UDK 808.63-22
Primož Jakopin
Filozofska fakulteta v Ljubljani

ALI SO ROJSTNA IMENA KRAJŠA OD DRUGIH SAMOSTALNIKOV?

Gre za primerjavo dveh vzorcev – 3.661 rojstnih imen iz slovenskega telefonskega imenika 1995 (mesto Ljubljana) in 51.448 samostalnikov iz Slovarja slovenskega knjižnega jezika (izdaja 1994: 21.823 m. sp., 21.427 ž. sp. in 8.198 srednjega spola). Povprečna dolžina pri imenih znaša 5.88 črke, pri občnih samostalnikih pa 8.49 črke in izkaže se, da so rojstna imena statistično zelo značilno krajša kot občni samostalniki. – Navedena je še primerjava dolžin samostalnikov po spolu (m povprečno 7.63, ž 8.95, s 9.59 črke) in tudi tu so pri vseh parih (m-ž, m-s in ž-s) razlike statistično značilne.

A comparison of two samples is presented: 3.661 first names from the Slovenian phonebook 1995 (Ljubljana area) and 51.448 nouns from the Dictionary of the Slovenian Literary Language (1994: 21.823 of masculine gender, 21.427 of f. g. and 8.198 of n.g.). The average Christian name length comes to 5.88 letters, compared to 8.49 letters with common nouns. The outcome is favourable: first names are significantly shorter than common nouns. – Noun lengths have also been compared by gender (average values: m 7.63, f 8.95 and n 9.59). Here again the differences within all pairs (m-f, m-n and f-n) are statistically significant.

Da bi naslovno vprašanje lahko potrdili ali ovrgli, si oglejmo podatke iz slovenskega telefonskega imenika 1995 (STI 95) in iz Slovarja slovenskega knjižnega jezika (SSKJ 95). Telefonski imenik Slovenije je uporabnikom že nekaj let na razpolago tudi v elektronski obliki in je podatke iz njega mogoče, po majhnih koščkih (za največ 500 naročnikov naenkrat) tudi prenesti v lastno datoteko, nato pa jo naprej obdelovati. Teoretično je mogoče tudi ves imenik odnesti v t.i. lastni telefonski imenik. To je datoteka, ki je sicer namenjena vašim pogosto uporabljanim številkam in spada v okvir programa za iskanje po imeniku, je pa njena zgradba dovolj preprosta, da jo je mogoče uporabiti za naš namen. Da bi bili podatki v telefonskem imeniku čim bolj varni pred kakim načinom uporabe, ki ga sestavljavci programa ne bi želeli, je prenos v lastni imenik narejen tako, da z rastjo tega imenika poteka čedalje počasneje. Poskus, da bi tako prenesli cel imenik, ni uspel, je pa razmeroma zelo sodoben in hiter osebni računalnik zmogel čez konec tedna, od petka popoldne do ponedeljka dopoldne, prenesti 89.823 telefonskih naročnikov, vseh iz Ljubljane. Imeli so 4.560 imen, med katerimi je po razbitju dvojnih imen, npr. *Brigita – Marija* ostalo 3.661 različnih, s skupno pogostnostjo 90.018. SSKJ obstaja v elektronski obliki že nekaj časa (Jakopin 95) in bo širši javnosti predvidoma dostopen še v letošnjem letu.

Poglejmo si na kratko oba vzorca:

Preglednica 1: Število imen in število drugih samostalnikov

3.661	različnih rojstnih imen
90.018	vseh rojstnih imen, pripadajočih 89.823 osebam
51.448	različnih drugih samostalnikov, od tega

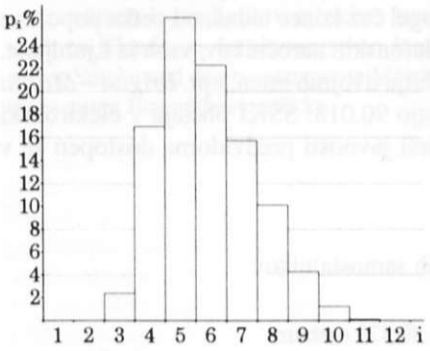
21.823 moškega spola,
21.427 ženskega in
8.198 srednjega spola

Preglednica 2: Črkovne dolžine pri imenih in pri občnih samostalnikih

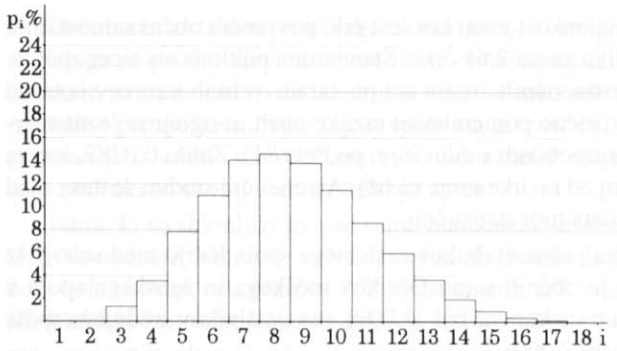
Dolžina	Imen	%	Občnih samost.	%
1	–	–	29	0.06
2	2	0.05	29	0.06
3	88	2.40	680	1.32
4	622	16.99	1826	3.55
5	968	26.44	4041	7.85
6	817	22.32	5676	11.03
7	577	15.76	7261	14.11
8	372	10.16	7523	14.62
9	159	4.34	7126	13.85
10	47	1.28	5920	11.51
11	5	0.14	4450	8.65
12	2	0.05	3069	5.97
13	2	0.05	1901	3.69
14	–	–	1009	1.96
15	–	–	510	0.99
16	–	–	221	0.43
17	–	–	107	0.21
18	–	–	42	0.08
19	–	–	17	0.03
20	–	–	7	0.01
21	–	–	1	–
22	–	–	1	–
24	–	–	2	–

Imena z dolžinami ne sežejo čez 13 črk, vrh 26.44% dosežejo pri dolžini 5 črk, pogostnost pa strmo pade pri dolžini 11. Najdaljši občni samostalnik je skoraj še enkrat daljši, 24 črk, vrh s 14.62% je šele pri 8, občutnejši padec je pa že skoraj na koncu, pri dolžini 18. Najdaljša štiri imena so *Christodoulos, Maksimilijana, Maksimilijan* in *Maksimiljana,* najdaljši trije samostalniki pa *stárocerkvénoslovánščina, vsèzavérodómcesárjevstvo* in *pétinsédemdesetlétnica.*

Še zgovornejše kot preglednica so slike. Na prvih dveh je porazdelitev imen in občnih samostalnikov, za vsake posebej (s p_i je označena pogostnost):

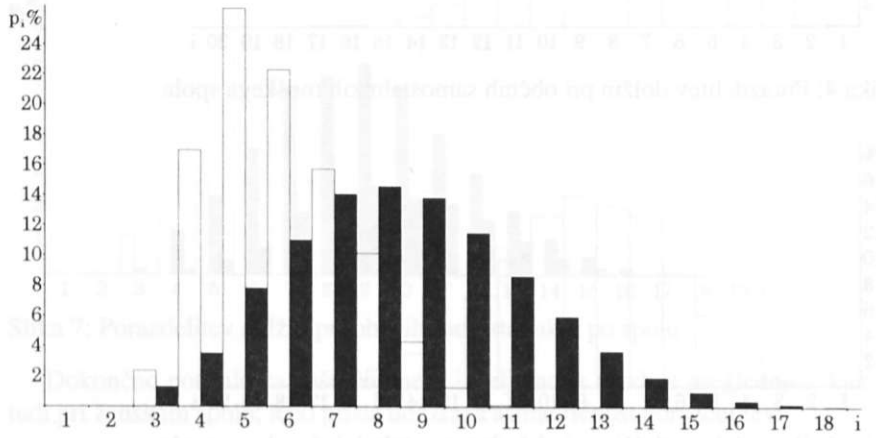


Slika 1: Črkovne dolžine imen



Slika 2: Črkovne dolžine občnih samostalnikov

Zgradba obeh porazdelitev je precej različna: pri imenih se skokovito dviguje proti vrhu (doseže ga že pri dolžini 5), od tam naprej pa sicer še vedno hitro, vendar manj odločno pada proti koncu. Pri občnih samostalnikih je krivulja dosti bolj umirjena: vrh pri 8 je kopast in asimetričnost porazdelitve v prid desni strani je manj izrazita. Zamik vrhov še lepše vidimo na sliki 3, kjer sta obe porazdelitvi na istem histogramu: stolpci imen so bele barve, samostalnikov pa črne:



Slika 3: Porazdelitev dolžin pri imenih in pri občnih samostalnikih

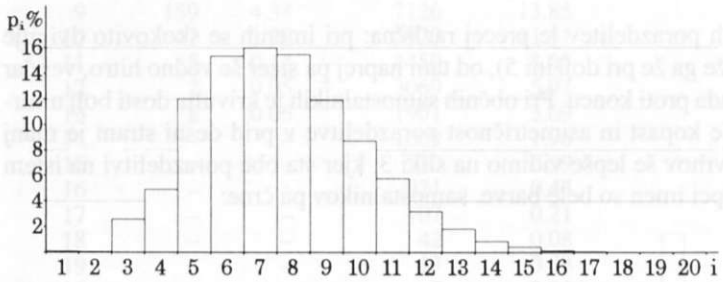
Da bi lahko ugotovili, ali je razlika med dolžinami imen in občnih samostalnikov tudi statistično pomembna, potrebujemo še osnovni statistični opis obeh spremenljivk:

Preglednica 3: Statistični opis dolžin imen in občnih samostalnikov

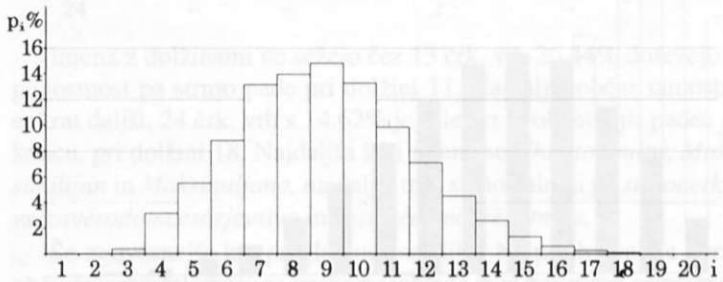
	Imena	Občni samost.
Število	3661	51448
Najmanjša vrednost	2	1
Največja vrednost	13	24
Aritmetična sredina	5.88	8.49
Standardni odklon	1.55	2.65
Standardna napaka aritmetične sredine	0.0256	0.0117

Ime ima v povprečju malenkost manj kot šest črk, povprečni občni samostalnik pa skoraj osem in pol; razlika znaša 2.61 črke. Standardna odklona sta sicer znatna, standardni napaki obeh aritmetičnih sredin sta pa zaradi velikih vzorcev izredno majhni. Da bi preverili statistično pomembnost razlike obeh, si oglejmo še standardno napako razlike obeh aritmetičnih sredin (npr. po Petz 81). Znaša 0.0282, kar za dva velikostna razreda manj od razlike same (2.61). Aritmetični sredini se torej med seboj zelo značilno razlikujeta tudi statistično.

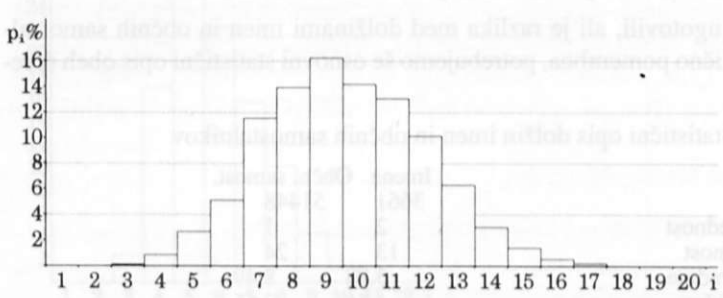
Ali se tudi dolžine občnih samostalnikov različnega spola ločijo med seboj? Iz preglednice 1 vidimo, da je občnih samostalnikov moškega in ženskega spola v SSKJ približno enako veliko (nekaj več kot 21.000), samostalnikov srednjega spola pa skoraj trikrat manj (dobrih 8.000).



Slika 4: Porazdelitev dolžin pri občnih samostalnikih moškega spola



Slika 5: Porazdelitev dolžin pri občnih samostalnikih ženskega spola

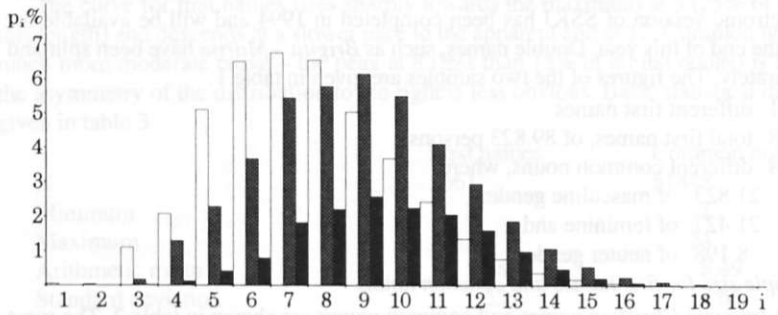


Slika 6: Porazdelitev dolžin pri občnih samostalnikih srednjega spola

Občni samostalniki moškega spola imajo vrh pri dolžini 7, ženskega in srednjega pa pri 9. Sklepamo torej, da so občni samostalniki m. sp. krajši kakor pri ženskem in srednjem spolu (slednji skupini se med seboj ne razlikujeta dosti). Porazdelitvi pri moškem in srednjem spolu sta asimetrični v desno, pri ženskem pa – presenetljivo – rahlo v levo. Zanimiv je tudi začetek diagramov: tri črke dolgi občni samostalniki so skoraj vedno moškega spola.

Takih, ki so sklonljivi in niso samo množinski, je srednjega spola samo 11: *dnò, igó, ílo, imé, ojé, okó, onè, psè, tlò, uhó* in *zlò*, ženskega pa 88: *ára, bél, bér, bíl, bít, bóa, ból, brv, cép, cév, čér, čŕn, čúd, êra, éta, gáz, gól, gós, góž, hčí, íba, íca, íva, jéd, jél, kàd, kál, káp, káp, kóp, kóv, krí, lás, lát, láž, lév, lóv, lúč, máz, méd, mél, mér, mìš, móa, móč, nát, nìt, nóč, óda, óka, óma, ôsa, óst, péč, péč, péd, póč, pót, rál, ráz, réč, réd, réz, réž, rìt, rja, róv, séč, séč, sév, slà, sól, srt, srž, súš, úra, úta, vás, véz, vós, vóž, vrv, zél, zév, zób, žál, žív* in *žrd*. Zanimivo je, da je med slednjimi mnogo takih, ki so pogosteje moškega spola in jih kot take tudi bolj poznamo, npr. gož, kop, las.

Pred preglednico z aritmetičnimi sredinami in standardnimi napakami za vse tri skupine si oglejmo še skupni diagram, kjer so navedene dolžine vseh treh spolov. Stolpci samostalnikov moškega spola so bele barve, ženskega so pisani, srednjega črni, odstotki pa se nanašajo na vse samostalnike skupaj:



Slika 7: Porazdelitev dolžin pri občnih samostalnikih po spolu

Dokončno potrdilo za naše domneve izpeljemo iz spodnje preglednice, kjer je, tudi pri ženskem spolu, lepo prišla do izraza asimetričnost porazdelitev:

Preglednica 4: Statistični opis dolžin občnih samostalnikov

	Moški spol	Ženski spol	Srednji spol
Število	21823	21427	8198
Najmanjša vrednost	1	2	3
Največja vrednost	21	24	24
Aritmetična sredina	7.63	8.95	9.59
Standardni odklon	2.46	2.66	2.43
Standardna napaka aritmetične sredine	0.0166	0.0182	0.0268

Razlike med aritmetičnimi sredinami dolžin za posamezne pare spolov so naslednje: moški samostalniki so povprečno 1.32 črke krajši od ženskih in 1.96 črke od srednjih, ženski od srednjih pa 0.64 črke. Ustrezne standardne napake razlik med aritmetičnimi sredinami posameznih parov so zaradi velikih vzorcev spet zelo maj-

hne: 0.0246 za par *m* – *ž*, 0.0316 za par *m* – *s* in 0.0324 za par *ž* – *s*. Iz navedenega lahko spet sklepamo, da so samostalniki moškega spola statistično zelo značilno krajši od samostalnikov ženskega in srednjega spola, ženskega pa tudi od tistih srednjega spola.

LITERATURA

Slovar slovenskega knjižnega jezika, 1994. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Državna založba Slovenije.

Telefonski imenik Republike Slovenije za leto 1995, 1995. Ljubljana: PTT Slovenije.

P. JAKOPIN, 1995: Nekaj števílk iz Slovarja slovenskega knjižnega jezika. *Slavistična revija* 43/3, 341–375.

B. PETZ, 1981: *Osnovne statističke metode za nematematičare*. Zagreb: Sveučilišna naklada Liber.

SUMMARY

To give an answer, probably an affirmative one, to the title question, two word samples are evaluated in the paper: first names from the Slovenian phonebook 1995 and nouns from the Dictionary of the Slovenian Literary Language (SSKJ, 1970–1991). The electronic version of the phonebook has been used (on CD ROM), with sample limited to Ljubljana area; the electronic version of SSKJ has been completed in 1994 and will be available for wide use by the end of this year. Double names, such as *Brigita – Marija* have been split and counted separately. The figures of the two samples are given in table 1:

3.661	different first names
90.018	total first names, of 89.823 persons
51.448	different common nouns, where
21.823	of masculine gender,
21.427	of feminine and
8.198	of neuter gender

Table 1: Sample size for first names and common nouns

The most frequent Christian names and common names are shown in table 2. The most popular name in the phonebook (and elsewhere) is *Marija*, followed by *Jože*, *Franc* and others; the top list is obviously male-biased, as could be expected, but in total the sample can be considered as quite complete. The order of common nouns, from *jêzik* to *okó* and *rôka* is very approximate – as no general frequency dictionary for words in Slovenian is available so far, the length of dictionary entry has been taken as the indirect measure of word frequency:

First names			Common nouns		
Marija *f*	Dušan *m*	Jožica *f*	jêzik *m*	réd *m*	móč *f*
Jože *m*	Peter *m*	Ivanka *f*	okó *n*	svét *m*	strán *f*
Franc *m*	Stane *m*	Vinko *m*	rôka *f*	úra *f*	šóla *f*
Janez *m*	Ana *f*	Majda *f*	gláva *f*	srcé *n*	sónce *n*
Anton *m*	Marko *m*	Slavko *m*	beséda *f*	síla *f*	sistém *m*
Ivan *m*	Bojan *m*	Irena *f*	léto *n*	vŕsta *f*	poglèd *m*
Marjan *m*	Boris *m*	Igor *m*	življênje *n*	mésto *n*	uhó *n*
Milan *m*	Drago *m*	Tatjana *f*	pót *f*	pravíca *f*	položáj *m*
Andrej *m*	Anica *f*	Nada *f*	kônec *m*	tóčka *f*	zvéza *f*
Alojz *m*	Branko *m*	Vladimir *m*	čàs *m*	vôda *f*	nôga *f*

Table 2: The most frequent first names and common nouns

The only obvious foreign word in the first 30 common nouns is *sistem* on position 25.

A better insight into lengths of words in both samples gives figure 1, where both distributions, of clearly different character, can be seen.

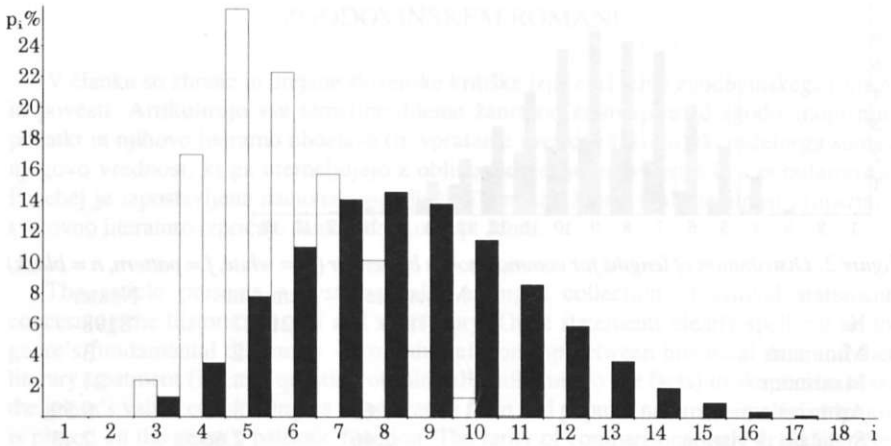


Figure 1: Distribution of lengths for first names (white) and common nouns (black)

The curve for first names rises sharply towards the maximum at 5 (25% of names have this length) and descends at a slower pace to the apparent end at 12. Common nouns show a much more moderate curve – the peak at 8 (less than 15% of all the nouns) is rounder and the asymmetry of the distribution to the right is less obvious. Basic statistical description is given in table 3.

	First names	Common nouns
N	3661	51448
Minimum	2	1
Maximum	13	24
Arithmetic mean	5.88	8.49
Standard deviation	1.55	2.65
Standard error of arithmetic mean	0.0256	0.0117

Table 3: Statistical description of lengths for names and nouns

An average first name is slightly shorter than 6 letters, while an average common noun is 2.5 letters longer. As both standard errors of arithmetic mean are two orders of magnitude smaller than the difference between both average lengths it can be concluded that first names are statistically significantly shorter than other nouns.

Another question which could also be answered using the above data and which comes as a natural extension of the problem is: *Do the lengths of nouns differ by gender?* The histograms in figure 2 support such assumption (the probabilities p_i in percents show the share of the total noun number – 51.448 and not of the respective totals of nouns by gender).

Masculine nouns peak at the length of 7 letters (with about 7% of all the nouns), the feminine (slightly over 6% of all) and neuter (about 2.5% of all) nouns at 9 letters. It can be seen that masculine nouns are shorter than feminine and neuter ones and that the latter two groups do not differ much. Distributions for nouns of masculine and neuter genders are, as could be expected for any word lists, asymmetrical to the right, while the distribution of lengths for feminine nouns is leaning, surprisingly, to the left. Interesting conclusion comes from the beginning of the histogram – three-letter nouns are almost always of masculine

gender. More detailed description, shown in table 4, gives additional weight to our hypothesis.

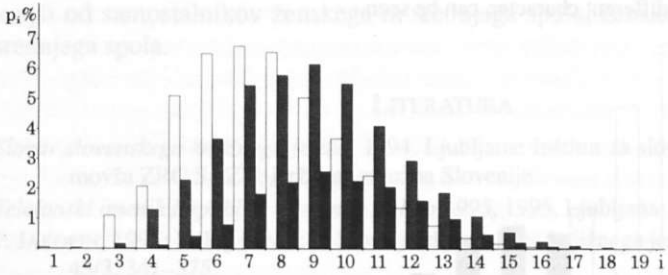


Figure 2: Distribution of lengths for common nouns by gender (m = white, f = pattern, n = black)

	Masculine	Feminine	Neuter
N	21823	21427	8198
Minimum	1	2	3
Maximum	21	24	24
Arithmetic mean	7.63	8.95	9.59
Standard deviation	2.46	2.66	2.43
Standard error of arithmetic mean	0.0166	0.0182	0.0268

Table 4: Statistical description of lengths for nouns by gender

Standard deviations are considerable; however, because of the large samples the standard errors of all the three arithmetic means are so small that the differences between average lengths for nouns of different gender, even for the feminine-neuter pair, are statistically significant.

When a good frequency dictionary of Slovenian is available it will be interesting to compare the average values and distributions of this paper to values obtained from weighted samples, where frequencies would also count. A shift to the left is obvious (weighted arithmetic mean for Christian names is 5.51 letters as opposed to 5.88 in the paper); its size for nouns would illuminate the other side of the question as well.