Delež minimalnih parov besed med besednimi oblikami in lemami

Primož Jakopin

Minimalni pari besed so pari, ki se med seboj razlikujejo samo v enem fonemu (*nika*, *bika*). V prispevku je s pomočjo besedilnega korpusa *Nova beseda* (za besedne oblike) in gesel v viru *Besede slovenskega jezika* (za leme) prikazan delež teh parov glede na sosednje, dve črki oddaljene pare in glede na vse možne pare enako dolgih besed. Izkaže se, da delež minimalnih parov glede na sosednje pare raste z dolžino in da je bistveno večji pri besednih oblikah kot pri lemah.

The Share of Minimal Pairs for Word Forms and Lemmas

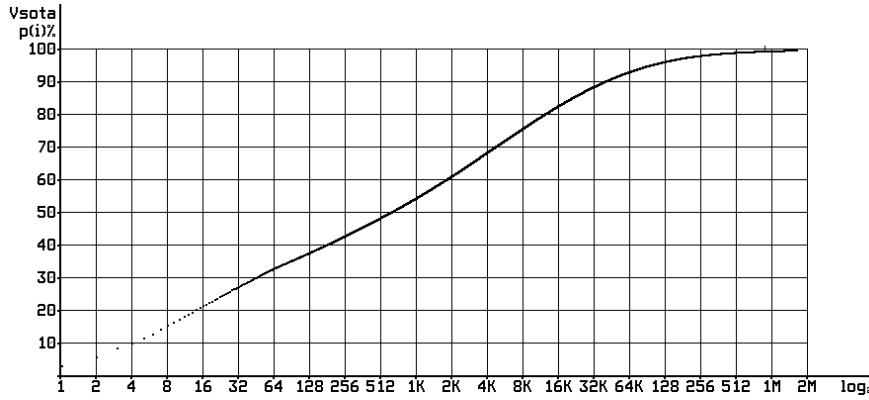
Minimal pairs differ by only a single phoneme (e.g., *pear/bear*). This article uses words from the index of the text corpus *Nova beseda* (New Word; 240 million running words) and lemmas from the web resource *Besede slovenskega jezika* (Slovenian Words; 356,000 entries) to calculate the share of minimal pairs with regard to near-minimal pairs in which words differ by two letters, and among all possible word pairs of equal length. The share increases with word length and is also significantly greater for word forms than for lemmas.

1 Uvod

Pri ugotavljanju pomenskorazločevalnih enot (fonemov) v jezikoslovju in z njimi povezanih raziskavah (npr. Orešnik 2008) imajo pomembno vlogo t. i. minimalni pari besed. To so pari besed, ki se med seboj razlikujejo samo v enem fonemu, primer je npr. par (*nika*, *bika*). Namen prispevka je osvetliti njihov delež glede na sosednje, dve črki oddaljene besedne pare in vse možne pare enako dolgih besed, delež tako med besednimi oblikami kot tudi med besednimi lemami. Ker ustrezno velikega fonemsko zapisanega vira za slovenski jezik še ni na razpolago, sta bila za odgovor na hipotetično vprašanje s programom EVA, orodjem za obdelavo jezikovnih virov (Jakopin 1995), obdelana dva besedna vira: indeks besedilnega korpusa *Nova beseda* (Jakopin – Michelizza 2009) ter gesla v viru *Besede slovenskega jezika* (Gložančev idr. 2009), oba si je mogoče ogledati na spletnem naslovu http://bos.zrc-sazu.si/.

2 Gradivo

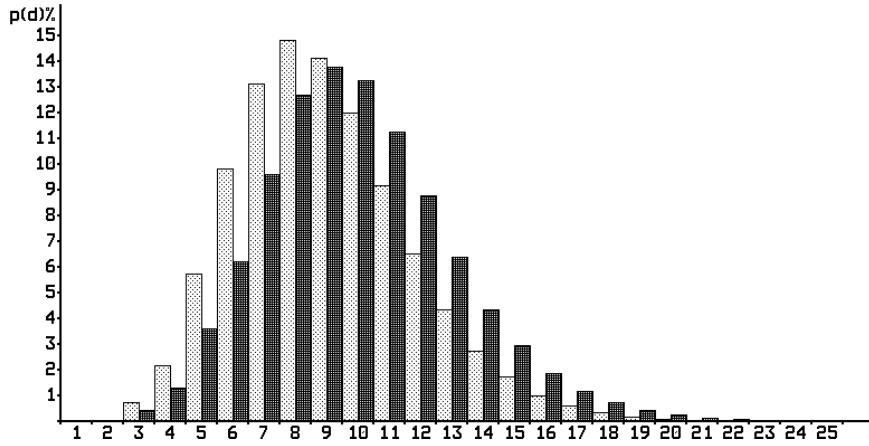
V obeh že v uvodu omenjenih besednih virih je bilo potrebno pred obdelavo opraviti ustrezen izbor. Odločiti se je treba za spodnjo in zgornjo mejo dolžine, do katere bi opazovali odnos med minimalnimi pari in med dve črki oddaljenimi pari. Spodnja meja se ponuja kar sama od sebe, to je dolžina treh črk, zgornja meja pa zahteva nekaj več premisleka. Po drugi strani pa je smiselno oba seznama omejiti glede na sestavo. Predvsem prvi ne vsebuje samo besed v običajnem pomenu, *jezikovnih enot iz glasov za označevanje pojmov* (SSKJ 1), ampak tudi nebesedne enote (Jakopin 2001), ki jih je posebno veliko med daljšimi enotami v indeksu. Tako je v njem med 6113 enotami z dolžino vsaj 30 znakov, najdaljša je dolga 249 znakov, le 61 takih, ki so sestavljene samo iz črk. Prevladujejo spletni in elektronski naslovi, skupaj jih je 4332, na osmem mestu je prvo število, 134 znakov dolgi googol, s katerim sta si pomagala Larry Page in Sergej Brin, ko sta iskala ime za svoj zdaj vodilni iskalnik, najdalješa prava beseda, na 859. mestu, je vrstilni številnik *šestmilijontidvestotriindvajsettisočtristodvaintrideseti*, dolg 56 znakov, prvi trije samostalniki, dolgi 32, 31 in 30 črk: *prapraprapraprapraprapravnukinja*, *klavstrofilofoboksenofilofobija* in *psihonevroendokrinoimunologija* so pa že bolj na repu te skupine.



Slika 1: Krivulja rasti za enote v indeksu *Nove besede*

Da bi bili rezultati bolj značilni za slovenski jezik, so bile upoštevane le enote v indeksu, sestavljene samo iz črk in s frekvenco vsaj 5, merilu, ki ga je, sicer za angleški jezik, predlagal Sinclair (1991); pri drugem viru pa le gesla iz črk. S slike 1 je razvidno, da najpogostejše 4 besedne oblike v besedilih (*je*, *v*, *in* in *na*) skupaj pokrijejo 10 % celote, najpogostejših 500 skupaj približno polovico korpusa, za 75-odstotno pokritost jih je potrebnih že 8000, za 90-odstotno pa dobrih 32.000. Omejitev na pogostnost 5 sicer res odreže proč dve tretjini bolj eksotičnih enot, ki pa pokrijejo le približno 0,75 % korpusa. Enkratnic, besednih oblik, ki se v korpusu pojavijo samo enkrat (angl. *hapax legomena*), je namreč 783.000, to je skoraj polovica (46,5 %) različnih enot. Za izbor zgornje meje dolžine, do katere bi opazovali

obnašanje deleža minimalnih parov si je vredno ogledati porazdelitev dolžin besednih enot v obeh virih, ki je prikazana na sliki 2.



Slika 2: Porazdelitev dolžin besed iz *Nove besede* in *Besed slovenskega jezika*

Vrednosti za besedne oblike iz indeksa Nove besede označene svetlosivo, za gesla iz seznama *Besede slovenskega jezika* pa temnosivo. Prve dosežejo vrh pri dolžini 8 črk, druge pri 9, in tudi upadanje proti večjim dolžinam je pri lemah dosti počasnejše. Avtor se je glede na prikazano odločil zgornjo mejo opazovane dolžine postaviti pri 17.

Preglednica 1: Obseg prvega vira, besednih oblik iz indeksa *Nove besede*

	Različnih	**Vseh**
Celoten indeks	1.684.465	239.786.693
Frekvenca vsaj 5	510.007	237.976.732
Samo enote iz črk	466.556	232.417.205
Dolžina 3–17	463.876	166.629.956

Iz zadnje vrednosti drugega stolpca je razviden velik delež oblik z dolžino 2. Že najpogostejših 12: *je*, *in*, *na*, *da*, *za*, *se*, *ki*, *so*, *pa*, *ne*, *bi* in *po* ima vsoto pogostnosti prek 40 milijonov.

Drugi vir, gesla iz seznama *Besede slovenskega jezika*, je bolj v skladu s pričakovanji, najdaljša beseda v njem je že videni števnik, sledita samostalnika *dvaalfahidroksibencilbenzimidazol* in *klavstrofilofoboksenofilofobija*, na naslednjih mestih pa sta prislov *primerjalnoliterarnozgodovinsko* ter pridevnik *filozofskoliterarnozgodovinski*. Pot do gradiva za raziskavo je v tem primeru krajša: vseh gesel je 356.912, ko upoštevamo le različna gesla iz črk, jih ostane 352.242, po dolžinski omejitvi na 3–17 pa 345.339.

3 Delež minimalnih parov

Za izračun tega podatka je treba najprej vedeti, koliko je vseh možnih besednih parov. Vzemimo za pomoč pri izpeljavi najpogostejše besedne oblike iz *Nove besede*, ki so dolge 5 črk: *lahko*, *nekaj*, *sicer*, *proti*, *potem*, *drugi* in *treba*. Če sta besedi dve, je možen en par: (*lahko*, *nekaj*). Če so besede 3, so pari trije: (*lahko*, *nekaj*), (*lahko*, *sicer*) in (*nekaj*, *sicer*). 4 besede dajo 6 parov, 5 besed 10, 6 besed 15 in 7 besed 21 parov: (*lahko*, *nekaj*), (*lahko*, *sicer*), (*lahko*, *proti*), (*lahko*, *potem*), (*lahko*, *drugi*), (*lahko*, *treba*), (*nekaj*, *sicer*), (*nekaj*, *proti*), (*nekaj*, *potem*), (*nekaj*, *drugi*), (*nekaj*, *treba*), (*sicer*, *proti*), (*sicer*, *potem*), (*sicer*, *drugi*), (*sicer*, *treba*), (*proti*, *potem*), (*proti*, *drugi*), (*proti*, *treba*), (*potem*, *drugi*), (*potem*, *treba*) in (*drugi*, *treba*). Gre za kombinacije (reda r med n elementi) brez ponavljanja (npr. Jamnik 1994: 241), v matematiki navadno označene kot

$$C(n, r) = n(n-1)(n-2) \ldots (n-r+1) = \frac{n!}{r(n-r)!} \quad (1)$$

V našem primeru je red *r* enak 2 in zveza se močno poenostavi:

$$C(n, 2) = \frac{n}{n!(n-2)!} = \frac{n(n-1)}{2} \quad (2)$$

Število besed v obeh opazovanih virih ni majhno, število možnih parov pa seveda zvezi (2) ustrezno večje. Pred desetletjem ali dvema bi ugotavljanje števila minimalnih parov in števila parov besed, ki se razlikujejo za dve črki za tehnologijo tistega časa predstavljalo znaten napor, danes pa je problem rešljiv v nekaj minutah procesorskega časa. Dobljene vrednosti so navedene v preglednici 2.

Preglednica 2: Pari glede na dolžino pri besednih oblikah iz *Nove besede*

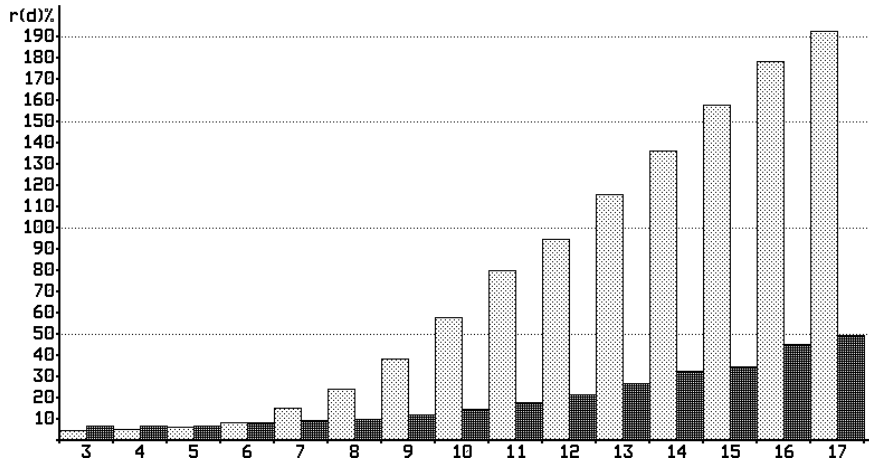
Dolžina	**n**	**Vseh parov**	**Minimalnih parov**	**Parov z razdaljo 2**
3	6.054	18.322.431	106.105	2.211.662
4	14.156	100.189.090	126.958	2.276.598
5	33.227	552.000.151	137.808	2.174.036
6	51.580	1.330.222.410	110.650	1.270.261
7	65.326	2.133.710.475	84.453	552.932
8	71.575	2.561.454.525	75.376	309.147
9	65.527	2.146.861.101	58.903	151.417
10	53.424	1.427.035.176	42.454	73.224
11	39.086	763.838.155	28.311	35.316
12	26.615	354.165.805	17.522	18.460
13	16.860	142.121.370	10.606	9.156
14	10.004	50.035.006	5.984	4.389
15	5.791	16.764.945	3.198	2.020
16	2.988	4.462.578	1.566	876
17	1.663	1.381.953	837	434
Skupaj	463.876	11.602.565.171	810.731	9.089.928

Po pričakovanju so deleži minimalnih parov in njihovih sosedov večji pri krajših dolžinah in potem padajo, skupaj je delež minimalnih parov glede na celoto (810.731 od 11.602.565.171) zaokroženo 0,00007 ali 0,07 ‰, največji, 7 ‰, je pri dolžini 3, najmanjši, 0,027 ‰, pa pri dolžini 9. Zanimiv je tudi odnos med minimalnimi pari in njihovimi sosedi, glede na dolžino. Če upoštevamo vse dolžine, je število minimalnih parov približno 9 % števila parov z razdaljo 2 ali enajstkrat manj. Pri parih kratkih besednih oblik je minimalnih parov v primerjavi s pari z razdaljo 2 malo, približno 5 % števila, potem pa se razmerje spreminja in pri dolžini 14 je minimalnih parov že več, pri dolžini 17 skoraj dvakrat več.

Preglednica 3: Pari glede na dolžino pri geslih v seznamu *Besede slovenskega jezika*

Dolžina	**n**	**Vseh parov**	**Minimalnih parov**	**Parov z razdaljo 2**
3	1.566	1.225.395	15.176	207.873
4	4.606	10.605.315	22.506	317.459
5	12.760	81.402.420	38.065	528.196
6	21.848	238.656.628	39.376	444.021
7	33.693	567.592.278	41.380	419.468
8	44.586	993.933.405	38.341	364.312
9	48.416	1.172.030.320	24.670	200.024
10	46.507	1.081.427.271	14.222	94.913
11	39.469	778.881.246	7.542	41.499
12	30.837	475.444.866	3.997	18.417
13	22.540	254.014.530	1.868	6.850
14	15.364	118.018.566	896	2.742
15	10.461	54.711.030	460	1.307
16	6.570	21.579.165	235	519
17	4.268	9.105.778	131	263
Skupaj	343.491	5.858.628.213	248.865	2.647.863

Pri geslih iz seznama *Besede slovenskega jezika*, kjer izpeljane besedne oblike ne nastopajo in kjer tudi ni imen, je minimalnih parov manj. Skupaj je delež minimalnih parov glede na celoto (248.865 od 5.858.628.213) zaokroženo 0,00004 ali 0,04 ‰. Največji, 12 ‰, je pri dolžini 3, najmanjši, 0,007 ‰, pa pri dolžini 13. Odnos med minimalnimi pari in njihovimi sosedi je zelo primerljiv: skupaj je prvih glede na druge spet približno 9 % ali enajstkrat manj. Pri nobeni dolžini število minimalnih parov ne preseže števila sosednjih parov, res pa je, da razmerje praktično monotono narašča, od 7 % pri dolžini 3 do 50 % pri dolžini 17.



Slika 3: Razmerje med minimalnimi pari in pari z razdaljo 2 pri besednih oblikah *Nove besede* in geslih *Besed slovenskega jezika*

Bolj nazorno je odnos med minimalnimi pari in pari z razdaljo 2 glede na dolžino besed razviden s slike 3. Prvi vir je označen s svetlosivo, drugi pa s temnosivo barvo.

4 Sklep

Jezik, besede in črke v njem, zabeležene v pisanem sporočilu, bi se komu, ki bi uporabljal drugačen način komunikacije, morda na drugi strani Hubblovega obzorja, le zelo na hitro in od daleč zdeli kot zaporedje naključno nabranih in s presledki razmejenih nizov črk in ločil. Že njihove pogostnosti razkrijejo nekaj osnovnih zakonitosti, množica pravil, ki se jo da razbrati iz njihovih odnosov, pa kaj kmalu preraste okvirje, ki smo jih vajeni pri opisu procesov v naravoslovnih znanostih.

Tako tudi v prispevku ugotovljeni nelinearen in nemonoton odnos med minimalnimi pari besed in pari, ki se razlikujejo v dveh črkah, odpira nova vprašanja za empirični premislek in pojasnitev.

Viri in literatura

Gložančev idr. 2009 = Alenka Gložančev idr. 2009, *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri)*, Ljubljana: Založba ZRC, 2009.

Jakopin 1995 = Primož Jakopin, EVA – a Textual Data Processing Tool, *TELRI Newsletter* 2, December 1995, 13.

Jakopin 2001 = Primož Jakopin, Words and nonwords as basic units of a newspaper text corpus, *COMPLEX 2001 / 6th Conference on Computational Lexicography and Corpus Research »Computational Lexicography and New EU Languages«*, University of Birmingham, 49–65

Jakopin – Michelizza 2009 = Primož Jakopin – Mija Michelizza, Besedilni korpus Nova beseda, *Mostovi* 41 (2007/08), št. 1–2, 165–176.

Orešnik 2008 = Janez Orešnik, Natural syntax: English reported speech, *Studia Anglica Posnaniensia* 44 (2008), 218–252.

Sinclair 1991 = John Sinclair, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press, 1991.

SSKJ 1 = *Slovar slovenskega knjižnega jezika* 1, Ljubljana: DZS, 1970.

The Share of Minimal Pairs for Word Forms and Lemmas

Summary

This article investigates the shares of minimal pairs (pairs of words that differ only in a single phoneme such as *nika/bika*) among near-minimal pairs, in which words differ by two letters and among all possible word pairs of equal length. Because no suitable language resource with phonemes in lemmas and word forms is available for Slovenian, two resources for the written language were used: the index of the text corpus *Nova beseda* (New Word; 240 million running words, 500,000 different words) and lemmas from the web resource *Besede slovenskega jezika* (Slovenian Words; 356,000 entries). They are both available at bos.zrc-sazu.si/index_en.html. The EVA language resource tool (http://www.laze.org/eva) was used for processing. The number of all possible equal-length word pairs is large but manageable: 12 billion for word forms and 6 billion for lemmas.

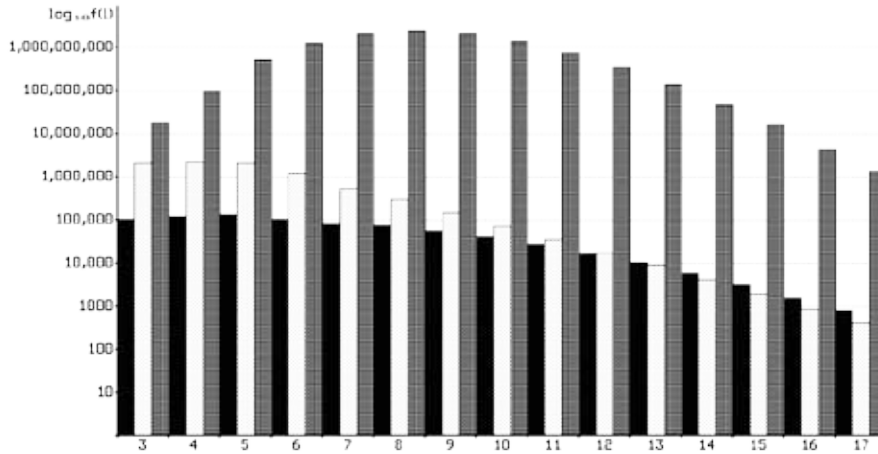


Figure 4: Frequencies of minimal word pairs (black), neighbouring word pairs (light grey) and all word pairs as related to word length for wordforms in Nova beseda

As can be concluded from Figure 4, the share of minimal pairs among all word pairs and among near-minimal pairs increases with word length. It is also worth noting that the number of minimal pairs is smaller by an order of magnitude than the number of near-minimal pairs that differ by two letters, for word lengths from three to five letters. For word lengths from six letters onwards, the difference between these two numbers steadily decreases, whereas with a word length of 13 letters or more the number of minimal pairs is even greater than the number of near-minimal pairs.

As could be expected, the share of minimal pairs is also substantially greater for word forms when compared to the share for lemmas.