

Modelling the frequency of loanwords in different semantic fields in core vocabularies (based on WOLD data)

Emmerich Kelih^{1*} 

¹ Affiliation University of Vienna

* Corresponding author's email: emmerich.kelih@univie.ac.at

DOI: https://doi.org/10.53482/2024_56_417

ABSTRACT

The present paper focuses on the statistical modelling of loanword frequencies in different semantic fields of ten selected languages of the world. The main question addressed in this paper is whether the occurrence of loanwords in different semantic fields (data are from the WOLD project) can be modelled by a (simple) power model or not. Within the framework of synergetic linguistics, the basic idea is that the transfer and integration of loanwords is a statistically well-regulated process. In the absence of deductively found models (theoretically, also a birth-and-death process could be a suitable starting point), the proposed modelling aims to shed some new light on the underlying statistical tendencies in linguistic borrowing.

Keywords: loanwords, quantitative modelling, birth and death process, synergetic linguistics, power models.

1 Introduction

According to the ideas of synergetic linguistics languages are open and dynamic systems, governed by principles of self-regulation and balancing forces of unification and diversification. The dynamics of linguistic systems can be observed at different levels. Among others, in particular at the lexical level (or, more generally, in the vocabulary), dynamic changes are quite obvious, since the lexical stock of a language is subject to continuous changes and adaptations to extra-linguistic circumstances. The incorporation of lexical material from other languages, usually terminologically referred as *foreign words*, *loanwords* or *borrowings*, plays an important role here. In language contact research, the focus is mainly on qualitative issues of borrowing, such as the determination of the donor language of loanwords, the type of loanwords, and particularly the phonetic, phonological and morphological adaptation of loanwords during the process of incorporation into the recipient language. However, there are also some quantitative aspects of outstanding interest, where, among other things, the number of different loanwords is an indicator of the extent and amount of the influence on the lexical system of a language. Our paper deals with the statistical modelling of loanword frequencies in different semantic fields (domains)

of ten selected languages. The data are taken from the handbook “Loanwords in the World’s Languages. A comparative handbook” (hereafter WOLD), edited by Haspelmath/Tadmor (2009). The main question addressed in this paper is whether the occurrence of loanwords in different semantic fields can be modelled by an appropriate statistical model, and to what extent the related parameters reflect a language-specific characteristic or not.

2 Loanword studies: Quantitative aspects

In language contact research (Matras 2009, Myers-Scotton 2002, Thomason 2001, Winford 2010, Zenner 2013 and many others) and historical linguistics (Hock/Joseph 1996) the analysis of loanwords plays a crucial role. Loanword research is attractive in many respects, e.g. for the reconstruction of (historical) contact situations, it provides important information about law-like sound changes and it helps to determine the chronology of language change processes. In quantitative linguistics, loanwords are usually discussed in the context of Piotrovski’s law (see Altmann 1983, 1985a, Best 2016 with a bibliography of recent literature). It is generally accepted, that chronologically the incorporation of loanwords follows a particular trend, usually in the form of a so-called S-curve, for which various models have been proposed in the past. From this point of view, diachronic phenomena in quantitative linguistics have already reached the level of models, whereas synchronic phenomena in particular, i.e. the question of how many loanwords from language *Y* occur in language *X*, seem to be limited for the time being to descriptive statistical facts.

A representative example of this state of the art is the handbook “Loanwords in the World’s Languages. A comparative handbook” (Haspelmath/Tadmor 2009), which provides a wealth of data on the number of loanwords in more than 40 languages of the world. The most important characteristic of this handbook is that the determination of loanwords is based on a word list with a core vocabulary of about 1400 lexical meanings. These meanings are already grouped into 22 semantic fields¹, such as the physical world, kinship terms, animals, parts of the body, clothing and grooming, etc. All 1400 lexical meanings are translated into the languages analysed. After doing this, it is checked whether among them there are loanwords or not. The choice of this method of analysing loanwords ensures that the data gained are indeed comparable and therefore quite appropriate for crosslinguistic studies. On the basis of the data obtained, an individual loanword profile can be created, taking into account the number of loanwords

¹ Haspelmath/Tadmor (2009) work with 24 semantic fields, as they have added two more to the original list of Buck (1949) and *The Intercontinental Dictionary Series* (cf. Borin/Comrie/Saxena 2013), namely “Modern world” and “Miscellaneous function words”. Both were added to cover previously missing terms of modern life and words used mainly for the grammatical organisation of texts. The “Modern world” group consists mainly of internationalisms (such as *radio*, *TV*, *bus*, etc.), which automatically increase the borrowing rate in most of the languages analysed. It is well known that function words are only borrowed in situations of strong language contact. Interestingly, the statistical modelling of the data (cf. section 3.4.) only works when the frequency data of “Modern world” and “Miscellaneous function words” are excluded. This could be a rough indicator that these two groups do not fully “correspond” to the other proposed semantic fields.

in the different semantic fields. Table 1 gives an overview of the proposed grouping (semantic fields). The main motivation is to cover the core vocabulary of a language², which contains the most important meanings within the proposed lexical-semantic fields.

Table 1: Semantic fields and number of meanings.

No.	Semantic field	Number of meanings
1	The physical world	75
2	Kinship	82
3	Animals	96
4	The body	157
5	Food and drink	81
6	Clothing and grooming	59
7	The house	48
8	Agriculture and vegetation	74
9	Basic actions and technology	78
10	Motion	82
11	Possession	46
12	Spatial Relation	75
13	Quantity	38
14	Time	57
15	Sense perception	49
16	Emotions and values	48
17	Cognition	51
18	Speech and Language	41
19	Social and political relations	36
20	Warfare and hunting	40
21	Law	26
22	Religion and belief	26
Sum		1365

The given framework allows for a complementary and systematic analysis of the number of loanwords. In particular, the data obtained help to determine precisely the depth and the intensity of borrowing processes, since it is possible to identify certain semantic “zones of loanword preference”, i.e. semantic fields in which a conspicuous number of loanwords (either many or very few) can be found. As already pointed out by Tadmor (2009: 64), there seems to be a clear tendency towards a specific distribution of loanwords within certain semantic fields:

„However, different languages display a remarkable degree of consistency with regard to which fields are more or less affected by borrowing. While there are certainly cross-linguistic differences, most languages tend to borrow more words into similar fields, and the same fields turn up again as the ones most resistant to borrowing.

From a linguistic point of view, the most interesting finding of this type of study is that there appear to be some semantic fields that are indeed “infected” by loanwords, while other domains of the core

² The proposed framework for core vocabulary analysis is closely related to glottochronology, where a smaller word list with 100 or 200 items (known in linguistics as the Swadesh-list) is usually used. For further details see Swadesh (1952) and Embleton (2000) for an overview on recent and past developments in lexicostatistics and glottochronology.

vocabulary seem to be more restrictive, and more or less immune³ to the incorporation of loanwords. In the light of these findings, it is quite reasonable to speak of different degrees and areas of the “risk of infection” by loanwords, which obviously depends on the particular semantic field to which a meaning belongs. However, it is usually assumed that the most important factor for the incorporation of new lexical items (meanings) is the filling of an existing semantic gap. Some other factors and motives why languages adopt loanwords are discussed in Chapter 2.

Coming back to the main results presented in the WOLD handbook (cf. Haspelmath/Tadmor 2009), based on 41 languages analysed (most of which are not genetically related and “... an effort was made to represent the world’s genealogical, typological, and sociolinguistic diversity” p. 3), it appears that there are indeed some “zones of preference” for loanwords. These zones of preference are (1) religion and belief, (2) clothing and grooming, and (3) house and living, i.e. all areas where there appears to be a high demand for dynamics and incorporation of “foreign” lexical material due to extra-linguistic factors. At the other end of the scale, there are indeed quite resistant semantic fields in which there are not many loanwords: (1) the body in general and parts of the body, (2) spatial relations and (3) sense perception, i.e. in these semantic fields of the core vocabulary of the languages analysed there are almost no or very few extra-linguistic changes can be obtained. The most resistant part of the core vocabulary, however, are function words with deictic function, expressions for time relations (*today, yesterday* etc.) or spatial relations (*below, above, outside*, etc.). Taken together, these results provide strong empirical evidence that there is no single homogenous core vocabulary. Therefore no “anything goes” seems to be at work, but obviously fine-tuned breaking mechanism in the loanword incorporation come into play and the transfer from one language to another seems to be regulated in its quantitative dimension too. In the next chapter, the question of a theoretical modelling of the frequency of loanwords in different semantic fields will be discussed in more detail.

3 Going further: Modelling the loanword frequency

The above findings of a supposed, but apparently uneven, distribution of loanwords in different semantic fields call for further in-depth methodological and theoretical research. Of particular interest is the embedding of quantitative loanword research in a synergetic linguistic context. Obviously, there are many factors that influence the number of loanwords in a language, and particularly the different number of loanwords in different semantic fields as outlined in section 2.

One leading factor is undoubtedly the basic need for encoding, specification and de-specification (cf. Köhler 2005). Speakers tend to transfer loanwords from a donor language during the encoding process,

³ The traditional claim is that the core (basic) vocabulary usually does not contain “foreign” lexical material (loanwords) (see, among others, Swadesh 1952, Hock/Joseph 1996: 257, Zenner/Speelman/Geeraerts 2014). Core vocabulary is generally regarded as a relatively stable, rather old and conservative layer of the vocabulary. It is therefore assumed not to be subject to major changes.

either to fill an existing semantic gap in their own language, or to use foreign items in a first step as a more prestigious synonym for an already existing item. In this way, synonymy increases and, in some cases, even native items can be completely replaced by the loanwords. Another realistic scenario is a semantic shift of either the loanwords or the indigenous element. In both cases, however, it is mainly up to a linguistic community to accept or reject the new coding possibilities offered. Moreover, at this stage of the borrowing process, a number of extra-linguistic factors come into play which can have a strong influence on the integration or rejection of a particular loanword. These include the cultural or political importance, the economic “value” and prestige of the potential donor language, the general intensity of language contact, the level of migration, the extent of purist movements, a general openness to foreign influence, the degree of bilingualism within a society and many other factors.

All these factors, taken together, shape the vocabularies and loanwords of languages, and it is a long tradition of linguistics to study each of these individual aspects. It is important to note that the transfer of lexical items from one language to another is by no means straightforward and unidirectional, but rather a multi-dimensional, constrained and controlled kind of process. In addition, a register-specific loanword transfer rate may be at work, e.g. in standardised languages the number of loanwords can be kept artificially low due to normative attitudes, whereas in sociolects and dialects of the same language many loanwords are used⁴ massively. In any case, one must bear in mind that all these factors have a more or less strong influence on the modelling of the number of loanwords in different semantic domains. Depending on the empirical realisation of hypothetical constraints and reinforcing factors, the integration of lexical items has to be taken into account. As a result, languages have a rather individual quantitative profile of how many loanwords occur in particular semantic fields, as will be shown in the next subsection.

3.1 Case study: Quantitative loanword profile of Slovene

To illustrate this, Figure 1 shows the relative number of loanwords in 22 semantic fields of Slovene, a South Slavic Indo-European language. The data have been compiled by the author (for details see the monograph in Kelih 2023) and the lexical input is based on the word list of meanings proposed by Tadmor/Haspelmath (2009) in the WOLD project.

⁴ A special case are pidgin and creole languages, where, for example, the age of the language and the age of the objects transferred can also have an impact. In any case, many factors need to be distinguished.

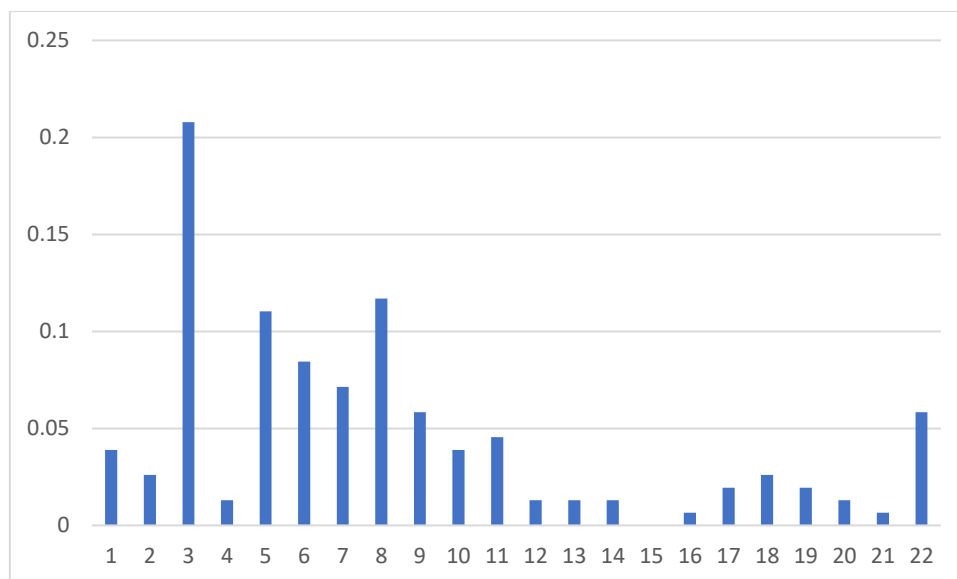


Figure 1: Relative amount of loanword frequencies in Slovene (22 semantic fields, as given in Table 1).

First of all, Slovene shows the expected uneven distribution of the number of loanwords in different semantic fields, i.e. there are fields with quite a lot of loanwords, but also others with no or very few loanwords. For example, a disproportionate high number of loanwords in Slovene can be found in the semantic fields (1) animals, (2) agriculture and vegetation, (3) food and drink, and (4) religion and belief. In fact, these are more or less the same fields that also Haspelmath/Tadmor (2009) found in their inductive generalisation of the analysis of loanwords in 41 languages of the world.

Secondly, on the other hand, relatively low loanword transfer rates can be obtained for Slovene in the semantic fields of (1) sense perception, (2) emotions and values, and (3) body and body parts. These findings, in turn, are again more or less in line with the tendencies of loanword resistance “zones” of a core vocabulary found in the above-mentioned WOLD project (for a detailed analysis and interpretation of the results for Slovene, cf. Kelih 2023).

While the quantitative analysis of loanwords could end at this descriptive stage, we now propose to go one step further and ask whether the loanword frequencies obtained follow a particular pattern or not. To do this, the next step is to transform the frequencies of loanwords in the semantic fields into a (simple) rank frequency distribution. This in turn allows the application of various statistical methods, and in the next step the statistical modelling of the data is discussed. The graphical representation of the rank-frequency distribution (based on the absolute number of loanwords within semantic fields) for Slovene is given in Figure 2; the corresponding raw data can be found in Table 2.

Table 2: Rank frequency distribution of Slovene loanwords in 22 semantic fields.

rank	no. of loanwords	rank	no. of loanwords
1	32	12	4
2	18	13	3
3	17	14	3
4	13	15	2
5	11	16	2
6	9	17	2
7	9	18	2
8	7	19	2
9	6	20	1
10	6	21	1
11	4	22	0

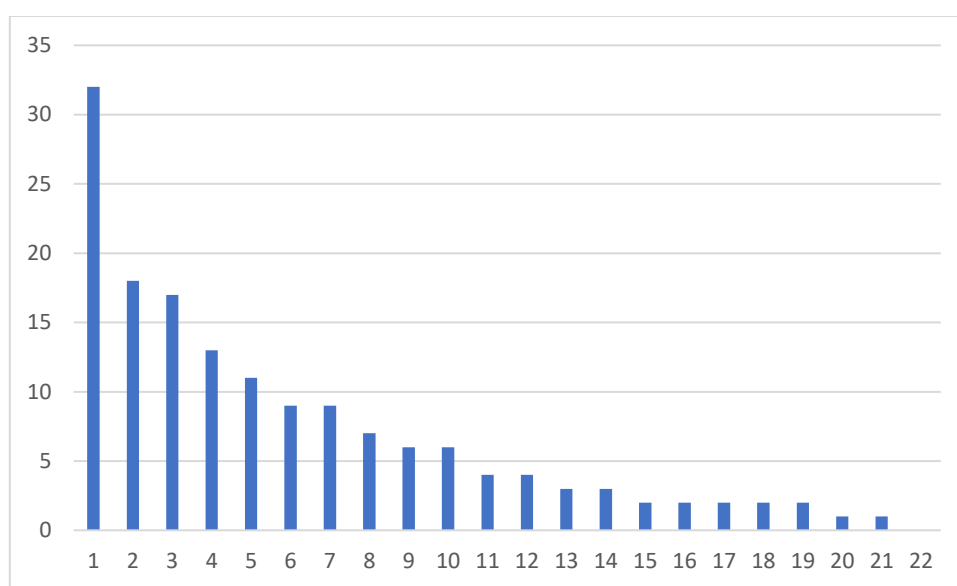
**Figure 2:** Absolute ranked frequencies of loanwords in Slovene in 22 semantic fields.

Figure 2 gives a new perspective on the distribution of loanword frequencies within particular semantic fields. As already mentioned above, they are distributed rather unevenly, i.e. there are obviously semantic fields in which we find many loanwords, and at the end of the “tail” we have a remarkable number of semantic fields in which no or very few loanwords occur. In other words, we are faced with an over- and under-utilisation of the analysed units, which is a well-known characteristic in quantitative linguistics. A similar picture can be seen in word frequency studies, which are the most intensively analysed area of quantitative “Zipfian” linguistics. The question to be discussed in the following section is whether the distribution obtained can be captured in a satisfactory way by an appropriate statistical model.

3.2 Extending the data base

Based on these preliminary observations in section 3.1. on a single language (Slovene), we will extend our data basis to some other languages that have already been analysed in the WOLD project. An overview of the languages analysed is given in Table 3 and all data, except for Slovene, are taken from Haspelmath/Tadmor (2009)⁵. The selection of languages is motivated by the language-specific borrowing rate, i.e. the average number of loanwords⁶ these languages have in their core vocabulary. Table 3 shows the calculated total relative number of loanwords in the analysed languages, which is calculated based on the absolute number of loanwords divided by the number of analysed meanings.

Tadmor (2009: 56-57) heuristically distinguishes languages with (1) very high, (2) high, (3) medium and (4) low numbers (amounts) of loanwords in their core vocabularies. For our analysis, a selection of languages with a very high, medium and low loanword borrowing rates was made.

Table 3: Languages analysed and relative number of loanwords (= borrowing rate)

No.	Language	borrowing rate	No.	Language	borrowing rate
1	Mandarin (Chinese)	0.04	6	Thai	0.21
2	Iraqw	0.11	7	Kanuri	0.26
3	Slovene	0.12	8	British English	0.38
4	Lower Sorbian	0.18	9	Tarifyt Berber	0.48
5	Sakha (Yakut)	0.23	10	Selice Romani	0.57

To provide a more detailed insight into the borrowing structure of the languages analysed, some brief linguistic comments are in order. Mandarin Chinese (cf. Wiebusch/Tadmor 2009) is of particular interest, as it tends to be a language with an exceptionally low number of loanwords. The borrowing rate of about four percent is the lowest in the entire WOLD sample, which motivates the classification of Mandarin as a truly “isolated” and “conservative” language. It should be noted that it is debatable whether even the identified loanwords are really borrowed, since in many cases the donor language and etymology remain unclear.

A representative of the “average” borrower group is Iraqw, a language spoken in Tanzania, typologically classified as Southern Cushitic (cf. Mous/Qorro 2009) and spoken by over one million people,

⁵ We are aware that the use of “foreign” data cannot exclude the possibility that different criteria were used to identify loanwords (the data were produced by different authors). However, the chosen word list approach at least provides the security of a high degree of comparability of the data.

⁶ The corresponding word lists of the analysed languages are published at <http://wold.livingsources.org/>. The lists were analysed manually and only clearly borrowed loanwords were counted for our purposes. In Haspelmath/Tadmor (2009) the authors distinguish five different degrees of borrowability: (1) no evidence of borrowing, (2) very little evidence of borrowing, (3) perhaps borrowed, (4) probably borrowed and (5) clearly borrowed. However, this scaling is not based on quantitative criteria, but is intuitive, since from our point of view “probably” and “perhaps” borrowed are hardly distinguishable by intersubjective criteria. Moreover, as Tadmor (2009: 54) points out, in the case of languages that are not well described linguistically, only marginal etymological and historical facts are known, which makes the identification of loanwords much more complicated. But even in well-documented languages, different etymological explanations are sometimes offered for the same word form.

most of whom are bilingual. Iraqw is predominantly a spoken language, with only liturgical texts available in written form. The Slovenian data were compiled by the author of this article. The main etymological dictionary of Slovene (Snoj 2003) and relevant references on loanwords from Slavic, Romance and Germanic languages (cf. Breznik 1909, Lägheid 1972, Striedter-Temps 1963, Jakopin 1975) were consulted for the identification of loanwords. The rather low rate of borrowing obtained (for further details see Kelih 2023) is rather surprising, given that Slovene is a Slavic language situated at the junction of the Romance, Germanic and Hungarian language areas, and that historically it has had very close contacts with German and Romance languages (especially Friulian and other northern Italian dialects). In this case, however, purist attitudes and normative pressures seem to have determined the low rate of borrowings in standard Slovene. The second Slavic language for which data are available goes back to Bartels (2009), who determined the loanwords in Lower Sorbian, a small West Slavic language spoken in eastern Germany and closely related to German for centuries. Both Slovene and Lower Sorbian can be considered as languages with a “medium” borrowing rate.

Sakha (Yakut), a Turkic language from Siberia (cf. Pakendorf/Novgorodov 2009), is mainly spoken by bilingual speakers, with Russian being used almost as the L2 language. However, it is reported that the communication sphere seems to be sharply divided, with Sakha dominating in private conversation and agricultural matters, and both Mongolian and Russian being important donor languages for Sakha. Thai, a standard language related to the Kra-Dai family, is the official standard language in Thailand, used mainly in official spheres of communication (politics, administration, school, mass media) and lexically strongly influenced by Chinese (cf. Suthiwan/Tadmor 2009). Kanuri, a language spoken by about four million people in Niger and Nigeria, belongs to the Saharan language family. It is mainly used orally, and the main donor languages are Arabic and Hausa (see Löhr/Ekkehard/Awagana 2009).

From the group of languages with rather high borrowing rates, English is analysed, where a borrowing rate of over 40% has been achieved. Clearly, the Latin linguistic heritage and loanwords from French seem to play an important role here (cf. Grant 2009). Tarifiyt Berber, an Afroasiatic language spoken in some parts of Morocco, also has a high borrowing rate. As a dialect of Berber, it is only used orally and does not have a generally accepted standard. Most speakers are bilingual and have a good knowledge of Arabic. The language mainly borrows words from Arabic, rather than from Spanish or French (cf. Kossmann 2009). The language with the highest rate of borrowing (67%), published in Haspelmath/Tadmor (2009), is Selice Romani, a Romani variety spoken in a single village in Slovakia. All speakers are either bilingual (Hungarian) or even multilingual, and the language is spoken only by a small group of people, with a lexicon heavily influenced by Slavic in general and Hungarian in particular (cf. Elšík 2009).

The given brief description of the languages given above shows first and foremost the “qualitative” heterogeneity of the data used. The raw data for these languages can be found in Appendix 1. To summarise,

1. languages from different genetic language families are analysed,
2. different varieties are considered, i.e. oral languages, dialects and established standard languages with a long historical tradition, and
3. languages spoken by bilingual or multilingual people are part of our pilot study.

Although this selection of languages cannot be seen as representative, our main motivation for the selection is to test languages with a rather different quantitative loanword profile in terms of their number of loanwords in different semantic domains. In addition, we believe that our choice of languages ensures rather complex conditions for modelling the rank-frequency distributions. By doing so (using complex and heterogeneous data), we believe to obtain additional support for the overall validity of the proposed statistical model presented and discussed in section 3.4.

3.3 Similarity of quantitative loanword profiles

However, before discussing the issue of statistical modelling, it is necessary to take a closer (statistical) look at the alleged heterogeneity of the data used. On the one hand, our selection of languages is deliberately focused on languages with a wide range of their quantitative loanword profile. On the other hand, the transformation of the obtained data into a rank-frequency distribution provides an opportunity to explore more deeply the idea of existing global and systematic patterns of loanword incorporation into core vocabularies. Considering the numerous factors that may influence the shape of the loanword profile, both similarity and dissimilarity of languages in this respect can be expected. However, as can be concluded from the numerous studies presented by Haspelmath/Tadmor (2009), there appear to be common preferences in the type of semantic fields in which loanwords are more likely to occur.

To get a better understanding of what is going on in the data obtained, we use the Kendall's B-correlation test, a commonly used non-parametric test that is perfectly suited to the comparison of the data used here, i.e. a closer look at which of the languages actually show similar preferences in terms of integrating loanwords into their core vocabularies. The test is based on comparing the order of rank frequencies. This means that languages with the same or similar rank profiles, based on the number of loanwords, show higher correlation coefficients than languages with a different behaviour.

In addition to the strength of the correlation, the statistical significance of the results⁷ can also be reported. In the Table 4 the results of the pairwise comparisons are given, and highly significant results ($p = 0.01$) are marked with ** and significant results ($p = 0.05$) are marked with *.

A closer look at the yielded results shows a rather clear tendency. In fact, most of the languages seem to have a rather similar rank-frequency profile, i.e. the clear tendency to accumulate loanwords in certain semantic fields and a clear tendency to avoid them in others can be recapitulated. Another

⁷ The null hypothesis tested is that there are no correlations (Kendall-B correlation = 0) between the languages compared.

interesting observation is that this tendency also holds true for quite different levels of borrowing activity (= average borrowing rates).

For details, see Table 4 with the results of the Kendall B-correlation test for Mandarin vs. Slovene, where Chinese is the language with the lowest number of loanwords and Slovene is an “average” borrower. The same is true when comparing Mandarin with Selice Romani, the language with the highest borrowing rate of the languages analysed. Therefore, our preliminary conclusion is that no matter how many loanwords are integrated in total, they are mostly systematically distributed within semantically determined sub-areas (= semantic fields) of the lexical stock.

Table 4: Kendall-B-correlation values for ten languages.

Language	Mandarin	Iraqw	Slov.	L.Sorb.	Sakha	Thai	Kanuri	BritischE	T.Berber	S.Romani
Mandarin		.446**	.403**	.394*	.385*	0.00	0.14	0.28	.377*	.394*
Iraqw			.628**	.584**	.680**	0.01	.403**	.351*	.377*	.463**
Slov.				.662**	.532**	0.06	.342*	.359*	.333*	.385*
L.Sorb.					.541**	0.22	.333*	.403**	.429**	.463**
Sakha						0.13	.325*	.463**	.645**	.645**
Thai							0.16	0.14	0.24	0.29
Kanuri								0.19	0.25	0.28
BritischE									.472**	.420**
T.Berber										.532**
S.Romani										

However, as can be seen from the results in Table 4, there does indeed seem to be one major outlier from the above-described mechanism of the distribution of loanwords between the semantic fields of the core vocabularies. Thai, shows a remarkable “individual” behaviour, since in all cases (in Table 4 these results are marked in bold) no (significant) correlations can be obtained. In fact, as Suthiwan/Tadmor (2009: 605) point out in their analysis, Thai is characterised by an “[...] unusually high loanword rate in function words (20.2%)” due to the massive borrowing of pronouns and numerals from Middle Chinese and/or Sanskrit and Pali. This rather specific feature is obviously one of the reasons why the loanword profile of Thai differs from all the other languages analysed here.

The same is true, but to a lesser extent, for Kanuri, whose rank frequency profile is quite different from Thai, but also from British English, Tarifiyt Berber and Selice Romani. In this case, the relatively high number of loanwords in the semantic field time could be a decisive factor, resulting in a rather specific loanword profile.

In conclusion, and without going into further details, the proposed analysis based on the Kendalls-Thau correlation coefficient allows to gain interesting insights into some tendencies and mechanisms of the distribution of loanwords among a predefined spectrum of semantic fields of the core vocabularies.

Although the above analysis is based only on a selection of ten languages, a general pattern of loanword integration within certain semantic fields is apparent, supposedly independent of the total number of loanwords. In this respect, the statistical re-analysis of the data on the one hand confirms some supposedly common cross-linguistic principles of loanword distribution, but on the other hand it also allows for the identification of individual and “deviant” tendencies. In the next chapter, an attempt at a statistical modelling of the obtained loanword rank frequencies will be presented.

3.4 Modelling the rank-frequencies of loanwords

A major goal of quantitative linguistics is the discovery of tendencies, regularities and laws of linguistic systems and phenomena. Here we ask whether there are systematic tendencies in the shape of the rank-frequency distributions of loanwords in the basic vocabulary. To get a visual impression, Figures 3a-3d show the relative rank-frequency profiles (ranked relative frequencies based on the number of loanwords per semantic field divided by the total number of loanwords in a language) of four selected languages. Starting with Mandarin (the languages with the lowest average borrowing rate) and ending with Selice Romani, the language with the highest number of loanwords in our case study.

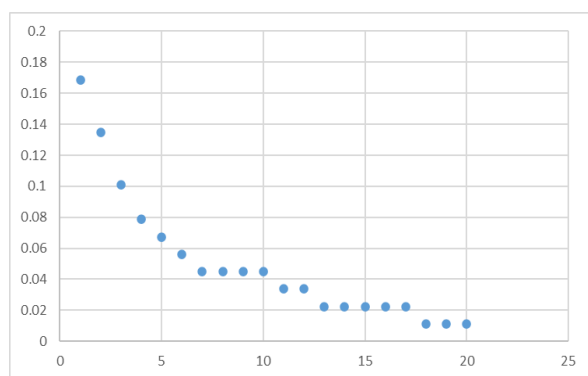


Figure 3a: Mandarin (Chinese).

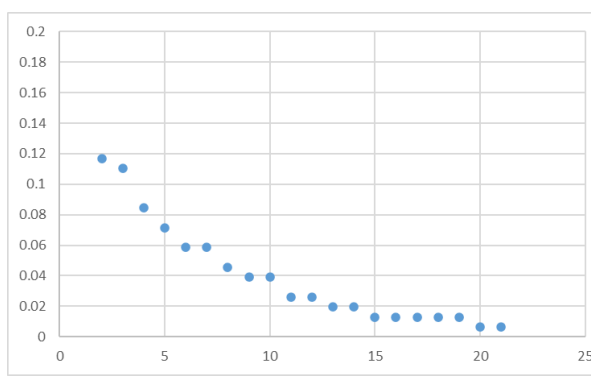


Figure 3b: Slovene.

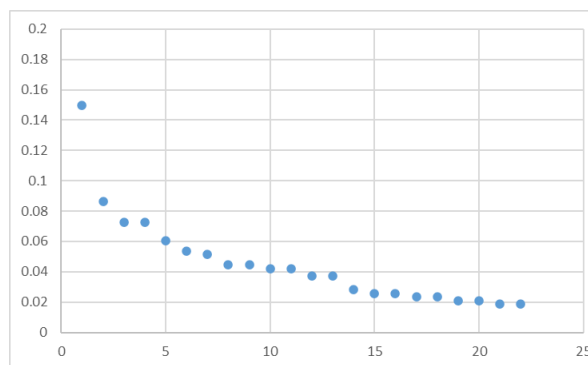


Figure 3c: Thai .

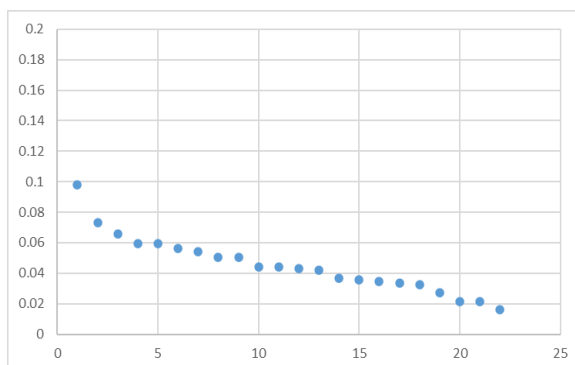


Figure 3d: Selice Romani.

A first visual interpretation of the rank-frequency distributions of the selected languages resembles a power-law distribution. Another observation regarding the most frequent rank is the decrease of the

most frequent loanwords (= rank 1) as the borrowing rate increases. I.e. Mandarin has a very low number of loanwords, but they are clearly concentrated in the front (left) part of the distribution, whereas Selice Romani (with over 55% loanwords) has a more “even” distribution, i.e. many loanwords occur in many different semantic fields, but the overall distribution (at least based on the graphical representation given in Fig. 3d.) is much more “even” than in the case of Mandarin Chinese.

Obviously, an increasing average borrowing rate goes hand in hand with a progressive spread of loanwords across different semantic fields. However, this is not accompanied by an overuse of loanwords in one particular semantic field, which seems to be the case only for languages with a rather small number of loanwords (as in the case of Mandarin Chinese, where 40% of the loanwords are concentrated in three semantic fields only). We therefore suggest that the quantitative incorporation of loanwords is a statistically well-regulated process in which there is no overburdening of any particular semantic field by loanwords, but rather a “scatter the burden” mechanism is at work. This mechanism is responsible for the successive filling of different semantic fields with an increasing overall borrowing rate.

For the modelling of the relative rank frequencies of loanwords, we limit⁸ ourselves here to two power models that play an important role in the statistical modelling of linguistic data:

$$(1) y = a * x^b \text{ and}$$

$$(2) y = a * e^{bx}.$$

In fact, for the data analysed, as can be seen from the results presented in Table 5, both models are quite adequate. However, the fitting results are slightly better for the exponential model (2), which clearly indicates a better coverage of the rank-frequency profiles of the loanwords.

⁸ Linguistically, we believe that (1) a birth-and-death process and (2) general epidemiological model are two (actually interrelated) prospective theoretical conceptions for modelling loanword data. A promising starting point is to interpret the transfer of loanwords to particular semantic fields as a birth-and-death process, where the rank frequency distribution is understood abstractly as the result of a continuous birth and death of lexical items (cf. Altmann 1985b, Beöthy/Altmann 1984, Krylov 2002 for further details). Indeed, at any given time, languages are confronted with the phenomenon that many loanwords attempt to be incorporated, but ultimately only a few survive, while many others fail to be actively used by a linguistic community. Moreover, even after successful integration, loanwords can become unattractive again, especially if they express phenomena and objects that are no longer relevant. Loanwords can potentially enter any semantic field, which corresponds to a birth in semantic field x ; at the same time, loanwords can also be eliminated within a semantic field (due to purist attitudes or simply because loanwords are no longer used by speakers), which in our model results in the death of an item. Both processes are in steady states, where the corresponding equilibrium equation can be solved mathematically. An epidemiological “interpretation” is something that is also at the core of the Piotrovskij law, which is based on the interaction between speakers and hearers. This approach is closely related with the well-known S-shaped chronological development of the incorporation of loanwords (cf. Altmann 1985a, Best 2015).

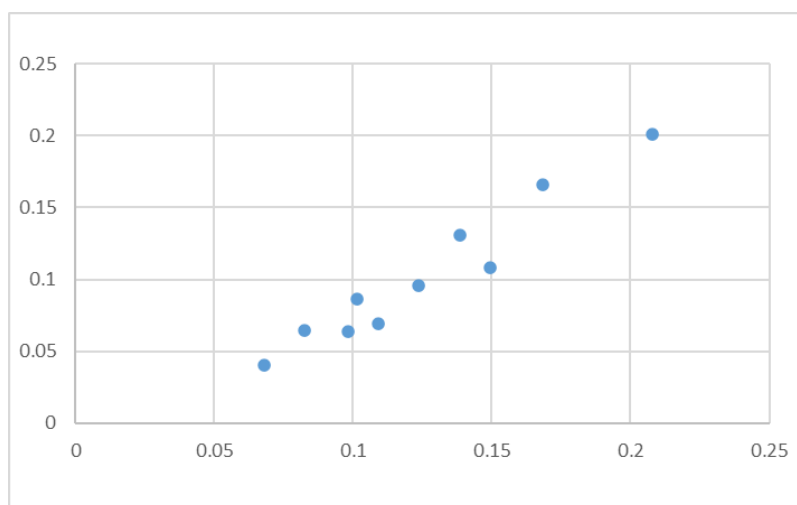
Table 5: Fitting results of rank-frequency distribution of loanwords.

Language	model (1)			model (2)		
	<i>a</i>	<i>b</i>	R ²	<i>a</i>	<i>b</i>	R ²
Mandarin	0.19	-0.71	0.93	0.18	-0.17	0.95
Iraqw	0.16	-0.63	0.85	0.15	-0.13	0.97
Slovene	0.22	-0.81	0.95	0.22	-0.20	0.95
Lower Sorbian	0.13	-0.51	0.83	0.12	-0.10	0.93
Sakha	0.12	-0.46	0.84	0.11	-0.09	0.98
Thai	0.15	-0.59	0.97	0.12	-0.11	0.85
Kanuri	0.11	-0.42	0.88	0.09	-0.07	0.88
English	0.10	-0.37	0.80	0.09	-0.06	0.92
Taryfit Berber	0.08	-0.25	0.80	0.07	-0.04	0.96
Selice Romani	0.10	-0.39	0.91	0.09	-0.06	0.92

For the exponential model (2), at least five languages show an excellent fit ($R^2 \geq 0.95$) and three languages a satisfactory fit ($R^2 \geq 0.90$). Among the languages with the poorest fit is Thai, a language that we have already identified as having a “special” rank-frequency behaviour of its loanwords compared to the other languages analysed here. The same is true, but to a lesser extent, for Kanuri.

Our limited selection of languages does not, of course, allow for sound generalisations, but what can be shown is a non-chaotic behaviour of the incorporation of loanwords into different semantic fields of the basic vocabulary.

While we do not claim to have found the “ultimate” correct model(s), a closer look at the parameter *a* is of model (2) in relation to the relative frequency of loanwords in rank 1 (= highest frequency *rel_1*), as shown in Figure 4, can be seen as a promising supporting indicator. The mutual relationship between the relative frequency in rank 1 and the parameter *a* is also reflected by a high correlation coefficient ($r = 0.98$).

**Figure 4:** Interrelation between *rel_1* (x-axis) and parameter *a* (y-axis).

Therefore, the iteratively gained parameter a could be replaced without any significant loss of information with rel_1 in the above given theoretical models. More important is the fact, that the frequencies of loanwords within semantic fields are not distributed chaotically but follow a certain statistical pattern. Moreover, the convincing fitting results give hope that (at least preliminarily) adequate models have been found. The idea of discovering such regularities should be further discussed in quantitative loanword research, taking into account data from many more languages than could be done here. However, it may be that no single model can be expected to be adequate for the distribution of loanwords, and also other reliable ways of modelling should be considered in the future. In any case, more data from many other languages will shed more (empirical) light on the specific behaviour of loanwords in different semantic domains of core language vocabularies.

4 Summary

This paper has shown that the frequency of loanwords in different semantic domains of the core vocabulary follows a particular statistical model. The transfer and selection (either integration or rejection) results in a statistically measurable specific loanword profile based on the corresponding rank frequency data. The results of the fit give rise to the hope that the borrowability and, more generally, the transfer of lexical material from one language to another is determined by a (probabilistic?) selection filter that regulates and directs loanwords towards certain semantic fields, which in turn leads to selected zones of loanword preference in the core vocabularies of the languages analysed. In any case, it could be shown that there is no unsystematic “flood” of loanwords into the core vocabularies of the languages, but rather a well-balanced, obviously self-regulated transfer. Moreover, an overall increase in the number of loanwords seems to be accompanied by an underlying “scatter the burden” mechanism, which results in a more even distribution of loanwords in different semantic fields, in order to avoid overuse and overburdening of certain fields.

Acknowledgments

Many thanks to the reviewers and editors for their valuable comments and suggestions for some methodological and theoretical improvements of the paper. Thanks are due to Gabriel Altmann for many years of discussion of the results presented here. The text was proofread with the help of DeepL Write (<https://www.deepl.com/de/write>).

References

- Altmann, G.** (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H. et al. (Eds.): *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte*, pp. 54–90. Göttingen: Herodot. (= Göttinger Schriften zur Sprach- und Literaturwissenschaft, 2).

- Altmann, G.** (1985a). On the dynamic approach to language. In: Ballmer, Thomas T. (ed.): *Linguistic Dynamics. Discourses, Procedures and Evolution*, pp. 181–189. Berlin [etc.]: de Gruyter. (= Research in text theory, 9).
- Altmann, G.** (1985b). Semantische Diversifikation. *Folia Linguistica*, 19 (1–2), pp. 177–200.
- Awagana, A., Wolff, H. E., Löhr, D.** (2009). Loanwords in Hausa, a Chadic language in West Africa. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 142–165. Berlin, New York: De Gruyter Mouton.
- Bartels, H.** (2009). Loanwords in Lower Sorbian, a Slavic language in Germany. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 304–323. Berlin, New York: De Gruyter Mouton.
- Best, K.-H.** (2015). Malay borrowings in English. *Glottometrics*, 31, pp. 50–53.
- Best, K.-H.** (2016). Bibliography – Piotrowski's law. *Glottology*, 7 (1), pp. 89–94.
- Beöthy, E., Altmann, G.** (1984). The diversification of meaning of Hungarian verbal prefixes. II. ki-. *Finnisch-Ungarische Mitteilungen* 8, pp. 29–37.
- Borin, L., Comrie, B., Saxena, A.** (2013). The Intercontinental Dictionary Series – a rich and principled database for language comparison. In: Saxena A., Borin, L. (Eds.). *Approaches to measuring linguistic differences*, pp. 285–302. Berlin: De Gruyter Mouton. (= Trends in Linguistics. Studies and Monographs, 265).
- Buck, C. D.** (1949). *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*. Chicago: The University of Chicago Press.
- Breznik, A.** (1909). *Slovanske besede v slovenščini*. Ljubljana: Katoliška Bukvarna.
- Embleton, Sh.** (2000). Lexicostatistics/Glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In: Renfrew, C., McMahon, A., Trask, L. (Eds.). *Time depth in historical linguistics. Volume 1*, pp. 143–165. Cambridge.
- Elšík, V.** (2009). Loanwords in Selice Romani, an Indo-Aryan language of Slovakia. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 260–303. Berlin, New York: De Gruyter Mouton.
- Grant, A.** (2009). Loanwords in British English. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 360–383. Berlin, New York: De Gruyter Mouton.
- Haspelmath, M., Tadmor, U.** (Eds.) (2009). *Loanwords in the world's languages. A comparative handbook*. Berlin, New York: De Gruyter Mouton.
- Hock, H. H., Joseph, B. D.** (1996). *Language history, language change, and language relationship. An introduction to historical and comparative linguistics*. Berlin, New York: Mouton De Gruyter (= Trends in Linguistics. Studies and Monographs, 93).
- Jakopin, F.** (1975). Ruske izposojenke v slovenščini. *Jezik in slovstvo*, 21 (4), pp. 134–137.
- Kelih, E.** (2023). *Entlehnungen im slowenischen Basiswortschatz - ein gebrauchsbasierter Ansatz*. Frankfurt a.M.: Lang. (= Sprach- und Kulturkontakte in Europas Mitte, 17).

- Kossmann, M.** (2009). Loanwords in Tarifit, a Berber language of Morocco. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 191–214. Berlin, New York: De Gruyter Mouton.
- Köhler, R.** (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Eds.). *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin, New York: De Gruyter, pp. 760–774. (= Handbücher zur Sprach- und Kommunikationswissenschaft, 27).
- Krylov, Ju. K.** (2002). Synergetic models and Methods in Quantitative Linguistics. *Journal of Quantitative Linguistics*, 9 (2), pp. 125–185.
- Lägreid, A.** (1973). *Die russischen Lehnwörter im Slovenischen. Die in der ersten Hälfte des 19. Jahrhunderts übernommenen Wörter*. München: Trofenik. München: Trofenik. (= Geschichte, Kultur und Geisteswelt der Slowenen, 12).
- Löhr, D., Wolff, H. E., Awagana, A.** (2009). Loanwords in Kanuri, a Saharan language. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 166–188. Berlin, New York: De Gruyter Mouton.
- Matras, Y.** (2009). *Language contact*. Cambridge: Cambridge University Press.
- Mous, M., Qorro, M.** (2009). In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 103–123. Berlin, New York: De Gruyter Mouton.
- Myers-Scotton, C.** (2002). *Contact linguistics. Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.
- Pakendorf, B., Novgorodov, I. N.** (2009). Loanwords in Sakha (Yakut), a Turkic language of Siberia. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 496–524. Berlin, New York: De Gruyter Mouton.
- Snoj, M.** (2003). *Slovenski etimološki slovar*. 2. pregledana in dopolnjena izdaja. Ljubljana: Modrijan.
- Suthiwan, T., Tadmor, U.** (2009). Loanwords in Thai. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 599–616. Berlin, New York: De Gruyter Mouton.
- Striedter-Temps, H.** (1963). *Deutsche Lehnwörter im Slovenischen*. Wiesbaden: Harrassowitz.
- Swadesh, M.** (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts. With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96 (4), pp. 452–463.
- Tadmor, U.** (2009). Loanwords in the world's languages: Findings and results. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 55–75. Berlin, New York: De Gruyter Mouton.
- Thomason, S. G.** (2001). *Language contact*. Edinburgh: Edinburgh University Press.
- Wiebusch, Th., Tadmor, U.** (2009). Loanwords in Mandarin Chinese. In: Haspelmath, M., Tadmor, U. (Eds.). *Loanwords in the World's Languages*, pp. 575–598. Berlin, New York: De Gruyter Mouton.
- Winford, D.** (2010). Contact and Borrowing. In: Hickey, R. (Eds.): *The handbook of language contact*. Malden, Mass.: Wiley-Blackwell, pp. 170–187.

Zenner, E., Kristiansen, G. (Eds.) (2013). *New Perspectives on Lexical Borrowing. Onomasiological, Methodological and Phraseological Innovations*. Boston: de Gruyter (= Language Contact and Bilingualism, 7).

Zenner, E., Speelman, D., Geeraerts, D. (2014). Core vocabulary, borrowability and entrenchment: A usage-based onomasiological approach. *Diachronica* 31 (1), pp. 74–105.

Appendix

Raw data used in this study (based on analyses taken from Haspelmath/Tadmor 2009)

Rank	Mandarin (Chinese)		Slovene		Iraqw		Sakha (Yakut)		Lower Sorbian	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
1	15	0.1685	32	0.2078	18	0.1385	35	0.1014	36	0.1237
2	12	0.1348	18	0.1169	16	0.1231	30	0.0870	23	0.0790
3	9	0.1011	17	0.1104	12	0.0923	30	0.0870	22	0.0756
4	7	0.0787	13	0.0844	12	0.0923	23	0.0667	22	0.0756
5	6	0.0674	11	0.0714	9	0.0692	23	0.0667	20	0.0687
6	5	0.0562	9	0.0584	9	0.0692	21	0.0609	19	0.0653
7	4	0.0449	9	0.0584	8	0.0615	19	0.0551	19	0.0653
8	4	0.0449	7	0.0455	6	0.0462	18	0.0522	16	0.0550
9	4	0.0449	6	0.0390	5	0.0385	18	0.0522	15	0.0515
10	4	0.0449	6	0.0390	5	0.0385	17	0.0493	12	0.0412
11	3	0.0337	4	0.0260	5	0.0385	16	0.0464	12	0.0412
12	3	0.0337	4	0.0260	5	0.0385	14	0.0406	12	0.0412
13	2	0.0225	3	0.0195	4	0.0308	12	0.0348	11	0.0378
14	2	0.0225	3	0.0195	4	0.0308	11	0.0319	10	0.0344
15	2	0.0225	2	0.0130	3	0.0231	10	0.0290	10	0.0344
16	2	0.0225	2	0.0130	3	0.0231	10	0.0290	8	0.0275
17	2	0.0225	2	0.0130	3	0.0231	9	0.0261	8	0.0275
18	1	0.0112	2	0.0130	2	0.0154	8	0.0232	6	0.0206
19	1	0.0112	2	0.0130	1	0.0077	7	0.0203	4	0.0137
20	1	0.0112	1	0.0065	0	0.0000	6	0.0174	3	0.0103
21	0	0.0000	1	0.0065	0	0.0000	6	0.0174	2	0.0069
22	0	0.0000	0	0.0000	0	0.0000	2	0.0058	1	0.0034
	89	1	154	1	130	1	345	1	291	1

Rank	Thai		Kanuri		British English		Tarifiyt Berber		Selice Romani	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
1	64	0.1495	37	0.1091	45	0.0826	52	0.0682	91	0.0982
2	37	0.0864	24	0.0708	45	0.0826	49	0.0642	68	0.0734
3	31	0.0724	22	0.0649	41	0.0752	45	0.0590	61	0.0658
4	31	0.0724	21	0.0619	39	0.0716	45	0.0590	55	0.0593
5	26	0.0607	21	0.0619	31	0.0569	45	0.0590	55	0.0593
6	23	0.0537	19	0.0560	28	0.0514	42	0.0550	52	0.0561
7	22	0.0514	17	0.0501	27	0.0495	41	0.0537	50	0.0539
8	19	0.0444	17	0.0501	27	0.0495	39	0.0511	47	0.0507
9	19	0.0444	16	0.0472	27	0.0495	36	0.0472	47	0.0507
10	18	0.0421	16	0.0472	25	0.0459	36	0.0472	41	0.0442
11	18	0.0421	16	0.0472	24	0.0440	35	0.0459	41	0.0442
12	16	0.0374	15	0.0442	23	0.0422	35	0.0459	40	0.0431
13	16	0.0374	14	0.0413	22	0.0404	32	0.0419	39	0.0421
14	12	0.0280	14	0.0413	22	0.0404	30	0.0393	34	0.0367
15	11	0.0257	13	0.0383	21	0.0385	29	0.0380	33	0.0356
16	11	0.0257	11	0.0324	20	0.0367	27	0.0354	32	0.0345
17	10	0.0234	10	0.0295	19	0.0349	27	0.0354	31	0.0334
18	10	0.0234	10	0.0295	17	0.0312	27	0.0354	30	0.0324
19	9	0.0210	8	0.0236	17	0.0312	27	0.0354	25	0.0270
20	9	0.0210	8	0.0236	10	0.0183	26	0.0341	20	0.0216
21	8	0.0187	6	0.0177	10	0.0183	22	0.0288	20	0.0216
22	8	0.0187	4		5	0.0092	16	0.0210	15	0.0162
	428	1	339	1	545	1	763	1	927	1