Emmerich Kelih
ORCID: 0000-0002-8315-8916

Peter Zörnig
ORCID: 0000-0002-1094-3972

Phoneme Frequencies in Slovene (Text vs. Dictionary)

Slovenski jezik / Slovene Linguistic Studies 14/2022. 64–95.

DOI: https://doi.org/10.3986/sjsls.14.1.03

ISSN tiskane izdaje: 1408-2616, ISSN spletne izdaje: 1581-127

https://ojs.zrc-sazu.si/sjsls

Slovenski jezik – Slovene Linguistic Studies 14 (2022): 64–95

Emmerich Kelih (ORCID: 0000-0002-8315-8916)
Institut für Slawistik, University of Vienna, Austria

Peter Zörnig (ORCID: 0000-0002-1094-3972)
Department of Statistics, University of Brasília, Brazil
DOI: https://doi.org/10.3986/sjsls.14.1.03

Phoneme Frequencies in Slovene (Text vs. Dictionary)

In this paper Slovene phoneme frequencies from a Slovene–German learner's dictionary are analysed. The structure of the dictionary allows the determination of phoneme frequencies on two distinct linguistic levels: the level of dictionary (analysis of headwords) and the level of text (example sentences, illustrating a prototypical context of a given headword). By applying various statistical significance tests it can be shown that no significant differences between the rank-frequency distributions are observable. The same holds true for testing the differences, based on the repetition rate of phoneme frequencies on the dictionary and text levels. In contrast to this, only dichotomised data (by grouping them into vowels and consonants) show a significantly different frequency behaviour. Overall it can be shown that based on the given empirical observations, the conceptual importance and relevance of the levels of dictionary vs. text for quantitative phoneme studies has to be reconsidered and critically reflected in future studies.

Keywords: Slovene, phoneme frequencies, repetition rate, frequency of vowels and consonants, statistical significance

V članku so analizirane frekvence slovenskih fonemov iz slovensko-nemškega tematskega (učnega) slovarja. Struktura slovarja omogoča ugotavljanje pogostosti fonemov na dveh različnih jezikovnih ravneh: na ravni slovarja (analiza gesel) in na ravni besedila (povedi, ki ponazarjajo prototipni kontekst določenega gesla). Z uporabo različnih statističnih testov pomembnosti je mogoče dokazati, da ni opaziti bistvenih razlik med porazdelitvami rangov pogostosti fonemov na ravni slovarja in besedila. Isto velja tudi za analizo glede stopnje ponavljanja fonemov. V nasprotju s tem dihotomizirani podatki (z razvrstitvijo v skupine na samoglasnike in soglasnike) pa vendarle kažejo bistveno drugačno »obnašanje« frekvenc. Na splošno je mogoče pokazati,

da je za kvantitativne študije fonemov treba preučiti in kritično reflektirati konceptualno sicer zelo važno razliko med slovarjem in besedilom.

Ključne besede: slovenščina, pogostost fonemov, stopnja ponavljanja, pogostost samoglasnikov in soglasnikov, statistična pomembnost

1 Introduction

This article deals with phoneme frequencies in Slovene. The basic ideas discussed here go back to the well-known Russian phonologist N. S. Trubetzkoy (Trubeckoj) (1890–1938). In his seminal monograph on the foundations of phonology (cf. Trubetzkoy 1939) he tackled the question of a statistical analysis of phoneme frequencies. He pointed out the conceptual need of distinguishing two fundamentally different levels, namely the language system as such and the text level, when counting phoneme frequencies. This theoretical assumption is taken as the starting point for an in-depth empirical and statistical analysis of Slovene phoneme frequencies. The empirical basis for our counts and quantitative analysis is a learner's dictionary of Slovene (cf. Kelih/Vučajnk 2018), where in addition to the dictionary entry (*headword*, *lemma*) the given sentences,[1] exemplifying the prototypical context of the headword, are analysed. Hence our quantitative analysis points at both the mentioned levels, the language system and text level. In the first section of the paper a short overview of the current state of the art in phoneme frequencies analysis is given. In the second section first the analysed material is presented, followed by the discussion and interpretation of the gained results. The focus is on the question whether a statistical difference between the stated levels can indeed be obtained, by using appropriate statistical tests, giving information about the statistical significance of the differences. In addition to the discussion of the raw phoneme frequencies, in a further step the repetition rate and the frequency of vowels and consonants will be examined. The third section summarises the main results of our

[1] The given sentences are not a running text in a usual sense, since only one or maximally two sentences are given. The used Slovene–German learner's dictionary is divided into two proficiency levels – 2,000 headwords which reflect the (beginners') core vocabulary and a further 2,500 dictionary entries focusing on a more advanced level.

empirical study, including some prospective and required next steps of the analysis of phoneme frequencies in the languages of the world.

2 Phoneme frequencies: Current state of the art

Phoneme frequencies play an important role in many linguistic and near-linguistic areas (for instance functional phonology, information theory, language typology, computational linguistics, cryptography etc.). In recent quantitative linguistics the main focus of interest has been on the statistical modelling of the frequencies of various linguistic entities, including phoneme frequencies. In particular, much effort has been put into the search for an appropriate theoretical (either continuous or discrete) statistical model which could capture observed empirical frequencies in an appropriate manner (Wilson and Mačutek 2020, Mitchell 2019 – both works are devoted to grapheme frequencies; Kelih 2018).

However, our basic idea discussed here is to shed (more) light on the linguistic and cognitive surroundings, related with the frequency of phonemes as a descriptive property of a language system. The occurrence of particular phonemes can be understood as a manifestation of various coding and decoding demands, where both the speaker and the hearer have a particular interest in balancing the stream of phonemes in the course of language communication (cf. Köhler 2005 for a more extended version of this synergetic approach in linguistics). To get a deeper understanding of such processes of self-organisation in a language system one can study the frequency behaviour of phonemes, where the discussion of various factors influencing the shape of phoneme frequency distributions as well as relevant properties interrelated with the frequency are of particular interest.

One crucial factor influencing phoneme distribution is the phoneme inventory size, i.e. how many different phonemes are present in the system and how they are utilised for structuring linguistic entities, like syllables, morphemes, words etc. (cf. Altmann and Lehfeldt 1980: 151f, Grzybek and Kelih 2005, Kelih 2012 for a full discussion of the

relevance of the inventory size). It is quite reasonable that depending on the phoneme inventory size, various phonotactic restrictions come into play. However, again, the inventory size influences the degree of utilisation, i.e. how many of the phonemes can be combined with other phonemes in the system. Hence, languages with a large inventory do not have a high pressure of using all possible phoneme combinations, whereas languages with a smaller inventory have to utilise the available phonemes to a greater extent. However, of course, the interrelation between phoneme inventory size and the number of phoneme combinations influences the shape of the phoneme frequency distribution.

In addition to the phoneme inventory, which is an important shaping factor, in Kelih (2012) it has been shown that also the word length is interrelated with the phoneme frequency distribution. This is explainable by referring to Menzerath's law, which states the longer the words, the shorter its syllables (for more details cf. Altmann 1980 and Coloma 2015). This basic mechanism is responsible for the regulation of word and syllable length, and affects in particular the frequency of vowels and consonants. Longer words have shorter syllables, and therefore with increasing word length the average number of vowels increases quite systematically and can be, as shown by Kelih (2016: 309ff), captured by appropriate power laws.

Leaving aside these interesting, but still not completely understood, relations and processes, the given article focuses, as already pointed out at the beginning, on some further factors which could have an influence on the shape of phoneme frequency distributions. Namely, what differences and similarities are obtained when one counts phoneme frequencies either on a systems level (dictionary) or on the text level. Usually terminologically, and owing to the structuralist paradigm, frequencies can be obtained on the paradigmatic (dictionary) and the syntagmatic level (text).

However, as often in linguistics and generally in empirical sciences, theoretical notions like dictionary or text have to be transformed into terms that open the door for empirical observation. One possible operationalisation of dictionary could be, though a rather rough

one, the analysis of bare dictionary entries (headwords, lemma). A dictionary entry is a linguistic unit, which is a normatively agreed form of recording lexical information. Usually it is a result of a commonly accepted lexicographic praxis; for instance, for most Slavic languages the infinitive of verbs, or the masculine and genitive singular of nouns etc., are cited as headwords in dictionaries. From a linguistic point of view the headword contains graphical, phonological, morphological and morphosyntactic information, but explicitly no information about the usage, for example when a noun is used in an utterance in its dative plural form or for a verb the third person masculine plural is required. Hence counting phonemes on this dictionary level is in some respects special and at the very least it should yield different results than counting phonemes on the text level. To illustrate this, at least one example has to be discussed. For instance, in the case of counting infinitives, one can expect that some particular phonemes have to occur with a frequency which is above the average. In the case of counting the Slovene[2] verbs *delati* ('to work'), *brati* ('to read') or *smejati se* ('to laugh'), one will find that due to the infinitive marker *-ti* the frequency of *t* and *i* should indeed be above the average. The same holds true for counting nouns, adjectives etc. based on their dictionary entries, with vowels such as -a, -e etc. required for marking the feminine, or the genitive form of nouns would increase the frequency of these phonemes.

In contrast to such kind of dictionary-based analysis it can be assumed that phoneme counting in running texts, containing morphosyntactic correctly formed utterances, results in a slightly different picture. From our point of view one key aspect for explaining such differences should be the frequency of synsemantic words (function words) in running texts. For this subset of word forms it is well-known that they occur quite often in running texts, and they are regulated by Zipf's law, a phenomenon of text organisation observed in many languages

[2] We are aware of the fact that in many languages headwords are given differently. For instance, in Indonesian affixed forms are part of the given displays, while other languages give the shortest form (for instance third person singular for verbs). Thus, analysing dictionary entries is to some extent arbitrary, which has to be considered in cross-linguistic studies in particular.

of the world (for some basic references about Zipfian linguistics cf. Popescu, Altmann and Köhler 2010, Pustet 2004).

The high frequency of function words in running texts is caused by the fact that they are required for the morphosyntactic organisation. This can be illustrated based on different forms of the auxiliary *biti* ('to be') in Slovene, which is usually the most frequent form when one analyses the word frequencies of running texts and followed by prepositions like *v*, *in*, *na* etc. For phoneme frequency counts, this means that in contrast to the dictionary entry *biti*, the 1st person sg. *sem*, 2nd person sg *si*, 3rd person sg. *je*, 1st person dual *sta*, 1st person pl. *smo* etc. trigger and increase the frequency of *s*, *e*, *j*, *m*, *t* etc. in running texts. Moreover, one has to consider that functional (synsemantic) words are usually shorter than non-functional (autosemantic) words. As a consequence, one can expect that the phoneme frequency distribution of running texts has to be organised in another way than the phoneme frequency distribution, based on the analysis of dictionary entries.

The assumed consequences of our basically simple idea, which in fact should hold true for a highly synthetic language[3] like Slovene, of obtainable differences between the system and text levels will be proven empirically and statistically in the next section.

3 Data used for the analysis

The basic material used for the empirical analysis is taken from a Slovene–German learner's dictionary (cf. Kelih and Vučajnk 2018), specifically the Slovene part of the dictionary only. It consists of 4,950 headwords (dictionary entries) and 5,095 example sentences, where an authentic utterance from contemporary Standard Slovene in a prototypical context is given for each headword. As an illustration, the headword imenovati se ('heißen', 'sich nennen', "to be named")

[3] For analytic languages phoneme frequency differences between dictionary entries and running texts should probably be smaller because of the lower degree of inflection in analytic languages.

looks as given in the Table 1 (English translation is added in brackets in smaller font).

Lemma (headword)	imenováti se -újem se *impf*
Beispielsatz (example sentence)	Kakó se imenúje vàš sodélavec?
Übersetzung (translation)	Wie heißt Ihr Mitarbeiter? (What is your employee's name?)

Table 1: Structure of the analysed learners dictionary

The verb is, as common in Slovene lexicography, given as infinitive *imenovati*, plus the ending of the 1st person sg. *-ujem*, including the specification of the aspect (in this case imperfect). The accompanying example sentence gives an authentic context (in our example as a question), which aims to facilitate the memorability of the given headword. As can also be seen from the given example, both the headword and the example sentence are marked by diacritics, as used in the monolingual dictionary of Slovene (*Slovar slovenskega knji*žnega jezika, 2nd edition, 2014, see <www.fran.si> for the online version).

The given phonetic annotation in respect to the suprasegmental features is a necessary precondition for the counting of the phoneme frequencies, since short, long and unstressed vowels can be distinguished. The starting point for counting the Slovene phoneme frequencies in the dictionary is the written form of Standard Slovene, which requires some preliminary remarks about Slovene phoneme–grapheme correspondences.

1. For the sake of simplicity the following 25 letters of the Slovene alphabet are understood as the basic graphemes of Slovene: < a, b, c, č, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, z, ž>. Kelih (2008) gives a detailed study of the phoneme–grapheme correspondences in Slovene. General information about the Slovene writing system is given in Rehder (2006) and Herrity (2010). For our analysis upper and lower case are not distinguished. Any foreign graphemes that may occur in the texts, such as <x, y, w>, are not counted, i.e. excluded. This also applies to punctuation marks and other numeric characters.

2. In the given context, all units listed below are counted as phonemes. In Slovenian, stress and vowel length are inherently related. In the learner's dictionary mentioned above, the following vowel notations are distinguished, according to the norm of Slovenian: six long-stressed vowels: /í, é, ə̀, á, ó, ú/, two long-stressed, open vowels: /ô, ê/ and five short-stressed vowels: /ì, è, à, ò, ù/. In addition, six unstressed vowels (/i, e, ə, a, o, u/) and twenty letters, marking corresponding consonants (<b, c, č, d, f, g, h, j, k, l, m, n, p, r, s, š, t, v, z, ž> are counted (cf. Srebot-Rejec (1988) and Tivadar (2004) for some disputed issues of Slovene tonemic vowel phonology). This results in a Slovene phoneme inventory of 39 units, containing segmental and suprasegmental phonemes. It is important to note that for the phoneme counts performed here no specific allophonic conditions (coarticulation) have been taken into consideration, e.g. regressive assimilations.

3. Although the Slovene phoneme–grapheme correspondence can be characterised as rather shallow, this does not hold true in respect to the graphemic representation of the semivowel /ə/, either in a stressed /ə̀/ or unstressed position. Toporišič (2000: 56-59) gives a detailed description of the graphemic representation of /ə/ from a normative point of view, which helps to determine the main realisations. The schwa can be represented either by <e> or in some selected cases by <è>. Moreover <ŕ> and <r>, word-initially, in pre-consonantal and inter-consonantal position, have to be taken into consideration to identify, based on the orthographic input, the corresponding realisations of the semivowel. The stressed semivowel /ə̀/ is limited to some single items, for instance in *bezèg*/*bèzeg* ('elder tree'), *pes* ('dog'), *stebèr*/*stèber* ('column, pillar'), *ves* ('entire') and frequently cited *meglà*/*mègla* ('fog'). As can be seen, even from a normative point of view alternative realisations are offered, where for the calculations the former items have been taken. Much more frequently stressed /ə̀/ can be derived from <ŕ>, which represents the phonemes /ə̀/ and /r/, as for instance in *pŕt* ('cloth'), *tŕg* ('square, market'), *prekŕšek* ('offence') etc. A more or less exact identification of the unstressed /ə/ based on the orthographic input is possible by taking into consideration morphological information. Based on the given list of frequently

occurring suffixes (Toporišič 2000: 57), for instance *ec*, *-el*, *-ek*, *-en*, *-er*, by semiautomatic filtering and subsequent manual control, mostly all relevant lexemes with unstressed /ə/ in word-final position could be extracted. In this way, out of over 4,000 instances of <e> and over 500 instances of /ə/ have been identified in the lemma list and over 800 units on the text level.

Furthermore, all given annotations (e.g. marking of the gender of nouns *f* (*feminine*), *m* (*masculine*), *n* (*neutrum*), *impf* (*imperfective aspect*), *pf* (*perfective aspect)*, *adj* (*adjective*) etc.) in the learner's dictionary have been excluded from the phoneme counts, which ensures that only the target language Slovene is analysed. In the case of inflected headwords (verbs, nouns, adjectives) however, not only the bare headword is analysed, but also the given formants, for instance the marking of the genitive singular for nouns, the first person singular for verbs, the feminine and neutral marking of adjectives etc. For example, for *imenováti se -újem se* ('to be named'), *Japónec -nca* ('Japanese'), *lep -a -o* ('beautiful') all parts and signs given here in italics are counted. All example sentences are counted fully; in some cases two example sentences are given for one headword, which explains the quantitative discrepancy of 4,950 headwords (dictionary entries) vs. 5,095 example sentences. The counts, which will be presented in detail in the next sections, were performed automatically.

3.1 Slovene phoneme frequencies on the dictionary level: Some descriptive aspects

Based on the given operationalisation criteria the phonemes in the headwords (henceforth dictionary) were counted. Since for Slovene only a few studies of phoneme frequency counts are available (cf. Hajnšek-Holz and Jakopin 1996, where some phoneme counts, based on the retrograde dictionary of Slovene, are given; for text data see Kolter 1994), a more detailed discussion of the obtained results is required. The raw data are given in Table 2, where absolute and relative frequencies are given in alphabetical order. In addition, one common kind of the presentation of phoneme frequencies, namely in ranked form, is also given. In this case the frequencies are sorted according to their rank, i.e. rank 1 = the most frequent

phoneme, rank 2 = the second-most frequent phoneme etc. The overall corpus of phonemes of the headwords (dictionary) consists of 50,413 phonemes.

Pho-neme	Abs. f	Rel. f.	% f.	Rank	Pho-neme	Abs. f	Rel. f.
a	5,686	0.1128	11.28	1	a	5,686	0.1128
á	1,704	0.0338	3.38	2	e	3,828	0.0759
à	88	0.0017	0.17	3	i	3,054	0.0606
b	729	0.0145	1.45	4	n	2,999	0.0595
c	709	0.0141	1.41	5	t	2,877	0.0571
č	841	0.0167	1.67	6	o	2,775	0.0550
d	1,319	0.0262	2.62	7	r	2,660	0.0528
e	3,828	0.0759	7.59	8	s	2,325	0.0461
é	999	0.0198	1.98	9	m	1,849	0.0367
è	161	0.0032	0.32	10	l	1,803	0.0358
ê	309	0.0061	0.61	11	k	1,786	0.0354
ə	594	0.0118	1.18	12	á	1,704	0.0338
ə̀	131	0.0026	0.26	13	p	1,689	0.0335
f	100	0.0020	0.20	14	v	1,643	0.0326
g	764	0.0152	1.52	15	í	1,613	0.0320
h	256	0.0051	0.51	16	j	1,406	0.0279
i	3,054	0.0606	6.06	17	d	1,319	0.0262
í	1,613	0.0320	3.20	18	é	999	0.0198
ì	11	0.0002	0.02	19	z	960	0.0190
j	1,406	0.0279	2.79	20	č	841	0.0167
k	1,786	0.0354	3.54	21	g	764	0.0152
l	1,803	0.0358	3.58	22	b	729	0.0145
m	1,849	0.0367	3.67	23	c	709	0.0141

n	2,999	0.0595	5.95	24	ó	685	0.0136
o	2,775	0.0550	5.50	25	ə	594	0.0118
ó	685	0.0136	1.36	26	š	470	0.0093
ò	103	0.0020	0.20	27	ú	428	0.0085
ô	344	0.0068	0.68	28	ž	358	0.0071
p	1,689	0.0335	3.35	29	u	354	0.0070
r	2,660	0.0528	5.28	30	ô	344	0.0068
s	2,325	0.0461	4.61	31	ê	309	0.0061
š	470	0.0093	0.93	32	h	256	0.0051
t	2,877	0.0571	5.71	33	è	161	0.0032
u	354	0.0070	0.70	34	ə̀	131	0.0026
ú	428	0.0085	0.85	35	ò	103	0.0020
ù	3	0.0001	0.01	36	f	100	0.0020
v	1,643	0.0326	3.26	37	à	88	0.0017
z	960	0.0190	1.90	38	ì	11	0.0002
ž	358	0.0071	0.71	39	ù	3	0.0001
	50,413	1	100				

Table 2: Slovene phoneme frequencies in the dictionary

Before going into further details regarding the individual frequency of the phonemes, it is first necessary to start with a global analysis of phoneme frequencies. The visual representation of the data, as ranked frequencies of the phonemes, is given in Figure 1.

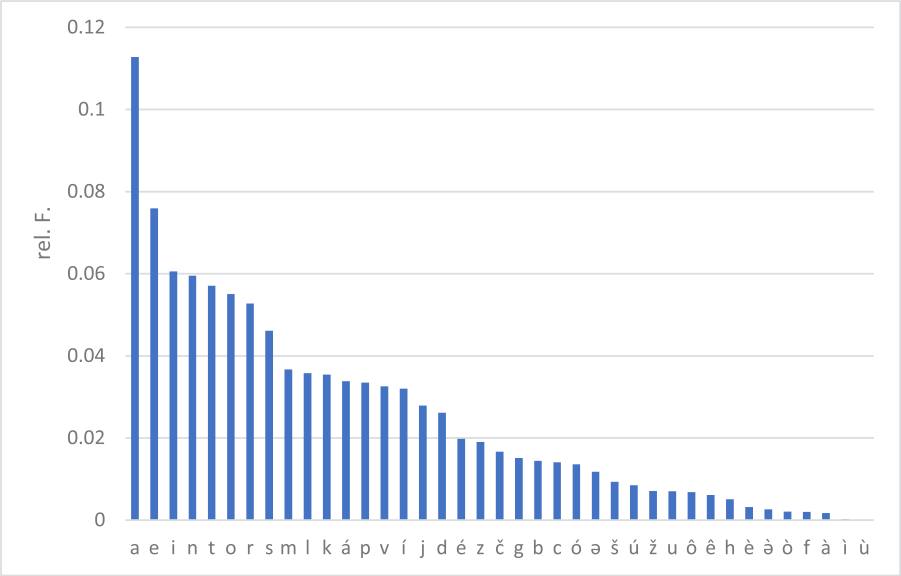

FIGURE 1: Phoneme frequencies in Slovene: Dictionary data

As can be seen, the three most frequent units in the dictionary are the vowels /a, e/ and /i/. Next, the fourth and fifth most frequent phonemes are /n/ and /t/, followed by an unstressed /o/. The front part of the rank-frequency distribution, i.e. the most frequent classes, is clearly dominated by vowels, only interrupted by the two consonant phonemes /n/ and /t/. Beyond this, another general tendency can be seen in the material analysed: obviously unstressed vowels occur most frequently, followed by long-stressed vowels, whereas short-stressed vowels only appear in the tail of the rank-frequency distribution. There seems to be a kind of frequency hierarchy of unstressed/stressed vowels, an observation which could be of interest when comparing the data to other typologically similar languages.

As regards the consonants, /n, t, r/ and /s/ are relatively frequent, however no clear tendencies related to subcategories as for sonorants, plosives, affricates etc. can be obtained. As expected, /f/ is indeed a rare, peripheral phoneme and occurs, as also in other Slavic languages, only in loanwords. The less frequent phonemes in the dictionary are /à, ì/ and /ù/, all of them short and stressed,

obviously having in Slovene a low degree of functional load. In fact /ì/ and /ù/ have a frequency close to zero, which indeed shows the absolute peripheral status of these units.

At this point, the corresponding input of our analysis must be recalled: the examined headwords contain, besides the respective root morphemes, above all possibly existing prefixes or the respective endings. In this respect, the headword analysis represents, in a certain way, system units of Slovene, which are now ordered according to their quantitative weight of their constituents. The relative dominance of vowels is worth noting – they are obviously the carriers of the morphological information in a narrow sense. Such a functional interpretation seems to us to be more plausible than attempts to relate the frequency of phonemes to the articulatory complexity (cf. Chitoran and Cohn 2009 for an overview of the related discussion). The interpretation and in-depth analysis of the ratios of unstressed and stressed vowels shall be reserved[4] for another article.

3.2 Phoneme frequencies on the text level: Some descriptive aspects

Now, in a second step, the corresponding data for the text analyses can be presented in more detail. The data presentation is done in analogy to the dictionary data, i.e. first the absolute and relative frequencies are presented and then the data are given according to their rank frequencies (cf. Table 3).

Pho-neme	Abs. f.	Rel. f.	% f.	Rank	Pho-neme	Abs. f.	Rel. f.
a	9,416	0.0627	6.27	1	e	10,924	0.0727
á	5,355	0.0356	3.56	2	o	10,305	0.0686
à	570	0.0038	0.38	3	a	9,416	0.0627
b	2,650	0.0176	1.76	4	n	9,028	0.0601

[4] The position of stressed vowels in Slovene is obviously influenced by the word length. Mačutek and Kelih (2021) show a clear tendency towards the centre (i.e. the stress is usually located in the middle of the word), but this is a tendency which is influenced by the corresponding word length.

c	1,363	0.0091	0.91	5	i	8,542	0.0568
č	2,177	0.0145	1.45	6	r	7,811	0.0520
d	4,872	0.0324	3.24	7	l	7,784	0.0518
e	10,924	0.0727	7.27	8	s	7,716	0.0514
é	3,662	0.0244	2.44	9	j	6,874	0.0457
è	512	0.0034	0.34	10	t	6,511	0.0433
ê	749	0.0050	0.50	11	v	6,103	0.0406
ə	886	0.0059	0.59	12	p	5,426	0.0361
ə̀	383	0.0025	0.25	13	á	5,355	0.0356
f	268	0.0018	0.18	14	k	5,251	0.0349
g	1,801	0.0120	1.20	15	d	4,872	0.0324
h	1,190	0.0079	0.79	16	m	4,721	0.0314
i	8,542	0.0568	5.68	17	í	4,172	0.0278
í	4,172	0.0278	2.78	18	é	3,662	0.0244
ì	84	0.0006	0.06	19	z	3,384	0.0225
j	6,874	0.0457	4.57	20	b	2,650	0.0176
k	5,251	0.0349	3.49	21	ó	2,490	0.0166
l	7,784	0.0518	5.18	22	č	2,177	0.0145
m	4,721	0.0314	3.14	23	g	1,801	0.0120
n	9,028	0.0601	6.01	24	u	1,767	0.0118
o	10,305	0.0686	6.86	25	š	1,743	0.0116
ó	2,490	0.0166	1.66	26	c	1,363	0.0091
ò	239	0.0016	0.16	27	ô	1,264	0.0084
ô	1,264	0.0084	0.84	28	h	1,190	0.0079
p	5,426	0.0361	3.61	29	ú	1,170	0.0078
r	7,811	0.0520	5.20	30	ž	1,073	0.0071
s	7,716	0.0514	5.14	31	ə	886	0.0059
š	1,743	0.0116	1.16	32	ê	749	0.0050

t	6,511	0.0433	4.33	33	à	570	0.0038
u	1,767	0.0118	1.18	34	è	512	0.0034
ú	1,170	0.0078	0.78	35	ə̀	383	0.0025
ù	21	0.0001	0.01	36	f	268	0.0018
v	6,103	0.0406	4.06	37	ò	239	0.0016
z	3,384	0.0225	2.25	38	ì	84	0.0006
ž	1,073	0.0071	0.71	39	ù	21	0.0001
	150,257	1	100				

TABLE 3: Phoneme frequency in the text

A closer look at the phoneme frequencies at the text level (cf. FIGURE 2) show, again in the case of the dictionary data, that the vowels /e, o/ and /a/ dominate the top three positions of the ranking. Rank 4 is occupied by /n/, followed by /i/ and /l/. Thus, compared to the dictionary data /l/ is now more prominently represented, while /t/, which dominates the dictionary data, slips back in terms of its relative rank frequency. This could be due to the relatively high occurrence of infinitives (usually marked by ending *-ati*) in the dictionary data, but this assumption has to be considered in more detail in the future.

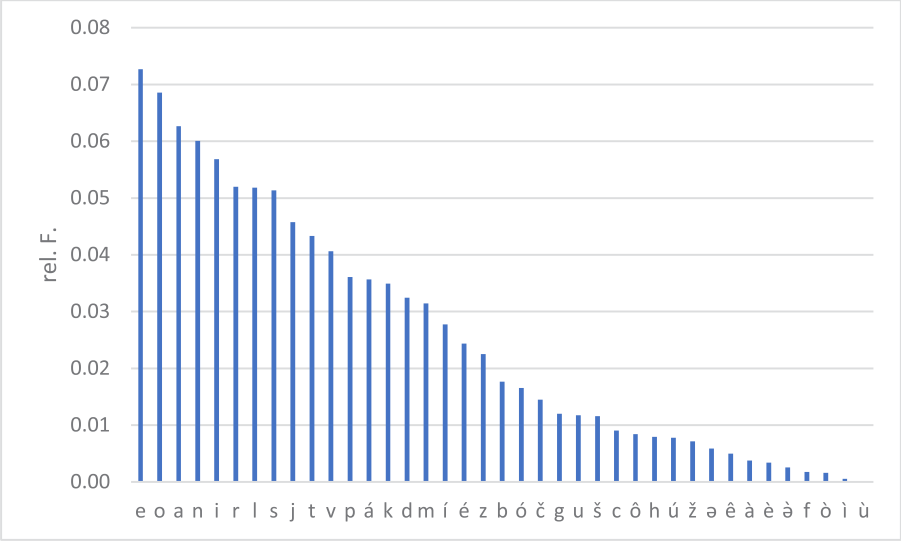

FIGURE 2: Slovene phoneme frequencies: Text data

However, what is even more striking is a shift in the ranking of consonants when one compares the dictionary and text data. In the case of the text data the five most frequent consonants are <n, l, s, r and j>, basically equally distributed (having a relative frequency of about 0.05), where four of them are sonorants. A striking shift compared to the dictionary data occurs in the case of the frequency of <j>. In the dictionary /j/ occurs with a relative frequency of 0.028, whereas in the text the frequency almost doubles to 0.046. This phenomenon can possibly be explained by the high frequency of synsemantic word forms in the text, where /j/ occurs quite often (e.g. as part of forms of the auxiliary *biti* ('to be') in the 3rd person sg. *je* ('is') or in pronouns like *jo* ('her') or *svoj* ('my', 'your', 'ours' etc.). This is an ad-hoc explanation which undoubtedly has to be examined more closely in the future.

In general, however, at first sight a rather similar picture emerges in both the dictionary and text data: the front part of the rank-frequency distribution is dominated by vowels (with an observable general order of unstressed, long-stressed, followed by short-stressed vowels), interrupted by some consonants, where in fact no absolutely clear and explainable ranking is observable. However, this overall finding must be complemented with two striking differences, obtained in the data when comparing them.

First of all, there is no common most frequent vowel in the dictionary and text data at all. In one case /a/ is the most frequent, in the other case it is /e/. There is a second noticeable quantitative difference: while the most frequent phoneme in the dictionary data occurs with a relative frequency of 0.1127, the equivalent relative frequency in the text data is only 0.0727. This is, without going into statistical details, a quite remarkable difference in regard to the functional exploitation of the vowels in the dictionary and the text. In particular, it can be seen from Figure 2 that the phonemes in texts are distributed more evenly, i.e. the frequencies in the text are more balanced, whereas in the dictionary data, the vowels /a/ and /e/ appear visually as outliers, dominating quantitatively the rank-frequency distribution, having together a frequency of 20%.

At first glance, the data presented so far may not reveal any serious differences between phoneme frequencies at the dictionary and text levels, except for the observations based on the available raw data and the corresponding graphical representation. However, our claims and observations have to be substantiated by methodologically sound procedures. Thus, in the next step, we shall analyse whether our finding of serious differences can be supported statistically or not, by applying proper[5] statistical methods.

3.3 Testing the correlation between dictionary and text: Tests of Spearman and Kendall

The frequencies in question can be compared by using appropriate statistical tests, which are available for the comparison of rank frequencies. For our kind of data, a parameter-free rank correlation test seems to be appropriate, where the related ranks are of relevance for the comparison. We apply two different statistical tests, the rank correlation test of Krüger and Spearman and the rank correlation test of Kendall, in order to check the relation between the phoneme ranks in the dictionary and those in the text. These tests can be found in many statistical textbooks; cf. Zöfel (2002: 126ff) among others.

Phoneme	Rank dictionary	Rank text	d	d^2
a	1	3	-2	4
e	2	1	1	1
i	3	5	-2	4
n	4	4	0	0

[5] In our approach we are applying parameter-free rank correlation tests, the χ^2 test for contingency tables, and test procedures developed for testing differences between repetition rates. For dealing with data on the nominal scale these seem to be the appropriate methods for testing the correlation and/or differences between the samples analysed. By doing this we can avoid the heated discussion (cf. Janda 2013 with an overview on statistical methods for testing differences of linguistic data and related pitfalls) about what in linguistics can be understood as dependent or independent samples (different texts, registers, languages etc.).

t	5	10	-5	25
o	6	2	4	16
r	7	6	1	1
s	8	8	0	0
m	9	16	-7	49
l	10	7	3	9
k	11	14	-3	9
á	12	13	-1	1
p	13	12	1	1
v	14	11	3	9
í	15	17	-2	4
j	16	9	7	49
d	17	15	2	4
é	18	18	0	0
z	19	19	0	0
č	20	22	-2	4
g	21	23	-2	4
b	22	20	2	4
c	23	26	-3	9
ó	24	21	3	9
ə	25	31	-6	36
š	26	25	1	1
ú	27	29	-2	4
ž	28	30	-2	4
u	29	24	5	25
ô	30	27	6	36
ê	31	32	-1	1

h	32	28	4	16
è	33	34	-1	1
ŕ	34	35	-1	1
ò	35	37	-2	4
f	36	36	0	0
à	37	33	4	16
ì	38	38	0	0
ù	39	39	0	0
				334

TABLE 4: Rank correlation test for phonemes in dictionary and text

TABLE 4 summarises the relevant data. Column 1 contains the n = 39 observed phonemes and column 2 gives the corresponding ranks in the dictionary, i.e. /a/ is the most frequent, /e/ the second-most frequent phoneme in the dictionary etc. Column 3 contains the ranks of the phonemes in the text, i.e. the phoneme with rank 3 in the dictionary is the most frequent phoneme in the text and the phoneme with rank 1 in the dictionary is the second-most frequent phoneme in the text etc. Column 4 contains the difference between the ranks and the last column gives the corresponding squares. The test statistic of Krüger and Spearman used to decide about the correlation between the ranks in columns 2 and 3 is given by

$$r = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \qquad \text{FORMULA 3.3.1}$$

This value lies between -1 and 1, where r=1 means that there is a perfect positive correlation between the ranks, i.e. the ranks are identical; r=-1 means that the ranks are perfectly negatively correlated, i.e. the first rank in column 2 corresponds to the last rank in column 3, the second rank in column 2 corresponds to the second-last rank in column 3 etc. For the given example the statistic (FORMULA 3.3.1) assumes the value

$$r = 1 - \frac{6 \cdot 334}{39 \cdot (39^2 - 1)} = 0.9662. \qquad \text{Formula 3.3.2}$$

The proximity of that value to the upper limit 1 indicates a high positive correlation between the ranks, i.e. high/low phoneme ranks in the dictionary usually correspond to high/low ranks in the text. In order to perform a legitimate statistical test, we consider the null hypothesis:

H_0: There is no correlation between the ranks of phonemes in the dictionary and phonemes in the text.

As required, we consider the transformed statistic

$$t = \frac{|r| \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9662 \cdot \sqrt{39-2}}{\sqrt{1-0.9662}} = 22.80 \qquad \text{Formula 3.3.3}$$

which is approximately distributed as a Student distribution with n-2 = 37 degrees of freedom. The p-value, i.e. the probability of obtaining a more extreme test value than the observed one in Formula 3.3.3, is given by

$$p = P(T < -22.80) + P(T > 22.80) = 2 \cdot \int_{22.80}^{\infty} f(x)\, dx = 2.2 \cdot 10^{-23}, \qquad \text{Formula 3.3.4}$$

where T is the Student random variable with 37 degrees of freedom and f is the corresponding density.

Since p is almost equal to 0, it is highly unlikely that the test result may be observed when the null hypothesis H_0 is valid. Thus, we must reject H_0, i.e. there is a high positive correlation between phoneme ranks in the text and phoneme ranks in the dictionary. Or in other words, the deviations between the ranks in columns 2 and 3 of Table 4 are not statistically significant, i.e. they are due to chance.

We will still consider the Kendall correlation test which results in the same conclusion, but this test is somewhat more intuitive. The null hypothesis is the same as above, but the test statistics are now

$$\tau = \frac{p_c - p_d}{p_c + p_d} = \frac{p_c - p_d}{\binom{n}{2}} \qquad \text{Formula 3.3.5}$$

where p_c and p_d are the numbers of concordant and discordant pairs of rank values in column 3 of Table 4. Let us assume that the values in this column are denoted by r_1, r_2, ...

The pair (r_i, r_j) is concordant if $r_i < r_j$ and discordant if $r_i > r_j$ (for i<j). For example, (r_1, r_5) = (3, 10) is concordant and (r_3, r_6) = (5, 2) is discordant.

It should be clear that $\binom{n}{2} = \frac{n \cdot (n-1)}{2}$ (Formula 3.3.6) is the number of all pairs which equal the sum p_c+p_d in Formula 3.3.5. The statistic (Formula 3.3.5) lies between -1 and 1. When there are many more concordant than discordant pairs, τ is close to 1; when p_d is much larger than p_c, τ is close to -1. Counting all the concordant and discordant pairs in column 3 we get p_c = 685 and p_d = 56, resulting in

$$\tau = \frac{685-56}{\binom{38}{2}} = 0.8489 \qquad \text{Formula 3.3.7}$$

The test result also indicates a high positive correlation between the considered phoneme ranks. Under the null hypothesis the τ statistic is approximately normally distributed with expected value 0 and variance

$$\sigma^2 = \frac{2(2n+5)}{9n(n-1)} = \frac{2(2\cdot 39+5)}{9\cdot 39\cdot 38} = 0.0124, \qquad \text{Formula 3.3.8}$$

resulting in the p-value

$$p = 2 \cdot \int_{0.8489}^{\infty} f(x)\ dx = 2.5 \cdot 10^{-14}, \qquad \text{Formula 3.3.9}$$

where f is now the normal density with expectation 0 and variance in Formula 3.3.8. With the same reasoning as before we reject H_0 and conclude that there is a high positive correlation between phoneme ranks in the text and phoneme ranks in the dictionary.

We have seen above that the rank distributions of phonemes in the dictionary and text are strongly related, i.e. frequent/rare phonemes in the dictionary tend to be frequent/rare in a text. The result, now statistically grounded, is linguistically quite surprising and counterintuitive, and obviously in this point does not confirm our basic ideas about the linguistically different input. Nevertheless, one possible explanation for the yielded statistical insignificance could be that the phoneme frequency in our example sentences (= running text) is determined, or predetermined, by the given headword. In other words, at least one syntagmatic realisation of the headword is in any case also part of the text phoneme frequency data and therefore a partial automatic doubling of phoneme frequencies is given. Seen through this prism, the relation dictionary vs. text is more complex than one would assume without having this empirical dimension in mind.

In the next section, a slightly different view on the data will be given, where it will be asked whether the overall utilisation of phonemes as manifested in the phoneme frequencies in the text and dictionary follows different mechanisms and regulations or not.

3.4 Comparison of the repetition rate: Dictionary vs. text

One fundamental characteristic of linguistic systems is the uneven frequency distribution of linguistic units (phonemes, graphemes, syllables, morphemes), and the over-usage of particular units of a given inventory. This is an effect as also seen in different forms of Zipf's law, being an inherent characteristic of linguistic systems and texts.

In quantitative linguistics there are many different approaches available to operationalise the concept of the functional load, functional burdening etc. of phonemes. One frequently applied metric in phonology (but also in lexicology) is the so-called repetition rate.

The repetition rate is a frequently studied quantitative phonological characteristic defined by

$$R = \sum_{r=1}^{n} p_r^{\,2}, \qquad \text{Formula 3.4.1}$$

where p_r denotes the relative phoneme frequencies.

The quantity (Formula 3.4.1) measures the degree of uniformity of a phoneme frequency distribution. The smaller the repetition rate, the more evenly distributed are the involved phonemes. Perfect uniformity

occurs when all frequencies are equal, i.e. when $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$ (Formula 3.4.2).

In this case the repetition rate is $R = n \cdot \left(\frac{1}{n}\right)^2 = \frac{1}{n}$ (Formula 3.4.3). This is the minimum

possible value of R. The characteristic (Formula 3.4.1) is at its maximum when only one phoneme occurs, i.e. when one of the p_k equals 1 and the others are 0. In this case the repetition rate is R=1. The expectation of the repetition rate is

$$R = \frac{2}{n} \quad . \qquad \text{Formula 3.4.4}$$

In Zörnig and Altmann (1983) it has been shown that the random variable (Formula 3.4.1) can be modelled by means of the Zipf–Mandelbrot law.

One could expect that the repetition rate for the dictionary is higher than that for the text. This can be justified by the fact that there is an

above-average utilisation of selected units in the dictionary, which are repeatedly necessary in coding for morphological reasons; in our study this would be the case for selected vowels. Based on Tables 1 and 2 we obtained the following repetition rates, which are given in Table 5.

Dictionary	Text
0.0490	0.0437

Table 5: Repetition rate

It can be observed that the value of the dictionary is very close to the expectation R=2/39=0.0513 given by Formula 3.4.4. . In order to decide whether the values in Table 5 are significantly different, we use a criterion proposed by Altmann and Lehfeldt (1980: 162).

Let R_d and R_t denote the observed repetition rate in the dictionary and in the text. Then R_t is significantly different from R_d if it is outside the confidence interval

$$[2/n - z_{1-\alpha/2}\sqrt{V(R_d)}, \quad 2/n + z_{1-\alpha/2}\sqrt{V(R_d)}],$$

Formula 3.4.5

where *n* is the number of phonemes and

$$V(R_d) = \frac{16(n^2-4)}{Nn^2(3n^2+4)}$$

Formula 3.4.6

is the variance; and N denotes the number of languages. In our example we have n=39 and only one language (Slovenian) is involved, but in two different manifestations (dictionary and text), thus we set N=2 in this preliminary test. We get the variance

$$V(R_d) = \frac{16(39^2-4)}{2 \cdot 39^2(3 \cdot 39^2+4)} = 0.00175.$$

Formula 3.4.7

For a level of significance of 5% the quantile in Formula 3.4.5 is given by $z_{1-\alpha/2} = 1.96$ (Formula 3.4.8). Thus the confidence interval obtains the form

$$[2/39 - 1.96\sqrt{0.00175}, \quad 2/39 + 1.96\sqrt{0.00175}] = [-0.0307, 0.1333].$$

Formula 3.4.9

Since the repetition rate is a positive quantity, any repetition rate smaller than 0.1333 can be considered not significantly different from the dictionary repetition rate of 0.0490. In particular, we can conclude that the values in Table 5 are not significantly different.

Again, the applied test does not confirm our linguistic assumptions in regard to a supposed different frequency behaviour in the dictionary and text, now based on the repetition rate, which gives information about the overall distribution and not only about the behaviour in the similarity/difference of the involved phonemes according to their rank. For the time being, our result holds, of course, only for the material used. Many further studies for other languages are required to see whether this is an overall trend or, what has to be excluded, caused by the data used here. In any case, our single observations and these first test results raise questions about the basic and fundamental linguistic concepts of dictionary vs. text.

3.5 Vowel and consonant frequencies — are they correlated?

Finally, another view on the data is possible, by dichotomising the obtained frequency data into two basic phonetic/phonological subgroups: vowels and consonants. As already pointed out, vowels play an active and important role in the encoding of endings in Slovene. This makes it likely to see vowels as the main carriers of grammatical information in a narrow sense. The dichotomisation of the data allows a more focused view on the functional load of vowels and consonants. As vocalic units < a, á, à, e, é, è, ê, ə, ə̀, i, í, ì, o, ó, ò, ô, u, ú, ù> have been counted; the absolute and relative frequencies of vowels and consonants for the dictionary and text data are given in Table 6. In the

case of Slovenian (as a representative of an Indo-European language from the family of Slavic languages) one obtains at the dictionary level (cf. Table 6 with the corresponding raw data) more than 45% vowels. In comparison, the frequency of vowels in the example sentences is only about 42%. In other words, at first glance, one is again dealing with a rather different degree of utilisation of vowels at the dictionary and text levels. This confirms our first conclusion that apparently the morphological information in Slovenian is mainly or more likely to be carried by vowels. This can also be deduced from the analysed material of the headwords (for example in the case of the genitive of feminine nouns, which is mainly expressed by <-e>, while for masculine nouns it is in the genitive <a>).

	Dictionary		Text	
	Abs. f.	Rel. f.	Abs. f.	Rel. f.
Vowels	22,870	0.4537	62,511	0.4160
Consonants	27,543	0.5463	87,746	0.5840
Sum	50,413	1	150,257	1

Table 6: Vowel and consonant frequency (dictionary, text)

We now apply the χ^2 test for contingency tables, which is suitable for the given situation. We want to decide whether the above-observed deviation between the vocalic part in the dictionary (45%) and that in the text (42%) is statistically significant.

	Dictionary	Text	Row sums
Vowels	O_{11} = 22,870 (E_{11} = 21,450)	O_{12} = 62,511 (E_{12} = 63,931)	s_1 = 85,381
Consonants	O_{21} = 27,543 (E_{21} = 28,963)	O_{22} = 87,746 (E_{22} = 86,326)	s_2 = 115,289
Column sums	t_1 = 50,413	t_2 = 150,257	n = 200,670

Table 7: χ^2 test results

Table 7 contains the observed frequencies O_{ij} and the expected ones E_{ij} (in brackets). The latter are calculated as $E_{ij} = \frac{s_i t_j}{n}$ (Formula 3.5.1), where s_i is the sum of the observed values of row i, t_j the sum of observed values of column j, and $n = s_1+s_2$ is the total number of observations. E_{ij} represents the frequencies that would occur under the null hypothesis H_0 that the vocalic part in the dictionary and that in the text are independent. The test statistic is

$$t = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \ , \qquad \text{Formula 3.5.2}$$

which has approximately a χ^2 distribution with (r-1)(c-1) degrees of freedom, where r and c denote the number of rows and columns of the contingency table. From Table 7 we obtain

$$t = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E} = 218.6,$$

Formula 3.5.3

where the number of degrees of freedom is (2-1)(2-1)=1. The p-value is therefore

$$\int_{218.6}^{\infty} \frac{1}{\sqrt{2\pi}} x^{-1/2} \, e^{-x/2} \, dx \approx 1.8 \cdot 10^{-49} \qquad \text{Formula 3.5.4}$$

This shows that it is highly unlikely that the deviation between the vocalic parts is due to chance. We have to reject the H_0 hypothesis and can state that there is a significant correlation between the vocalic proportions of dictionary and text. As a matter of fact, only on the dichotomic level of vowels and consonants can a considerable difference between dictionary and text be observed. Whether this tendency also holds true for other languages is, for the time being, unclear, and must be examined in detail in the future. Interestingly

enough, based only on a rough scaling an overall difference in the organisation of phoneme frequencies can be observed, whereas the individual behaviour of phonemes seems to play only a secondary role. Our findings can be directly linked to older ideas by Skalička (1966: 114), who claimed that vowels not only have an acoustic and syllabic function, but also appear as carriers of morphological information, which obviously is more present on the dictionary than text level. In any case, further investigations are necessary, in particular the analysis of various parts of words with separate phoneme frequency counts of individual morphological segments. This could be of help for better assessment of the morphological–grammatical dimension of vowels and consonants.

4 Conclusion

The present investigation brings some interesting findings about the behaviour of phoneme frequencies in Slovene and in general. From a linguistic point of view, there can't be any doubt about the conceptual and heuristic meaningfulness of the distinction between the dictionary and the text. However, in terms of an operationalisation, the doors are open for various possibilities and interpretations. In our attempt at one practical possibility the analysis of headwords and running text from a learner's dictionary was selected. Admittedly, based on the used material no universal generalisations can be drawn, but at least some of the observations are relevant for the future analysis of phoneme frequencies.

Our initial hypothesis was that phoneme frequencies based on dictionary data provide information about the system components of a language, whereas text data reflect the text structure conditioned by the repetition of synsemantic words. As such, differences in the rank-frequency distribution of phonemes are expected, but it was very clear that this couldn't be confirmed with the data used here. First of all, a significant correlation between dictionary and text could be observed regarding the distribution of the rank frequencies. Secondly, the same holds true if one takes into consideration the overall distribution by utilising the repetition rate of phonemes. And thirdly, only in the

case of a dichotomisation of the data and by comparing the vowel and consonant frequencies can some differences in the frequency behaviour be obtained. In this aspect, vowels and consonants obviously have a different function in the dictionary, since in this case they are clearly overrepresented, which confirms their relevance for marking morphological endings, suffixes etc. in Slovene. However, the yielded results require a cautious interpretation, because of the specificity of the used material (learner's dictionary), where in the running text the given headword appears in any case, which causes a deterministic partial doubling of the phoneme frequencies. Seen through this prism, the relation dictionary vs. text is more complex than one would assume without having this empirical dimension in mind. Hence, for the time being there is no need to question N.S. Trubetzkoy's basic idea, although in detail some specifications are required. In sum, one swallow does not make a summer, and it is obvious that one has to ask to what extent or in which way the dichotomy of system vs. text can be operationalised in quantitative phonological studies in the future. Particular attention has to be paid to the sample sizes of considered tests and to a cautious interpretation of the p-values. Furthermore, all classical questions and problems of quantitative phoneme studies, like the appropriate sample size and the analysis of influencing factors (such as inventory size of phoneme systems, phonotactic constraints, syllables and word structure) and their mutual interrelations, remain as prospective desiderata.

References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In: R. Grotjahn (ed.): *Glottometrika 2.* Bochum: Brockmeyer. 1–10.

Altmann, Gabriel. Lehfeldt, Werner. 1980. *Einführung in die quantitative Phonologie.* Bochum: Brockmeyer.

Chitoran, Ioana. Cohn, Abigail C. 2009. Complexity in phonetics and phonology: gradience, categoriality, and naturalness. In: F. Pellegrino, E. Marsico, I. Chitoran and Ch. Coupé (eds.): *Approaches to Phonological Complexity.* Berlin: Mouton De Gruyter. 21–46.

Coloma, Germán. 2015. The Menzerath-Altmann Law in a Cross-Linguistic Context. *Sky Journal of Linguistics* 28: 139–159.

Grzybek, Peter. Kelih, Emmerich. 2005. Häufigkeiten von Buchstaben/ Graphemen/Phonemen: Konvergenzen des Rangierungsverhaltens. *Glottometrics* 9: 62–74.

Hajnšek-Holz, Milena. Jakopin, Primož. 1996*. Odzadnji Slovar Slovenskega Jezika po Slovarju Slovenskega Knjižnega Jezika.* Ljubljana: ZRC. SAZU.

Herrity, Peter. 2010. *Slovene. A comprehensive grammar.* London: Routledge.

Janda, Laura A. 2013. Quantitative methods in Cognitive Linguistics: An introduction. In: L. A. Janda (ed.): *Cognitive linguistics. The quantitative turn. The essential reader.* Berlin: de Gruyter Mouton. 1–32.

Kelih, Emmerich 2018. Phonemfrequenz. In: R. Köhler, S. Naumann and P. Grzybek (eds.): *Quantitative und Formale Linguistik.* Berlin: de Gruyter. DOI: https://doi.org/10.1515/wsk.9.0.phonemfrequenz

Kelih, Emmerich. 2016. *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie. Morphologie und Syntax.* Frankfurt am Main: Lang.

Kelih, Emmerich. 2012. Systematic interrelations between grapheme frequencies and word length: Empirical evidence from Slovene. *Journal of Quantitative Linguistics,* 19, 3, 205–231.

Kelih, Emmerich 2008. The phoneme-grapheme relationship in Slovene. In: G. Altmann and F. Fengxiang (eds.): *Analyses of Script. Properties of Characters and Writing Systems.* Berlin. New York: Mouton De Gruyter. 61–74.

Kelih, Emmerich. Vučajnk, Tatjana. 2018. *Slovensko-nemški tematski slovar: osnovno in razširjeno besedišče : 4500 gesel, frazemov in stavčnih primerov. Grund- und Aufbauwortschatz Slowenisch-Deutsch. 4500 Lemmata Phrasen und Satzbeispiele.* Klagenfurt/Celovec, Wien/Dunaj: Hermagoras/Mohorjeva.

Köhler, Reinhard. 2005. Synergetic linguistics. In: R. Köhler, G. Altmann and R.G. Piotrowski (eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin, New York: de Gruyter. 760–774.

Kolter, Franz. 1994. *Kvantitativna fonološkoa analiza slovenščine. Kosmačev roman Pomladni dan.* Klagenfurt: Diploma thesis.

Mačutek, Ján. Kelih, Emmerich. 2022. Free or not so free? On stress position in Russian, Slovene, and Ukrainian. (in press, *Proceedings of Qualico. Tokyo 2021*).

Mitchell, David. 2019. Models of Lithuanian Grapheme Frequencies. *Journal of Quantitative Linguistics,* 26, 2. 172–185.

Popescu, Ioan-Iovitz. Altmann, Gabriel. Köhler, Reinhard. 2010. Zipf´s law—another view. *Quality and Quantity* 44, 4. 713–731.

Pustet, Regina. 2004. Zipf and his heirs. *Language Sciences* 26, 1. 1-25.

Rehder, Peter. 2006. Das Slovenische. In: P. Rehder (ed.): *Einführung in*

die slavischen Sprachen. (mit einer Einführung in die Balkanphilologie). 5. Auflage. Darmstadt: Wissenschaftliche Buchgesellschaft. 230–245.
Skalička, Vladimír. 1966. Konsonatenkombinationen und linguistische Typologie. *Travaux Linguistiques de Prague* 1: 111–114.
Srebot-Rejec, Tatjana. 1988. *Word acent and vowel duration in standard Slovene. An acoustic and linguistics investigation.* München: Sagner.
Tivadar, Hotimir 2004. Fonetično-fonološke lastnosti samoglasnikov v sodobnem književnem jeziku. *Slavistična revija* 52, 1, 31–48.
Toporišič, Jože 2000. *Slovenska slovnica.* Maribor: Obzorja.
Trubetzkoy, N.S. 1939. *Grundzüge der Phonologie.* Göttingen: Vandenhoeck & Ruprecht (= Travaux du Cercle Linguistique de Prague, 7). (citation based on 7th edition, 1989)
Wilson, Andrew. Mačutek, Ján. 2020. A Classification of the Celtic Languages Based on Grapheme Frequencies. In: Emmerich Kelih and Reinhard Köhler (eds.): *Words and Numbers. In Memory of Peter Grzybek (1957-2019).* Lüdenscheid: Ram-Verlag. 53–68.
Zöfel, Peter. 2002. *Statistik verstehen. Ein Begleitbuch zur computergestützten Anwendung.* München u.a.: Addison-Wesley.
Zörnig, Peter. Altmann, Gabriel. 1983. The Repeat Rate of Phoneme Frequencies and the Zipf-Mandelbrot Law. *Glottometrika* 5: 205-211.

Received October 2021, accepted February 2022.
Prispelo oktobra 2021, sprejeto februarja 2022.

Acknowledgements

We are grateful to two anonymous readers for helpful advices and remarks, which helped to improve the quality of the paper.

Summary

Phoneme Frequencies in Slovene (Text vs. Dictionary)

In the given article an in-depth statistical and linguistic analysis of Slovene phoneme frequencies is given. The gained data are based on a Slovene–German learner's dictionary, which gives the possibilities of counting Slovene phoneme on two different levels. One the one side by counting phonemes in lemmas the dictionary level is covered. In addition, on the other side, phoneme counts in example sentences, illustrating a prototypical context of a given headword, are performed, what represents an analysis on the level of text. For both levels the frequency of every single phoneme of Slovene,

having a phoneme inventory of 39 units, is given and briefly commented. The comparison of the data coming from the dictionary and the text level shows, that there are no statistically significance differences between them, although in regard to the individual position of phonemes in the rank-frequency distribution differences are observable. The same holds true for the analysis of the repetition rate, which usually is considered as measure of the degree of uniformity of a phoneme frequency distribution. But, as a matter of fact, only on the dichotomic level of vowels and consonants a considerable difference between the dictionary and text data can be observed. Our results by no means question the important and well-known differentiation of dictionary vs. text, but they should be understood as further motivation for a critical reflection of a future operationalization of basic linguistic notations in quantitative phonology.

Frekvenca fonemov v slovenščini (besedilo in slovar)

V članku je podana poglobljena statistična in jezikoslovna analiza pogostosti slovenskih fonemov. Pridobljeni podatki temeljijo na slovensko-nemškem tematskem (učnem) slovarju, ki daje možnosti štetja slovenskih fonemov na dveh različnih ravneh. Na eni strani je s štetjem fonemov v lemah pokrita slovarska raven. Poleg tega se na drugi strani izvaja štetje fonemov v danih vzorčnih povedih, ki ponazarjajo prototipni kontekst geselske besede, kar predstavlja analizo na ravni besedila. Za obe ravni je podana in na kratko komentirana frekvenca vsakega posameznega fonema v slovenščini, ki ima fonemski inventar od 39 enot. Primerjava podatkov s slovarske in besedilne ravni pokaže, da ni mogoče potrditi statistično značilnih razlik, čeprav so glede na individualni položaj fonemov v rang-frekvenčni porazdelitvi razlike možne. Enako velja za analizo stopnje ponavljanja, ki jo običajno obravnavamo kot merilo stopnje enakomernosti porazdelitve pogostosti fonemov. Vendar je dejansko le na dihotomični ravni samoglasnikov in soglasnikov mogoče opaziti statistično razliko med slovarjem in besedilom. Naši rezultati nikakor ne postavljajo pod vprašaj pomembnega in znanega razlikovanja med slovarjem in besedilom, temveč jih je treba razumeti kot dodatno motivacijo za kritičen razmislek o prihodnji operacionalizaciji osnovnih jezikovnih enot v kvantitativni fonologiji.