Toldalékolós baseline

Rácz

27/24/2022

80 ember válaszolt fejenként 162 szóra, így lett kb 25 válasz egy szóra.

Feltéve, hogy a kísérlet működik, két lehetőség volt: (i) a válaszokat a célszavak jól kontrollált alapvető tulajdonságai határozzák meg (szótagszám, végződés, ilyesmi), (b) a válaszokat az határozza meg, hogy a célszavak általában mennyire hasonlítanak bizonyos módon viselkedő létező szavakra. (Ez a két dolog persze átfed.)

Végigmegyek a három szókategórián, megnézem, hogy egyáltalán hasonlítottak-e a kísérletben adott válaszok arra, ahogyan a kategóriába tartozó létező szavak viselkednek az új webkorpuszban (jellemzően igen), és hogy ezt specifikus tulajdonságok vagy valamilyen általános hasonlósági metrika határozta-e meg (is!).

Dzsungelban/dzsungelben

Célszavak

A célszavak főnevek, a válaszlehetőségek a -ban, -nak, -nál elöl vagy hátulképzett alakjai. Egy szóhoz egy toldalék tartozik, amit kidobtam kockával: tehát ban, nak, vagy nál.

Azt, hogy a célszó főnév, és a válaszlehetőségek posztpozíciós alakok, a szövegkörnyezet tette tisztába:

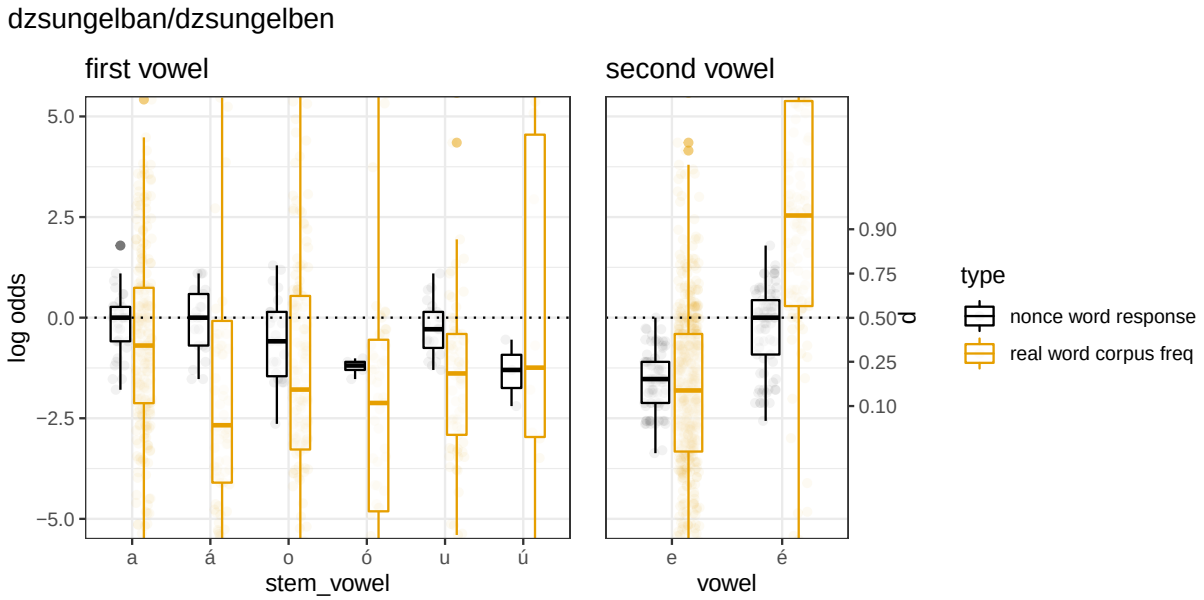
- "Ez itt egy kalréz."
- "Elneveztem a kutyámat..." / "Ott vagyunk a..." / "Nincs is jobb egy jó..."
- "kraléznak - kraléznek" / "kralézban - kralézben" / "kraléznál - kraléznél"

A két variáns sorrendje trialonként változott, összevissza módon.

Minden célszó kétszótagos, az első magánhangzó hátul képzett vagy i/í, a második e/é. i/í szóból csak néhányat akartam eredetileg, kontrollnak (erősen azt várjuk, hogy ezek elölképzett toldalékot választanak), de aztán kicsit több lett végül, mint kéne, nem baj.

Mintázatok

Először nézzük meg, mennyire hasonlítanak a kísérletben adott válaszok az új webkorpuszban látott mintázatokra. Ezt mutatják a lenti ábrák:



A baloldali ábrán azt látjuk, hogy az első magánhangzó (mgh), a másodikban azt, hogy a második mgh mennyiben befolyásolja a toldalék hátulképzettségét. Az y tengelyen a hátul és elölképzett válasz log oddsa látható (a jobb oldalra tettem egy valószínűségeket mutató tengelyt is segítségnek, de a mértékegység a log odds). Tehát magasabb érték = több ban-nak-nál. Alacsonyabb érték = több ben-nek-nél. Nulla = fele-fele arány.

A fekete eloszlás a kísérletben, a sárga a korpuszban kapott arányok. Mindent pont egy szópár. Az ábrára rá van nagyítva, mert a korpuszos eloszlások sokkal zajosabbak, és nem lehetne látni jól a kontrasztokat, ha benne lenne az egész mindenség.

Valószínűbb az elölképzett alak, ha az első mgh hosszú u vagy o, illetve ha a második mgh e, nem pedig é. Ez igaz a korpuszra is és a kísérletre is. Az e-é kontraszt nem meglepő, a hosszú u-o valszeg a szó összetettszó hangulatát erősíti, erre később visszatérek.

Na jó, szóval bizonyos alapvető paraméterekben hasonlít a kísérlet a korpuszra, de miért van ez?

Lexikon

A Rebrus Péter szerint a hotel/dzsungel/fotel szavak a magyarban három osztályba sorolhatók.

1. Egyértelműen idegen eredetű, művelt regiszterben használt szavak (kódex, modell, koncert, projekt), ezek jellemzően *elölképzettek* lesznek.
2. Szlengszavak, sőt argó (matek, haver, kolesz), ezek jellemzően *hátulképzettek* lesznek.
3. A többiek (dzsungel, bunker, motel), ezeknek kb mindegy.

Az új magyar webkorpuszban található vegyes hangrendű főnévi tövek közül kiválasztottam azokat, amelyek elég gyakoriak, szoktak variálni szépen hátul- és elölképzett alakok között, aztán számoltam mindegyikre egy átlagos hátul-elölképzettségi arányt, és ez alapján három egyenlő elemszámú részre osztottam a szavakat, nagyon hátulképzett, nagyon elölképzett, és a langyosak (őket kiköpi az úr).

Ezek a számalapú osztályok elég jól hozzák a Rebrus Péter intuícióit. A (főleg) hátulképzett kategóriában ilyen szavak vannak:

```
##  [1] "pancser"  "oszét"    "flaszter" "kóter"    "bakter"   "mutter"
```

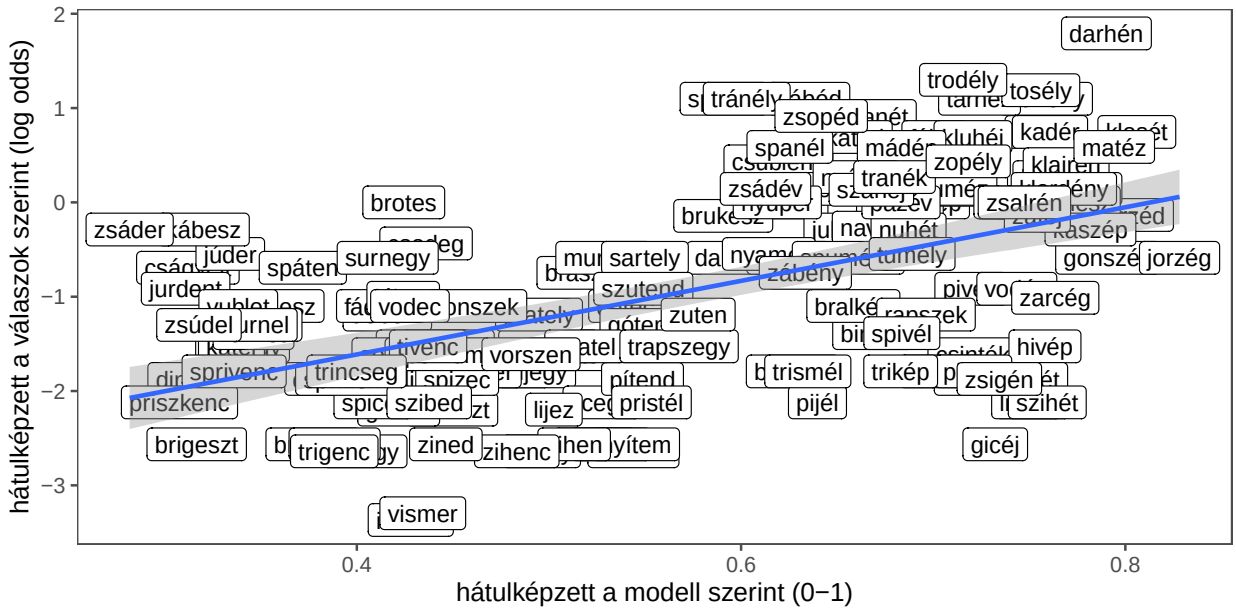
```
##  [7] "bohém"    "samesz"   "szlovén"  "arzén"    "malter"   "macesz"
## [13] "patent"   "klozet"   "muter"    "krapek"   "szolfézs" "balek"
## [19] "honvéd"   "fater"    "kadét"    "ankét"    "norvég"   "affér"
## [25] "haver"    "maszek"   "trapéz"   "matek"    "konkrét"  "koktél"
## [31] "poén"     "árnyék"   "fazék"    "tányér"   "húsvét"   "pótlék"
## [37] "szomszéd" "kastély"  "játék"
```

A (főleg) elölképzettben ilyenek:

```
##  [1] "tanév"     "háttér"    "koncert"   "projekt"   "modell"    "szoftver"
##  [7] "korrekt"   "duett"     "sláger"    "partner"   "zsáner"    "bróker"
## [13] "kvartett"  "hardver"   "poszter"   "kábel"     "transzfer" "póker"
## [19] "kráter"    "rubel"     "dzsungel"  "konvent"   "advent"    "kókler"
## [25] "rulett"    "moped"     "roller"    "púder"     "amper"     "raszter"
## [31] "makett"    "puffer"    "páter"     "parkett"   "hotel"     "pasztell"
## [37] "fráter"    "kartell"   "sólet"     "bankett"
```

Van pár de facto összetett szó is, de emellett a Rebrus Péter kategóriái jól látszanak szerintem: argó (pancser, flaszter, muter, krapek) és művelt (partner, koncert, modell, parkett).

Írtam egy sima kategorizációs algoritmust, amely kiszámolgatja, hogy egy adott célszó inkább az első vagy a második csoportra hasonlít, és ez alapján ad neki egy hasonlósági súlyt. Ez a súly valamennyire korrellál azzal, mennyire nézték az emberek elöl/hátulképzettnek a célszavakat a kísérletben:



Ez persze annyiban csal, hogy a két létező szavak alkotta célkategóriát (nagyon elölképzett és nagyon hátulképzett) teljesen diszkrétnek képzeli el, ami nyilván abstart nem igaz, de szerintem elég messze vannak egymástól az eloszlások ahhoz, hogy ez ne legyen komoly gond.

Csináltam egy regressziós modellt, ami azt jósolja meg, mekkora lesz az elölképzett válaszok aránya, abból, hogy az adott célszó:

- mennyire hasonlít az elölképzett szavakra a kategorizációs modell szerint
- mi az első magánhangzója
- mi a második magánhangzója

- mi a toldalék amivel szerepelt
- egynél több mássalhangzó van-e a két magánhangzó között ("iszkenc" vs "trodély") – ez az összetettszó-szerűségnek egy primitív mérőszáma (a trodély tuti lehet monomorfemikus, az iszkenc nem biztos)

A modell r^2-je .8. Itt vannak a becslések a prediktorokra:

term	estimate	std.error	statistic	conf.low	conf.high
(Intercept)	-0.62	0.33	-1.86	-1.26	0.03
cat1_weight	0.01	0.59	0.02	-1.15	1.17
long_clusterTRUE	-0.28	0.09	-3.13	-0.45	-0.10
stem_vowelá	0.01	0.15	0.05	-0.28	0.29
stem_voweli	-2.00	0.13	-15.58	-2.26	-1.75
stem_vowelí	-1.38	0.29	-4.73	-1.99	-0.84
stem_vowelo	-0.08	0.11	-0.68	-0.30	0.15
stem_voweló	-0.39	0.25	-1.59	-0.89	0.08
stem_vowelu	-0.16	0.12	-1.31	-0.40	0.08
stem_vowelú	-0.47	0.31	-1.51	-1.10	0.12
vowelé	1.17	0.17	7.03	0.85	1.50
suffixnak	-0.42	0.09	-4.44	-0.60	-0.23
suffixnál	-0.29	0.09	-3.17	-0.47	-0.11

Mit látunk?

- Az intercept nem értelmezhető önmagában.
- Nem túl meglepő, hogy az első magánhangzó (ezt hívom stem vowel-nek) és a második magánhangzó (ez simán vowel) baromi fontosak: Nyilván, ha az első mgh i vagy í, akkor nagyon elölképzett lesz a válasz (negatív értékek). A hosszú ó és ú is elölképzettebb szót eredményez, valszeg azért, mert ezektől a célszónak összetett szó hangulata támad (ld lentebb)
- Szintén nem túl meglepő, hogy ha a második magánhangzó é, akkor az alapértelmezett e-hez képest jóval hátulképzettebb lesz a válasz (pozitív értékek: tehát az é átlátszóbb).
- Ha több msh van a két mgh között, akkor inkább lesz elölképzett a szó (ha összetettszó-szerű, akkor úgy is fog viselkedni, mint egy összetett szó: tanév, látkép).
- Ha a toldalék -nak vagy -nál, elölképzettebb lesz a szó, mint az itt alapértelmezett -ban. Passz. Mondjuk ez nem egy óriási különbség.
- És a legérdekesebb: ezek a tulajdonságok fontosabbak, mint a hasonlósági kategóriasúly (amely itt a nem nagyon egyértemű cat1_weight nevet kapta). Ez nem ad hozzá semmit az eredményhez.

Tehát úgy tűnik, hogy a résztvevők a válaszaikat az alapján hozzák meg, hogyan néznek ki a célszavak. Ez jó. De itt nem holisztikusan nézik, hogy hogyan néz ki a szó, hanem bizonyos fix dolgokat néznek, és az alapján "döntenek": milyen a mgh, összetett-e a szó vajon intuitíve.

Én ennek az ellenkezőjét vártam, a Rebrus-féle sejtéssel párhuzamosan: ha egy szó kódex-hangulatú, akkor elölképzett lesz, ha stukker-hangulatú, akkor hátulképzett. Ez nem jött be.

Lakok/lakom

Célszavak

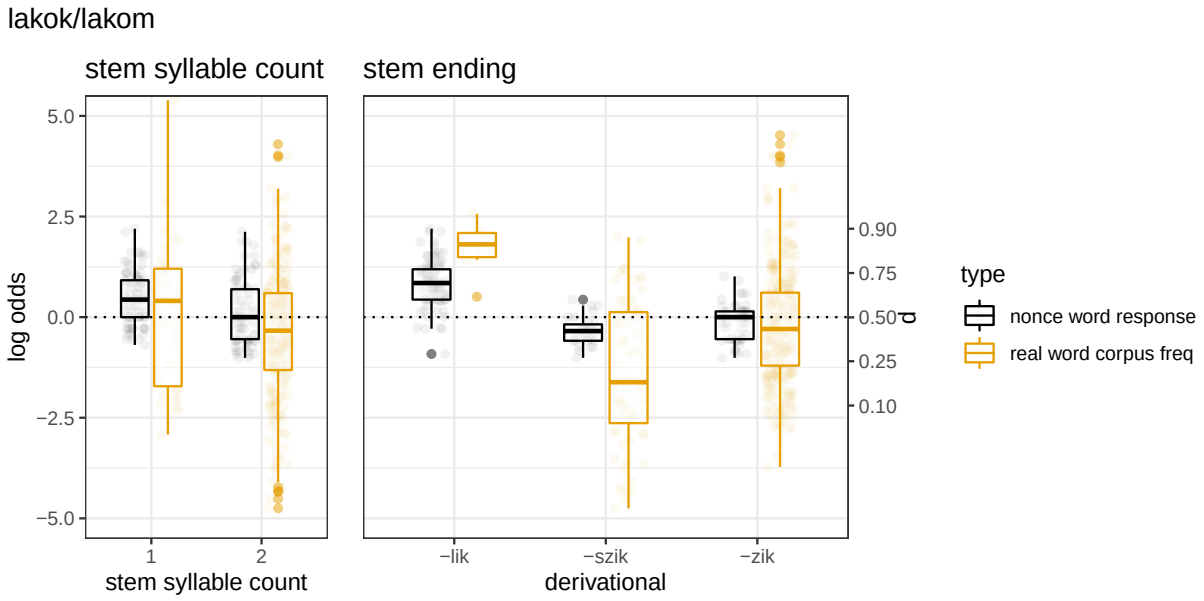
Ezek a célszavak két vagy háromszótagosak voltak (tehát egy vagy kétszótagos volt az igető kvázi) és három képző valamelyikével végződtek (-szik, -zik, -lik).

A két válaszlehetőség minden esetben az indefinit alak jelölt -m és a jelöletlen -k változatai voltak.

- “Te bizony sokat pratánylasz.”
- “Én is sokat. . . ”
- “pratánylok - pratánylom”

Mintázatok

Nézzük, mennyire hasonlítanak a kísérletben adott válaszok az új webkorpuszban látott mintázatokra. Ezt mutatják a lenti ábrák:



A baloldali panelon azt látjuk, hogy az egyszótagú tövek inkább lesznek -k-sak, mint a kétszótagúak.

A jobboldali panel azt mutatja, hogy a -lik sokkal -k-sabb, mint a másik két képző. A korpuszban ide kevés, de olyan jó ikes, tárgyatlan ige tartozik, mint a botlik, csuklik, hanyatlik. A másik két osztály jóval nagyobb, mert az -ozik és a -szik elég produktívok.

Az új webkorpuszon végigment egy morfológiai egyértelműsítő. Tehát elvileg meg tudja különböztetni az “eszem egy kenyeret” alakokat az “eszem a kenyeret” alakoktól a mondatszerkezet alapján. De ebben nem lehetünk teljesen biztosak. A hibák azoknak a szavaknak fognak “kedvezni”, amelyeknek eleve van tárgyas alakjuk, mivel ezek fognak feltűnni -m végződéssel a korpuszban.

Elképzelhető például, hogy az indef “egész nap dolgozom” és def “végigdolgozom a napot” típusú alakokat az elemző keresztbe fogja címkézni. Olyan igéknél, amelyek szívesen tárgyasak, ez a hiba megdobja az “indef” (valójában rosszul címkézett) -m alakok arányát.

Ez a Rebrus Péter félelme, én itt két okból nem aggódnék ezen nagyon. Ha megnézzük a végződéseket fent jobbra, akkor egyrészt az emberek hozzák azt, amit a korpuszban találunk, és ők nem tudnak ilyen értelemben rosszul parse-olni. Abban az értelemben persze tudnak, hogy inkább tudnak analógiázni létező gyakori paradigmacellákból. Másrészt ezek a képzett igealakok nem olyan nagyon tárgyaznak, a definit alakok előfordulása intuitíve alacsony lesz. Erre lentebb mutatok példát.

Lexikon

Ugyanúgy fogtam azokat a szavakat, amik az új webkorpuszban 1sg.indef alakban váltakozást mutatnak, és megnéztem, hogy melyik mennyire preferálja az -m végződést a -k-val szemben.

A főleg -k kategóriában ilyen szavak vannak:

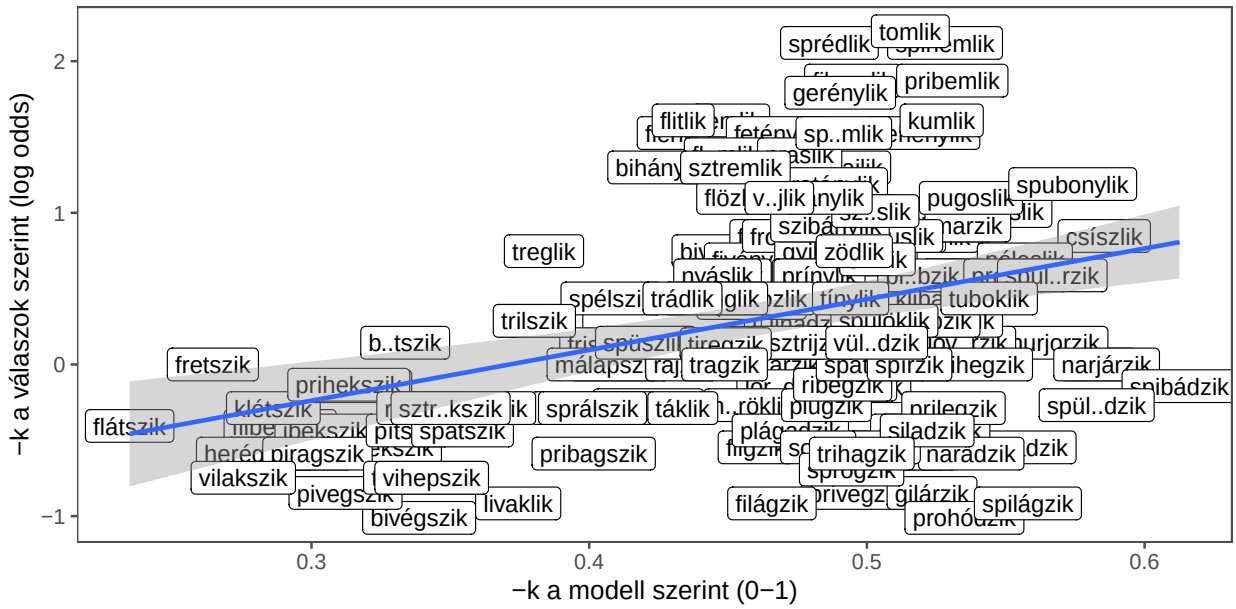
```
##  [1] "gúnyolódik"        "emlékezik"         "lélegzik"
##  [4] "elmélázik"         "játszódik"         "hízik"
##  [7] "taktikázik"        "kézilabdázik"      "kávézik"
## [10] "nekiveselkedik"    "kerekedik"         "tétlenkedik"
## [13] "csalódik"          "oldódik"           "sopánkodik"
## [16] "kapálódzik"        "alakoskodik"       "autókázik"
## [19] "kezdődik"          "ügyeskedik"        "csigázik"
## [22] "közlekedik"        "ejtőzik"           "ordítozik"
## [25] "háborúzik"         "időzik"            "nyűglődik"
## [28] "okoskodik"         "élősködik"         "húzódzkodik"
## [31] "leselkedik"        "építkezik"         "túlórázik"
## [34] "rendezkedik"       "karikázik"         "szabálytalankodik"
## [37] "teljesedik"        "epekedik"          "jóllakik"
## [40] "törölközik"        "pipázik"           "falatozik"
## [43] "mentegetőzik"      "hazudozik"         "döglődik"
## [46] "tántorodik"        "teniszezik"        "nedvesedik"
## [49] "fejlődik"          "bújócskázik"
```

A főleg -m kategóriában ilyenek:

```
##  [1] "húzódozik"       "betegszik"       "tetszik"         "viaskodik"
##  [5] "fogadkozik"      "átkozódik"       "vágyakozik"      "erőszakoskodik"
##  [9] "alkalmazkodik"   "kopaszodik"      "avatkozik"       "öltözködik"
## [13] "álmodik"         "jelentkezik"     "imádkozik"       "kacérkodik"
## [17] "iszonyodik"      "hagyatkozik"     "vállalkozik"     "tisztálkodik"
## [21] "csókolózik"      "tartózkodik"     "öregszik"        "bízik"
## [25] "gondolkodik"     "szándékozik"     "vadászik"        "nyugszik"
## [29] "zongorázik"      "kutyázik"        "szomorkodik"     "jógázik"
## [33] "vétkezik"        "szomjúhozik"     "bizakodik"       "hősködik"
## [37] "fukarkodik"      "foglalatoskodik" "udvariaskodik"   "származik"
## [41] "cigarettázik"    "fohászkodik"     "töltődik"        "törekszik"
## [45] "szűkölködik"     "bocsátkozik"     "fázik"           "akaszkodik"
## [49] "osztozkodik"     "bűnhődik"
```

Ugye ha erre úgy ránéz az ember, akkor nincsenek kiugró különbségek. Valami gyakorisági izé van, hogy az iszik meg a fázik az inkább m-es, az aszik meg nem.

Ha megépítjük az előző körből ismert tanulóalgoritmust, és megnézzük, hogyan kategorizálgatja a kísérlet célszavait, azt kapjuk, hogy a kategóriahasonlóság itt is jósolja azt, hogy milyen válaszokat kapunk a kísérletben:



Ha pedig regressziós modellt építünk, amibe betesszük ezt a kategóriasúlyt, és azt, hogy a szótő hány szótagos, és milyen képzővel dolgozunk, akkor a modell r^2-e .57 lesz (tehát jóval pontatlanabb, mint a főneves, jóval kevésbé tudja megmondani, mit fognak csinálni az emberek), és így néznek ki a prediktorok:

term	estimate	std.error	statistic	conf.low	conf.high
(Intercept)	-0.30	0.37	-0.82	-1.04	0.43
cat1_weight	3.28	0.78	4.21	1.76	4.81
nsyl	-0.30	0.07	-4.50	-0.42	-0.17
derivational-szik	-0.70	0.14	-4.86	-0.98	-0.42
derivational-zik	-1.02	0.07	-13.67	-1.17	-0.87

- Az intercept megint nem túl értelmes magába.
- A kategóriasúly nagyon zajos prediktor, de eléggé pozitív, tehát valamennyire hasznosak a hasonlóságalapú jóslatok: mennél jobban hasonlít a célszó k-s szavakra, annál több k választ kap. Persze az, hogy ez a prediktor itt ilyen nagyot megy, az azért is van, mert a szavak belsejéről itt mást nem tudunk mondani (az előző modellban ott volt rögtön mind a két magánhangzó, ami egy harmonikus váltakozásnál azért elég fontos).
- A hosszabb szavak inkább -k-sak (azt tippeltük, hogy ez valszeg azért van, mert Kedvenc Jelölt Szavaink, mint az eszek, a lakok, az iszok, stb, mind rövidebbek).
- A szik és a zik jóval -m-esebb, mint a -lik. Ezt láttuk a nyers adatokban is.

És hát itt a legfontosabb az, hogy bőven nem használták kategorikusan az emberek az -m szabályt az új szavakra, de közben jónéhány -m válasz becsúszott, és viszonylag szisztematikus módon.

Ez.. fura? A webkorpusz legjelöltebb -m-es alakjai jellemzően művelt regiszterbe tartozó olyan klasszikusok, mint az esküszöm, tolakszom, törekszem, származom. A legkevésbé -m-esek olyan gyakori képzett alakok, mint a bűnözök, távozok, szűnök, tartozok, emlékezek. Az lik/szik különbség itt ránézésre egyáltalán nem feltűnő például.

Én azt gondolnám, hogy van a magyarban egy elég zárt belvárosi halálzóna szókincs (eszek, bízok, játszok, iszok, lakok, alszok, stb), és ez képes lehet valamennyire produktívan kiterjeszteni az indef m-et hasonló új alakokra. Na itt nem ez látszik: nagyjából egy alaktani hasonlóságokra épülő papírforma érvényesül, ahol a valahogyan viselkedő létező alakokhoz hasonló célszavak követik a megfelelő viselkedési mintát, egy bizonyos mértékig.

cselekedik / cselekszik

Célszavak

Ez volt a legelborultabb kategória. Három váltakozás volt a célszavakon belül: l-z (bribváglik/brivágozik), sz-d (brivepszik/brivepedik), és z-z (bratárzik/bratározik).

Ezek közül az l-z nem nagyon van a valóságban, illetve van, de nem ikesen (szkvasol-szkvasozik). De be akartam rakni, mert nagyon tetszett. A másik két típus azok aránylag gyakoriak (cselekszik-cselekedik és hullámzik-hullámozik).

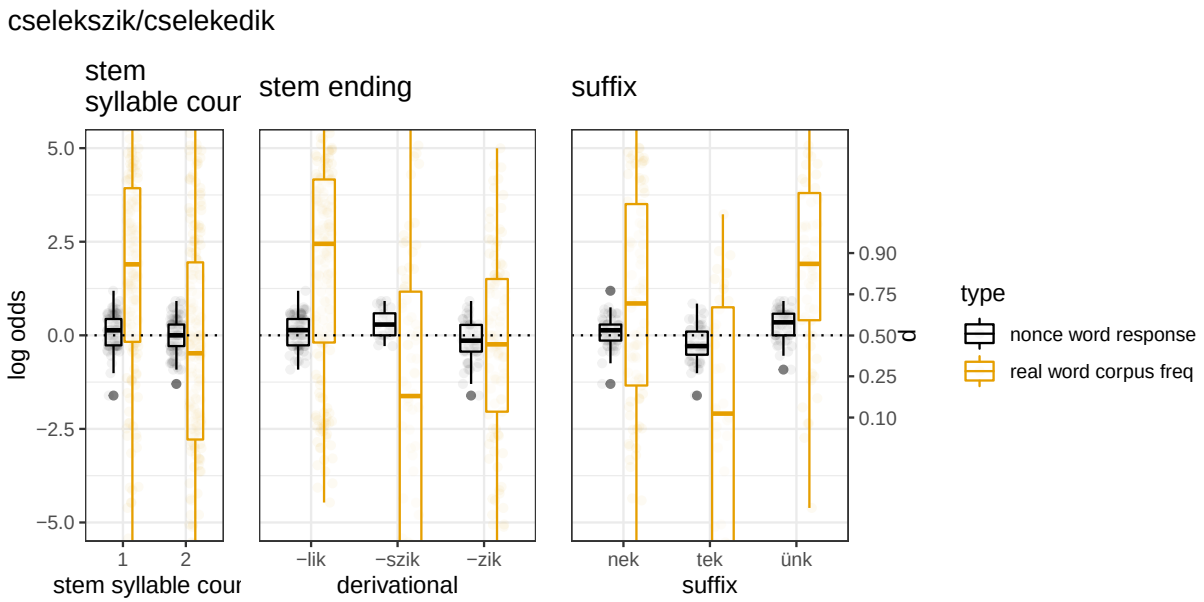
Ezen belül voltak egy és kétszótagos tövek, és ezt három toldalékkal néztem meg, ez a T3 (áramlanak-áramolnak), a T2 (áramlotok-áramoltok), és a T1 (áramlunk-áramolunk). Ezekben az a jó, hogy minden hangkivető típus mindkét alakja létezik velük, ami más, gyakoribb végződésekre nem igaz (ld. ő cselekedne - *ő cselekszene), illetve ketten mássalhangzó (msh), egy pedig mgh kezdetű.

A prompt így nézett ki:

- "János imád prihegedni, ezért sokat prihegszik."
- "Ti/mi/ők ritkán. . . "
- "Prihegedtek-prihegszetek/prihegedünk-prihegszünk/prihegednek-prihegszenek.

A célmondatban mindig a cvc alak van előbb.

Mintázatok



Mit látunk a válaszokban és a korpuszban? A bal panelon azt látjuk, hogy a kétszótagos tövek (pl cselekszik) jóval markánsabban választják a cvc alakot (cselekedik) mint az egyszótagos tövek (hajlik). Akinek van, annak megadatik, ugye, és ezt a mintázatot a kísérletben is szépen hozták a válaszok, ha nem is ilyen erősen.

A középső panelon a tővégződések összehasonlítása annyiban trükkös, hogy a kéklenek-kékeznek típusú, a valóságban azért nem nagyon létező minta egybeesik a lélegzenek-lélegeznek mintával. Talán ezzel is magyarázható, hogy itt a kísérletben a válaszok arányai nem követik a korpuszban talált arányokat.

A jobb panelon az igeragok hatását látjuk. Itt ismét, mondja a Rebrus Péter, az indef-def eszem alakok félrecímkézéséhez hasonló zajforrást jelent az, hogy a T2 cvc alakja egybeesik a múltidejű T3 alakjával, igaz,

csak elölképzett töveknél: melegszetek-melegedtek vs áramlotok-áramoltok-áramoltak. A magyar pro drop nyelv, egy elemzőnek bizonyos esetekben a tágabb szövegkontextusból kéne megmondania, hogy itt most ti melegedtek, vagy ők melegedtek, és a hibalehetőség viszonylag magas. Ráadásul a T3 jóval gyakoribb, mint a T2.

Ennek ellentmond, hogy a kísérletben, ahol a T2 egyértelműsített (ti ritkán...), az emberek kb ugyanazt a mintát mutatják, mint amit a korpuszban látunk: T3 jelen cc-sebb (cselekszenek/cselekednek), mint a T2 jelen (cselekszetek/cselekedtek). A legcc-sebb persze a T1 jelen, ami mgh-ra végződik (cselekszünk/cselekedünk), ezért bármilyen ikes tővel kötőhangzó nélkül is elvan (csuklotok, hanyatlotok), nyilván ez fontos része ennek.

Lexikon

Részben intuitív, részben a próbálkozásaim is alátámasztják azt, hogy a cvc-cc célszavak viselkedését a fenti példákkal szemben nem az határozza meg, hogy milyen vacilláló szavakhoz hasonlítanak, hanem az, hogy milyen stabil CVC és CC igékhez.

Ez szerintem oké: a hotelban és lakom osztályban minden olyan alak variálni fog, amelyre az alaktani leírás ráillik (az ikes igéknél van pár jelölt kivétel, mint a *válom és a *szarom). Ha a leírást vesszük alapul (cvc-ik vagy cc-ikre végződő ige), akkor összességében CVC és CC igékkel kell összehasonlítanunk a célszavaink. Mi több, az egyszerűség kedvéért ki is szűrtem innen azokat a szavakat, amelyek variálnak a webkorpuszban.

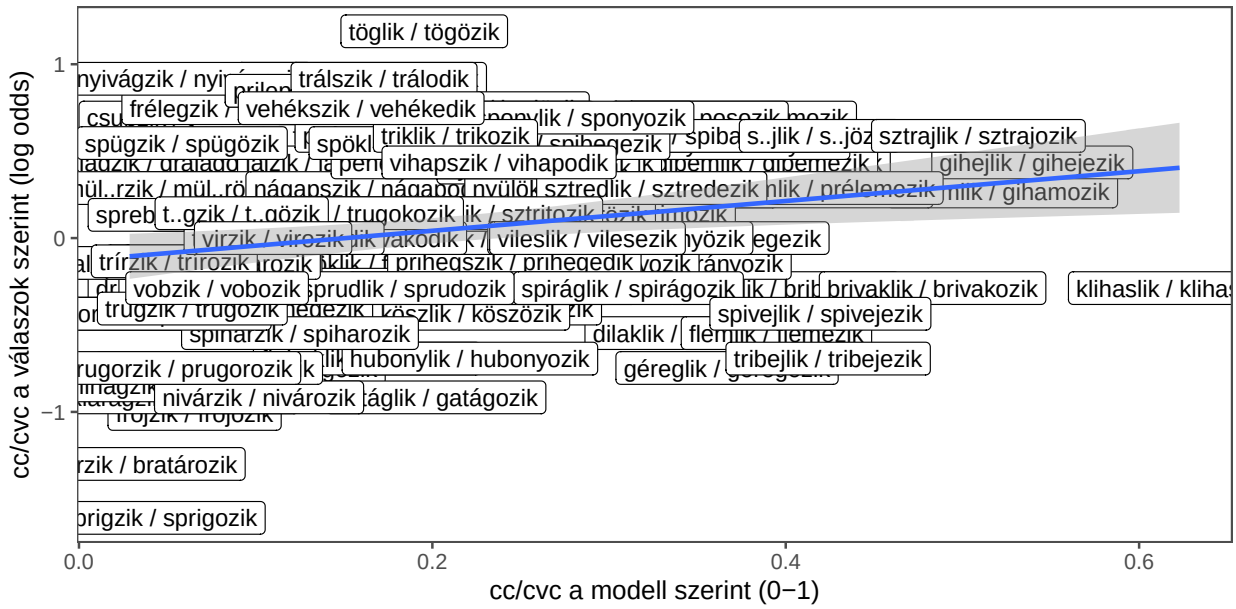
A cvc kategóriában végül tehát ilyen szavak voltak:

```
##  [1] "álmodozik" "ágaskodik" "tanyázik"  "kárhozik"  "dohosodik" "folyik"
##  [7] "bunyózik"  "malmozik"  "hajhászik" "gombázik"  "erősbödik" "őrizkedik"
## [13] "kamuzik"   "ütközik"   "gürizik"   "sokadozik" "költözik"  "rühelődik"
## [19] "kócolódik" "fújódik"   "zsűrizik"  "tétovázik" "hárfázik"  "tusakodik"
## [25] "huzakodik" "bumlizik"  "morzézik"  "nyalizik"  "ötvöződik" "hajcsizik"
## [31] "óvakodik"  "ígérkezik" "sápadozik" "packázik"  "áhítozik"  "kosarazik"
## [37] "aszalódik" "epéskedik" "szolizik"  "adatik"    "hazudik"   "tonik"
## [43] "narkózik"  "kenuzik"   "zónázik"   "szarozik"  "tetőzik"   "öröklődik"
## [49] "tintázik"  "jojózik"
```

A cc kategóriában ilyenek:

```
##  [1] "hanyatlik" "fingik"    "bűzlik"    "eléugrik"  "betegszik" "gyüremlik"
##  [7] "kihajlik"  "látszik"   "toronylik" "vérengzik" "díszlik"   "vedlik"
## [13] "légzik"    "szülemlik" "átsejlik"  "hólyagzik" "fogamzik"  "mosdik"
## [19] "szüremlik" "sejlik"    "alapszik"  "behajlik"  "lehajlik"  "öregszik"
## [25] "parázslik" "fehérlik"  "szólamlik" "játszik"   "porlik"    "homálylik"
## [31] "vöröslik"  "aláhajlik" "ivarzik"   "meghajlik" "korábbik"  "robajlik"
## [37] "türemlik"  "nyaklik"   "rémlik"    "elhajlik"  "aránylik"  "rejlik"
## [43] "morajlik"  "bicsaklik" "hajlik"    "tetszik"   "csetlik"   "pattogzik"
## [49] "örvénylik" "hallszik"
```

A kísérletben adott válaszok korrellálnak azzal, hogy a szó inkább cvc vagy cc szavakra hasonlít, de azért nem olyan rettenetesen:

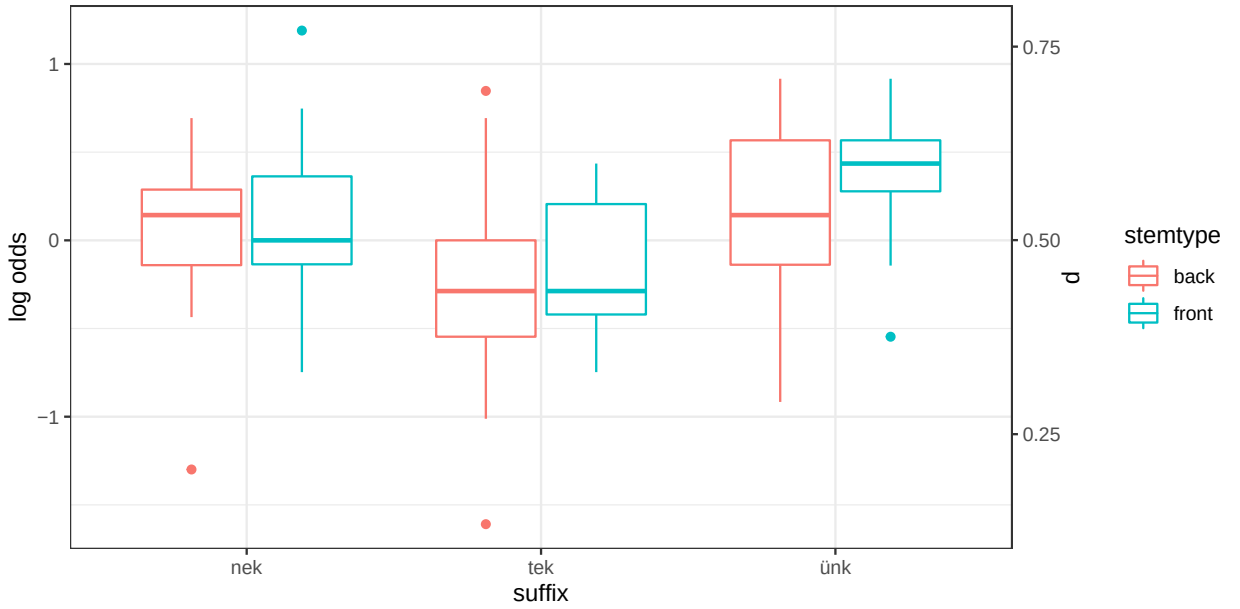


A regressziós modell a leggyengébb a három közül (r^2 = .36).

term	estimate	std.error	statistic	conf.low	conf.high
(Intercept)	0.21	0.16	1.31	-0.10	0.51
cat1_weight	0.76	0.45	1.70	-0.12	1.63
consonantsß/d	0.38	0.11	3.43	0.16	0.59
consonantsz/z	-0.11	0.12	-0.90	-0.34	0.13
nsyl	-0.20	0.06	-3.04	-0.32	-0.07
suffixtek	-0.33	0.08	-4.35	-0.48	-0.18
suffixünk	0.21	0.08	2.84	0.07	0.36

- Az általános hasonlóság fontos valamennyire, még ha a nulla bőven benne is van a 95%-os konfidenciaintervallumban. Ez rezonál Rácz Rebrus Törkenczy eredményére, akik azt találták, hogy a korpuszban vacilláló cvc/cc igéknél a variációt kismértékben jósolja az, hogy az ige inkább fix cvc vagy fix cc igékre hasonlít. Valószínűleg ha kicsit még tekergetném az algoritmust, akkor fel lehetne rugdosni ezt az együtthatót, de a lényeg kb az, hogy van, de nem erős.
- A cselekszik/cselekedik párok egyértelműen a cselekszik alakot preferálják az emlékszik-emlékezik illetve hullámlik-hullámozik típusú párokhoz képest.
- A rövdebb igék inkább cc-sek, tehát alszotok inkább mint aludtok.
- Az áramlunk típusú alakok jóval gyakoribbak, mint az áramolunk típus, ami nem meglepő, viszont érdekes módon az áramoltok típusok népszerűbbnek az áramlotok típusnál.

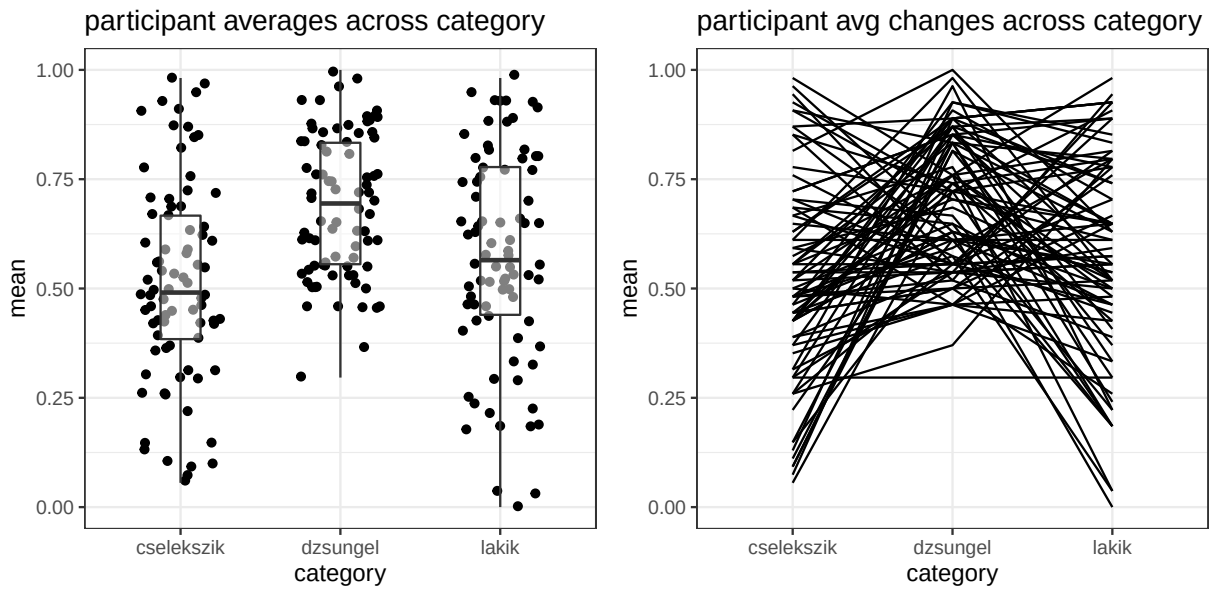
Ez valszeg nem azért van, mert az emberek kerülik az elölképzett cvc T2 alakokat, mert azok egybeesnének a T3 múltidejű alakokkal ("cselekedtek"). Ha megnézzük, hogyan választottak a cc és cvc alak közül annak függvényében, hogy mi volt a toldalék ÉS hogy a tő elöl- vagy hátulképzett volt, nem látunk ilyen eltérést (lenti ábra, középső oszlop):



Mi a tanulság itt összességében? Hát. Ha azt gondoljuk, hogy van egy mögöttes modell, amely meghatározza általában az ikes igék viselkedését, akkor azt mondanám, hogy ennek a hangkivetés tünete, de nem következménye: vannak valamilyen lexikális megszorítások arra, hogy milyen tőben lehet epentetikus mgh és milyenben nem, és vannak morfofonológiai megszorítások arra, hogy milyen toldalékhoz kell epentetikus mgh, és aztán vannak alakok, amelyek ebbe belepusztulnak (mint a csuklik), és van néhány, amelyek instabilok, és bizonyos helyeken mindkét alak lehetséges (ez a dohányozik-dohányzik osztály), és itt totális a káosz abban, hogy melyik alak milyen végződéssel hogyan variálhat (dohányzanak nincs de dohányzunk van, cselekszenek van, de cselekszene nincs, és így tovább). Persze ezt nem neked kell magyaráznom!

Összefoglalás

A résztvevők válaszai elég jól lekövetik a korpuszban talált arányokat. Ez alsó hangon valami olyasmit jelent, hogy a korpuszban látott eloszlás nem teljesen lexikalizált (hogy nem tudom az eszik imád indef-ben is eszem lenni), és nem is a mintát konzisztensen használó emberek rakták össze (ahol az egyik mindik eszek-et ír, a másik mindig eszem-et). Ez részben a fenti eredményekből is látszik, részben pedig abból is, hogy az egyes résztvevők még a részkategoriákon belül sem viselkednek kategorikusan. Ezt mutatja a lenti ábra, ami a szavak helyett az alanyokra számol átlagokat, és ezeket ábrázolja kategoriánként, kétféle módon:



A következő kérdés az, hogy mi bújik meg a kísérletben és a korpuszban tapasztalt eloszlások mögött. Van egy (akár stochasztikus) nyelvi szabály, vagy minden a lexikális hasonlóságon alapul?

- A magánhangzóharmónia (vh) esetén valszeg könnyű kiabsztrahálni a szabályszerűséget: a vh eleve csak mgh-kon működik, ezekből a célszavakban véges kombinációk lehetségesek, a hatások pedig aránylag additívak: e ilyen, é olyan. Ehhez képest minimálisan van jelen az, hogy mennyire elemzik az emberek összetett szónak a kevert töveket (tanév vs nem tudom, malév). Ez utóbbihoz kéne esetleg valamilyen tudás arról, hogy a magyarban a szavak hogyan szoknak kinézni.
- A másik véglet a hangkivető ikes igék esete. Valszeg itt is van egy szabály, ami a magyarban eltűrt msh csoportokról mond valamit, ez keveredik egy hangsúlyosabb lexikális elemmel, hogy ti. bizonyos igék baromira utálják az epentetikus mgh-t (csuklik, hanyatlik), másokban pedig mindig ott van, pedig nélküle is meglenne az ige (vonatozik,vonatoznak). A csavar ebben az, hogy összességében véve a beszélőknek nincs közvetlen tapasztalatuk arról, hogy egy igető hogyan néz ki "valójában", hanem a különféle toldalékolt alakokból próbálják ezt retrofitelni. És ennek az egésznek valójában egy viszonylag kis szelete az, hogy a cselekszik-típusú igék mennyire variálnak – ebben a leosztásban még mindig a defektív igék az igazán érdekesek, csak azok nem fognak variálni persze.
- Az eszek-eszem váltakozás pedig teljesen ortogonális erre az egészre, mert semmilyen strukturális vetülete nincs. Bármilyen igének lehet -k-t vagy -m-et rakni a végére. Az E1 def-indef neutralizációnak annyiban van strukturális része, hogy kevésbé agresszívan tárgyas igéknél kisebb ár az, ha E1-ben a def és az indef egybeesik. De a világ összes nyelve elég jól elvan def-indef megkülönböztetés nélkül, tehát nem hiszem, hogy a magyar ne élné ezt túl. Itt talán az van, hogy valamilyen történeti folyamat kb meghatározta azokat az ikes igéket, amelyek E1-ben mindenhol -m-et szeretnek használni, főleg gyakoriság, kisebb részben pedig valencia és alaktan alapján, a beszélők pedig ehhez a fejben tartott listához hasonlítgatják az álszavakat, amikor el akarják dönteni, hogy ezek "ikesek", vagy sem. (Arról aztán nem is beszélve, hogy a kitöltők szociolingvisztikai profilja korántsem biztos, hogy homogén!)