

Rank-frequency distribution of phonemes in texts and corpora across languages: One model to capture them all?

Anonymous

ABSTRACT

We present phoneme rank-frequency distributions from 24 languages, covering 15 language families. The negative hypergeometric distribution is used as a theoretical model, it fits all the data sufficiently well. We provide some hints towards an interpretation of the model parameters – their fitted values closely correlate with the relative frequency of the most frequent phoneme and with the relative mean rank. The parameters also closely correlate between themselves. The study reveals a universal pattern of behaviour across typologically different languages.

Keywords:
phoneme,
rank-frequency
distribution,
quantitative
linguistics,
negative
hypergeometric
distribution

INTRODUCTION

1

The term *phoneme frequency* often refers in scientific literature not only to the frequency of phonemes themselves. It has been used also to denote numbers of occurrences of closely related linguistic units such as, e.g., phones, allophones, and speech sounds, as well as classes of phonemes grouped together (e.g., the frequency of vowels, consonants, or consonant subclasses). Furthermore, it often encompasses also the frequency of segmental and suprasegmental properties (such as stressed vs. unstressed vowels, specific analyses of vowel

quantity, etc.). Understood in this broad terminological sense, the study of phoneme frequency has a long-standing tradition. Notable pioneers in this field include Förstemann (1846, 1852), Schleicher (1852), Budilovič (1883), Dikarev (1891), and Peškovskij (1925). Most of works were primarily descriptive in nature, interpreting the frequencies of phoneme subclasses (such as vowels, consonants, etc.) as typological characteristics (for instance, in comparative-historical linguistics). Relevant examples range from the abovementioned early studies through later structuralist works (e.g., Burlakova *et al.* 1962, Segal 1972) to more recent publications (Majewicz 1989, Sawicka and Holvoet 1991, Sawicka 2008).

This approach has for a long time essentially remained unchanged. However, quantitative linguistics (cf. Altmann 1972, 1973, 1978, 1997 and Köhler 1986, 2005) has increasingly recognized that the goal should not merely be the descriptive recording of quantifiable linguistic features; rather, the focus should shift to the interrelations and interactions between linguistic levels and properties. For phoneme frequencies in particular, a systematization of principles of mathematical modelling can be found in Altmann and Lehfeldt (1980). The evaluation and statistical analysis of phoneme frequencies are inextricably linked to questions of operationalization, to general methodological issues regarding counting methods, and, ultimately, to the actual linguistic interpretation of phoneme counts.

With regard to the quantitative analysis of phoneme systems, older findings indicate that the quantitative makeup of the phoneme inventory can significantly influence syllable and word structure (Köhler 1986, Fenk-Oczlon and Fenk 2008) as well as the number of phoneme combinations (Kelić 2009). A series of studies on grapheme frequencies in Slavic languages has demonstrated that the size of the inventory directly influences the shape of the rank-frequency distribution of graphemes (see, e.g., Grzybek and Kelić 2005b and Grzybek *et al.* 2006b). This reveals that a paradigmatic property (such as the size of the phoneme or grapheme inventory) exerts an empirically demonstrable influence on syntagmatic properties, such as the frequency of phonemes or graphemes, the number of phoneme combinations, and so forth.

The paper is organized as follows. In Section 2, methodological problems and choices one has to make when studying phoneme fre-

quencies are discussed. Section 3 provides an overview of attempts to model phoneme frequencies from the past. Section 4 introduces the negative hypergeometric distribution as a candidate for a general model for phoneme rank-frequency distributions. Language material used in the paper is described in Section 5. Section 6 contains results of fitting, together with some hints towards possible parameter interpretations. Finally, Section 7 concludes the paper and also presents some topics for future research.

METHODOLOGICAL ISSUES

2

A crucial linguistic issue in phoneme counting concerns the adoption of a suitable and operationalizable definition of the phoneme. This question—which has been the subject of intense linguistic debate—must be viewed in light of the specific hypothesis being pursued regarding phoneme frequency counts. Like linguistic units in general, phonemes are theoretical constructs whose relevance emerges only through their integration into a broader theoretical framework (Altmann and Lehfeldt 1980, p. 104).

In addition to choice of an operational definition, determining phoneme frequencies necessitates other fundamental linguistic decisions, such as distinguishing between segmental and suprasegmental properties. Regarding segmental units, one must also decide how to treat diphthongs, affricates, palatalization correlations, geminates, and the like. As for suprasegmental properties such as stress, accent, length, tone contours, and similar features, different counting methods may need to be considered. For instance, suprasegmental properties can be viewed either as components of segmental phonemes or as distinct suprasegmental phonemes. In any case, these decisions often entail significant differences in the resulting phoneme inventory. Various approaches to phoneme counting have been extensively discussed in the literature. For practical proposals regarding different counting methods, see Moulton (1956) for German and Hill (1987) for Slavic languages. Grzybek and Rusko (2009) discuss, on the example of the Slovak language, how different counting methods affect the phoneme

inventory and some properties of phoneme frequencies (e.g., their repetition rate).

Before counting phoneme frequencies, the scope of the investigation must be precisely defined. Trubetzkoy (1939, p. 231ff.) had already pointed out the necessity of distinguishing between counts based on texts and those based on dictionaries, as differences are to be expected depending on the chosen text base.¹ Drawing on this postulate—well known in structuralist linguistics—Altmann and Lehfeldt (1973, p. 78) distinguish in this regard between systemic and pragmatic samples. A systemic sample is understood as the counting of phonemes in a dictionary² that is, the extent to which phonemes are utilized in the formation of linguistic units. Within pragmatic samples, a distinction can be made between different functional styles or text types. However, the possibility that phoneme frequency may also be an individual characteristic of authors cannot be ruled out. Overall, it can be assumed that varying phoneme frequencies may be observed depending on the specific linguistic level (dictionary vs. text) under investigation.

The question of sample size also plays an important role. While older studies on phoneme and grapheme statistics implicitly assumed that increasing the sample size would lead to a stabilization of individual phoneme frequencies, the prevailing view today is that any increase in sample size leads to greater heterogeneity in the data (see Grzybek and Kelih 2003, pp. 157–159 for a presentation of a method for estimating an adequate sample size based on the normal distribution). Against this backdrop, a linguistically grounded selection of research material is preferable to a study focused exclusively on large, extensive samples. For a clear overview of sampling methods in phoneme counting, see Altmann and Lehfeldt (1980, p. 120ff.). Kubáček (1994) provides an overview of current debates and the calculation of sample sizes using confidence intervals in quantitative

¹Trubetzkoy was a pioneer of structural phonology but also of phonological statistics.

²Either literary in a dictionary, or in a list of words occurring in a text, a collection of texts, a corpus etc. (i.e., in word types as opposed to tokens). If a list of words is used, one must make a decision whether to analyze word lemmas or all word forms.

phonology. Ultimately, the decision should be guided less by the sample size alone and more by the underlying linguistic hypothesis.

From a methodological perspective, it is necessary to distinguish between different aspects of phoneme frequencies. On the one hand, they can be viewed at a descriptive-analytical level, and on the other, at a modeling level (involving mathematical models). At the descriptive-analytical level, the focus can be on the frequency of a particular phoneme (or several phonemes). Various descriptive characteristics can be calculated on this basis, enabling the application of diverse statistical methods—for instance, to compare frequencies across different samples or languages. Counting phonemes and determining phoneme frequencies should be viewed as an important intermediate step in quantitative phoneme analysis. The goal of any such counting and frequency analysis is to provide a linguistic interpretation of the resulting data or metrics. Ranked frequencies (i.e., phoneme frequencies ordered from the highest to the lowest one) represent another form of presentation. Considering rank-frequency distributions (as opposed to frequencies of individual phonemes) makes it possible to compare phoneme frequency behaviour in languages with different phoneme inventories (see Koščová *et al.* 2016 for a short discussion on advantages of using rank-frequency distribution in grapheme frequency modelling). At the modeling level, the focus shifts to the mathematical modeling of phoneme ranked frequencies using appropriate discrete distribution models or continuous functions. A necessary step in an evaluation of goodness-of-fit of a theoretical model (very loosely speaking, it must be checked if the model is "close enough" to the data), One can choose either a statistical test or an established rule of thumb (both approaches have their advantages and drawbacks, see Mačutek and Wimmer 2013).

RANK-FREQUENCY DISTRIBUTIONS: MODELS AND ATTEMPTS AT INTERPRETATIONS

3

To begin with, we refer to a selection of key studies in this field. Swedish linguist Bengt Sigurd made an early attempt to model phoneme

frequencies using a continuous function. Based on phoneme frequency data from six different languages (American English, Bengali, Kaiwa, Samoan, and Swedish), Sigurd (1968, pp. 5ff.) demonstrated that neither Zipf's law—so commonly used in quantitative linguistics—nor its mathematical modifications are able to adequately model the data. A geometric series was proposed as a model, which—based on the respective graphical representation—was deemed suitable for the phoneme frequencies under investigation. Further proposals for modeling phoneme frequencies using continuous distributions can be found in Tuldava (1995) and Altmann (1993). Martindale *et al.* (1996) and Grzybek *et al.* (2004) provide an overview of functions and probability distributions used to model phoneme and grapheme frequencies.

Needless to say, if a mathematical model (not only of phoneme rank-frequency distributions) should be relevant for linguistics, not only a well-fitting distribution (or function, etc.) must be found, but also its linguistic interpretation. A key desideratum in the modeling of phoneme frequencies is a plausible interpretation and explanation of the parameters of the model—or, more generally, the identification of factors that can influence phoneme frequencies. Another desideratum in this field is the integration of the model into a general mathematical model of linguistic phenomena (such as, e.g., the one proposed by Wimmer and Altmann 2005).

Despite the wide range of distribution models and functions proposed, a satisfactory linguistic rationale regarding the suitability of individual models is still lacking. Phoneme frequencies are discussed in various contexts; considering them could potentially yield a linguistic interpretation of the models. In the following we list the most significant approaches.

One can assume that phoneme frequency is linked to the complexity of a phoneme's articulation. A prominent proponent of this view is Zipf (1935), who proceeds from the postulate that the easier a phoneme is to pronounce, the more frequently it occurs (p. 69). For instance, he examined the frequency of aspirated consonants in the Chinese and Danish phoneme systems, using his findings as the basis for the following generalization: "[...] that the magnitude of complexity of a phoneme tends to bear an inverse relationship to its relative frequency [...]". However, this approach—which addresses the

complicated interplay between frequency and complexity in general—has been criticized by various scholars (cf. Trubetzkoy 1939, p. 235 and Nikonov 1963, p. 262). The concept of articulatory difficulty or phoneme complexity is difficult to quantify and must be regarded as a relative term. Altmann and Lehfeldt (1980, p. 115) acknowledge the possibility of a connection between the frequency of sounds and their phonetic characteristics, yet they also recognize theoretical difficulties in determining phoneme complexity, articulatory difficulty, and the like. However, a concrete result of analyzing phoneme frequencies is the finding that certain manners or modes of articulation are indeed utilized more than average across the world's languages (see also Lindblom 1983, p. 232 and Lindblom and Maddieson 1988). The human speech apparatus tends to avoid producing "extreme" sounds, aiming instead to produce phonemes or sounds in a manner that facilitates successful communication. There is also reason to suspect that phoneme frequencies might be linked to the perceptual intelligibility of sounds or phonemes (cf. Tarnóczy 1961, p. 76). Janson (1986) and Kawasaki-Fukumori (1992) also link (albeit indirectly) the existence of a perceptual contrast between individual phonemes that is necessary for communication to their frequency of occurrence and the formation of phoneme sequences. In general, these considerations hint at a close connection of phoneme frequencies with the concept of language economy.

Martinet (1955) adopts a functional perspective influenced by the Prague School in his interpretation of phoneme frequencies. For him, phonemes are meaning-bearing units that serve as "building blocks" for the formation of words. Therefore, an interrelationship can be assumed between the frequency of phonemes and, for example, their participation in word formation. Thus, there is a reciprocal relationship between the frequency of occurrence of word forms—determined, for instance, by stylistic, morphological, or syntactic factors—and the phoneme frequency in those word forms. It is well known that the utilisation of specific linguistic units is uneven across all linguistic levels; this applies equally to the individual phonemes comprising a phoneme system (see Nikonov 1963 and Hug 1979, pp. 168ff.). The thesis that phonemes should be viewed primarily as building blocks of syllables, morphemes, and word forms proves productive for a broader typological interpretation of phoneme frequencies. It explains, for ex-

ample, why languages utilise their phoneme inventories differently based on their morphological characteristics, necessitating an above-average degree of utilisation for certain phonemes in each case. In a comparison of phoneme frequencies in Russian and Bulgarian—both Slavic languages—Nikonov (1963) demonstrates, for instance, that the phoneme /m/ occurs more frequently in Russian than in Bulgarian because it plays a crucial role in affix formation in Russian. Conversely, the phoneme /t/ exhibits a higher frequency in Bulgarian, since this consonant is employed, among other things, to mark the postpositive article. Phoneme frequency is closely linked to the functional load of phonemes (see, e.g., Francescato 1962).

An older approach posits that the well-known variation in phoneme frequencies across the world's languages stems from the physiological nature of the speech organs and the varying ways in which they are utilized (van Ginneken 1932, 1933; similarly Segal 1972). Starting from two distinct bases of articulation (anterior and posterior), various mixtures—or "sound crossings"—develop; consequently, the specific proportioning of different modes of articulation—whether by the individual speaker or within a language—is cited as a cause for the observed differences in frequency distributions. Overall, this attempt should be viewed as an early—albeit unsuccessful—effort to link linguistics, anthropology, and biology. A detailed contextualization and discussion of the reception of this approach, including references to phoneme statistics that were conducted, can be found in van der Stroom (1995). More recent approaches in phonetics and phonology that deal with the utilization of manners or modes of articulation—such as the "Quantal Theory of Speech" (cf. Stevens 1972, Stevens and Keyser 2010)—have not yet demonstrated any links to phoneme frequency. This also applies to "Maximal Dispersion Theory", which, based on striking commonalities in the world's sound and phoneme systems, posits that these systems are shaped by certain general principles. According to one such principle, a prerequisite for successful communication is that sounds must exhibit "maximal perceptual contrast" in order to be successfully recognized by a listener. This contrast is, in turn, responsible for the varying qualities of vowels (cf. Liljencrants and Lindblom 1972, Ohala 1983). Here, too, potential links to usage frequency remain unexplored.

Some scholars assumed a connection between phoneme frequen-

cies and the relative "age" of the phonemes. For instance, Lotz (1952, p. 234) points out in his comparative study of Hungarian vowel frequencies that vowels emerging later in diachronic terms exhibit very low syntagmatic frequency. Nikonov (1963) presents a similar argument, namely, that observed phoneme frequencies are the result of historical processes. These remarks are thus to be understood primarily as a pointer towards an explanation of the differences in phoneme frequencies observable across languages.

In the next Section 4, we follow an empirical approach to modelling phoneme rank-frequency distribution. We present the distribution which is a general model in the sense that it fits sufficiently well all data presented in this paper. The results of fitting and hints towards an interpretation of its parameters are then presented in Section 6.

NEGATIVE HYPERGEOMETRIC DISTRIBUTION AS A GENERAL MODEL

4

We use the negative hypergeometric distribution (the NHGD henceforth) as a model phoneme ranked frequencies. The distribution is defined as

$$(1) \quad P_x = \frac{\binom{M+x-1}{x} \binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}, \quad x = 0, 1, \dots, n,$$

see Wimmer and Altmann (1999, pp. 465--468). Given that the item with the highest frequency is (conventionally) assigned the rank of 1, we shift the distribution from (1) to the right by 1, which yields

$$(2) \quad P_x = \frac{\binom{M+x-2}{x-1} \binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \quad x = 1, 2, \dots, n+1.$$

The choice of the NHGD as the model is motivated by two facts.

First, it was shown to fit well rank-frequency distributions of graphemes in languages that use alphabetic scripts.³ There are extensive studies conducted on Slavic languages (Grzybek *et al.* 2004, Grzybek and Kelih 2005a,b, Grzybek *et al.* 2006a,b, Grzybek 2007b, Grzybek and Rusko 2009, Grzybek *et al.* 2009, Koščová *et al.* 2016). Some other languages, although investigated far less systematically in this aspect, display the same behaviour - the NHGD is a model that achieves a good fit (Celtic languages in Wilson 2013 and Wilson and Mačutek 2020, English in Pande 2021, German in Best 2005 and Grzybek 2007a, Maninka-Mori in Rovenchak and Vydrin 2010). These results provide a clear indication that rank-frequency distributions of graphemes are regularly organised. No other model can claim to be so general.

Moving from graphemes to phonemes, we assume that the phoneme-grapheme relationship is not arbitrary; rather, there is a tendency towards a systematic graphemic representation of phonemes. Naturally, there are languages with shallow orthographies and those with deep orthographies (i.e., the phoneme-grapheme relationship can be close to 1 : 1, or it can be relatively far away from this ratio, see Coulmas 1996, p. 380). We mention Slavic languages as examples of shallow orthographies, and English and French as languages with deep ones. But a text in a language with a very irregular graphemic representation of phonemes would be very difficult to read. Therefore, we assume that the rank-frequency distributions of graphemes and phonemes are similar—similar enough to be modelled by the same probability distribution.

Furthermore, linguistics (and science in general) aims at formulating general laws and theories, as opposed to language specific. According to Altmann (2025, p. 291), linguistic laws are "general statements in the sense that they apply to all languages (with the same boundary conditions)". The fact that the NHGD is a well-fitting model in several different languages makes it a preliminary candidate for a math-

³In languages that use other scripts, power law-like models are more realistic, see, e.g., the rank-frequency distribution of graphemes in the Garhwali language in Riyal *et al.* (2016). The language uses the Devanagari script, i.e., an abugida (see Daniels 1996, p. 4), which has much larger inventory (247 graphemes according to Riyal *et al.* 2016, as opposed to several tens typical for an alphabet).

ematical formulation of another linguistic law, that of the phoneme rank-frequency distributions.

Second, Wimmer and Altmann (2005) derived a very general model within the theoretical framework of so-called synergetic linguistics (Köhler 1986, 2005). From a mathematical point of view, the assumption is that the probability of a class x is a function of its neighbouring class $x - 1$, i.e., that

$$(3) \quad P_x = g(x)P_{x-1}.$$

The functional relation (3) can be seen as a difference equation. Wimmer and Altmann (2005) provide a solution of this equation for

$$(4) \quad g(x) = 1 + a_0 + \frac{a_1}{(x + b_1)^{c_1}} + \frac{a_2}{(x + b_2)^{c_2}},$$

with a_0 , a_1 , a_2 , b_1 , b_2 , c_1 , and c_2 being parameters. Solutions for particular choices of the parameters are models for different properties of language units, such as length and frequency. The parameters are interpreted (in very general terms) as requirements of speaker and hearer⁴ that meet at a point of an equilibrium. Distributions (1) and (2) are also special cases of this approach, see Wimmer and Altmann (2005, p. 800). The NHGD is thus embedded into a general framework of mathematically expressed linguistic laws (or of empirically corroborated regularities in language, at the very least).

The goodness of fit of the NHGD was evaluated in terms of the determination coefficient

$$(5) \quad R^2 = 1 - \frac{\sum_{i=1}^n (f_i - NP_i)^2}{\sum_{i=1}^n (f_i - \bar{f})^2},$$

with n being the number of categories (phonemes in this case), f_i the observed frequency in the i -th category, $\bar{f}$ the mean of the observed frequencies (i.e., $\bar{f} = \sum_{i=1}^n f_i / n$), P_i the probability of the i -th category predicted by the model, and N the sample size. The advantage of R^2 is that it is not sensitive to the sample size if relative frequencies are fixed,⁵ and the Belarusian, Macedonian, and Russian data were

⁴ Again, words speaker and hearer must be understood very generally (e.g., the interpretation applies also to written texts).

⁵ It differs in this aspect from the χ^2 statistic used in the Pearson χ^2 goodness-of-fit test.

published as proportions, not as absolute frequencies (see Section 5). The fit is usually considered satisfactory if $R^2 \geq 0.9$. See Mačutek and Wimmer (2013) for an overview of the properties of R^2 and other rules of thumb used in quantitative linguistics to evaluate goodness of fit of mathematical models.

5

LANGUAGE MATERIAL

The NHGD is fitted to phoneme frequencies from 24 languages, representing 15 different language families (one of them being language isolate). In addition, there is one English-based creole language. The list of languages is presented in Table 1. Together with information on language, its Glottolog code,⁶ and the language family, we also add the source of data. The data are either taken from older publications (if the last column of Table 1 contains only the reference), or we use our own data from translations of parts of the Bible (the New Testament or the Gospel of Mark) available online (in that case, we refer to literature for phoneme inventories and the grapheme-to-phoneme transcription rules, see tables in Appendix B). Links to the Bible translations can be found in Appendix A (if the webpage has only links to individual chapters, the link to the first chapter of the Gospel of Matthew is given). Languages with simple grapheme-to-phoneme relations are chosen to minimize the risk of errors in the grapheme-to-phoneme conversions.

As older sources contain data from texts (as opposed to dictionaries), we consider in this paper exclusively word tokens, not word types. Data from already published sources are taken, with some exceptions, as they appear in the cited works. The exceptions are as follows.

First, phoneme frequencies in Belarusian, Macedonian, and Russian are presented as percentages, not as absolute frequencies. We adjusted the data so that we have artificial samples with the size of approximately 10,000 (due to rounding, the sums of the relative frequencies are not equal exactly to 1).

Second, Slovak has both syllabic and non-syllabic /l/ and /r/. Bosák (1965) takes the syllabicity of /l/ and /r/ into account, i.e.,

⁶<https://glottolog.org/>

Table 1: List of languages

language (Glottolog)	language family	source of data
Aymara (ayma1253)	Aymaran	New Testament
Bilua (bilu1245)	Central Solomon	New Testament
Belarusian (bela1254)	Indo-European	Kelih (2013)
Bola (bola1250)	Austronesian	New Testament
Bulgarian (bulg1262)	Indo-European	Marinova and Marinov (1964)
Cabécar (cabe1245)	Chibchan	González Campos (2024)
Czech (czec1258)	Indo-European	Ludvíková and Königová (1967)
Finnish (finn1318)	Uralic	Gospel of Mark
Gagauz (gaga1249)	Turkic	New Testament
Hixkaryana (hixk1239)	Cariban	New Testament
Jamamadi (jama1281)	Arawan	New Testament
Kadiwéu (kadi1248)	Guaicuruan	New Testament
Kashibo (cash1251)	Pano-Tacanan	New Testament
Macedonian (mace1250)	Indo-European	Džukeski (1955)
Pitjantjatjara (pitj1243)	Pama-Nyungan	New Testament
Rapanui (rapa1244)	Austronesian	New Testament
Rotokas (roto1249)	North Bougainville	New Testament
Russian (russ1263)	Indo-European	Kučera and Monroe (1968)
Slovak (slov1269)	Indo-European	Bosák (1965)
Slovene (slov1268)	Indo-European	Kelih and Zörnig (2022)
Tok Pisin (tokp1240)	creole (English-based)	New Testament
Udmurt (udmu1245)	Uralic	Gospel of Mark
Warao (wara1303)	language isolate	New Testament
Yaqui (yaqu1251)	Uto-Aztecan	New Testament

he has two categories for both of the two consonants (syllabic and non-syllabic /l/, syllabic and non-syllabic /r/). We disregarded this distinction and merged syllabic and non-syllabic positions of /l/ and /r/ (i.e., we have only one category for /l/ and one for /r/).

For Udmurt, we have two versions of phoneme frequencies, as there are Russian loanwords with Russian phonemes (that do not occur in Udmurt) in the Udmurt Bible translation. We decided to consider both the "pure" version (without foreign phonemes; whole words containing the foreign phonemes are disregarded) and the version as it appears in the translations. This approach gives us the opportunity to check the influence of the phoneme inventory size on the parameters of the model (see Section 6).

Complete datasets (from older sources as well as the ones analyzed for the first time in this paper) are in Tables 3-27 in Appendix B.

6

RESULTS

If we accept the usual threshold ($R^2 \geq 0.9$), the NHGD achieves a satisfactory fit for all languages under analysis. The fitted parameter values of the NHGD and sample sizes are presented in Table 2. The NHGD thus can be considered a candidate for a new statistical language universal, at least in a very general sense (see, e.g., Zhou *et al.* 2025, where the distribution of sentence length is spoken of as a language universal).

Russian clearly has the lowest value of the determination coefficient, with $R^2 = 0.9165$. However, there are two facts to notice. First, two vowels, /i/ and /i/, are merged into one category in the data from Kučera and Monroe (1968). Second, /i/ ranks second in the rank-frequency distribution, and there is a huge gap between the second and third most frequent phoneme (see Table 20). If /i/ and /i/ are considered two different phonemes as e.g. in Yanushevskaya and Bunčić (2015), the frequency of /i/ would be lower and the decrease in the rank-frequency distribution smoother. Consequently, one can expect a better fit in that case.

Table 2: Values of fitted parameters of the negative hypergeometric distribution (K , M , n), determination coefficient R^2 , sample size N (RF if source presents only relative frequencies), relative frequency of most frequent phoneme fr_1 , and relative mean rank RMR . Udmurt (RUS) denotes Udmurt with three Russian phonemes

language	K	M	n	R^2	N	fr_1	RMR
Aymara	2.569	0.474	31	0.9763	800,561	0.254	0.348
Belarusian	1.843	0.457	42	0.9519	RF	0.190	0.479
Bilua	4.345	0.720	39	0.9914	1,043,938	0.165	0.340
Bola	1.901	0.515	17	0.9855	823,757	0.241	0.543
Bulgarian	4.898	0.827	41	0.9871	18,212	0.138	0.343
Cabécar	3.326	0.893	26	0.9928	111,710	0.111	0.531
Czech	2.964	0.799	37	0.9879	156,580	0.097	0.542
Finnish	5.404	1.010	36	0.9921	58,479	0.101	0.328
Gagauz	4.109	0.802	40	0.9941	669,905	0.120	0.399
Hixkaryana	3.850	0.880	22	0.9796	1,603,302	0.146	0.460
Jamamadi	2.621	0.587	19	0.9903	936,438	0.238	0.443
Kadiwéu	3.797	0.814	22	0.9862	1,244,533	0.167	0.458
Kashibo	4.601	0.861	20	0.9776	1,129,601	0.193	0.421
Macedonian	3.840	0.802	30	0.9869	RF	0.139	0.428
Pitjantjatjara	2.664	0.543	22	0.9821	1,245,660	0.247	0.412
Rapanui	2.734	0.692	19	0.9892	977,691	0.192	0.506
Rotokas	3.924	0.871	15	0.9721	1,322,879	0.214	0.434
Russian	2.326	0.509	37	0.9165	RF	0.176	0.492
Slovak	3.450	0.745	43	0.9514	19,842	0.104	0.455
Slovene	3.867	0.989	38	0.9954	150,257	0.073	0.517
Tok Pisin	3.873	0.909	26	0.9985	1,043,460	0.125	0.472
Udmurt	3.967	0.981	32	0.9934	54,043	0.090	0.496
Udmurt (RUS)	4.366	1.000	35	0.9997	54,947	0.089	0.457
Warao	1.720	0.493	15	0.9574	891,775	0.255	0.567
Yaqui	3.120	0.765	24	0.9657	1,274,595	0.157	0.492

We present the results of fitting for Bulgarian and Hixkaryana also graphically in Figures 1 and 2, respectively. The sample sizes for these languages are the smallest (Bulgarian) and largest (Hixkaryana) among the languages under study.

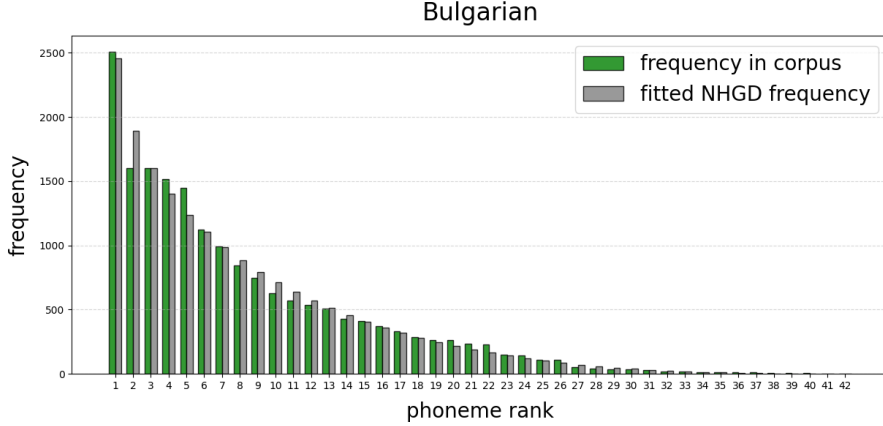


Figure 1: Phoneme rank-frequency distribution in Bulgarian (data and model)

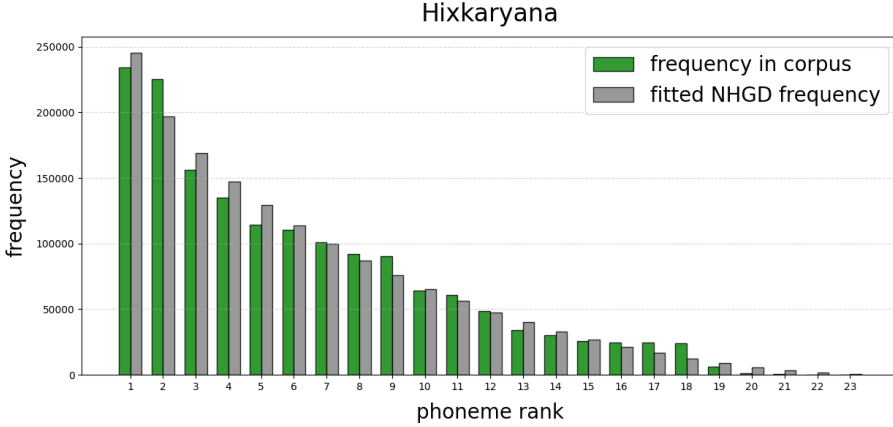


Figure 2: Phoneme rank-frequency distribution in Hixkaryana (data and model)

For the time being, only parameter n can be interpreted without any doubt, $n + 1$ is the inventory size (the shift by 1 is caused by the fact that according to the usual definition, the NHGD has the support $0, 1, \dots, n$ rather than $1, \dots, n$). However, Table 2 contains

also two empirical statistical characteristics that provide some preliminary hints towards a possible linguistic interpretation of the other two model parameters. The relative frequency of the most frequent phoneme is obvious. We define the relative mean rank *RMR* as

$$(6) \quad RMR = \frac{\sum_{i=1}^I \frac{f_i}{N} i - 1}{\frac{I+1}{2} - 1},$$

with N being the sample size, $\frac{f_i}{N}$ the empirical relative frequency of the i -th most frequent phoneme, and I the phoneme inventory size. The *RMR* can, in theory, take values from 0 to 1. These extreme values of *RMR* represent two extreme cases of a rank-frequency distribution. If a text contains only one phoneme and all other phonemes from the inventory have zero frequencies, the mean rank (i.e., $\sum_{i=1}^I \frac{f_i}{N}$) is equal to 1 and it holds $RMR = 0$. On the other extreme,⁷ if all frequencies are equal, we have the uniform distribution with the support $1, 2, \dots, I$; its mean is $\frac{I+1}{2}$ and in such a case $RMR = 1$. The *RMR* characterizes the uniformity of a distribution; or, from the opposite point of view, dominance of the most frequent phonemes. If there are several phonemes with high frequencies and some marginal phonemes, the value of the *RMR* will be closer to 1. If a language exploits its phonemes relatively uniformly, the value will be small.

One can observe a strong correlation between the frequency of the most frequent phoneme f_{r_1} and the fitted values of the parameter M (the Pearson correlation coefficient $r = -0.802$) in Figure 3.

The correlation between the *RMR* and the fitted values of the parameter M is admittedly weaker ($r = -0.594$), but the decreasing trend is still visible (see Figure 4).

There is also a strong correlation between the fitted values of the parameters K and M , with $r = 0.842$ (see Figure 5). It suggests that the fit could remain good if one of the parameters could be expressed as a linear function of the other. However, that would bring into play another two parameters, those of the linear regression. Such an approach

⁷ We remind that we consider rank-frequency distributions, i.e., the sequence of frequencies must be non-decreasing (consequently, the mean rank cannot be greater than $\frac{I+1}{2}$).

Figure 3:
Relationship
between highest
frequency fr_1
and parameter
 M of NHGD
distribution

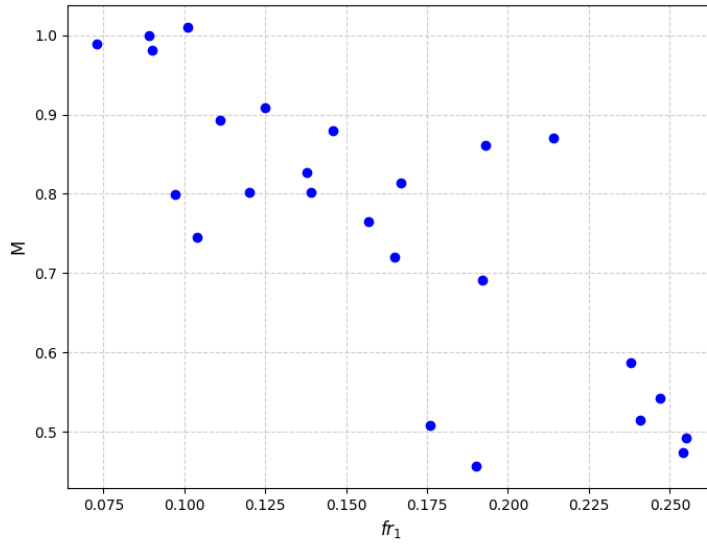
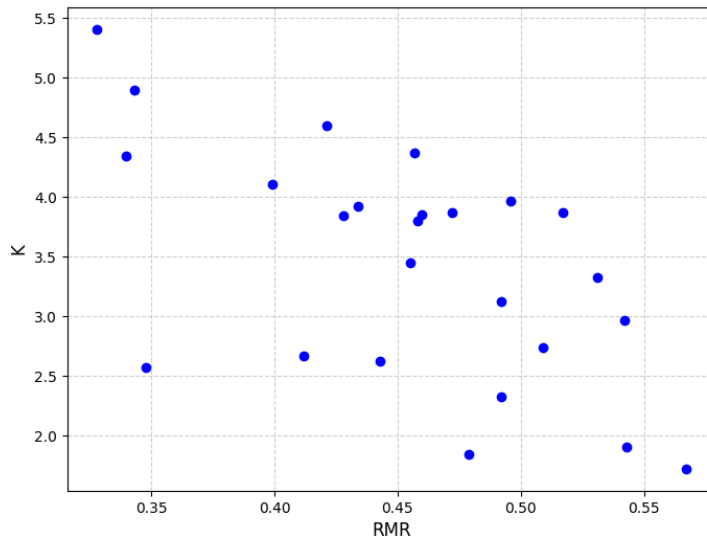


Figure 4:
Relationship
between RMR
and parameter K
of NHGD
distribution



can contribute to the scientific explanation only if the parameters of the linear regression themselves can be interpreted linguistically.

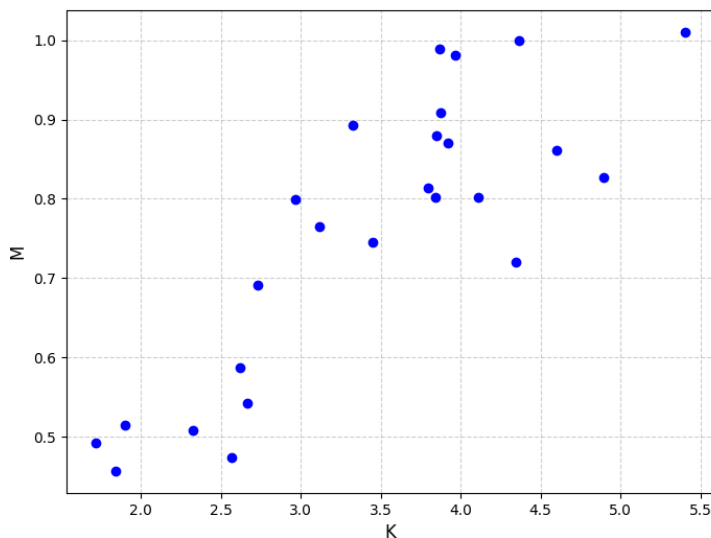


Figure 5:
Relationship
between
parameters K
and M of NHGD
distribution

Phoneme inventory size also has an influence, although its character is not yet clear (there is no obvious relationship between inventory sizes and the fitted parameter values). But the two versions of Udmurt ("pure" and with additional three Russian phonemes) differ quite clearly in the values of the parameter K . See Grzybek (2007b) for a similar study of Russian graphemes. Officially, <ë> is a letter of the Russian alphabet, but it is almost always replaced with <e>. There are thus two versions of grapheme frequencies that differ in their inventory sizes by 1. Grzybek (2007b) showed that this difference is mirrored with a difference of the fitted parameter values in the NHGD.

CONCLUSION AND DISCUSSION

7

Rank-frequency phoneme distributions from 24 typologically diverse languages display one pattern of statistical behaviour, all of them can

be modelled by the NHGD. The NHGD (applied also to modelling grapheme ranked frequencies, a "close relative" of phonemes) can be, at least preliminarily, considered a general model and a candidate for a new statistically understood language universal.

Naturally, phoneme frequencies from more languages must be investigated before one can be more certain. While we include in this paper some languages with very small phoneme inventories (Bola, Rotokas, and Warao), languages from the other side of the spectrum must be studied in future (e.g., recently extinct Ubykh had 84 consonant and 3 vowel phonemes, see Fenwick 2011; some analyses of Taa suggest even more than 160 consonant phonemes, depending on how consonant clusters containing clicks are treated, see Naumann 2016).

In this paper, we present some regularities in the parameter behaviour. However, in order to be able to speak of a language law, a more linguistically straightforward interpretation of the model parameters must be achieved. The interpretation does not have to be limited to the characteristics of the phoneme frequencies themselves. Phoneme frequencies are not isolated from other language properties (see, e.g., a synergetic language model by Köhler 2005 in which language properties are interconnected). Phonotactics (which limits frequencies of phonemes that are not allowed to combine with many others) and distinctive features (very similar phonemes, such as short and long vowels of the same quality, should not have similar frequencies, as they are easy to be confused) come to mind in the first place. Properties of syllables and words will have to be considered as well.

Finally, the question of a sufficient sample size must be solved. Most probably, it will strongly depend on the phoneme inventory size. Homogeneity of samples could serve as a useful criterion, e.g., if increasing a sample would not lead to significant changes in phoneme frequencies, such a sample can be considered sufficient.

ACKNOWLEDGEMENT

Funded by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04- 00748

APPENDIX A: LINKS TO BIBLE
TRANSLATIONS

A

Aymara

<https://live.bible.is/bible/AYMBSB/MAT/1>

Bilua

<https://www.bible.com/en-GB/bible/2979/MAT.1.BLBNT>

Bola

<https://www.bible.com/bible/333/MAT.1.BNP>

Finnish

<https://www.wordproject.org/bibles/fi/41/1.htm>

Gagauz

<https://www.bible.com/en-GB/bible/2554/MAT.1.GAGNTL>

Hixkaryana

<https://www.bible.com/bible/587/MAT.1.HIX>

Jamamadi

<https://live.bible.is/bible/JAAJAM/MAT/1>

Kadiwéu

<https://www.bible.com/bible/584/MAT.1.KBC>

Kashibo

https://ebible.org/pdf/cbrNT/cbrNT_all.pdf

Pitjantjatjara

https://ebible.org/pdf/pjt/pjt_a4.pdf

Rapanui

<https://www.bible.com/bible/2164/MAT.1.RAP>

Rotokas

<https://live.bible.is/bible/R00WBT/MAT/1>

Tok Pisin

<https://www.bible.com/bible/3481/MAT.1.TPIPNG>

Udmurt

http://finugorbib.com/bible/udmurt/40_Mat01_na.html

Warao

<https://live.bible.is/bible/wbabiv/mat/1>

Yaqui

<https://www.bible.com/bible/799/MAT.1.YAQ>

B APPENDIX B: PHONEME FREQUENCIES

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	a	203,003	17	tʃ	10,640
2	i	73,937	18	ɲ	10,616
3	u	55,171	19	l	7,484
4	k	53,944	20	t'	4,752
5	n	53,871	21	i:	4,463
6	t	51,749	22	k ^h	4,409
7	s	40,680	23	q ^h	3,136
8	p	35,857	24	λ	2,420
9	r	30,768	25	tʃʳ	1,823
10	m	27,888	26	k'	1,391
11	x	27,542	27	t ^h	1,273
12	χ	23,801	28	q'	946
13	w	22,437	29	p ^h	943
14	j	18,722	30	tʃ ^h	675
15	q	12,950	31	u:	473
16	a:	12,436	32	p'	361

Table 3:
Phoneme frequencies
in Aymara (phoneme
inventory from Hardman
et al. 2001)

Table 4:
Phoneme frequencies
in Belarusian (phoneme
inventory from Burlyka
et al. 1989)

rank	phoneme	frequency	rank	phoneme	frequency
1	a	1,879	23	s ^j	158
2	i	547	24	l	143
3	ε	448	25	ʂ	142
4	ɨ	445	26	z	140
5	ɔ	432	27	x	129
6	j	425	28	v ^j	106
7	r	412	29	b	105
8	n	386	30	(ts	105
9	u	364	31	k ^j	79
10	t	359	32	z _ɫ	74
11	k	302	33	m ^j	74
12	d	283	34	(dz ^j	59
13	n ^j	259	35	p ^j	53
14	s	249	36	z ^j	43
15	m	247	37	b ^j	38
16	p	246	38	ɣ ^j	16
17	ɫ ^j	214	39	f	13
18	v	208	40	ɸ ^j	8
19	w	195	41	x ^j	6
20	ɣ	177	42	(dʒ _ɫ	3
21	(ts ^j	170	43	(dz	2
22	(tʂ	160			

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	a	172,700	21	d	10,907
2	o	106,616	22	ia	8,898
3	k	100,152	23	ou	7,174
4	e	87,465	24	j	5,931
5	i	63,918	25	oa	4,929
6	t	52,524	26	ui	4,274
7	l	50,844	27	ae	3,727
8	m	47,670	28	ea	3,377
9	u	41,289	29	ue	3,194
10	β	39,720	30	eu	2,671
11	s	37,920	31	au	2,448
12	n	35,700	32	oi	1,736
13	r	33,308	33	ie	1,553
14	p	25,180	34	iu	1,523
15	b	16,606	35	ua	1,402
16	ɲ	16,473	36	oe	1,323
17	io	13,140	37	ei	1,011
18	g	11,949	38	ao	684
19	ai	11,872	39	eo	646
20	z	11,150	40	uo	334

Table 5:
Phoneme frequencies
in Bilua (phoneme
inventory from Obata
2003)

Table 6:
Phoneme frequencies
in Bola (phoneme
inventory from van den
Berg and Wiebe 2019)

rank	phoneme	frequency
1	ɑ	198,418
2	i	107,846
3	u	59,914
4	r	56,165
5	ɛ	53,695
6	n	46,938
7	ɔ	44,564
8	t	38,271
9	k	36,480
10	m	36,118
11	l	32,588
12	v	29,314
13	ɣ	25,957
14	β	14,109
15	g	13,215
16	ŋ	12,472
17	p	10,790
18	d	6,903

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	a	2,508	22	g	230
2	i	1,601	23	f	146
3	ε	1,600	24	x	140
4	ɔ	1,518	25	ʒ	109
5	t	1,450	26	ts̃	106
6	n	1,125	27	kʲ	51
7	s	994	28	lʲ	38
8	r	843	29	nʲ	35
9	v	744	30	tʲ	35
10	k	626	31	gʲ	27
11	l	569	32	rʲ	18
12	d	538	33	vʲ	16
13	p	507	34	bʲ	13
14	ə	425	35	dʲ	11
15	m	409	36	xʲ	11
16	z	368	37	mʲ	9
17	u	328	38	sʲ	5
18	ʃ	285	39	pʲ	4
19	j	264	40	ts̃ʲ	4
20	b	263	41	zʲ	2
21	tʃ̃	235	42	dʒ̃	2

Table 7:
Phoneme frequencies
in Bulgarian (phoneme
inventory from Scatton
1993)

Table 8:
Phoneme frequencies
in Cabécar (phoneme
inventory from González
Campos 2024)

rank	phoneme	frequency
1	a	12,393
2	k	10,242
3	d	8,942
4	i	8,558
5	ã	7,870
6	t	6,580
7	u	6,265
8	b	5,821
9	ɣ	5,795
10	s	5,355
11	h	4,909
12	ɛ	4,660
13	l	4,237
14	ĩ	3,360
15	ɾ	3,321
16	ḍ͡ʒ	3,124
17	ẽ	2,197
18	ʃ	1,963
19	ũ	1,556
20	ʊ	1,027
21	ɔ	999
22	õ	878
23	p	854
24	k ^h	574
25	t ^h	147
26	p ^h	65
27	ŋ	18

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	e	15,236	20	z	3,027
2	o	10,739	21	b	2,856
3	a	10,379	22	ts	2,145
4	ɪ	9,358	23	fi	1,982
5	t	7,415	24	e:	1,939
6	s	7,316	25	ʃ	1,919
7	n	7,092	26	r̥	1,878
8	l	6,760	27	x	1,861
9	i:	6,275	28	tʃ	1,495
10	k	6,260	29	ou	1,366
11	v	6,038	30	c	1,331
12	m	6,333	31	ʒ	1,301
13	r	5,627	32	f	1,139
14	p	4,838	33	u:	905
15	j	4,502	34	ʝ	747
16	u	4,289	35	g	633
17	d	4,263	36	au	392
18	a:	3,733	37	o:	103
19	ɲ	3,108	38	eu	0

Table 9:
Phoneme frequencies
in Czech (phoneme
inventory from Šimáčková
et al. 2012)

Table 10:
Phoneme frequencies
in Finnish (phoneme
inventory from Karlsson
2015)

rank	phoneme	frequency	rank	phoneme	frequency
1	ɑ	5,930	20	t:	716
2	n	5,702	21	y	642
3	i	5,585	22	e:	463
4	e	4,914	23	d	430
5	t	4,259	24	i:	399
6	s	4,100	25	æ:	382
7	æ	3,378	26	s:	252
8	k	2,755	27	u:	239
9	o	2,570	28	n:	228
10	u	2,295	29	ŋ	225
11	h	2,149	30	ø	218
12	j	1,843	31	k:	146
13	l	1,796	32	m:	126
14	m	1,514	33	o:	73
15	v	1,514	34	p:	66
16	p	981	35	y:	35
17	l:	971	36	r:	33
18	r	802	37	ø:	12
19	ɑ:	768			

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	ɑ	80,665	22	g	7,966
2	n	66,857	23	tʃ	6,649
3	i	57,383	24	p	6,399
4	r	45,662	25	dʒ	5,655
5	e	39,191	26	v	5,422
6	l	38,008	27	ø	5,397
7	d	34,301	28	ɑ:	4,553
8	ʊ	29,969	29	i:	3,352
9	s	27,658	30	f	2,297
10	k	26,755	31	e:	2,277
11	m	22,571	32	y:	1,732
12	u	18,222	33	ə	1,478
13	o	18,078	34	o:	1,122
14	æ	18,036	35	æ:	884
15	t	17,968	36	ə:	620
16	b	16,717	37	ʊ:	518
17	z	14,243	38	u:	359
18	j	14,239	39	ø:	59
19	h	9,503	40	ts	28
20	y	8,908	41	ʒ	4
21	ʃ	8,200			

Table 11:
Phoneme frequencies
in Gagauz (phoneme
inventory from Pokrovskaja
1964)

Table 12:
Phoneme frequencies
in Hixkaryana (phoneme
inventory from Derbyshire
1979)

rank	phoneme	frequency
1	ɔ	233,984
2	æ	225,519
3	h	156,116
4	e	135,226
5	k	114,563
6	n	110,274
7	m	101,177
8	ɾ	92,065
9	uu	90,428
10	t	64,069
11	j	60,557
12	w	48,533
13	tʃ	34,052
14	ʃ	30,013
15	u	25,772
16	ɲ	24,767
17	s	24,465
18	ɭ	23,764
19	p	5,935
20	ϕ	1,082
21	b	613
22	d	268
23	ʒ	60

rank	phoneme	frequency
1	a	223,096
2	i	135,092
3	o	84,421
4	n	71,162
5	e	68,856
6	m	54,991
7	j	47,777
8	t	45,275
9	d	33,357
10	k	31,565
11	h	28,689
12	b	27,677
13	s	25,417
14	r	24,185
15	w	21,410
16	ϕ	10,308
17	a:	2,200
18	e:	637
19	i:	290
20	o:	33

Table 13:
Phoneme frequencies
in Jamamadi (phoneme
inventory from Dixon and
Vogel 2004)

Table 14:
Phoneme frequencies
in Kadiwéu (phoneme
inventory from Griffiths
2002)

rank	phoneme	frequency
1	i	208,153
2	a	169,180
3	o	128,836
4	e	115,851
5	n	90,415
6	ɔ	83,581
7	d	83,306
8	t	62,400
9	g	36,820
10	l	36,694
11	m	34,473
12	ɗ̥	33,866
13	k	31,050
14	a:	25,656
15	w	25,535
16	q	19,122
17	e:	12,585
18	b	12,406
19	j	10,056
20	p	9,411
21	i:	6,702
22	tʃ̥	5,133
23	o:	3,302

rank	phoneme	frequency
1	a	217,779
2	i	166,157
3	n	161,474
4	u	99,072
5	k	92,032
6	ɨ	77,624
7	ʔ	46,350
8	m	46,183
9	β	34,426
10	ɕ	34,392
11	t	27,584
12	s	26,008
13	k ^w	20,422
14	r	20,189
15	p	18,073
16	o	9,984
17	ɲ	9,631
18	ʈs	7,829
19	ʃ	7,196
20	ʈʃ	3,924
21	e	3,272

Table 15:
Phoneme frequencies
in Kashibo (phoneme
inventory from
Zariquiey Biondi 2018)

Table 16:
Phoneme frequencies
in Macedonian (phoneme
inventory from Friedman
1993)

rank	phoneme	frequency	rank	phoneme	frequency
1	a	1,389	17	ʃ	208
2	ɛ	1,107	18	z	175
3	ɔ	1,018	19	b	145
4	i	790	20	j	145
5	t	705	21	tʃ	108
6	n	636	22	f	69
7	s	443	23	c	61
8	k	404	24	ʒ	52
9	r	400	25	ts	52
10	d	361	26	ʝ	12
11	v	353	27	ɲ	12
12	ʎ	340	28	dʒ	8
13	m	296	29	x	3
14	p	249	30	l	2
15	u	243	31	dʒ̥	2
16	g	222			

rank	phoneme	frequency
1	e	352,414
2	u	219,446
3	k	112,103
4	ɪ	100,634
5	p	75,244
6	ŋ	73,884
7	c	69,600
8	l	60,616
9	n	59,109
10	ɲ	58,358
11	r	45,924
12	m	37,443
13	w	35,875
14	ɹ	29,517
15	t	25,695
16	j	19,163
17	ʈ	15,997
18	ŋ	13,980
19	ɭ	11,776
20	λ	7,161
21	e:	1,063
22	ɪ:	437
23	ʊ:	221

Table 17:
Phoneme frequencies
in Pitjantjatjara (phoneme
inventory from Tabain and
Butcher 2014)

Table 18:
Phoneme frequencies
in Rapanui (phoneme
inventory from Du Feu
1996)

rank	phoneme	frequency
1	a	188,063
2	e	112,758
3	t	88,175
4	i	87,691
5	ʔ	77,607
6	o	67,930
7	u	52,716
8	r	50,409
9	h	48,424
10	k	44,925
11	m	32,311
12	n	32,218
13	a:	21,915
14	ŋ	16,849
15	p	16,736
16	o:	13,079
17	β	8,056
18	u:	7,448
19	i:	5,590
20	e:	4,791

rank	phoneme	frequency
1	a	283,232
2	o	184,277
3	r	176,758
4	i	147,568
5	v	133,923
6	e	91,555
7	t	89,541
8	u	85,978
9	p	68,301
10	k	22,383
11	g	14,643
12	a:	11,000
13	i:	7,856
14	o:	2,686
15	u:	1,905
16	e:	1,273

Table 19:
Phoneme frequencies
in Rotokas (phoneme
inventory from Stebbins
et al. 2018)

Table 20:
Phoneme frequencies
in Russian (phoneme
inventory from
Yanushevskaya and Bunčić
2015)

rank	phoneme	frequency	rank	phoneme	frequency
1	a	17,586	20	d	1,665
2	i	14,132	21	tʃʲ	1,633
3	t	4,266	22	ʃ	1,564
4	j	4,137	23	rʲ	1,381
5	n	4,100	24	z	1,367
6	o	4,026	25	g	1,304
7	u	3,495	26	b	1,096
8	k	3,175	27	dʲ	1,038
9	s	3,093	28	vʲ	1,016
10	v	2,978	29	x	987
11	r	2,907	30	ʒ	951
12	l	2,664	31	f	948
13	e	2,614	32	mʲ	805
14	m	2,317	33	ʦ	562
15	p	2,309	34	kʲ	534
16	nʲ	2,297	35	pʲ	475
17	ʎʲ	2,081	36	bʲ	368
18	tʲ	1,885	37	zʲ	322
19	sʲ	1,863	38	ʃʲ	59

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	a	2,058	23	c	334
2	ɔ	2,056	24	i:	322
3	ɛ	1,656	25	tʃ	242
4	i	1,196	26	ʃ	288
5	t	818	27	ʒ	254
6	s	786	28	ʒ	252
7	v	760	29	ie	176
8	m	710	30	x	158
9	r	700	31	ts	142
10	n	680	32	u:	126
11	l	676	33	ia	88
12	k	654	34	g	76
13	p	616	35	ɛ:	72
14	d	530	36	f	58
15	u	478	37	dʒ	28
16	ɲ	458	38	uo	26
17	λ	440	39	æ	16
18	b	414	40	r:	10
19	j	398	41	dʒ	8
20	fi	390	42	iu	2
21	z	354	43	l:	0
22	a:	336	44	ɔ:	0

Table 21:
Phoneme frequencies
in Slovak (phoneme
inventory from Hanulíková
and Hamann 2010)

Table 22:
Phoneme frequencies
in Slovene (phoneme
inventory, including
graphemic representations
of vowels, from Kelih and
Zörnig 2022)

rank	phoneme	frequency	rank	phoneme	frequency
1	e	10,924	21	ó	2,490
2	o	10,305	22	tʃ	2,177
3	a	9,416	23	g	1,801
4	n	9,028	24	u	1,767
5	i	8,542	25	ʃ	1,743
6	r	7,811	26	ts	1,363
7	l	7,784	27	ô	1,264
8	s	7,716	28	h	1,190
9	j	6,874	29	ú	1,170
10	t	6,511	30	ʒ	1,073
11	v	6,103	31	ə	886
12	p	5,426	32	ê	749
13	á	5,355	33	à	570
14	k	5,251	34	è	512
15	d	4,872	35	è	383
16	m	4,721	36	f	268
17	í	4,172	37	ò	239
18	é	3,662	38	ì	84
19	z	3,384	39	ù	21
20	b	2,650			

rank	phoneme	frequency
1	i	130,169
2	a	107,950
3	l	95,709
4	o	89,608
5	m	79,908
6	e	67,203
7	n	64,335
8	s	57,540
9	t	51,911
10	p	46,978
11	ŋ	38,385
12	k	37,222
13	b	32,118
14	u	25,687
15	r	21,122
16	ai	20,726
17	g	16,691
18	d	14,169
19	j	12,670
20	w	9,128
21	h	6,462
22	au	5,931
23	v	5,866
24	ɖ͡ʒ	3,151
25	ia	1,485
26	f	762
27	oi	574

Table 23:
Phoneme frequencies
in Tok Pisin (phoneme
inventory from Smith
2008)

Table 24:
Phoneme frequencies
in Udmurt (phoneme
inventory from Vakhrushev
and Denisov 1992)

rank	phoneme	frequency	rank	phoneme	frequency
1	i	4,853	17	d	1,242
2	o	4,500	18	v	975
3	a	4,041	19	p	892
4	e	3,871	20	ʃ	881
5	s	3,392	21	g	532
6	i	3,279	22	b	420
7	k	3,025	23	ʌ	397
8	n	2,871	24	ə	391
9	t	2,612	25	ɲ	356
10	u	2,564	26	z'	336
11	l	2,395	27	t͡ɕ	260
12	z	2,334	28	ʒ	182
13	m	1,797	29	d͡ʒ	157
14	ɕ	1,774	30	d'	133
15	r	1,768	31	t͡ʃ	98
16	j	1,611	32	t'	70
			33	d͡ʒ̥	34

Rank-frequency distribution of phonemes across languages

rank	phoneme	frequency	rank	phoneme	frequency
1	i	4,868	19	p	905
2	o	4,551	20	ʃ	881
3	a	4,145	21	g	537
4	e	3,972	22	b	420
5	s	3,450	23	ʌ	401
6	i	3,367	24	ə	391
7	k	3,028	25	ɲ	367
8	n	2,897	26	z'	336
9	t	2,625	27	ʈʂ	262
10	u	2,574	28	ʒ	182
11	l	2,414	29	ɖʒ	157
12	z	2,345	30	d'	134
13	r	1,854	31	ʈʃ	98
14	m	1,815	32	t'	70
15	ç	1,803	33	x	59
16	j	1,715	34	f	44
17	d	1,254	35	ɖʒ	34
18	v	989	36	ʈs	3

Table 25:
Phoneme frequencies
in Udmurt (with three
Russian phonemes /x/, /f/,
and /ʈs/)

Table 26:
Phoneme frequencies
in Warao (phoneme
inventory from
Romero-Figueroa and
Rybka 2023)

rank	phoneme	frequency
1	a	227,003
2	i	81,693
3	ɔ	73,817
4	ɛ	73,225
5	t	64,780
6	n	60,538
7	ɾ	54,316
8	k	51,205
9	h	44,087
10	u	42,871
11	m	35,888
12	b	32,691
13	s	20,748
14	j	17,180
15	w	8,857
16	k ^w	2,876

rank	phoneme	frequency
1	a	200,075
2	e	116,026
3	i	109,317
4	t	98,685
5	ʔ	86,413
6	u	84,003
7	n	81,543
8	k	80,684
9	o	68,106
10	m	64,041
11	h	64,013
12	β	33,749
13	w	29,983
14	j	26,268
15	s	26,241
16	p	21,153
17	a:	16,641
18	tʃ	16,080
19	l	15,408
20	r	14,382
21	e:	9,863
22	b ^w	4,749
23	o:	3,062
24	i:	2,798
25	u:	1,312

Table 27:
Phoneme frequencies
in Yaqui (phoneme
inventory from Trueman
et al. 2018)

REFERENCES

- Gabriel ALTMANN (1972), Status und Ziele der quantitativen Sprachwissenschaft, in Siegfried JÄGER, editor, *Linguistik und Statistik*, volume 6 of *Schriften zur Linguistik*, pp. 1–9, Vieweg, Braunschweig, https://doi.org/10.1007/978-3-322-85868-9_1.
- Gabriel ALTMANN (1973), Mathematische Linguistik, in Walter A. KOCH, editor, *Perspektiven der Linguistik*, pp. 208–232, Kröner, Stuttgart.
- Gabriel ALTMANN (1978), Towards a theory of language, in Gabriel ALTMANN, editor, *Glottometrika 1*, volume 1 of *Quantitative Linguistics*, pp. 1–25, Brockmeyer, Bochum.
- Gabriel ALTMANN (1993), Phoneme counts. Marginal remarks to Pääkkönen's article, in Gabriel ALTMANN, editor, *Glottometrika 14*, volume 53 of *Quantitative Linguistics*, pp. 54–68, WVT, Trier.
- Gabriel ALTMANN (1997), The art of quantitative linguistics, *Journal of Quantitative Linguistics*, 4(1-3):13–22, <https://doi.org/10.1080/09296179708590074>.
- Gabriel ALTMANN (2025), Fashion and truth in linguistics, in Emmerich KELIH, Ján MAČUTEK, and Michaela KOŠČOVÁ, editors, *Quantification in Linguistics and Text Analysis: Selected Papers of Gabriel Altmann*, pp. 287–294, De Gruyter Mouton, Berlin & Boston, <https://doi.org/10.1515/9783111351605-020>.
- Gabriel ALTMANN and Werner LEHFELDT (1973), *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren*, Fink, Munich.
- Gabriel ALTMANN and Werner LEHFELDT (1980), *Einführung in die quantitative Phonologie*, volume 7 of *Quantitative Linguistics*, Brockmeyer, Bochum.
- Karl-Heinz BEST (2005), Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten, *Glottometrics*, 11:9–31, <https://glottometrics.iqla.org/wp-content/uploads/2021/06/g11zeit.pdf>.
- Ján BOŠÁK (1965), Frequency of phonemes and letters in Slovak and numerical expression of some phonemic relations, *Jazykovedný časopis*, 16(2):120–135, https://www.juls.savba.sk/ediela/jc/1965/2/JC_1965_2_LQ.pdf.
- Anton S. BUDILOVIČ (1883), *Načertanie cerkovnoslavjanskoj grammatiki, priměnitel'no k obščej teorii russkago i drugich rodstvennyh jazykov* [An outline of Church Slavonic grammar, in relation to the general theory of Russian and other related languages], Koval'skij, Warsaw.
- Margarita I. BURLAKOVA, Tat'jana M. NIKOLAEVA, Dmitrij M. SEGAL, and Vladimir N. TOPOROV (1962), Strukturnaja tipologija i slavjanskoe jazykoznanie [Structural typology and Slavic linguistics], in Tat'jana N. MOLOŠNAJA, editor, *Strukturno-tipologičeskie issledovanija* [Structural-typological studies], pp. 3–18, Nauka, Moscow.

Vasil' R. BURLYKA, Liliya C. VYGONNAJA, Georgij V. LOSIK, and Aljaksandr I. PADLUŽNY (1989), *Fanetyka belaruskaj litaraturnaj movy [Phonetics of Belarusian literary language]*, Navuka i tehnika, Minsk.

Florian COULMAS (1996), *The Blackwell encyclopedia of writing systems*, Blackwell, Oxford, <https://doi.org/10.1002/9781118932667>.

Peter T. DANIELS (1996), The study of writing systems, in Peter T. DANIELS and William BRIGHT, editors, *The World's Writing Systems*, pp. 3–17, Oxford University Press, New York & Oxford.

Desmond C. DERBYSHIRE (1979), *Hixkaryana*, volume 1 of *Lingua Descriptive Studies*, North-Holland Publishing Company, Amsterdam.

Mitrofan A. DIKAREV (1891), Očerok voronežskogo meščanskogo govora sravnitel'no s ukraino-russkim narečiem (opyt statističeskogo issledovanija zvukovyh javlenij) [A study of the Voronezh townsman dialect compared to the Ukrainian-Russian dialect (an attempt at a statistical analysis of sound phenomena)], in Mitrofan A. DIKAREV, editor, *Voronžeskij etnografičeskij sbornik [Voronezh Ethnography Volume]*, pp. 1–68, Izdanie Voronežskago Gubernskago Statističeskago Komiteta, Voronezh.

Robert M. W. DIXON and Alan R VOGEL (2004), *The Jarawara language of Southern Amazonia*, Oxford University Press, Oxford, <https://doi.org/10.1093/oso/9780199270675.001.0001>.

Veronica DU FEU (1996), *Rapanui*, Descriptive Grammars, Routledge, London & New York.

Aleksandar DŽUKESKI (1955), Procentualnata sostojba na glasovite vo makedonskiot literaturnen jazik [Percentual composition of sounds in the Macedonian literary language], *Makedonski jazik*, 6(2):165–200.

Gertraud FENK-OCZLON and August FENK (2008), Complexity trade-offs between the subsystems of language, in Matti MIESTAMO, Kaius SINNEMÄKI, and Fred KARLSSON, editors, *Language Complexity: Typology, contact, change*, pp. 43–65, Benjamins, Amsterdam & Philadelphia.

Rhona S. H. FENWICK (2011), *A grammar of Ubykh*, volume 19 of *LINCOM Studies in Caucasian Linguistics*, LINCOM Europa, Munich.

Giuseppe FRANCESCATO (1962), Notes on relevant features with low functional yield, *Lingua*, 11:118–127, [https://doi.org/10.1016/0024-3841\(62\)90018-9](https://doi.org/10.1016/0024-3841(62)90018-9).

Victor I. FRIEDMAN (1993), Macedonian, in Bernard COMRIE and Greville G. CORBETT, editors, *The Slavonic Languages*, Routledge Language Family Descriptions, pp. 249–305, Routledge, London & New York.

Ernst FÖRSTEMANN (1846), Über die numerischen Lautverhältnisse im Deutschen, *Germania. Neues Jahrbuch der Berliner Gesellschaft für Deutsche Sprache und Altertumskunde*, 7:83–90.

Ernst FÖRSTEMANN (1852), Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen, *Zeitschrift für Vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen*, 1:163–179.

Guillermo GONZÁLEZ CAMPOS (2024), *Determinación de las frecuencias fonémicas del idioma cabécar mediante un corpus electrónico*, volume 5 of *Cuadernos de investigación de la Sede del Atlántico*, Universidad de Costa Rica (Sede del Atlántico), Turrialba.

Glyn GRIFFITHS (2002), *Dicionário de língua kadiwéu*, SIL, Cuiabá.

Peter GRZYBEK (2007a), On the systematic and system-based study of grapheme frequencies. A re-analysis of German letter frequencies, *Glottometrics*, 15:82–91, <https://glottometrics.iqla.org/wp-content/uploads/2021/06/g15zeit-1.pdf>.

Peter GRZYBEK (2007b), What a difference an «E» makes. Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen, in Peter DEUTSCHMANN, editor, *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag*, pp. 105–128, Praesens, Vienna.

Peter GRZYBEK and Emmerich KELIH (2003), Graphemhäufigkeiten (am Beispiel des Russischen)). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen, *Anzeiger für Slavische Philologie*, 31:131–162, <https://unipub.uni-graz.at/download/pdf/10353284.pdf>.

Peter GRZYBEK and Emmerich KELIH (2005a), Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph ('), in Gabriel ALTMANN, Viktor LEVICKIJ, and Valentina PEREBEJNIS, editors, *Problems of Quantitative Linguistics*, pp. 159–179, Ruta, Černivci.

Peter GRZYBEK and Emmerich KELIH (2005b), Towards a general model of grapheme frequencies for Slavic languages, in Radovan GARABÍK, editor, *Computer Treatment of Slavic and East European Languages*, pp. 73–87, Veda, Bratislava.

Peter GRZYBEK, Emmerich KELIH, and Gabriel ALTMANN (2004), Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung, *Anzeiger für Slavische Philologie*, 32:25–54, <https://unipub.uni-graz.at/download/pdf/10309880.pdf>.

Peter GRZYBEK, Emmerich KELIH, and Gabriel ALTMANN (2006a), Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen, in Ružena KOZMOVÁ, editor, *Sprache und Sprachen im mitteleuropäischen Raum. Vorträge der Internationalen Linguistik-Tage, Trnava 2005*, pp. 641–664, University of Ss. Cyril and Methodius, Trnava.

Peter GRZYBEK, Emmerich KELIH, and Ernst STADLOBER (2006b), Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein

- Beitrag zur theoretischen Begründung der sog. Schriftlinguistik, *Anzeiger für Slavische Philologie*, 34:41–74,
<https://unipub.uni-graz.at/download/pdf/10531599.pdf>.
- Peter GRZYBEK, Emmerich KELIH, and Ernst STADLOBER (2009), Slavic letter frequencies: A common discrete model and regular parameter behavior?, in Reinhard KÖHLER, editor, *Issues in Quantitative Linguistics*, volume 5 of *Studies in Quantitative Linguistics*, pp. 17–33, RAM-Verlag, Lüdenscheid,
https://ram-verlag.de/wp-content/uploads/studies_open_access/studies5.pdf.
- Peter GRZYBEK and Milan RUSKO (2009), Letter, grapheme and (allo-)phone frequencies: The case of Slovak, *Glottology*, 2(1):30–48,
<https://doi.org/10.1515/glott-2009-0004>.
- Adriana HANULÍKOVÁ and Silke HAMANN (2010), Slovak, *Journal of the International Phonetic Association*, 40(3):373–378,
<https://doi.org/10.1017/S0025100310000162>.
- Martha J. HARDMAN, Juana VÁSQUEZ, and Juan de Dios YAPITA (2001), *Aymara: Compendio de estructura fonológica y gramatical*, volume 4 of *Biblioteca lengua y cultura andina*, Instituto de Lengua y Cultura Aymara, La Paz.
- Marc HUG (1979), *La distribution des phonèmes en français : études statistiques*, volume 5 of *Travaux de linguistique quantitative*, Slatkine, Geneva.
- Tore JANSON (1986), Cross-linguistic trends in the frequency of CV sequences, *Phonology*, 3:179–195, <https://doi.org/10.1017/S0952675700000634>.
- Fred KARLSSON (2015), *Finnish: An essential grammar*, Routledge Essential Grammars, Routledge, London & New York, 3rd edition.
- Haruko KAWASAKI-FUKUMORI (1992), An acoustical basic for universal phonotactic constraints, *Language and Speech*, 35(1-2):73–86,
<https://doi.org/10.1177/002383099203500207>.
- Emmerich KELIH (2009), Phonemverbindungen und Inventarumfang: Empirische Evidenz und Modellentwicklung, *Glottology*, 2(1):60–74,
<https://doi.org/10.1515/glott-2009-0006>.
- Emmerich KELIH (2013), Phonemhäufigkeiten in slawischen Sprachen: Empirisches Verhalten und Modellierung, *Anzeiger für Slavische Philologie*, 41:75–108,
<https://unipub.uni-graz.at/anzeiger/download/pdf/10269006>.
- Emmerich KELIH and Peter ZÖRNIG (2022), Phoneme frequencies in Slovene (text vs. dictionary), *Slovenski jezik / Slovene Linguistic Studies*, 14:64–95,
<https://doi.org/10.3986/sjsls.14.1.03>.
- Michaela KOŠČOVÁ, Ján MAČUTEK, and Emmerich KELIH (2016), A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies, *Journal of Quantitative Linguistics*, 23(2):177–190,
<https://doi.org/10.1080/09296174.2016.1142327>.

Lubomír KUBÁČEK (1994), Confidence limits for proportions of linguistic entities, *Journal of Quantitative Linguistics*, 1(1):56–61, <https://doi.org/10.1080/09296179408589999>.

Henry KUČERA and George K. MONROE (1968), *A comparative quantitative phonology of Russian, Czech and German*, Elsevier, New York.

Reinhard KÖHLER (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*, volume 31 of *Quantitative Linguistics*, Brockmeyer, Bochum.

Reinhard KÖHLER (2005), Synergetic linguistics, in Reinhard KÖHLER, Gabriel ALTMANN, and Rajmund G. PIOTROWSKI, editors, *Quantitative Linguistik / Quantitative Linguistics. Ein internationales Handbuch / An International Handbook*, volume 27 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pp. 760–774, de Gruyter, Berlin & New York.

Johan LILJENCANTS and Björn LINDBLOM (1972), Numerical simulation of vowel quality systems: The role of perceptual contrast, *Language*, 48(4):839–862, <https://doi.org/10.2307/411991>.

Björn LINDBLOM (1983), Economy of speech gestures, in Peter F. MACNEILAGE, editor, *The Production of Speech*, pp. 217–245, Springer, New York, https://doi.org/10.1007/978-1-4613-8202-7_10.

Björn LINDBLOM and Ian MADDIESON (1988), Phonetic universals in consonant systems, in LARRY M. HYMAN and Charles N. LI, editors, *Language, speech and mind. Studies in honour of Victoria A. Fromkin*, pp. 62–78, Routledge, London.

John LOTZ (1952), Vowel frequency in Hungarian, *WORD*, 8(3):227–235, <https://doi.org/10.1080/00437956.1952.11659435>.

Marie LUDVÍKOVÁ and Marie KÖNIGOVÁ (1967), Quantitative research of graphemes and phonemes in Czech, *Prague Bulletin of Mathematical Linguistics*, 7:15–29.

Alfred F. MAJEWICZ (1989), *Języki świata i ich klasyfikowanie [Languages of the world and their classification]*, Państwone Wydawnictwo Naukowe, Warsaw.

M. MARINOVA and A. MARINOV (1964), Statističeski izsledvanija na fonemite v bălgarskija knižoven ezik [Statistical studies of phonemes in the Bulgarian standard language], *Bălgarski ezik*, 1964(2--3):173–179.

Colin MARTINDALE, Sabir M. GUSEIN-ZADE, Dean MCKENZIE, and Mark Y. BORODOVSKY (1996), Comparison of equations describing the ranked frequency distributions of graphemes and phonemes, *Journal of Quantitative Linguistics*, 3(2):106–112, <https://doi.org/10.1080/09296179608599620>.

André MARTINET (1955), *Économie des changements phonétiques. Traité de phonologie diachronique*, Francke, Bern.

Ján MAČUTEK and Gejza WIMMER (2013), Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics, *Journal of Quantitative*

Linguistics, 20(3):227–240,

<https://doi.org/10.1080/09296174.2013.799912>.

Christfried NAUMANN (2016), The phoneme inventory of Taa (West !Xoon dialect), in Rainer VOSSEN and Wilfrid H. G. HAACKE, editors, *Lone Tree: Scholarship in the Service of the Koon: Essay in the memory of Anthony T. Traill*, pp. 311–351, Köppe, Cologne.

Vladimir A. NIKONOV (1963), Interpretacija fonetičkih častot [An interpretation of phonetic frequencies, *Učenyje zapiski instituta slavjanovedenija AN SSSR*, 27:259–270.

Kazuko OBATA (2003), *A grammar of Bilua: A Papuan language of the Solomon Islands*, volume 540 of *Pacific Linguistics*, The Australian National University, Canberra, <https://doi.org/10.15144/PL-540>.

John J. OHALA (1983), The origin of sound patterns in vocal tract constraints, in Peter F. MACNEILAGE, editor, *The Production of Speech*, pp. 189–216, Springer, New York, https://doi.org/10.1007/978-1-4613-8202-7_9.

Hemlata PANDE (2021), Mathematical modeling of the frequencies of letters for their occurrence in corpora, words (types) and in the initial positions of words of corpora, *Glottology*, 12(1):57–69, doi:doi:10.1515/glot-2020-2010, <https://doi.org/10.1515/glot-2020-2010>.

Aleksandr M. PEŠKOVSKIJ (1925), Desjat tysjač zvukov. Opyt zvukovoj charakteristiki russkogo jazyka kak osnovy dlja eufoničeskich issledovanij [Ten thousand sounds. An attempt at the sound characteristization of the Russian language as a basis for euphony studies], in Aleksandr M. PEŠKOVSKIJ, editor, *Metodika rodnogo jazyka, lingvistika, stilistika, poëtika [Methodology of the native language. linguistics, stylistics, poetics]*, pp. 167–191, Gosudarstvennoje izdatel'stvo, Leningrad & Moscow.

Ljudmila A. POKROVSKAJA (1964), *Grammatika gagauzskogo jazyka: fonetika i morfologija [A grammar of the Gagauz language: Phonetics and morphology]*, Nauka, Moscow.

Manoj K. RIYAL, Nikhil K. RAJPUT, Vinod P. KHANDURI, and Laxmi RAWAT (2016), Rank-frequency analysis of characters in Garhwali text: Emergence of Zipf's law, *Current Science*, 110(3):429–434, <https://www.currentscience.ac.in/Volumes/110/03/0429.pdf>.

Andrés ROMERO-FIGUEROA and Konrad RYBKA (2023), Warao, in Patience EPPS and Lev MICHAEL, editors, *Amazonian Languages. Language Isolates II: Kanoé to Yurakaré. Volume 2. An International Handbook*, pp. 1243–1282, De Gruyter Mouton, Berlin & Boston, <https://doi.org/10.1515/9783110432732-015>.

Andrij ROVENCHAK and Valentin VYDRIN (2010), Quantitative properties of the Nko writing system, in Peter GRZYBEK, Emmerich KELIH, and Ján

MAČUTEK, editors, *Text and Language: Structures, Functions, Interrelations, Quantitative Perspectives*, pp. 171–181, Praesens, Vienna.

Irena SAWICKA (2008), The phonetics of Macedonian, *STUF - Language Typology and Universals*, 61(2):147–152, <https://doi.org/10.1524/stuf.2008.0014>.

Irena SAWICKA and Axel HOLVOET, editors (1991), *Studies in the phonetic typology of the Slavic languages*, Omnitech Press, Warsaw.

Ernest A. SCATTON (1993), Bulgarian, in Bernard COMRIE and Greville G. CORBETT, editors, *The Slavonic Languages*, Routledge Language Family Descriptions, pp. 188–248, Routledge, London & New York.

August SCHLEICHER (1852), *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt*, H.B. König, Bonn.

Dmitrij M. SEGAL (1972), *Osnovy fonologičeskoj statistiki (na materiale pol'skogo jazyka) [Fundamentals of phonological statistics (based on Polish language material)]*, Nauka, Moscow.

Bengt SIGURD (1968), Rank-frequency distributions for phonemes, *Phonetica*, 18(1):1–15, <https://doi.org/10.1159/000258595>.

Geoff P. SMITH (2008), Tok Pisin in Papua New Guinea: phonology, in Kate BURRIDGE and Bernd KORTMANN, editors, *Varieties of English 3: The Pacific and Australasia*, pp. 188–209, De Gruyter, Berlin & New York, <https://doi.org/10.1515/9783110208412.1.188>.

Tonya STEBBINS, Bethwyn EVANS, and Angela TERRILL (2018), The Papuan languages of Island Melanesia, in Bill PALMER, editor, *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*, volume 4 of *The World of Linguistics*, pp. 775–894, De Gruyter Mouton, Berlin & Boston, <https://doi.org/10.1515/9783110295252-007>.

Kenneth N. STEVENS (1972), The quantal nature of speech: Evidence from articulatory-acoustic data, in Edward E. DAVID and Peter B. DENES, editors, *Human communication: A unified view*, volume 15 of *Inter-University Electronics Series*, pp. 51–66, McGraw-Hill, New York.

Kenneth N. STEVENS and Samuel J. KEYSER (2010), Quantal theory, enhancement and overlap, *Journal of Phonetics*, 38(1):10–19, <https://www.sciencedirect.com/science/article/pii/S0095447008000600>.

Marija TABAIN and Andrew BUTCHER (2014), Pitjantjatjara, *Journal of the International Phonetic Association*, 44(2):189–200, doi:10.1017/S0025100314000073.

Thomas TARNÓCZY (1961), Phonetische Gesichtspunkte bei der Zusammenstellung von Texten für Verständlichkeitsmessungen. Ein Beispiel für die ungarische Sprache, *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14(1):74–87, <https://doi.org/10.1524/stuf.1961.14.14.74>.

Nikolaj S. TRUBETZKOJ (1939), *Grundzüge der Phonologie*, volume 7 of *Travaux du Cercle Linguistique de Prague*, Cercle Linguistique de Prague, Prague.

Alex TRUEMAN, Louise ST. AMOUR, and Heidi HARLEY (2018), Implementing orthography, in Megan KEOUGH, Natalie WEBER, Andrei ANGHELESCU, Sihwei CHEN, Erin GUNTLY, Khia JOHNSON, Daniel REISINGER, and Oksana TKACHMAN, editors, *Proceedings of the Workshop on Structure and Constituency in the Languages of the Americas 21*, volume 46 of *University of British Columbia Working Papers in Linguistics*, pp. 254–265, University of British Columbia, Vancouver, https://lingpapers.sites.olt.ubc.ca/files/2018/01/WSCLA21-19-Trueman_etal.pdf.

Juhan TULDAVA (1995), Quantitative analysis of the phonemic system of the Estonian language, in Juhan TULDAVA, editor, *Methods in Quantitative Linguistics*, volume 54 of *Quantitative Linguistics*, pp. 161–187, WVT, Trier.

Vasilij M. VAKHRUSHEV and Viktor N. DENISOV (1992), *Sovremennij udmurtskij jazyk: Fonetika. Grafika i orfografija. Orfoepija. [Contemporary Udmurt language: Phonetics. Graphics and orthography. Orthoepy]*, Udmurtija, Izhevsk.

René VAN DEN BERG and Brent WIEBE (2019), *Bola grammar sketch*, volume 63 of *Data Papers on Papua New Guinea Languages*, SIL-PNG Academic Publications, Ukarumpa, https://www.sil.org/system/files/reapdata/69/02/40/69024055998614577724075211921224072564/Bola_grammar_20_03_2019_final.pdf.

Gerrold VAN DER STROOM (1995), *The reception of Jac. van Ginneken's language biology*, Stichting Neerlandistiek VU, Amsterdam, Münster.

Jacobus VAN GINNEKEN (1932), *De ontwikkelingsgeschiedenis van de systemen der menselijke taalklanken [The history of the development of human speech sound systems]*, N.V. Noord-Hollandsche Uitgevers-Maatschappij, Amsterdam.

Jacobus VAN GINNEKEN (1933), La biologie de la base d'articulation, *Journal de Psychologie Normale et Pathologique*, 30:266–320.

Andrew WILSON (2013), Probability distributions of grapheme frequencies in Irish and Manx, *Journal of Quantitative Linguistics*, 20(3):169–177, <https://doi.org/10.1080/09296174.2013.799919>.

Andrew WILSON and Ján MAČUTEK (2020), A classification of the Celtic languages based on grapheme frequencies, in Emmerich KELIH and Reinhard KÖHLER, editors, *Words and Numbers: In Memory of Peter Grzybek (1957-2019)*, pp. 53–68, RAM-Verlag, Lüdenscheid.

Gejza WIMMER and Gabriel ALTMANN (1999), *Thesaurus of univariate discrete probability distributions*, STAMM, Essen.

Gejza WIMMER and Gabriel ALTMANN (2005), Unified derivation of some linguistic laws, in Reinhard KÖHLER, Gabriel ALTMANN, and Rajmund G. PIOTROWSKI, editors, *Quantitative Linguistik / Quantitative Linguistics. Ein*

internationales Handbuch / An International Handbook, volume 27 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pp. 791–801, de Gruyter, Berlin & New York.

Irena YANUSHEVSKAYA and Daniel BUNČIĆ (2015), Russian, *Journal of the International Phonetic Association*, 45(2):221–228,
<https://doi.org/10.1017/S0025100314000395>.

Roberto ZARIQUIEY BIONDI (2018), *A grammar of Kakataibo*, volume 75 of *Mouton Grammar Library*, Mouton de Gruyter, Berlin & Boston,
<https://doi.org/10.1515/9783110765816>.

Yikai ZHOU, Jingyang JIANG, and Haitao LIU (2025), Language universals in sentence length: Comparing sentence length distributions of 10 languages, *Cognitive Science*, 49(9):e70115, <https://doi.org/10.1111/cogs.70115>.

George K. ZIPF (1935), *The psycho-biology of language: An introduction to dynamic philology*, Houghton-Mifflin, New York.

Šárka ŠIMÁČKOVÁ, Václav J. PODLIPSKÝ, and Kateřina CHLÁDKOVÁ (2012), Czech spoken in Bohemia and Moravia, *Journal of the International Phonetic Association*, 42(2):225–232,
<https://doi.org/10.1017/S0025100312000102>.

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

 <http://creativecommons.org/licenses/by/4.0/>