagos megbízhatósággal jellemezni. Az eredetileg csak viszonyítási alapnak szánt 50 leggyakoribb szóhoz való hasonlítás alapján azonban láthatjuk, hogy a gyakoriságnak van a legkiugróbb szerepe a szavak viszonyrendszerében. Ha csak a számunkra fontos szavak 20%-ához van gyors hozzáférésünk, már akkor egészen jól tudjuk leírni a maradék 80% viselkedését. Ha szavainkat a komplex jegymérték alapján azonosítható hasonlósági csoportok[4] leggyakoribb szavaihoz hasonlítjuk a komplex jegymértékkel, akkor ismét közepesen erős korrelációt tudunk kimutatni ($r(280) = 0,4$, $t = 7,31$, $p > 0,001$). Ez alapján láthatjuk, hogy ha a gyakoriságot lokálisan értelmezzük egy adott csoporton belül, akkor is képesek vagyunk az egyes szavak hangkivetési mértékével kapcsolatban együttjárásokat megfigyelni. Ha komplex jegymérték (0,5-ös küszöbérték) által kiválasztott prototípusainkból és a leggyakoribb szavakból alkotott csoporthoz hasonlítjuk hangkivető szavainkat, akkor némileg még szorosabb együttjárást ($r(280) = 0,485$, $t = 9,27$, $p < 0,001$) figyelhetünk meg a halmaz szavaiból kiválasztott leghasonlóbb prototípusok hasonlóságértéke és a hangkivetési mértékek közt. Ebből arra következtethetünk, hogy ha a prototípus kiválasztásában alkalmazott szempontjainkat még jobban optimalizálnánk, akkor a hangkivetési mértéket vagy akár bármilyen más viselkedési mutatót jobban tudnánk megragadni.

[4] Az összes hangkivető főnév viszonyait megragadó hasonlósági gráfban 50 hasonlósági csoportot tudunk azonosítani, ha csak a legszorosabb kapcsolatokat vesszük figyelembe.



3. ábra. Komplex tengelymérték leghasonlóbb prototípusaihoz való hasonlóság és a hangkivetési mérték összefüggése a *Szószablya Korpusz*ban.

Bibliográfia

1. Bybee, J. L., Eddington, D.: A usage-based approach to Spanish verbs of 'becoming.' Language Vol. 82 (2006) 323–355
2. Frisch, Stefan A.: Similarity and Frequency in Phonology. PhD-disszertáció (1996) http://www.cas.usf.edu/~frisch/Frisch96.pdf (2010.07.01.)
3. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: A Szószablya projekt. In: Alexin Z., Csendes D. (szerk.): Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). (2003)
4. Kálmán L., Rebrus P., Törkenczy M.: Lehet-e az analógiás nyelvelmélet szinkrón? A magyar nyelvészeti kutatások újabb eredményei II., Kolozsvár. 2010. április 16. http://budling.nytud.hu/~tork/KRT/bbte10_slides_print.pdf (2010.07.01.)
5. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics - Doklady Vol.10, No. 8 (1966) 707–710
6. Nosofsky, R. M. Exemplar based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition Vol. 14 (1988) 700–708
7. Rung A.: Determining word similarity in the Hungarian language. In: Kálmán L. (szerk.): Papers from the Mókus Conference. Tinta Kiadó, Budapest (2008) 112–118
8. Rung A.: Szóhasonlóság mérése analógiás megközelítésben. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2009. Szegedi Tudományegyetem, Szeged (2009) 104–113

9. Skousen, R., Lonsdale, D., Parkinson, D. B. (szerk.): Analogical Modeling. John Benjamins, Amsterdam (2002)
10. van den Bosch, A.: Expanding k-NN analogy with instance families. In: Skousen, R., Lonsdale, D., Parkinson, D. B. (szerk.): Analogical Modeling. John Benjamins, Amsterdam (2002) 209–223