Text and Language

Structures · Functions · Interrelations
Quantitative Perspectives

Edited by
Peter Grzybek
Emmerich Kelih
Ján Mačutek

Language

praesens

Peter Grzybek
Emmerich Kelih
Ján Mačutek
(eds.)

Advisory Editor
Eric S. Wheeler

Text and Language

Structures · Functions · Interrelations.
Quantitative Perspectives

SONDERDRUCK

prae
sens

Contents

Preface

The present volume unites twenty-three contributions from international scholars, all renowned experts in the field of quantitative linguistics. The contributions were presented at the Quantitative Linguistics Conference (QUALICO 2009), standing in a tradition of previous meetings organized by the International Quantitative Linguistics Association IQLA (`www.iqla.org`) in Trier (Germany), Moscow (Russia), Helsinki (Finland), Praha (Czechia), and Athens (Georgia, USA). QUALICO 2009 was organized in co-operation with the University of Graz (Austria), particularly the Institute for Slavic Studies, Sept. 17–20, 2009; without substantial support from Graz University, particularly the Faculty of Arts and Humanites, and the Office for the Government of the Province of Styria (Department of Science), the conference and, as a consequence, this volume, would not have been possible, and it is our wish to express our gratitude to all these benevolent sponsors.

Generally speaking, issues of quantitative linguistics in a broad understanding of this term were the focus of the conference, and they thus shape the general profile of the present book. As a discipline, quantitative linguistics typically follows a specific scientific paradigm: in this theoretical framework, (qualitative) linguistic hypotheses are 'translated' into quantitative terms and tested by means of statistical procedures. The results are first quantitatively interpreted, which leads to either the rejection or the retainment of the hypothesis; only then are they, after some kind of 're-translation' into linguistic terms, qualitatively interpreted and embedded into theoretical concepts. The application of mathematical and statistical methods thus is no self-contained aim or objective in a quantitative linguistics framework, but one necessary step in the logic of science.

In detail, against the background of this general approach, the complex relations between 'text' and 'language' are specifically focused in the contributions to this volume. Over the last decades, a number of laws and law hypotheses have been developed, and often we do not have sufficient knowledge about the boundary conditions of them, that is, if they are relevant for language as an abstract construct, or if they are (also) valid for concrete individual texts, or groups, or types of texts, etc. This fact explains why some contributions, in this point rather following the tradition of linguistics in a narrower understanding, focus on language as a whole (be it conceived of as a set of rules, as an abstract model of concrete or possible utterances, or even differently), whereas others concentrate on individual texts; why some contributions focus questions of text typology, and others are concerned with problems of individual texts.

Given such a broad horizon of quantitative linguistics, it is not astonishing that there are many implicit or explicit points of contact with, or even technical references to neighboring disciplines – not only to mathematics, statistics, or information sciences, but also to computer linguistics, corpus linguistics, literary scholarship including individual and inter-individual stylistics, and others. After all, quantitative linguistics turns out to be genuinely interdisciplinary, and the fact that many contributions are authored by more than one person is a clear indication of this circumstance. It seems that times of individual and separate work in ivory towers is passing by also in the humanities, and the need for interdisciplinary and international co-operation becomes more and more obvious.

Such common endeavors cannot be realized without personal communication, notwithstanding the rapid developments and changes of global communication techniques which, without a douobt, clearly alleviate everyday work. Personal contacts are indispensable, and it is self-evident that they can be realized only if the scientific work is supported as, in our case, by the institutions mentioned above. In this context, we are well aware of the fact that 'behind' all these institutions there are always individual persons to whom our gratitude should be explicitly extended here. Additionally, we want to mention some persons, without whom the present volume would not have gotten the final shape: Christoph Eyrich and Werner Lemberg have been helpful in various issues of LaTeX 2_ε problems, Eric Wheeler has done a wonderful job in editing all texts, and Veronika Koch has done perfect work in supporting the layout work and in preparing the author and subject indices at the end of this volume.

Peter Grzybek
Emmerich Kelih
Ján Mačutek

Quantitative analysis of Keats' style: genre differences

Sergej Andreev

1 Introduction

One of the main objects of stylometric analysis – retrieval of the information to characterize individual style (author's stylome, fingerprints) – is connected with an important issue of style stability. If, on the one hand, research, aimed at authorship detection, classification of styles, etc., is usually based on implicitly accepted assumption of the existence of constant, unchangeable style markers, on the other hand, there does not seem to be any doubt now about the variability of style of an author over time and in different genres (Goldfield and Hoover 2008; Grzybek et al. 2005; Juola 2007; Kelih et al. 2006; Rudman 2006: 613). Of the two types of variability we shall focus on genre style differences.

Quite a number of studies have presented convincing arguments in favour of the variation of style in different genres (Grzybek et al. 2005; Kelih et al. 2006; Martynenko 1988; Rudman 1998). Nevertheless, there exists one aspect, which makes the question of genre differentiation far from being trivial. The problem may be formulated as follows – can such differences disappear over time and, secondly, are there any factors which can obscure or, vice versa, intensify stylistic genre differences? Baevskij shows the possibility of the disappearance of genre style distinctions. In his study (Baevskij 2001: 185ff.) he found that stylistic genre differences between odes, elegies, idylls, epigrams, epistles and some others existed in the poems by Lomonosov, Derzhavin, Karamzin, Krylov and other Russian poets in the second half of the 18th century, but completely disappeared in Russian verse texts during 1813–1820.

Our paper deals with the second of the above-mentioned questions, concerning the factors which can influence genre differences. Among such factors we shall analyze the 'cultural relevance', or, rather, 'success rate' of poetic works. This study is based on the analysis of the poems by English romantic poet John Keats whose posthumous influence on English poetry is considered now to be very strong and whose works, though severely criticized by his contemporary critics, are now recognized as an important element in the cultural system of present-day life.

2 Data sources

One important aspect, which is usually ignored, consists of working out the strategy of selecting the data sources for such a study, and more generally, for any stylometric research. We consider it reasonable to distinguish between the works by an author that received general recognition of the reading public and critics, or were distinguished by the author himself, and, on the other hand, poems that are less prominent, less culturally significant. In most cases this issue is not raised at all and texts by an author are taken indiscriminately, regardless of their 'popularity rate' or the author's estimation.

In this research we make such a distinction between poems that are more 'relevant' and works that are less 'relevant' for the author's creative activity judging by the author's opinion. In this case, information about whether the author included a given poem for publication in his collections of works or not may be used as objective criteria. We assume that the choice of the author in such cases reveals his priorities in matters of expression, his vision of what he considers as his proper style.

For the present analysis two lyrical genres were chosen: sonnets and odes written in iambic pentameter. Lyrics reveal most vividly the individual world of an author, his specific traits of expression. As for the genres, it should be noted that many of Keats' sonnets and odes are considered to be among the best works written by romantic writers. The restriction that we imposed on the meter by analyzing poems of the same metric scheme (iambic pentameter verses) is due to the necessity of achieving homogeneous linguistic material for comparison. It should be noted that Keats wrote in iambic pentameter most of his works belonging to these two genres. Keats' creative period lasted for about six years from 1814 till 1819. In the above-mentioned meter during his life Keats published in two collections (in 1817 and 1820) 17 sonnets (afterwards referred to in this paper as 'Sonnets [1]') and 5 odes (referred to as 'Odes'). The list of these poems is given in Appendix A.

Besides these works Keats wrote a number of sonnets which he did not include in his collections. Some of them were not published during his life at all and became known only after his death. Out of these sonnets we chose at random 18 poems (about 50% of such sonnets), forming one more class 'Sonnets [2]'. The list of these sonnets is also presented in Appendix A.

3 Characteristics

In choosing characteristics for the analysis we were guided by the following criteria. Characteristics must be well distinguished formally; their number must be less than the number of the analyzed texts, they must be frequent enough in the texts. Besides, which we consider as very important, these characteristics

must reflect essential features of lyrical poems, and be poetically and linguistically relevant. This last condition, of course, introduces certain restrictions into research, changing its perspective from looking for any feature that possesses discriminant force to testing the discriminant value of an *a priori* formed feature scheme.

Due to these requirements and based on our preliminary empirical data experiments the following list of characteristics was formed. It should be mentioned that this feature set on the whole is based on the scheme proposed by Baevskij (Baevskij 1993), though a number of changes were introduced. The methodology for rhythm description is used according to Tarlinskaja (1976). Our feature set includes rhythmic characteristics (absence of stress on ictuses 1–5; stressed anacrusis; feminine and dactylic clausula); one characteristic of rhyme (inexact rhyme); syntactic characteristics (beginning of enjambement, syntactic pauses, lines which end in exclamation or question marks – "emphatic end", inversion – partial and complete). A more detailed explanation of the characteristics may be found in Andreev (2006). Examples from Keats' poems are given in Appendix B.

4 Method

In order to establish whether there is any difference between the poems of these two genres we used discriminant analysis (Warner 2008: 650ff.). This method has been successfully used by linguists in data mining research, including problems of classification and author attribution (Grzybek et al. 2005; Kelih et al. 2006; Mikros 2006; Tambouratzis et al. 2004).

5 Results

At the first stage of our analysis the aim was to see if there are any characteristics which differentiate between two groups of texts: sonnets, included into the author's collections (Sonnets [1]) and odes, and if there are such variables, to find out how they contribute to the discrimination. As a result of the discriminant analysis it was found that the two groups of texts differ significantly. As seen in the histogram (see Figure 1, p. 4) the opposition of the sonnets and odes is clearly marked.

The following characteristics from the list of features given above were included in the discriminant model: inexact rhyme, omission of stress on ictuses 1, 3, 4, 5, feminine and dactylic clausula, emphatic end. It was also found out that four features were more characteristic of sonnets (omission of the stress on ictuses 1, 3, 4 and feminine rhyme) and the other four – of odes. From the point of view of their localization, the characteristics possessing discriminant

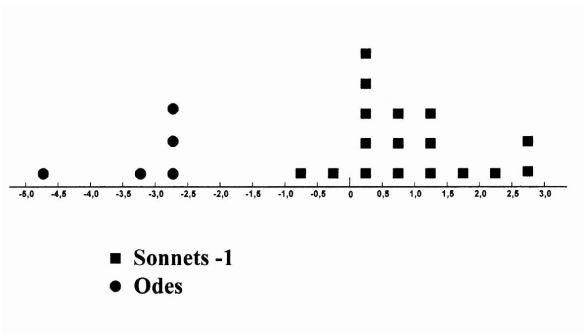


Figure 1: Histogram for two text groups (Sonnets [1] vs. Odes)

force express mostly the properties of the end of the line. The results of post hoc test are very good, as seen from Table 1, in which rows are observed and columns are predicted classifications.

Table 1: Classification matrix: Sonnets – Odes

	Percent correct	Sonnets	Odes
Sonnets	100	17	0
Odes	100	0	5
Total	100	17	5

At the first stage of analysis each ode was taken as a whole unit. On the other hand, since odes consist of a number of stanzas, it is also possible to replace complete odes by stanza sequences. Inclusion of separate stanzas into the analysis changes the number of members of the ode class from 5 (complete odes) to 23 (stanzas), bringing the total number of all cases of the two classes to 40. Discriminant analysis in this case again shows a considerably good separation of sonnets and odes (Table 2). The results of this test are much better than could be expected. Indeed, it was difficult to expect that stanzas would be so homogeneous in the odes with only two of them deviating radically from their class centroids.

Table 2: Classification matrix: Sonnets – Ode stanzas

	Percent correct	Sonnets	Ode stanzas
Sonnets	88.23	15	2
Ode stanzas	91.30	2	21
Total	90.00	17	23

At the next stage of analysis the sonnets which Keats did not include into his collections were added (Sonnets [2]), bringing the number of classes to three: Sonnets [1] (17 texts), Sonnets [2] (18), Odes (5). The results of the discriminant analysis are displayed in Figure 2. This diagram shows a clearly marked separation between Sonnets [1] and the rest of the poems (Function 1). The opposition between Sonnets [2] and odes is almost completely neutralized.

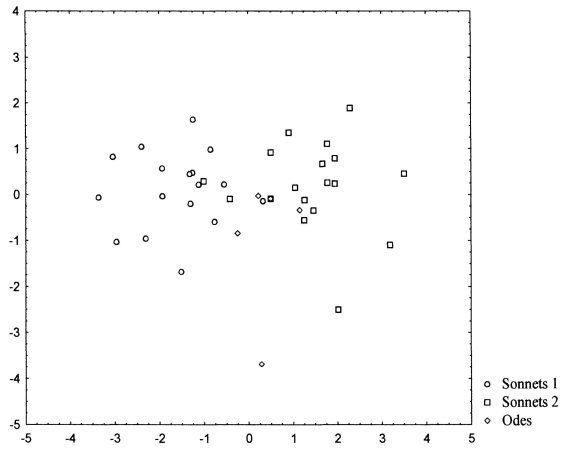


Figure 2: Discriminant analysis: three text groups (Sonnets [1] – Sonnets [2] – Odes

The discriminant model consists of the following characteristics: feminine and dactylic clausula, omission of stress on ictuses 3, 4, 5, inexact rhyme, partial inversion. The characteristics which are more typical of each of the three classes are presented in Table 3 (elements of Function 1) and Table 4 (Function 2). The second function, as it was mentioned above, has very little relevance in differentiating classes of Sonnets [2] and Odes.

Table 3: Elements of Function 1

Sonnets [1]	Sonnets [2] and Odes
Feminine rhyme	Dactylic rhyme
Unstressed ictus 4	Unstressed ictus 5
Partial inversion	Inexact rhyme

The members of this model are almost the same as the characteristics of the model differentiating Sonnets [1] and Odes. Partial inversion in this new model replaced Ictus 1 and Emphatic end from the previous one. Here, too, most of the features, except Ictus 3 and Partial inversion, reflect the properties of the end

Table 4: Elements of Function 2

Odes	Sonnets [2]
Inexact rhyme	Unstressed ictus 3
Unstressed ictus 5	Dactylic rhyme

of the line. But the efficiency of this new discriminant model is considerably lower than of the model which was obtained for two classes.

Post hoc classification (Table 5) is, generally speaking, not bad for three classes (80%), but it reflects the opposition of poems of the same genre: instead of the opposition of sonnets and odes here the opposition of Sonnets [1] and Sonnets [2] takes place. In other words the style difference in sonnets, selected by the author for his collection of works, and the rest of the sonnets is much more pronounced than style difference of different genres.

Table 5: Classification matrix: Sonnets [1] – Sonnets [2] – Odes

	Percent correct	Sonnets [1]	Sonnets [2]	Odes
Sonnets [1]	94.12	16	1	0
Sonnets [2]	83.33	2	15	1
Odes	20.00	1	3	1
Total	80.00	19	19	2

Figure 3 reflects the results of the analysis of the same three groups, when instead of complete odes their stanzas are taken. Here again the main opposition takes place between Sonnets [1] and other poems (discriminant Function 1). Discriminant Function 2 introduces a very slight distinction, if any at all, between Sonnets [2] and ode stanzas.

6 Conclusion

The results obtained prove the correctness of the hypothesis that in the lyrics by Keats there exist genre differences. The latter were revealed with the help of formal linguistic features. It should be stressed that none of these features, used here for the genre distinction, has ever been directly associated with either of the genres.

The results showed that odes are opposed to the sonnets mostly by the end of the line parameters. Genre differences are observed when we compare the texts that were selected by the author for his collections, whereas in other cases this distinction is obscured. A possible explanation is that Keats, choosing poems for his collections, was more conservative in matters of genre distinctions

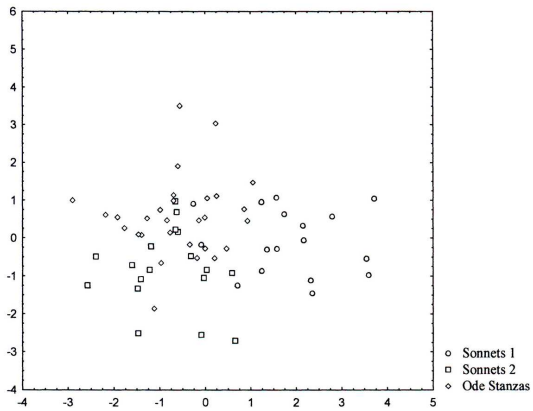


Figure 3: Discriminant analysis: three text groups (Sonnets [1] – Sonnets [2] – Ode stanzas

than when he wrote them. His preference for certain text style organization coincided with subconscious attempts at preserving traditional canons.

This phenomenon of distinction between works preferred by an author and works, which he values less, needs, of course, further investigation by means of enlarging the list of parameters (including, e.g. word and sentence length, parts of speech, syntactic relations characteristics), as well as increasing data sources. Different reasons may be offered for such discrepancies, but whatever explanation is suggested, if this phenomenon is observed for other writers, stylometric studies will have to consider carefully the selection of data sources.

References

Andreev, S.
2006 "A diachronic study of the style of Longfellow". In: Grzybek, P.; Köhler, R. (eds.), *Festschrift in honor of Professor Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin, New York: de Gruyter, 1–12.

Baevskij, V.
1993 *Pasternak – Lirik.* [= Pasternak the lyric poet]. Smolensk: Trust-Imakom.

2001 *Lingvističeskie, matematičeskie, semiotičeskie i kompjuternye modeli v istorii i teorii literatury.* [= Linguistic, mathematical, semiotic and computer models in the history and theory of literature]. Moskva: Jazyki slavjanskoj kul'tury.

Goldfield, J.; Hoover, D.L.
2008 "Homebodies and Gad-Abouts: A Chronological Stylistic Study of 19th Century French and English Novelists". In: Opas-Hanninen, L.L.; Jokelainen, M.; Juusso, I.; Seppanen, T. (eds.), *Digital Humanities. Conference Abstracts*. Oulu: University of Oulu, 117–120.

Grzybek, P.; Stadlober, E.; Kelih, E.; Antić, G.
2005 "Quantitative Text Typology: The Impact of Word Length". In: Weihs, C.; Gaul, W. (eds.), *Classification – The Ubiquitous Challenge*. Berlin: Springer, 53–64.

Juola, P.
2007 "Becoming Jack London", in: *Journal of Quantitative Linguistics*, 14; 145–147.

Kelih, E.; Grzybek, P.; Antić, G.; Stadlober, E.
2006 "Quantitative Text Typology: The Impact of Sentence Length". In: Spiliopoulou, M.; Kruse, R.; Nürnberger, A.; Borgelt, Ch.; Gaul, W. (eds.), *From Data and Information Analysis to Knowledge Engineering*. Berlin: Springer, 382–389.

Martynenko, G.
1988 *Osnovy stilemetrii.* [= Foundations of Stylometry]. Leningrad: Leningrad University.

Mikros, G.
2006 "Authorship attribution in Modern Greek newswire corpora". In: Uzuner, O.; Argamon, S.; Karlgren, J. (eds.), *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval, Seattle, USA, August 10, 2006*. 43–47.

Rudman, J.
1998 "Non-traditional Authorship Attribution Studies in the Historia Augusta: Some Caveats", in: *Literary and Linguistic Computing*, 13; 151–157.

2006 "Authorship Attribution: Statistical and Computational Methods". In: Brown, R. (ed.), *Encyclopedia of Language & Linguistics. Vol. 1*. Second Edition, Oxford: Elsevier, 611–617.

Tambouratzis, G.; Markantonatou, S.; Hairetakis, N.; Vassiliou, M.; Carayannis, G.; Tambouratzis, D.
2004 "Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the Feature Vector to Optimize Author Discrimination", in: *Literary and Linguistic Computing*, 19/2; 221–242.
Tarlinskaja, M.
1976 *English Verse: Theory and History.* The Hague, Paris: Mouton.
Warner, R.
2008 *Applied Statistics*. Los Angeles, London: Sage Publications.

Appendix A

Sonnets [1] (17 sonnets)

*To My Brother George; To ******; Written on the day that Mr. Leigh Hunt left Prison; "How many Bards gild the Lapses of Time"; To a Friend who sent me some Roses; To G.A.W.; "O Solitude! if I must with thee dwell"; To My Brothers; "Keen, fitful gusts are whisp'ring here and there"; "To one who has been long in City pent"; On first looking into Chapman's Homer; On leaving some Friends at an early hour; Addressed to Haydon; Addressed to the Same; On the Grasshoper and Cricket; To Kosciusko; "Happy is England I could be content".*

Sonnets [2] (18 sonnets)

To Lord Byron; "As from the darkening gloom a silver dove"; To Chatterton; Written in Disgust of Vulgar Superstition; On the Sea; "After dark vapours have oppressed our plans"; To Leigh Hunt, Esq.; On Seeing the Elgin Marbles; To Mrs Reynolds's Cat; On sitting Down to Read King Lear Once Again; "When I have fears that i may cease to be"; The Human Seasons; To Homer; To Sleep; "If by dull rhymes"; On Fame (I); On Fame (II); To — .

Odes (5 odes – 23 stanzas)

Ode to a Nightingale (8 stanzas); *Ode on a Grecian Urn* (5 stanzas); *Ode to Psyche* (4 stanzas); *Ode on Melancholy* (3 stanzas); *To Autumn* (3 stanzas).

Appendix B

Examples, illustrating charateristics, used in the analysis

Ictus 1:	After dark vapours have oppressed our plains / For a long dreary season, comes a day ("After dark vapours have oppressed our plains")
Ictus 2:	That scarcely will the very smallest shell (*On the Sea*)
Ictus 3:	Wasting of old Time – with a billowy main (*On Seeing the Elgin Marbles*)
Ictus 4:	I am no happy shepherd of the dell (*To* *****)
Ictus 5:	And, as I feasted on its fragrancy (*To a Friend who sent me some Roses*)
Stressed anacrusis:	Listen awhile ye nations, and be dumb (*Addressed to the Same*)
Feminine clausula:	I shall as soon pronounce which grace more neatly (*To G.A.W.*)
Dactylic clausula:	For I am brimfull of the friendliness ("Keen, fitful gusts are whisp'ring here and there")
Inexact rhyme:	Regions of peace and everlasting love; / [. . .] / Taste the high joy none but the blest can prove ("*As from the darkening gloom a silver dove"*)
Enjambement:	That is the Grasshopper's – he takes the lead / In summer luxury – he has never done / With his delights (On the Grasshopper and Cricket)
Syntactic pause:	Minion of grandeur! think you he did wait? (Written on the day that Mr. Leigh Hunt left Prison)
Partial inversion:	E'en now, Dear George, while this for you I write (*To My Brother George*)
Complete inversion:	Happy is England! I could be content / To see no other verdure than its own ("*Happy is England! I could be content"*)
Emphatically marked end of the line:	For what a height my spirit is contending! (*On leaving some Friends at an early Hour*)

Word-length-related parameters of text genres in the Ukrainian language. A pilot study

Solomija Buk, Olha Humenchyk, Lilija Mal'tseva, Andrij Rovenchak

1 Introduction

Text styles and genres are described in various linguistic fields – stylistics, communicative linguistics, gender linguistics, hermeneutics, rhetorics, statistical linguistics, etc. – from different points of view. Different parameters are applied to attribute the genres. But in general they do not contradict but rather supplement each other. However, none of these approaches, except statistical linguistics, proposes an automatic way to differentiate the genres. Presently, machine text processing is becoming more and more important, and a correct genre attribution is significant for automated translations.

The objective of the present paper is to check the possibility of genre attribution of Ukrainian texts using automatically obtainable parameters. The methods based on part-of-speech (PoS) or morpheme analysis (cf. Perebyjnis 1967; Karlgren and Cutting 1994) are not suitable as only tagged texts can be processed in such a way. The Ukrainian language is an inflectional one, so the PoS annotation is basic for it. A correct lemmatization of words in texts is, however, quite a complicated and long procedure.

Word-length studies are a good alternative because "raw" texts, with only little work on preprocessing, can be analyzed.

2 Parameters for genre attribution

While many text genres have been identified, only some of them have been subjected to a more detailed analysis, e.g., fiction (belles letters), journalistic or scientific texts. In this work, we focus on some less-studied genres: private letters, open letters, cooking recipes, sermons, sonnets, and parliamentary speeches. A couple of scientific texts are involved for comparison.

For the texts, various parameters connected with word length counted in syllables were calculated, in particular: mean word length, second central moment, dispersion quotient, fraction of multisyllabic words (i.e., those having four and more syllables), etc. In finding the set of variables providing the best separation to prescribe a correct genre attribution we rely on the results of Kelih et al. (2005).

The number of syllables was defined by counting the vowels, which correspond to the following graphemes: <а, е, и, і, о, у, я, є, ї, ю>. It is worth noting that auxiliary words having no vowel (б, в, з, й) were treated as zero-syllable words, not as clitics of the respective full-meaning words (cf. Grzybek and Altmann 2002; Buk and Rovenchak 2007). The following parameters were calculated:

1. mean word length in syllables m_1:

$$m_1 = \frac{1}{N}\sum_i x_i \,,$$

   where N is the number of words in a given text, x_i is the length of the i-th word;
2. dispersion of word length (second central moment) m_2:

$$m_2 = \frac{1}{N}\sum_i (x_i - m_1)^2 \,;$$

3. dispersion quotient d:

$$d = \frac{m_2}{m_1 - 1} \,;$$

4. fraction of four-syllabic words p_4:

$$p_4 = N_4/N \,,$$

   where N_4 is the number of four-syllabic words in the text;
5. fraction of five-syllabic words p_5; and some others.

Two variables, the dispersion quotient (d) and the fraction of four-syllabic words (p_4), make a pair of variables adequate for the separation of specific genres in Russian, which is also an East-Slavic language (see Kelih et al. 2005). The points corresponding to texts are plotted on the (d; p_4) plane (see Figure 1). We have analyzed 30 private letters, 20 open letters, 49 sermons, 30 parliamentary speeches, 29 sonnet wreaths, and 31 cooking receipts. Data for some genres exhibit a good concentration with respect to the dispersion quotient and the fraction of four-syllabic words variables, allowing separation, e.g., open letters versus private letters, epistolary genres versus scientific texts, etc. Sermons appear to occupy an intermediate region between open and private letters. Sonnets are well discriminated from cooking recipes. The high dispersion of sonnets is a bit unexpected and will be studied in more detail later.

The centers of distributions for studied genres shown in the figure are calculated as simple arithmetic means of the coordinates of the respective data-points. The centers of parliamentary speeches and open letters as well as the centers of sermons and private letters are closely located on the plane. Due to the specifics of these texts, this fact seems quite logical.

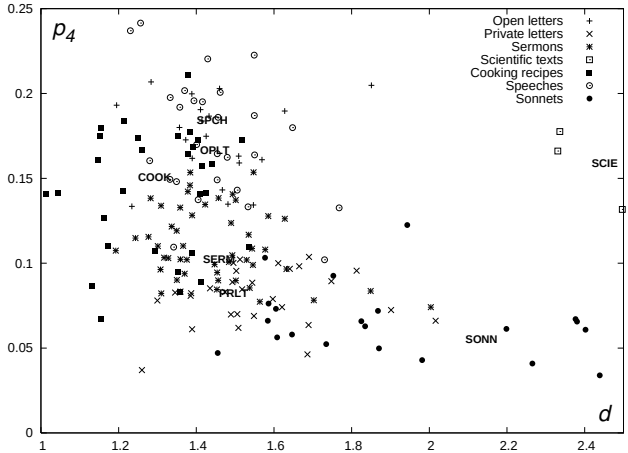


Figure 1: Texts of different genres on the (d; p_4) plane

The issue of homogeneity of texts on the intra-genre level (namely, authorship, subject, etc. differences) arises in the study of genres. We consider this problem for the particular case of sermons. Figure 2 demonstrates no special difference between sermons of different denominations (confessions) given by the following abbreviations: AC (Orthodox, Autocephalous), GK (Greek-Catholic), KP (Orthodox, Kyiv Patriarchate), MP (Orthodox, Moscow Patriarchate), RK (Roman-Catholic). The sample thus appears homogeneous.

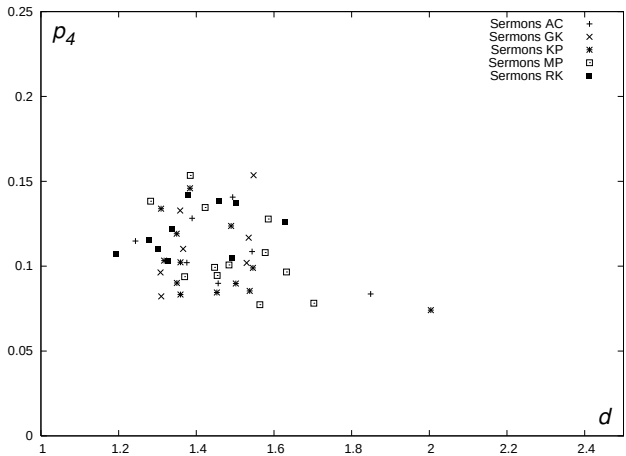


Figure 2: Sermons by denomination (using d and p_4)

The whole analysis shows that a third parameter might be necessary to achieve better separations. Again, it is convenient to search for such a parameter within the quantities obtainable automatically, which excludes methods based on part-of-speech or morphemic data as not quite suitable. Analysis of graphemic and phonemic behavior of text may be a promising alternative; in any case, more elaborate statistical methods must be applied.

3 Phoneme frequencies

All the texts were processed to obtain the phonemic data, according to the grapheme-to-phoneme scheme described by Buk et al. (2008). The phoneme distribution is obtained for particular text genres as well as for the whole corpus. Table 1 shows the results for the first six ranks.

Table 1: Most frequent phonemes, by genres

	Sermons		Cooking recipes		Open letters		Parliamentary speeches		Private letters	
r	P	f_r	P	f_r	P	f_r	P	f_r	P	f_r
1	о /ɔ/	0.10	о /ɔ/	0.10	а /a/	0.10	а /a/	0.10	а /a/	0.11
2	а /a/	0.10	а /a/	0.09	о /ɔ/	0.09	о /ɔ/	0.10	о /ɔ/	0.09
3	и /ɪ/	0.06	и /ɪ/	0.06	і /i/	0.07	і /i/	0.07	е /ɛ/	0.06
4	і /i/	0.06	і /i/	0.06	и /ɪ/	0.06	и /ɪ/	0.06	и /ɪ/	0.06
5	в /v/	0.06	у /u/	0.06	н /n/	0.06	е /ɛ/	0.06	і /i/	0.06
6	е /ɛ/	0.05	е /ɛ/	0.04	в /v/	0.05	н /n/	0.05	в /v/	0.05

The obtained rank-frequency dependencies (Figure 3) allow checking the hypothesis if the negative hypergeometric distribution (Wimmer and Altmann 1999: 465ff.) yields a good fit for phonemes.

We confirmed this fact obtaining the following values of the distribution parameters: $K = 3.2317; M = 0.8003$ ($C = 0.0085$, the whole text collection), $K = 3.1397; M = 0.7813$ ($C = 0.0140$, for a particular subcorpus of sermons). The results are shown in Figure 4.

4 Phonemes-related parameters

From the rank-frequency phonemic distributions, the following variables can be defined in particular: line slope between first and second most frequent phonemes with relative frequencies f_1 and f_2 : $s_{12} = f_1 - f_2$ or, more generally, line slope between *i*- and *j*-ranked phonemes: $s_{ij} = f_i - f_j$. The slopes s_{12}, s_{23}, and s_{45} are the most pronounced ones, cf. similar data on Polish (Rocławski 1981: 77ff.).

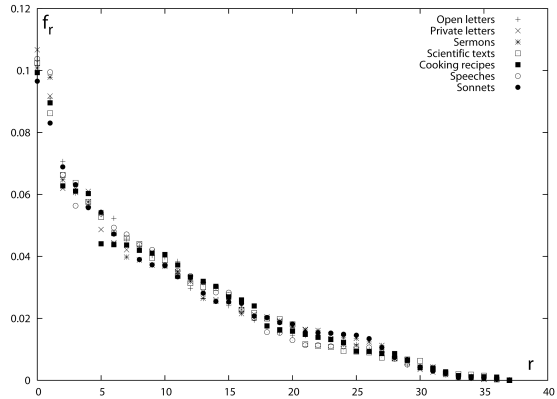


Figure 3: Distribution of phonemes in different genres

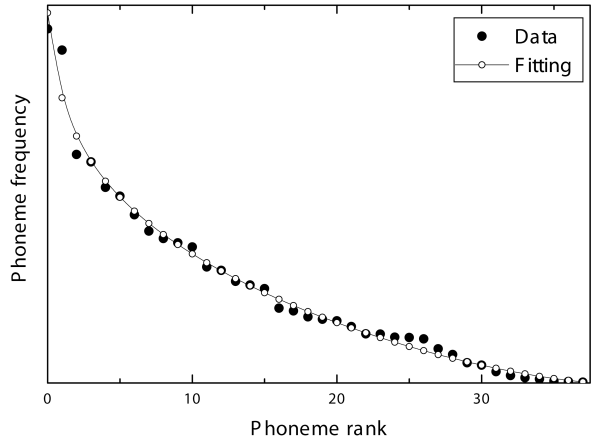


Figure 4: Fitting rank-frequency dependence for phonemes by the negative hypergeometric distribution (data analysis with Altmann Fitter 2.1)

One can see from Table 2 that parameters d and p_4 are not sufficient to distinguish some genres (e.g., open letters from parliamentary speeches, or private letters from sermons) as their values appear to be quite close. Other parameters related to word length, such as m_1 and m_2, do not help to solve this problem as they have a similar behavior. If the parameter s_{12} is considered, a better result for an automatic genre attribution can be achieved. Indeed, its mean value differs about twice in magnitude for the genres where other values are close. The sign of the parameter s_{12} corresponds to the slope "direction" and depends on which phoneme is most frequent, /ɔ/ or /a/. Further studies can establish if this sign is relevant.

Table 2: Discrimination by phonemes-related parameters (mean parameter values for genre discrimination)

Genre	m_1	m_2	d	p_4	p_5	s_{12}
Open letters	2.61	2.33	1.45	0.17	0.0814	−0.01064
Parliamentary speeches	2.54	2.18	1.43	0.18	0.0643	−0.00436
Private letters	1.96	1.49	1.56	0.08	0.0258	−0.01493
Sermons	2.15	1.66	1.45	0.11	0.0386	0.02979
Cooking recipes	2.33	1.71	1.29	0.15	0.0425	0.00977

5 Conclusions

From the presented material, we conclude that phoneme distribution can be a good addendum to word-length-related parameters in genre attribution. The task is to relate the parameters to genres properly, defining the domains of parameter variation for genres. Detailed analysis is required to achieve this goal, with more texts and genres involved. Other automatically calculated parameters might be necessary to obtain a better genre attribution. Multivariate discriminant analysis with respect to the calculated parameters, including word-length and phonemic frequency data, is yet to be applied.

Acknowledgments. We appreciate discussions with Emmerich Kelih on the issues presented. This research is done as a part of a joint Austrian-Ukrainian program (Project No. M/6-2009 from the Ministry of Education and Sciences of Ukraine and WTZ Project UA 05/2009 from ÖAD)

References

Buk, S.; Mačutek, J.; Rovenchak, A.
2008 "Some properties of the Ukrainian writing system", in: *Glottometrics*, 16; 63–79.
Buk, S.; Rovenchak, A.
2007 "Statistical parameters of Ivan Franko's novel *Perekhresni stežky [= The Cross-Paths]*". In: Grzybek, P.; Köhler, R. (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday.* Berlin, New York: Mouton de Gruyter, 39–48.
Grzybek, P.; Altmann, G.
2002 "Oscillation in the frequency-length relationship", in: *Glottometrics*, 5; 97–107.
Karlgren, J.; Cutting, D.
1994 "Recognizing text genres with simple metrics using discriminant analysis", in: *Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan, Vol. 2*, 1071–1075.
Kelih, E.; Antić, G.; Grzybek, P.; Stadlober, E.
2005 "Classification of Author and/or Genre? The Impact of Word Length". In: Weihs, C.; Gaul, W. (eds.), In: *Classification – The Ubiquitous Challenge.* Heidelberg: Springer, 498–505.
Perebyjnis, V.
1967 *Statystyčni parametry styliv.* [= Statistical parameters of styles]. Kyjiv: Naukova dumka.
Rocławski, B.
1981 *System fonostatystyczny współczesnego języka polskiego. [= Phonostatistical system of modern Polish]*. Wrocław: Zakład Narodowy imienia Ossolińskich, Wydawnictwo Polskiej Akademii Nauk.
Wimmer, G.; Altmann, G.
1999 *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

On the quantitative analysis of verb valency in Czech

Radek Čech, Ján Mačutek

1 Introduction

It is a matter of common knowledge in linguistics that verb valency is a verbal property which governs the other parts of a sentence. Although valency has been analysed in detail for more than fifty years (cf. Agel et al. 2004), some fundamental problems have not been resolved so far. For instance, no common criteria or tests for the distinguishing complements and adjuncts have been found, despite the fact that a distinction between them plays a crucial role in any valency approach (see Section 2). Since the absence of these criteria seriously undercuts the whole conception of valency, the question about the validity or the suitability of the valency approach emerged.

The goal of the present study is not to solve any of the fundamental problems of valency. We just decided to test empirically whether valency, in spite of the mentioned problems, reflects some important language property or mechanism. The only attempt, to our knowledge, to analyse valency empirically was presented in Köhler (2005a), where some properties of verb valency in German were observed: specifically the distribution of valency frames of each verb, the distribution of unique valency patterns, and the distribution of complement variants (a variant being the possibility to express a given complement of the verb by different semantic roles). Also the relationship between the number of complements of each verb and the number of complement variants was observed. In all cases regular distributions were detected which means that the distribution of observed entities could be viewed as a result of a diversification process (cf. Altmann 2005). In the present study we follow Köhler's methodological approach; we examine the distribution of valency frames in Czech and test the hypothesis concerning a relationship between the number of valency frames and word length.

The article is organized as follows: a very short overview of the main valency properties, in the "traditional" sense, is given in Section 2; valency hypotheses which were tested are presented in Section 3; Section 5 is focused on a methodology and language material used for the hypotheses testing; the results are presented in Section 4; and the article is closed by further research proposals.

2 Valency properties

Valency is usually viewed as a kind of a lexico-syntactic property which "involves the relationship between, on the one hand, the different subclasses of a word-class (such a verb) and, on the other, the different structural environments required by the subclasses, these environments varying both in the number and in the type of elements. Valency is thus seen as the capacity a verb has for combining with particular patterns of other sentence constituents" (Allerton 2005: 4878). In other words, valency "denotes the property of the verb to claim or to admit, respectively, particular kinds and forms of complements. The verb opens up slots, in which the complements enter as arguments" (Heringer 1993: 303). More concretely, valency determines

(1) the number of complements, compare monovalent verb *sleep*:
   a. Baby – sleeps
      versus bivalent verb *write*
   b. Mary – writes – the letter
      versus trivalent verb *give*
   c. Peter – gave – Mary – the book

(2) the form of the complements, compare verb *look* claiming adverbial complementation:
   a. Mary looks nice
      NOUN VERB ADVERB
      versus verb *bring* claiming nominal complementation
   b. Peter brought the book,
      NOUN VERB NOUN

(3) the meaning of the complements, compare the subject of the verb *see* which is assigned as the experiencer:
   a. Mary saw the house
      EXPERIENCER PATIENT
      versus the subject of the verb *kick* which is assigned as the agent
   b. Peter kicked the ball
      AGENT PATIENT

As we noted in Section 1, in any valency theory, a distinction between obligatory complements and facultative (optional) complements (they are usually called adjuncts) of the verb plays a crucial role. However, despite a huge endeavour (for more details see Buysschaert 1982, Herbst 2007, Panevová 1974, Storrer 1992, Van Valin & LaPolla 1997) to find common criteria or tests for distinguishing complements and adjuncts, a satisfying outcome has not been reached yet (Comrie 1993: 906ff.). So, some authors admit that "[t]he state of distinction into *C* [complement] and *A* [adjunct] and the position of valency

theory suggests that an intuitively substantiated basis (…) has not yet been sufficiently justified by theory. The different relational criteria – as far as they are methodically applicable in a controlled way – yield similar results in the majority of cases but also opposite ones. There are no adequate criteria to evaluate the quality of the results. (…) *It seems likely, however, that valency is a semantic phenomenon of which we have not yet found a clear view or which we perhaps have not even understood properly*" (Heringer 1993: 307; emphasis added by the authors).

It is clear that this fact seriously undermines the conception of valency in general. In other words, how can one seriously talk about "valency theory" without clear criteria for determining one of the most important properties of verb valency? Consequently, is not valency the notion which, although it fits one's intuition, does not reflect any important language mechanism? Or even, is it not just a matter of tradition?

Of course, the fact that the criteria have not been found yet does not necessarily mean that valency is an "empty" or senseless notion. However, if valency indeed reflects some important language property or mechanism, it is necessary, according to us, to prove the validity of this notion empirically. Therefore we tested two hypotheses concerned with (1) a regular distribution of valency frames in a language and (2) the relationship between the number of valency frames and the word length (several hypotheses on valency can be found in Köhler and Altmann 2009: 16ff.). So, if these hypotheses are not rejected, it seems reasonable to consider valency as a linguistically meaningful notion. Moreover, it will be possible to integrate valency to the synergetic linguistic framework (Köhler 2005b).

3 Valency hypotheses

3.1 Regular distribution of verb valency

Let us assume that valency, contrary to all problems related to the notion, reflects some important language mechanism and it could be considered as a verb classification enabling hypotheses testing and the exploration of relationships between valency and other language properties. One of the ways of evaluation of any classification scheme is an observation of rank-frequency distribution. It has been shown that "linguistic classification is 'good', 'useful' or 'theoretically prolific' if the taxa follow a 'decent rank-frequency distribution' " (Altmann 2005: 647). The regular distribution is viewed as a consequence of a diversification process and there is an assumption which says "that if an entity diversifies on one direction, the frequencies of the resulting classes are not equal but can be ordered according to decreasing frequency" (Altmann 2005: 646). So, if valency represents a "theoretically prolific" class, it should have a regular distribution.

3.2 The shorter the verb, the more verb valency frames

A relationship between the length of the verb and the number of valency frames of the given verb should be a consequence of the relationship between frequency and length. In other words, the shorter the verb, the more frequent the verb, and so the more frequent the verb occurs in more contexts, i.e. in more valency frames.

3.3 Language material and methodology

The crucial aspect of the testing of the hypotheses lies in both the choice of language material and the clear definition of valency. As for language data, we have used the Czech valency lexicon Vallex 1.0 (Lopatková et al. 2003) which contains about the 1400 most frequent Czech verbs.[1] Vallex 1.0 is based on Sgall's theoretical approach known as the Functional Generative Description (Sgall et al. 1986, Hajičová et al. 1998) and is closely related to the Prague Dependency Treebank project (Hajič et al. 2006).

As for definition of valency, we follow the Prague Dependency Treebank approach and we use the Vallex 1.0 annotation. In this study, we take into account only those verb modifications assigned as obligatory. The obligatoriness of a verb modification is determined by means of a so-called dialogue test in Vallex 1.0. The main principle of the dialogue test is defined as follows: "If [speaker] *A* uses a sentence *S* and [speaker] *B* asks him *wh*-question concerning the participant *P*, *A*'s answer might be "I don't know" (without disturbing the dialogue) if and only if the participant *P* is not semantically obligatory in *S*" (Panevová 1974: 15). More concretely, in the dialogue (4) the answer "I don't know" is unacceptable, so the verb *come* has assigned obligatory complementation "direction-to" and it is taken as bivalent in Vallex 1.0, although it is properly used as monovalent in the "surface" sentence structure.

(4) A: My friends have come.
B: Where to?
A: *I don't know.

On the contrary, in the dialogue (4) the answer "I don't know" is acceptable, so the complementation "direction-from" is optional.

1. Concretely, verbs were selected as follows: the 1000 most frequent Czech verbs, according to their number of occurrences in a part of the Czech National Corpus, were taken at the beginning and then their perfective or imperfective aspectual counterparts were added, if they were missing. For more details, see Vallex's 1.0 official web pages: `http://ufal.mff.cuni.cz/vallex/1.0/` and the technical report (Lopatková et al. 2006).

(5) A: My friends have come.
B: Where from?
A: I don't know.

For the hypotheses testing we counted verb valency frames which consist just of obligatory complementation (Vallex 1.0 comprises also other types of complementation; these ones we omit in this study). It is necessary also to note that we just counted formally unique valency frames; this means that if the verb has, for instance, two identical valency frames (as a consequence of a semantic shift), we count only one.

4 Results

4.1 Distribution of valency frames

As it can be seen in Table 1, the distribution of valency frames is indeed regular – in fact, so regular that there are many distributions with a very good fit.

Table 1: Distribution of valency frames

x – Number of valency frames	Number of verbs with x valency frames
1	815
2	319
3	152
4	73
5	38
6	17
7	7
8	7
9	4
10	2
11	1
14	1
17	1

Tentatively, we present the fit of the Good distribution (cf. Wimmer and Altmann 1999: 219ff.),

$$P_x = C\frac{p^x}{x^a} \tag{1}$$

where a, p are parameters and C is a normalization constant. We obtain an excellent fit (in terms of the chi square goodness of fit test, with $P = 0.9693$,

$a = 0.6562$, $p = 0.6034$). We do not claim that the Good distribution should be a general model; here only the 'smoothness' or 'regularity' of the distribution is demonstrated. Most probably the model would have to be modified or generalized when data from more languages are available.

4.2 Relationship between verb length and number of valency frames

The hypothesis "The shorter the verb, the more valency frames" is also corroborated, see Table 2. We note that the verb length was measured in syllables and the infinitive form of verbs was considered.

Table 2: Mean length of valency frames

x – Number of valency frames	Mean length (in syllables) of verbs with x valency frames
1	3.40
2	3.14
3	2.97
4	2.71
5	2.45
6	2.41
7	2.00
8	2.57
9	1.50
10	1.50
11	2.00
14	1.00
17	1.00

Again only tentatively, we suggest the function $y = Cx^a e^{-bx}$ as a model. The suggested model is a special case of a very general scheme derived by Wimmer and Altmann (2005). The goodness of fit, although not so excellent as for the distribution of valency frames, is still satisfying ($R^2 = 0.8959$, with $C = 3.6675$, $a = 0.0308$, $b = 0.0834$). Some discrepancies (the observed values are not decreasing) can be caused by relatively small numbers of verbs with many valency frames (e.g., we have only one verb with 11 valency frames, which is one of two problematic cases).

5 Further research

The corroboration of the hypotheses presented in this study allows us to consider valency as an important property of the language, despite many obscurities associated with this notion in linguistics. Nevertheless, further analyses

should be done: first, it is necessary to observe valency properties in other languages; next, hypotheses predicting relationships between valency and synonymy, polysemy, frequency and the other language characteristics should be tested. A fresh view to valency could be achieved by analyses focused on valency "in use", meaning that the distribution of valency frames given by both obligatory and optional complements in actual language usage are the subject of the analysis.

Acknowledgments. Radek Čech was supported by the Czech Science Foundation GAČR (grant no. 405/08/P157: "Components of transitivity analysis of Czech sentences (emergent grammar approach)". Ján Mačutek was supported by the Austrian FWF Lise Meitner Program.

References

Agel, V.; Eichinger, L.M.; Eroms, H.-W.; Hellwig, P.; Heringer, H.J.; Lobin, H.
2004 *Dependenz und Valenz / Dependency and Valency: Ein Internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter.

Allerton, D.J.
2005 "Valency Grammar." In: Brown, E.K. (ed.), *The Encyclopedia of Language and Linguistics.* Amsterdam: Elsevier Science Ltd., 4878–4886.

Altmann, G.
2005 "Diversification processes." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 646–659.

Buysschaert, J.
1982 *Criteria for the Classification of English Adverbials.* Brussels: Koninklijke Academie.

Comrie, B.
1993 "Argument Structure". In: Jacobs, J.; Stechow, A.; Sternefeld, W.; Vennemann, T. (eds.), *Syntax. An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter, 905–914.

Hajič, J.; Panevová, J.; Hajičová, E.; Pajas, P.; Štěpánek, J.; Havelka, J.; Mikulová, M.
2006 *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Hajičová, E.; Partee, B.H.; Sgall, P.
1998 *Topic-Focus Articulation, Tripartite Structures, and Semantic Content.* Dordrecht: Kluwer.

Herbst, T.
2007 "Valency complements or valency patterns?" In: Herbst, T.; Götz-Votteler, K. (eds.), *Valency: Theoretical, Descriptive and Cognitive Issues.* Berlin, New York: de Gruyter, 15–36.

Heringer, H.J.
1993 "Basic Ideas and the Classical Model?" In: Jacobs, J.; Stechow, A.; Sternefeld, W.; Vennemann, T. (eds.), *Syntax. An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter, 297–316.

Köhler, R.
2005a "Quantitative Untersuchungen zur Valenz deutscher Verben", in: *Glottometrics*, 9; 13–20.
2005b "Synergetic Linguistics." In: Köhler, R.; Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 760–775.

Köhler, R.; Altmann, G.
2009 *Problems in Quantitative Linguistics 2.* Lüdenscheid: RAM-Verlag.

Lopatková, M.; Žabokrtský, Z.; Skwarska, K.; Benešová, V.
2003 "Vallex 1.0. Valency lexicon of Czech Verbs." Prague: Center of Computational Linguistics. `http://ufal.mff.cuni.cz/vallex/1.0/`

Lopatková, M.; Žabokrtský, Z.; Benešová, V.
2006 "Valency lexicon of Czech verbs VALLEX 2.0. Technical Report 34." Prague: ÚFAL MFF UK. `http://ufal.mff.cuni.cz/vallex/2.0/publ/06-techrep.pdf`

Panevová, J.
1974 "On verbal frames in functional generative description I.", in: *Prague Bulletin of Mathematical Linguistics*, 22; 3–40.

Sgall, P.; Hajičová, E.; Panevová, J.
1986 *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht: Reidel Publishing Company.

Storrer, A.
1992 *Verbvalenz. Theoretische und methodische Grundlagen ihrer Beschreibung in Grammatikographie und Lexikographie.* Tübingen: Niemeyer.

Van Valin, R.D.; LaPolla, R.J.
1997 *Syntax: Structure, Meaning, and Function.* Cambridge: Cambridge University Press.

Wimmer, G.; Altmann, G.
1999 *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.
2005 "Unified derivation of some linguistic laws." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 791–807.

A link between the number of set phrases in a text and the number of described facts

Łukasz Dębowski

1 Introduction

In this communication, we announce recent developments of a new probabilistic explanation of the Zipf law and the Herdan law. This direction of research was introduced by Dębowski (2006) and the following thesis, formalized by Dębowski (2009b), constitutes the core of the novel interpretation:

(1) If an n-letter long text describes n^β independent facts in a repetitive way, where $0 < \beta < 1$, then the text contains at least $n^\beta / \log n$ different set phrases.

This paper covers our results very briefly. A longer discussion of the linguistic relevance of our results, the detailed definitions, and the proofs of the theorems can be found in Dębowski (2009b, 2009c). We hope that our constructions provide a more plausible explanation of the Zipf law and the Herdan law for natural language texts than the well known monkey-typing explanations (Mandelbrot 1954; Miller 1957).

2 The main result

A formal version of the statement (1) was proved by Dębowski (2009b) as a mathematical theorem for the computational model of set phrases and the probabilistic model of texts as follows.

The definition of set phrases in texts. The set phrases contained in a text will be understood as letter chunks that are repeated within the text sufficiently many times. Formally, the letter chunks are defined as distinct nonterminal symbols in the *shortest grammar-based compression* of the text. Empirical correspondence between such letter chunks and set phrases or words in the linguistic sense was observed to a certain extent (Wolff 1980; de Marcken 1996; Nevill-Manning 1996; Kit and Wilks 1999).

Grammar-based codes are uniquely decodable codes which compress a string by transforming it into a special context-free grammar and then encoding the grammar as a less redundant string. An example of such a grammar is

$$\left\{\begin{array}{l} A_1 \mapsto A_2A_2A_4A_5\textbf{dear\_children}A_5A_3\textbf{all.} \\ A_2 \mapsto A_3\textbf{you}A_5 \\ A_3 \mapsto A_4\textbf{\_to\_} \\ A_4 \mapsto \textbf{Good\_morning} \\ A_5 \mapsto \textbf{,\_} \end{array}\right\}. \tag{1}$$

The constraint is that the grammar must generate only one string as its production. If we start the derivation with the symbol A_1 and follow the rewriting rules, we obtain the compressed text of a song:

Good morning to you,
Good morning to you,
Good morning, dear children,
Good morning to all.

In the shortest grammar-based compression of a longer text, nonterminals A_i often correspond to words or set phrases in the linguistic sense (like *New York*), especially if it is additionally required that the nonterminals were defined as strings of only terminal symbols (Kit and Wilks 1999). In the following, a lower bound for the number of distinct nonterminal symbols in the shortest grammar-based compression of a text will be given in terms of the number of independent elementary facts described by the compressed text.

The definition of facts and texts. Both the corpus of texts and the state of affairs repeatedly described in the corpus will be modeled as random variables. Let $Z_1, Z_2, Z_3, \ldots$ be the logical values (true or false), with respect to the random state of affairs, of certain systematically enumerated logically independent propositions. We assume that Z_k, when interpreted as random variables Z_k, are equidistributed and probabilistically independent. Such variables exist if the space of possible states of affairs is sufficiently complex, namely, if the possible states of affairs generate a nonatomic σ-field (Dębowski 2009a). Z_k's will be called (elementary) facts. On the other hand, let $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ be the consecutive letters of the corpus, X_i. We suppose that each elementary fact Z_k can be ultimately inferred from the corpus if we start reading it from an arbitrary position, i.e., given $X_{m+1}, X_{m+2}, X_{m+3}, \ldots$ for any m.

Formally, we shall assume that variables $X_i : \Omega \to \mathbb{X}$ take a finite number of distinct values and form a stationary strongly nonergodic finite-energy process. A *nonergodic process* $(X_i)_{i\in\mathbb{Z}}$ is a process such that there exist functions $s_k : \mathbb{X}^* \to \{0,1\}$ and IID variables $Z_k : \Omega \to \{0,1\}$ where $P(Z_k = z) = 1/2$, $z \in \{0,1\}$ and

$$\lim_{n\to\infty} P\left(s_k\left((X_i)_{i=m+1}^{m+n}\right) = Z_k\right) = 1 \tag{2}$$

holds for all $m = \ldots, -2, -1, 0, 1, 2, \ldots$ and $k = 1, 2, 3, \ldots$. On the other hand, a *finite-energy processes* is a process with exponentially dumped conditional

block probabilities (Shields 1997). Such a condition is satisfied for processes dithered with a small amount of IID noise (Shields 1997), so it also appears plausible in natural language modeling.

In contrast, it has been assumed in the monkey-typing explanations (Mandelbrot 1954; Miller 1957) that the corpus of texts is generated by an IID process or a finite-state hidden Markov process. We can distinguish among these kinds of processes by the value of excess entropy $E = I((X_i)_{i\leq 0}; (X_i)_{i\geq 1})$, which is the mutual information between the past and future of the process (Crutchfield and Feldman 2003). We have (i) $E = 0$ for an IID process, (ii) $E < \infty$ for a finite-state hidden Markov process, and (iii) $E = \infty$ for a strongly nonergodic process (Dębowski (2009a). We can say informally that strongly nonergodic processes convey an infinite amount of information in a repetitive way. On the other hand, IID processes and finite-state hidden Markov process convey only a very limited amount of repeated information.

The theorem proved as a formalization of the thesis (1) reads:

Theorem 1
Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary strongly nonergodic finite-energy process over an alphabet $\mathbb{X}$, i.e., $X_i : \Omega \to \mathbb{X}$. Assume that $\mathbb{X}$ is finite and that

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \tag{3}$$

holds for the set of well predictable facts

$$U_\delta(n) := \{k \in \mathbb{N} : P\,(s_k\,((X_i)_{i=1}^n) = Z_k) \geq \delta\}\,,$$

where $\delta \in (1/2, 1)$ and $\beta \in (0, 1)$.

Consider the vocabulary size $\mathbf{V}[\Gamma((X_i)_{i=1}^n)]$ of a $(|B(\cdot)|, \mathscr{G})$-minimal grammar transform $\Gamma : \mathbb{X}^+ \to \mathscr{G}$, where $B : \mathscr{G} \to \mathbb{Y}^+$ is an appropriate local grammar encoder. Then we have

$$\limsup_{n\to\infty} \boldsymbol{E} \left(\frac{\mathbf{V}[\Gamma((X_i)_{i=1}^n)]}{n^\beta (\log n)^{-1}} \right)^p > 0, \quad p > 1. \tag{4}$$

This theorem follows from two bounds for the entropy of a finite text $(X_i)_{i=1}^n$. The vocabulary size $\mathbf{V}[\Gamma((X_i)_{i=1}^n)]$ is the number of distinct nonterminal symbols in the shortest grammar-based compression $\Gamma((X_i)_{i=1}^n)$ of the text $(X_i)_{i=1}^n$. For the definitions of appropriate local grammar encoders see Dębowski (2009b). The bound in Theorem 1 is given only for the expectation of the vocabulary size and it holds only for certain kinds of grammar-based codes. The codes that are good are closer to the ideas of de Marcken (1996) than to the ideas of Kieffer and Yang (2000).

3 An example of a process

In the following we shall construct an example of a process to which Theorem 1 can be applied. According to this proposition, we obtain (4) if the process $(X_i)_{i\in\mathbb{Z}}$ satisfies four conditions:

(a) $(X_i)_{i\in\mathbb{Z}}$ is a process over a finite alphabet $\mathbb{X}$,
(b) $(X_i)_{i\in\mathbb{Z}}$ is stationary,
(c) $(X_i)_{i\in\mathbb{Z}}$ has finite energy, and
(d) $(X_i)_{i\in\mathbb{Z}}$ is strongly nonergodic and (3) is true.

We shall exhibit a very simple example of a stochastic process that satisfies these conditions. Although this example cannot be called a realistic stochastic model of texts in natural language, it can be given an intriguing linguistic reading.

First, observe the following. Properties (b)–(d), but not (a), are satisfied by the process $(S_i)_{i\in\mathbb{Z}}$ introduced below, if we put $\alpha = \beta^{-1}$:

Example 1
Let $(S_i)_{i\in\mathbb{Z}}$ be a stochastic process where variables

$$S_i = (K_i, Z_{K_i}) \tag{5}$$

assume values from an infinite alphabet $\mathbb{N}\times\{0,1\}$, variables K_i and Z_k are probabilistically independent, K_i are distributed according to a power law,

$$P(K_i = k) = k^{-\alpha}/\zeta(\alpha), \qquad \alpha > 1, \qquad \zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}, \tag{6}$$

and Z_k are equidistributed, $P(Z_k = z) = 1/2$, $z \in \{0,1\}$.

Variables $S_i = (K_i, Z_{K_i})$ can be given some formal semantic interpretation. Imagine that $(S_i)_{i\in\mathbb{Z}}$ is a sequence of consecutive statements extracted from a random collection of texts which describe the random state of affairs $(Z_k)_{k\in\mathbb{N}}$ at random but consistently. Each statement $S_i = (k,z)$ asserts that the value of a random k-th bit of the state of affairs is z, i.e., it affirms that $Z_k = z$ in such way that both the bit address k and its value z can be identified. Logical consistency of the description is reflected in the following property: If two statements $S_i = (k,z)$ and $S_j = (k',z')$ happen to describe bits of the same address ($k = k'$) then they always assert the same bit value ($z = z'$).

We supposed that a suitable process over a finite alphabet can be constructed as the stationary mean of a certain encoding of the process $(S_i)_{i\in\mathbb{Z}}$. The role of this encoding is to represent abstract statements S_i as strings $f(S_i)$ which consist of letters X_i. The following proposition has been proved:

Theorem 2
Let $\mu = P((S_i)_{i\in\mathbb{Z}} \in \cdot)$ *be the distribution of the process given by (5)–(6) and put* $\mathbb{X} = \{0,1,\square\}$. *Consider a coding function* $f : \mathbb{N} \times \{0,1\} \mapsto \mathbb{X}^+$ *given as*

$$f(k,z) = b(k)z\square, \tag{7}$$

where $b(k) \in \{0,1\}^+$ *is the binary representation of a natural number* k *stripped of the initial digit* 1. *The process* $(X_i)_{i\in\mathbb{Z}}$ *distributed according to the stationary mean* $P((X_i)_{i\in\mathbb{Z}} \in \cdot) = \overline{\mu \circ (f^{\mathbb{Z}})^{-1}}$ *satisfies conditions (a)–(d) for* $\alpha = \beta^{-1}$ *and* $\zeta(\alpha) > 4$ *(i.e., for* $\beta > 0.7728\ldots$*).*

Notation $\mu \circ \left(f^{\mathbb{Z}}\right)^{-1}$ denotes the measure of the process

$$\ldots\square b(K_{-1})Z_{K_{-1}}\square b(K_0)Z_{K_0}\square.b(K_1)Z_{K_1}\square b(K_2)Z_{K_2}\square\ldots, \tag{8}$$

where the coded statements $b(K_i)Z_{K_i}$ are separated by symbols $\square$. Measure $\overline{\mu \circ (f^{\mathbb{Z}})^{-1}}$ is the stationary mean of that measure (cf. Gray and Kieffer 1980). Informally speaking, we obtain the sequence $\ldots X_{-1}X_0.X_1X_2\ldots$ by shifting the sequence (8) with a certain random shift. Because of the separators $\square$, there is no problem in reading the addresses K_i and the values of Z_k from the shifted sequence. The constraint $\zeta(\beta^{-1}) > 4$ comes from satisfying the condition (c).

4 Conclusion

According to our theoretical results, the vocabulary size of certain grammar-based compressions provides an upper bound on the total amount of information repetitively expressed in the text, if we follow the compression procedures by de Marcken (1996) rather than those by Kieffer and Yang (2000). However, the preliminary experimental data by Dębowski (2007) indicate a stronger relationship. Namely, the number of distinct nonterminal symbols in de Marcken's compressions of texts in natural language is several orders larger than in similar compressions of monkey-typing texts. It appears that the vocabulary size of de Marcken's style of grammar-based compression can be actually used to discriminate between texts that convey different amounts of repeated information. The plausibility of this conjecture should be investigated systematically in an experimental way.

References

Crutchfield, J.P.; Feldman, D.P.
2003 "Regularities unseen, randomness observed: The entropy convergence hierarchy", in: *Chaos*, 15; 25–54.

de Marcken, C.G.
1996 *Unsupervised language acquisition.* Ph.D. thesis, Massachussetts Institute of Technology.

Dębowski, Ł.
2006 "On Hilberg's law and its links with Guiraud's law", in: *Journal of Quantitative Linguistics*, 13; 81–109.
2007 "Menzerath's law for the smallest grammars." In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text*. Berlin: Mouton de Gruyter, 77–85.
2009a "A general definition of conditional information and its application to ergodic decomposition", in: *Statistics and Probability Letters*, 79; 1260–1268.
2009b "On the vocabulary of grammar-based codes and the logical consistency of texts." `http://arxiv.org/abs/0810.3125`
2009c "Variable-length coding of two-sided asymptotically mean stationary measures." `http://arxiv.org/abs/0911.5318`

Gray, R.M.; Kieffer, J.C.
1980 "Asymptotically mean stationary measures", in: *The Annals of Probability*, 8; 962–973.

Kieffer, J.C.; Yang, E.
2000 "Grammar-based codes: A new class of universal lossless source codes", in: *IEEE Transactions on Information Theory*, 46; 737–754.

Kit, C.; Wilks, Y.
1999 "Unsupervised learning of word boundary with description length gain." In: Osborne, M.; Sang, E.T.K. (eds.), *Proceedings of the Computational Natural Language Learning ACL Workshop*. Bergen, 1–6.

Mandelbrot, B.
1954 "Structure formelle des textes et communication", in: *Word*, 10; 1–27.

Miller, G.A.
1957 "Some effects of intermittent silence", in: *American Journal of Psychology*, 70; 311–314.

Nevill-Manning, C.G.
1996 *Inferring sequential structure.* Ph.D. thesis, University of Waikato.

Shields, P.C.
1997 "String matching bounds via coding", in: *The Annals of Probability*, 25; 329–336.

Wolff, J.G.
1980 "Language acquisition and the discovery of phrase structure", in: *Language and Speech*, 23; 255–269.

Modeling word length frequencies by the Singh-Poisson distribution

Gordana Đuraš, Ernst Stadlober

1 Introduction

Throughout history the problem of modeling the distribution of word length was not only the interest of linguists, but also of scientists from other areas such as physics, mathematics and statistics. In 1851 the English mathematician and logician Augustus De Morgan was the first to point out the relevance of the length of a linguistic unit. He mentioned word length as a possible style characteristic which may be helpful as an indicator in determining authorship (cf. Lord 1958). Several other scientists have dealt with the same topic counting even the frequency with which words of a given length occur in a text. Using preferably graphical methods to represent the results obtained, they noticed that the word length is not only influenced by the individual style of an author, as De Morgan stated, but may also depend on other factors such as genre.

With regard to word length studies, the first probability model was constructed in the 1940s. Observing the distribution of word length measured in the number of syllables, S.G. Čebanov, a Russian military doctor, found the Poisson distribution to be the most appropriate general model for the Indo-European group of languages (cf. Best 2001, 2005: 261, Grzybek 2006: 26). Independently, the German physicist Wilhelm Fucks (1955, 1956a, 1956b) came to a similar conclusion. He aimed at a mathematical description of word formation through syllables by introducing a mixture of Poisson probabilities, known as Fucks' Generalized Poisson distribution, where the Poisson distribution is a special case of this general distribution model (cf. Fucks 1956a, 1956c). Under particular conditions also the Dacey-Poisson distribution can be derived as a two-parameter special case of the proposal of Fucks (cf. Antić et al. 2005).

An intensive examination of word length began with Grotjahn's (1982) modification of Fucks' approach. He introduced mixtures of distributions where the parameter of the Poisson distribution is considered as a random variable following the gamma distribution. The resulting marginal model is the well-known negative binomial distribution and provides a good fit, at least for German texts.

Since the texts studied contained no zero-syllable words,1-displaced versions of the models mentioned above became relevant for further analysis. Trying to figure out under which empirical conditions 1-displaced Poisson

models may be adequate for word length frequencies, Grotjahn (1982: 53) proposed considering the index of dispersion (i.e., the variance to mean ratio) $\delta = \sigma^2/(\mu - 1)$ and estimated it by the empirical value $d = s^2/(\bar{x} - 1)$. Since the 1-displaced Poisson distribution has $\delta = 1$, this model provides adequate fit only for empirical samples with $d \approx 1$, i.e. for count data where the sample mean-1 is near to the sample variance. However, quite frequently count data are over-dispersed ($d > 1$) or under-dispersed ($d < 1$). Following Grotjahn's approach, it is obvious that $\delta > 1$ holds for the 1-displaced negative binomial models and $\delta \leq 1$ for 1-displaced Dacey-Poisson distributions. This means that negative binomial distributions are likely to be adequate for empirical samples with over-dispersion, while Dacey-Poisson models are suitable in case of under-dispersion. As to the problem of Poisson over-dispersion many models have been suggested in the literature using mixtures of Poisson distributions and Poisson-stopped-sum distributions (cf. Johnson et al. 2005). Contrary to this, under-dispersion has been given much less attention in the literature.

In this paper, our aim is rather to find a general model for word length frequency distributions that covers the whole d range. The model should be unique for all texts under study and should have at most two parameters. Therefore, we propose the Singh-Poisson model which is applicable to count data with over-dispersion, equi-dispersion and under-dispersion. In Section 2 the frequency distribution of word length is introduced. Section 3 gives an overview of data used in this paper. The definition and main properties of the Singh-Poisson model are given in Section 4. Section 5 discusses the estimation of Singh-Poisson parameters by three different methods. In Section 6 the performance of the estimation procedures is evaluated by a simulation study. The goodness of fit and discrepancy index C for verifying appropriateness of the Singh-Poisson model are examined in Section 7. Finally, in Section 8, the Singh-Poisson model is applied to 120 Slovenian texts selected. Section 9 summarize the results of the study.

2 Word length frequency distribution

Let $w_1, w_2, \ldots, w_n$ denote different words in a linguistic text of size n, where w_j refers to the j-th word in the text. The length of the word w_j, denoted by $l(w_j)$ is measured by the number of syllables per word consistent with the principles of automatic text analysis developed in the framework of the Graz project on quantitative text analysis (cf. Antić et al. 2006). In the process of automatically counting word lengths each text is submitted to certain tagging procedures, as for example treatment of titles, subtitles, numbers, etc.[1] According to this principles zero-syllable words as a part of the subsequent word

1. For further details as to tagging procedures see `http://www-gewi.uni-graz.at/quanta`.

do not exist as a separate word class, rather they are "lost" in the process of word length determination. Therefore, a certain linguistic text can be seen as a collection of one, two, three, or maximally k syllable words. By counting the number of words of the same length $l(w_j) = i,\ i = 1,\ldots,k$, we observe frequency f_i with which words of a certain length i appear in a given text. The number of elements f_i that belong to the same frequency class i is called 'absolute frequencies'. Relative frequencies, denoted by p_i, are calculated as f_i/n. The total number of words in a given text (text length), n, the absolute frequencies, f_i and the relative frequencies, p_i, are related through $\sum_{i=1}^{k} f_i = n$ and $\sum_{i=1}^{k} p_i = 1$. Collecting the words of the same length into one class leads to the frequency distribution of the word length or frequency distribution of i-syllable words. The random variable X denotes hereby the number of syllables per word and has the range $\{1,2,\ldots,k\}$. According to the fact that we have to deal with words that have at least one syllable, the model considered will be 1-displaced.

3 Data base of the study

The 120 Slovenian texts which serve as a basis for this study represent four different text types (journalistic, poems, private letters and prose), thirty texts of each text type being analyzed.[2] These texts have been systematically selected based on findings from recent word length studies published elsewhere (cf. Antić et al. 2006). Table 1 represents the composition of the sample with the two characteristic statistical measures, mean word length ($\bar{x}$) and sample variance (s^2).

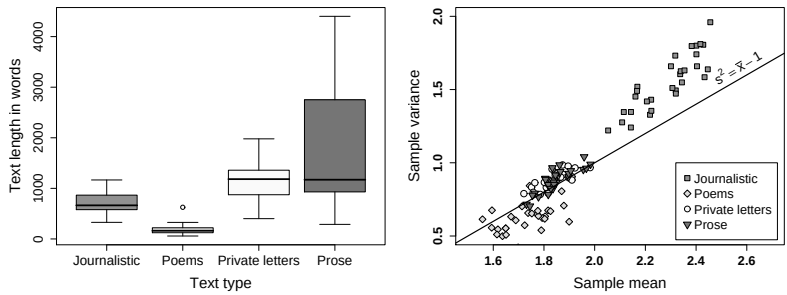
Table 1: Overview of 120 Slovenian texts

Author	Text type	Amount	$\bar{x}$		s^2	
			min	max	min	max
Journal *Delo*	Journalistic	30	2.05	2.46	1.22	1.96
S. Gregorčić, V. Vodnik	Poems	30	1.48	1.90	0.37	0.84
I. Cankar	Private letters	30	1.72	1.98	0.78	0.98
I. Cankar	Prose	30	1.73	1.98	0.70	1.04

Although not very different with respect to mean word length and sample variance, prose texts seem to be more dispersed than private letters regarding text length as Figure 1a clearly shows. Evidently, the shortest are the poems, followed by journalistic texts. Figure 1b shows the results of plotting sample

2. The text basis of this study is part of the text data base developed in the Graz research project "Word Length Frequencies in Slavic Texts", mentioned above.

mean versus sample variance for all 120 Slovenian texts. The solid black line is the Poisson reference line where equality of mean -1 and variance is present. Obviously, all journalistic texts lie above the Poisson line, private letters and prose texts above and below this line, while most of the poems lie below the Poisson line. Being aware of this fact and based on Grotjahn's (1982) findings, we calculated the index of dispersion, d, for each of the 120 texts under study. It turned out that $d > 1$ for all 30 journalistic texts. 27 out of 30 poems had $d < 1$, while in the case of private letters and prose texts all three possibilities for d were observed.



(a) Text length differences *(b)* Sample mean vs. sample variance

Figure 1: Differences regarding text type for 120 Slovenian texts

In searching for a suitable model for the word length frequency distribution of our sampled texts that covers the whole d range and has only two parameters, we found the Singh-Poisson model as the most appropriate one.

4 The Singh-Poisson model

The Singh-Poisson (S-P) distribution is a simple alternative to the Poisson distribution applicable in situations where the observed count data have $d \neq 1$ indicating that there is some deviation from the Poisson distribution. The S-P distribution is a special case of a finite mixture, known also as zero-modified Poisson distribution where the Poisson distribution is combined with a one-point (degenerate) distribution concentrated at zero (cf. Johnson et al. 2005: 351). It has two parameters denoted by α and θ. An important feature of the S-P distribution is its ability to model both over- and under-dispersion. Moreover, for $\alpha = 1$ the S-P distribution reduces to the standard Poisson distribution with parameter θ, the equi-dispersed case.

In its 1-displaced form, the probability mass function of a discrete random variable X having S-P distribution is given by

$$\pi_x(x|\alpha,\theta) = P(X=x) = \begin{cases} 1-\alpha+\alpha e^{-\theta}, & x=1 \\ \alpha\theta^{x-1}e^{-\theta}/(x-1)!, & x=2,3,\ldots \end{cases} \tag{1}$$

where where $\theta > 0$ and $0 < \alpha \leq \alpha_{\max} = 1/(1-e^{-\theta})$. Here, $\alpha_{\max}$ denotes the maximal possible value of α for given $\theta > 0$. As long as $0 < \alpha < 1$, the probability $P(X=1)$ is greater than $e^{-\theta}$, the 1-displaced Poisson probability at one and hence we have an *excess of ones* compared to the parent 1-displaced Poisson distribution. The higher proportion of ones (one-inflation) results in a reduction of the remaining frequencies by a corresponding amount. For $1 < \alpha \leq \alpha_{\max}$ there is one-deflation (cf. Table 2 and 3).

The first two moments of the distribution (1) are $\mathrm{E}(X) = 1+\alpha\theta$ and $\mathrm{var}(X) = \alpha\theta(1+\theta-\alpha\theta)$. The parameter α measures dispersion and hence tunes the type of the distribution. Justification is based on the index of dispersion, δ, given by

$$\delta = \frac{\mathrm{var}(X)}{\mathrm{E}(X)-1} = \frac{\alpha\theta(\theta+1-\alpha\theta)}{\alpha\theta} = 1+\theta(1-\alpha)\,. \tag{2}$$

Clearly, under-dispersion or over-dispersion is governed only by parameter α, as θ is positive. Obviously, for $\alpha = 1$ there is equi-dispersion ($\delta = 1$). For $0 < \alpha < 1$, δ is strictly greater then 1 and we have over-dispersion with respect to Poisson variation. The degree of over-dispersion increases as α decreases to zero. When α increases from 1 to $\alpha_{\max}$ the distribution becomes under-dispersed. For $\alpha = \alpha_{\max}$ we have $\delta = 1-\theta/(e^{\theta}-1) < 1$, which is obtained when $P(X=1) = 0$, i.e. when the first probability class disappears (cf. Table 2 and 3).

Table 2: 1-displaced case: Singh-Poisson vs. Poisson probabilities ($\theta = 0.8$)

$\theta = 0.8$	π_1	π_2	π_3	π_4	π_5	$\geq \pi_6$
$\alpha = 0.1$	0.945	0.036	0.014	0.004	0.001	0.000
$\alpha = 0.9$	0.504	0.324	0.129	0.035	0.007	0.001
Poisson	0.449	0.360	0.144	0.038	0.008	0.002
$\alpha = 1.1$	0.394	0.395	0.158	0.042	0.008	0.003
$\alpha_{\max} = 1.82$	0.009	0.647	0.259	0.069	0.014	0.002

In Tables 2 and 3 the probabilities of the 1-displaced S-P distribution are compared to those of the 1-displaced Poisson distribution for $\theta = 0.8$ and $\theta = 1.4$, respectively. As α is coming closer to zero, the higher is the proportion of ones in S-P compared to the Poisson case, while decreasing all remaining

frequencies. The degree of over-dispersion decreases as α approaches 1, also evident in the proportion of ones, which are now much closer to the Poisson probabilities. As soon as the parameter α is between 1 and $\alpha_{\max}$ we have one-deflation in the S-P which causes higher frequencies in the remaining classes compared to the Poisson case.

Table 3: 1-displaced case: Singh-Poisson vs. Poisson probabilities ($\theta = 1.4$)

$\theta = 1.4$	π_1	π_2	π_3	π_4	π_5	$\geq \pi_6$
$\alpha = 0.1$	0.925	0.035	0.024	0.011	0.004	0.001
$\alpha = 0.9$	0.322	0.311	0.217	0.101	0.036	0.013
Poisson	0.247	0.345	0.242	0.113	0.039	0.014
$\alpha = 1.1$	0.171	0.380	0.266	0.124	0.043	0.016
$\alpha_{\max} = 1.33$	0.021	0.449	0.314	0.147	0.051	0.018

5 Parameter estimation

In this section the three most common methods for finding estimators are discussed: method of moments (MM), maximum likelihood (ML) method and estimation based on sample mean and first frequency class (FF).

5.1 Estimation by method of moments

This method yields almost always some sort of estimates, provided the theoretical moments exist. By equating $\mu = 1 + \alpha\theta$, the theoretical mean and $\mu_{(2)} = 2\alpha\theta + \alpha\theta^2$, the second theoretical factorial moment, to their empirical counterparts $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} i f_i$ and $m_{(2)} = \frac{1}{n}\sum_{i=1}^{k} i(i-1) f_i$, respectively, one gets simple MM estimates of the parameters α and θ as

$$\hat{\alpha}_{\text{MM}} = \frac{\bar{x}-1}{\hat{\theta}_{\text{MM}}} \qquad \text{and} \qquad \hat{\theta}_{\text{MM}} = \frac{m_{(2)}}{\bar{x}-1} - 2\,.$$

5.2 Estimation by maximum likelihood

Consider a random sample of size n from a population with probability mass function $\pi_i(i|\Theta) = P_\Theta(X = i)$. Let f_i denote the observed frequency of class i such that $\sum_{i=1}^{k} f_i = n$, where k is the largest frequency class. The ML estimator of parameter vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$ is the value that maximizes the

likelihood function $L(\Theta|f_1,\ldots,f_k)$ given by

$$L(\Theta|f_1,\ldots,f_k)=\prod_{i=1}^{k}[P_\Theta(X=i)]^{f_i}.$$

Hence, the log-likelihood function $\ell(\Theta|f_1,\ldots,f_k)=\log L(\Theta|f_1,\ldots,f_k)$ is

$$\ell(\Theta|f_1,\ldots,f_k)=f_1\log P_\Theta(X=1)+\sum_{i=2}^{k}f_i\log P_\Theta(X=i). \tag{3}$$

Since in our case $\Theta=(\alpha,\theta)$, equation (3) consequently results in

$$\ell(\alpha,\theta|f_i)=f_1\log(1-\alpha+\alpha e^{-\theta})+\sum_{i=2}^{k}f_i[\log\alpha+(i-1)\log\theta-\theta-\log(i-1)!].$$

The score equations are obtained by equating the partial derivatives of $\ell(\alpha,\theta|f_i)$ with respect to the parameters α and θ to zero. These equations are given by

$$\frac{\partial\ell(\alpha,\theta|f_i)}{\partial\alpha} = \frac{(e^{-\theta}-1)f_1}{1-\alpha+\alpha e^{-\theta}}+\frac{1}{\alpha}\sum_{i=2}^{k}f_i=0, \tag{4}$$

$$\frac{\partial\ell(\alpha,\theta|f_i)}{\partial\theta} = \frac{-\alpha e^{-\theta}f_1}{1-\alpha+\alpha e^{-\theta}}+\frac{1}{\theta}\sum_{i=2}^{k}f_i(i-1-\theta)=0. \tag{5}$$

After a few algebraic simplifications we obtain the ML estimator $\hat{\theta}_{\text{ML}}$ of parameter θ as a solution of the transcendental equation

$$\frac{\hat{\theta}(n-f_1)}{n(\bar{x}-1)}+e^{-\hat{\theta}}-1=0.$$

After solving equation (4) for α, the resulting ML estimator $\hat{\alpha}_{\text{ML}}$ of parameter α is

$$\hat{\alpha}_{\text{ML}}=\frac{n-f_1}{n(1-e^{-\hat{\theta}_{\text{ML}}})}.$$

5.3 Estimation based on sample mean and first frequency class

This method is useful for situations in which the frequency of the first class in the sample is much larger than the other frequency classes or if the graph of the sample distribution is approximately *L*-shaped (cf. Anscombe 1950). The approach is to equate the sample mean $\bar{x}$ and the relative frequency of the first class f_1/n to the population mean $\mu=1+\alpha\theta$ and the probability of the first class $\pi_1=1-\alpha+\alpha e^{-\theta}$, respectively, in order to give more weight to this large frequency class. After some algebra, it can be shown that the corresponding parameter estimates $\hat{\alpha}_{\text{FF}}$ and $\hat{\theta}_{\text{FF}}$ are identical to the ML estimates $\hat{\alpha}_{\text{ML}}$ and $\hat{\theta}_{\text{ML}}$ given in Section 5.2.

6 A simulation study

To investigate the behavior of the Singh-Poisson model we performed a simulation study where all three situations were taken into account: (i) over-dispersion ($\delta > 1$), (ii) equi-dispersion ($\delta = 1$), (iii) under-dispersion ($\delta < 1$). As model parameters we choose (i) $(\alpha, \theta) = (0.82, 1.58)$, (ii) $(\alpha, \theta) = (0.92, 0.91)$ and (iii) $(\alpha, \theta) = (1.14, 0.63)$. These choices coincide with the ML estimates of each text aggregation for journalistic texts, private letters (i.e. prose) and poems, respectively in order to get *representative texts* of each text type. For each of the three situations $M = 500$ Monte Carlo samples of size $n = 500$ and $n = 1000$ are drawn. To generate S-P random variates we apply the inversion method (cf. Stadlober 1989: 7ff.) where the probabilities of the S-P distribution are computed using $\pi_1 = 1 - \alpha + \alpha e^{-\theta}$, $\pi_2 = \alpha\theta e^{-\theta}$ and the recurrence formula

$$\pi_x = \frac{\theta}{x-1}\pi_{x-1} \quad \text{for } x \geq 3.$$

The whole procedure was implemented by the public domain software *R*. For both sample sizes $n = 500$ and $n = 1000$ the results of the simulation study are summarized in Table 4. For each of the three data situations MM and ML estimates have been calculated. The corresponding mean values of $M = 500$ estimated parameters $\hat{\alpha}$ and $\hat{\theta}$ are displayed in the second and fourth column of Table 4. For both sample sizes, we observed similar results, independently of the estimation procedure applied. However, the standard errors of the estimated parameters are smaller for ML estimates and decrease with increasing sample size.

Table 4: Estimation results for over-, equi- and under-dispersed data situations

		$(\alpha, \theta) = (0.82, 1.58)$		
$\delta > 1$	$(\bar{\alpha}_{MM}; \bar{\theta}_{MM})$	$(se_{\bar{\alpha}_{MM}}; se_{\bar{\theta}_{MM}})$	$(\bar{\alpha}_{ML}; \bar{\theta}_{ML})$	$(se_{\bar{\alpha}_{ML}}; se_{\bar{\theta}_{ML}})$
$n = 500$	$(0.824; 1.579)$	$(0.041; 0.098)$	$(0.822; 1.581)$	$(0.034; 0.083)$
$n = 1000$	$(0.823; 1.578)$	$(0.029; 0.067)$	$(0.821; 1.579)$	$(0.022; 0.059)$
		$(\alpha, \theta) = (0.92, 0.91)$		
$\delta = 1$	$(\bar{\alpha}_{MM}; \bar{\theta}_{MM})$	$(se_{\bar{\alpha}_{MM}}; se_{\bar{\theta}_{MM}})$	$(\bar{\alpha}_{ML}; \bar{\theta}_{ML})$	$(se_{\bar{\alpha}_{ML}}; se_{\bar{\theta}_{ML}})$
$n = 500$	$(0.925; 0.909)$	$(0.070; 0.079)$	$(0.924; 0.909)$	$(0.061; 0.070)$
$n = 1000$	$(0.923; 0.910)$	$(0.048; 0.057)$	$(0.921; 0.911)$	$(0.042; 0.051)$
		$(\alpha, \theta) = (1.14, 0.63)$		
$\delta < 1$	$(\bar{\alpha}_{MM}; \bar{\theta}_{MM})$	$(se_{\bar{\alpha}_{MM}}; se_{\bar{\theta}_{MM}})$	$(\bar{\alpha}_{ML}; \bar{\theta}_{ML})$	$(se_{\bar{\alpha}_{ML}}; se_{\bar{\theta}_{ML}})$
$n = 500$	$(1.155; 0.627)$	$(0.103; 0.066)$	$(1.150; 0.629)$	$(0.094; 0.062)$
$n = 1000$	$(1.152; 0.625)$	$(0.075; 0.047)$	$(1.148; 0.627)$	$(0.068; 0.044)$

7 Goodness of fit

Whenever count data are considered and one wants to test the fit of a certain model, the standard procedure is to compare the expected frequencies with the observed frequencies. This means testing the null hypothesis

$$H_0 : p_i = \pi_i(\Theta), \;\; i = 1, 2, \ldots, k$$

that the observed frequency distribution is consistent with a particular theoretical distribution. For that purpose we calculate Pearson's chi square test statistics

$$X^2 = \sum_{i=i}^{k} \frac{(f_i - n\pi_i(\Theta))^2}{n\pi_i(\Theta)}$$

and reject the null hypothesis whenever the value of X^2 is large (within the critical region of the test).

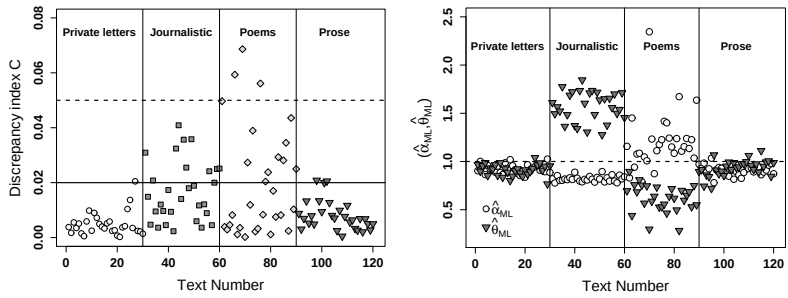
Here, $\Theta = (\alpha, \theta)$ and $\pi = (\pi_1, \pi_2, \ldots \pi_k)$ denotes the hypothesized S-P probability vector. The probability of the greatest word length class, π_k, is calculated as 1 minus the sum of all remaining classes, since there are no infinitely long words in language. However, the goodness of fit test has the disadvantage that it indicates only a statistical significance of a possible deviation from the model, but not the order of the departure. Hence, a direct comparison of goodness of fit values in the case of different sample sizes may be misleading. To avoid this problem we suggest instead the standardized discrepancy index $C = X^2/n$. Based on empirical investigations, we consider the fit of the model (a) as extremely good if $C \leq 0.01$, (b) as good if $0.01 < C \leq 0.02$ and (c) as acceptable if $0.02 < C \leq 0.05$.

8 Application to Slovenian texts

The results of fitting the Singh-Poisson model to 120 Slovenian texts are given in Figure 2a. The solid black line in the graphic is the reference bound $C = 0.02$, while the dashed line refers to $C = 0.05$. Obviously, the S-P model provides a good fit for the majority of the texts. It seems not to be appropriate just for the three poems of Gregorčić, which are indeed short texts. It would be interesting to study if and how goodness of fit may be influenced by text length, and in particular to consider the characteristics of short texts.

For all 120 Slovenian texts ML estimates of both parameters were computed and each pair of parameters $(\hat{\alpha}_{\text{ML}}, \hat{\theta}_{\text{ML}})$ was plotted versus the corresponding text, as shown in Figure 2b. The estimated parameters $\hat{\alpha}_{\text{ML}}$ are represented by circles, while estimated parameters $\hat{\theta}_{\text{ML}}$ are signified by triangles. It is evident that each group of texts leads to a different pattern of parameters. In case of private letters both parameters are very close to each other, the

same holds for prose texts although reversed in respect to the order. In contrast to this, in journalistic texts and poems parameters are quite distant from each other. The $\hat{\alpha}_{\mathrm{ML}}$ outlier in Figure 2b refers to Gregorčić's poem "Njega ni". This text has only 106 words, $\hat{\alpha}_{\mathrm{ML}} = 2.34$ and $\hat{\theta}_{\mathrm{ML}} = 0.3$ ($\alpha_{\max} = 3.88$). However, the C value of 0.0002 indicates here rather an extremely good fit.



(a) Discrepancy index *(b)* $(\hat{\alpha}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}})$ parameter regions

Figure 2: Results of fitting Singh-Poisson model to 120 Slovenian texts

9 Summary

The Singh-Poisson model with the two parameters α and θ is a simple generalization of the Poisson distribution with parameter θ. The new parameter α tunes the type of dispersion. It allows the modeling of under-dispersion ($1 < \alpha \leq \alpha_{\max}$), equi-dispersion (Poisson case $\alpha = 1$) and over-dispersion ($0 < \alpha < 1$). The estimation relies on maximum likelihood which leads in case of the Singh-Poisson distribution to the same estimates as the method based on the sample mean and the first frequency class. The proposed model offers a unified approach for all cases of under-, equi- and over-dispersion. In a simulation study we demonstrated the usefulness of the parameter estimates under three data-driven dispersion scenarios. Finally, the Singh-Poisson model is applied to 120 Slovenian texts and in all cases we obtained reasonable and stable estimates.

References

Anscombe, F.J.
1950 "Sampling theory for the negative binomial and logarithmic series distributions", in: *Biometrika*, 37; 358–382.

Antić, G.; Grzybek, P.; Stadlober, E.
2005 "Mathematical aspects and modifications of Fucks' Generalized Poisson Distribution (GPD)." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin: de Gruyter, 158–180.

Antić, G.; Kelih, E.; Grzybek, P.
2006 "Zero-Syllable Words in Determining Word Length." In: Grzybek, P. (ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 117–156.

Best, K.-H.
2001 "Biographische Notiz: Sergej Grigor'evič Čebanov (1897–1966)." In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt, 284–310.

2005 "Wortlänge." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin: Walter de Gruyter, 260–273.

Fucks, W.
1955 "Theorie der Wortbildung." In: Behnke, H.; Lietzmann, W.; Süss, W. (eds.), *Mathematisch-Physikalische Semesterberichte.* Göttingen: Vandenhoeck & Ruprecht, 195–212.

1956a "Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen", in: *Nachrichtentechnische Fachberichte*, 3; 7–21.

1956b "Statistische Verteilungen mit gebundenen Anteilen", in: *Zeitschrift für Physik*, 145; 520–533.

1956c "Mathematical theory of word formation." In: Cherry, C. (ed.), *Information Theory.* London: Butterworth, 154–170.

Grotjahn, R.
1982 "Ein statistisches Modell für die Verteilung der Wortlänge", in: *Zeitschrift für Sprachwissenschaft*, 1; 44–75.

Grzybek, P.
2006 "History and Methodology of Word Length Studies." In: Grzybek, P. (ed.), *Contributions to the Science of Language.* Dordrecht: Springer, 15–90.

Johnson, N.L.; Kemp, A.W.; Kotz, S.
2005 *Univariate Discrete Distributions.* New Jersey: Wiley & Sons.

Lord, R.D.
1958 "Studies in the history of probability and statistics. VIII: De Morgan and the statistical study of literary style", in: *Biometrika*, 45; 282.

Stadlober, E.
1989 *Sampling from Poisson, binomial and hypergeometric distributions: Ratio of uniforms as a simple and fast alternative.* (Bericht No. 303) Graz: Mathematisch-statistische Sektion.

How do I know if I am right? Checking quantitative hypotheses

Sheila Embleton, Dorin Uritescu, Eric S. Wheeler

1 Introduction

When we obtain a quantitative result in language study, there is always the nagging doubt that perhaps the technique (usually something fairly sophisticated) may have obscured the real nature of language. This is particularly so for someone more trained in language than statistics, but to some extent it should be everyone's concern. It is all too easy to accept the (often unstated) assumptions of a chosen technique without seeing how well they apply to the subject at hand.

In our use of Multidimensional Scaling (MDS) applied to geolinguistic data, we have been asked to justify the selection and completeness of our data – a fair question. After all, perhaps a little more (or less) data could give a quite different MDS analysis. We have tried to respond to this request with a small study of the stability of the MDS technique (Embleton et al. 2009). In it, we were able to show that the technique was "stable": a small change in the data (whether by choice or accident) would have minimal impact on the resulting MDS map. However, we also discovered in our study of dialect variation in the North-West region of Romania (Embleton et al. 2007, 2008) that the MDS technique did not demonstrate the range and type of variation that was commonly accepted as existing in that area. Sub-regions that were supposed to be quite distinct turned out not to be so, although some of the expected distinctions were clearly there. The existence of the conflicting view made us think more deeply about what the MDS technique (and the alternative) was actually measuring, and this has led us to propose a new concept of dialect variation.

The lesson we draw from these experiences is a simple one: it pays to have multiple approaches to a result, and it is wise not to accept the conclusions of any analysis, especially when the technique is quantitatively sophisticated, unless there is a methodologically independent way of getting the same answer. It may not be just a question of whether or not the mathematics was done correctly, but more importantly, whether our understanding of the results is appropriate.

2 Background

The Romanian Online Dialect Atlas (RODA) is a digitalized form of the first two volumes of a dialect atlas of North-West Romania (Stan and Uritescu 1996, 2003) and consists of software that permits the user to select files, search for patterns, view the search results, count the occurrences of a pattern by location and view the results as a map, create interpretive maps, hear samples of the data, and apply analytic tools to the data.

Our first analytic tool uses multidimensional scaling (MDS), a statistical technique for viewing a large number of relationships as a two-dimensional picture. We measure the linguistic similarity of all pairs of locations in our data set; the result is a 120-dimensional space in which each location is exactly its linguistic distance from every other location. But to visualize such a data structure, we apply MDS to produce a two-dimensional picture (a kind of "shadow" of the actual data structure) that displays the original relationships as closely as possible.

3 The stability of MDS

When we presented the results of some of our MDS analyses, we got questions about what would happen if the underlying data set had been more selective (e.g. had included only phonological data), or if the data set had errors or editorial judgements that others might question. What happens when the data set changes even a little? Does the MDS picture change dramatically?

In a simulation of such a situation (Embleton et al. 2009), we looked at randomly selected subsets of the underlying data, and compared the resulting distance matrices (the input to the MDS procedure). We used 10 test runs at each level ranging from 98% to 10% of the original data. There was some variation among the 10 test runs at each level, and we took the variation between the extremes as a measure of variation for that test level. As expected, the variation at each level was greater as the percentage of the original data decreased. However, the difference between the 100% case (the one picture that used all the data) and any of the test cases was always much less than the variation at that test case level.

In other words, the 100% picture was always a close representation of the other pictures. Our interpretation is that small changes to an MDS data set will not change the picture dramatically, and that we do not need to worry about potential data errors, or editorial changes altering the picture dramatically. It is possible that data selected according to some purpose such as a theoretical principle will have a distinct picture, but even then, it will have some resemblance to the full picture.

Our doubts about the "correctness" of our work, then, have been allayed by approaching the problem in a different way (in this case, a "simulation") and finding confirmation of what we originally expected.

4 The nature of dialect

A more challenging question for our work came when we compared our MDS analysis of the total Romanian data with earlier analyses of English and Finnish. The earlier analyses had shown that the dialect picture was close to the geographic map. That is, we saw distinct dialect areas, and the areas were arranged more or less according to geography. Geographically compact areas were also linguistically homogeneous. Where geography did not match linguistics, there was usually some reasonable explanation.

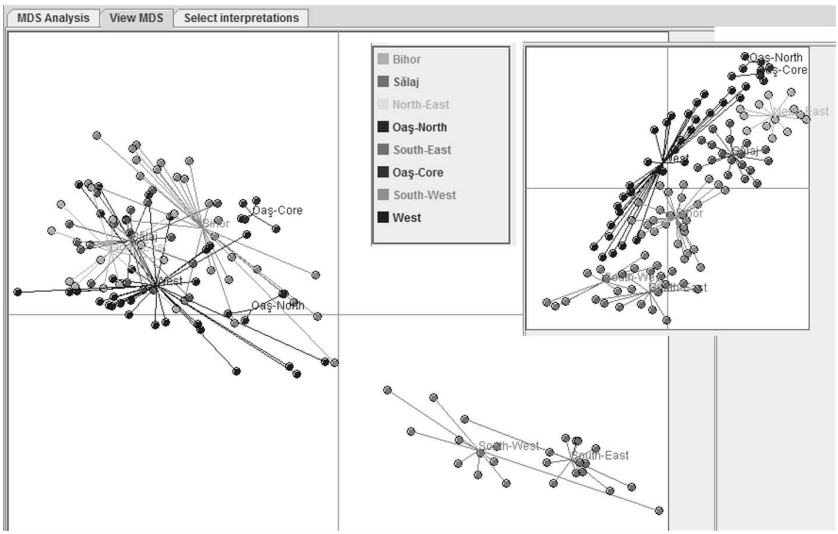


Figure 1: MDS picture of all data with a geographic map in the inset

But, the situation with the Romanian data was different. In our MDS analysis of the total data set (see Figure 1), we found that there were indeed two regions in the south that were linguistically distinct (as expected). The remaining dialect areas, however, were not distinct. Not only did most areas mix together what we anticipated would be separate dialect regions based on geography and traditional analysis, but also areas that we expected to be homogeneous were not. In particular, the region of Oaş in the far north of our area is in an isolated mountain valley; we expected it to be distinct from the rest of our area. In the

linguistic picture, it showed up as two areas, separated linguistically from one another by locations from further south.

However, in addition to our full data set, we have a collection of interpretive maps. Each interpretive map reflects the judgement of a trained scholar looking at a particular set of data. The interpretive map shows which dialect features are or are not present. When we run our MDS procedure on a large set of interpretive maps (237 in total), and on subsets reflecting phonology, morphology or lexicon, the MDS pictures show distinct dialect areas more clearly, and are more in line with the expected analysis of the data (see Figure 2). Note that the Finnish and English data was also of this interpretive type.

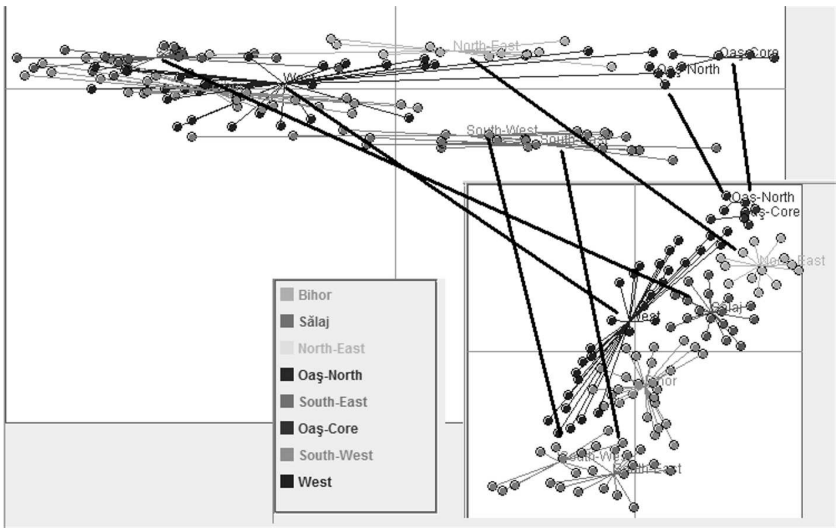


Figure 2: MDS picture of 237 interpretive maps, with a geographic map in the inset

So, which picture is correct? Unlike the stability of MDS, this question does not have a simple answer. On the one hand, we can argue that every piece of data is important, and that all of it should be considered in any quantitative analysis. If some of the data suggests a dialect division and some not, perhaps the two balance one another out. On the other hand, we can argue that there are some kinds of evidence that are more important than others, and that an educated and well-considered selection of data is more valuable for our understanding than a "brute-force" tallying of everything. But such a view can also be seen as "circular reasoning" in which we find distinctions because we selected only the data elements that show the distinctions.

Our response to this dilemma has been to say that the concept of dialect is more complex than simply dividing the geographic landscape between the "have"s and the "have-not"s. In a proposal made recently (Embleton, Uritescu

and Wheeler 2008c), we have suggested that there are a multitude of views possible for any dialect situation:

- At ground level, every linguistic pattern that can be used as a RODA search string will generate a dialect picture. For example, the alternation between /t/ and /ty/ gives a dialect picture; /s/ and /sy/ gives another; "pail" and "bucket" yet another. There are innumerable possibilities here.
- By some linguistic theory or other, ground level pictures can be aggregated, for example into a more general "phonological", "lexical" or "morphological" pattern. In a different sense, there are also innumerable possibilities here, because of the many choices of theory.
- There is conceivably one or more "top" level views that use as much data as possible.

Furthermore, there is a quantitative element to each picture because we count the number of occurrences of the pattern at each location. Thus, there is a picture for each chosen threshold for a dialect feature. In addition, for some of the patterns, there can be a quantitative element related to qualitative decisions, e.g. when is a vowel simply a "raised" variant of another vowel, versus a "different" vowel? There is no fixed answer to these questions. It is merely a case of having more parameters to consider, each choice providing a different "view" of the underlying dialect situation.

One imagines that a dialect situation is like a piece of marble, with dark and light bands running through it. Which pattern one sees will depend on where and how you slice into the marble. Of special interest are the bands that are persistently there in most or all slices. Some of the divisions will reflect real differences in the social and geographic structure of the area: for example, a syntactic innovation that spreads but does not cross a phonological dialect boundary could be reflecting a real social divide. Or, a set of lexical boundaries that are still transparent to phonological and morphological innovations may simply reflect an accident of history, and the boundary may no longer be real. It is the challenge of teasing out all the possibilities that makes it useful to think of a multifaceted "dialect situation", and each dialect map as a selected "view" of the underlying dialect situation.

5 Lesson learned

Quantitative analyses should and do raise doubts. Have I got it right? Some of the doubts can be settled by approaching the subject in a simulation, to actually "see" what is happening. By implementing the analysis this way, one gets a logically consistent and executable version that one can explore. The explorations can show not only the expert but also the interested non-expert what the analysis says. In this way, we were able to convince ourselves that the MDS method would not fall apart simply because of small changes in the source data

set. Some of the doubts will be more fundamental. Does this analysis really reflect what is happening in the real world? Why do the results not reflect our earlier expectations? There may be no simple answer to these questions, but in asking them, we may be led to a different perspective on our subject, and to the recognition that our analyses are not wrong, but rather that they are innovative.

References

Embleton, S.; Wheeler, E.S.

1997a "Multidimensional Scaling and the SED Data". In: Viereck, W.; Ramisch, H. (eds.), *The Computer Developed Linguistic Atlas of England.* 2nd ed., Tübingen: Max Niemeyer, 5–11.

1997b "Finnish Dialect Atlas for Quantitative Studies", in: *Journal of Quantitative Linguistics*, 4; 99–102.

2000 "Computerized Dialect Atlas of Finnish: Dealing with Ambiguity", in: *Journal of Quantitative Linguistics*, 7; 227–231.

Embleton, S.; Uritescu, D.; Wheeler, E.

2003 "Romanian Online Dialect Atlas." *International Colloquium of IQLA – International Quantitative Linguistics Association, University of Georgia, Athens, Georgia, May, 2003.*

2004 "Romanian Online Dialect Atlas. An exploration into the management of high volumes of complex knowledge in the social sciences and humanities", in: *Journal of Quantitative Linguistics*, 11/3; 183–192.

2006 "Seeing Words Change using the Romanian Online Dialect Atlas". *Presentation to International Linguistics Association. Annual Meeting. Toronto. April 2006.*

2007a *Online Romanian Dialect Atlas.* `http://vpacademic.yorku.ca/romanian`, now at `http://pi.library.yorku.ca/dspace/` under the "dialectology" community, "RODA" collection)

2007b "Romanian Online Dialect Atlas: Data Capture and Presentation." In: Grzybek, P.; Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the occasion of his 75th birthday* Berlin, New York: Mouton de Gruyter. 87–96.

2008a *Digitalized Dialect Studies: North-Western Romanian.* Bucharest: Romanian Academy Press.

2008b "Defining User Access to the Romanian Online Dialect Atlas", in: *Dialectologia et Geolinguistica*, 16; 27–33.

2008c "Identifying Dialect Regions: Specific features vs. overall measures using the Romanian Online Dialect Atlas and Multidimensional Scaling." *Leeds, UK: Methods XIII Conference. August 2008.* (to be published)

2008d "Data management and linguistic analysis: Multidimensional scaling applied to Romanian Online Dialect Atlas." *Trier Symposium on Quantitative Linguistics, Trier, Germany, December 2007.* In: Köhler, R. (ed.), *Issues in Quantitative Linguistics.* Lüdenscheid: RAM-Verlag, 10–16.

2008e "Lessons from Digitizing a Dialect Atlas." *International Conference on Linguistic, Literary and Ethnolinguistic Communication in the New European Context, Iasi, Académie roumaine, Institut de Philologie roumaine 'Al. Philippide', September 2008.*

2009 "The Stability of Multidimensional Scaling over Large Data Sets: Evidence from the Digitized Atlas of Finnish." In: Havu, E.; Helkkula, M.; Tuomarla, U. (eds.), *Du côté des langues romanes. Mélanges en*

l'honneur de Juhani Härmä. Helsinki: Société Néophilologique, 207–214.

Stan, I.; Uritescu, D.

1996 *Noul Atlas lingvistic român. Crişana.* Vol. I. Bucharest: Academic Press.

2003 *Noul Atlas lingvistic român. Crişana.* Vol. II. Bucharest: Academic Press.

Text difficulty and the Arens-Altmann law

Peter Grzybek

1 Introduction

The study of text difficulty is considered to be an important issue for many branches of applied research. In the fields of journalism or education, for example, it is particularly important to know if (or to what degree) a given text is likely to cause difficulties for a recipient, or a group of recipients, i.e., if it is likely to be on an adequate (intended) level of difficulty or beyond.

In order to achieve this goal, a specific line of text difficulty research has developed over the last decades, beginning in the 1920s, which attempts to combine linguistic analysis with informants' ratings of text difficulty.[1] Text difficulty thus is a double-faced kind of empirical research in two directions, either of which may be emphasized in individual studies: it is text-based, on the one hand, and informant-oriented, on the other. Due to this dual perspective, alternative terms such as 'text readability'[2] or 'text comprehensibility' have been used to refer to the related area(s) of research, the first term emphasizing the predominantly written (rather than oral) basis of communication, the second being broader in its understanding. Compared to these alternatives, 'text difficulty' as a term primarily refers to the analysis of linguistic structures, aiming at the identification and characterization of linguistic factors rendering a given text more or less easily comprehensible to a given person (or a group of persons), and at the (cor)relation of these structures to informants' ratings about text difficulty. Such a definition is in line with research from the last decades: Klare (1963: 1), for example, understands this term as referring to "the ease of understanding or comprehension due to the style of writing", and DuBay (2004: 3), more recently, has defined the overall aim of text difficulty research as the study of "what makes some texts easier to read than others".

Given this general orientation, research in this field, from its beginnings on, has continually tried to develop, modify and improve formulae to predict text difficulty and, by way of it, prognose comprehension ability. This is to say that attempts have been undertaken to develop measures of text difficulty, including formulae which combine quantitative (or quantified) linguistic characteristics in such a way that these characteristics serve as (possibly combined and

1. Informants may be either be recipients, mainly readers, or experts in the given field, such as teachers, librarians, publishers, lecturers of publishing houses, etc.
2. 'Text readability' in turn should not be confused with 'text legibility' which concerns factors such as typeface and layout of texts.

specifically weighted) factors for an optimized prediction of text difficulty. In the history of research[3] starting in the early 1920s, a number of relevant phases can be distinguished, in which researchers have tried to identify linguistic factors to be good indicators and predictors of text difficulty[4]. Early work as e.g. by Lively and Pressey (1923) mainly concentrated on lexical analysis; here, two major approaches can be distinguished: research concentrated on either the (relative) number of different words in a given text[5], or on references to frequency lists.[6] Subsequent work attempted to enlarge the linguistic spectrum and identify further factors, guided by the principle «The more, the better»: thus, authors like Gray and Leary (1935) already used a collection of 64 linguistic variables. Later, possible interactions between different linguistic factors became focused, in order to arrive at higher levels of correlation between attributed text difficulty and the combination of a set of linguistic variables. In this direction, two important results were obtained: first, many linguistic variables were highly intercorrelated, and second, an increase of the number of linguistic variables did not generally raise the correlation coefficient. Since, therefore, the use of more variables may be only minutely more accurate, but much more difficult to measure and apply, the next step included the reduction of variables and the identification of maximally predictive factors.

As a consequence, many different formulae were developed over the following years; Klare (1981) noted there were over 200 published formulae to measure text difficulty. All of these formulae have been developed by inductive-empirical approaches, typical for research in this field. Most of these formulae differ less as to the linguistic factors included, rather than how they are weighted. Among those factors re-occuring most frequently in all these formulae, are factors such as word frequency, amount of different words, average sentence length, average word length, and others (cf. Amstad 1978: 48f.).

From the perspective of quantitative linguistics in general, and synergetic linguistics, in detail, the high degree of relatedness between the various linguistic factors is not surprising; after all, it is well-known that both frequency and length characteristics of linguistic units on all analytical levels are closely

3. Since there are a number of informative surveys on this topic, this need not be presented here in detail.
4. Klare (1963: 4), for example, has distinguished between four phases of development: according to him, the early 'pioneer phase' (1921–1934) was followed be the development of detailed (1934–1938), efficient (1938–1953) and specialized (1953ff.) formulae.
5. This approach is well-known today as the study of 'lexical richness', usually including some kind of lexical type-token ratio. As we know today, there are quite a number of theoretical problems with this approach as, e.g., the dependence of the type-token ratio on text length. Additionally, it should be mentioned that in these early studies, no specific definition of 'word' has been used and, as a consequence, no distinction between 'word' and 'word form' (or lemma) has been made.
6. The early studies were mainly based on E.L. Thorndike's (1921, 1932), or Thorndike's and Lorge's (1944) lexical frequency analyses; later studies rather referred to G.K. Zipf's works as a reference line, which are better known today.

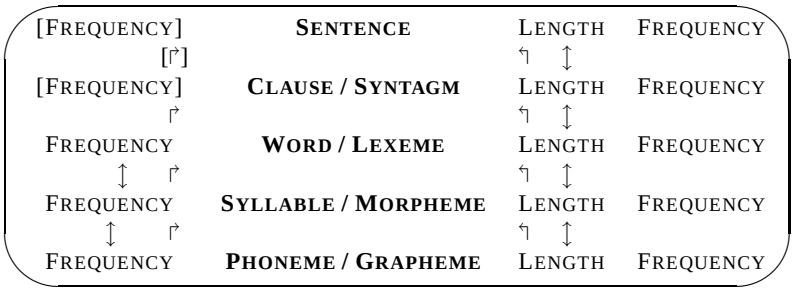
related and mutually interwoven. As a consequence, it is almost self-evident that if text difficulty is to be measured by reference to linguistic characteristics, it is sufficient to concentrate on only a few factors.

In this respect, the Flesch Reading Ease Index (*REI*), developed by Flesch (1948) with regard to English texts, is probably the most quoted and one of the easiest to apply. It is the result of a "simple" linear regression, i.e. combination of the average word length (*WoL*) and average sentence length (*SeL*) of a given text as the only two relevant factors (in addition to a constant):

$$REI_{engl} = 206.835 - (1.015 \cdot SeL) - (84.6 \cdot WoL) \quad (1)$$

Although quite simple at first sight, this formula is still today considered to be very efficient[7] and probably it is just due to its easy application that it is continuing to be one of the most widely used to measure text difficulty. Last not least, it is just the ambivalence between simplicity and efficiency of this formula which has given rise to skepticism, partly motivated by the lack of the formula's theoretical foundation. In this context, the validity of this formula has been generally called into question emphasizing the fact that "isolated linguistic units" are no adequate means for measuring text difficulty.

This view contradicts, of course, the above-mentioned synergetic interrelations between linguistic units, the relevance of which for text difficulty research have hardly ever been theoretically reflected in the whole research area. Therefore Best (2006), in his critical analysis of this discussion, is fully correct in objecting and countering that there are no isolated units in language. Particularly the word may be seen in the center of 'horizontal' and 'vertical' interrelations; as is well-documented, the word is part of a complex control circuit, the most basic factors of which are word length, semantic complexity, cotextuality, and word frequency (cf. Köhler and Altmann 1986: 261). Other relevant elements of this self-regulating dynamic system are syllable/morpheme length, clause length, sentence length, etc., and their respective frequencies. The following schema illustrates some basic synergetic processes; it makes clear that frequency and length characteristics of linguistic units stand in close self-regulating relations:



7. In comparative studies, the Flesch formula has repeatedly turned out to be the most efficient of those which need no word list (cf. Amstad 1978: 64).

As a result, Best (2006) correctly concludes: "Readability formulae, based on sentence and word length, indirectly measure substantially more than is expressed in these formulae, due to the manifold interactions between linguistic units." This view contains, of course, no theoretical foundation as to the question which specific factors influence text difficulty in what way or to what degree; yet it offers a theoretically based post-hoc answer to the question why the reduction to only a couple of seemingly elementary factors has made this concept to have such a success story.

Notwithstanding this insight, there is a whole bunch of crucial questions which continue to be unsolved. A major problem is the language-specific character of Flesch's *REI*: as was pointed out above, formula (1) was originally devoloped for English texts in the late 1940s. In later attempts to apply this formula to other languages, it soon turned out that language-specific adaptations were necessary, mainly due to the interest of having results on a scale from 0 to 100 in each language. Thus, for example, for Dutch, French, Spanish, German and Ukrainian the following adaptations were suggested[8], all following the general expression $REI = C - a \cdot WoL - b \cdot SeL$:

$$REI_{dutch} = 195 - (0.66 \cdot WoL) - (2 \cdot SeL), \quad (1a)$$

$$REI_{french} = 207 - (73.6 \cdot WoL) - (1.015 \cdot SeL), \quad (1b)$$

$$REI_{german} = 180 - (58.5 \cdot WoL) - SeL, \quad (1c)$$

$$REI_{spanish} = 206.84 - (77 \cdot WoL) - (0.93 \cdot SeL), \quad (1d)$$

$$REI_{ukrainian} = 206.84 - (28.3 \cdot WoL) - (5.93 \cdot SeL). \quad (1e)$$

As can be seen, the language-specific differences between these formulae consist in different weights for *WoL* and *SeL*, i.e. in different parameter values for *a* and *b*. *WoL* and *SeL* thus represent two crucial factors in measuring text difficulty across languages; yet, either their importance as separate factors, or their specific interrelation (i.e., the relation between *WoL* and *SeL*), clearly differs for individual languages.

Unfortunately, no systematic cross-linguistic studies are availabe which might explain what causes, or motivates, the observed differences in weighting. From a theoretical perspective, Best's (2006) reference to the synergetic specifics of language offers a good starting point for research in this direction. In this context, particularly the $WoL - SeL$ relation has recently been studied in detail, both from an inter-textual and intra-textual perspective; whereas the first concentrates on relations within a given text (or groups of texts), the second compares more than one textual object and studies the relation between them. For both perspectives, law-like regularities have been postulated and demon-

8. Cf. Kandel and Moles (1958), Fernández Huerta (1959), Brouwer (1963), Amstad (1978), Partiko (2001: 257).

strated to exist. From an *intra-textual perspective*, we are concerned with the *Menzerath-Altmann law*, relevant for the relation between a given construct and its constituting components within a given text (notwithstanding the possibly intervening level of clauses coming into play, on an intermediary level between sentence and word). As compared to this, the *inter-textual relation* is covered by the *Arens-Altmann law*, based on the calculation of the mean length of words ($\bar{x}$) and sentences ($\bar{y}$) in a series of text samples, resulting in two vectors of arithmetic means ($\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$).

In order to gain insight into the specific role *SeL* and *WoL* play for text difficulty in the individual languages, it seems reasonable, therefore, to study relevant data on the background of the Arens-Altmann law. Since such a systematical approach has never been undertaken before, a first approach into this direction should start with one language only. But even with this restricting focus, it is of utmost importance to pay due attention to yet another circumstance: as recent analyses have shown (Grzybek et al. 2007, Grzybek and Stadlober 2007, Grzybek et al. 2008), both *WoL* and *SeL* are not constant within a given language (i.e., are not 'typical' of a given language as a whole); rather, they differ for specific discourse types within a language. It seems likely that this finding is also relevant for the $WoL - SeL$ relation, but this possibility, too, has never been submitted to systematical reflection.

In the following analyses, these objectives shall be pursued, using German language material, strictly controlling text type. Since the perspective should be cross-linguistic right from the beginning, it seems reasonable to immediately provide a meta basis adequate for comparison. In this respect, suggestions developed by Estonian scholar Tuldava in a series of articles (1993a,b), turn out to be of utmost importance, since they contain a language-independent formula of measuring text difficulty (TD), also based on *WoL* and *SeL*, only:

$$TD = WoL \cdot \ln(SeL) \; . \tag{2}$$

This formula has remained rather unknown in the field of text difficulty research. As a consequence, its efficiency has never been generally tested; specifically, no systematic comparisons with Flesch's *REI* or any one of its language-specific adaptations have ever been undertaken. Tuldava himself applied his formula (2) to a sample of 20 German texts of different types (text books, journalistic, literary prose, scientific). Comparing the results obtained to Flesch's original *REI* formula (1), rather than to Amstad's German adaptation (1c), Tuldava (1993a: 78) found a close rank correlation of $\varphi = -0.97$ between these two measures.[9] Tuldava did not attempt to establish a detailed regression equation, which would allow for the transformation of one measure to the other.

9. As a re-analysis of his data shows, this correlation is highly significant ($p < 0.001$), with the linear regression $TD^* = 7.16 - 0.051 \cdot REI$.

If this result were confirmed on a broader basis of linguistic material, this would mean that Tuldava's formula (2) could indeed serve as a basis for cross-linguistic comparisons, for which no language-specific parameter estimations would be needed. More importantly, this would be an important step in the direction outlined above, both in practical and theoretical respects:

- From a *practical* point of view, the application of Tuldava's parameter-free formula would not only imply the option of measuring text difficulty without knowlege of language-specific parameters (i.e., weights), but, in addition to this, the results obtained might easily be transformed to fit one of the 'established' Flesch measures mentioned above.
- From a *theoretical* perspective, insight might be gained as to the question how *WoL* and *SeL*, either as individual factors or as a complex combination in their self-regulating interrelation, influence text difficulty.

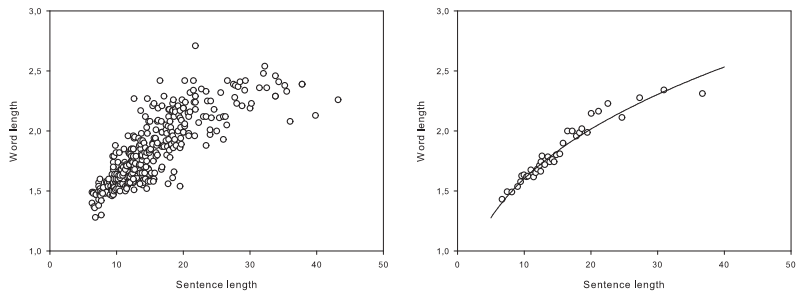
The detailed study of the $WoL - SeL$ relation is of utmost importance in yet another respect for text difficulty research: If *WoL* can be characterized to depend on *SeL*, as predicted by the Arens-Altmann law, then Tuldava's formula (2) might even be further reduced to one linguistic variable, only. At first sight, it might be equally plausible to substitute either the *WoL* or the *SeL* variable by the theoretical value to be expected according to the Arens-Altmann law; however, with *WoL* being the dependent variable, rather than *SeL*, it seems more appropriate to substitute the *WoL* variable, the more since the latter displays much less variation than *SeL* in a given text. In fact, the idea to substitute *WoL* has been brought forth by Tuldava (1993a), but it has never been empirically tested, due to insufficient research on the Arens law.

2 Analysis

As to appropriate data serving as material for our study, German texts analysed by Bamberger and Vanecek (1984) in their study on readability of school texts seem to be adequate. The authors investigated the readability of 380 texts from primary and lower secondary level textbooks; in detail, they analyzed 240 special texts [Sachtexte], and 120 literary prose texts for adults (i.e., youth literature). These texts were evaluated by an expert team according to their appropriateness for different school grades, each text being attributed to a particular difficulty level (*DL*). The authors then applied a variety of readability formulae, taking into account a large number of different linguistic factors which were tested for different levels from grades four through twelve. The linguistic characteristics of these factors are not relevant here; for our purposes, it may suffice to say that among others, average values for *WoL* and *SeL* were calculated for all texts, and these data shall serve for the subsequent re-analysis.

Figures 1a and 1b show the relation between *WoL* and *SeL* for the 380 texts: Figure 1a shows the original data points, in Figure 1b the latter are pooled in

groups of ten each, in order to make the overall tendency appear more transparent. As can be seen, there is an obvious trend of *WoL* to increase with increas-



(a) Unpooled data *(b)* Data pooled by 10

Figure 1: Dependence of *WoL* on *SeL* for 380 German texts from Bamberger and Vanecek (1984)

ing *SeL*; this tendency is particularly clearly expressed in Figure 1b. According to the Arens-Altmann law, this relation may be modeled by the function $WoL = a \cdot SeL^b$: in fact, with parameter values $a = 0.75$ and $b = 0.33$, the fit turns out to be very good ($R^2 = 0.95$), as can also be seen from the regression curve added in Figure 1b.

These findings are in accordance with the Arens-Altmann law and the hitherto undoubted assumption that, within a given language, the *WoL*-*SeL* relation on the inter-textual level can be modeled without distinction of text types. However, extending the data base of 380 texts by adding the above-mentioned 117 data sets from the original Arens (1965) study, analogically pooled by items of ten each, radically changes this view. Figure 2 clearly shows that the literary prose texts studied by Arens display the same overall trend of *WoL* increasing with an increase of *SeL*, but in a different way as compared to the schoolbook texts. This finding asks for a differentiated analysis of all three text types separately.

Figure 2 shows the resulting tendencies in detail: Quite obviously, there is an increase of *WoL* with an increase of *SeL* for all three text types.

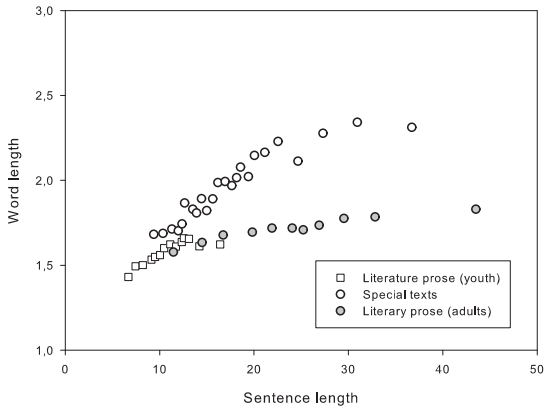


Figure 2: Dependence of *WoL* on *SeL* for 497 German texts (data pooled by 10)

Yet, the kind of increase differs for each of them; this fact is corroborated by the divergent parameter values, which are represented in Table 1.[10]

Table 1: Fitting results for three text types

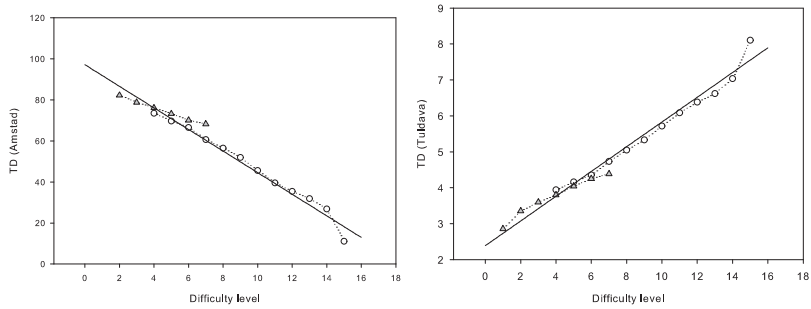
Texts	*N*	*a*	*b*	R^2
Youth literature	140	0.89	0.24	0.9956
Adult literature	117	1.23	0.11	0.9658
Special texts	240	0.69	0.37	0.9562

Summarizing, we can say that no simple substitution of either the *WoL* or the *SeL* variable is possible for Tuldava's *TD* formula, since the relation between *SeL* and *WoL* is not constant within a given language, but differs for text types. It is a task for future research to find out which and how many text types must be distinguished in this respect; it seems to be reasonable, however, to assume that we are concerned with the same kind of discourse types which have been identified to be relevant for the discrimination of discourse type on the basis of 'simple' *WoL* and *SeL* studies (cf. Grzybek et al. 2005; Kelih et al. 2006).

10. Interestingly enough, youth literature and adult literature seem to follow an identical kind of increase, though at different ends of the regression curve: joining the pooled data points for the 257 literary texts in a common type of 'literature' results in a good fit ($R^2 = 0.92$); in this case, we obtain parameter values for $a = 1.25$ and $b = 0.10$, which come very close to those adult literature. Nevertheless, the two literary text groups shall be treated separately in the subsequent analyses.

3 Text difficulty and text types

With these relations established, we can now come back to the question of text difficulty, separately for each of the two text types (i.e., 120 texts from youth literature and 240 special texts). Figures 3a and 3b present the results for Amstad's and Tuldava's formulae (1c) and (2), respectively.



(a) TD according to Amstad formula (1c) *(b) TD* according to Tuldava formula (2)

Figure 3: Text difficulty (*TD*) for 380 German texts (pooled by 20)

An inspection of Figure 3 allows for a number of important observations:

1. As expected, there is a clear major tendency of *DL* and *TD* being closely correlated; this tendency holds for both formulae, though with opposite directions. Ignoring text type specifics, the dependency is of clearly linear kind, with a high correlation coefficient of $r = 0.99$ in both cases.
2. Whereas there seem to be clear differences in the kind of relation between word and sentence length for the two text types – at least this was the result of the analyses discussed above (cf. Figure 2) –, the corresponding *TD* values seem to follow a common tendency (notwithstanding difficulty differences, of course). Obviously, particular text types have their own specific mechanisms of ruling *TD*, which allows, as a consequence, for a common analytical procedure. As long as no additional data change the picture, or further interpretations are available, it seems reasonable to consider the relation between *DL* and *TD* to be linear, across text types as well as within a given text type (with $r > 0.97$) in all four cases). Still, it remains an open question whether or not *TD* can be reasonably defined without taking into account text typological specifics.
3. Regardless of possible text typological specifics, it turns out that, at least for German, the language-independent measure for *TD* according

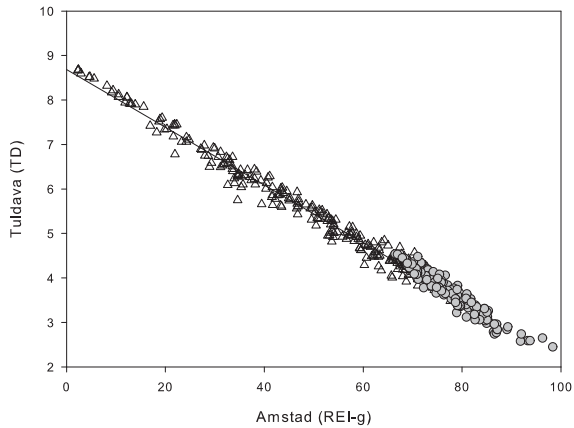


Figure 4: Comparison of Amstad's *REI* and Tuldava's *TD* indices for text difficulty

to Tuldava's formula (2) is equally efficient in predicting *DL* as is Amstad's language-specific adaptation (1c) of Flesch's *REI* to German, the correlation between both measures being highly significant ($r = 0.99$). Figure 4 shows the correlation between both measures, combined for both text types, but with distinct marks. This confirms Tuldava's above-mentioned observations on a broader data basis; additionally, it is based on the specific German adaptation of Flesch's *REI*, rather than on the original developed for English texts. In fact, both formulae turn out to measure in principle the same, though on different scales; as a consequence, they can be transformed one into the other. With regard to the 380 texts analyzed here, for example, the transformation from Tuldava's *TD* value to Amstad's scale might be easily calculated by way of the equation $REI_{german^*} = 133.09 - 15.24 \cdot TD$; alternatively, the transformation from Amstad's scale to Tuldava's value can easily be achieved by calculating $TD^* = 8.68 - 0.065 \cdot REI_{german}$. It goes without saying that, before generally applying these transformations to German texts, more text types must be studied, covering the whole textual spectrum. It is highly probable that this will result in a more or less considerable modification of these transformational procedures; by way of a comparison, the transformation from Flesch's original *REI* into Tuldava's *TD* would result in the equation $TD^* = 6.72 - 0.05 \cdot REI$, which also slightly differs from the figures given in footnote 9.

4 Results and conclusions

A first major result of the present study is the finding that, at least for German, text difficulty can be measured without any language-specific adaptation. As compared to Amstad's German adaptation of Flesch's *REI*, Tuldava's *TD* provides practically the same exactness of predictability, practically without loss of information. This finding is of relevance not only for German: if it can be corroborated for further languages, no language-specific adaptations, and no parameter estimations, will be necessary in future. Tuldava's formula may be considered to be universally valid; but this is a matter of boundary conditions in the individual languages; at present, we have no idea as to this point which represents an interesting linguistic question in its own right, namely, to what extent the formula works in which way (i.e., with which parameters) for which languages.

A second major result is that possibly no text typological specifics need to be taken into account when measuring text difficulty with Tuldava's *TD*: Since word length and sentence length are the only two characteristics taken into consideration in this formula, their interrelation has been submitted to a detailed analysis in this study. This analysis results in the observation that, within a given language, text typological differences do exist, but might not play a crucial role for measuring *TD*; rather, it seems possible that *TD* is the result of a language-intrinsic control mechanism, which allows for the application of a common (unique) procedure in text difficulty analysis.

Given these overall results, a number of important tasks remain to be tackled by future research:

1. As compared to the history of text difficulty research, much more systematic study is necessary; this concerns both cross-linguistic comparisons and intra-lingual specifics of text types:
   (a) Within a given language, attention must be paid to (the comparability of) different text types; for each of them the specific relation between word and sentence length must be studied.
   (b) As to cross-linguistic studies, the application of Tuldava's formula and its comparison with language-specific formulae seems to be an extremely promising way; in these inter-lingual comparisons too, of course, due attention must be paid to text typology to compare only like to like.
2. As suggested by Tuldava (1993a), the value for either word length or sentence length may be substituted, theoretically, one for the other. A necessary pre-condition for this substitution is, of course, knowledge about the specific relation between word lengthand sentence length (be it for a given language, in general, or for specific text groups, in particular). In this respect, it has not been considered sufficiently thus far that, within

a given language, this relation may differ across text types; therefore, before such substitutions, much more systematic study on the *WoL-SeL* relation along the Arens-Altmann law and its text type specific boundary conditions is necessary.

3. Tuldava's formula and its efficiency remain almost unexplained; it is obvious that the logarithm included leads to a weight reduction of sentence length, but for the time being, there is no explanation in sight why this weight reduction should be logarithmic. It seems reasonable to assume that controlling the relation between word and sentence length will yield relevant insight into this question, the logarithm possibly turning out to be but a good approximation. In any case, it would be desirable either to strive for a theoretical explanation of the logarithmic weight or to replace the logarithm by a parametric model, the parameters of which, in turn, are then open to be interpreted.

References

Arens, H.
1965 *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute.* Düsseldorf: Pädagogischer Verlag Schwann.

Amstad, T.
1978 *Wie verständlich sind unsere Zeitungen?* Diss., University of Zürich.

Bamberger, R.; Vanecek, E.
1984 *Lesen – Verstehen – Lernen – Schreiben.* Wien: Jugend und Volk.

Best, K.-H.
2006 "Sind Wort- und Satzlänge brauchbare Kriterien zur Bestimmung der Lesbarkeit von Texten?" In: Wichter, S.; Busch, A. (eds.), *Wissenstransfer.* Frankfurt/M.: Lang; 21–31.

Brouwer, R.H.M.
1963 "Onderzoek naar de leesmoeilijkheid van Neerlands prosa", in: *Pedagogische Studiën*, 40; 454–464.

DuBay, W.H.
2004 *The Principles of Readability.* Costa Mesa, CA: Impact Information.

Fernández Huerta, J.
1959 "Medidas sencillas de lecturabilidad", in: *Consigna* 214; 29–32.

Flesch, R.
1948 "A New Readability Yardstick", in: *Journal of Applied Psychology*, 32/3; 221–233.

Gray, W.S.; Leary, B.
1935 "What makes a book readable." Chicago: Chicago University Press.

Grzybek, P.; Stadlober, E.
2007 "Do We Have Problems With Arens' Law? A New Look at the Sentence-Word Relation." In: Grzybek, P.; Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann on the Occasion of His 75th Birthday.* Berlin / New York: Mouton de Gruyter; 205–218.

Grzybek, P.; Stadlober, E.; Kelih, E.
2007 "The Relation of Word Length and Sentence Length: The Inter-Textual Perspective." In: Decker, R.; Lenz, H.-J. (eds.), *Advances in Data Analysis.* Berlin etc.: Springer; 611–618.

Grzybek, P.; Kelih, E.; Stadlober, E.
2008 "The relation between word length and sentence length. An intra-systemic perspective in the core data structure", in: *Glottometrics*, 16; 111–121.

Grzybek, P.; Stadlober, E.; Kelih, E.; Antić, G.
2005 "Quantitative Text Typology: The Impact of Word Length." In: Weihs, C.; Gaul, W. (eds.), *Classification. The Ubiquitous Challenge.* Heidelberg, New York: Springer, 53–64.

Kandel, L.; Moles, A.
1958 "Application de l'indice de Flesch à la langue français", in: *Cahiers d'Etudes de Radio-Television*, 19; 253–274.
Kelih, E.; Grzybek, P.; Antić, G.; Stadlober, E.
2006 " Quantitative Text Typology. The Impact of Sentence Length." In: Spiliopoulou, M.; Kruse, R.; Nürnberger, A.; Borgelt, C.; Gaul, W. (eds.), *From Data and Information Analysis to Knowledge Engineering.* Heidelberg, Berlin: Springer, 382-389.
Klare, G.R.
1963 *The measurement of readability.* Ames, Iowa: Iowa State University Press.
1981 "Readability indices: do they inform or misinform?", in: *Information design journal* 2; 251–255.
Köhler, R.; Altmann, G.
1986 "Synergetische Aspekte der Linguistik", in: *Zeitschrift für Sprachwissenschaft*, 5; 253–265.
Lively, B.A.; Pressey, S.L.
1923 "A method for measuring the 'vocabulary burden' of textbooks", in: *Educational administration and supervision* 9; 389–398.
Mikk, J.
2000 *Textbook: Research and Writing.* Frankfurt/M. etc.: Lang.
Partiko, Z.V.
2001 *Zagal'ne redaguvannja. Normativni osnovi.* L'viv: Afiša.
Thorndike, E.L.
1916 "An improved scale for measuring ability in reading", in: *Teachers college record*, 17; 40–67.
1921 *The teacher's word book.* New York: Bureau of Publications, Teachers College, Columbia University.
1932 *A teacher's word book of 20,000 words.* New York: Bureau of Publications, Teachers College, Columbia University.
Thorndike, E.L.; Lorge, I.
1944 *The teacher's word book of 30,000 words.* New York: Bureau of Publications, Teachers College, Columbia University.
Tuldava, J.
1993a "Measuring text difficulty." In: Altmann, G. (ed.), *Glottometrika 14.* Trier: Wissenschaftlicher Verlag wvt; 69–81.
1993b "The statistical structure of a text and its readability." In: Hřebíček, L.; Altmann, G. (eds.),*Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag wvt; 215–227.

Parameter interpretation of the Menzerath law: evidence from Serbian

Emmerich Kelih

1 Introduction

The law-like relation between word and syllable length as part of the Menzerath law has been corrobated empirically in many different languages. As to South Slavic languages, we have the studies by Gajić (1950) and Grzybek (1999) on Croatian, and by Grzybek (2000) on Slovene. The aim of the present paper is first of all to provide empirical evidence of the Menzerath law for another South Slavic language, namely Serbian, distinguishing different text types in our analysis. Second, a linguistic interpretation of the usually iteratively derived parameters of the Menzerath law is offered. Furthermore it will be shown that some parameters of the Menzerath law can be replaced by empirically obtainable quantitative features.

2 Word and syllable length: theoretical background

The Menzerath law is one the most important insights of quantitative linguistics – cf. Altmann (1980), Altmann and Schwibbe (1989), Hřebíček (1990) from recent years. It contains some law-like statements of interrelations between language constituents and their components, such as the relation between the sound duration and the syllable length, between the word and the syllable length, between word and sentence length etc. In this paper special attention is paid to the relation between word and syllable length. According to the Menzerath law, it is expected that with increasing word length (*WoL*), measured by the number of syllables, the mean syllable length (*SyL*), measured in number of graphemes, phonemes or sounds, decreases. Mathematically this can be expressed as $SyL = a \cdot WoL^{-b}$. Usually the parameters a and b are derived iteratively by means of statistical software. The meaning of these parameters is as follows: Parameter a determines the shift on the y-axis and can be understood as the "starting value" of the fitting curve, while parameter b is responsible for the steepness and "speed" of the decrease of the curve. Before a more detailed analysis of the parameters of the Menzerath law can be carried out, the Serbian texts used and the behaviour of word and syllable length in Serbian first have to be discussed.

2.1 The Menzerath law in different text types

A corpus of Serbian texts of different text types and functional styles is used for the analysis of word and syllable length. It consists of ten chapters from diploma dissertations, 32 sermons, seven prose texts by Miloš Ćrnjanski (*Dnevnik o Čarnojeviću*) and 30 journalistic texts.[1] It is noteworthy that it is not the individual texts that are analysed, but rather the sub-corpora of the different text types already mentioned. Additionally a whole corpus was created, which includes all sub-corpora. This structure allows both the analysis of a broad spectrum of different texts types and – in terms of the whole corpus – the influence of text mixtures on the relation of word and syllable length.

The average text length of the sub-corpora used is[2] approximately 4900 word form types. The literary texts are the longest (ca. 5500 types), whereas the sermons consist of only 4365 types. The whole corpus has a text length of 16461 types; see Table 1 for an overview of the texts used.

Table 1: Analysed texts and text length

Text type	Number of texts	Word form types
Scientific texts	10 chapters	4948
Literary prose	7 chapters	5216
Journalistic texts	30	5436
Sermons	32	4365
whole corpus		16461

To obtain the necessary data for the measurement of the word and syllable length these linguistic operations were performed:

1. Serbian has, as proposed in text books and academic grammars (cf. Rehder 2006), 30 graphemes: <а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, у, ф, х, ц, ч, џ, ш>. The texts have been analysed in their orthographical form.
2. The word length (length of word form types) is measured by the number of syllables, and <а, е, и, о, у> are treated as syllabic graphemes. However, to take into consideration the phonetical/phonological level, the <р> – if located between two consonants – is also treated as a syllabic grapheme. For further information on the automatically performed word length analysis cf. Antić et al. (2006).
3. In every sub-corpus and in the whole corpus the word length and mean syllable length – the two sets of data needed for the analysis of the Menzerath law – were determined by the number of graphemes.

1. All texts used are part of the research project on Quantitative Text Analysis (QuanTA) located in Graz; cf. `http://quanta-textdata.uni-graz.at/`.
2. We applied orthographical criteria for the identification of word form types, cf. Kelih 2007.

2.2 Empirical results

Using the basic power model $SyL = a \cdot WoL^{-b}$ it was ascertained that in all analysed sub-corpora the validity of the Menzerath law can be confirmed. The R^2 is in all cases > 0.94. For the scientific texts we even attained an $R^2 = 0.9864$, which can generally be understood as a very well-fitting result. Table 2 gives the empirical data (SyL), the theoretical values (SyL^*), the parameter values for a and b and the R^2 values.[3]

Table 2: Word length – Syllable length in Serbian text types and the whole corpus

	Whole corpus		Scientific texts		Literary prose		Journalistic texts		Sermons	
WoL	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*	*SyL*	SyL^*
1	3.18	3.08	2.96	2.93	3.09	3.01	3.18	3.08	2.96	2.93
2	2.53	2.64	2.54	2.58	2.44	2.54	2.53	2.64	2.54	2.58
3	2.32	2.42	2.38	2.40	2.2	2.30	2.32	2.42	2.38	2.40
4	2.21	2.27	2.24	2.28	2.11	2.15	2.21	2.27	2.24	2.28
5	2.17	2.16	2.19	2.19	2.08	2.03	2.17	2.16	2.19	2.19
6	2.15	2.08	2.14	2.12	2.06	1.94	2.15	2.08	2.14	2.12
7	2.10	2.01	2.10	2.06			2.10	2.01	2.10	2.06
a		3.08		2.93		3.01		3.08		2.93
b		−0.22		−0.18		−0.24		−0.22		−0.18
R^2		0.94		0.99		0.95		0.94		0.99

Figure 1 (p. 74) demonstrates the relation between word and syllable length in the whole corpus. For our Serbian texts the Menzerath law is confirmed and furthermore the mixing of texts (i.e., the whole corpus) clearly has no negative impact on the the fit.

Thus, it can be concluded that the word and syllable length in different text types – as predicated a priori – is regulated by the Menzerath law. In the following section we analyse whether there are significant differences between the coefficients of regression of the different text sub-corpora.

2.3 Significant differences of parameter b?

As can be seen from Table 2, the parameter b clearly depends on the text type: The smallest value is found for literary prose ($b = -0.2430$), whereas for journalistic texts the parameter b has the highest value ($b = -0.1761$). All other

3. All 8, 9 and 10-syllable words have been excluded from our analysis. These words occur extremely rarely, e.g. ten-syllable-words occur twice and nine-syllable words occur only three times in the whole corpus. The mean syllable length of these word lengths shows a slightly abnormal behaviour and they do not fit the commonly obtained tendency. Hence, they are treated here as outliers. It is unclear whether the low frequency or the relatively high word length is responsible for this unusual behaviour.

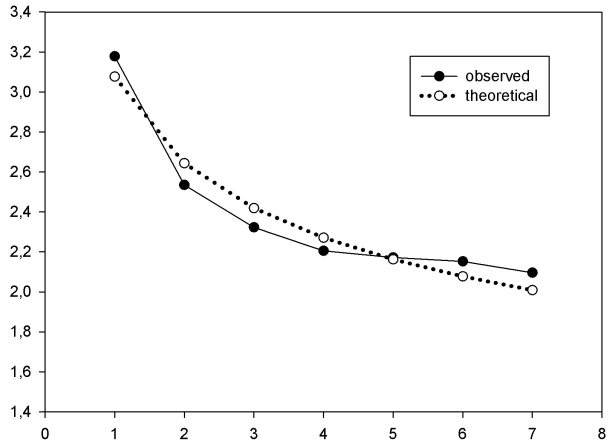


Figure 1: Word length vs. syllable length: the whole corpus

text types and the whole corpus can be located between these two poles. It has to be clarified whether or not these differences are statistically significant.

To do this, the formula $SyL = a \cdot WoL^{-b}$ is transformed by logarithmization to the linear model $log(SyL) = \log(a) + b \cdot \log(WoL)$. The statistical test used is applied in Grzybek et al. (2006) and Zöfel (2002: 146), and hence does not need to be presented again in detail here.

It is not necessary to test all possible differences systematically, but it is sufficient to present the comparison of the lowest b (literary prose) with all other sub-corpora and the whole corpus. Table 3 represents the t-values and p for the performed test and it remains clear that there are no significant differences (in all cases $p > 0.05$) between the compared pairs.

Table 3: Results for the t-distributed test statistics

Pairs of comparison		t-value	DF	p
Literary prose	Journalistic texts	0.23	9	0.8200
	Scientific texts	0.22	10	0.8302
	Sermons	0.90	10	0.3874
	Whole corpus	0.08	10	0.9364

As a result, it can be stated that there are no statistically significant differences in the "steepness" of the fitting curves, and thus a common statistical mechanism seems to organise the relation of word and syllable length in our Serbian texts.

3 Interpretation of parameter *a*

A systematic interpretation of the parameters of the Menzerath law, i.e. the length of a component is a function of the length of the construct, has been proposed by Köhler (1984, 1989). In regard to linguistic systems, it is suggested that human language processing is a sequential process and that language components are processed term by term linearly.

Furthermore, it is assumed that there is some kind of capacity limit in language processing, especially in regard to the length of linguistic components. For the Menzerath law, the parameter *a* represents, as proposed by Köhler (1984, 1989), the mean length of a language construct, consisting of one component. For a detailed re-analysis of the parameters *a* and *b* from various language levels (syllable, word, and sentence length) of the Menzerath law, see Cramer (2005).

If this interpretation holds true, then the parameter *a* approximately equals the mean syllable length (measured here in the number of graphemes) of one-syllable words. Thus, parameter *a* can be replaced by the mean syllable length of one syllable words (henceforth SyL_1). However, such a replacement can be performed only if this leads to no substantial worsening of the fit of the results in general, i.e., the fitting results should not be worse than the others when iteratively determined parameters are used.

Replacing parameter *a* with the mean syllable length of one syllable words indeed does not cause a substantial worsening of the results that fit; see Table 4 for the detailed results and the R^2 calculated on the basis of the replaced parameter *a*.

Table 4: Replacing parameter *a* and new results

Text types	SyL_1	New parameter *b*	New R^2	R^{2*}
Scientific texts	2.9640	-0.1894	0.9830	0.9864
Literary prose	3.0902	-0.2636	0.9456	0.9463
Journalistic texts	2.9595	-0.1919	0.9543	0.9487
Sermons	2.9708	-0.1998	0.9543	0.9554
Whole corpus	3.1784	-0.2413	0.9288	0.9439

* Based on iteratively determined parameters

There is of course a worsening of the R^2, but a satisfying $R^2 > 0.92$ can still be obtained for all text types and the whole corpus. It has thus been shown that a replacement of the iteratively determined parameter *a* by an empirical characteristic (mean syllable length of one-syllable words) causes no substantial worsening of the results that fit. Thus, the interpretation proposed above, that parameter *a* represents the upper limit of a language construct consisting of one component, seems to hold true for the Serbian texts analysed here.

3.1 Dependency of parameter *a* on *b*

As commonly known from synergetic linguistics, there are hardly any isolated language characteristics. This also holds true for parameter *a*, which is in a systematic interrelation with parameter *b*. As already pointed out by Köhler (1984: 181 and 1989: 110), under ideal circumstances these parameters should be in a linear interrelation. According to our interpretation and the replacement of parameter *a*, the relation between SyL_1 and parameter *b* can indeed be captured by a simple linear relation. As can be seen from Figure 2, both characteristics can be modelled by the simple linear equation $b = -0.2869 \cdot SyL_1 + 0.6528$ with an $R^2 = 0.7109$. This is of course not a perfect fit ($p = 0.07$), but at least a common tendency can be obtained, which globally supports the interpretation mentioned above.

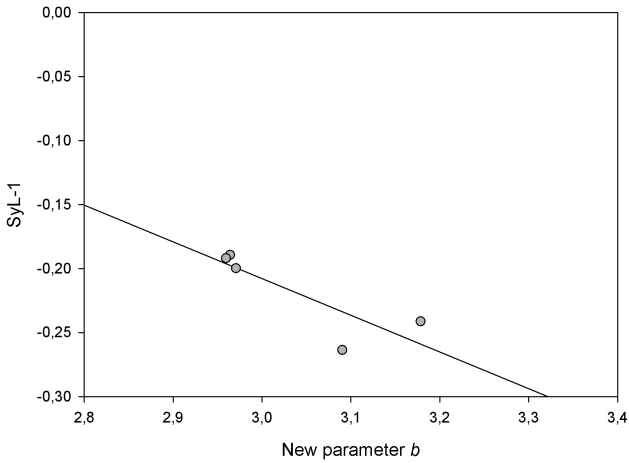


Figure 2: Relation between $SyL - 1$ and parameter *b*

Finally, with this linear interrelation in mind, the original model of the Menzerath law can be "simplified": Replacing parameter *b* with the linear model $b = -0.2869 \cdot SyL_1 + 0.6528$, we arrive at the final equation of $SyL = SyL_1 \cdot WoL^{0.2869 \cdot SyL_1 + 0.6528}$. Therewith, both formerly iteratively determined parameters are replaced by empirical characteristics, namely the mean syllable length of one-syllable words and systematically related characteristics of this value. This replacement is particularly reasonable, because for our analysed texts types it holds true that the longer the one-syllable words, the faster the shortening in longer (i.e. $2, 3, 4, \ldots x$ syllables) words.

This replacement is justified due to the fact that again, despite the replacement of the parameters, no substantial worsening of the fitting results is obtain-

able; see Table 5 for an overview on these results with iteratively determined and replaced parameters.

Table 5: Comparison of results

Text types	R^2 (iterative parameters)	R^2 (replaced a and b)
Scientific texts	0.99	0.98
Literary prose	0.95	0.89
Journalistic prose	0.95	0.94
Sermons	0.96	0.95
Whole corpus	0.94	0.91

Naturally, the replacement of the parameters by empirical characteristics leads to slightly worse fitting results, such as, for instance, for the whole corpus ($R^2 = 0.94 \rightarrow 0.90$) and the literary prose ($R^2 = 0.94 \rightarrow 0.88$). But for the remaining three text types a satisfying $R^2 > 0.94$ is obtainable. However, this result has to be interpreted as a good result, especially because of the fact that ultimately the "meaning" of the parameters used remains quite clear now.

4 Summary

The results of the present paper can be summarised as follows: In Serbian texts the relation of word and syllable length is organised quite systematically according to the Menzerath law. Moreover, it has been shown that the usually iteratively determined parameters can be replaced by empirical characteristics of the word and syllable length, namely by the mean syllable length of one-syllable words. Due to an empirically derived mutual interrelation of the parameters and the mean syllable length, a model with interpreted parameters can be used. Lastly the replacement and reduction of parameters causes no substantial worsening of the fitting results and thus the proposed simplification of the Menzerath law seems to be justified for the texts analysed in this paper.

References

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

Altmann, G., Schwibbe, M.H.
1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Zürich, New York: Hildesheim.

Antić, G.; Kelih, E.; Grzybek, P.
2006 "Zero-syllable Words in Determining Word Length." In: Grzybek, P. (ed.), *Contributions to the Science of Language. Word Length Studies and Related Issues*. Boston: Kluwer, 117–156.

Cramer, I.M.
2005 "The Parameter of the Altmann-Menzerath Law", in: *Journal of Quantitative Linguistics*, 12/1; 41–52.

Gajić, D.M.
1950 *Zur Struktur des serbokroatischen Wortschatzes. Die Typologie der serbokroatischen mehrsilbigen Wörter.* Bonn, Dissertation.

Grzybek, P.
1999 "Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen." In: Tošović, B. (ed.), *Die grammatischen Korrelationen*. Graz: Institut für Slawistik, 67–77.
2000 "Pogostnostna analiza besed iz elektronskego korpusa slovenskih besedil", in: *Slavistična Revija*, 48; 141–157.

Grzybek, P.; Kelih, E.; Stadlober, E.
2006 "Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik", in: *Anzeiger für Slavische Philologie*, 33; 41–74.

Hřebíček, L.
1990 "The Constants of Menzerath-Altmann's law." In: Hammerl, R. (ed.), *Glottometrika 12.* Bochum: Brockmeyer, 61–71.

Kelih, E.
2007 "Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen." In: Kaliuščenko, V.; Köhler, R.; Levickij, V. (eds.), *Problems of Typological and Quantitative Lexicology.* Chernivtsi: Ruta, 91–105.

Köhler, R.
1984 "Zur Interpretation des Menzerathschen Gesetzes." In: Boy, J.; Köhler, R. (eds.), *Glottometrika 6.* Bochum: Brockmeyer, 177–183.
1989 "Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus." In: Altmann, G.; Schwibbe, M.H. (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Zürich, New York: Hildesheim, 108–116.

Rehder, P.
2006 "Das Serbische." In: Rehder, P. (ed.), *Einführung in die slavischen Sprachen.* Darmstadt: Wissenschaftliche Buchgesellschaft, 279–295.

Zöfel, P.
2002 *Statistik verstehen. Ein Begleitbuch zur computergestützten Anwendung.* München: Addison-Wesley.

A syntagmatic approach to automatic text classification. Statistical properties of *F*- and *L*-motifs as text characteristics

Reinhard Köhler, Sven Naumann

1 Introduction

Most approaches to text classification are based on paradigmatic information only, i.e. they apply a "bag-of-words" model of text or they use, as Herdan (1966: 423) put it, "language in the mass" studies to obtain text features for classification (and other) purposes. Classical corpus-linguistics and information retrieval techniques have in common that documents are represented by term weight vectors based on word frequency information. There are a few attempts to use syntactic or statistic phrases instead of words (Caropreso et al. 2001, Fuhr and Buckley 1991, Schütze et al. 1995 and Tzeras et al. 1993), but they have not proved to be superior to simple word-based models. Researchers in the field of quantitative linguistics try to contribute to text classification on the basis of, e.g. word length and sentence length distributions looking for statistical properties of these quantities that could be typical of text genres or text sorts in general. Similar investigations are known from stylometrics and, in recent years, from forensic linguistics (where, in the first place, word frequency distributions play a crucial role).

All these methods ignore syntagmatic information: The organization of the linguistic elements and of the property values of these elements in the course of the text is not evaluated. This is true not only for applications such as text classification but also for quantitative linguistics in general. Only a few attempts to include syntagmatic information have been published yet (Andersen 2005, Hřebíček 2000, Köhler 1999, Köhler 2000, Pawłowski 2001 and Uhlířová 2007). Although term occurrence contributes only a fraction of a text's meaning the above-mentioned methods are rather successful (cf. the usefulness of search engines and other document retrieval applications). Nevertheless, investigations as to how much simple and machine-operable techniques to determine and processing syntagmatic information may improve classification results seem to be worthwhile. Therefore, the present study presents one such simple approach to utilizing "language-in-line" (Herdan 1966: 423) features of texts and first results.

As an appropriate unit of investigation, the *motif* was chosen (originally called *sequences* or *segments*, cf. Köhler and Naumann 2008):

Definition 1
A motif is defined as the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit.

As a refinement of this general definition, we obtain two special ones:

Definition 2
An L-motif is a continuous series of equal or increasing length values (e.g. of morphs, words or sentences).

Definition 3
An F-motif is a continuous series of equal or increasing frequency values (e.g. of morphs, words or syntactic construction types).

An example of a *L*-motif segmentation is the following. The sentence "Word length studies are almost exclusively devoted to the problem of distributions." is, according to the above-given definition, represented by a sequence of five *L*-motifs: *(1-1-2) (1-2-4) (3) (1-1-2) (1-4)*, if the definition is applied to word length measured in the number of syllables. Similarly, motifs can be defined for any linguistic unit (phone, phrase [type], clause [type], etc.) and for any linguistic property (polysemy/polyfunctionality, polytextuality, etc.). Variants of investigations based on motifs can be generated by changing the direction in which these units are segmented, i.e. beginning from the first unit in the text/discourse and proceeding forward or beginning from the last item and applying the definition in the opposite direction and by replacing "increasing" by "decreasing" values of the given property in the definition of the motif.[1] We do not expect statistically significant differences in the results. In contrast, different operationalisations of properties will affect the results in many cases, e.g. if word length is measured in the number of letters or in the average duration in *ms* in speech. Some of the advantageous properties of the new units are the following:

1. Segmentation in motifs is always exhaustive, i.e. no unsegmented input will remain.
2. Motifs have an appropriate granularity; they can always be operationalised in a way that segmentation takes place in the same order of magnitude as the phenomena under analysis.
3. Motifs are scalable with respect to granularity. One and the same definition can be iteratively applied: It is possible to form motifs on the basis of length or frequency values etc. of motifs.
4. Following the definition, any text or discourse can be segmented in an unambiguous way.

1. It may be, e.g. appropriate to go from right to left when a language with syntactic left branching preference is analyzed.

5. Motifs display a rank-frequency distribution of the Zipf-Mandelbrot type (cf. Figure 1), i.e. they behave in this respect in a way similar to other, more intuitive units of linguistic analysis.

2 Method and data

A relatively small text corpus was collected for the present analysis, consisting of 55 documents from five different text sorts (10 poems, 10 narrative texts, 10 juridical, 10 scientific, and 15 journalistic texts). Text lengths varied from 90 to 1500 (poems) and 350 to 8500 (prose) running word forms.

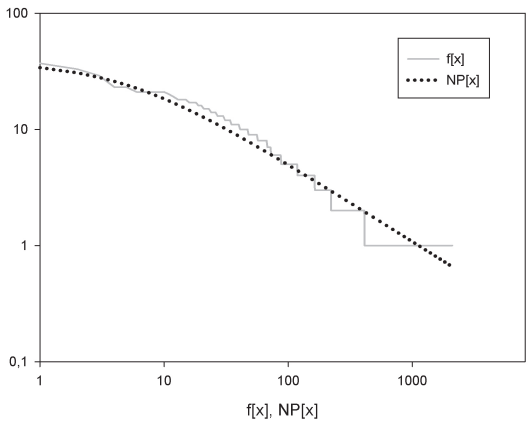


Figure 1: The rank-frequency distribution of L- and F-motifs can be modeled by the Zipf-Mandelbrot distribution. Here, $df = 1607, P[X^2] \approx 1.0, N = 3623, n = 2102$

Each of these documents was analyzed, according to the definitions of L- and F-motifs as given above, first on the word level and then on the level of the motifs themselves. If L-motifs are formed from such a sequence of motifs a second order L-motif (or a LL-motif) is obtained. Thus, in the example *(1-1-2) (1-2-4) (3) (1-1-2) (1-4)*, we have two L-motifs of length 3 followed by one of length 1 etc. The corresponding LL-motif sequence is *(3-3) (1-3) (2)*.

Analogously, LLL-motifs etc. can be formed but also LF-, FL,- FLL-, FLF- etc. motifs are possible, depending on the hypotheses under study. For the purposes of the present paper, only first and second order motifs were determined and evaluated. Then, the rank-frequency distributions of all the motifs were determined for all the documents and fitted by the Zipf-Mandelbrot distribution.

For text classification, the 55 documents in the corpus were represented by vectors of 11 and alternatively 9 attributes (cf. Table 1). Then all pairs of

attributes were scrutinized; finally a Best-First Decision Tree classification algorithm was applied to the complete set of vectors.

Table 1: Attributes formed on the basis of the frequency and probability distributions of the motifs

A1	$hapax/V[L]$	Proportion of hapax legomena of the L-motifs in relation to the inventory size of the motif tokens on the basis of words
A2	$hapax/L[LL]$	Proportion of hapax legomena of the LL-motifs in relation to the inventory size of the motif types on the basis of L-motifs
A3	$hapax/L[LF]$	Proportion of hapax legomena of the LF-motifs in relation to the inventory size of the motif types on the basis of F-motifs
A4	$hapax/L[FF]$	Proportion of hapax legomena of the FF-motifs in relation to the inventory size of the motif types on the basis of L-motifs
A5, A6	Ord I, S [F_glob]	Ord's criteria I and S of the rank-frequency distributions of F-motifs in the individual texts on the basis of word frequencies in the corpus
A7	b [F_{glob}]	Parameter b of the Zipf-Mandelbrot d. of F-motifs (F: corpus)
A8	b [F]	Parameter b of the Zipf-Mandelbrot d. of F-motifs (F: text)
A9	hpx/L [W]	Proportion of the hapax legomena of the word-forms
(A10)	a [F_{glob}]	Parameter a of the Zipf-Mandelbrot d. of F-motifs (F: corpus)
(A11)	a [F]	Parameter a of the Zipf-Mandelbrot d. of F-motifs (F: text)

F-motifs in texts from a corpus can be determined with respect to two different methods of frequency count: The frequency values for words can be determined on the basis of the occurrences of the words in the given text or with respect to the complete corpus. This difference is shown in Table 1 in the following way: "[F]" refers to frequency counts in the individual texts whereas "[F_{glob}]" indicates that the frequency values were taken from the complete corpus.

The last two attributes (the parameter a of the Zipf-Mandelbrot distribution) were not used in the final analysis because, as is well known, the parameters of this distribution are not independent of each other and because a depends also on text length.

3 Results

Pairwise examination of the attributes revealed that some of them are able to separate one of the text sorts from all the others. Figures 2–4 show examples.

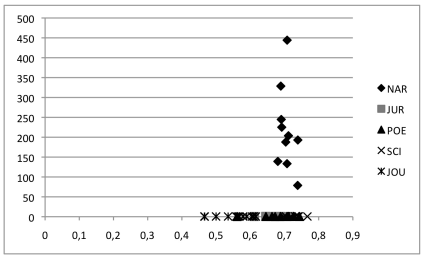


Figure 2: The relative number of hapax legomena of *F*-motifs and parameter *b* of the Zipf-Mandelbrot distribution of motifs on the basis of corpus frequencies separate narrative texts from the others

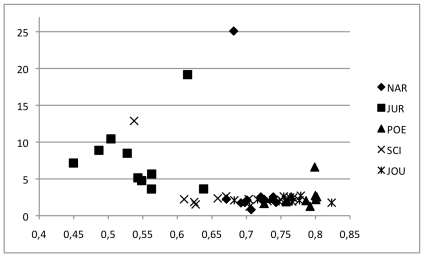


Figure 3: The relative number of hapax legomena of word-forms and parameter *b* of the Zipf-Mandelbrot distribution of *F*-motifs (on the basis of text frequencies) separate juridical texts from the others

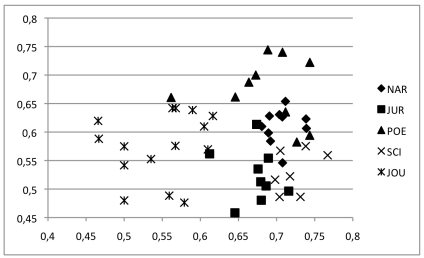


Figure 4: The relative number of hapax legomena of *F*-motifs formed on the basis of *F*-motifs and the relative number of hapax legomena of *L*-motifs on the basis of *L*-motifs separate journalistic texts from the others

These findings suggest that it should be possible to obtain a fairly good classification on the basis of the selected properties. Moreover, it should be possible to reduce the number of features as there are only five categories to classify in. Therefore, the Best-First Decision Tree method (Shi 2007) was applied.

Figure 5 displays the best composition of attributes from Table 1 and the classificatory keys (i.e. the threshold values). The numbers in parentheses indicate how many of the texts were correctly assigned to the given class and how many belong to another class.

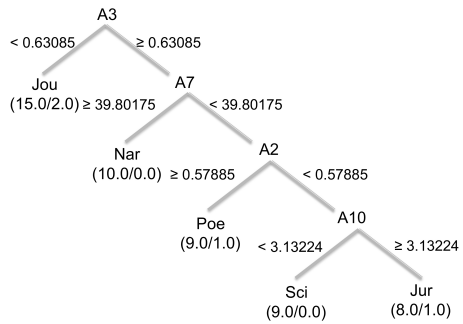


Figure 5: Four attributes separate the text sorts in the corpus

Table 2 shows the confusion matrix.

Table 2: Confusion matrix for the classification

	NAR	JUR	POE	JOU	SCI
NAR	10	0	0	0	0
JUR	0	6	1	1	2
POE	0	0	8	1	1
JOU	0	0	1	14	0
SCI	0	1	1	0	8

Table 3 shows the evaluation matrix for the classification.

4 Conclusion

Our experiments encourage approaches that are based on purely formal and automatically determinable properties of texts:

1. The properties used here are absolutely independent of topics and domains (vocabulary-independent).

Table 3: Evaluation matrix for the classification

TP rate	FP rate	Precision	Recall	*F*-Measure	ROC area	Class
1.000	0	1.000	1.000	1.000	1.000	NAR
0.600	0.022	0.857	0.600	0.706	0.689	JUR
0.800	0.067	0.727	0.800	0.762	0.832	POE
0.983	0.050	0.875	0.933	0.903	0.920	JOU
0.800	0.067	0.727	0.800	0.762	0.853	SCI
0.836	0.042	0.841	0.836	0.834	0.864	Weighted average

T(rue)**P**(ositive), **F**(alse)**P**(ositive), **R**(eceiver)**O**(perating)**C**(haracteristic)

2. Only very few properties are used whereas most other approaches need long vectors of term weights etc.
3. Only four properties suffice to obtain *F*-values around 0.90.

Future studies comprising
- large numbers of texts
- more text sorts
- exploration of other functions of motif properties

are necessary in order to determine whether syntagmatic features like the ones we proposed provide a useful tool for (text) classification tasks and might help to reveal theoretically interesting text properties.

References

Andersen, S.
2005 "Word length balance in texts: Proportion constancy and wordchain-lengths in Prousts longest sentence", in: *Glottometrics*, 11; 32–50.

Caropreso, M.F.; Matwin, S.; Sebastiani, F.
2001 "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization". In: Chin, A.G. (ed.), *Text Databases and Document Management: Theory and Practice.* Hershey, PA: Idea Group Publishing, 78–102.

Fuhr, N.; Buckley, C.
1991 "A Probabilistic Learning Approach for Document Indexing", in: *ACM Transactions on Information Systems*, 9/3; 223–248.

Herdan, G.
1966 *The Advanced Theory of Language as Choice and Chance*. Berlin etc.: Springer.

Hřebíček, L.
2000 *Variation in sequences. Contributions to general text theory.* Prague: Oriental Institute.

Köhler, R.
1999 "Syntactic Structures. Properties and Interrelations", in: *Journal of Quantitative Linguistics*, 6; 46–57.

2000 "A study on the informational content of sequences of syntactic units." In: Kuz'min, L.A. (ed.), *Jazyk, glagol, predloženie. K 70-letiju G.G. Sil'nitskogo.* Smolensk, 51–61.

2006 "The frequency distribution of the lengths of length sequences." In: Genzor J.; Bucková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa.* Bratislava: Slovak Academic Press, 145–152.

Köhler, R.; Naumann, S.
2008 "Quantitative text analysis using *L*-, *F*- and *T*- segments." In: Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R. (eds.), *Data Analysis, Machine Learning and Applications.* Berlin, Heidelberg: Springer, 637–646.

Pawłowski, A.
2001 *Metody kwantytatywne w sekwencyjnej analizie tekstu.* [= Quantitative methods in the sequential analysis of text]. Warszawa: Universytet Warszawski, Katedra lingwistyki formalnej.

Schütze, H.; Hull, D.A.; Pedersen, J.O.
1995 "A comparison of classifiers and document representations for the routing problem", in: *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval (Seattle, WA, 1995)*, 229–237.

Shi, H.
2007 "Best-first decision tree learning." MSc Thesis, Hamilton, NZ.

Tzeras, K.; Hartmann, S.
1993 "Automatic indexing based on Bayesian inference networks", in: *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval (Pittsburgh, PA, 1993)*, 22–34.
Uhlířová, L.
2009 "Word frequency and position in sentence." In: Popescu, I.-I. et al., *Word Frequency Studies.* Berlin, New York: Mouton de Gruyter, 203–230.
Wimmer, G.; Altmann, G.
1999 *Thesaurus of Univariate Discrete Probability Distributions.* Essen: Stamm.

Probabilistic reading of Zipf

Jan Králík

1 Introduction

The Zipf law has become the basic formula of quantitative linguistics. Numerous verifications have been followed by numerous interpretations. These modifications were intended to find a better fit with the measured reality in different types of linguistic situations. Linguists got used to the Zipf formula generally and its influence became so strong and popular that nearly no one asks for more explanation than the one offered by Zipf originally: the principle of least effort (Zipf, 1949).

The essential problem in an attempt to explain or find the mechanism leading to what is being described by Zipf´s formula lies in the conception of rank r in its relation to frequency f:

$$f_r = k/r \tag{1}$$

or in the Mandelbrot (1954) version:

$$f_r = k(r+\sigma)^{-B} \tag{2}$$

In these formulae, algebraically, *rank* plays the role of one of the variables (k and σ are constants). But rank does not represent any natural variable. It does not express any property, any feature, nor any characteristics that could acquire values and which, therefore, could be measurable.

On the other hand, rank is not an independent phenomenon. Rank appears as a result of different frequencies, and different frequencies are caused by different possibilities of words to occur (or "to be used"). Frequencies of words vary according to the utility of words, and rank is the function of resulting frequencies. Therefore, the conception of rank involves also an important statistical and probabilistic aspect.

2 Utility

The existence of the frequency of any linguistic element is caused by many factors, which could be projected into the commonly understandable conception of utility or usefulness. Thus, all the grammatical, topical, stylistical, terminological, educational, mental, plus many other, aspects can be involved.

Experimental observation of such *utility* or *usefulness* differs in various situations and times. It even changes within one author's texts. As has been already shown elsewhere, such a conception of *utility* is that very essential property, for which (in terms of the axioms of The Probability Theory by Kolmogorov) the simple additive measure can be defined over all the sets created upon the field of elementary linguistic events. Such an additive measure then, according to the axioms, is called *probability* (Kolmogorov 1933). When the Law of Large Numbers is applied, this axiomatical *probability* can be commonly identified with the limit value of relative frequencies, computed in a long row of observations made under comparable circumstances.

This explanation is not a mere play with words. The term *probability* refers both to the property (*utility*, *usefulness*, *usage*) and to some number (from the interval $< 0;1 >$). The number, however, is one of many possible realizations of the influence of the property. Not only in linguistic research, *probabilities* are often represented by relative frequencies, and the original property of *usefulness* or of *chance* to be used or of utility is forgotten. Such simplification, substituting the measurable random variable directly by its values, is comfortable and practical, but it nearly excludes any further considerations.

In contrast, when the random variable called *utility* is discussed, we can go on thinking about its distribution and density function and we can use it with further models. For example, let us express the random variable *utility* by θ. Then, its distribution function

$$F(x) = P(\theta < x) \tag{3}$$

expresses the probability that the *utility* θ does not exceed the value x. Both the *utility* θ and x are from the interval <0;1>. The supplementary function

$$N(x) = 1 - F(x) \tag{4}$$

describes the probability that θ reaches or exceeds the value x. This $N(x)$ has its real representation in every frequency list of words in the ratio

$$N(x) \approx \frac{\text{(Number of words with the utility equal to or greater than x)}}{\text{(Vocabulary extent of the text)}} \tag{5}$$

where instead of *word*, the terms *element* or *item* can be used for immediate generalization (Králík 1997). If the sense of this ratio is considered for the beginning of the frequency list, it can be seen that: For any *utility* level x, the nominator (number of words with the measure of *utility* equal to or greater than x) equals to the *rank* corresponding to level x.

If the sense of ratio (5) in the middle and at the end of the frequency list is considered, it can be seen that: Following the reverse direction of the frequency word list, there are numerous sets (long "intervals") of words, in which

the measure of *utility* is equal to *1*, *2*, *3*, etc. In such sets the non-statistical alphabetic ordering is used, so that the *rank* becomes random. In fact even within such sets the values of *utility* of words differ. In the case of a single frequency list, many different values are projected into the same integer. If many frequency lists were confronted, or great corpora analysed, a real chance appears to distinguish between any two words (elements, items), as to their measure of *utility*. Subsequently, it would be possible to establish their *rank*. So, again, it could be seen that: For any *utility* level x, the *number of words* with the measure of *utility* equal or greater than x, equals the *rank* corresponding to level x. Thus, for the whole vocabulary, we could generally write:

$$N(x) = 1 - F(x) = r_x/V \tag{6}$$

where r_x symbolises the rank of the word with measure of *utility* equal to x and V symbolises the number of all different words (vocabulary). Analogically, the real representation of x can be found in every frequency list or frequency dictionary in the form of the ratio

$$x \approx \frac{\text{(Number of occurrences of the word with the measure of utility equal to x)}}{\text{(Number of all current words)}}$$

The numerator of this ratio is usually symbolised by f_x (absolute frequency), the denominator is usually symbolised by N (the current text extent):

$$x = f_x/N \tag{7}$$

Here again, as is the rule, x is generally different for different words at the beginning of the frequency list, and, at the end of the frequency list, x is usually expressed by some common value for more or even many words. Again it is clear that, as a general view which would deal with many frequency lists, it would be possible to distinguish each corresponding value of *utility* for each word as precisely as requested. Using the introduced points of view of representations of $N(x)$ and x, we could write

$$\begin{aligned} r_x/V &= 1 - F(f_x/N) \\ F(f_x/N) &= 1 - r_x/V \end{aligned} \tag{8}$$

These equations show how the conception of *rank* is connected with the corresponding measure of *utility*, or, in other words, with the *probability* or *relative frequency*. This natural connection is existential. The property of utility (the ability to be used, measured by probability) causes and influences the phenomenon of *frequency*, and frequencies form *rank*.

3 Density

The above written formulae open the way to further considerations. The first logical question concerns the explicit form of the unknown distribution func-

tion of utility $F(x)$ as described in (3). It opens endless space for sophisticated suggestions, hypotheses, estimations or subjective trials.

Let us discuss the density function $F'(x)$, the derivative of the distribution function $F(x)$ for vocabulary, so that we can use the common knowledge about word frequencies. The density function $F'(x)$ of the variable x called *utility*, should give information about the distribution of *utility*. It should express the common experience, that in normal text, for every level of *utility*, more repeated words do form less numerous sets than more rare words do. Even when only this principle of reverse proportionality is accepted, an interesting model can be constructed. It is based on the idea, that according to the aforementioned principle, some *hierarchy* of *utility* exists.

Let us consider re-ordering the vocabulary (set of all different words, lexicon) based on dividing it into categories according to *utility* in the following way: the category assigned to index i involves some number y_i of elements, being represented by one common value of utility x_i. Categories can be constructed in a way so that

$$x_i = \alpha \cdot x_{i-1} \tag{9}$$

for $i = 1,2,3,\ldots,k$ and with "coefficient of hierarchization" $\alpha > 1$. Each next category is characterised by *utility* on the level that is α-times higher (stronger). This implies that also some dependence between the numbers of elements inside the categories should exist. The discrete construction allows the generally linear expression:

$$y_i = \tau_i \cdot y_{i-1}$$

The previously mentioned empirical experience from frequency dictionaries (more repeated words do form less numerous sets than more rare words do) indicates $\tau_i < 1$. Thus, at the same time, we could write

$$x_i = \alpha \cdot x_{i-1} = \alpha^2 \cdot x_{i-2} = \ldots = \alpha^{i-1} \cdot x_1$$
$$y_i = \tau_i \cdot y_{i-1} = \ldots \left(\prod_{r=2}^{i} \tau_r \right) \cdot y_1$$

A further step consists in fixing the second parameter τ.

Presumption. Let us discuss the simplest case in which the decreasing proportionality values $\tau_r (r = 2,3,\ldots,i)$ do not differ too much, so that they could be approximated by one common value $\mu < 1$ so that the following simplification could hold:

$$\prod_{r=2}^{i} \tau_r = \mu^{i-1}$$

Then, the number of words in the category i could be counted as

$$y_i = \mu^{i-1} \cdot y_1 \tag{10}$$

where y_1 is the number of words in the first category after hierarchization has been done, which equals the number of words with the lowest *utility* expressed by the chosen level x_1; let us remark that usually $x_1 = 1$ and $y_1 =$ [number of hapax legomena]). This presumption tells us in other words, that the *utility* of a word from category i is proportional to the sum of utilities of all words in the previous category $i-1$. The proportionality is given by values α (coefficient of hierarchization) and μ (number-of-words reduction).

An important algebraic simplification can be done in following way. Let us re-write the exponent $(i-1)$ from $x_i = \alpha^{(i-1)} \cdot x_1$ by means of logarithms $i - 1 = (\log \alpha)^{-1}(\log x_i - \log x_1)$ and let us use this in the expression of y_i

$$y_i = y_1 \cdot \mu^{(\log x_i - \log x_1)/\log \alpha}$$

Other algebraic steps can be performed, as follows:

$$\begin{aligned} y_i =& y_1 \cdot (exp \log \mu)^{(\log x_i - \lg x_1)/\log \alpha} \\ =& y_1 \cdot exp\{\log \mu \cdot [(\log x_i - \lg x_1)/\log \alpha]\} \\ =& y_1 \cdot exp\{[(\log x_i \lg \mu)/\log \alpha] + [(\log x_1 \cdot \log \mu)/(-\log \alpha)]\} \\ =& y_1 \cdot exp\{(\log x_i \cdot \log \mu)/(-\log \alpha)\} \cdot (exp \log x_i)^{\log \mu/\log \alpha}. \end{aligned}$$

To reach the readable explicit form of dependence between y_i and x_i:

Because we know that α and μ, as well as y_1 and x_1 are supposed to be constants, concluding simplification can be written

$$y_i = B \cdot x_i^{\beta} \tag{11}$$

where $B = y_1 \cdot x_1^{-\beta}$ and $\beta = (\lg \mu / \lg \alpha) < 0$, still knowing that $\alpha > 1$ is the coefficient of hierarchization and $\mu < 1$ is the decreasing proportion of number of words in the neighbouring *utility* categories.

By means of words we could summarize that within the suggested hierarchy, the number of words in every category of utility can be expressed by means of the exponential (power) function of *utility*. We already mentioned that the characterization of words by utility can be done as accurately as requested. This enables us to presume that the discrete formula (11) can be satisfactorily replaced by the continuous equation

$$y = B \cdot x^{\beta}$$

which offers one of the possible analogies with the density function y^* for the distribution of words (lexical units) according to their utility. For such an idea it would hold

$$y^*(x) = 0 \qquad \text{for} \quad x \leq 0 \tag{12}$$

$$y^*(x) = B^* \cdot x^{\beta} \quad \text{for} \quad x > 0 \tag{13}$$

where B^* involves the necessary norming and $\beta < 0$ (we could write even more didactically $\beta = -b; b > 0$). The most important issue of all these steps is that the corresponding distribution function can be expressed in the form

$$F(x) = B^* \int_{-\infty}^{x} t^{\beta} dt = B^* \int_{0}^{x} t^{\beta} dt = [B^*/(\beta+1)] \cdot x^{\beta+1} = E \cdot x^{\varepsilon} \tag{14}$$

and that this integral, as the function, again possesses the form of a power function:

$$F(x) = [B^*/(\beta+1)] \cdot x^{\beta+1} = E \cdot x^{\varepsilon}$$

Let us remind you that all the parameters and variables can be explained from the hierarchy of the increasing levels of utility and from decreasing number of words on such levels (Králík 1983).

4 Conclusion

From the first part of our considerations we already know this possible interpretation of the supplementary function

$$N(x) = 1 - F(x) = r_x/V \tag{15}$$

We also know that the value of utility can be approximated by relative frequency (6) (f/N), $f =$ absolute number of occurrences within $N =$ current text length) and we supposed the most simple form of distribution function (14), so that following final steps can be done:

$$\begin{aligned} r_x/V &= 1 - E \cdot (f_r/N)^{\varepsilon} \\ E \cdot (f_r/N)^{\varepsilon} &= 1 - r_x/V \\ (f_r/N)^{\varepsilon} &= -1/E \cdot (r_x/V - 1) \\ f_r &= N(-1/E)^{1/\varepsilon} \cdot \left\{ (r_x - V)^{1/\varepsilon} / V^{1/\varepsilon} \right\} \\ &= N(-1/E \cdot V)^{1/\varepsilon} \cdot (r_x - V)^{1/\varepsilon} \, . \end{aligned} \tag{16}$$

Within finite texts for comparable N (text extents) and V (vocabulary extents) the above written last equation can be formally understood as:

$$f = k(r + \sigma)^{-B} \tag{17}$$

which equals the well known form of Mandelbrot's (1954) correction of original formulae by Estoup (1916) and Zipf (1949). It can be objectively stated that this is no new knowledge. It is not new from the point of view of the necessity of verification. Hundreds, maybe thousands of confirmations have been

published. What is new, however, is the insight based on the above performed deductions: the construction of each parameter preserves real meaning, so that parameters can be analysed, measured and interpreted.

But not only this direction can be followed for concluding considerations. The starting point of this construction was a natural idea about the type of hierarchy in the distribution of *utility*. All the known verifications of the result indicate that our nearly random presumption about hierarchy (the higher the utility is, the lower the number of its words) given by some α, and about the decreasing proportionality of numbers of elements in the inducted categories by an equal μ, must have been very near to the reality.

Economists will see similarities with the distribution of financial incomes as has been described by Pareto. The Pareto parallel, among others, leads to a special type of Pearson curves, which corresponds with the Pólya scheme for a special urn model with a changeable measure of return. To this, however, a linguistic sense can be given, by what we know about the human brain and about the author's text creation. Independently, the suggested hierarchy is – maybe surprisingly – very near to the idea of the neuron organization of the human brain, its memory and functioning.

A very similar type of hierarchy is well known from quantifications of the Menzerath-Altmann Law (the longer the construct is, the shorter are its constituents). And last, but not least, the suggested conception of utility hierarchy is not in contradiction with the Principle of Least Effort. The *utility hierarchy* offers an explanation on which bases and why the Principle of Least Effort works.

References

Estoup, J.B.
1916 *Gammes sténographiques.* Paris: Institut sténographique.
Kolmogorov, A.N.
1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer.
Králík, J.
1983 "Some Notes on the Frequency", in: *Prague Studies in Mathematical Linguistics*, 8; 67–80.
1997 "On the Probability of Probabilities." In: *Qualico 1997. Third International Conference on Quantitative Linguistics.* Helsinki: Monila, 77–82.
Mandelbrot, B.
1954 "Structure formelle des texts et communication", in: *Word*, 10; 1–27.
Zipf, G.K.
1949 *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* New York: Hafner.

Revisiting Tertullian's authorship of the *Passio Perpetuae* through quantitative analysis

Jerónimo Leal, Giulio Maspero

1 Introduction

The *Passio Perpetuae* is one of the oldest *Acta Martyrum*, written in Africa at the begining of the III century. It is an amalgam of texts surely not written by the same author: Introduction (321 words), the Diary of the martyrium (that is not interesting for our research) and the Conclusion (1121 words). We must suppose one or two compilators for the introduction and the conclusion. These two parts, by style analysis, have been attributed to Tertullian.[1]

This hypothesis is traditionally related to d'Alès (1907: 7); but also Ruinart, Robinson, Zahn, Harnack, Franchi de' Cavalieri, Bonwetsch and Krüger (Leal 2009: 62ff.) are among its proponents. De Labriolle (1913: 126ff.) is the only one against the authorship of Tertullian. More recently authorship has been attributed to Pomponius (Braun 1979: 117), a demi-lettré (Fontaine 1968: 73), a bishop or presbyter (d'Alès 1907: 7), somebody of the entourage of Tertullian (Lanata 1973: 160), or two different compilators (Amat 1996: 67).

The aim of this contribution is to test the validity of a quantitative method to analyze the works of Tertullian and apply this method to the question of Tertullian's authorship of the *Passio Perpetuae*. In particular, our aim is to test numerically the hypothesis that the first and the final part of the work were written by Tertullian and to check the attribution of all the works transmitted under the name of this author, together with those considered spurious by critical philological studies.

2 Our approach

Our idea is to test the method for quantitative authorship attribution on a real problem. From our perspective the issue is not to find a general method for authorship attribution, but to decide whether the very one text we are working on is or is not of the same author. The present work is based on the paper by Basile and Lana on Gramsci's articles (Basile and Lana 2008). We try to apply their method to our case, modifying the procedure according to the results that

1. We have 31 conserved writings of Tertullian, an African Christian writer who lived between ca. 150 and ca. 220, of different styles and word number.

we find. Because of that, our approach in purely quantitative. According to Basile and Lana's procedure, we have recourse to two different text-distances and we try to study the Tertullian works with them. The first metric is based on the *n*-grams and is a modified version of the pseudo distance described by Kešelj's et al. (2003):

$$D_n(A,B) = \frac{1}{|S_n(a)| + |S_n(b)|} \sum_{\omega \in S_n(a) \cup S_n(b)} \frac{(\phi_n(a) - \phi_n(b))^2}{(\phi_n(a) + \phi_n(b))^2}, \tag{1}$$

where A and B are two authors, a and b are two texts composed by the authors, $S_n(a)$ is the set of n-grams ω in a and $S_n(b)$ the corresponding set of ω in b, $\phi_n(a)$ is the normalized density of the n-grams in a, and $|S_n(a)|$ the dimension of its set. We get the best results from the bigram distribution D_2. This is different with respect to Basile et al. (2008: 177), where the best results were found for D_8.

The second metric is the entropic one. It is defined as:

$$\Delta(A,B) = \frac{\delta(a,b) - \delta(b',b)}{\delta(b',b)} + \frac{\delta(b,a) - \delta(a',a)}{\delta(a',a)}, \tag{2}$$

where $\delta(a,b) = (\ell_c(a+b) - \ell_c(a))/|b|$, i.e. the difference between the compressed length of the concatenation of a and b and the compressed length of a, divided by the number of bits of b. If $b \ll a$, then $\delta(a,b)$ is an approximation of the relative entropy of the two texts. The validity of the approximation depends on a transition length that we have taken into account, dividing the second file into small fractions, each of size ℓ. Because of that, in order to compute $\Delta(A,B)$, we have divided the file from source A, using the first half as a and the second one as a'. We have done the same with the file from source B, but we have also divided it in units of a size $\ell \ll |a| + |b|$, averaging the results. In this way we can compare texts of any size. In our computation we have used $\ell = 450$.

In our computations, we have used the zlib Python library and we have preprocessed out texts, stripping from them the punctuation, transforming them to small letters and changing all the *v*'s to *u*'s, as is usual in ancient Latin.

3 Sets of data

We have divided the 34 remaining works of Tertullian into two sets of 17, each chosen at random: the first one to be used as a sample set (TT) and the second one to be used as a test set (TP). The total number of works that are known for sure under the authorship of Tertullian are the following (with the corresponding length in bytes):

Ad Martyres (10149, TP-1); Ad Nationes (53973, TP-2); Ad Scapulam (10072, TT-3); Adversus Hermogenem (70996, TT-4); Adversus Iudaeos (75587, TT-5); Adversus Marcionem i (68404, TP-6i); Adversus Marcionem ii (69954, TP-6ii); Adversus Marcionem iii (71561, TP-6iii); Adversus Marcionem iv (222855, TT-6iv); Adversus Marcionem v (121898, TP-7v); Adversus Praxean (89188, TT-7); Adversus Valentinianos (44884, TT-8); Apologeticum (135501, TT-9); De Anima (162910, TP-10); De Baptismo (30166, TP-11); De Carne Christi (62019, TT-12); De Corona Militis (33758, TT-13); De Cultu Feminarum (12705, TP-14); De Exhortatione Castitatis (26264, TT-15); De Fuga in Persecutione (35396, TP-16); De Idololatria (47373, TT-17); De Ieiunio Adversus Psychicos (41084, TT-18); De Monogamia (46505, TP-19); De Oratione (31201, TP-20); De Pallio (23718, TP-21); De Paenitentia (29245, TT-22); De Patientia (32149, TT23); De Praescriptione Haereticorum (58107, TT-24); De Pudicitia (92827, TP-25); De Resurrectione Carnis (153484, TP-26); De Spectaculis (43941, TT-27); De Testimonio Animae (15100, TP-28); De Virginibus Velandis (37794, TP-29); Scorpiace (53672, TT-30).

We have divided *Adversus Marcionem* into five different files, each one corresponding to a book, in order to have a bigger and more homogeneous set of sample works. Our aim is to verify the attribution of the spurious works (TS), which are the following:

Adversus Omnes Haereses (17809, TS-1); Carmen Ad Senatorem (3733, TS-2); Carmen De Iona Propheta (4634, TS-4); Carmen De Iudicio Domini (17180, TS-5); Carmen Genesis (7040, TS-6); De Execrandis Gentium Diis (4667, TS-7); Passio Perpetuae et Felicitatis (22988, TS-8).

The last one is the *Passio Perpetuae*, that, as already said, is composed of three parts: the beginning and the final are suspected to really be by Tertullian. We are interested in verifying this hypothesis. For this reason, we had divided the *Passio Perpetuae et Felicitatis* into three parts, as follows: incipit (2270, TS-9); explicit (7881, TS-10); body (12874, TS-11).

To compare the spurious works with the authentic ones, we have to check the effectiveness of our measures in distinguishing non Tertullian works. Hence we had taken recourse to two sets of works: the first one (NT) includes 17 works, all composed in a time span of one century with respect to our author:

Caesar, De Bello Gallico (144111, NT-1); Cicero, Adversus Catilinam (21915, NT-2); Cicero, Oratio i (22075, NT-3); Cicero, Oratio ii (20765, NT-4); Cicero, Oratio iii (21633, NT-5); Cicero, Oratio iv (19787, NT-6); Alexander Aphrodisiensis, De Intellectu (20926, NT-7); Arnobius Afrus, Disputationum Adversus Gentes (464871, NT-8); Commodianus, Carmen de Duobus Populis (43813, NT-9); Cyprianus Carthaginensis, De Catholicae Ecclesiae Unitate (36489, NT-10); Ovidious, Fasti (31572, NT-11); Quintillianus, Institutiones Oratoriae

(83093, NT-12); Suetonius, De Poetis (26098, NT-13); Suetonius, Vitae Caesarum (508453, NT-14); Tacitus, Annales (648460, NT-15); Tacitus, Historiae (379321, NT-16); Virgilius, Aeneis (186300, NT-17).

The second one (*NN*) is a more heterogeneous sample, with authors of different centuries and with more works devoted to religious subjects, including a *Passio* (NN-15):

Virgilius, Bucoliche (100411, NN-1); Horatius, Ars Poetica (22088, NN-2); Cato, De Agricultura (100569, NN-3); Commodianus, Intructiones (50079, NN-4); Lactantius, Ad Donatum (77149, NN-5); Ambrosius, Ep XX (14380, NN-6); Frontinus, De arte mensoria (4216, NN-7); Ambrosius, Ep XVII (9230, NN-8); Ambrosius, Ep XVIII (20359, NN-9); Apuleius, De mundo (45607, NN-10); Boethius, Tract Theologici 2 (3550, NN-11); Hieronymus, Vita Malchi (14094, NN-12); Isidorus, Sententiarum liber I (81982, NN-13); Martialis, Liber spectaculorum (9352, NN-14); Anonimous, Passio Scillitanorum (2635, NN-15); Phaedrus, Phabulae (14915, NN-16); Salvianus, Ep. I (4373, NN-17).

We expect that the second set (NN) of non Tertullian works will be harder to distinguish from the original works than the first set (NT).

4 Results

We present in Figure 1 our results, obtained using TT as a reference sample of Tertullian works.

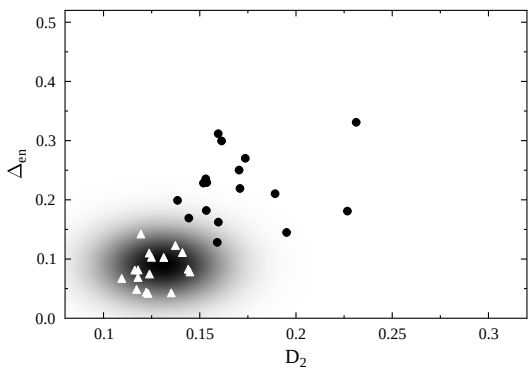


Figure 1: On the *y* axis the entropic distance Δ_{en} and on the *x* axis the bigram distance D_2: there is a clear divinsion into two different groups, with circular points designating the non Tertullian works and the triangles the Tertullian ones

We try to distinguish the test corpus (TP) from the easy non Tertullian works (NT). We plot the average distance between the analyzed work and the 17 sample Tertullian works. The density plot represents the bidimensional Gaussian centered on the average distance between the different works in the reference sample that is composed of $n = 17$ works. Because of that we have $n(n-1)/2$ distances, the one for $n = 17$ give a total set of 136 internal distances, which are distributed according to a Gaussian area.

We observe that the results show a clear distinction between two sets, which correspond to the Tertullian and the non Tertullian works. If we try with the second more difficult set of non Tertullian works (NN), we see that the two groups are still well separated, but we have a false positive, as NN-12 lies in the area of the Gaussian area (cf. Figure 2). It is to be noted that the method does not give a false negative because the work is by Jerome. If we apply the present

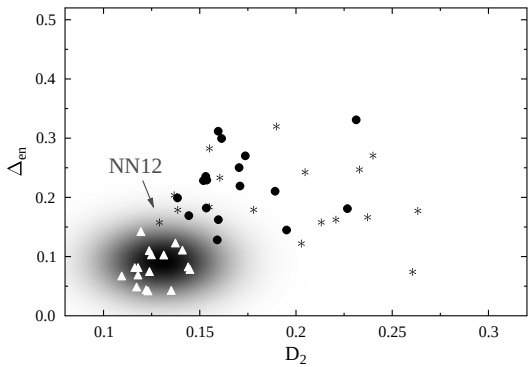


Figure 2: On the *y* axis the entropic distance Δ_{en} and on the *x* axis the bigram distance D_2: in the case of *NN* (asterisks), we get a false positive

method to the spurious works, including the *Passio Perpetuae* considered as a whole (TS8) and divided into the incipit (TS9), explicit (TS10) and central part (TS11), we get the results in Figure 3.

It seems that TS7, i.e. *De Execrandis Gentium Diis*, a work that most recent studies (Turcan and Turcan-Verkerk 2000: 205–271) are trying to attibute to Tertullian as excerpts of a lost treatise, could be a Tertullian work. The other works are clearly non Tertullian. As regards *Perpetua* (TS-8), it lies near Tertullian. It is remarkable that the three parts of the work have different behaviour: the incipit (TS-9) has a very low entropy distance and is a good candidate to be accepted, while TS11, i.e. the central part, lies at a greater distance from the Tertullian area, finally TS-10, i.e. the end of the work, is excluded by the first metric.

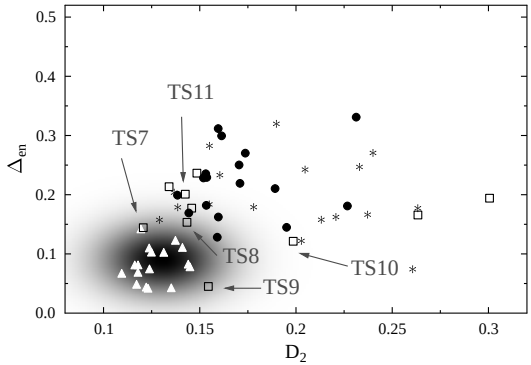


Figure 3: On the *y* axis the entropic distance Δ_{en} and on the *x* axis the bigram distance D_2: in the case of TS (square points), TS7 lies in the Tertullian area, TS9 and TS8 very near, TS10 and TS11 fall outside

The present results do not seem satisfactory, as it is difficult to say anything sure about the authorship of the work under study. One way to improve the method consists of introducing voting for the attribution: this means to compute an index of *Tertullianity* that could offer a way of discriminating the authorship of the studied texts.

In Basile et al. (2008: 176f.) an index was defined, based on the position of the test text in the list of works attributed to the author and of the works in another list of different authors. The problem with this kind of index is that it depends on the choice of the different authors. This being the case, we have defined an index n_T as the percentage of distances between the analysed work and the Tertullian sample works inferior to the average internal distance plus the standard deviation:

$$n_T = \frac{1}{N} \sum_m f(D(m)) \tag{3}$$

where $f(D) = 1$ if $D \leq (\bar{D} + \sigma)$ and zero otherwise. In our case, the sum includes 17 distances.

The index can offer some kind of estimate of the precision of the data. We expect that the 17 distances between the analysed work and the Tertullian sample works are distributed according to the Gaussian distribution in Figures 2 and 3. That means that $n_T \geq 0.85$ should be the normal result for the probability of one true Tertullian value to have a distance less than or equal to the mean, plus one standard deviation. Because of that, we reject the works with $n_T \leq 0.75$, we consider those values between 0.75 and 0.85 to indicate 'almost by Tertullian', but doubtful, and we accept them for $n_T \geq 0.85$. The results are

the following (results which deviate from the average behavior are printed in bold face):

TP-1: **0.53**/0.94	TP-10: 0.88/1.00	TP-20: 0.94/1.00
TP-2: 0.94/0.94	TP-11: 0.88/0.94	TP-21: 0.76/**0.65**
TP-6i: 0.94/0.94	TP-11: 0.88/0.94	TP-25: 0.94/0.76
TP-6ii: 0.94/1.00	TP-14: **0.71**/0.94	TP-26: 0.94/**0.41**
TP-6iii: 1.00/1.00	TP-16: 0.94/1.00	TP-28: **0.71**/0.82
TP-6v: 0.88/0.94	TP-19: 0.94/1.00	TP-29: 0.94/1.00

Both methods give false negatives and their precision is similar in the case of the Tertullian works. We get two false negatives for the entropic distance and three of them for the bigram distance. The result is different for the non Tertullian works:

NT-1: 0.23/0.00	NT-7: 0.00/0.00	NT-13: 0.76/0.18
NT-2: 0.23/0.00	NT-8: 0.06/0.47	NT-14: 0.00/0.18
NT-3: 0.41/0.00	NT-9: 0.23/0.59	NT-15: 0.06/0.00
NT-4: 0.35/0.12	NT-10: 0.23/0.23	NT-16: 0.06/0.00
NT-5: 0.35/0.00	NT-11: 0.06/0.06	NT-17: 0.29/0.00
NT-6: 0.41/0.00	NT-12: **0.88**/0.00	

On the set of easier non Tertullian works (NT), the entropic distance gives better results, without any false positives. The bigram distance gives one false positive. It seems that the entropic method is more precise:

NN-1: 0.35/0.00	NN-7: 0.00/0.00	NN-13: 0.00/**1.00**
NN-2: **0.82**/0.06	NN-8: 0.00/0.12	NN-14: 0.29/0.00
NN-3: 0.06/0.00	NN-9: 0.00/0.65	NN-15: 0.00/0.06
NN-4: 0.00/0.12	NN-10: 0.71/0.12	NN-16: 0.12/0.12
NN-5: 0.00/0.29	NN-11: 0.00/0.00	NN-17: 0.00/0.29
NN-6: 0.41/0.06	NN-12: **0.94**/0.35	

This impression is confirmed by the test on the more difficult set of non Tertullian works (NN), where we get one false positive (N-13, i.e. Isidorus' Sententiarum liber I). The bigram distance gives more false positives. The overall result is that the bigram method fails to detect 3/17 true Tertullian works, i.e. 18% of the sample, while the entropic method fails in 2/17 cases, i.e. 12%. In the case of non Tertullian works, the bigram method fails three times over 34 works (9%), while the entropic method fails only in one case (3%). We get similar results interchanging TT and TP, i.e. using the second set of Tertullian works as sample texts and trying to guess the Tertullianity of TT.

If we apply the two methods to the spurious works, we get the following results:

TS-1: 0.65/0.00	TS-6: **0.71**/0.00	TS-9: 0.47/**1.00**
TS-2: 0.00/0.29	TS-7: **0.88**/0.35	TS-10: 0.00/**0.71**
TS-4: 0.00/0.12	TS-8: 0.59/0.29;	TS-11: **0.71**/0.00
TS-5: 0.53/0.00		

The result is different according to the two methods. The bigram distance recognizes as Tertullian TS-7. Moreover it gives results near Tertullian for TS-6 and TS-11, i.e. for *Carmen Genesis* and for the central part of the *Passio Perpetuae*, that is known for sure to be non Tertullian. In contrast, the entropic distance rejects all the works, except the incipit of the *Passio Perpetuae*. It is interesting to point out, that the explicit is near the limits of the Tertullian authorship, explaining perhaps its attribution to Tertullian or pointing out that Tertullian was only the compiler and that he used a previous traditional text. But according to our data, we have to conclude that the author of the incipit is different from the author of the final part, even if that latter is near the style of Tertullian.

We trust more the entropic method not only because it is more precise. A test can be done on different *Acta Martyrum*; these should be the more difficult texts to be distinguished, as there is a great proximity with *Passio Perpetuae* at the level of vocabulary. We get the following results:

Acta S. Cipriani 1 (4370):	0.00/0.00
Acta S. Cipriani 2 (3691):	0.00/0.00
Passio S. Crispinae (4826):	0.00/ 0.06
Passio S. Marcelli Tingitani (2526):	0.00/ 0.00
Passio S. Mariani et Iacobi (18054):	**0.82**/0.06
Passio S. Maximiliani (3797):	0.00/0.06
Passio S. Lucii, Montani et aliorum (22611):	**0.94**/0.53

These data suggest that the bigram distance recognizes the two Passions as Tertullian. On the contrary, the entropic method gives much better results, without false positives.

5 Conclusion

We have come to two sets of conclusions:

1. on the philological problem, our result, in the first place, demonstrates that the incipit and explicit belong to two differents authors; and, in the second place, it supports the hypothesis of Tertullian's authorship of the incipit of the *Passio Perpetuae*; as a complement of the more recent studies, the *De Execrandis Gentium Diis* is to be retained very near the lost Tertullian if not attributed to him;

2. at the computational level, we have shown that the entropic distance is more effective than the bigram distance in attributing these kinds of texts and, on the other hand, we have suggested two very easy implementations of the computation of the entropic distance and an index for voting.

References

Amat, J.
1996 "Passion de Perpétue et de Félicité, suivi des Actes." In: *Introduction, texte critique, traduction, commentaire et index par Jacqueline Amat.* Paris: Sources Chrétiennes.

Basile, C.; Benedetto, D.; Caglioti, E.; Degli Esposti, M.
2008 "An example of mathematical authorship attribution", in: *Journal of Mathematical Physics*, 49; 1–20.

Basile, C.; Lana, M.
2008 "L'attribuzione di testi con metodi quantitativi: riconoscimento di testi gramsciani", in: *AIDAinformazioni*, 26; 165–183.

Braun, R.
1979 "Nouvelles observations linguistiques sur le rédacteur de la 'Passio Perpetuae"', in: *Vigiliae Christianae*, 33; 105–117.

d'Alès, A.
1907 "L'auteur de la Passio Perpetuae", in: *Revue d'Histoire Ecclésiastique*, 8; 5–18.

de Labriolle, P.
1913 "Tertullien, auteur du prologue et de la conclusion de la passion de Perpétue et de Félicité", in: *Bulletin d'ancienne littérature et d'archéologie chrétiennes*, 3; 126–132.

Fontaine, J.
1968 "Tendances et difficultés d'une prose chrétienne naissante: l'esthétique composite de la Passio Perpetuae." In: Fontaine, J. (ed.), *Aspects et problèmes de la prose d'art latine au III^e siècle. La genèse des styles latins chrétiens*, Torino: Bottega d'Erasmo, 69–97.

Kešelj, V.; Peng, F.; Cercone, N.; Thomas, C.
2003 "N-gram based author profiles for authorship attribution", in: *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*. Halifax: Dalhousie University, 255–264.

Lanata, G.
1973 *Gli atti dei martiri come documenti processuali.* Milano: Giuffrè.

Leal, J.
2009 "Actas latinas de mártires africanos." In: *Introducción, traducción y notas de Jerónimo Leal.* Madrid: Ciudad Nueva.

Turcan, M.; Turcan-Verkerk, A.-M.
2000 "Faut-il rendre à Tertullien l'Ex libris Tertulliani de execrandis gentium diis du manuscrit Vatican latin 3852?" In: *Revue des Études Augustiniennes*, 46; 205–271.

Textual typology and interactions between axes of variation

Sylvain Loiseau

1 Introduction

This article aims at bringing some aspects of variationist frameworks into text typology and corpus-based analyses of variation. Textual typology is an approach for describing language variation specific to corpus linguistics. The question of text typology is rooted in an old philologic and literary tradition; however, in the methodological context of corpus linguistics, it has become a method for studying the general linguistic question of language variation. It may be described with four properties. First, it considers the text as a key unit, and it accounts for regularities and correlations at the level of the text. Second, it uses statistical analyses in order to make texts comparable and summarise large amounts of data. Third, it implies using corpus methodologies in order to build and search large corpora. Last of all, it uses well-established notions as categories of variation – genres, registers, or text types – giving them a slightly different meaning.

Variationist linguistics has proposed a distinction between several axes of variation. For instance, Flydal (1952), Weinreich (1954) and Coseriu (2001) – see also Völker 2009 for a recent presentation – have distinguished between up to four axes of variation: variation across space, time, socio-cultural background and situational position: "[…] a natural language is not a homogeneous system: it is a collection of different systems, which are more or less overlapping […]. In a language, there are well known differences according to space (diatopic), according to sociological and cultural groups of the community (diastratic differences), and differences according to expressivity, following the situation type and the way of speaking, differences that I call diaphasic" (Coseriu 2001: 112)[1]. Other types of variation have been analyzed, such as the "conceptional" variation (Koch and Oestereicher 2001), which accounts for the degree of spontaneity/personal implication of the speaker. Finally, variationist linguistics includes the concept of genre: "in letters, commercial negotiations,

1. "[…] une langue historique n'est pas un système homogène: c'est une collection de systèmes différents qui coïncident en partie et en partie se distinguent les uns des autres […]. Dans une langue historique, il y a les différences bien connues dans l'espaces, ou diatopiques, et aussi des différences entre les couches socio-culturelles de la communauté (différence diastratique) et des différences entre les modalités expressives déterminées par les types de situations de l'activité de parler, différences que j'appelle diaphasique."

poetry or scientific text, speakers are recycling pieces of previous utterances belonging to the same textual genre and use a large inventory of prefabricated linguistic materials." (Glessgen 2007 : 104)[2]. In sum, in this framework, "every utterance is simultaneously localized in three dimensions: variationist, conceptional, and textual" (Glessgen 2007: 106)[3].

In this article I will argue that text typology based on large corpora and statistical methods may benefit from the notion of the plurality of axes of variation as described in variationist frameworks. In the last two decades numerous statistical text classification experiments have been proposed, starting mainly with Biber (1988). These experiments have shown that there is variation across several levels of analysis: lexicon, morphosyntax, but also morphological (Baayen 1994) or prosodic features (Obin et al. 2008). They have also shown that variation occurs across several descriptive categories: texts have been shown to vary according to genre, discourse, domain, author style, but also according to socio-geographic variables (van Keune and Baayen 2006), modality, i.e. speech/writing (Biber 1988, Plag et al. 1999), "text type" (Biber 1988, Baayen 1994), etc. Morphosyntactic tagsets of different granularity have been used and a high number of statistical methods have been tested and evaluated.

The limits of automatic text classification for textual typology, however, are well known. Little consensus has emerged as to how to define and stabilize these descriptive categories (text types or genres). Many general, common sense, broad categorisations may be illustrated with statistics based on large corpora. This is especially true if the corpus is organized in several very different groups of texts. For instance, Obin et al. (2008) succeeded in automatically classifying texts into five different "discourses". But these discourses were very different: "radio news", "task map", "political discourse", "life story" and "radio interviews". It mixes oral and written texts (or oralized texts, cf. Koch and Oesterreicher 2001), different degrees of spontaneity, genre, theme… The result of the experiments does support the hypothesis that prosodic features are discriminatory (this was the aim of the paper). However it does not entail any better understanding of linguistic variation. In such a classification, one may argue that we find the categories we have put in the corpus.

More generally, the fact that the statistical classification is successful does not entail that the typology is scientifically grounded or that we gain better knowledge of the units (genres) from it. Kilgarriff (2005) showed that virtually every statistical textual classification experiment, whatever the parameters may be, shows that the distribution of features is non-random, without giving evidence that it is non-arbitrary: "the probability model, with its assumptions

2. "Pour une lettre ou une conversation d'achat comme pour une poésie ou un texte scientifique, les énonciateurs reproduisent le modèle d'autres discours semblables appartenant au même genre textuel et ils puisent dans un vaste inventaire d'élèments de langue préfabriqués."
3. "Tout énoncé s'inscrit donc parallèlement dans les trois dimensions, variationnelles, conceptionnelle et textuelle, qui sont à tout moment co-présentes."

of randomness, is inappropriate, particularly where counts are high (e.g., thousands or more)" (2005 : 268). Using a text-typology experiment, the author has shown that "given enough data, H_0 [the hypothesis that two subcorpora are distinguishable from two subcorpora which have been randomly generated on the basis of the frequencies in the joint corpus] is almost always rejected however arbitrary the data" (2005: 268). Hence, "There is no a priori reason to expect words to behave as if they had been selected at random, and indeed they do not. It is in the nature of language that any two collections of texts, covering a wide range of registers (and comprising, say, less than a thousand samples of over a thousand words each) will show such differences." (2005: 269f.). Inductive typology and typology using mainly internal properties of texts are particularly concerned by this drawback. In sum, if everything seems to vary, whatever the corpora, the linguistic features, and the statistical methods may be, are we identying and characterising a variation in a linguistic sense?

2 Hypothesis

In this paper I will explore the hypothesis that taking into account the plurality of axes of variation may be useful in textual typology. In a certain sense, the frequency of a single feature (say, the frequency of a modal verb) cannot be assigned to an axis of variation while disregarding the other axes. There is no way that we know if a feature is characteristic of, say, an author, without taking into account the properties of the genre or the discourses this author uses. An axis of variation cannot be described in isolation. The question of the relation and interaction between axes of variation is well known in variationist linguistics. For instance there is a well known relation between diastratic and diachronic variation (under some circumstances, the further you are from innovative centers, the more archaic your variety is), and between diastratic and diaphasic variation (Finegan and Biber 2001, Dufter and Stark 2002: 89, Gadet 2003). To a certain extent, classifying texts into one axis of variation – for instance, the genre – without taking into account other relevant axes of variation – such as authorship or domain, is like trying do describe diastratic variation without taking into account the age of the speakers, diatopic or diaphasic properties.

In corpora, a homogeneity regarding an axis of variation can never be "isolated" from heterogeneity through other axes. For instance, representing an idiolect in a corpus requires sampling many genres, dates, or conversational parameters. Manning (2002: 294) stresses that "there is no easy answer to the problem of getting sufficient data of just the right type: language changes across time, space, social class, method of elicitation, etc. There is no way that we can collect a huge quantity of data (or at least a collection dense in the phenomenon of current interest) unless we are temporarily prepared to ride roughshod over at least one of these dimensions of variation." This entails that

there is no "homogeneous" corpus, and, moreover, that taking into account interactions between axes of variation is required for characterising an idiolect or a genre.

3 Methodology

In order to investigate these inter-relations, I have focused on the question of which features are specific to one axis of variation, and which features are common to several axes of variation, in a corpus where several dimensions of variation are known. I have performed four independent automatic classifications, corresponding to four axes of variation on a corpus. I use a family of statistical methods, the decision trees, allowing an easy extraction the sets of discriminant morpho-syntactic features on each axis. I analyse the intersection between the sets of features. Some features are common to several sets of features and, then, shared between several axes of variation, while other are specific to one axis of variation. Each feature may be analyzed in the light of the axes it helps to predict. Can we distinguish between features specialized in one axis of variation, and features varying according to several axes of variation? Does considering features specific to one axis of variation help for characterising this axis of variation? Does analysing intersection between sets of features of each axis of variation help determine the correlation between axes of variation?

The corpus has been carefully designed in order to allow for this experiment. I used articles from the French daily newspaper *Le Monde*. Thanks to the meta-information available for each article, many dimensions can be observed:

- date: the articles available range from 1987 to 2002
- author
- genre (interview, biography, analysis, etc.)
- section (international, national, sport, opinion, enterprise, etc.)

In order to create a balanced corpus, representing various genres, authors, sections and spans of time, I selected a subset of 840 articles. This implied much pre-processing in order to deal with changes in section or genre names: a section or a genre may have different names across time due to the editorial evolution of the newspaper (see Figure 1, left). Such renaming was identified using a Hierarchical Ascendant Classification of subcorpora of articles of each section (Figure 1, right). This classification shows that some pairs of sections that are consecutive in time are also very close regarding their lexical content.

Moreover, in order to observe relationships between features and axes of variation, I have designed a corpus with as little attraction between categories as possible. For instance, I have selected articles by authors writing from 1987 to 2002, articles belonging to genres that are not specialized in only one section, etc. Eliminating as far as possible correlation between categories leads to

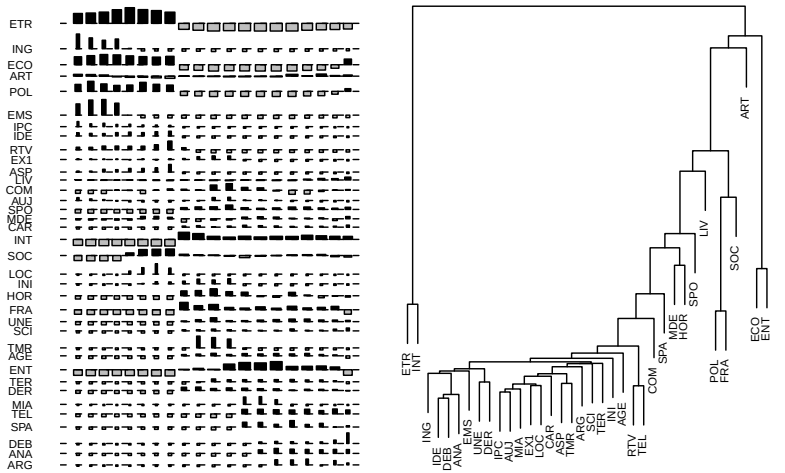


Figure 1: Left: Association plot of the sections (only sections with more than 1000 articles are kept) over the 16 years of the corpus. The section «ETR» (étranger: foreign affairs; first line) is associated with years until 1994; then the section named «INT» (international, international, middle) is associated with the following years.– *Right:* Dendrogram of a clustering of the subcorpora representing these sections (using lemma as features). Consecutive sections in the association plot are also grouped in the dendogram (ETR and INT at the far left of the plot, ECO (economy) and ENT (entreprise) at the far right, for instance).

selecting a very restricted subset of articles (840 out of 950 000 articles) containing only some authors, genres and sections of the newspaper. The 16 years were regrouped in three groups of years in order to increase the number of articles in each diachronic category. Eventually, the following categories may be observed in the corpus:

- three periods: (from 1987 to 1991 (147 articles), from 1992 to 1996 (247 articles), from 1997 to 2002 (446 articles).
- 17 authors, ranging from 4 to 120 articles.
- 4 genres (interview, biography, portrait, obituary), ranging from 56 to 350 articles. There is a strong bias due to the fact that only genres closely related to portraying poeple are represented. This is due to the fact that these genres are the only ones spreading across sections and authors.
- 6 sections (ART (art), ECO-ENT (economy and enterprise), HOR (opinion), INT-ETR (international news), POL-FRA (national news), SPO (sport), ranging from 13 (SPO) to 280 articles (INT-ETR).

Of course variables (year, genre, section, author) are not independent from each other: the chi square test reveals some attraction between every pair of tasks. The mosaic plot (Figure 2) shows associations between authors and sections on the one hand, and authors and periods, on the other hand. Nevertheless, selected authors are not completely specialized in one section or one period, and so it was the optimal "cross-balanced" corpus than could be extracted from the whole corpus.

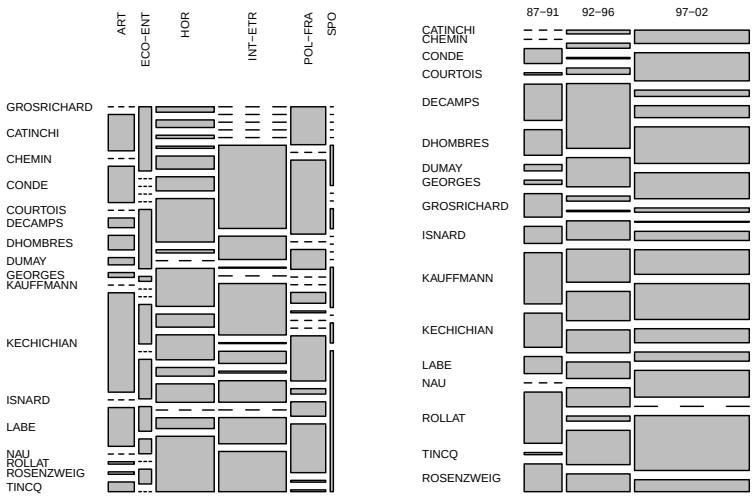


Figure 2: Mosaic plot of the associations between author and sector (left) and author and year (right)

Our corpus has been analysed using the Syntex parser (Bourigault et al. 2005). The morphosyntactic tagset of this analyser contains only 93 tags; these coarse-grained categories allow for strong robustness.

4 Results

The classification tree algorithm (Ripley 1996) is used for extracting the features making these axes predictable for each of the four axes of variation. Figure 3 shows the classification trees for two tasks: genres and sections.

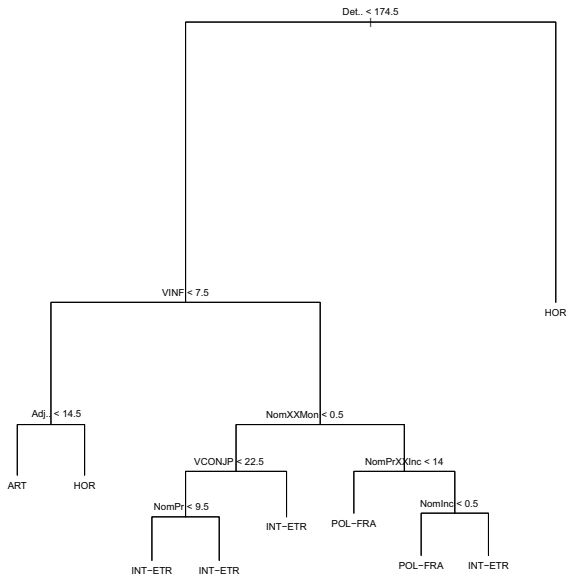


Figure 3: Classification trees for classifying articles into sections

Intersection between sets of features may be summarised as follows:

a. 73 features are never used;
b. 13 features are used for one task only (specific):
   – *Author:* CCoordAdj, CCoordPrepDe, PpaMS, Prep, PrepDet, Typo
   – *Genre:* Det, ProRel
   – *Section:* Adj.., Det.., NomInc, VINF, NomXXMon
   – *Year:* none
c. Seven features are shared between two tasks:
   – *Genre and author:* CSub, NomPrXXPrenom
   – *Genre and year:* Elim, NomFS, NomXXDate
   – *Genre and section:* NomPrXXInc
   – *Author and section:* NomPr
d. One feature, VCONJP, is shared by three tasks (genre, section, year).

Without much surprise, all features related to proper nouns (begining with "Nom-", noun) are used mainly for discrimination between section and genre, i.e. the most thematic axes of variation, and less frequently for discriminating author and year. Section and genre use different proportions of nouns. Eight features are necessary in order to distinguish between authors, while only four features are necessary to distinguish between years. Nine features are necessary in order to distinguish between genres, few of them being specific to genre: this axis of variation is mainly associated with features shared with other axes of

variation (this does not necessarily entail that it is strongly correlated to other axes of variation). Author variation, on the contrary, is the most specific axis of variation.

5 Conclusion

This experiment supports the hypothesis that some features are discriminatory for several axes of variation. Discriminatory features are not specific or "specialized" into one axis of variation; on the contrary, the same restricted set of features are used by many axes of variation. This implies that one cannot use automatic text classification for text typology, where the classes are different values along one axis of variation, without taking into account the interaction of axes of variation: there is no set of features that is discriminatory for one axis without being influenced by the others.

Further studies for a better understanding of the interaction between axes, as well as for a better understanding of the way of using these interactions for textual typology, may benefit from some machine learning algorithm supporting the classification into several sets of classes simultaneously, such as the machine learning algorithm family called "multi task learning". More generally, the analysis of these interactions seems to be a fruitful avenue for future research.

Acknowledgments. This work has been supported by TGE Adonis, CNRS.

References

Baayen, R.H.
1994 "Derivational Productivity and Text Typology", in: *Journal of Quantitative Linguistics*, 1; 16–34.

Biber, D.
1988 *Variation across speech and writing.* Cambridge: Cambridge University Press.
1993 "Using Register-Diversified Corpora for General Language Studies", in: *Computational Linguistics*, 19/3; 219–241.
1990 "Methodological issues regarding corpus-based analyses of linguistic variation", in: *Literary and Linguistic Computing*, 5/4; 257–270.

Bourigault, D.; Fabre, C.; Frérot, C.; Jacques, M.-P.; Ozdowska, S.
2005 "Syntex, analyseur syntaxique e corpus". In: *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan.

Brunet, É.
2006 "Le corpus comme une boule", in: *Proceedings Albi 2006 conference.* [Electronic source: `www.revue-texto.net/Parutions/Livres-E/Albi-2006/Brunet.pdf`]

Coseriu, E.
2001 *L'homme et son langage.* Paris: Peters.

Dufter, A.; Stark, E.
2002 "La variété des variétés: combien de dimensions pour la description?", in: *Romanistisches Jahrbuch*, 53; 81–108.

Finegan, E.; Biber, D.
2001 "Register Variation and Social Dialect Variation: the Register axiom." In: Eckert, P.; Rickford, J. R. (eds.), *Style and Sociolinguistic Variation.* Cambridge: Cambridge University Press, 235–267.

Flydal, L.
1952 "Remarques sur certains rapports entre le style et l'état de langue", in: *Norsk Tidsskrift for Sprogvidenskap*, 16; 241–258.

Gadet, F.
2003 "La signification sociale de la variation", in: *Romanistisches Jahrbuch*, 54; 98–114.

Glessgen, M.-D.
2007 *Linguistique romane. Domaines et méthodes en linguistique française et romane.* Paris: Armand Colin.

Kilgarriff, A.
2005 "Language is never ever ever random", in: *Corpus Linguistics and Linguistic Theory*, 1/2; 263–276.

Koch, P.; Oesterreicher, W.
2001 "Langage parlé et langage écrit." In: Holtus, G.; Metzeltin, M.; Schmitt, C. (eds.), *Lexikon der Romanistischen Linguistik.* Tübingen: Max Niemeyer Verlag, 584-627.

Loiseau, S.
2008 "Corpus, quantification et typologie textuelle", in: *Syntaxe et sémantique*, 9; 73–85.
Manning, C.D.
2002 "Probabilistic syntax." In: Bod, R.; Hay, J.; Jannedy, S. (eds.), *Probabilistic Linguistics.* Cambridge: The MIT Press, 289–341.
Obin, N.; Lacheret, A.; Veaux, C.; Rodet, X.; Simon, A.-C.
2008 "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features." In: *Interspeech 2008*, Brisbane, Australia.
Plag, I.; Dalton-Puffer, C.; Baayen, R.H.
1999 "Morphological productivity across speech and writing", in: *English Language and Linguistics*, 3; 209–228.
Ripley, B.
1996 *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.
van Keune, H.; Baayen, R.H.
2006 "Socio-geographic variation in morphological productivity in spoken Dutch: a comparison of statistical techniques", in: *Actes des 8es journées d'analyse des données textuelles (JADT 2006)*, 571–581.
Völker, H.
2009 "La linguistique variationnelle et la perspective intralinguistique", in: *Revue de linguistique romane*, 73/289; 27–76.
Weinreich, U.
1954 "Is a Structural Dialectology Possible?", in: *Word*, 10; 388–400.

Rank-frequency distributions: a pitfall to be avoided

Ján Mačutek

1 Introduction

The advantages – and we dare say necessity – of quantification in linguistic research have been highlighted in several papers (e.g., Köhler and Altmann 2005); yet, one should be warned that results obtained by mathematical and statistical methods are no ultimate truth. Statistics is a powerful apparatus, but it requires a proper application. A scientist must know the properties of the tools he is using – most often statistical tests yield reliable output only if certain conditions are met.

Next, even with a proper application of proper methods he must keep some critical distance, because no statistical method can be free of some possibility of error. And even if the error probability is negligible, one must be careful not to misunderstand the results obtained. For example, a satisfactory goodness of fit measure – which itself is not without problems, cf. Kvålseth (1985) and Grotjahn (1992) – does not guarantee that all features of a mathematical model match the data. As Box and Draper (1987: 424) wrote, "essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind".

We present an example of a model which fits the data excellently in terms of the chi square goodness of fit test, but the model mean and variance are significantly different from the respective empirical moments computed from the data.

2 Data and model

The data we use for the analysis are artificial. They are a "modification" (the tail of the rank-frequency distribution is made much longer, so that some discrepancies are more conspicuous) of the word rank-frequency distribution from the poem *Erlkönig* by J.W. von Goethe. In the study by Altmann and Altmann (2008), *Erlkönig* serves as the material for investigations in several linguistic areas, including word frequencies.

In spite of the artificiality of the data, they are not unrealistic for a language which is more synthetic than German and hence has more word forms (see Popescu et al. 2009 for the relation between word frequencies and analytism/synthetism measures).

In Table 1, all frequencies with ranks from 9 to 15 are 4, all frequencies with ranks from 16 to 21 are 3, etc.

Table 1: Artificial data

Rank	Frequency
1	11
2	9
3	9
4	7
5	6
6	6
7	5
8	5
9-15	4
16-21	3
22-39	2
40-225	1

The right truncated zeta distribution (cf. Wimmer and Altmann 1999: 577ff.) fits the data very well:

$$P_x = cx^{-a}, \quad x = 1,2,\ldots,n\,. \tag{1}$$

We obtain the test statistic value $\chi^2 = 14.31$, with 171 degrees of freedom and p-value $P = 1$. The optimal (with respect to the minimization of the chi square statistic) value of the parameter is $a = 0.5202$, n is usually the maximum rank with a non-zero frequency (225 for our data). Anyone who sees the p-value only must say that in terms of the chi square goodness of fit test the right truncated zeta distribution is a very good model for the data from the table.

Before we proceed to further considerations, we note that equivalent distributions must have the same moments. As to the first two moments, for the model (the right truncated zeta distribution with the parameters $a = 0.5202$ and $n = 225$) we have the mean $\mu_{zeta} = 77.36$ and the variance $\sigma^2_{zeta} = 4426.49$. For the data one obtains the empirical mean $\bar{x}_{data} = 81.71$ and the empirical variance $s^2_{data} = 5126.38$.

As will be shown in the next section (which contains technical details of statistical tests and computer simulations and can be totally or partially skipped by a reader without some basic knowledge of the two areas), the differences between the theoretical (i.e., model related) and empirical moments are significant. The results of the statistical tests we performed – the very good fit of the right truncated zeta distribution and the significant differences between the theoretical and empirical moments – are counterintuitive, but not contradictory. We discuss them in the conclusion of the paper.

3 Simulation study

The data presented in Table 1 can be considered a random sample from the right truncated zeta distribution. However, our data are not normally distributed (symmetry is obviously out of the question), hence we cannot perform the t-test directly, i.e., we cannot test the null hypotheses $\mu = \mu_{zeta}$ and $\sigma^2 = \sigma^2_{zeta}$. Next, as we are working with a parametric distribution model, non-parametric tests do not help much.

Very generally speaking, what we can do is generate a sufficient amount of similar samples and create a new sample from their means (see Robert and Casella 2004 for general random numbers generation algorithms). According to the central limit theorem, the new sample will be normally distributed (nevertheless one should test it if the number of generated samples is relatively modest) and the classical statistical tests can be applied.

The method which was chosen to approach the problem can be described as follows. We generate 100 random samples from the right truncated zeta distribution with the parameters $a = 0.5202$ and $n = 225$. The sample size is always 326, which is also the size of our artificial sample (cf. Table 1). We evaluate the mean and the variance of each generated sample. In this way we obtain two new samples – 100 means and 100 variances. We apply the Shapiro-Wilk test to check the normality of the two samples; the results confirm our expectation (the t-test can be used). Finally, we test whether the mean of the 100 means is equal to the empirical mean of the original sample (81.71) and whether the mean of variances is equal to the empirical variance of the original sample (5126.38). In both cases, the p-value is smaller than 0.01, which means that we reject both hypotheses. For control purposes we run two more tests, namely, whether the mean of means and the mean of variances are equal to their theoretical counterparts (77.36 and 4426.49, respectively). Now both answers are positive.

Naturally, one cannot exclude the possibility that the results were caused by a strange choice (very improbable, but not impossible) of the random numbers. However, the possibility is all but eliminated, as we repeated the above described process (random number generation, creating two new samples consisting of means and variances, testing) 50 times. We obtained almost the same results 50 times. The only exception is one rejection (out of 50) in the control tests for variance. As a possible alternative to our approach we mention bootstrapping (cf. Davison and Hinkley 1997), which is based on resampling with replacement from the data.

4 Conclusion

The results obtained strongly indicate that there are significant differences between the theoretical and empirical moments. At the same time, in terms of the

chi square goodness of fit test there is no reason to reject the model (the right truncated zeta distribution). And in addition, as we stated above, equivalent distributions must have the same moments.

However, one must realize that the chi square goodness of fit test does not tell us that the observed distribution and the model are the same. What it says is that the distributions are – in a way – not "too far away" from each other (we emphasize that the statement is relative, another test can give a different answer). Nothing else can be taken for granted. One cannot automatically proclaim the properties of a set of data to be the same as the properties of the model. Moreover, the chi square test is not based on moments (either empirical or theoretical). Once more we cite Box and Draper (1987: 74): "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." In this case, the model is not reliable as far as the mean and the variance are concerned. The same model can be useful for many other purposes.

It remains to explain why we focus on the rank-frequency distributions. We are convinced that they are more prone to the mentioned discrepancies, although those discrepancies (and some others) are not limited to the distributions of rank-frequency type (i.e., frequencies ordered from the highest to the lowest one). The reason is that rank-frequency distributions exhibit two "non-natural" properties. First, often there are no observed zero frequencies, especially if there is no fixed inventory for the investigated phenomenon (e.g., words). The fact can be the reason of the significantly higher sample mean (when compared with the theoretical one), as demonstrated above. The longer the distribution tail, the greater the difference that can be expected. Second, the ordering of observed frequencies creates a distribution structure which can be quite different from the one of a "natural" distribution (i.e., without ordering the frequencies) especially in the tail region, where one would expect some fluctuations in the frequency sequence (like, e.g., 0,0,2,0,1,0).

We do not claim that rank-frequency distributions should not be used as models. With frequencies of nominal data (like words, graphemes, etc., which have no obvious ordering; it cannot be said that the grapheme "m" must stand before "z", and it does not in the Russian alphabet – in contrast to, e.g., word length measured in syllables, where a word with the length 1 certainly is shorter than a word with the length 2) there are not many other possibilities left, but one should be aware of some inexactness.

By no means do we want to discourage anyone from using quantitative methods in linguistics. We only emphasize that one cannot blindly accept everything which mathematics and statistics offer. There is always some space (although not unlimited) to interpret results. If a statistical test rejects a hypothesis, one should take it as a suggestion, not as a dogma (especially if the test was performed on one sample only). Even higher cautiousness is recommended if one finds or derives a well fitting model. The words "all models are

wrong" and "all models are of the approximate nature" should sound like a refrain in every scientist's ears.

Acknowledgments. The author was supported by the Austrian FWF Lise Meitner Program.

References

Altmann, V.; Altmann, G.
2008 *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM-Verlag.
Box, G.E.P.; Draper, N.R.
1987 *Empirical Model Building and Response Surfaces.* New York: Wiley.
Davison, A.C.; Hinkley, D.V.
1997 *Bootstrap Methods and Their Applications.* Cambridge: Cambridge University Press.
Grotjahn, R.
1992 "Evaluating the adequacy of regression models: some potential pitfalls." In: Rieger, B. (ed.), *Glottometrika 13.* Bochum: Brockmeyer, 121–172.
Köhler, R.; Altmann, G.
2005 "Aims and methods of quantitative linguistics." In: Altmann, G.; Levickij, V.; Perebyinis, V. (eds.), *Problems of Quantitative Linguistics.* Chernivtsi: Ruta, 12–41.
Kvålseth, T.O.
1985 "Cautionary note about R^2", in: *American Statistician*, 39; 279–285.
Popescu, I.-I.; Mačutek, J.; Altmann, G.
2009 *Aspects of Word Frequencies.* Lüdenscheid: RAM-Verlag.
Robert, C.P.; Casella, G.
2004 *Monte Carlo Statistical Methods.* New York: Springer.
Wimmer, G.; Altmann, G.
1999 *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

Measuring lexical richness and its harmony

Gregory Martynenko

1 Introduction

The mathematics of harmony based on the theory of the golden section and Fibonacci numbers has intensively developed in last decades. Recurrent sequences, the theory of proportions and progression, iterated radicals and continued fractions, symmetry theory, number theory, combinatorial analysis, etc. are forming the nucleus of a mathematics of harmony. Ideas from the mathematics of harmony become actively embedded in philological studies, e.g., quantitative linguistics, stylometrics, the theory of poetry, prosody studies, the semiotics of mathematical languages, etc. (Grinbaum 2008; Martynenko 2010).

2 Research tasks and methodology

In lexical statistics and statistical lexicography various indices are used to measure lexical richness (diversity) of text vocabulary (Tuldava 1987; Woronczak 1965). The necessity to use these indices is partially caused by the fact that the vocabulary size strongly depends on corpus size. The utilization of standard methods is only valid to compare vocabularies in the case of equal text samples. Therefore, it is important to find parameters which do not depend on sample size. For that purpose various analytic dependencies such as "vocabulary size – sample size" are investigated and indices of lexical richness are built. In our research we pursue the same approach.

The investigation was made using frequency lists of prose texts by Anton Chekhov, Leonid Andreev, and Aleksandr Kuprin. Frequency lists for Chekhov and Andreev were made on text samples of 200000 graphic words each, and the frequency list for Kuprin is based on a sample of about 300000 words. The dependency of "vocabulary size – sample size" was analyzed using a special methodology, which is based on the least squares technique with considerable modifications caused by the specifics of research material. The quantitative processing of corpus for each author was made step-by-step, by calculating the vocabulary size each time a new story was added. It was found that the vocabulary size for each author increases at a declining rate. However, this raises some questions:

1. Does vocabulary size tend to some upper limit or not?
2. What is the analytic form of these tendencies?

3 Approximation of dependency "vocabulary size – sample size"

For an approximation of the empirical dependencies between sample size (x) and vocabulary size (y) a number of asymptotic and non-asymptotic growth functions were used. We used three groups of asymptotic growth functions. The first group contains different power functions (k is an asymptote):[1]

$$y = k - \frac{a}{x^b} \qquad \text{power function ,} \tag{1}$$

$$y = k - ke^{-ax^b} \qquad \text{exponential function (Weibull function) ,} \tag{2}$$

$$y = k - \frac{a}{(\ln x)^b} \qquad \text{logarithmic power function .} \tag{3}$$

The second group contains variants of the fractional exponential function:

$$y = \frac{k}{e^{\frac{a}{x^b}}} \qquad \text{exponential power function ,} \tag{4}$$

$$y = \frac{k}{e^{\frac{a}{e^{bx}}}} \qquad \text{double exponential function ,} \tag{5}$$

$$y = \frac{k}{e^{\frac{a}{(\ln x)^b}}} \qquad \text{exponential logarithmic function .} \tag{6}$$

The third group is formed by logistic functions:

$$y = \frac{k}{1 + \frac{a}{x^b}} \qquad \text{power (delayed) logistic function ,} \tag{7}$$

$$y = \frac{k}{1 + \frac{a}{e^{bx}}} \qquad \text{exponential logistic function ,} \tag{8}$$

$$y = \frac{k}{1 + \frac{a}{(\ln x)^b}} \qquad \text{logarithmic logistic function .} \tag{9}$$

Each of these nine functions of asymptotic growth may also be transformed into a linear dependency by means of taking the single or repeated logarithm. A system of normal equations for a linear dependency is simple, so it should not be a problem to solve it. However, in our case one of the parameters (k = asymptote) is included in the dependent variable. This fact does not allow us to use the least squares technique in its standard form. A method of approximation is described in Martynenko (1988: 77ff.). The source data is given in the first and second columns of Table 1, the number of stories being denoted by n, sample size by N.

1. For $b = 1$, the Weibull function becomes the exponential function.

Table 1: Values for short stories by Anton Chekhov

n	N	VS_{emp}	VS_{th}
10	4250	1593	1517
20	8735	2600	2603
30	13304	3488	3528
40	19413	4550	4588
50	24938	5344	5424
60	32340	6188	6408
70	38343	7006	7114
80	52139	8316	8502
90	59940	9103	9171
100	74670	10221	10257
110	88823	11355	11128
120	105084	12121	11966
130	127428	13360	12897
140	167076	14103	14105
150	198066	14610	14776
	FORECAST		
189	250000		15554
379	500000		16838
757	1000000		17088
1514	2000000		17100

These data set was approximated with the use of all the theoretical functions listed above. The results of the approximation are as follows: The match to the empirical data was obtained by asymptotic growth functions, and the Weibull function ($y = k - ke^{-ax^b}$) was the best. Its linear variant has the following form:

$$\ln\ln\frac{k}{k-y} = \ln a - b\ln x\,. \tag{10}$$

Empirical and theoretical function values for different text sample sizes are given in Tables 1–3, empirical vocabulary size being denoted by VS_{emp}, theoretical vocabulary size by VS_{th}. The values of constant coefficients are given in Table 4 (p. 129).

Table 2: Values for short stories by Leonid Andreev

n	N	VS_{emp}	VS_{th}
5	16255	3976	3964
10	40244	6504	6575
15	74789	9324	9173
20	115802	11296	11498
25	133444	12420	12348
30	175281	14205	14118
35	198592	14942	14988
	FORECAST		
44	250000		16689
88	500000		22482
176	1000000		28839
352	2000000		34777

Table 3: Values for short stories by Aleksandr Kuprin

n	N	VS_{emp}	VS_{th}
5	18340	4952	4844
10	42825	8408	8305
15	51244	9004	9269
20	62779	10222	10471
25	79192	11883	11998
30	91851	13099	13060
35	102680	13070	13903
40	129278	15830	15763
45	143391	16670	16651
50	156834	17416	17442
55	171239	18274	18239
60	183538	18981	18881
65	208193	20184	20075
70	225079	20959	20830
80	268150	22549	22560
90	310372	24103	24032
	FORECAST		
101	35000		25249
145	500000		28819
290	1000000		34789
580	2000000		38303

Correspondent curves for these data are shown in Figure 1.

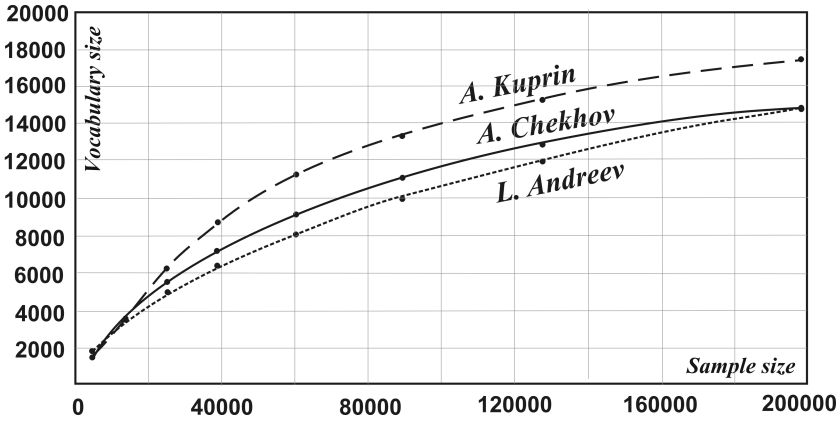


Figure 1: Empirical dependency "vocabulary size – sample size" for Chekhov, Andreev, Kuprin

In Figure 1 you can see theoretical increasing curves for the three authors. Notice that Kuprin's vocabulary size for any sample size is considerably greater than that of Chekhov and Andreev. We may observe that Chekhov's vocabulary is slightly larger than Andreev's for the analyzed samples sizes. However, at the end of the analyzed interval Andreev's curve approaches Chekhov's curve and, as will be shown below, Andreev's curve becomes higher than Chekhov's one.

This difference between Chekhov's and Andreev's tendencies may be explained by the fact that Chekhov's stories are considerably shorter than Andreev's. The diversity of Chekhov's topics gives rise to the curve in the initial part of the chart. However later this effect is no longer working as Andreev is more prone to detailed description than Chekhov. This means that the correlation between sample size and vocabulary size has powerful diagnostic potential.

Table 4: Constant coefficients of the Weibull function in the dependency of "vocabulary size – sample size" for Chekhov, Andreev, Kuprin

	Chekhov	Andreev	Kuprin
k	17100	42172	39612
a	$2.18 \cdot 10^{-4}$	$3.04 \cdot 10^{-4}$	$2.52 \cdot 10^{-4}$
b	0.80	0.60	0.65

4 Extrapolation and forecast of vocabulary size

Having determined theoretical parameters for all three curves we may forecast the values of parameters for sample sizes considerably exceeding the analyzed one and even for all literary works by some given writer. The theoretical forecast is given in Tables 1–3. The prognostic curve is shown in Figure 2.

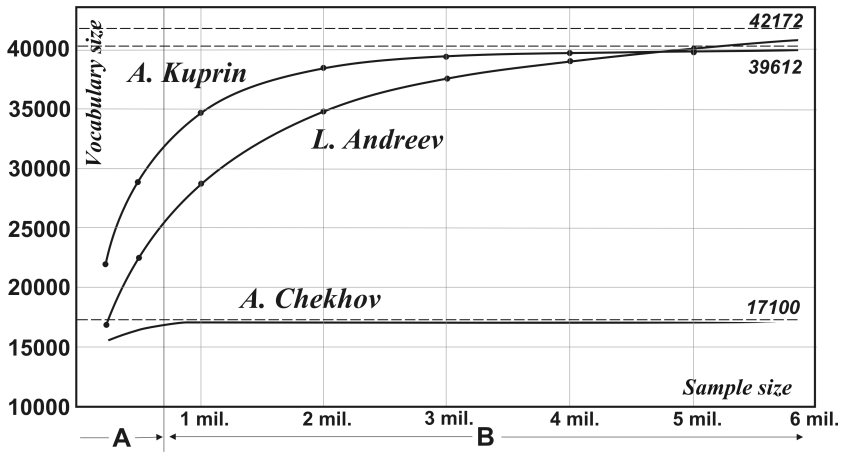


Figure 2: Prognostic (forecast) curves for dependency "vocabulary size – sample size"

Figure 2 presents an even greater difference between the writers' styles. First of all we notice that Chekhov has a very poor theoretical vocabulary. This may be explained by the fact that Chekhov, who was prone to making general conclusions, was very restrained in detailed description: he simply preferred not to go into details. If he used a detailed description, it was one single detail.

Later Chekhov's impressionism was gradually displaced by writers who were modernists. One of these modernist leaders was Leonid Andreev and he was prone to expressionism. After Andreev a new wave of naturalist writers rolled into Russian literature. Among them the bright and colorful figure of Aleksandr Kuprin was dominate. Kuprin really was a master of details.

We may suppose that these three authors delineate boundaries of stylistic "corridor" for pre-revolutionary prose writers. This is clearly seen in Figure 2, where the prognostic curve of vocabulary is increasing starting from the empirical sample size to the potential size which can be achieved from an endless writing process. With regards to actual numbers, we observe the amazing difference in asymptotic levels between Chekhov's vocabulary size and that of the two other writers.

5 Golden section of asymptotic levels for vocabulary size

Many literary critics and philologists agree that Leonid Andreev was extraordinary in his style. Our prognostic results have confirmed that he was an extraordinary writer even in the size of his vocabulary. In this aspect he was the direct opposite of Anton Chekhov, who used a comparatively small vocabulary due to his inclination to make extraordinary generalizations. The ideal short story for Chekhov was the following: "They loved one another".

We may suppose that other writers of the beginning of the 20th century are located between these two stylistic poles. Now we can use the idea of the golden section and its two fundamental properties: multiplicativity and additivity for three numbers – one central number (k_s) and two framing ones (k_{min}) and (k_{max}). The multiplicative property for maximal (Andreev's) and minimal (Chekhov's) sizes of vocabulary may be presented as following

$$k_s = \sqrt{k_{\min} \cdot k_{\max}} = \sqrt{17100 \cdot 42172} = 26854\,. \tag{11}$$

This formula allows us to calculate the vocabulary size of an "average" writer in the beginning of the 20th century. In this case addition property is also valid:

$$k_{\max} = k_{\min} + k_s = 43954\,. \tag{12}$$

Then the golden ratio for the sequence of three numbers (k_{min}, k_s, k_{max}) equals to $k_{max}/k_s = 1.618$ $k_s/k_{min} = 1.570$. The first ratio coincides with the golden section, and the second ratio is slightly less. However, we may consider that both ratios tend to the golden ratio. We may suppose that there was a writer in the beginning of the 20th century, whose maximal vocabulary size (e.g., $26854/1.618 = 1650$) is less than Chekhov's (1710). In this case the whole literary system of the given epoch may be considered to be harmonious.

6 Conclusions

A Weibull-approximation of the empirical dependencies "vocabulary size – sample size" allows us to outline hypothetical borders of the lexical diversity in Russian artistic prose in the beginning of the 20th century. It allows us to find a rule for the harmonious organization in vocabulary size of the literary epoch based on the golden ratio.

References

Grinbaum, O.
2008 *Garmonija stikha Pushkina.* [= The harmony of Pushkin's poetry]. Sankt Petersburg: Sankt Petersburg State University.

Martynenko, G.
2010 *Vvedenie v teoriju matematicheskoj garmonii teksta.* [= Introduction to the theory of mathematical harmony of text]. Sankt Petersburg: Sankt Petersburg State University.

Martynenko, G.
1988 *Osnovy stilemetrii.* [= Fundamentals of stylometry]. Sankt Petersburg: Sankt Petersburg State University.

Tuldava, J.
1987 *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki.* [= Problems and methods of the quantitative-systematic study of lexicon]. Tallinn: Valgus.

Woronczak, J.
1965 "Metody obliczania wskaźników bogactwa słownikowego tekstów". In: Mayenowa, M.R. (ed.), *Poetyka i matematyka.* Warszawa: PIW, 145–163.

Measuring semantic relevance of words in synsets

Ivan Obradović, Cvetana Krstev, Duško Vitas

1 Introduction

When delivering a query to an information retrieval (*IR*) system, a user is typically interested in information related to a particular topic, available in texts stored in electronic form. The result of this query is a selection of texts the *IR* system determines as relevant to the query. The information the user is interested in can generally be expressed in terms of *concepts*, abstract ideas or mental symbols that denote objects in a given category or class of entities, interactions, phenomena, or relationships between them. On the other hand, concepts are lexicalized by one or more synonymous words (simple or compound). For example, the concept of a "housing that someone is living in" is lexicalized by the word "house", but also by "dwelling", "home", "domicile", "abode", "habitation" or "dwelling house". Hence, the concept an IR query pertains to is in practice very often formalized by a Boolean OR combination of words, which the user believes best describe the concept in question, e.g. "house OR home OR domicile".

It goes without saying that the choice of words used in a query is of crucial importance for the relevance of the result delivered by the *IR* system. At first glance, the main problem lies in the fact that the user, when composing a query, might omit some words related to the concept, thus reducing system *recall*, the ratio of the number of relevant texts retrieved to the total number of relevant texts available. A simple query expansion by adding the omitted words would seemingly resolve this problem. However, the expansion of the set of words describing a concept in a query, although contributing to the recall in general, has an adverse effect. Namely, due to the fact that many words are homonymous or polysemous, adding new words to the query might reduce *precision*, the ratio of the number of relevant documents retrieved to the total number of (irrelevant and relevant) documents retrieved. Given this trade-off between recall and precision, words used in a query have to be very carefully selected in order to attain an optimal balance between the two.

Lexical resources such as electronic thesauri, ontologies and wordnets offer various possibilities for automatic or semi-automatic refinement of queries by adding new words to the set of words initially specified by the user. However, as we have already pointed out, this query expansion should not be performed blindly, or else it might seriously jeopardize precision. We argue that measures of *semantic relevance* of a word to a concept this word relates to in a particular language can be defined, and that they should be taken into account in

query formulation. This semantic relevance is twofold, based on the following assumptions:

1. Synonymous words, which denote a particular concept, are not used with the same frequency to denote this concept. Hence, they bear different semantic relevance to that concept. For instance, the word "home" is more frequently used to denote the concept defined as "housing that someone is living in" than the word "abode", and thus has a greater semantic relevance to this concept.
2. A homonymous or polysemous word, which can be used in more than one sense, to denote totally or partly different concepts, is more frequently used to denote one concept than another. Hence, it bears different semantic relevance to each of them. For example, the word "pen" is more frequently used to denote the concept defined as "a writing instrument which applies ink to a surface, usually paper" than it is used for the concept defined as "an adult female swan", and thus bears greater semantic relevance to the former.
3. In both cases the semantic relevance of a word to a concept can be quantified. It should be noted, however, that measures of semantic relevance we propose here should be distinguished from the mathematical model for computing the importance of a semantic feature in concept identification (Sartori and Lombardi 2004: 440) and the semantic relevance of a word in a given lexical context (Mattys et al. 2005: 486).

We can now conclude that the selection of words in a query with the aim of establishing an optimal balance between recall and precision in an *IR* system is far from a simple task. In this paper our focus is on wordnets as a means for refining queries in *IR* tasks. We propose a set of very simple and natural relevance indices to be used for tuning the query formulation process.

In Section 2 a brief overview of wordnets and the process of development of the Serbian wordnet are described, in Section 3 we describe the construction and possible use of the indices proposed, and in Section 4 some examples are given, followed by a conclusion in Section 5.

2 The development of Serbian wordnet

Wordnets were conceived in 1985 by George Miller and his associates at Princeton University who started to develop the Princeton WordNet (*PWN*), or simply WordNet, a linguistic database that maps the way the mind stores and uses language. Its aim was to serve as some sort of a mental lexicon that can be used in the scope of psycholinguistic research projects (Fellbaum 1998: 3). *PWN* was conceived as a semantic network of concepts, where each concept is represented by a set of synonymous English word-sense pairs which, accompanied by a definition of the concept, form the synset for this concept. Concepts

are interconnected by semantic relations, such as hypernym/hyponym (kind of, e.g. animal/dog) or holonym/meronym (part of, e.g. hand/finger). This database now contains about 150000 words organized in over 115000 synsetsfor a total of 207000 word-sense pairs.

The EuroWordNet project introduced multilingualism into the semantic network of concepts by building wordnets for seven European languages in a manner similar to *PWN*, and aligning them by interconnecting synsets representing the same concept in different languages by an Inter-Lingual-Index, or *ILI* (Vossen 1998: 75). Along the same lines, the BalkaNet project set as its goal the development of aligned semantic networks for Bulgarian, Greek, Romanian, Serbian and Turkish, while at the same time extending the existing network for Czech, initially developed within EuroWordNet (Tufiş et al. 2004: 11). Thirteen scientific and research institutions from Bulgaria, Greece, Romania, Serbia, Turkey, France, the Netherlands and Czech Republic gathered within the project consortium. Six teams were formed, each responsible for the development of a wordnet in one of the six languages. The core of the Serbian team was the Human Language Technologies (HLT) group at the Faculty of Mathematics, University of Belgrade (Krstev et al. 2004: 147).

The initial development of wordnets for the six BalkaNet languages was planned and realized synchronously. Namely, the core of each monolingual wordnet was built from several commonly agreed sets with a total of 8516 concepts selected from *PWN*. Beyond these sets the network for each language has been developed independently, but always within the framework set by *PWN*. This approach generated some specific problems. Namely, during the work on the development of the network the following questions have often been raised: are concepts linguistically independent or not, are the lexicalization patterns for concepts universal, is the structure of *PWN* valid for other languages as well, is the set of semantic relations built in *PWN* sufficient for all languages (Vossen 2004: 5). Although the work on the development of specific networks for Balkan languages often pointed to a negative answer to these questions, the initially established procedure has not been abandoned. The main reason was to preserve the mapping of Balkanet wordnets to *PWN*, thus making them more applicable in multilingual *IR* tasks. After the termination of the BalkaNet project the development of monolingual networks continued, and at present the Serbian wordnet contains more than 25000 words and about 15000 synsets.

Since wordnets represent concepts by means of synsets, they can be used in various ways for tuning user queries to obtain better recall and precision. The most straightforward is the detection of synonymous words omitted in a query which can improve recall. Through semantic relations wordnets also point to closely related concepts, (e.g. more general or more specific), which could also be candidates for query expansion. However, as we have already pointed out, the addition of words from synsets to a query needs to be scrutinized in some way. The relevance parameters we define in the next section could be used

as a straightforward assessment mechanism for candidate words offered by a wordnet within a query refinement task.

3 Relevance indices

In order to assess the relevance of each word in a synset for the lexicalization of the concept it is used for, we will now define a set of very simple and natural indices as numerical measures of this relevance. The semantic relevance of words in the *IR* context is best assessed by observing the way they are used in a corpus of written texts for a particular language. Thus we define our indices in direct relation to the occurrences of words in the corpus. Although the proposed indices were tested using Serbian wordnet synsets and the corpus of Serbian written texts, the methodology can be applied to any other language without modification, provided that both the wordnet and a relevant corpus for that language exist. Let **S** be the finite set of all synsets within a wordnet:

$$\mathbf{S} = \{S_i | S_i \text{ is a synset describing a specific concept, } i = 1, 2, \ldots, n_S\} ,$$

where n_S equals the total number of synsets within a wordnet; we shall also denote by $S_i \geq 1$ the total number of words within a nonempty synset S_i. Let **W** be the finite set of all words used as lexicalizations for one or more concepts:

$$\mathbf{W} = \{W_j | W_j \text{ is a word in at least one synset, } j = 1, 2, \ldots, n_W\}$$

where n_W equals the total number of different words in the wordnet. When a word $W_j \in \mathbf{W}$ is used as a lexicalization of a specific concept, described by synset S_i, it is used in a specific sense (a sense tag is attached to it), thus yielding a word-sense pair. We shall denote by $w_j \geq 1$ the total number of senses the word W_j is used in, or words-sense pairs for that word within the wordnet.

As we have already mentioned, we build the numerical parameters of a selected word W_j on the occurrences of this word, together with its inflected forms, in the corpus of written texts. We shall denote the total number of these occurrences of W_j as t_j, and the number of times the word W_j is used for lexicalization of a concept described by synset S_i as c_{ij}. In general, the equation

$$\sum_{i=1}^{w_j} c_{ij} = t_j \tag{1}$$

holds. However, given the fact that the wordnet might be incomplete, namely that all senses the word occurs in within the corpus might not be covered by the wordnet, it is also possible that

$$\sum_{i=1}^{w_j} c_{ij} \leq t_j \, . \tag{2}$$

We need to point out that, simple as it may seem at first glance, the establishing of the number of times the word W_j is used for lexicalization of a concept described by synset S_i, that is c_{ij}, can be a tedious task. Namely, unless the corpus has previously been semantically annotated using wordnet word-sense pair codes, the sense in which a word has been used in the corpus must be established manually. In that case, lexicographers have to be involved to determine the sense a word was used in each occurrence, before the corresponding numbers $c_{ij}(i = 1,2,\ldots,w_j)$ can be established.

We will now proceed to the definition of two types of indices. As one word may appear in different synsets, we will first construct the indices which express the relevance of a particular word W_j to different synsets the word appears in. The most natural way to construct such an index for a particular synset S_i is to compare the number of occurrences of this word in the corpus denoting the concept represented by synset S_i, that is c_{ij}, to the total number of occurrences of this word within the corpus, namely t_j. Thus we define the *wordnet relevance index* of the word W_j to the synset S_i as the ratio of the number of occurrences where this word has been used to denote the concept represented by the synset S_i and the total number of occurrences of this word in the corpus, namely: $WI_{ij} = c_{ij}/t_j$. It is obvious that the index range is $0 < WI_{ij} \leq 1$, where $WI_{ij} = 1$ holds if the word W_j is used in one and only one sense ($w_j = 1$), and that is to lexicalize the concept described by the synset S_i.

It is easy to prove that the sum of all wordnet relevance indices for a given word W_j is:

$$\sum_{i=1}^{w_j} WI_{ij} \leq 1 \,, \tag{3}$$

where the inequality holds only in the case that all senses the word occurs in within the corpus are not covered by the wordnet. On the other hand, as a synset may be composed of several words, we will now construct an index that expresses the relevance of a particular word W_j within synset S_i in comparison to other words in that synset. In order to construct such an index we need to calculate the total number of occurrences of all words within the corpus which denote the concept represented by synset S_i, namely:

$$a_i = \sum_{j=1}^{s_i} c_{ij} \,. \tag{4}$$

We can now define the ratio of the number of occurrences where the word W_j has been used to denote the concept represented by the synset S_i and the total number of occurrences of all words within the corpus denoting the concept represented by the synset: $SI_{ij} = c_{ij}/a_i$ as the synset relevance index of the word W_j to synset S_i. It should be noted that the range of this index is also $0 < SI_{ij} \leq 1$, where $SI_{ij} = 1$ holds when either synset S_i consists of only one

word ($s_i = 1$), and that is word W_j, or other words from that synset have not appeared in the corpus. It is obvious that the sum of synset relevance indices for all words in a given synset S_i is

$$\sum_{j=1}^{s_i} SI_{ij} = 1 \,. \tag{5}$$

Let us now take a look at a possible interpretation of the two indices. As we have already pointed out, each new word added to a query as a possible lexicalization of a concept generally increases recall and reduces precision. The indices we defined here can point to the possible impact the addition of a word will have on both recall and precision. They also indicate whether a word is synonymous as well as whether it is homonymous or polysemous.

The wordnet relevance index WI_{ij} clearly indicates whether the word W_j is used in the wordnet in one ($WI_{ij} = 1$) or more senses ($WI_{ij} < 1$), namely whether it is a homonymous or polysemous word or not. Further, for homonymous and polysemous words, it indicates the semantic relevance of the word to different concepts it relates to. Given the fact that all wordnet relevance indices of a word sum to a value less or equal to one, the higher the index for one concept, the lower for all the others. For example, a wordnet relevance index $WI_{ij} > 0.5$ indicates that the word W_j is more closely related to the concept denoted by synset S_i, than to all other concepts it also relates to. The higher the wordnet relevance index of a word, the smaller the impact on precision caused by the addition of this word in a query pertaining to the concept denoted by synset S_i. Simply put, the addition of words with high wordnet relevance indices will not considerably decrease precision. However, this index does not give any information as to the possible effect of the addition of the word W_j on recall.

On the other hand, the synset relevance index SI_{ij} indicates whether the word W_j is synonymous when it relates to the concept denoted by synset S_i. Namely, $SI_{ij} = 1$ means that only the word W_j is used to lexicalize the concept denoted by S_i, whereas $SI_{ij} < 1$ means that the synset S_i contains at least two words. As synset relevance indices for all words in a synset sum to 1, a relevance index $SI_{ij} > 0.5$, indicates that the word W_j is more related to the concept denoted by synset S_i, than all other words within the synset. Adding a word with such a relevance index in a query pertaining to the concept denoted by S_i should considerably raise the recall. On the other hand, the index does not give any information as to the possible effect of the addition of the word W_j on precision.

Hence, the assessment of the effects the addition of a word will have should be made by observing both indices. The "ideal candidate" to be added to a query pertaining to the concept lexicalized by words in synset S_i would be a word W_j from this synset with both a high wordnet and a high synset relevance

index. Conversely, a word that has a low value for both indices is a poor candidate and should be omitted in query expansion. If the user him/herself has already inserted the word in the query he/she should be advised to eliminate it.

The two indices can be combined in several different ways. We propose here a *global relevance index* GI_{ij} of the word W_j to the concept denoted by S_i the word belongs to, as a weighted arithmetic mean of the two indices:

$$GI_{ij} = \alpha WI_{ij} + \beta SI_{ij}\,, \tag{6}$$

where $\alpha + \beta = 1$. In case the user cannot decide which is more important, precision or recall, the values of α and β should be both equal to 0.5; if, however, s/he gives priority to recall, the value of β should be raised at the expense of α, whereas if the user is more concerned with precision, then a greater value should be given to α than to β.

We believe that the simple measures of relevance proposed in this section could be of value to the user when deciding which words offered by the wordnet should be considered for query expansion.

Finally, since we have based our approach on the idea of extending a query using a wordnet, we should point out that another index exists that measures the extent to which the wordnet covers all possible senses of a word as indicated by the corpus (Obradović et al. 2004: 183). Namely, due to the fact that all senses of a word that appear in the corpus are not necessarily covered by the wordnet, which we have already mentioned, a *wordnet coverage index* for the word W_j can be defined as the ratio

$$CI_j = \frac{\sum_{i=1}^{w_j} c_{ij}}{t_j}\,. \tag{7}$$

This index does not give any information pertaining to recall or precision but rather the "quality" of the wordnet with respect to word W_j. The index ranges between 0 and 1, and in the case of full coverage is equal to 1.

4 The validation procedure

The proposed approach was validated using the Serbian wordnet and different corpora of Serbian written texts. For validation purposes a set of words that we called pivotal words was chosen among the nouns and verbs that generate the largest number of word-sense pairs in Serbian wordnet. In the next step all synsets in which the pivotal words appeared were analyzed, and the words that appear in these synsets with the pivotal words were identified, and named *supporting words*. The pivotal and supporting words formed the "lexical sample" as defined by the SENSEVAL project (Kilgarriff and Rosenzweig 2000). The main objective of the validation procedure was to assess whether the initial

presumptions on the twofold semantic relevance of the words to corresponding concepts, and the relevance indices defined, are supported by experimental data.

The first corpus of approximately 1.7 million words used in the validation procedure consisted of contemporary newspaper texts. Using the available lexical tools concordances were produced for all inflectional forms of both pivotal and supporting words. Since the corpus was not semantically tagged using wordnet word-sense pair codes, the concordances of around 10000 items had to be manually analyzed by lexicographers. The senses of pivotal and supporting words were identified and marked using word-sense pair codes from the Serbian wordnet. Cases where senses of the word were not covered by the wordnet were marked as "other". On basis of the results obtained indices introduced in Section 2 were calculated. Before proceeding to an analysis of a few examples of relevance indices it should be noted that the wordnet coverage indices pointed out that the coverage of the corpus by the wordnet still varies considerably. Namely, for the words analyzed the wordnet coverage index ranged from 0.246 to 1. Only 3 out of 12 pivotal words that have been chosen had the value of the wordnet coverage index equal to 1, which means that only for these words have all the senses identified in the corpus been included in the Serbian wordnet.

Data for the Serbian noun *lice* and verb *proizvesti* obtained from the newspaper corpus are given in Table 1 and Table 2. The first column is the concept number, the second its definition, and the third the sense in which the pivotal word is used to describe the concept. Column four gives the frequencies of the appearance of the pivotal word in different senses within the corpus and the following columns give the frequencies for supporting words. In the last three columns the total number of occurrences of all words within the corpus which denote the concept is given, followed by the wordnet and synset relevance indices. In the bottom row of the table the total number of occurrences of both the pivotal and supporting words within the corpus is given.

The pivotal word *lice* has eight possible senses, and thus belongs to eight different synsets. In six of them, it is the only synset word, whereas in two of them supporting words *uloga*, *lik* and *strana* also appear. However, in the newspaper corpus this word was identified in only three out of eight possible senses (concepts 1, 2 and 3). Concept 4 was added to the table because of the appearance of the supporting word *strana* in the corpus. Cases when the synset relevance index of a word is 1 are not of great interest for query expansion, since this is the only word denoting the concept and it has to be used in any case. We will thus only point out that data from Table 1 show that *lice* has the greatest ordnet relevance index for concept 2. However, it is interesting to observe the effect of this word to queries pertaining to concepts 3 and 4. Both of its indices for concept 4 are 0, which means that adding this word to a query pertaining to this concept is not advisable, since it would not improve

Table 1: Relevance indices for the word *lice* obtained from newspaper corpus

	Concept	Sense	c_{ij}	uloga	lik	strana	a_i	WI_{ij}	SI_{ij}
1	The front of the human head	1a	33	*	*	*	33	0.063	1.000
2	A part of a person that is used to refer to a person	2a	353	*	*	*	353	0.675	1.00
3	An actor's portrayal of someone in a play	2b	1	34	3	*	38	0.002	0.026
4	A surface forming part of the outside of an object	5a	0	*	*	5	5	0.000	0.000
	Other		136						
		t_j	523	208	20	861			

recall and would have a detrimental effect on precision. The same is basically true for concept 3, since both indices are also very low. Finally, the wordnet coverage index for *lice* is $CI_j = 0.740$, which indicates that around 26% of the meanings of this word are not yet covered by the wordnet.

As for the pivotal word *proizvesti*, its wordnet coverage index $CI_j = 0.985$, which means that less than 2% of the meanings of this word are not covered by the wordnet. Table 2 indicates that this word has the greatest wordnet relevance index to concept 3, with the corresponding synset relevance index being moderately low. However, expanding the query pertaining to concept 3 with this word could be recommended: recall should be moderately raised, but precision should not be significantly affected.

In order to test the impact of the nature of the corpus to the values of relevance indices an additional validation was performed on a small literary corpus of 0.5 million words for a selected set of words. As indicated by Table 3, showing data for the word *lice* obtained from the literary corpus, index values can be largely affected by the nature of the corpus. Thus, for example, the wordnet relevance index of the noun *lice* has dramatically changed for senses 1a and 2a. This does not come as too much of a surprise since the concept that the meaning 2a refers to is more used in newspaper texts, whereas the concept that the meaning 1a refers to is more a literary concept. The changes seem to be far less dramatic for the synset relevance indices, but in order to draw some final conclusions, the impact of the nature of the corpus on relevance indices should be more systematically tested on larger corpora.

In general, the order of words within a synset is arbitrary. However, once the indices are calculated, they provide for an ordering of words in the synset.

Table 2: Relevance indices for the word *proizvesti* obtained from newspaper corpus

	Concept	Sense	c_{ij}	prouzrokovati	potaknuti	iznedriti	proizvoditi	napraviti	a_i	WI_{ij}	SI_{ij}
1	Cause to occur or exist	1a	6	31	1	*	*	*	38	0.09	0.16
2	Be the cause or source of	1b	1	*	*	0	*	*	1	0.02	1
3	Create or manufacture a man-made product	3	59	*	*	*	106	21	186	0.88	0.32
	Other		1								
		t_j	67	31	1	99	114	159			

Several possibilities exist, but a natural ordering would be in decreasing order of the global relevance index with parameters α and β chosen according to the preferences of the user. In order to optimize query expansion, the candidate words for expansion could then be offered to the user in this order.

Table 3: Relevance indices for the word *lice* obtained from newspaper corpus

	Concept	Sense	c_{ij}	uloga	lik	strana	a_i	WI_{ij}	SI_{ij}
1	The front of the human head	1a	380	*	*	*	380	0.936	1
2	A part of a person that is used to refer to a person	2a	3	*	*	*	3	0.007	1
3	An actor's portrayal of someone in a play	2b	3	6	1	*	10	0.007	0.300
4	A surface forming part of the outside of an object	5a	2	*	*	4	6	0.005	0.333
	Other		18						
		t_j	406	22	25	287			

Besides query expansion, the indices defined in this paper can also be used for wordnet refinement. Namely, if the value of the synset relevance index SI_{ij}

for the word W_j is close to 0, it can indicate that the word has been misplaced in synset S_i, especially in the case when at the same time both its total occurrence in the corpus t_j and the total number of occurrences of all words within the corpus which denote the concept represented by synset S_i, namely a_i, are considerably greater than 0. For instance, that could be the case for the word *napraviti* in the synset denoting concept 3 in Table 2. The total number of occurrences of the word *napraviti* is relatively big ($t_j = 159$) and the total number of occurrences of all words within the corpus in the synset denoting concept 3 is also considerably high ($a_i = 186$). However, if the synset relevance index for *napraviti* is calculated for the synset denoting concept 3, a relatively low value ($SI_{ij} = 0.113$) is obtained. Thus, the synonymy of the word *napraviti* with the pivotal word *proizvesti* should be reconsidered.

5 Conclusion

The wordnet and synset relevance indices proposed in this paper as a measure for semantic relevance of a word to a concept the word denotes have been applied on a small sample of chosen words and corpora for validation purposes. The results obtained indicate that the rationale for their definition rests on solid grounds. However, further analysis and testing on larger and balanced corpora are needed for their proper assessment. The problem within the validation procedure is the determination of senses a word is used in the corpus. Namely, a prerequisite for this validation is the tagging of the words in the corpus with senses used in the wordnet. To that end, automatic or semi-automatic procedures are needed in order to alleviate the time-consuming task of manual sense assignment. The indices can be useful in query expansion for determining the impact of the addition of a word on the precision and recall of the query. The calculation and assignment of indices should be focused on the most frequently used words in the corpus in the initial phase. The sensitivity of indices to the type of texts they are drawn from has been noted, but it also needs further investigation. Relevance indices can be used for wordnet refinement as well, since the determination of synsets for a given concept is not always a simple task.

References

Fellbaum, C.
1998 "Introduction". In: Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press, 1–19.

Kilgarriff, A.; Rosenzweig, J.
2000 "English SENSEVAL: Report and Results". In: *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC-2000.* Athens, 1239–1244.

Krstev, C.; Pavlović-Lažetić, G.; Obradović, I.; Vitas, D.
2004 "Using Textual and Lexical Resources in Developing Serbian Wordnet", in: *Romanian Journal of Information Science and Technology*, 7/1-2; 147–161.

Mattys, S.L.; White, L.; Melhorn, J.F.
2005 "Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework", in: *Journal of Experimental Psychology: General*, 134/4; 477–500.

Sartori, G.; Lombardi, L.
2004 "Semantic relevance and semantic disorders", in: *Journal of Cognitive Neuroscience*, 16/3; 439–452.

Obradović, I.; Krstev, C.; Pavlović-Lažetić, G.; Vitas, D.
2004 "Corpus Based Validation of Wordnet Using Frequency Parameters". In: Sojka, P.; Pala, K.; Smrz P.; Fellbaum C.; Vossen P. (eds.), *Proceedings of the Second International WordNet Conference, GWC 2004.* Brno: Masaryk University, 181–186.

Tufiş, D.; Cristea, D.; Stamou, S.
2004 "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview", in: *Romanian Journal of Information Science and Technology*, 7/1-2; 9–43.

Vossen, P.
1998 "Introduction to EuroWordNet", in: *Computers and the Humanities*, 32/2-3; 73–89.

Vossen, P.
2004 "Introduction to the Special Issue on the BalkaNet Project", in: *Romanian Journal of Information Science and Technology*, 7/1-2; 5–6.

Distribution of canonical syllable types in Serbian

Ivan Obradović, Aljoša Obuljen, Duško Vitas, Cvetana Krstev, Vanja Radulović

1 Introduction

If a canonical syllable type in a given language is denoted by a combination of the letter V, which stands for the "nucleus" of the syllable, usually a vowel, and one or more letters C, representing consonants, which surround the nucleus, forming its "periphery", then each syllable belongs to a specific canonical syllable type. It has been argued by Zörnig and Altmann (1993: 190) that the number of different syllables within a given canonical syllable type is neither chaotic nor deterministic, but rather follows a stochastic distribution. This opens the problem of finding a model, namely an adequate probability distribution that would fit the empirical data obtained by extracting syllables from texts of a given language and grouping them into canonical syllable types. A related issue to be solved is whether each language requires a specific model or more general models exist for languages belonging to the same group, such as Slavic languages; maybe even a universal model can be found.

The first result in solving this complex problem was presented by Zörnig and Altmann (1993). The essence of their approach to modeling canonical syllable types can be summarized in three steps. The first step is to propose a model with several parameters, the second is to estimate parameter values based on empirical data, namely a sample of canonical syllable types, and the third to apply the model with estimated parameters and compare the results obtained by the model and the empirical data. Although aware that this approach can be criticized for estimating parameters from a sample and then comparing the results obtained by this estimation to the same sample, we nevertheless decided to follow the same approach in our research.

Following the aforementioned procedure Zörnig and Altmann proposed a particular mathematical model and validated that model on a sample from Indonesian. The basis for the model was the discrete two-dimensional approach (Wimmer and Altmann 2005: 334) to the application of a truncated Conway-Maxwell-Poisson distribution (Conway and Maxwell 1962). Starting from a sample of 610 Indonesian syllables grouped into 12 canonical syllable types, they estimated the four model parameters, applied the model, and then further adjusted the results with two weight factors, to finally obtain satisfactory results.

Given the successful application of the Zörnig-Altmann model to Indonesian, this model presents a natural starting point for modeling canonical syllable types for other languages. However, to the best of our knowledge, no results were reported as to the validity of this model in any other language, although the Zörnig-Altmann Indonesian sample has been used, yet in another context, namely for the distribution of the average number of phonemes per syllable in the function of the number of syllables per lexical unit, in comparison with an English sample (Rousset 2004: 95, 110). For that matter, we are also unaware of a comparable model proposed for any other language. Thus our initial step in investigating the distribution of the number of different syllables within canonical syllable types in Serbian was to retrace the procedure outlined by Zörnig and Altmann. To that end we have extracted syllables from two Serbian texts generating two samples both of a size comparable with the Indonesian sample. As the Zörnig-Altmann model failed to produce acceptable results for Serbian, we proceeded by investigating another possibility, but it also failed to capture the stochastic distribution of canonical syllable types in Serbian, if such distribution indeed exists.

In Section 2 we outline the procedure we used for creating the two samples of canonical syllable types in Serbian. In Section 3 results of the application of the Zörnig-Altmann to Serbian are given. In Section 4 we discuss the results obtained by the alternative model, and in the final Section we give our conclusions.

2 Collecting syllable data for Serbian

There are five vowels in Serbian: ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’, and each of them can function both as a syllable by itself or as a syllable nucleus accompanied by one or more consonants. In addition to that, the consonant ‘r’ can also function as a syllable nucleus in Serbian. However, as opposed to the five vowels, this “syllabic” consonant cannot be a syllable all by itself, but only accompanied by one or more other consonants as in the words “prst” (*finger*) or “vrt” (*garden*). Nevertheless, we still have six canonical forms of the “V” type in Serbian. Namely, the consonant “s”, although unable to perform the “syllabic” function the way “r” can, may appear in texts all by itself as the abbreviated form of the preposition “sa” (*with*), and hence be considered as the sixth canonical “V” type syllable.

In order to investigate possible models of canonical syllable type distribution in Serbian, we have extracted syllables from sample texts coming from two sources: a monograph on the University of Belgrade and the literary magazine *Književne novine*. The first text, extracted from the monograph, consisted of around 10700 word tokens, whereas the other, from the literary magazine, consisted of about 13200 word tokens. Thus their size was comparable to the

Indonesian sample used by Zörnig and Altmann, which had around 15000 word tokens. Syllables were extracted from words following a semiautomatic procedure. Namely, we used a software product named RAS consisting of a spellchecker for Serbian and a hyphenator (Stojanović 2001). This software handles all relevant coding schemes, both alphabets used in Serbian (Cyrillic and Latin), as well as the "ekavian" and "ijekavian" dialect. However, both sample texts were in "ekavian". The hyphenator breaks word forms into syllables by inserting optional hyphens between two syllables within a word following a set of rules and a library of exceptions. However, the hyphenation rules for Serbian prohibited some words to be completely broken into syllables by RAS. Namely, according to these rules, a word can never be hyphenated after its first letter even if this letter is a vowel representing a syllable by itself. Conversely, a word cannot be hyphenated before its last letter even if it is, again, a vowel representing a syllable. Thus for example the word "ugao" (*angle*) with as much as three vowels, and hence three syllables: "u", "ga" and "o", cannot be broken into syllables by hyphenation, and hence RAS does not insert a single optional hyphen between the three syllables of this word. As a consequence of these rules, results obtained by RAS had to be manually checked and corrected in order to complete the procedure of extracting all syllables from word forms. Once this had been accomplished, we grouped the syllables into canonical syllable types and counted them.

When we completed the aforementioned procedure, the first text of 10700 word tokens generated nearly 29000 syllables, of which 964 were different, within 11 canonical syllable types; the data are represented in Table 1).

Table 1: Number of syllables within canonical syllable types for the *UM* sample

	V	VC	VCC
V	6	34	3
CV	128	424	26
CCV	176	126	7
CCCV	23	11	

The other text had 13200 word forms, which also generated around 29000 syllables, but this time 1378 of them different, within 12 canonical syllable types (cf. Table 2).

We decided to keep the two samples apart, and we will further refer to them as the *UM* (University Monograph) and *LM* (Literary Magazine) samples. The majority of syllables in both samples definitely belong to the CVC type, which is a feature Serbian shares with many other languages, including Indonesian, the language Zörnig and Altmann used for testing their model. On the other hand, the syllable CVCCC type had only one representative, namely the single-syllable word "tekst" (*text*), which appeared only in the *LM* sample (although

Table 2: Number of syllables within canonical syllable types for the *LM* sample

	V	VC	VCC	VCCC
V	6	44	7	
CV	133	620	38	1
CCV	253	221	10	
CCCV	33	12		

five times), but not once in the *UM* sample, thus equalling the lack of the fourth column in Table 1.

Although the majority of syllables belong to the CVC type followed by the CCV type as the second largest, if we look at Tables 3 and 4, which give the five most frequent syllables in both samples, we will notice that none of them belong to the largest CVC syllable type.

Table 3: Five most frequent syllables in the *UM* sample

Syllable	Frequency	Type
u	1028	V
na	873	CV
o	784	V
ni	754	CV
i	748	V

Table 4: Five most frequent syllables in the *LM* sample

Syllable	Frequency	Type
o	1047	V
je	917	CV
i	871	V
na	695	CV
u	674	V

Even more, once we ordered the syllables by the frequency of their appearance in the sample, the first CVC syllable type in the *UM* sample ("ver") appeared in place 21 with 307 occurrences, most probably due to the frequently used word university ("u-ni-ver-zi-tet") in the University monograph, whereas the rank of the first CVC syllable type in the *LM* sample ("nog") was down all the way to 73, with only 93 occurrences. Hence, we should keep in mind that we are dealing here with the numbers of different syllables of a certain type rather than frequencies of particular syllables, which might, naturally, also be a subject of a similar research.

3 Applying the Zörnig-Altmann model to Serbian

As we have already mentioned, the successful application of the Zörnig-Altmann model to Indonesian made this model a natural starting point in our attempt to find a model for canonical syllable types in Serbian. We will now briefly outline the model and parameter estimation procedure followed by Zörnig and Altmann, which we have retraced for Serbian.

Denoting the probability of a canonical syllable type with i consonants before and j consonants after the nucleus as P_{ij}, the authors proposed the following distribution:

$$P_{ij} = \frac{a^i b^j}{(i!)^k (j!)^m} P_{00} \quad i, j = 0, 1, \ldots, 4 \tag{1}$$

where P_{00} results from normalization, namely

$$P_{00} = \left[\sum_{i=0}^{4} \sum_{j=0}^{4} \frac{a^i b^j}{(i!)^k (j!)^m} \right]^{-1} .$$

The authors justified the restriction of $i, j \leq 4$ by arguing that the syllable periphery cannot be infinite. This is an obvious fact, and the periphery limits were indeed corroborated by experimental data both for Serbian and Indonesian. Even more, in both cases i and j never exceeded 3. As for the four parameters, a, b, k and m, the authors proposed that they be estimated from corresponding frequency types from experimental data. If the number of different syllables belonging to the canonical syllable type with i consonants before and j consonants after the nucleus in the sample is denoted as n_{ij}, the following parameter estimations follow:

$$\begin{aligned} a &= \frac{n_{10}}{n_{00}} , \\ b &= \frac{n_{01}}{n_{00}} , \\ k &= \frac{\ln\left(a \cdot \frac{n_{10}}{n_{20}} \right)}{\ln 2} , \\ m &= \frac{\ln\left(b \cdot \frac{n_{01}}{n_{02}} \right)}{\ln 2} . \end{aligned} \tag{2}$$

In addition to that, arguing that every language prefers one or more syllable types, the authors also proposed that the probabilities obtained by the aforementioned distribution be weighted by two weight factors a and b, proposing

for their Indonesian sample the following modification of the initial distribution:

$$P'_{ij} = \begin{cases} \beta \cdot P_{ij} & \text{for} \quad i = j = 1 \\ \alpha \cdot P_{ij} & \text{for} \quad i, j = 0, 1, \ldots, 4, \quad \text{if} \quad i \neq 1 \text{ or } j \neq 1 \end{cases} \tag{3}$$

Finally, they suggested that the weight factors again be estimated from experimental data as follows:

$$\begin{aligned} \alpha &= 1 + \frac{n_{10} \cdot b - n_{11}}{N} \, , \\ \beta &= \frac{\alpha \cdot n_{11}}{n_{10} \cdot b} \, , \end{aligned} \tag{4}$$

where N stands for the sum of all different syllables within canonical syllable types appearing in the sample:

$$N = \sum_{i=0}^{4} \sum_{i=0}^{4} n_{ij} \, .$$

The authors then proceeded to estimate the four model parameters and two weight factors from the sample of canonical syllable types for Indonesian given in Table 5.

Table 5: Number of syllables within canonical syllable types for the Indonesian sample

	V	VC	VCC	VCCC
V	6	36	7	
CV	36	391	44	2
CCV	9	61	13	
CCCV	1	4		

They further applied their model and the weight factors, and obtained a model prediction for the same sample size, which they assessed as obviously acceptable without test (Zörnig and Altmann 1993: 196). Model prediction is given in Table 6, but we must note that the results slightly differ from those in the original Zörnig and Altmann paper. Namely, as more than 15 years have passed from its publication, we were now able to recalculate all values with greater precision without too much effort. A comparison of Tables 5 and 6, however, corroborates the conclusion reached by Zörnig and Altmann.

We applied the Zörnig-Altmann approach on the two samples of Serbian canonical syllable types independently, following the outlined steps, with a slight modification we will mention shortly. However, the initial brief comparison of Serbian samples with the Indonesian sample already showed that syllables types follow a substantially different pattern in the two languages. Namely, numbers of syllables within Indonesian syllable types display a considerable symmetry when consonants are added to the syllable type on the left

Table 6: Number of syllables within canonical syllable types for Indonesian obtained by the model

	V	VC	VCC	VCCC
V	6.2	37.2	7.2	0.2
CV	37.2	404.1	43.4	1.1
CCV	9.3	55.8	10.9	0.3
CCCV	0.4	2.2	0.4	0

and right sides of the nucleus. This feature is, essentially, compliant to the symmetry of the model itself along the two dimensions. However, this is not the case with Serbian, indicating possible problems in model application. This difference can best be observed on the V-VC-VCC and V-CV-CCV syllable type sequences, which are especially important since they serve as the basis for estimating model parameters. In case of Indonesian these sequences are almost identical (6-36-7) and (6-36-9), whereas in Serbian they differ significantly, namely (6-34-3) and (6-128-176) for the *UM* sample, and (6-44-7) and (6-133-253) for the *LM* sample. Although the V-VC-VCC patterns in Serbian an Indonesian are similar, the V-CV-CCV pattern is completely different due to a very high number syllables belonging to the CCV type in both samples.

When we applied the model to two Serbian samples, without the weight factors, we obtained results presented in Tables 7 and 8.

Table 7: Number of syllables within canonical syllable types for *UM* obtained by the model

	V	VC	VCC	VCCC
V	2.2	12.7	1.1	0
CV	48.0	271.9	24.0	0.2
CCV	66.0	373.8	33.0	0.3
CCCV	18.2	103.4	9.1	0.1

Table 8: Number of syllables within canonical syllable types for *LM* obtained by the model

	V	VC	VCC	VCCC
V	1.7	12.6	2	0
CV	38.0	278.8	44.4	0.8
CCV	72.3	530.3	84.4	1.4
CCCV	32.7	239.9	38.2	0.6

If they are compared with the initial samples given in Tables 1 and 2 it is obvious that the difference between empirical and theoretical results is too big to justify the model. It should be noted that we have refrained from the weight factors, as it turned out that they only further enlarge the difference between empirical and theoretical results.

In order to illustrate the difference in results for Indonesian and Serbian we used a simple measure of estimation error, namely the square root of the mean squared difference between the number of syllables within canonical syllable types obtained from the sample (n_{ij}) and the one obtained by the model for a sample of the same size (n'_{ij}):

$$e = \sqrt{\frac{\sum_{i=0}^{3}\sum_{j=0}^{3}\left(n_{ij} - n'_{ij}\right)^2}{16}}\,. \tag{5}$$

In the case of Indonesian the error was 3.6, which equals only 0.59% of the sample size, whereas for Serbian the error amounted to as much as 84.0 (*UM*) and 140.0 (*LM*), equaling 8.71% and 10.16% of the sample size, respectively.

4 Investigating the alternative model

Although the results we have obtained clearly indicated that the Zörnig-Altmann model cannot be applied to predict the number of different syllables within canonical syllable types in Serbian, this did not necessarily mean that this number does not follow a stochastic distribution. Indeed, if we compare the frequency distribution of different syllables within canonical syllable types in two independent Serbian samples, given in Table 9, we will observe that they do follow a similar pattern, which is also obvious from the accompanying Figure 1.

Table 9: Frequency distribution of syllables within canonical syllable types in two Serbian samples (in %)

	V	CV	VC	CCV	CVC	VCC
UM	0.62	13.28	3.53	18.26	43.98	0.31
LM	0.44	9.65	3.19	18.36	44.99	0.51
	CCCV	CCVC	CVCC	CCCVC	CCVCC	CVCCC
UM	2.39	13.07	2.70	1.14	0.73	0
LM	2.39	16.04	2.76	0.87	0.73	0.07

Thus, further models, based on the same general hypothesis of stochastic distribution of different syllables within canonical syllable types, were worth

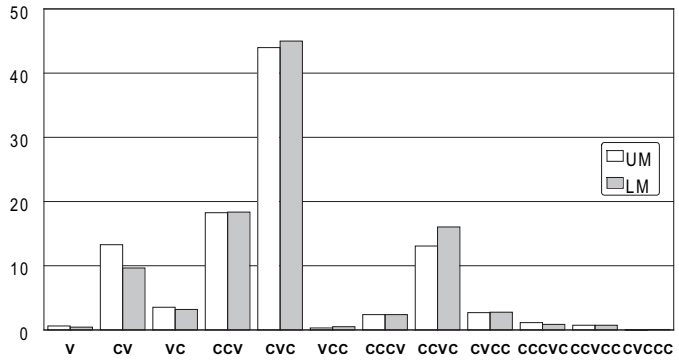


Figure 1: Frequency distribution of syllables within canonical syllable types in two Serbian samples

investigating. The alternative model we tried to apply to Serbian syllable types, similar to the approach Beőthy and Altmann (1984) used for semantic diversification of Hungarian verbal prefixes, was the two-dimensional negative binomial distribution, namely

$$P_{ij} = \left(\begin{array}{c} a+i-1 \\ i \end{array} \right) \left(\begin{array}{c} b+j-1 \\ j \end{array} \right) c^i d^j P_{00} \,, \tag{6}$$

where a, b, c and d are model parameters, and P_{00} is the sum of all values yielding the normalizing constant:

$$P_{00} = \left[\sum_{i=0}^{4} \sum_{j=0}^{4} \left(\begin{array}{c} a+i-1 \\ i \end{array} \right) \left(\begin{array}{c} b+j-1 \\ j \end{array} \right) c^i d^j \right]^{-1} . \tag{7}$$

The first problem we encountered with this model was that parameter estimations from sample values by analogy to the Zörnig-Altmann model yielded negative parameter values, and could thus not be applied. We then resorted to different approaches to the estimation of model parameters. We first tried to obtain the parameters by the minimization of the sum of squared differences between the theoretical (model) and empirical (sample) frequencies, namely

$$\sum_{i=0}^{3} \sum_{j=0}^{3} \left(P_{ij} - \frac{n_{ij}}{N} \right)^2 . \tag{8}$$

Parameter values obtained in this manner were now acceptable, but the results obtained by applying the model with these parameters were again unsatisfactory. The error measure e that we have used to assess the Zörnig-Altmann model was 85.0 for *UM* and 126.9 for *LM*, or 8.82% and 9.21% of the sample

size, respectively. Thus the alternative binomial model generated errors close to those obtained by applying the initial Zörnig-Altmann model to the two Serbian samples.

In order to rule out the possibility that the alternative binomial model keeps failing in the case of Serbian due to inappropriate parameter estimation, we made yet another attempt to estimate model parameters, this time by using maximum likelihood estimation, namely by maximizing the expression

$$\log\left[\prod_{i=0}^{3}\prod_{i=0}^{3} P_{ij}^{n_{ij}}\right] . \tag{9}$$

Parameter values obtained by this estimation were again acceptable, but the model produced results with an even greater error of 90.6 for *UM* and 129.5 for *LM*, accounting for 9.40% of the sample size in both cases.

In order to justify the two alternative approaches to parameter estimation, we decided to the test their results by estimating parameters in the initial Zörnig-Altmann model for Indonesian by both approaches and compare model results based on alternative parameter estimations with the results obtained by the parameter estimation approach used by Zörnig and Altmann. When applied in the initial Zörnig-Altmann model for Indonesian, parameters estimated by the minimization of squared differences between theoretical and empirical frequencies yielded the results presented in Table 10. If these results are compared with the original sample in Table 5, they can be assessed as quite satisfactory.

Table 10: Number of syllables within canonical syllable types for Indonesian obtained by the Zörnig-Altmann model with parameters estimation by minimization of the sum of squared differences

	V	VC	VCC	VCCC
V	4	39.4	4.8	0
CV	39.6	392.9	48.1	0.5
CCV	6.5	64.9	7.9	0.1
CCCV	0.1	1	0.1	0

This is especially true given the fact that they were obtained without the application of the two weight factors *a* and *b*. Although parameter values were slightly different from the values estimated by the original Zörnig-Altmann approach, the model generated results with an error of only 2.7, or 0.44% of the sample size, which is less than the error obtained by estimating parameters according to the original approach.

Using maximum likelihood estimation for parameters in the Zörnig-Altmann model for Indonesian yielded the results presented in Table 11. The error was this time 5.9, or 0.97% of the sample size, which is more than in the two

previous cases, but still acceptable, as the error still remained under 1% of the sample size. Besides, it should be noted that the results were once again obtained without the application of the two weight factors *a* and *b*.

Table 11: Number of syllables within canonical syllable types for Indonesian obtained by the Zörnig-Altmann model with maximum likelihood estimation of parameters

	V	VC	VCC	VCCC
V	4.4	40.8	5.6	0.1
CV	40.3	374.0	51.3	0.6
CCV	7.9	73.0	10.0	0.1
CCCV	0.1	1.5	0.2	0

Hence, parameter estimation by minimization of squared differences between theoretical and empirical frequencies and maximum likelihood estimation for parameters proved to be fully acceptable as alternatives to the parameter estimation used by Zörnig and Altmann. Failure to obtain successful results for the alternative model for Serbian thus could not be attributed to parameter estimation, but rather to the model itself.

Wrapping up this research we made two more experiments. First, in order to confirm that failure to obtain successful results for the initial Zörnig-Altmann model for Serbian could also not be attributed to parameter estimation, we used both minimization of the sum of squared differences and maximum likelihood estimation to obtain parameters for Serbian syllables, but to no avail. Second, to verify whether the alternative binomial model fails for Serbian only, we tried both parameter estimation approaches to fit this model to Indonesian syllables, but that did not yield satisfactory results either.

Hence, our research confirmed that neither the Zörnig-Altmann model nor the alternative model can be applied for modeling canonical syllable types in Serbian. On the other hand, it also confirmed that the Zörnig-Altmann model fits Indonesian data, no matter which of the three methods for parameter estimation is applied (Table 12). Finally, it also confirmed that the alternative Altmann model not only fails when applied to Serbian, but fails also on Indonesian.

5 Conclusions

Modeling the distribution of canonical syllable types in a given language turns out to be an extremely challenging problem in quantitative linguistics, as witnessed by our attempt to find such a model for Serbian. Our research results outlined in this paper, involving two languages, two models and three approaches to model parameter estimation indicate that a search for a universal

Table 12: Comparing approaches to parameter estimation for the Zörnig-Altmann model for Indonesian

	Parameter value				Error without weighting	Weight factors		Error after weighting
	a	b	k	m	e_1	α	β	e_2
Original	6	6	4.59	4.95	21.25	0.71	1.29	3.63
LSE	9.97	9.92	5.92	6.34	2.66			
MLE	9.17	9.28	5.55	6.08	5.87			

model does not look like a promising task. Hence, models should be investigated for a particular language, possibly language groups of kin languages. However, we failed to reach even this moderate goal in the case of Serbian, and the problem remains open. We would like to point out that in our pursuit for an adequate model we have tried several other options, but so far without success, and we refrained from burdening this paper with more negative results.

Another interesting research direction that we might take in the future would be to investigate possible models for the frequency distribution of all syllables, not only different syllables within canonical syllable types. Namely, as we have already noted, in the case of Serbian the most frequent syllables do not belong to the most frequent canonical syllable types, and the distribution of syllables follows an entirely different pattern from canonical syllable types. Thus further research in this area might take two different directions: searching for a model of the distribution of frequencies of canonical syllables types and searching for a model of distribution of frequencies of single syllables.

References

Beőthy, E.; Altmann, G.
1984 "Semantic diversification of Hungarian verbal prefixes. III. 'föl-', 'el-', 'be-'." In: Rothe, U. (ed.), *Glottometrika 7*. Bochum: Brockmeyer, 73–100.

Conway, R.W.; Maxwell, W.L.
1962 "A queuing model with state dependent service rates", in: *Journal of Industrial Engineering*, 12; 132–136.

Rousset, I.
2004 *Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles.* Thèse pour obtenir le grade de docteur de l'Université Grenoble III.
[Electronic source: `http://tel.archives-ouvertes.fr/docs/00/25/01/54/PDF/These_I.Rousset_10-06-04.pdf`]

Stojanović, B.
2001 "RAS u zemlji slogova", in: *PCPress*, 68.
[Electronic source: `www.pcpress.rs/arhiva/tekst.asp?broj=68\&tekstID=3111`]

Wimmer, G.; Altmann, G.
2005 "Towards a Unified Derivation of Some Linguistic Laws." In: Grzybek, P. (ed.), *Contributions to the Science of Language*. Dordrecht: Springer, 329–337.

Zörnig, P.; Altmann, G.
1993 "A model for the distribution of syllable types." In: Köhler, R; Rieger, B. (eds.), *Glottometrika 14*. Trier: wvt, 190–196.

Statistical reduction of the feature space of text styles

Vasilij V. Poddubnyj, Anastasija S. Kravcova

1 Introduction

The style of a text is characterized by a random feature set that can include syntactic words, high frequency words, bigrams, etc. Every feature is measured by a relative frequency of the occurrence in the text. These frequencies specify the feature space of text styles. Every frequency set can be presented geometrically as a point in a multidimensional feature space. A number of different texts form a point "cloud", or a text scatter plot. However, these features are not of the same value. Some features describe better the style of the author or genre: they have greater frequency variance and better distinguish texts of different authors or genres. Others have smaller frequency variance and less discrimination. Besides there are some "noise" features. In most cases, these features are statistically related to each other. This means that a random feature set has redundancy. This paper considers the transformation of the feature space that allows one to find a minimal set of statistically independent latent features.

2 Principal component analysis

A widely used statistical method of feature space transformation is that of Principal Components Analysis – PCA (Afifi and Azen 1979). This method consists of the orthogonal linear transform of data to a new coordinate system in which the greatest variance of any projection of the data lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. As a result, new factors (principal components) are uncorrelated, and the first few components almost completely define the whole scatter of points; so the components with small variances can be omitted. In the space of the first two principal components, the scatter of text-points is maximal. New features (factors) are defined by factor loadings which are the coefficients at the initial factors. Principal component analysis requires only regularity of the correlation matrix of frequency features. The frequency distribution may be arbitrary and not necessarily Gaussian. However, the probabilistic approach to principal component analysis is substantially based on normal feature distribution (Lawrence 2005). Normalization of features requires a nonlinear transformation of the initial feature space.

3 Discriminant analysis

Another method of dimensional reduction is that of Discriminant Analysis (cf. Klecka 1980, Kendall and Stuart 1979. This method consists of a linear transformation of the coordinates of a feature space which leads to the maximization of the discrepancy of the average values of new features in different classes. The deviations of new features from their average values are uncorrelated and have equal variances within the classes. In the case of text analysis, the classes are the groups of texts that differ either in author, or in genre, or in gender of the author, or in age of the author, etc. Hence, the number of classes equals the number of authors, genres, etc. The direction of the first axis of the new feature space (coordinate axis of the first discriminant functions – DF) is chosen so that the centers of classes have maximum difference from each other on this axis (for the first DF). The second axis (coordinate axis of the second DF) is directed at a right angle to the first axis so that centers of classes have maximum difference from each other on this axis (for the second DF). The third axis is directed at a right angle to the plane of the two above mentioned axes, etc. The dimension of the new feature space (of DF) is less than the lesser of the dimension of the initial space and the number of classes minus one. The discrimination property of discriminant function decreases monotonically as the number of DF grows (in the space of the first two DF, the centers of classes differ from each other in maximal degree).

4 Ranking and normalization of frequency features of text style

Formally, discriminant analysis does not require the feature distribution to be normal, the same as principal component analysis. But, it needs non-degeneracy of the correlation feature matrices within and between the classes. However, evaluation of the quality of the discriminant function method (statistical significance of DF) is based on normality of features distribution. The normalization of features presumes a proper nonlinear transformation of the initial feature space (reduction to the Gaussian distribution).

Most methods for solving discrimination, classification, and recognition problems (such as discriminant analysis, Bayes classification, recognition methods, etc.) are based on the normal (Gaussian) feature distribution (Klecka 1980, Kendall and Stuart 1979). At the same time, relative frequencies of the initial feature system of the text style not always correspond to normal distribution. By this reason the application of the well-known parametric methods of mathematical statistics to text analysis is questionable.

As for the implementation, these methods are not always mathematically correct. Therefore two approaches are possible. The first approach consist of developing non-parametric (distribution-free) methods of discriminant analysis, classification, and recognition. The second approach is to find a nonlinear

transformation of the absolute and relative frequencies of initial features that ensures the normality of both the feature distribution and the principal component and discriminant functions related to them.

This paper proposes a method of the second approach. Let us consider an ordered series of the relative frequencies for each feature in the analyzed text. Let n texts of different (in general) volumes N_i, $i = 1,\ldots,n$, be examined. We select m features of text style (for example, m syntactic words or bigrams). Each j-th feature ($j = 1,\ldots,m$) occurs in the i-th text v_{ij} times. The numbers v_{ij} are absolute frequencies of the occurrence of the j-th feature in the i-th text and can be presented in a table where the columns correspond to the features and the rows to the texts. It is obvious that the sum of absolute frequencies v_{ij} gives the whole number v_i of occurrence of the features set in the i-th text: $\sum_{j=1}^{m} v_{ij} = v_i$, $i = 1,\ldots,n$. Then $p_{ij} = v_{ij}/v_j$ is the relative frequency of the j-th feature in i-th text; $\sum_{j=1}^{m} p_{ij} = 1$ for all $i = 1,\ldots,n$. Thus the relative frequencies show the relative parts of features and assume values in the interval from 0 to 1, so they cannot be modeled in general by the normal distribution. The set of frequencies in the i-th row (the i-th text) forms a vector-row that specifies the coordinates of the i-th point-text in the feature space. We order the relative frequencies of each j-th feature in all the texts (across the j-th column, the j-th sample) in ascending order. The place of each element of a sample in the ordered series is called its rank. Thus, the vector-column $p_j = (p_{1j}, p_{2j},\ldots,p_{ij},\ldots,p_{nj})^T$ of relative frequencies of the j-th feature corresponds to the column-vector $r_j = (r_{1j}, r_{2j},\ldots,r_{ij},\ldots,r_{nj})^T$ of their ranks, $j = 1,\ldots,m$. It will be noted that equal frequencies must have the same rank which is the arithmetic mean value of ranks in a bunch of equal frequencies. In this case, the row-vector $p_i = (p_{i1}, p_{i2},\ldots,p_{ij},\ldots,p_{im})$ of relative features frequencies will be matched by the row-vector $r_i = (r_{i1}, r_{i2},\ldots,r_{ij},\ldots,r_{im})$ of their ranks, $i = 1,\ldots,n$.

It is a well known fact (Hollander and Wolfe 1999) that, under general conditions, the ranks have the uniform probability distribution in the interval from one to the sample size n. It follows from the fact that the empirical integral distribution function of ranks is the uniformly increasing step function on the interval $[0,n]$.

Let us divide each element of the column-vector of ranks r_j by $n+1$. Then the range of ranks will be the unit interval $[0,1]$ with step $1/(n+1)$, so ranks from 1 to n will be transformed to the relative ranks from $1/(n+1)$ to $n/(n+1)$. Nonexistent ranks 0 and $n+1$ will correspond to the boundary values 0 and 1 of the interval $[0,1]$. A vector of ranks obtained in this way will be called a vector of relative ranks. Then every column-vector of relative ranks will be transformed by making use of the function inverse to the integral function of the standard normal distribution. As a result, we get the set of column-vectors $x_j = (x_{1j}, x_{2j},\ldots,x_{ij},\ldots,x_{nj})^T$, $j = 1,\ldots,m$, that are correlated and have the standard normal distribution function. The column-vector

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$, $i = 1, \ldots, n$, will characterize the i-th text by the set of normally distributed new features that are correlated and have zero mean and unit variance. New features are normally distributed relative to the ranks of the relative frequencies of initial features.

As a result, we come to the nonlinear transformation of an initial feature space of non-normally distributed relative frequencies v in a new feature space of normally distributed relative ranks x. This allows one to use the parametric methods of discriminant analysis and classification (Klecka 1980, Kendall and Stuart 1979).

5 Mathematical tools of principal component analysis

Now we will find the n-column-vectors $y_l = (y_{l1}, y_{l2}, \ldots, y_{lm})^T$, $l = 1, \ldots, m$, of principal the components of the normalized data $\{x_j\}$ by the linear transformations $y_l = xU_l - y_{l0}$. Here $y_{l0} = x_{..}U_l$ are scalars (average values of principal components), $x_{..}$ – the m-row-vector of the average value n-column-vectors $\{x_j\}$, $x_{..} = \sum_{i=1}^{n} x_{ij}/n$, $j = 1, \ldots, m$; the m-column-vectors of coefficients $\{U_l\}$ are eigenvectors of the following symmetric positive definite empirical covariance $m \times m$-matrix K of vectors $\{x_j\}$. The coefficient vectors correspond to nonnegative eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_l, \ldots, \lambda_m$ that decrease monotonically with the growth of index l. These nonnegative eigenvalues define the variances of the principal components. Thus we have:

$$KU_l = \lambda_l U_l, K_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - x_{.j})(x_{ij'} - x_{.j'}), \tag{1}$$

$$My_l = 0, Dy_l = \lambda_l, \; j, j', l = 1, \ldots, m, \tag{2}$$

where M is the mathematical expectation, D is the variance, eigenvalues $\{\lambda_l\}$ are the roots of characteristic equation $\det(K - \lambda I) = 0$. In this equation I is the identity diagonal matrix. We choose $k < m$ of the first principal components as new features from the principal component system. The new features correspond to the first eigenvalues, greater than unity. As a result, we get the nonlinear statistical reduction of the feature space of texts style. The space obtained has a smaller dimension than the initial one.

6 Mathematical tools of discriminant analysis

Now we will find the n-column-vectors of the discriminant functions (DF) $z_l = (z_{l1}, z_{l2}, \ldots, z_{ln})$, $l = 1, \ldots, q$, $q = \min(m, g-1)$, of the normalized data $\{x_j\}$ by applying the linear transformations $z_l = xV_l - z_{l0}$, where $z_{l0} = x_{..}V_l$

are scalars, $x_{..}$ is the m-row-vector of average values of n-column-vectors $\{x_j\}$; $x_{..}j = \sum_{i=1}^{n} x_{ij}/n$, $j = 1,\ldots,m$; the m-row-vectors $\{V_l\}$ are eigenvectors that correspond to the matrices B and W and obey the equation $BV_l = \lambda_l WV_l$, $l = 1,\ldots,q$. The set $\{\lambda_l > 0\}$ is composed of their first q eigenvalues, that satisfy the equation $\det(B - \lambda W) = 0$ (Klecka 1980). Here $B = T - W$, where $T/(n-1)$ is the total covariance $m \times m$-matrix of vectors $\{x_j\}$:

$$T_{jj'} = \sum_{k=1}^{g} \sum_{i_k=1}^{n_k} (x_{i_k j} - x_{..j})(x_{i_k j'} - x_{..j'}),\ j, j' = 1,\ldots,m\,. \tag{3}$$

Inner summation is taken by the indices (rows) that correspond to the k-th class, $i_k = 1,\ldots,n_k$, where n_k is the number of the elements (rows) of the k-th class; $\sum_{k=1}^{g} n_k = n$; $W/(n-g)$ is the within-group covariance $m \times m$-matrix of vectors $\{x_j\}$:

$$W_{jj'} = \sum_{k=1}^{g} \sum_{i_k=1}^{n_k} (x_{i_k j} - x_{.kj})(x_{i_k j'} - x_{.kj'}),\ j, j' = 1,\ldots,m\,. \tag{4}$$

Here $x_{.kj} = \sum_{i_k=1}^{n_k} x_{i_k j}/n_k$, $k = 1,\ldots,g$, $j = 1,\ldots,m$ are elements of the $g \times m$-matrix of average values of vectors $\{x_j\}$ in the group (class). When average values of vectors $\{x_j\}$ for different classes (centers of classes) are equal ($x_{.kj} = x_{..}j$, $k = 1,\ldots,g$), then matrices T and W coincide, and all elements of the matrix B are zero. But if the averages for different classes differ from each other, then the values of elements of matrix B specify the discrepancy measure between the groups (classes). The maximization of expression $\lambda_l = (V_l^T BV_l)/(V_l^T WV_l)$, $l = 1,\ldots,q$, with respect to the weight vectors V_l provides the maximum discrimination ability of DF and leads to equation $BV_l = \lambda_l WV_l$, $l = 1,\ldots,q$, that defines the eigenvectors of the matrix $W^{-1}B$. Variables $\{\lambda_l\}$ are eigenvalues of this matrix. They give the discrepancy measure between the classes for each DF, in the order of decreasing eigenvalues.

The utility of each l-th DF (for every new feature that is obtained in this way from the initial features) can be evaluated by means of the canonical correlation coefficient (Klecka 1980) $R_l = \sqrt{\lambda_l/(1+\lambda_l)}$, $0 \le R_l < 1$, $l = 1,\ldots,q$. This coefficient expresses the level of statistical relationship of the l-th DF with its classes. The nearer the coefficient of canonical correlation is to 1, the higher is its relationship with its classes, and the greater and more secure is its discrimination of the class centers. This allows one to answer the question how many discriminant functions from the maximum number $q = \min(m, g-1)$ ensure the statistically significant discrimination of the class centers.

Let $j < q$ be the number of the first calculated DF. In discriminant analysis, Wilks' Lambda statistic Λ is used to estimate the total discriminative power of the remaining DF ("remainder discrimination"; cf. Klecka 1980):

$$\Lambda_j = \prod_{i=j+1}^{q} \frac{1}{1+\lambda_i},\ j = 0,\ldots,q-1. \tag{5}$$

If $j = 0$, one has the highest remainder discrimination because all $\{\lambda_l\}$ are nonnegative. The remainder discrimination is the lowest when $j = q-1$. So, Wilks' Λ-statistic is the "inverse" measure of class discrimination. A value of Λ close to zero indicates high discrimination of classes (it means that the centers of classes are well divided and differ greatly from each other with respect to the value of point scattering within the classes). As Λ increases to its maximum value (one) there is a gradual deterioration of class differentiation (the centers of classes fail to be significantly different with respect to the point scattering within classes).

For an estimation of the statistical significance of the discriminative power of the first j discriminant functions, Pearson's chi square test is used. It is based on the statistic

$$\chi_j^2 = -(n - \frac{m+g}{2}) \ln \Lambda_j,\ j = 0,\ldots,q-1. \tag{6}$$

This statistic has the probability density function χ^2 with $\nu_j = (m-j)(g-j-1)$ degrees of freedom under the condition that hypothesis H_0 is true (Klecka 1980). That means the remaining $q-j$ DF don't improve the discrimination ability of the first j DF (they don't increase the distance between the centers of classes). It allows one to calculate the significance level P (p-level) of the chi square test that has been reached (the actual probability of an error of the first kind to reject by mistake the null hypothesis when it is true): $P_j = 1 - F(\chi_j^2|\nu_j)$, where $F(\chi^2|\nu)$ is the integral function of the chi square distribution with ν degrees of freedom.

The interpretation of the discriminant functions as hidden parameters that determine the differences of classes can be achieved by correlation coefficient analysis (factor loadings analysis) of the column-vector z_l of the discriminant functions with column-vectors x_j of the normalized relative ranks:

$$\rho_{ij} = \frac{1}{n-1} \sum_{i=1}^{n} z_{il}(x_{ij} - x_{\cdot\cdot j}) / \sqrt{Dz_l Dx_j},\ l = 1,\ldots,q,\ j = 1,\ldots,m\,. \tag{7}$$

It is well known (Sachs 1972) that statistic $t = \rho\sqrt{(n-2)/(1-\rho^2)}$ has Student's t-distribution with $\nu = n-2$ degrees of freedom, provided that z_l and x_j have normal distribution and the null hypothesis (the correlation coefficient $\rho = 0$) is true. This enables one to find the critical value of Student's statistics as quantile $t_{crit} = t_{n-2,1-P_{crit}/2}$ of level $1 - P_{crit}/2$ of this distribution. The critical significance level of Student's t-test should be fixed, e.g., $P_{crit} = 0.05$. From here one easily gets the critical value $\rho_{crit} = t_{crit}/\sqrt{(n-2)+t_{crit}^2}$ of the correlation coefficient which specifies $(1-P_{crit})\cdot 100\%$-th interval $[-\rho_{crit}, \rho_{crit}]$ of

the statistical insignificance of the correlation coefficient. The values of the correlation coefficient outside this interval are statistically significant on $P \leq P_{crit}$ level of significance.

7 Example of feature space reduction on the basis of methods of principal components and discriminant analysis

The above described procedures of constructing m-row-vectors r_i, m-row-vectors x_i, m-row-vectors y_i, and q-row-vectors z_i ($i = 1, 2, \ldots, n$) from the original m-row-vectors p_i of relative frequencies of text style features for each textual work are implemented in the *StyleAnalyzer* software (Shevelyov and Poddubnyj 2010) which is intended for the complex statistical analysis of textual work styles of different authors, genres, etc. Figures 1–3 give examples of using the described approach to ranking, normalization and reduction of a feature space on the basis of the methods of principal components and discriminant analysis.

Textual material is represented by 80 large works of fiction by 11 Russian authors of the 19th century (11 works by N.V. Gogol', 3 by I.A. Goncharov, 18 by F.M. Dostoevskij, 2 by I.A. Kuprin, 3 by M.Ju. Lermontov, 7 by N.S. Leskov, 9 by A.S. Pushkin, 2 by M.E. Saltykov-Shchedrin, 8 by L.N. Tolstoj, 13 by I.S. Turgenev, 4 by A.P. Chekhov).

We used 55 syntactic words as text style attributes.[1] Absolute frequencies of their occurrence in the text are the text style features. These frequencies are being presented in *StyleAnalyzer* in the form of a spreadsheet with indication of authors and texts in rows and that of style attributes in columns. Figure 1 shows the connection between the original attributes – the relative frequencies of one in 55 features (namely, the forth one) in 80 texts (*init-data*), the ranks of relative frequencies (*rank-data*) and the relative ranks (normally distributed after the non-linear transformations) of relative frequencies (*gauss-data*).

Eigenvalues of the covariance matrix K are the variances of the principal components. The calculation of them for *init-data* and *gauss-data* variables shows that several first principal components are responsible for the majority of text variability. For example, the first six principal components (10.1% of its total number) explain 51.4% of the feature variability for *gauss-data* and 49.6% for *init-data*.

Eigenvalues of matrix $W^{-1}B$ for *init-data* and *gauss-data* variables are the variances of the discriminant functions of these variables. One can see that only $q = \min(m, g-1) = 10$ of them are other then zero; here $m = 55$ is the number

1. These syntactic words are: в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под, и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто, не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо.

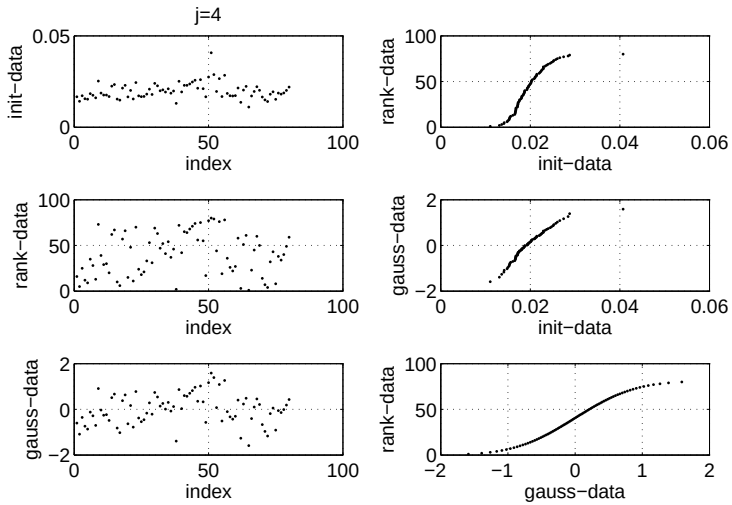


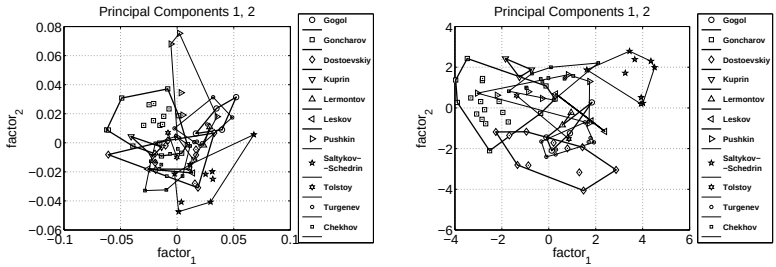
Figure 1: Nonlinear transformation of the data for the fourth feature

of original features (syntactic words), $g = 11$ is the number of classes (writers, authors of works). The calculation of the significance levels (p-levels) of the discriminant functions shows that almost all DF are statistically significant ($P < 0.05$ for the first nine DF).

The points with the markers of different types in Figure 2 represent 80 fiction works of 11 writers of the 19th century in the coordinates of the first two principal components (factors 1 and 2) for *init-data* (see Figure 2a) and *gauss-data* (see Figure 2b) variables. Convex hulls of sets of work-points for each author are shown by the closed broken lines. One can see that the normalized relative ranks of relative frequencies (*gauss-data*) distinguish between writers better than the relative frequencies.

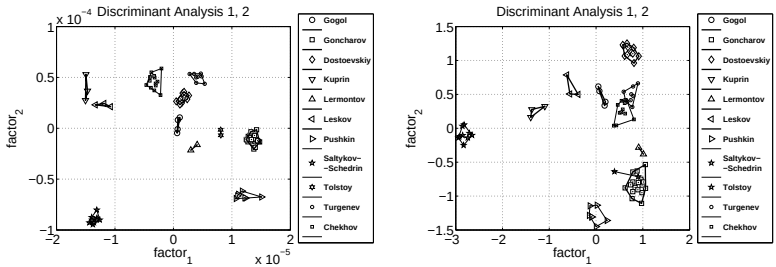
The points with the markers of different types in Figure 3 refer to the same 80 fiction works of 11 writers of the 19th century in the coordinates of first two discriminant functions (factors 1 and 2) for *init-data* (see Figure 3a) and *gauss-data* (Figure 3a) variables. Convex hulls of sets of the work-points for each author are shown by the closed broken lines.

If one compares Figures 2 and 3, one sees that discriminant analysis provides full discrimination of classes by relative frequencies (*init-data*) and almost full discrimination by their normalized relative ranks (*gauss-data*), whereas the author classes overlap significantly in the course of principal component analysis.



(a) init-data *(b)* gauss-data

Figure 2: Text representation in the coordinates of the first two principal components (features are 55 syntactic words)



(a) init-data *(b)* gauss-data

Figure 3: Text representation in the coordinates of the first two discriminant functions (features are 55 syntactic words)

8 Conclusion

Thus, discriminant analysis ensures a considerably better discrimination of authors in terms of 55 syntactic words as compared with the analysis of principal components, though both methods provide graphical representation of the whole Russian fiction literature of the 19th century by sets of dots (representing texts) in the plane. This is to be expected since the discriminant analysis provides a transformation of the original attribute space of text styles that maximally increases the mean-square distance between the class centers fixing the distance variance between the elements (dots-texts) inside the classes on a constant level.

In other words, discriminant analysis makes author classes equally compact and maximally discriminated from each other. Residual overlapping of classes indicates the proximity of text styles of different authors that appears in the overlapping classes in the corresponding feature space.

In conclusion, it will be noted that close results could be obtained when the method of principal components and discriminant analysis is applied directly to ranks of frequencies rather than the normalized relative ranks of relative frequencies of attributes. This is due to the fact that gaussianity of data is no longer significant when the indicated methods are used for the multidimensional analysis of texts, though the calculations of statistical significance of the results may turn out to be incorrect.

References

Afifi, A.A.; Azen, S.P.
1979 *Statistical Analysis: A Computer Oriented Approach.* New York etc.: Academic Press.
Hollander, M.; Wolfe, D.A.
1999 *Nonparametric Statistical Methods.* New York etc.: Wiley.
Kendall, M.G.; Stuart, A.
1979 *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship.* New York: Oxford University Press.
Klecka, W.R.
1980 *Discriminant analysis, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-019.* Beverly Hills–London: Sage Publications.
Lawrence, N.
2005 "Probabilistic Non-Linear Principal Component Analysis with Gaussian Process Latent Variable Models", in: *Journal of Machine Learning Research*, 6; 1783–1816.
Sachs, L.
1972 *Statistische Auswertungsmethoden.* Berlin etc.: Springer.
Shevelyov, O.G.; Poddubny, V.V.
2010 "Complex investigation of texts with the system «StyleAnalyzer»". In: This volume, pp. 207–212.

Quantitative properties of the Nko writing system

Andrij Rovenchak, Valentin Vydrin

1 Introduction

Nko (ߒߞߏ) is an indigenous writing system invented in 1949 by a Guinean encyclopedist and enlightener Sòlomána Kántɛ (1922–87). The script was intended as a writing for the Manding languages of West Africa (see map represented in Figure 1).[1]

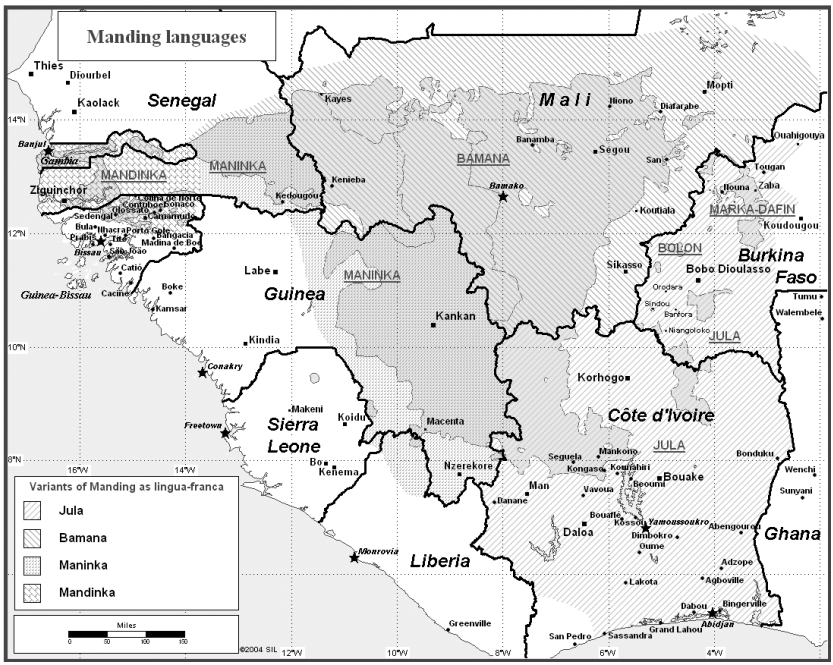


Figure 1: Manding languages as lingua franca

Nko gained some popularity among the Manding speakers not only in Guinea, but also in Liberia, Mali, Côte d'Ivoire, and in diaspora in Nigeria, Egypt, and elsewhere. According to the ideology of Nko, the Nko writing is to be used for all Manding variants, but in reality, it promotes the establishment of the Guinean Maninka (Maninka-Mori) as the common literary norm.

1. www.sil.org/SILESR/2000/2000-003/Manding/MandingLinguaFranca.htm

In this paper, we focus on the quantitative properties of the Nko script, including character complexities and their correlation with frequencies, analysis of the grapheme–phoneme correspondence and phoneme distribution. This work is meant to extend the quantitative data on indigenous African scripts limited so far to the Meroitic script (cf. Smith 2007) and the Vai syllabary (cf. Rovenchak et al. 2009).

2 Nko alphabet

Nko script is a right-to-left running alphabet consisting of 29 letters: 7 vowels, syllabic nasal, and 19 consonants; two characters are used to denote combinatory phonetic transformations (cf. Dalby 1969; Vydrin 1999); see Table 1. The script includes seven tonal diacritics, the nasalization mark (see Table 2), several additional diacritics for foreign sounds, as well as 10 digits and some other special marks (punctuation and related). The characters within a word are written continuously, joined by a horizontal bar resting on the baseline. The nasalization mark is a dot placed below the respective vowel, and the tone marks are placed above the vowels.

For the complexity analysis, we will use the isolated shapes of the letters. These are playing in Nko the part of capitals to some extent: they are used in abbreviations, titles, etc. The changes into initial, medial, and final shapes are unique and systematic for all the characters, unlike the Arabic script.

3 Nko script complexity

In calculating character complexities we adhere to the approach of Altmann (2004). Namely, a point is given the weight 1, a straight line evaluates to 2, an arc 3. A continuous connection produces the weight 1, a crisp gives 2, and a crossing gives 3.

Some modifications by Mačutek (2008b) apply to calculate the number of connections. Complexities of the Nko characters and diacritical marks are given in Tables 1 and 2.

Table 1: Complexities of the Nko characters*

		Point	Straight line	Arc	Continuous connection	Crisp	Crossing	Complexity
		1	2	3	1	2	3	
ߊ	a		1					2
ߍ	e			2	2			8
ߎ	i		3			3		12
ߏ	ɛ		2			1		6
ߐ	u		3			2		10
ߑ	o		3			2		10
ߒ	ɔ		2	2	2	1		14
ߓ	N		1	2	1	1	1	14
ߔ	b		3			2		10
ߕ	p		3			2		10
ߖ	t		1	1		2		9
ߗ	j		1	1		1		7
ߘ	c		2			1		6
ߙ	d		1	2		5		18
ߚ	r		2				1	7
ߛ	rr		3				2	12
ߜ	s		4			4		16
ߝ	gb		3			3		12
ߞ	f		3	1		4	1	20
ߟ	k		3			2		10
ߠ	l		1	1		1	1	10
ߡ	*dental*→n		2			1		6
ߢ	m		3			3		12
ߣ	ɲ		3	2	2	3		20
ߤ	n		3			2		10
ߥ	h		3			2		10
ߦ	w		5			5		20
ߧ	y		2	2	2	2		16
ߨ	y→ɲ		3	2	2	3		20
ߩ	g	1	3			3		13

* The letter ߐ (u) might look like an arc (∪) in this typeface, while it is treated as a ⊔ shape; the letter ߩ (g) does not belong to the original alphabet, see comment on this issue in Section 5.

Table 2: Complexities of the Nko diacritics (shown attached to < ߊ> /a/)

	Point	Straight line	Arc	Continuous connection	Crisp	crossing	Complexity
	1	2	3	1	2	3	
◌̄ /á/*		1					+2
◌̃ /à/		3			2		+10
◌̇ /ă/	1						+1
◌̂ /á:`, áa`/		2			1		+6
◌́ /á:, áa/		2			1		+6
◌̀ /à:, àa/		3			2		+10
◌̆ /ă:, ăa, àá/		2			1		+6
◌̣ /ã, an/	1						+1

* In a word-final position only, otherwise left unmarked

The distribution of complexities is presented in Table 3.

Table 3: Distribution of complexities (f_C = number of characters with complexity C)

C	f_C	C	f_C	C	f_C	C	f_C
		6	**3**	11	**0**	16	**2**
2	**1**	7	**2**	12	**4**	17	**0**
3	**0**	8	**1**	13	**1**	18	**1**
4	**0**	9	**1**	14	**2**	19	**0**
5	**0**	10	**8**	15	**0**	20	**4**

The uniformity hypothesis can be tested by the Wald–Wolfowitz runs test. Let I denote the inventory size and R is the range of complexities. For Nko one has $I = 30$ and $R = 18$. The uniform distribution of data means that all expected frequency values equal $E = I/(R+1)$. A run is a sequence of frequencies which are either all greater than E or all smaller than E. We have $E = 30/(18+1) \simeq 1.58$ and $r = 12$ runs, namely $[\underline{1,0,0,0},\underline{3,2},\underline{1,1},\underline{8},\underline{0},\underline{4},\underline{1},\underline{2},\underline{0},\underline{2},\underline{0,1,0},\underline{4}]$. Let $n = R+1$, and n_1 is the number of frequencies smaller than E while n_2 denotes the number of frequencies greater than E (we have $n = 19$, $n_1 = 12$, $n_2 = 7$). The number of runs is considered random (meaning that the distribution is uniform) if

$$z = |r - E(r)| - 0.5\sigma_r < 1.96,$$

where

$$E(r) = 1 + \frac{2n_1n_2}{n} \quad \text{and} \quad \sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}\,.$$

The calculation yields $z \simeq 0.84 < 1.96$, and thus the uniformity hypothesis is confirmed for complexities. Performing the same test for the characters with all possible combinations of diacritics we obtained $z \simeq 0.91 < 1.96$ confirming the uniformity hypothesis in this case as well.

4 Complexity vs. frequency

The analysis of frequency distribution of the Nko characters was made on a small sample of 2881 tokens consisting of five subsamples, three of them taken from the *Sinjiya* journal (December 2005, no. 66, page 2 and 3), *Yelen* journal (Oct/Nov 2002) and an entry from the *Nko Kodofolan Kanjamadi* dictionary. The frequency list with complexities of the respective characters is given in Table 4.

Table 4: Frequency list of the Nko characters compared to complexities (f_i vs. C_i)

i			f_i	C_i	i			f_i	C_i	i			f_i	C_i
1	ߊ	a	0.1972	2	11	ߏ	b	0.0403	10	21	ߣ	n*	0.0104	6
2	ߋ	k	0.0736	10	12	ߐ	u	0.0385	10	22	ߤ	N	0.0101	14
3	ߌ	l	0.0725	10	13	ߑ	s	0.0361	16	23	ߥ	w	0.0042	20
4	ߍ	i	0.0725	12	14	ߒ	m	0.0333	12	24	ߦ	gb	0.0035	12
5	ߎ	d	0.0573	18	15	ߓ	f	0.0271	20	25	ߧ	h	0.0024	10
6	ߏ	ɛ	0.0517	6	16	ߔ	t	0.0267	9	26	ߨ	c	0.0010	6
7	ߐ	ɔ	0.0482	14	17	ߕ	r	0.0222	7	27	ߩ	g	0.0007	13
8	ߑ	n	0.0451	10	18	ߖ	y	0.0180	16	28	ߪ	ɲ*	0.0007	20
9	ߒ	e	0.0420	8	19	ߗ	j	0.0118	7	29	߫	p	0.0003	10
10	ߓ	o	0.0413	10	20	ߘ	ɲ	0.0111	20	30	߬	rr	0.0000	12

* An asterisk (*) denotes combinatory phonetic transformation

The correlation coefficient between frequencies and complexities has a small but negative value of $r = -0.39$. It means that simpler shapes occur with rather greater frequency (as expected) but this statement holds only roughly. The best fit for both phoneme and grapheme frequencies is given by the negative hypergeometric distribution (cf. Grzybek et al. 2006; Mačutek 2008a) with parameters $K = 2.7447$, $M = 0.6696$, $n = 26$ (for phonemes, $C = 0.0399$), and $K = 2.9932$, $M = 0.6727$, $n = 29$ (for graphemes, $C = 0.0438$; see Figure 2).

5 Phonology of Maninka-Mori

The phonological system of Maninka-Mori comprises seven vowel phonemes: /a, e, ɛ, i, o, ɔ, u/, one syllabic nasal: /N/ [n̩], and 18 consonant phonemes: /b, p, t, d, r, s, gb, f, k, l, m, ɲ, n, h, w, j [ʤ], c [ʧ], y [j]/ giving in total 26 phonemes.

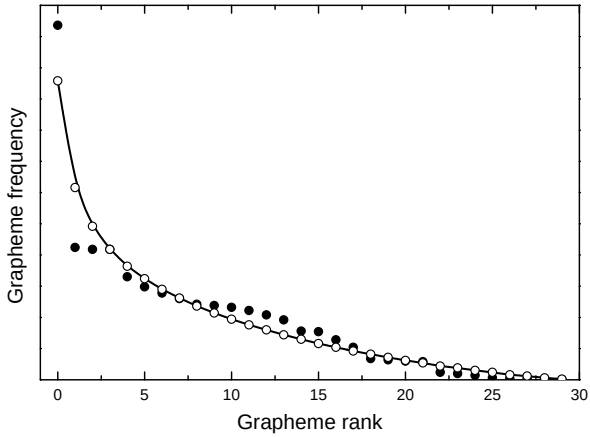


Figure 2: Fit for grapheme frequencies. Filled circles denote observed frequencies, open circles are the values of the fitting function

Note that in loanwords found in Nko texts foreign phonemes occur sometimes (/g/, /z/, etc.). To represent them in writing, letters of the original Nko alphabet are supplemented by several types of diacritics, e. g., <߭̇> for /g/ (a dotted sign for /gb/ <߭>), <ߍ̇> for /z/ (from /c/ <ߍ>), etc. In the present study, such foreign phonemes are not taken into consideration when the grapheme–phoneme relations are analyzed. In a more detailed consideration, all the vowels can be short, long, short nasalized and long nasalized. Occasionally, syllabic nasal can be long as well. Such an approach gives $7 \times 4 + 2 = 30$ vowel phonemes (counting here the nasal as well).

Maninka is a tonal language. However the nature of its tones differs from that of, e. g., Chinese or Vietnamese, and the tonal distinctions cannot be considered as phoneme-differentiating features. All the vowels can have four tones and the syllabic nasal has two tones for the short one and two tones for the long one. This produces $7 \times 4 \times 4 + 2 + 2 = 116$ units for vowel phonemes with tonal distinction and thus counting the vowels together with the 18 consonants gives 134 units.

6 Script peculiarities

The Nko script was designed specifically for the Manding languages and the orthography is quite ‘shallow’ (cf. Coulmas 2004). Anyway, some peculiarities make the script different from an ideal variant with a one-to-one correspondence between phonemes and graphemes. The deviations are mostly due to the so-called *gbàralí* (contraction) rule: if the same short (oral) vowel with the

same tone appears in two consecutive syllables, only the second one is written: *bóló*̀ 'hand' is written as <ßßß> <b(ó)ló>.

This rule does not apply to short nasal vowels and additionally has some exceptions. If the same vowel with the same accent is repeated after two identical consonants (except /r/), both are written: *bùbùlí*̀ 'fidelity' <ßßßßß>. In polysyllabic words, vowels are combined into pairs, e. g., in a three-syllable word with the same vowel repeated thrice, the first two are combined: *súlúkú*̀ 'hyena' <ßßßß> <s(ú)lúkú>. To avoid this omission (especially in foreign words) a special mark <ß> is used, e. g., the name Christophe (Kristof) is written as <ßßßßßß> otherwise without this mark it would read /kírísɔ́tɔ́f/.

Short vowels thus have at least two graphemic representations, one of which is an empty grapheme: <>. Three consonant phonemes have two graphemic representations each, namely, /r/: <ß>, <ßß>; /n/: <ß>, <ß>; /ɲ/: <ß>, <ß>. The second symbol for /n/ is used to denote assimilatory changes of dentals after /N/ or a nasalized syllable, while the second symbol for /ɲ/ denotes a similar transformation of /y/. A duplicated symbol for /rr/ occurs rarely and is used in a limited set of words. Otherwise, the "one phoneme = one grapheme" principle holds. Graphemic representations of the Maninka phonemes are listed in Table 5.

7 Phonemes and graphemes

If one does not consider diacritics and, consequently, does not take into account length, nasalization or tones, 26 phonemes are counted in Maninka-Mori (cf. Section 5. Sixteen phonemes have one graphemic representation and ten phonemes have two graphemic representations, which yields the mean orthographic uncertainty

$$U = \frac{1}{N} \sum_k f_k \log_2 k = 0.38, \tag{1}$$

where N is the total number of phonemes and f_k is the number of phonemes having k graphemic representations. The mean orthographic uncertainty of an ideal alphabet, with one-to-one phoneme–grapheme correspondence, equals zero. Real alphabets deviate from this ideal variant to different extents: the Italian (Roman) alphabet has $U = 0.56$, the Ukrainian (Cyrillic) alphabet has $U = 1.12$, etc., cf. Buk et al. (2008). One should expect even smaller values for new orthographies created for a specific language, and Nko belongs to this category.

If length and nasalization of vowels is taken into account, the number of phonemes amounts to 48. In this case, however, it is hard to find an appropriate treatment for the calculation of the orthographic uncertainty as there are no separate diacritics for length, but they are coupled with tone. Therefore, if one

Table 5: Graphemic representations of the Maninka phonemes

Phoneme	Nko	Roman	Examples		
/a/	< ߊ >	<a>	ߘߊ	dá`	mouth
	< >		ߓߊߘߊ	bádá`	home
/e/	< ߋ >	<e>	ߛߋ߫	sé	to reach
	< >		ߜߋߙߋ	gbéré`	dry land
/ɛ/	< ߍ >	<ɛ>	ߞߍ߫	kɛ́	to do
	< >		ߥߍߟߍ	wɛ́lɛ́`	ear (*respectful*)
/i/	< ߌ >	<i>	ߟߌ	lí`	honey
	< >		ߞߌߟߌ	kílí`	egg
/o/	< ߏ >	<o>	ߝߏ߬	fò	to greet
	< >		ߞߏߟߏ	kóló`	bone
/ɔ/	< ߐ >	<ɔ>	ߞߐ	kɔ́`	back
	< >		ߞߐ߬ߟߐ߬ߟߐ	kɔ̀lɔ̀lɔ́`	rainbow
/u/	< ߎ >	<u>	ߞߎ	kú`	tail
	< >		ߡߎߙߎߕߌ	múrútí`	revolt
/N/ [n̩]	< ߒ >	<n>	ߒ߫	ń	I
			ߒ߬	ǹ	we
/b/	< ߓ >	<b>	ߓߊ	bá`	river
/p/	< ߔ >	<p>	ߔߊ߬ߕߊ߬ߞߎߙߊ	pàtàkúrá`	loincloth
/t/	< ߕ >	<t>	ߕߋ߬ߟߋ	tèlé`	sun
/j/ [ʤ]	< ߖ >	<j>	ߖߎ߯	júú`	enemy
/c/ [ʧ]	< ߗ >	<c>	ߗߋ߲߱	cèén	beautiful
/d/	< ߘ >	<d>	ߘߎ߱	dùú`	earth
/r/	< ߙ >	<r>	ߟߍߙߊ	lɛ́rá`	book
	< ߚ >		ߓߏ߫ߚ	bórr	*expressive adverb*
/s/	< ߛ >	<s>	ߛߊ߱	sàá`	sheep
/gb/	< ߜ >	<gb>	ߜߏ߬ߟߏ	gbòló`	skin
/f/	< ߝ >	<f>	ߝߎ߬ߘߎ	fùdú`	stomach
/k/	< ߞ >	<k>	ߞߊߓߊ	kábá`	stone
/l/	< ߟ >	<l>	ߟߍ߭	lɛ̌`	pig
/m/	< ߡ >	<m>	ߡߍ߲߫	mɛ́n	to hear
/ɲ/	< ߢ >	<ɲ>	ߢߊ	ɲá`	eye
	< ߧ >	<y>	ߒ߫ ߧߋ߫ ߊ߬ ߝߍ߬	ń **y**é à fɛ̀	I want it; I like it
/n/	< ߣ >	<n>	ߣߊ߬	nà	to come
	< ߠ >		ߘߊ߲ߠߊ߫	dán-**n**á	fabricate + *imperfective suffix*
			from ߘߊ߲ + ߟߊ߫ (dán + **l**á)		
/h/	< ߤ >	<h>	ߤߊߟߌ߬	hálì	even
/w/	< ߥ >	<w>	ߥߊߙߊ	wárá`	lion; any wild feline
/y/ [j]	< ߦ >	<y>	ߦߌߙߌ	yírí`	tree

counts the number of graphemes considering only basic symbols (without diacritics), the value of U given by equation (1) would be underestimated, while the inclusion of diacritics leads to the overestimation of this quantity.

If the tones are taken into account, the system comprises 134 units, of which 109 units are with graphemic representation, 18 units with two graphemic representations, and seven units with three graphemic representations. Orthographic uncertainty yields $U = 0.22$, as calculated from equation (1) with N standing for the total number of units and f_k for the number of units with k graphemic representations. The best fits for the number of graphemic representations are given by several distributions, including the geometric distribution

$$P(n) = p(1-p)^n\,, \tag{2}$$

which is a special case of the Shenton-Skees-geometric distribution known as a model for the grapheme–phoneme relation in several languages. Table 6 represents the fitting results for for the number of graphemic representations in Nko.

Table 6: Fit for the number of graphemic representations

i	f_i	Geometric d.
1	109	106.6
2	18	21.8
3	7	5.6
	$\lambda = 0.2091$	$p = 0.7955$
	$C = 0.0001$	$C = 0.0082$

8 Final remarks

In this paper, some results on the quantitative behavior of the Nko script are presented. Complexities of the Nko characters are calculated and their correlation with frequencies is analyzed. The grapheme–phoneme correspondence was studied, yielding the mean orthographic uncertainty of the script $U = 0.38$ or $U = 0.22$ depending on the approach to the treatment of diacritics and tonal distinctions. Further tasks in this direction include: finding optimal fits for components and connections, analysis of grapheme and phoneme frequencies on larger text samples, study of grapheme/complexity correlation on larger text samples, etc. The applied way to handle tone distinctions must be verified in application to other scripts, including Roman orthographies for African languages. The proposed treatment of diacritics is not unique, and comparisons to other approaches are required.

Acknowledgments. The Nko typeface used in this paper is a TEX adaptation of the JG Nko typeface © Glavyfonts. A.R. appreciates discussions with Ján Mačutek on some details of the statistical analysis. Data fitting was done with Altmann-Fitter, Version 2.1.

References

Altmann, G.
2004 "Script complexity", in: *Glottometrics*, 8; 68–74.
Altmann, G.; Fan, F. (eds.)
2008 *Analysis of Script. Properties of Characters and Writing Systems.* Berlin: de Gruyter.
Buk, S.; Mačutek, J.; Rovenchak, A.
2008 "Some properties of the Ukrainian writing system", in: *Glottometrics*, 16; 63–79.
Coulmas, F.
2004 *The Blackwell encyclopedia of writing systems.* Blackwell Publishing.
Dalby, D.
1969 "Further indigenous scripts of West Africa: Manding, Wolof and Fula alphabets and Yoruba 'holy' writing", in: *African Language Studies*, 10; 161–181.
Grzybek, P.; Kelih, E.; Stadlober, E.
2006 "Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik", in: *Anzeiger für Slavische Philologie*, 34; 41–74.
Mačutek, J.
2008a "A generalization of the geometric distribution and its application in quantitative linguistics", in: *Romanian Reports in Physics*, 60; 501–509.
Mačutek, J.
2008b "Runes: complexity and distinctivity", in: *Glottometrics*, 16; 1–16.
Rovenchak, A.; Mačutek, J.; Riley, C.
2009 "Distribution of complexities in the Vai script", in: *Glottometrics*, 18; 1–12.
Smith, R.
2007 "Investigation of the Zipf-plot of the extinct Meroitic language", in: *Glottometrics*, 15; 53–61.
Vydrin, V.
1999 *Manding–English Dictionary (Maninka, Bamana): Vol. 1.* St. Petersburg: Dimitry Bulanin Publishing House.

Distribution of motifs in Japanese texts

Haruko Sanada

1 Introduction

In Japanese there are two possibilities to measure word length: in terms of syllables and in terms of morae. Morae are associated with writing, syllables are phonetic entities. Long vowels, double consonants, and nasal sounds which are counted as a mora are not considered for measuring length in terms of syllable numbers. We employ the pronunciation of the Tokyo dialect which is a model of standard Japanese. One of the characteristics of the Tokyo dialect is the devoicing of vowels. The devoicing of vowels often occurs with k(i), k(u), sh(i), s(u), ch(i), ts(u), h(i), f(u), p(i), and p(u):

1. at the end of the sentence, e.g. *dekimas'(u)* [possible], and
2. before unvoiced consonants (k, s, t, h, f, or p) in a word or in a phrase, e.g. *watak'(u)shi* [I, myself].

Some other cases can be found with /a/ or /o/ if the same unvoiced consonant is repeated, as in *k′(a)karu* [price] or *k′(o)koro* [heart], and those with /i/ or /u/ followed by voiced consonants, e.g *s′(u)gi* [tree of cedar]. However, devoiced vowels are not followed by devoiced vowels like in *f'(u)kusô* [dress].

2 Motifs in Japanese texts

Transcribing the text in one of these ways, we obtain word-length sequences which can be considered time-series or can be grouped into "motifs", introduced in linguistics by Köhler (cf. Köhler 2006, Köhler and Naumann 2008). Motifs are non-decreasing sequences of lengths. However, in Japanese which is a strongly postpositional language it is more appropriate to consider non-increasing word length sequences, e.g. 2-1-1, 3-2-2-1-1, etc. which better simulate both the word and the rhythm of the language.

Our aim is to show whether or not there are regularities or tendencies associated with motifs in such a highly agglutinating and postpositional language as Japanese. For the analysis we use the text *Jinseiron Note* [Essay on Life] (Miki 1941) which is written in modern Japanese without any spoken language. The text has a total of 1987 sentences and 45809 words, symbols, and punctuation marks. This essay was originally written by a philosopher Kiyoshi Miki in 1941, and it has several chapters on aspects of our life, e.g. death, doubt, customs, vanity, solitude, jealousy, success, hypocrisy, etc. The text is included in

Shinchôsha (1995) which contains 100 novels and essays as a magnetic compiliation, and this text is one of a few works written in modern Japanese without any spoken language in Shinchôsha (1995). In novels with spoken language, it is difficult to divide sentences into words and to classify words into parts of speech because contracted forms are often found.

3 How to count motifs

Since Japanese texts have no word boundaries, we used some software called *ChaSen* and *Unidic*, developed by the Nara Institute of Science and Technology and by the National Institute for Japanese Language and Linguistics, to partition the sentences into words. The software made errors almost 8% of the time, and these errors were corrected by hand. Finally we got 45809 words, symbols, and punctuation marks, including some compound words e.g. cases of a verb used as a prefix. Analyzing the text *Jinseiron Note* we wanted to study the following properties:

(1) the structural diversity of motifs (types),
(2) the rank-frequency distribution of motifs of motifs (tokens),
(3) the frequency spectrum of motifs (tokens),
(4) the relation between the length and the frequency of motifs.

In all cases we want to compare the standard language with the Tokyo dialect and the two ways of counting (syllable vs. mora). The motifs are not exploited uniformly. It can be shown that the longer the first element in the motif, the smaller the number of motifs formed in this way. In this contribution we shall present only problems (1), (2), and (4). The examination of motifs can be performed in three different ways:

1. The motif cannot extend past the given punctuation.
2. The motif cannot extend past the end of sentence.
3. All punctuations are ignored.

All types of counting were performed using the Tokyo dialect. In Table 1 we present the different kinds of segmentation. “B” indicates a motif boundary, “P” indicates a punctuation which functions as a motif boundary; a circumflex in the reading means a long vowel. Differences are denoted by D.

4 Results of the analysis

4.1 The number of types

The number of possible motifs with decreasing structure can be computed easily if we fix the length. Otherwise a non-increasing motif can be prolonged to

Table 1: Examples of motifs with three variants

Original text*	Reading	Part of speech†	Number of syllables‡	Type (i)	Type (ii)	Type (iii)
健康	kenkô	N	2			
感	kan	N	1			
は	wa	PP	1	B	B	B
自覚	jikaku	N	2			
的	teki	S	2			
で	de	AV	1	B	B	B
あり	ari	VA	2	B	D	D
、		PU	0	P	D	D
不	fu	PR	1	B	B	B
安定	antê	N	2			
で	de	AV	1	B	B	B
ある	aru	VA	2	B	B	D
。		PU	0	P	P	D
健康	kenkô	N	2			
と	to	PP	1			
いう	yû	V	1			
の	no	PP	1			
は	wa	PP	1	B	B	B
Obtained motifs				2-1-1, 2-2-1, 2, 1, 2-1, 2, 2-1-1-1-1	2-1-1, 2-2-1, 2-1, 2-1, 2, 2-1-1-1-1	2-1-1, 2-2-1, 2-1, 2-1, 2-2-1-1-1-1

* The original text means: “The conception of health is subjective, and it is not stable. Health is. . . "

† N = noun, PP = postposition, PR = prefix, PU = punctuation, S = suffix, V = verb, VA = verb (as an auxiliary verb), AV = auxiliary verb

‡ according to Tokyo dialect

the complete text length. However, the number of motif types actually used in a language differs from the theoretically possible one and this holds even more when the text is short. Nevertheless, it can be shown that in our text the number of motifs (types) beginning with a word of length x is distributed binomially as shown in Table 2, f_x being the number of different motifs.

Since the iterative fitting yielded $n = 7,9$ and 10, we can conclude that in Japanese there will be no motifs beginning with a word longer that 10 syllables. This length is extreme even for strongly agglutinating languages. The longest first word in our text had 8 syllables. As can be seen in Table 2, the fitting is adequate.

Table 2: Fitting the binomial distribution to motif types beginning with length x

Beginning number x	Type (i) f_x	Type (i) NP_x	Type (ii) f_x	Type (ii) NP_x	Type (iii) f_x	Type (iii) NP_x
1	7	8.36	7	9.85	2	6.00
2	33	29.42	34	32.84	38	26.17
3	46	46.01	50	49.27	52	48.91
4	43	41.98	45	43.80	48	50.78
5	20	24.62	24	25.56	23	31.64
6	9	9.63	9	10.22	9	11.83
7	4	2.51	5	2.84	5	2.46
8	1	0.46	1	0.61	1	0.22
Total	163		175		178	
	$p = 0.2811, n = 9$ $X^2_{DF=4} = 2.97$ $P = 0.56$		$p = 0.25, n = 10$ $X^2_{DF=4} = 3.03$ $P = 0.55$		$p = 0.3839, n = 7$ $X^2_{DF=2} = 4.63$ $P = 0.10$	

4.2 Distribution of motifs

Just as with other linguistic units, motifs have their rank-frequency distribution of motifs and a parallel frequency spectrum. Here we shall consider the rank-frequency as a simple ranked sequence, otherwise the number of pooled classes would distort the picture. We shall adhere to the proposal of Popescu et al. (2009: 14) who propose the theoretical rank-frequency sequence in the form

$$f(r) = 1 + a\exp(-br)\,. \tag{1}$$

The results of the fitting are presented in the appendix (see Table 4, pp. 190ff.); as can be seen, the fitting is very good, the determination coefficient being in all cases greater than $R^2 > 0.95$.[1]

4.3 The relationship between length and frequency of motifs

In all the tables, "length" means the number of elements in a motif. We assume that in all cases (i) to (iii) the same regularity holds but the parameters of the given functions differ according to the boundary conditions. Further, we assume that very short motifs are rather intermediary states between two long motifs or at the beginning of the sentence, hence the function expressing

1. For type (i), with parameter values for $a = 6922.14, b = 0.42$, we obtain a determination coefficient of $R^2 = 0.97$; for type (ii), $R^2 = 0.95$ ($a = 6362.71, b = 0.40$), and for type (iii), $R^2 = 0.96$ ($a = 5808.58, b = 0.41$).

this dependence will have its maximum not at $x = 1$ but at $x = 2$. We do not know whether this is a specificity of Japanese but in any case it differs from the usual monotonically decreasing length-frequency functions of other linguistic units. One can capture this fact either by considering a function consisting of two parts ($x = 1$ and $x = 2,3,4,\ldots$) or simply a function capturing both the increasing and the decreasing parts by a multiplication of two functions. For our purposes we used

$$y = ax^b \exp(-cx) \,, \tag{2}$$

which can be derived from the differential equations of Wimmer-Altmann's (2005) general theory or from Köhler's (1986, 2005) self-regulation cycles. The results of the three variants are presented in Table 3, y being the motif frequency, and $\hat{y}$ the predicted values according to (2).

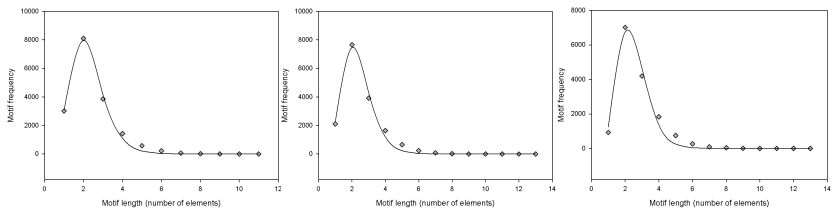
Table 3: The length-frequency relationship with three variants of motifs

Motive	Type (i)		Type (ii)		Type (iii)	
length x	y	$\hat{y}$	y	$\hat{y}$	y	$\hat{y}$
1	3017	3093.5421	2107	2253.3602	928	1255.7037
2	8102	7976.6739	7660	7472.7249	7014	6755.7732
3	3862	4150.6125	3919	4312.0173	4193	4646.6733
4	1434	1121.5643	1642	1216.0421	1840	1376.7816
5	582	211.6426	662	231.6555	757	257.1465
6	219	31.8237	238	34.4327	269	35.8702
7	66	4.091	85	4.3125	98	4.0908
8	20	0.4689	27	0.4765	40	0.4026
9	3	0.0492	5	0.0478	10	0.0354
10	2	0.0048	2	0.0045	1	0.0028
11	2	0.0004	2	0.0004	3	0.0002
12			1	0	1	0
13			1	0	1	0
Total	17309	$a = 56730.88$ $b = 5.56$ $c = 2.91$ $R^2 = 0.9942$	163351	$a = 45920.79$ $b = 6.08$ $c = 3.01$ $R^2 = 0.9898$	15155	$a = 33150.12$ $b = 7.15$ $c = 3.27$ $R^2 = 0.9832$

As can easily be seen from Figures 1a to 1c, the regularity is very rigorous.

5 Conclusions

Our problem does not concern only Japanese but has a general aspect: What is the situation in other postpositional languages? Is the situation parallel to



(a) Type (i) *(b)* Type (ii) *(c)* Type (iii)

Figure 1: Motif length and frequency

prepositional languages? Do motifs abide by general laws of length and distribution? Are they "legal" linguistic units or only very high abstractions? In order to answer these questions a number of examinations in other languages will be necessary. Nevertheless, our results show that motifs – just as any other linguistic units – are "correct" conceptual abstractions abiding by the same laws as all other linguistic units. The specificities of motifs, e.g. the parameters in the laws, must be scrutinized step by step using different texts in different languages.

References

Köhler, R.
1986 *Zur sprachlichen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

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

Köhler, R.
2006 "The frequency distribution of the lengths of length sequences." In: Genzor J.; Bucková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa.* Bratislava: Slovak Academic Press, 145–152.

Köhler, R.; Naumann, S.
2008 "Quantitative text analysis using *L*-, *F*- and *T*-segments." In: Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R. (eds.), *Data Analysis, Machine Learning and Applications.* Berlin, Heidelberg: Springer, 637–646.

Popescu, I.-I.; Mačutek, J.; Altmann, G.
2009 *Aspects of word frequencies.* Lüdenscheid: RAM-Verlag.

Wimmer, G.; Altmann, G.
2005 "Unified derivation of some linguistic laws." In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin–New York: de Gruyter, 791–807.

Text Material

Miki, K.
1941 *Jinseiron Note* [= Essay on Life]. Tokyo: Sôgensha.

Shinchôsha
1995 *Shinchô Bunko no 100 satsu.* [CD-ROM edition of 100 paperbacks extracted from the Shinchô Bunko series). Tokyo: Shinchôsha.

Software

ChaSen: The Nara Institute of Science and Technology.
`http://en.sourceforge.jp/projects/chasen-legacy/`

Unidic: The National Institute for Japanese Language and Linguistics [In Japanese].
`http://download.unidic.org/`

Table 4: The rank-frequency distribution of motif tokens of motifs

Rank	Type (i)			Type (ii)			Type (iii)		
	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$
1	2-1	4917	4546.62	2-1	4714	4238.88	2-1	4218	3840.10
2	2	2313	2986.01	3-1	1939	2823.64	3-1	1928	2538.40
3	3-1	1970	1961.19	2	1912	1881.02	2-2-1	1578	1678.06
4	2-2-1	1269	1288.22	2-2-1	1280	1253.19	2-1-1	1084	1109.43
5	2-1-1	1102	846.29	2-1-1	1139	835.02	2	858	733.60
6	3-2-1	460	556.08	2-1-1-1	483	556.50	3-2-1	502	485.20
7	4-1	430	365.51	3-2-1	474	370.99	2-1-1-1	470	321.02
8	2-1-1-1	429	240.37	4-1	425	247.43	2-2-2-1	430	212.52
9	3-1-1	359	158.19	3-1-1	390	165.14	4-1	425	140.80
10	1	350	104.22	2-2-2-1	323	110.32	3-1-1	390	93.40
11	2-2	318	68.78	2-2	278	73.81	2-2-1-1	284	62.07
12	3	293	45.51	2-2-1-1	207	49.50	2-2	242	41.36
13	2-2-2-1	270	30.23	2-1-1-1-1	193	33.30	4-2-1	197	27.68
14	1-1	211	20.19	4-2-1	192	22.51	2-1-1-1-1	195	18.63
15	4-2-1	186	13.60	3-1-1-1	171	15.33	3-1-1-1	168	12.65
16	2-1-1-1-1	184	9.28	3	151	10.54	2-2-1-1-1	154	8.70
17	2-2-1-1	180	6.44	3-2-2-1	131	7.36	3-2-2-1	146	6.09
18	3-1-1-1	158	4.57	4-1-1	124	5.23	4-1-1	124	4.36
19	3-2-2-1	119	3.34	2-2-1-1-1	118	3.82	3-3-1	110	3.22
20	4-1-1	119	2.54	3-2	99	2.88	3-2-1-1	100	2.47
21	3-2	114	2.01	3-3-1	99	2.25	3-2	89	1.97
22	2-2-1-1-1	107	1.66	3-2-1-1	93	1.83	3-1-1-1-1	76	1.64
23	1-1-1	97	1.44	1-1	81	1.56	2-2-2-2-1	74	1.42
24	3-3-1	91	1.29	3-1-1-1-1	76	1.37	3	67	1.28
25	3-2-1-1	79	1.19	2-1-1-1-1-1	64	1.25	2-1-1-1-1-1	63	1.19
26	3-1-1-1-1	65	1.12	5-1	58	1.16	2-2-2-1-1	59	1.12
27	2-1-1-1-1-1	64	1.08	4-2-2-1	57	1.11	4-2-2-1	58	1.08
28	5-1	59	1.05	4-3-1	49	1.07	5-1	58	1.05
29	4-2-2-1	48	1.04	3-2-1-1-1	45	1.05	2-2-2	52	1.04
30	4-3-1	47	1.02	2-2-1-1-1-1	42	1.03	4-3-1	52	1.02
31	4	47	1.02	2-2-2-1-1	42	1.02	2-2-1-1-1-1	50	1.02
32	2-2-1-1-1-1	41	1.01	2-2-2-2-1	41	1.01	3-2-1-1-1	47	1.01
33	3-2-1-1-1	40	1.01	1-1-1	40	1.01	2-2-2-1-1-1	32	1.01
34	2-2-2	35	1	2-2-2	33	1.01	3-3-2-1	32	1
35	2-2-2-1-1	33	1	4-1-1-1	30	1	4-1-1-1	31	1
36	2-2-2-2-1	31	1	1	26	1	3-2-2	28	1
37	3-2-2	27	1	3-3-2-1	26	1	3-1-1-1-1-1	25	1
38	4-1-1-1	26	1	3-1-1-1-1-1	25	1	3-2-2-2-1	24	1
39	4-2	25	1	2-2-2-1-1-1	23	1	3-3-1-1	24	1
40	3-1-1-1-1-1	23	1	3-2-2	23	1	3-2-2-1-1	23	1
41	3-3-2-1	22	1	3-2-2-2-1	22	1	4-2-1-1	22	1
42	2-2-2-1-1-1	21	1	3-3-1-1	21	1	4-1-1-1-1	20	1

(continued on next page)

Table 4 (continued from previous page)

Rank	Type (i)			Type (ii)			Type (iii)		
	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$
43	4-2-1-1	21	1	4-2-1-1	21	1	4-2	18	1
44	3-3	18	1	3-2-2-1-1	20	1	5-2-1	18	1
45	3-2-2-2-1	17	1	4-1-1-1-1	20	1	2-2-2-2	16	1
46	4-1-1-1-1	17	1	4-2	18	1	2-2-1-1-1-1-1	15	1
47	3-2-2-1-1	16	1	5-2-1	17	1	4-2-1-1-1	15	1
48	3-3-1-1	16	1	3-3	16	1	2-2-2-2-2-1	14	1
49	5-2	16	1	4	16	1	2-2-2-2-1-1	13	1
50	4-2-1-1-1	15	1	4-2-1-1-1	15	1	3-1-1-1-1-1-1	13	1
51	1-1-1-1	14	1	1-1-1-1	14	1	5-2	13	1
52	5-2-1	12	1	2-1-1-1-1-1-1	13	1	2-2-2-1-1-1-1	12	1
53	5-1-1	11	1	5-2	13	1	5-1-1	12	1
54	5	11	1	2-2-2-1-1-1-1	12	1	2-1-1-1-1-1-1	11	1
55	2-2-1-1-1-1-1	10	1	3-1-1-1-1-1-1	12	1	3-2-1-1-1-1	11	1
56	6-1	10	1	5-1-1	12	1	6-1	11	1
57	2-1-1-1-1-1-1	9	1	2-2-2-2-2-1	11	1	4-1-1-1-1-1	10	1
58	3-1-1-1-1-1-1	9	1	6-1	11	1	3-3-1-1-1	9	1
59	3-2-1-1-1-1	9	1	2-2-1-1-1-1-1	10	1	4-2-2-1-1	9	1
60	4-1-1-1-1-1	9	1	2-2-2-2	10	1	4-3-2-1	9	1
61	4-2-2	9	1	3-2-1-1-1-1	10	1	2-1-1-1-1-1-1-1	8	1
62	4-3	9	1	4-1-1-1-1-1	10	1	3-2-2-1-1-1	8	1
63	4-4-1	9	1	2-1-1-1-1-1-1-1	9	1	4-2-2-2-1	8	1
64	1-1-1-1-1	8	1	3-2-2-1-1-1	8	1	4-4-1	8	1
65	2-2-2-2	8	1	4-2-2-1-1	8	1	2-2-1-1-1-1-1-1	7	1
66	2-2-2-1-1-1-1	7	1	4-2-2-2-1	8	1	2-2-2-2-2-1-1	7	1
67	4-2-2-1-1-1	7	1	4-2-2	8	1	3-2-2-2-2-1	7	1
68	3-2-2-1-1-1	6	1	4-3-2-1	8	1	3-3-3-1	7	1
69	4-2-2-1-1	6	1	4-4-1	8	1	3-3	7	1
70	4-2-2-2-1	6	1	1-1-1-1-1	7	1	4-2-2-1-1-1	7	1
71	4-3-1-1-1	6	1	3-3-1-1-1	7	1	4-2-2	7	1
72	1-1-1-1-1-1	5	1	4-2-2-1-1-1	7	1	4-3-1-1-1	7	1
73	2-2-1-1-1-1-1-1	5	1	2-2-2-2-2-1-1	6	1	2-2-2-1-1-1-1-1	6	1
74	2-2-2-2-2-1-1	5	1	3-2-1-1-1-1-1	6	1	3-2-1-1-1-1-1	6	1
75	2-2-2-2-2-1	5	1	3-2-2-2-2-1	6	1	4-2-1-1-1-1	6	1
76	3-2-1-1-1-1-1	5	1	4-2-1-1-1-1	6	1	5-2-2-1	6	1
77	3-3-2-1-1	5	1	4-3-1-1-1	6	1	2-2-1-1-1-1-1-1-1	5	1
78	4-2-1-1-1-1	5	1	4-3	6	1	2-2-2-2-1-1-1-1	5	1
79	4-3-2-1	5	1	5-2-2-1	6	1	2-2-2-2-1-1-1	5	1
80	4-3-3-1	5	1	2-2-1-1-1-1-1-1	5	1	3-2-2-1-1-1-1	5	1
81	5-2-2-1	5	1	2-2-2-2-1-1	5	1	3-3-1-1-1-1	5	1
82	6-2-1	5	1	3-3-2-1-1	5	1	3-3-2-1-1	5	1
83	2-1-1-1-1-1-1-1	4	1	3-3-3-1	5	1	3-3-2-2-1	5	1
84	2-2-2-1-1-1-1-1	4	1	4-3-1-1	5	1	3-3-3-2-1	5	1
85	2-2-2-2-1-1	4	1	4-3-2	5	1	4-3-1-1	5	1
86	3-2-2-1-1-1-1	4	1	4-3-3-1	5	1	4-3-2	5	1
87	3-3-1-1-1-1	4	1	6-2-1	5	1	4-3-3-1	5	1
88	3-3-1-1-1	4	1	2-2-2-1-1-1-1-1	4	1	6-2-1	5	1
89	3-3-2-1-1-1	4	1	3-2-2-1-1-1-1	4	1	2-2-2-2-2-2-1	4	1

(continued on next page)

Table 4 (continued from previous page)

Rank	Type (i)			Type (ii)			Type (iii)		
	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$
90	3-3-3-1	4	1	3-3-1-1-1-1	4	1	3-3-2-1-1-1	4	1
91	4-1-1-1-1-1-1	4	1	3-3-2-1-1-1	4	1	4-1-1-1-1-1-1	4	1
92	4-3-1-1	4	1	3-3-2-2-1	4	1	5-3-1	4	1
93	4-3-2	4	1	3-3-3-2-1	4	1	6-1-1-1	4	1
94	5-3-1	4	1	4-1-1-1-1-1-1	4	1	7-2-1	4	1
95	6-1-1-1	4	1	6-1-1-1	4	1	2-2-2-2-1-1-1-1-1	3	1
96	7-2-1	4	1	7-2-1	4	1	3-2-2-1-1-1-1-1	3	1
97	3-2-2-2-2-1	3	1	2-2-1-1-1-1-1-1-1	3	1	3-3-1-1-1-1-1	3	1
98	3-3-2-2-1	3	1	2-2-2-2-1-1-1	3	1	4-2-2-2-2-1	3	1
99	3-3-3-2-1	3	1	3-2-2-2	3	1	4-3-2-1-1	3	1
100	4-3-2-1-1	3	1	3-3-2	3	1	4-4-1-1-1	3	1
101	5-1-1-1	3	1	4-2-2-2-2-1	3	1	4	3	1
102	5-2-1-1	3	1	4-3-2-1-1	3	1	5-1-1-1	3	1
103	6-1-1	3	1	4-4-1-1-1	3	1	6-1-1	3	1
104	6	3	1	5-1-1-1	3	1	2-1-1-1-1-1-1-1-1-1-1	2	1
105	7-1	3	1	5-3-1	3	1	3-1-1-1-1-1-1-1	2	1
106	1-1-1-1-1-1-1	2	1	6-1-1	3	1	3-2-2-2-1-1-1-1	2	1
107	2-1-1-1-1-1-1-1-1-1-1	2	1	2-1-1-1-1-1-1-1-1-1-1	2	1	3-2-2-2-2-1-1	2	1
108	2-2-1-1-1-1-1-1-1	2	1	2-2-2-2-2-2-1	2	1	3-2-2-2	2	1
109	2-2-2-2-1-1-1	2	1	3-2-2-1-1-1-1-1	2	1	3-3-2-2-1-1-1-1	2	1
110	3-2-2-1-1-1-1-1	2	1	3-3-1-1-1-1-1	2	1	3-3-2	2	1
111	3-2-2-2	2	1	3-3-2-2-1-1-1-1	2	1	4-3-1-1-1-1	2	1
112	3-3-1-1-1-1-1	2	1	4-3-1-1-1-1	2	1	4-3-2-1-1-1-1	2	1
113	3-3-2	2	1	4-3-2-1-1-1-1	2	1	4-3-2-2-1	2	1
114	4-3-1-1-1-1	2	1	4-4-2-1	2	1	4-3	2	1
115	4-3-2-1-1-1-1	2	1	5-2-1-1-1	2	1	4-4-2-1	2	1
116	4-4-1-1-1	2	1	5-2-1-1	2	1	5-2-1-1-1	2	1
117	5-3-1-1	2	1	5-2-2	2	1	5-2-1-1	2	1
118	7-1-1	2	1	5-3-1-1	2	1	5-2-2	2	1
119	2-1-1-1-1-1-1-1-1-1	1	1	5	2	1	5-3-1-1	2	1
120	2-2-2-1-1-1-1-1-1-1	1	1	6-2-1-1-1	2	1	6-2-1-1-1	2	1
121	2-2-2-2-2-2-1	1	1	7-1-1	2	1	7-1-1	2	1
122	2-2-2-2-2	1	1	7-1	2	1	7-1	2	1
123	3-1-1-1-1-1-1-1	1	1	1-1-1-1-1-1-1-1	1	1	1-1-1	1	1
124	3-2-2-1-1-1-1-1-1	1	1	1-1-1-1-1-1-1	1	1	1-1	1	1
125	3-2-2-2-1-1-1-1	1	1	2-1-1-1-1-1-1-1-1-1	1	1	2-1-1-1-1-1-1-1-1-1	1	1
126	3-2-2-2-2-1-1	1	1	2-2-2-1-1-1-1-1-1-1	1	1	2-1-1-1-1-1-1-1-1	1	1
127	3-2-2-2-2	1	1	2-2-2-2-1-1-1-1-1	1	1	2-2-2-2-1-1-1-1-1-1-1	1	1
128	3-3-2-2-1-1-1	1	1	2-2-2-2-2	1	1	2-2-2-2-2-1-1-1	1	1
129	3-3-2-2-1-1	1	1	3-1-1-1-1-1-1-1	1	1	2-2-2-2-2-2-1-1	1	1
130	3-3-2-2-2-1-1-1	1	1	3-2-1-1-1-1-1-1-1-1-1-1-1	1	1	2-2-2-2-2	1	1
131	3-3-3-2-2-2-2-1	1	1	3-2-2-1-1-1-1-1-1-1-1-1	1	1	3-2-1-1-1-1-1-1-1-1-1-1-1	1	1
132	3-3-3-2	1	1	3-2-2-2-1-1-1-1	1	1	3-2-2-1-1-1-1-1-1-1-1-1	1	1
133	3-3-3	1	1	3-2-2-2-2-2-1-1	1	1	3-2-2-2-2-1-1	1	1
134	4-2-2-1-1-1-1-1	1	1	3-2-2-2-2-2	1	1	3-2-2-2-2-2	1	1
135	4-2-2-2-1-1-1	1	1	3-3-1-1-1-1-1-1-1	1	1	3-3-1-1-1-1-1-1-1	1	1
136	4-2-2-2-1-1	1	1	3-3-2-2-1-1-1	1	1	3-3-2-1-1-1-1	1	1

(continued on next page)

Table 4 (continued from previous page)

Rank	Type (i)			Type (ii)			Type (iii)		
	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$	y	f_y	$\hat{y}$
137	4-2-2-2-2-1	1	1	3-3-2-2-2-1-1-1	1	1	3-3-2-2-1-1-1	1	1
138	4-3-2-2-1-1-1	1	1	3-3-3-1-1	1	1	3-3-2-2-2-1-1-1	1	1
139	4-3-2-2-1	1	1	3-3-3-2-1-1-1	1	1	3-3-3-1-1	1	1
140	4-4-1-1	1	1	3-3-3-2-2-1	1	1	3-3-3-2-1-1-1	1	1
141	4-4-2-1-1-1	1	1	3-3-3-2-2-2-2-1	1	1	3-3-3-2-2-1	1	1
142	4-4-2-1	1	1	3-3-3-2	1	1	3-3-3-2-2-2-2-1	1	1
143	4-4-2-2-1	1	1	4-2-2-1-1-1-1-1	1	1	3-3-3-2	1	1
144	4-4-3-2-1	1	1	4-2-2-2-1-1-1	1	1	4-2-2-1-1-1-1-1	1	1
145	4-4-3-3-1	1	1	4-2-2-2-1-1	1	1	4-2-2-2-1-1-1	1	1
146	4-4-3	1	1	4-2-2-2-2-1-1	1	1	4-2-2-2-1-1	1	1
147	4-4	1	1	4-2-2-2-2-2	1	1	4-2-2-2-2-1-1	1	1
148	5-1-1-1-1	1	1	4-3-2-2-1-1-1	1	1	4-2-2-2-2-2	1	1
149	5-2-1-1-1-1	1	1	4-3-2-2-1	1	1	4-2-2-2	1	1
150	5-2-1-1-1	1	1	4-4-1-1	1	1	4-3-2-1-1-1	1	1
151	5-2-2-2-1-1	1	1	4-4-2-1-1-1	1	1	4-3-2-2-1-1-1	1	1
152	5-2-2-2-2-1	1	1	4-4-2-2-1	1	1	4-3-2-2	1	1
153	5-2-2	1	1	4-4-3-1	1	1	4-4-1-1	1	1
154	5-3-2-2-1	1	1	4-4-3-2-1-1-1	1	1	4-4-2-1-1-1	1	1
155	5-3-2	1	1	4-4-3-3-1	1	1	4-4-2-2-1	1	1
156	5-3	1	1	4-4-3	1	1	4-4-3-1-1-1-1	1	1
157	5-4-2	1	1	5-1-1-1-1	1	1	4-4-3-1	1	1
158	6-2-1-1-1	1	1	5-2-1-1-1-1	1	1	4-4-3-2-1-1-1	1	1
159	6-2-1-1	1	1	5-2-2-1-1-1-1	1	1	4-4-3-3-1	1	1
160	6-3-1-1	1	1	5-2-2-2-1-1	1	1	5-1-1-1-1	1	1
161	6-3-2-1	1	1	5-2-2-2-1	1	1	5-2-1-1-1-1	1	1
162	7-2-1-1-1	1	1	5-2-2-2-2-1	1	1	5-2-2-1-1-1-1	1	1
163	8-1-1-1	1	1	5-3-2-1	1	1	5-2-2-2-1-1	1	1
164				5-3-2-2-1	1	1	5-2-2-2-1	1	1
165				5-3-2	1	1	5-2-2-2-2-1	1	1
166				5-4-2	1	1	5-3-2-1	1	1
167				5-5-2	1	1	5-3-2-2-1	1	1
168				5-5-3-1	1	1	5-3-2	1	1
169				6-2-1-1	1	1	5-4-2	1	1
170				6-3-1-1	1	1	5-5-2	1	1
171				6-3-2-1	1	1	5-5-3-1	1	1
172				6-4-1-1-1	1	1	6-2-1-1	1	1
173				7-1-1-1	1	1	6-3-1-1	1	1
174				7-2-1-1-1	1	1	6-3-2-1	1	1
175				8-1-1-1	1	1	6-4-1-1-1	1	1
176							7-1-1-1	1	1
177							7-2-1-1-1	1	1
178							8-1-1-1	1	1
	Total	17309		Total	16351		Total	15155	

Quantitative data processing in the ORD speech corpus of Russian everyday communication

Tatiana Sherstinova

1 Introduction: the ORD corpus

The main aim of creating the ORD corpus is to collect recordings of normal speech which is used in everyday communication and made in natural conditions. For this purpose subjects spent one day with dictaphones that hang around their necks and record all their communication. The abbreviation *ORD* stems from the Russian *Odin Rečevoj Den'*, literally translated as "one day of speech" (Asinovskij et al. 2009: 251f.).

At present the ORD corpus contains recordings made by a demographically balanced group of 40 subjects (20 men and 20 women) representing various social and age strata of the St. Petersburg population: students, military students, engineers, managers, scientists, doctors; also an IT technologist, seller, builder, psychologist, photographer, baby-sitter, drawing teacher, etc. (for more details see Table 2). The subjects' ages range from 16 to 70 years. Though the recordings were made under conditions of full anonymity, all subjects filled in sociological questionnaires, passed psychological testing, and kept diaries of their "day of speech", noting basic conditions of communication. Beside subjects' speeches, their 600 interlocutors were also recorded. Interlocutors were people of different ages (from 3 to 68 years) and occupations that were in friendly, family, professional or other relations with the subjects.

The recorded material contains diverse genres and styles of speech: conversations at home with relatives (at breakfast, at dinner, in the evening, at leisure, at home parties, etc.), professional and informal conversations with colleagues, communication during studies (lectures, practical lessons, informal students' conversations), communications with friends in different places and under different circumstances, e.g. visiting doctors, shopping, child-rearing, holidays events, entertainment and leisure-time, all kind of telephone talks, etc. Recordings were made in a variety of places: at home, in the office, while traveling by public transportation, walking outside, at universities, in military college, coffee shops, bars, restaurants, shopping centers, amusement parks, etc. (Asinovskij et al. 2009: 253).

As a result more than 320 hours of recording were obtained, from which 268 hours contained speech data quite suitable for further linguistic analysis and more than 90 hours of recordings were good enough for further phonetic

analysis. All recordings were audited by experts, the fragments without speech longer than several minutes were cut from ORD files. The recordings were split into files according to communicative episodes (Sherstinova 2009: 259). Currently the corpus contains 994 communicative episodes.

The corpus is being annotated on multiple linguistic and paralinguistic levels: main communication episodes, mini-episodes (within larger ones), orthographic transcripts of phrases, non-language audio events, speakers, voice quality, different comments. ORD annotation principles are described in (Sherstinova et al. 2009). By December 2009, more than 33 hours of recording (125 communicative episodes) have been annotated on the main eight levels. In addition, segmentation into words has been selectively made as well as segmentation and annotation of some affixes.

Figure 1 presents a fragment of annotation of one phrase on eight levels: *Frase* (transcript of speech), *Speaker* (Speaker's code), *Words*, *Morphems* (real phonetic transcription of morphemes in IPA), *Morphems-gram* (grammatical type of morphemes), *Morphems-orth* (spelling of morphemes), *Voice* (quality of speaker's voice), *FraseComment* (general comment about the phrase). This fragment is taken from communication episode #35–20 (at home with a cat) and mini-episode "The subject is coming home from work and receives an enthusiastic welcome from his cat. Nobody else at home". Speaker's code is "И35" (male, 70 years of age, engineering inspector, higher education, Russian, born in Buryatia, lives in St. Petersburg since 1970). Code *Л on Voice level means that the phrase translated as «*He is meeting daddy, that's an attaboy!*» is spoken in a tender mode. FraseComment explains that the phrase is addressed to a cat.

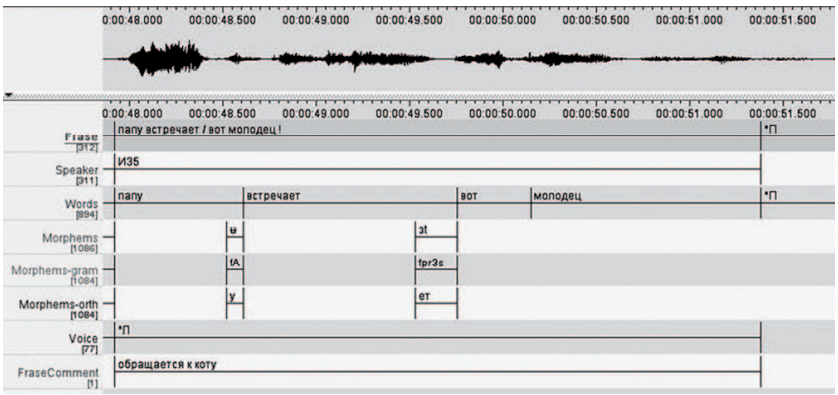


Figure 1: A fragment of annotation for one phrase in the ORD corpus

The corpus presents unique linguistic material, allowing us to perform fundamental research in many aspects: study of real spontaneous speech, phonetics and grammar of spoken language, psycholinguistics, communication studies, etc. Special interest is invoked by the possibility to compare the real speech (spoken texts) of a person in different communicative situations. At the same time these absolutely natural recordings may be used for practical purposes: for example, for verification of many scientific hypotheses, for adjustment and improvement of speech synthesis and recognition systems, etc.

The creation of the ORD corpus requires quantitative data processing on different levels of linguistic and paralinguistic annotations: frequency lists of words in different communication situations, frequencies of syntactic structures, grammatical forms, morphemes, real phonetic transcription of words and morphemes, prosody models, rhythm patterns, etc. Quantitative data processing is based mainly on statistical methods.

2 ORD processing tools

The ORD annotation is being made with the use of two professional annotation programmes: ELAN (EUDICO Linguistic Annotator) developed at the Max Planck Institute for Psycholinguistics (Nijmegen, NL) – see Hellwig et al. (2009) – and PRAAT created by Paul Boersma and David Weenink (the Institute of Phonetics Sciences, University of Amsterdam, The Netherlands) – see Boersma and Weenink (2009). Most of the annotation levels are made in ELAN, whereas phonetic transcripts are performed in PRAAT. For quantitative data processing the following standard software tools are used:

1. ELAN procedures for processing of linguistic annotations (Hellwig et al. 2009).
2. VISUAL BASIC utilities and macros created in MS ACCESS for the processing of tabular data and linguistic annotations converted into *.mdb format.
3. STATISTICA 8 for statistical processing of all data types.

Besides, a number of corpus-oriented software packages were created for ORD processing:

1. The E-CAR programme that allows us to perform various lexical and morphological analysis (in particular, to built frequency lists and concordances, carry out full-fledged lexical analysis, perform automatic extraction of morphemes, classify morphemes, etc.).
2. The E-LEX software was specially created for compiling digital dictionaries. It is used for compilation of a Russian everyday speech dictionary, and for integration and comparison with academic dictionaries of the Russian language.

3. Specialized programs and utilities were created for data conversion optimization between processing software and for specialized processing tasks not handled by the standard tools, in particular, utilities for fragmentary playback of recordings in different applications, the programs for calculating various text data indices, utilities for determining most frequent valency of lexical units, etc.

3 General statistics of transcripts

Orthographic transcripts of phrases, which are the main units of description, are kept on the **Frase** level of annotation files. Currently transcripts have been made for 33 hours of speech. Annotation on the Frase level contains 244075 "annotation words", from which 205009 are proper linguistic words or discourse particles. Table 1 shows the distribution of transcribed communication episodes grouped by social role of interlocutors. From Table 1 one may see that a quarter of all transcripts refer to communication between friends, about 20% of transcripts describe conversations between relatives, and the same amount of transcripts were obtained for communication between colleagues.

Table 1: Duration of transcribed communication episodes grouped by the social role of the interlocutors

N	Interlocutors of communication episodes	Duration (min)	Duration (hours)	%
1	Friends	524	8.73	25.79
2	Relatives	424	7.07	20.87
3	Colleagues	415	6.92	20.42
4	Neighbors or acquaintances	156	2.60	7.68
5	Service staff/Clients	124	2.07	6.10
6	Doctors/Patients	98	1.63	4.82
7	Teachers/Students	97	1.62	4.77
8	Oneself	89	1.48	4.38
9	Classmates	74	1.23	3.64
10	Strangers	19	0.32	0.94
11	Owner/Animal	12	0.20	0.59
$\sum$		2032	33.87	100.00

Table 2 contains information about the amount of transcribed speech obtained for each subject (and her/his interlocutors).

Table 2: The ORD subjects and transcripts in numbers (G = Gender; A = Age)

Speaker (Code)	G	A	Occupation	Episodes	Duration (min)	Duration (hours)	Annotation words (f)
S01	F	33	Baby-sitter	5	93	1.55	11206
S02	M	32	Engineer	1	23	0.38	2234
S03	F	33	Market analyst	4	38	0.63	5675
S04	F	34	Linguist (Ph.D.)	5	125	2.08	20892
S05	F	27	Psychologist (Ph.D.)	5	47	0.78	6173
S06	F	40	Housewife, nurse	6	57	0.95	5242
S07	M	45	Warrant officer	4	97	1.62	10962
S08	F	16	Schoolgirl	3	64	1.07	7290
S09	F	27	Structural designer	2	31	0.52	4865
S10	M	28	Engineer	2	36	0.60	3993
S11	F	28	Guide	4	65	1.08	6636
S12	F	26	Purchase manager	2	57	0.95	6526
S13	F	22	Secretary	3	55	0.92	8126
S14	F	33	Lecturer (phil., Ph.D.)	2	6	0.10	294
S15	M	20	Military student	2	23	0.38	2440
S16	M	22	Military student	1	13	0.22	1506
S17	M	17	Military student	3	30	0.50	2827
S18	F	19	Student	1	23	0.38	3172
S19	F	41	Market analyst	3	75	1.25	8858
S20	F	23	Economist	2	31	0.52	4254
S21	M	27	Business manager	6	56	0.93	5938
S22	F	35	Spanish teacher	3	64	1.07	7905
S23	F	23	Shop assistant	1	10	0.17	1502
S24	F	63	Meteorologist. D.S.	7	120	2.00	9405
S25	M	35	Chemist (techn.)	1	29	0.48	4303
S26	M	44	IT-specialist	2	18	0.30	1929
S27	F	20	Student	4	66	1.10	7767
S28	M	19	Student	4	67	1.12	7592
S29	M	22	Archaeologist	1	31	0.52	2594
S30	F	20	Student	2	62	1.03	6821
S35	M	70	Engineering inspector	8	114	1.90	11512
S36	M	40	Builder	6	74	1.23	10489
S37	F	59	Painting teacher	4	37	0.62	5357
S38	M	58	Businessman	1	27	0.45	4280
S39	M	53	Photographer	5	73	1.22	12671
S40	M	40	Pediatrician	2	41	0.68	3858
S41	M	63	Engineer	1	29	0.48	3050
S42	M	56	Project consultant. Ph.D.	4	78	1.30	7322
S43	M	60	University prof. (Ph.D.)	1	24	0.40	1655
S44	M	41	Pediatrician	2	23	0.38	4954
∑				125	2032	33.87	244075

Thus, transcripts for male subjects cover 15.10 hours of speech and contain 106109 annotation words, whereas for 20 female subjects 18.77 hours were transcribed (137966 annotation words). Though initially it was planned to transcribe the same amount of speech material for each subject, when the annotation had started it turned out that the recordings of some subjects are "better" than that of others (e.g., less background noise, more types of communicative episodes, interesting and clearly audible interlocutors). Because of that it was decided to transcribe first the most interesting episodes of higher quality. Annotating of the corpus is still in progress.

Besides transcripts the level Frase contains references to pauses (*П) (cf. Figure 1). Neither periods nor commas were used for transcribing speech; phonetic symbols '//' and '/' were used instead. The division of the real speech stream into fragments – sentences or syntagmas – turned out to be a serious problem, which does not have a unique solution in many cases; for more details see (Ryko and Stepanova 2008). Therefore, though we have the exact numbers of how often the symbols marking the end of the phrase ('//', '/', '?' and '!') were used in ORD transcripts, we should consider these data to be approximate (cf. Table 3).

Table 3: Frequency list of sentences and syntagmas

Types of speech fragment	Frequency	%
Sentences (all)	34213	100.00
Declarative sentences	23504	68.70
Interrogative sentences	6627	19.37
Exclamatory or imperative sentences	2080	6.08
Unfinished (interrupted) phrases	2002	5.85
Syntagmas (all)	65052	100.00
Syntagmas (not-finite)	32841	50.48

The average length of sentences in the ORD corpus is six words (5.99) or 1.9 syntagmas, and the average length of syntagmas is 3.15 words.

Speech is written in standard orthography (phonetic declinations are to be noted on another level). Besides symbols of sentence/syntagma division, transcripts may contain other symbols (see Table 4). The percentage given in Table 4 is calculated with respect to the total number of annotation words (244075).

Table 4: Frequency list of auxiliary annotation symbols

Symbol	Meaning	Frequency	%
*П	pause	24816	10.17
*Н	fragment of unintelligible speech	5869	2.40
*С	laughter	1204	0.49
*В	sigh or audible breath	1517	0.62
*К	cough	128	0.05
*Ш	noise	1275	0.52
()	short hesitation pause	1504	0.62
(. . .)	long hesitation pause	2465	1*l*.01
(м-м), (э-э), (а-а), (а-м), etc.	fillers (hesitation pause filled by different sounds)	1839	0.75
по. . . , канна. . . , etc.	interrupted words	3010	1.23
(?)	questionable or ambiguous transcript	910	0.37
#	change of a speaker in overlapping speech	3759	1.54
@	remark inserted by another speaker in overlapping speech fragments	3485	1.43

4 Frequency lists for speech transcripts

Table 5 presents the top of the frequency list for the transcribed part of the corpus compiled on the base of 205005 "linguistic" words and discourse particles. These data are not lemmatized, as POS-tagging of the ORD corpus is still not finished. Moreover, it turned out that in many cases the border between lexical and discourse meaning of the word is rather vague. For example, the word form *говорю* (#68 in the frequency list) in the utterance "*Я ему говорю, что. . .*" [I told him that . . .] has pure lexical meaning, whereas in "*Говорю тебе, что. . .*" [I am telling you that. . .] it has more discourse meaning: "I know what I am speaking about", "I insist on my opinion".

Many words in this top list are polyfunctional: depending on context they may act either as a "normal" lexeme with direct semantic meaning or as a discourse particle, having rather a pragmatic function or being just a filler. In Table 5 these words are denoted by an asterisk (*) together with discourse particles. These words deserve special attention and further investigation.

The highest rank in the frequency list is the personal pronoun *я* (*I*) taking 2.63% of all tokens. For oral speech this kind of result was expected. The three top ranking items in the frequency list of the spoken component of the British

Table 5: Most frequent Russian word forms in the ORD corpus

R	Word	N	%	R	Word	N	%	R	Word	N	%
1	я	5398	2.63	49	к	484	0.24	97	вам	264	0.13
2	не	4924	2.40	50	когда	481	0.23	98	сегодня	253	0.12
3	вот*	4800	2.34	51	будет	480	0.23	99	тогда	251	0.12
4	ну*	4727	2.31	52	как бы*	479	0.23	100	о	247	0.12
5	да*	4370	2.13	53	очень	478	0.23	101	от	243	0.12
6	а*	4187	2.04	54	тебе	464	0.23	102	думаю*	241	0.12
7	и	3637	1.77	55	его	463	0.23	103.5	быть	237	0.12
8	что	3625	1.77	56	ой*	448	0.22	103.5	из	237	0.12
9	в	3516	1.72	57	тут*	447	0.22	105	как-то	231	0.11
10	это*	3365	1.64	58	может*	442	0.22	106	пока*	228	0.11
11	там*	2931	1.43	59	значит*	438	0.21	107	них	227	0.11
12	у	2684	1.31	60	такой*	434	0.21	108.5	ж*	226	0.11
13	так*	2510	1.22	61	или	428	0.21	108.5	почему	226	0.11
14	на	2293	1.12	62	хорошо*	428	0.21	110	была	224	0.11
15	ты	1674	0.82	63	потому	422	0.21	111	нам	221	0.11
16	с	1590	0.78	64	за	416	0.20	112	такие*	220	0.11
17	то*	1550	0.76	65	давай*	415	0.20	113	сколько	218	0.11
18.5	всё*	1508	0.74	66	тебя	414	0.20	114	него	217	0.11
18.5	нет	1508	0.74	67	знаешь*	410	0.20	115	этого*	208	0.10
20	(э-э)*	1484	0.72	68	говорю*	403	0.20	116	блядь*	206	0.10
21	он	1441	0.70	69	только*	401	0.20	117	ему	204	0.10
22	как	1366	0.67	70	чего	399	0.19	118.5	пять	199	0.10
23	мне	1207	0.59	71	бы*	392	0.19	118.5	типа*	199	0.10
24	она	1134	0.55	72	этот*	372	0.18	121	время	197	0.10
25	угу*	1116	0.54	73	можно*	369	0.18	121	короче*	197	0.10
26	есть	1073	0.52	74	где	367	0.18	121	понятно*	197	0.10
27	меня	1045	0.51	75	ага*	366	0.18	123.5	(м)*	196	0.10
28	мы	947	0.46	76	ничего*	361	0.18	123.5	слушай*	196	0.10
29	они	944	0.46	77	конечно*	355	0.17	125	эта	190	0.09
30	сейчас	934	0.46	78	такая*	352	0.17	126	какой	189	0.09
31	ещё*	917	0.45	79	что-то	350	0.17	127.5	могу	182	0.09
32	уже*	867	0.42	80	раз	330	0.16	127.5	наверное*	182	0.09
33	но	863	0.42	81	её	314	0.15	130	двадцать	180	0.09
34	надо*	846	0.41	82	такое*	312	0.15	130	ей	180	0.09
35	же*	812	0.40	83	эти*	312	0.15	130	нормально*	180	0.09
36	просто*	739	0.36	84	даже*	307	0.15	132.5	туда	178	0.09
37	по	725	0.35	85	блин	305	0.15	132.5	хочу	178	0.09
38	вообще	689	0.34	86	до	301	0.15	134	в общем	176	0.09
39	нас	615	0.30	87	вас	293	0.14	135	эту	174	0.08
40	знаю*	608	0.30	88	два	292	0.14	136.5	нужно*	172	0.08
41	если	585	0.29	89	ладно*	288	0.14	136.5	этом	172	0.08
42	тоже	583	0.28	90	был	285	0.14	138	много	170	0.08
43	все	526	0.26	91	их	277	0.14	139.5	(н-н)*	164	0.08
44	было	521	0.25	92	кто	274	0.13	139.5	понимаешь*	164	0.08
45	вы	511	0.25	93	ли*	272	0.13	141.5	буду	157	0.08
46	здесь	496	0.24	94.5	для	271	0.13	141.5	делать	157	0.08
47	говорит	494	0.24	94.5	чтобы	271	0.13	143	больше	155	0.08
48	потом	490	0.24	96	один	267	0.13	144	этой	153	0.07

National Corpus are similarly taken by pronouns: *I*, *you* and *it* (Leech et al. 2001). After being lemmatized, the lexeme *я* has the frequency 7686 (3.75% of tokens). As for the pronoun *ты* (*you*), its frequency is more modest, making way for a row of discourse particles: the word form *ты* (the nominative case of *you*) has rank 15 (0.82%). After being lemmatized, *ты* has a frequency of 2624 (1.28%), but still has a lower rank than "you" has in British English frequency list.

The negative particle *не* (*not*) has the second rank, taking practically the same percentage as the pronoun *я* (2.40%). It is an interesting phenomenon, which may lead to a hypothesis that in oral communication Russians prefer to use negative phrases. However, if we refer to traditional Russian frequency dictionaries, we may see that particle *не* always takes high ranks: third rate in Zasorina (1977) and fourth rate in Steinfeldt (1963).

What is totally surprising is the great number of discourse particles: *вот* (rank 3, 2.34%), *ну* (rank 4, 2.31%), *да* (both a positive particle *yes* and a discourse particle *yet*) (rank 5, 2.13%), *там* (rank 11, 1.43%), *так* (rank 11, 1.22%) and many others. Discourse particles, called in traditional Russian linguistics "parasitic words", are known to be a widespread feature of Russian spoken language. Earlier these particles were considered to be indicators of poor speech skills. This is likely to be the case when the percentage of these idle "words" is excessive as in the example below. However, in recent studies the meaning of discourse particles is reconsidered in a more pragmatic sense, e.g., in Shmelev (2008).

The ORD corpus provides opportunity to study the functioning of lexemes and particles in real speech. For example, the following phrase belongs to speaker S39 (photographer) explaining to a client in his studio the differences in anthropological features of different nations. From 48 word-like elements of this utterance only 10 (which are underlined in wave form) have lexical meaning. All the other words are discourse particles or just fillers:

а вот (э-э) на. . . н. . . вот наше вот это вот (э-э) вот это вот / вот тут / тут сложнее гораздо / да // потому что / значит / я вот / вот (э-э) вот эти / ну в принципе / значит / ну / п. . . по моим / понятиям значит / я же не отличу так скажем / таджика от узбека что называется / да да да // да ?

This sentence means:

[. . .] it's much more difficult here / [. . .] I can not distinguish [. . .] / Tajik from Uzbek [. . .] / can I ?

5 Conclusion

The results presented should be considered to be preliminary, as annotation of the corpus is still in progress. Evidently, the expansion of the empirical base and the processing of new transcripts will in a way modify the given results, though general tendencies should remain. Our next task is to built frequency lists and concordances for individual speakers, for groups of speakers and for communicative situations that are similar. It should allow us to describe more precisely the vocabulary and grammar of the Russian modern spoken language, and throw more light on Russian communication strategies.

Acknowledgments. The first recordings and database creation of the ORD corpus was supported by the Russian Foundation for Humanities within the framework of the project "Speech Corpus of Russian Everyday Communication *One Day of Speech* (ORD)" (Project #07–04–94515е/*Я*). Creating and processing of the ORD corpus is supported by the special program of the Russian Ministry of Education entitled "Sound Form of Russian Grammar System in Communicative and Informational Approach" and by a grant of the Russian Foundation for Humanities "The development of an information system for monitoring of Russian spoken language" (Project #09–04–12115v).

References

Asinovskij, A.; Bogdanova, N.; Rusakova, M.; Ryko, A.; Stepanova S.; Sherstinova, T.
2009 "The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation". In: Matoušek, V.; Mautner, P. (eds.), *TSD 2009, LNAI 5729.* Berlin, Heidelberg: Springer, 250–257.

Boersma, P.; Weenink, D.
2009 "Praat: Doing Phonetics by Computer."
[Electronic source: `http://www.praat.org`]

Hellwig, B.; van Uytvanck, D.; Hulsbosch, M.
2009 "ELAN – Linguistic Annotator. Version 3.8."
[Electronic source: `http://www.mpi.nl/corpus/manuals/manual-elan.pdf`]

Leech, G.; Rayson, P.; Wilson, A.
2001 *Word Frequencies in Written and Spoken English: based on the British National Corpus.* London: Longman.

Ryko, A.; Stepanova, S.
2008 "Problemy vychleneniya edinic analiza spontannogo ustnogo teksta." In: *Materialy XXXVII mezhdunarodnoj filologicheskoj konferencii*, V. 21. Sankt Peterburg: Sankt Petersburg State University, 71–80.

Sherstinova, T.
2009 "The Structure of the ORD Speech Corpus of Russian Everyday Communication". In: Matoušek, V.; Mautner, P. (eds.), *TSD 2009, LNAI 5729.* Berlin, Heidelberg: Springer, 258–265.

Sherstinova, T.; Stepanova, S.; Ryko, A.
2009 "Sistema annotirovaniya v zvukovom korpuse russkogo yazyka", in: *Materialy XXXVIII mezhdunarodnoj filologicheskoj konferencii: 'Formal'nye metody analiza russkoj rechi'*. St. Petersburg: St. Petersburg Gos. Universitet; 6–75.

Shmelev, A.
2008 "Parasitic words as discourse markers: the case of Russian."
[Electronic source: `http://www.aatseel.org/100111/pdf/program/2008/28a8_1shmelev_alexei.pdf`]

Steinfeldt, E.
1963 *Chastotnyj slovar' sovremennogo russkogo literaturnogo yazyka.* Tallinn: Estonian Research Pedagogical Institute.

Zasorina, L. (ed.)
1977 *Chastotnyj slovar' russkogo jazyka*. Moskva: Russkij Yazyk.

Complex investigation of texts with the system "StyleAnalyzer"

O.G. Shevelyov, V.V. Poddubnyj

1 Introduction

Discovering regularities of stylistic features of texts (genres, authors, author's gender, topics, etc.), clustering and classification, and vocabulary analysis involves the processing of textual and numeric data. Common mathematical and statistics packages, and word processors are too unwieldy to work with large amounts of such mixed data. They have a lot of irrelevant methods and a lack of specific ones dealing with text processing peculiarities. Most specialized quantitative linguistics software are single-purpose implementations of particular methods that do not provide functionality for multi-faceted computational linguistics investigations like preparing and filtering the data, choosing method parameters, the testing and comparison of different methods. At the same time the development of quantitative linguistics highly depends on the availability of powerful tools for versatile and mass text analysis. Those tools have to be convenient enough to be used not only by software developers and mathematicians, but, what is even more important, by classical scholars (e.g. linguists, historians).

In 2004 the project named "StyleAnalyzer" was started at the Computer Science Department of Tomsk State University. In 2005 a group of linguists from Moscow State University joined the project. The idea was to create a versatile multiple document interface (MDI) software tool for researchers to carry out various computational linguistics investigations. The process of research in "StyleAnalyzer" can be divided into three main stages:

1. preprocessing of texts,
2. transforming texts to numeric data and preprocessing it,
3. analysis of numeric data.

Every stage is independent and provides intermediate data that can be saved and used in other systems if StyleAnalyzer lacks some method. This article outlines the stages of processing in the tool developed, the methods included in the tool, and the possibilities that the tool brings to the researchers.

2 Preprocessing of texts

Unfortunately there is no public data base of electronic texts covering many text styles, languages and authors where all the texts are free of formatting errors. Texts for experiments usually come from different sources, and have different formatting. Analysis of them without any preprocessing can cause unpredictable and distorted results. Moreover, preprocessing is useful for some specific analyses of texts (e.g. narrator speech). StyleAnalyzer works with simple text files. It receives a list of texts which can be processed with given parameters. Preprocessing includes:

- cleaning texts and the unification of their formatting in batch mode,
- special processing (e.g. removing dialogs, splitting into fragments),
- the correction and transforming of Russian grammar markup,
- the substitution of words by given vocabulary (e.g. replacing all the words with their roots).

StyleAnalyzer also works with so called "vertical texts". They are text files where each running word is located on its own line together with its normal form and grammar codes. The current version of StyleAnalyzer does not create vertical texts itself, but imports them from the format of the system DicTUM-1, developed at MSC (cf. Kukushkina and Polikarpov 1996). DicTUM-1 adds grammar information to words. On the stage of the preprocessing, the service information is manually added to texts. The information can include different attributes like title, short title, author name, and short author name, gender of author, genre type, and genre, dates of creation and publication, period of creation (for example, 60th), subject. It is used first to create uniform subsets of texts (for example, to investigate text only in a specific genre to get rid of gender differences), and second for analysis (for example, to classify by author, or gender, or to show detailed information about texts on graphs).

3 Extraction of numeric features

The majority of analysis methods are feature-based. Therefore one first has to extract numeric data from texts and only then can analyze them. StyleAnalyzer has a built-in query language for the extraction of values of various user-defined features. It allows the extraction of frequency values of specific combinations and chains of text elements: letters, morphemes, words, categories of words and sentences. One can set the grammar properties of words in a query, providing processed texts are vertical. For a given set of texts a user can choose a set of features. They can do basic operations with queries (e.g. adding a new feature, deleting one), save and load presets of features to a file. Before starting the extraction the user has to choose a representation of results (e.g. feature table, list of text elements fitting to features, sequences of features in text).

Texts can be automatically broken into equal size fragments if need be. There are macros buttons that allow adding elements and their properties to a query. During feature value extraction the user sees a growing table of features and detailed statistical information about the process. Any extracted data maintains all its source information, so every line in a result table is linked to a particular text. The results are written to two files:

1. a text file with pure numeric data,
2. an XML file with additional information about the fragments processed (e.g. their sizes, titles, authors).

The size of a text in different elements (sentences, words, and letters) is one of the most important pieces of information that is used for normalization and by text processing algorithms. One special type of text transformation is the extraction of suffix arrays. A suffix array is an array of integers holding the starting positions of suffixes of a string (text) in lexicographical order. In StyleAnalyzer suffix arrays are used for creating suffix trees representing suffixes of texts in a special graph that is convenient for compression-based analysis methods.

4 Analysis

There are three types of data analysis in StyleAnalyzer:

1. structure analysis,
2. feature-based analysis,
3. compression-based analysis.

Structure analysis deals with source texts without any feature extraction. StyleAnalyzer has rather limited abilities in this field right now. It allows for extracting vocabularies of texts (words and normal forms), calculates phonosemantic values of Russian words and texts (to estimate how a word or text sounds, for example, bad or good, great or misery, weak or strong, cold or hot) (Žuravlev 1974). StyleAnalyzer can also generate pseudo-words with given phonosemantic characteristics. Suffix arrays and suffix trees can also be used for structure analysis.

Feature-based analysis uses the data frequency tables built at the stage of feature extraction. StyleAnalyzer includes hierarchical cluster analysis (by different measures, cluster distances), statistical hypothesis testing, classification (decision trees, feed-forward neural networks, entropy-based Khmelev method, *C*-, *R*-measures), feature space analysis and reduction (entropy- and classification-based, factor analysis), visualization (graphs of feature values, self-organizing maps). Those methods provide all runtime data (e.g. distance or cluster linkage tables, decision tree text rules) for reporting or debugging. Results data (tables, graphs, maps) can be adjusted in the graphical user-interface

to show the information of interest to the user (e.g. only text authors and titles, only gender and topic for texts or distances values for clusters). Additional statistics such as corpus size in megabytes or words, the number of texts by text characteristics (gender, genre, etc.) are available. The resulting classification and clustering data can be verified by state-of-the-art testing methods (k-fold, leave-one out) and estimated by tried-and-tested measures (recall, precise, F-measure). The results of testing and estimations can be shown in an MDI window with corresponding tables or marked directly on a graph (e.g. F-measure for best clusters on a dendrogramm) (Poddubnyj et al. 2006: 121).

Compression-based analysis works with suffix trees. There are no particular features the user has to set for compression-based methods. Those methods use texts themselves as as sets of strings. StyleAnalyzer uses suffix arrays as an input and creates suffix trees on-the-fly while using a compression-based method. Currently only one compression-based approach is implemented in the system that is used for clustering texts by CS-, RS- or TS-measures. They are symmetrical modifications of the C-, R-, T-measures (Shevelyov 2008: 113) developed by Hunnisett and Teahan (2004).

5 Conclusion and future work

StyleAnalyzer has been extensively used by the Laboratory of Computer Linguistics and Lexicology at Moscow State University. Employees of the laboratory have been doing different experiments with big corpora. Hundreds of texts by many authors were clustered with different parameters by different text styles. Some experiments were made in StyleAnalyzer using classification methods (Kukushkina et al. 2007: 391). The main goal of the experiments was to find sets of features that could be reliably used to distinguish different types of texts.[1]

It has been noted by Moscow State University linguists that StyleAnalyzer is very convenient for different kinds of text analysis, though some methods and procedures that are necessary for better understanding of the results are still missing. The linguists are not interested in "black box" methods. They always ask for clear explanations of how a specific method works, which rules or limitations are used, and whether or not it is possible to see additional graphs and tables that clarify why a text falls into a certain cluster or was recognized as a particular class. The collaboration of the programmers from Tomsk and the linguists from Moscow, is mainly an effort to create a common understanding. The main results of the collaboration are improved reporting, configuration and visualization systems of StyleAnalyzer.

It was found that the query language's powerful abilities to create sophisticated queries was not very useful. Most of the feature sets are quite simple and

1. Russian Foundation for Basic Research, grant 06-07-89320 for 3 years

it is the speed of feature value extraction that is important for users. Storing text as files on a local disk is not so practical as well. Users tend to create a lot of different versions of corpora and then it is really hard to merge them. Researchers create subsets of the data for their needs, copy texts and feature sets here and there, and it becomes very complicated to follow the actual state of the corpora when researchers have different geographical locations. There are problems with access rights. Many third-party researchers ask to try the system, and students of the philological faculty need it for their scientific work, but the StyleAnalyzer research group currently cannot share the tool. The only way to distribute it is to copy the whole program that is written in *C#* programming language which provides limited abilities to protect the algorithms implemented. Another problem is the speed of calculations. Processing large amounts of text data in a reasonable time may require parallel algorithms.

In the end, we decided to start the development of the next generation of the StyleAnalyzer. This time the idea is to create a web-tool that will work with texts in a database to facilitate distributed access and different access levels researches. The tool is going to be developed in the open source program language Java, use open source database and state-of-art technologies like JSP, Ajax, dependency injections, and the Google Web Toolkit. A security system, and parallel processing are being considered from the very beginning. The query language will be changed by using regular expressions to make feature extraction more standardized, faster and simpler. The main efforts will be put into the user interface, corpus statistics, and explanatory features. Most of the analysis algorithms will be imported from the old StyleAnalyzer.

References

Hunnisett, D.; Teahan, W.J.
2004 "Context-based methods for text categorization." [Electronic source: `http://www.philol.msu.ru/~lex/articles/dictum.htm`]

Kukushkina, O.; Polikarpov, A.
1996 "DicTUM-1. A system for dictionary-text universal manipulations and analysis." [Electronic source: `http://www.philol.msu.ru/~lex/articles/dictum.htm`]

Kukushkina, O.; Polikarpov, A.; Poddubnyj, V.; Shevelyov, O.
2007 "Avtomaticheskaya klassifikaciya tekstov korpusa russkikh gazet konca XX veka po zhanrovym tipam i istochnikam [= Automatic classification of corpora of texts of Russian newspapers of the end of XX century by the genre type and newspaper sources]", in: *Russkij yazyk: istoricheskie sud'by i sovremennost': III Mezhdunarodnyj kongress issledovatelej russkogo yazyka (Moskva, MGU im. M.V. Lomonosova, filologicheskij fakul'tet, 20–23 marta 2007 g.*; 391–392.

Poddubnyj, V.; Shevelyov, O., Bormashov, D.
2006 "Sravnenie kachestva pokhodov k klasterizacii tekstov na osnove gipergeometricheskogo kriteriya [= The comparison of quality of texts clustering approaches based on hypergeometrical criterion]", in: *Vestnik Tomskogo gosudarstvennogo universiteta*, 293; 120–125.

Shevelyov, O.
2008 *Metody avtomaticheskoj klassifikacii tekstov na estestvennom yazyke.* [= Methods of automatic classification of natural language texts.] Tomsk: TML-Press.

Zhuravlev, A.P.
1974 *Foneticheskoe znachenie.* [= Phonetic meaning]. Leningrad: LGU.

Retrieving collocational information from Japanese corpora: its methods and the notion of "circumcollocate"

Tadaharu Tanomura

1 Introduction

Although Japanese may be said to be one of the best-studied languages of the world, the history of Japanese corpus linguistics has been very short unfortunately. There has been no publicly available balanced corpus of the language to this day, and the number of researchers and consequently the number of relevant works as well has been limited. The situation started to change recently, however. Most notably, a five year nation-level project of Japanese corpus linguistics started in 2006. It is a collaborative project of the National Institute for the Japanese Language and a few dozen researchers from other institutions. The principal goal of the project is to construct a balanced corpus of contemporary written Japanese. A corpus of a hundred million words, which is now under construction, will be completed by the spring of 2011. The corpus will interest researchers of Japanese, and corpora will play an increasing role in a variety of areas of Japanese linguistics in the future.

In what follows, I will discuss issues of the retrieval of collocational information from Japanese corpora, one of my recent attempts at making effective use of Japanese corpora. The most important area of expected application of corpus-based collocational analysis I have in mind is the creation of a dictionary of Japanese collocations, either in printed or electronic form.[1]

2 Preliminary remarks

Before starting the main discussion, a few preliminary remarks will be in order.

2.1 Japanese grammar and writing system

First, we will briefly sketch the grammar and writing system of Japanese, so that the main discussion later on may be understood better by the reader who is not familiar with the language.

1. A more detailed discussion of the topics dealt with in Section 3 of this paper, as well as a few created sample entries of a possible collocational dictionary of Japanese, may be found in Tanomura (2009). The topic to be taken up in Section 4 is discussed here for the first time.

2.1.1 Grammar

Japanese is a consistent SOV, head final language, as may be exemplified by the following simple sentence:[2]

(1) watasi-wa ongaku-o aisu-ru.
I-TOPIC music-ACC love-PRES
'I love music.'

Another important characteristic of Japanese grammar is its agglutinative morphology. Grammatical elements are appended to nouns and verbs as shown in Example (2):

(2) a. watasi-ga ('I-NOM')
watasi-o ('I-ACC')
watasi-ni ('I-OBL', 'to me')
watasi-ni-mo ('I-OBL-also', 'also to me')
watasi-ni-sae ('I-OBL-even', 'even to me')

b. tabe-ru ('eat-PRES', 'eat')
tabe-ta ('eat-PAST', 'ate')
tabe-nai ('eat-NEG', 'do not eat')
tabe-rare-ru ('eat-PASS-PRES', 'be eaten')
tabe-rare-ta ('eat-PASS-PAST', 'was/were eaten')
tabe-rare-nai ('eat-PASS-NEG', 'be not eaten')
tabe-sase-ru ('eat-CAUS-PRES', 'make *sb* eat')
tabe-sase-rare-ta ('eat-CAUS-PASS-PAST', 'was/were made to eat')

2.1.2 Writing system

As for the writing system of Japanese, there are two major kinds of characters used in Japanese, called *kana* and *kanji* respectively. Each *kana* basically denotes a mora. There are about a hundred of them, including the following.

(3) あ (a), い (i), う (u), え (e), お (o),
か (ka), き (ki), く (ku), け (ke), こ (ko)

Each *kanji* basically represents a combination of sound and meaning. There are tens of thousands of them, and they were borrowed from Chinese in older times with a small number of exceptions.

2. Japanese examples will be transcribed roughly according to the *Kunrei-shiki* romanization system.

(4) 一 (iti, ‘one’), 二 (ni, ‘two’), 三 (san, ‘three’), 四 (si, ‘four’)
私 (watasi, ‘I’), 愛 (ai, ‘love’), 音楽 (on gaku, ‘music’)

From a computational linguistics point of view, a crucial fact about conventional Japanese writing is that words are not separated by spaces. A sentence written in *kana* and *kanji* will look like:

(5) 私たちはみな音楽を愛する。
‘We all love music.’

which may be decomposed as follows according to the phrasal and morphological boundaries:

(6) 私-たち-は みな 音楽-を 愛す-る。
watasi-tati-wa mina ongaku-o aisu-ru
I-plural-TOPIC all music-ACC love-PRES

The fact that words are not separated by spaces presents a challenge for almost any kind of computational processing of Japanese texts, including the analysis of collocation. We will return to this issue later.

2.2 Web corpus used in this study

A large amount of linguistic data is required for collocational analysis. In this study, a Web corpus constructed by the author in 2008 will be used. It is a collection of some ten million Web pages, and consists of about 75 billion characters. This amounts to about 45 billion words, or 150 Giga bytes in file size.

The Web corpus was constructed by the following five-step procedure:

1. Make a large list of (sets of) words.
2. Search with *Yahoo*! using those (sets of) words as keywords.
3. Acquire the first hundred URLs in each search result.
4. Acquire the documents referred to by the URLs.
5. Eliminate HTML tags, etc. from the documents.

Although a number of problems were encountered and dealt with in the actual processing, they will not be mentioned here. The nature of the acquired documents may differ depending on the keywords given to the search engine in the second of the five steps mentioned above. Basically I employed the method of using a set of keywords which were expected to be unbiased as a whole, but I also attempted a second method of using a set of keywords characteristic of a particular topic or style. In the first method, the set of keywords were prepared by means of a mechanical segmentation of various types of Japanese texts, whereas in the second method, a few sets of keywords characteristic of

a particular topic or style were prepared manually. Of the constructed Web corpus of 150 Giga bytes, the ratio of the amounts of the texts collected by the two methods is 2:1. In this paper, the text data of 100 Giga bytes acquired by the first method will be used for analysis of collocation.

3 Retrieving collocational information

Now we turn to the main topic of this paper, i.e., the retrieval of collocational information from Japanese corpora.

3.1 Three possible approaches

The first question we need to address is what type of collocates we should count. Since words are not separated by spaces in Japanese writing, we have to decide what to count in the first place. I tried three methods:

1. counting co-occurring *words*
2. counting co-occurring *word sequences*
3. counting co-occurring *character sequences*

My conclusion is that the second method of counting co-occurring word sequences, namely, co-occurring *N*-grams on the word level, is the most effective one. The third method does produce useful results similar to those obtainable by the second method, but is not as precise as the latter. Due to the limitation of space, we will limit our discussion below to the first and second methods.

In the first and second methods of collocational analysis, we need to identify words. Identification of words, or morphemes to be more exact, will be done with the help of a morphological analyzer named *MeCab* (a broad IPA transcription of its pronunciation in Japanese is [mekabu]), which is distributed at `http://mecab.sourceforge.net/`.[3]

3.2 Method of counting co-occurring words

In the first method of collocational analysis, words which co-occur with a given expression are counted. Below are three of the verbs which were found to most frequently follow the noun *netui* (‘zeal, enthusiasm’) in the Web corpus.

3. Mechanical morphological analysis cannot be without errors. But fortunately, that does not cause a serious problem as far as practical application such as dictionary making is concerned. In making a collocational dictionary, statistical facts are only one of the elements which need to be taken into consideration. Thus, for example, if a collocate sometimes fails to be counted in corpus analysis, or if something which need not be counted is sometimes counted, it does not matter unless such mistakes occur frequently. For a more detailed discussion on this point, see Tanomura (2009).

(7) a. aru (‘exist’)[4]
b. motu (‘have’)
c. makeru (‘succumb’)

Actually we get a large list of such verbs by analyzing a huge corpus, but we will show only a small number of them here, so that the reader who is not familiar with Japanese can easily concentrate on the meaning of the examples. Since this kind of information concerning the co-occurrence of words is not easily accessible to the native speaker’s introspection, we can ascertain the effectiveness of this approach.

However, the list of verbs thus obtained is not informative enough. This is because the three verbs in list (7) are not grammatically related to the noun *netui* (‘zeal, enthusiasm’) in the same way. As is shown in list (8), in the case of the verb *aru* (‘exist’), the noun *netui* takes on the nominative case marker *ga* and functions as the subject; in the case of *motu* (‘have’), *netui* is followed by the accusative case marker *o* and functions as the object; in the case of *makeru* (‘succumb’), *netui* is followed by the oblique case marker *ni*.

(8) a. netui-ga aru (‘zeal-NOM exist’, ‘(*lit.*) a zeal exists, *sb* has a zeal’)
b. netui-o motu (‘zeal-ACC have’, ‘have a zeal’)
c. netui-ni makeru (‘zeal-OBL succumb’, ‘succumb to *sb’s* zeal’)

Thus we should not limit our attention to the co-occurring relation between the bare noun *netui* and the verbs. Rather, we need to observe the relation between the noun followed by each of the case markers on the one hand and the verbs on the other.

The following shows part of the results which were obtained by way of such a modified procedure. Here are listed four of the verbs which most frequently co-occur with the noun followed by the oblique case marker *ni*.

(9) a. makeru (‘succumb’)
b. kotaeru (‘respond’)
c. utu (‘strike’)
d. ugokasu (‘move’)

Note that verbs such as *aru* (‘exist’) and *motu* (‘have’) have been properly excluded from this list of verbs. Obviously, (9) and similar lists of the verbs co-occurring with the noun *netui* followed by other case markers will be more useful than an indiscriminative verb list of which (7) is a part.

4. In fact, it is possible to analyze the basic verbal forms *aru*, *motu* and *makeru* as *ar-u* (‘exist-PRES’), *mot-u* (‘have-PRES’) and *make-ru* (‘succumb-PRES’) respectively as we did with the predicate forms listed in Examples (1) and (2-b). For the sake of simplicity of exposition, however, we will hereafter regard the present tense suffix *-(r)u* as a part of the unanalyzed monomorphemic verbal forms, rather than as an independent morpheme. Morphological complications will be ignored at the sacrifice of exactness and consistency.

Nevertheless, list (9) is not informative enough, either. The reason is that the four verbs in (9), in actual contexts of occurrence, are not all related to *netui-ni* ('zeal-OBL') in the same way. In the case of the first and second verbs, *makeru* ('succumb') and *kotaeru* ('respond'), they appear in the active voice, as shown in (10-a) and (10-b), whereas in the case of the third and fourth verbs, *utu* ('strike') and *ugokasu* ('move'), they must be followed by the suffix *-areru*, forming passive predicates, as in (10-c) and (10-d).

(10) a. netui-ni makeru ('zeal-OBL succumb', 'succumb to *sb*'s zeal')
b. netui-ni kotaeru ('zeal-OBL respond', 'respond to *sb*'s zeal')
c. netui-ni ut-areru ('zeal-OBL strike-PASS', 'be struck by *sb*'s zeal')
d. netui-ni ugokas-areru ('zeal-OBL move-PASS', 'be moved by *sb*'s zeal')

To summarize the point, although the method of counting co-occurring words in fact provides us with collocational information which is hard to obtain through the native speaker's introspection, information obtainable by this method is not informative enough. This method therefore needs to be replaced in favor of the second method of counting not only co-occurring single words but also co-occurring sequences of words or morphemes (including verb-suffix sequences such as *ut-areru* ('strike-PASS') and *ugokas-areru* ('move-PASS') among others), to which we now turn.

3.3 Method of counting co-occurring word sequences

In the second method of analysis, sequences of words, or morphemes, which co-occur with a given expression are counted.

The following list shows a few of the word sequences co-occurring with *netui-ni* ('zeal-OBL') which were acquired by analyzing the Web corpus by this method.

(11) a. makeru ('succumb')
b. kotaeru ('respond')
c. ut-areru ('strike-PASS', 'be struck')
d. ugokas-areru ('move-PASS', 'be moved')

These word sequences are in fact what we listed in (10) above and wanted to obtain by corpus analysis.

We will see another example of the results of an analysis by the second method. Example (12) is a small portion of a large list of word sequences co-occurring with the adverb *mekkiri* ('noticeably, to an obvious degree'). This adverb is felt by the author to tend to be used with a predicate denoting either a decrease of something observable, a change to cold weather or a decline in health or vigor.

(12) a. nat-ta (‘become-PAST’, ‘became (cold, few, etc.)’)
b. het-ta (‘decrease-PAST’, ‘decreased’)
c. sukunaku nat-ta (‘few become-PAST’, ‘became few, decreased’)
d. samuku nat-ta (‘cold become-PAST’, ‘(the weather) turned cold’)
e. samuku nari-masi-ta (‘cold become-POLITE-PAST’, ‘turned cold’)
f. otoroe-ta (‘decline-PAST’, ‘(*sb*’s health) declined’)

This result again conforms to the native speaker’s intuition, and thus we may ascertain the effectiveness of the method of counting co-occurring word sequences.

However, with respect to the notion of “word sequence”, there is still something more to observe, which we will discuss in the next section.

4 The notion of “circumcollocate”

There would be no reason to restrict the notion of collocation to a relation between two words, as is commonly done in discussions of collocation. Rather, collocation should be regarded as a relation between multiple words in general, as we saw in Section 3.2 especially if our goal is to make a collocational dictionary. What counts from the practical viewpoint of dictionary making are the overall patterns in which a given expression habitually occurs, rather than simple relations between the expression and co-occurring single words.

To advance this understanding further, we may introduce the notion of “circumcollocate” as a particular type of co-occurring word sequence. A “circumcollocate” is a collocate which habitually sandwiches (i.e. occurs on both sides of) a given expression. In other words, a “circumcollocate” is a combination of what we would call for convenience a “precollocate” and a “postcollocate”, each of which may be a word sequence instead of a single word.

We will illustrate the significance and usefulness of the notion of circumcollocate in the analysis of Japanese collocation by taking two examples.

First, *tukiru* is an intransitive verb denoting ‘run out, be exhausted’. The nouns which frequently appear as the subject of this verb include *tikara* (‘power’), *bansaku* (‘all measures’), *aisoo* (‘patience’), *kyoomi* (‘interest’), *gimon* (‘doubt’), *nayami* (‘worry’). But these nouns fall under two groups depending on the overall expression in which they co-occur with the verb *tukiru*.

(13) a. {tikara/bansaku/aisoo-ga} tuki-ta
{power/all measures/patience-NOM} be exhausted-PAST
‘{power/all measures} was/were exhausted’

b. {kyoomi-ga/gimon-wa/nayami-wa} tuki-nai
{interest-NOM/doubt-TOPIC/worry-TOPIC} be exhausted-NEG
‘{interests/doubts/worries} will never be exhausted’

Note that in the first group of nouns shown in (13-a), they co-occur with the affirmative, past tense form of the verb *tukiru*. On the other hand, in the second group of nouns shown in (13-b), they co-occur with the negative, non-past tense form of the verb. The nouns in question might sometimes appear in a different configuration as well, but there is an overwhelming tendency for them to be used in the ways indicated above.

So it does not suffice to simply say that the verb *tukiru* frequently follows these nouns, or for that matter that it sometimes precedes the suffixes of past tense or negation. We need to pay attention to the combination of the noun as a precollocate and the verbal suffix as a postcollocate. My proposal is that we identify discontinuous word sequences such as *tikaral…ta* ('power…PAST') and *kyoomi-ga…nai* ('interest-NOM…NEG') as circumcollocates of the verb *tukiru*. Such information about circumcollocates of a given expression in a collocational dictionary will be no less useful than information concerning precollocates and postcollocates.

Let us see another example of circumcollocate with *ooki* ('(be) many'), an adjective of an archaic style. The adjective *ooki* is typically used in the phrase pattern *N1 ooki N2*, where *N1 ooki* is a relative clause modifying *N2*, as exemplified by the following.

(14) koi ooki onna
love be many woman
'(*lit.*) woman with whom loves are many, woman with many love affairs'

What is worth noting about this phrase pattern is that the habitual combination of *N1* and *N2* is rather restricted. The expressions which frequently occur in the Web corpus include the following.

(15) a. koi ooki {onna/otoko/otome}
love be many {woman/man/maiden}
'{woman/man/maiden} with many love affairs'

b. {nayami/yume} ooki {tosigoro/zinsei/hibi}
{worries/dreams} be many {age (of a person)/life/days}
'{age/life/days} with many {worries/dreams}'

The phrase pattern *N1 ooki N2* is not idiomatic when saying, for example, 'man with a lot of money' or 'days with a lot of rain', in spite of the grammaticality of the expressions. Hence the need to identify *koi…onna* ('love…woman'), *nayami…tosigoro* ('worry…age'), etc. as circumcollocates of the adjective *ooki*, and to include such information in the entry for *ooki* in a collocational dictionary.

5 Conclusion

The aim of this paper has been to search for an effective method of retrieving collocational information from Japanese corpora, and to argue for the significance of looking at the collocational relation between multiple words, including that of "circumcollocate".

Japanese corpus linguistics has been lagging behind, but the situation is changing. Corpora will begin to be applied to a variety of research topics in Japanese linguistics and yield fruitful results in the near future.

References

Tanomura, T.
2009 "Retrieving collocational information from Japanese corpora: An attempt towards the creation of a dictionary of collocations", in: *Handai Nihongo Kenkyu [Osaka University]*, 21; 21–41. [In Japanese].

Diachrony of noun-phrases in specialized corpora

Nicolas Turenne

1 Introduction

Nowadays some fields like biology or nanotechnology produce a large number of scientific papers. It is possible to build specialized text collections for content analysis. Documents are also time-stamped, and so features extracted from them are able to be processed by time. Systematic analysis of language properties benefits from a distribution description defining linguistic laws. Lots of laws have been discovered since the pioneering work of Zipf (1929, 1932, 1935). In this paper, we explore the impact of time on the distribution of "content-word" occurrences in texts, sometimes called named entities, and the impact of time on the distribution of clusters of "content-words" where "content-words" are linked because they share common contexts and so reveal context dependencies. We made our analysis of single noun-phrases using two different tools for named-entity extraction (names of proteins and genes) from two different text collections. But single noun-phrases do not give real semantic information about content, the factor most representative is association. In our study we suppose it is given by simultaneous presence (co-presence or co-occurrence) in the same context and with repetition (at least two different contexts). In our definition a context is a document, which is larger than the classical linguistic context defined by a sentence. In the second part we used three corpora, two in biology and one in computing. The method to extract noun-phrases is robust and not field-dependent (a sequence extraction of strings which do not contain a function word such as a verb, conjunction, adverb, preposition, pronoun or number). We plot distributions for a number of patterns (i.e. associations or clusters) by size for any given period settled a priori, and the distribution of the number of patterns by time interval for any given size of pattern.

Section 2 presents an overview of the state of the art of linguistic approaches for diachronic analysis, linear distribution and distribution of word component position in text. Section 3 presents our distribution analysis for single noun-phrases, and noun-phrase co-occurrences over time.

2 Related works

In this section, we review formal approaches focused on diachrony, linear laws and complexity. The earliest publication on this theme is probably Jespersen

(1929), who investigated the distribution of French loan words in the English lexicon as a function of time. He developed the "ease theory", as a systematic approach to show the evolution of language change driven by a principle of least effort. In the same year, G.K. Zipf published his doctoral dissertation, which dealt with changes of sounds (phonemes).

There are two ways to study the invariance through time of word productivity: by surveying successive editions of dictionaries (Dubois, 1962; Neuhaus, 1973) and by the analysis of corpora within different periods. Our interest is more focused on the latter. Baayen and Renouf (1996) call attention to the hapax legomena with certain affixes. Their study shows that hapax appear more often as time increases, which might indicate that their productivity is increasing. A category-conditioned degree of productivity provides a lower bound for the rate at which new formations may be expected. Lots of studies try to examine language creolization (Labov, 1980). Early approaches were concerned with the determination of the genetic proximity of languages with respect to their vocabularies by the reconstruction of "family trees". Corresponding methods were proposed by Kroeber and Chrétien in 1937. Similar approaches were developed during the following decades (Embleton, 1986). Almost as famous as Zipf's laws is glottochronology (Swadesh 1952, 1955). With his method, Swadesh tried to calculate the number of words which disappear from a lexicon to predict and to date the moment of separation of two related languages on the basis of the observed proportion of common words in their lexicons. In analogy to the well-known method of dating objects in archaeology, it uses a mathematical model of the radioactive decay of the carbon isotope 14C. The concrete course of lexical change processes is explained by the Piotrowski Law. Altmann coined this term for the approach he presented (Beőthy and Altmann, 1984) which describes and predicts the observed trend of the changes on the time axis. The history of success of a new linguistic phenomenon always begins slowly, then speeds up and finally slows down again. Formerly, several authors made assumptions about the nature of the corresponding curves. A first systematic consideration was by Piotrowskij (1968) from which Altmann derived his mathematical form of the law starting from an interactionist approach and set up a differential equation whose solution provides appropriate models for three different types of increase dynamics. The Piotrowski Law represents the development (increase and/or decrease) of the portion of new units or forms over time. The law means that a word's usage is initially low, grows exponentially and finally reaches a stable state. In the same way Polikarpov (1993) developed a word life cycle, and built a theory of the organisation and historical development of language systems as a whole (where each century represents a period). Saussure (1916) introduced the notion of a syntagmatic axis, analogous to the time axis and reflecting a text sequence. Another pillar in the development of sequential text analysis was the use of the probabilistic techniques based on the theory of Markov chains (Markov 1913; Baum and Petrie 1966). The third

pillar that supports analysis of sequential structures in text is the Information Theory which calculates possible arrangements in a sequence (Shannon, 1948). The fourth pillar that lies at the basis of contemporary research into sequential structures in text includes techniques of time-series analysis (Pawłowski, 1997, 1999) based on spectral and ARIMA (Autoregressive Integrated Moving Average) methods (Box and Jenkins 1970; Nurius 1983). The Pawłowski law predicts the autocorrelation of phonemes when a text segment is interpreted as a time series of units with a seasonal lag.

Though it is not common some laws take a linear form. For instance, the Oono law (Oono 1956) states that the ratios among the numbers of all parts of speech stay the same over time in the lexicon of a language, although, typically, the lexicon size grows. Hřebíček (2005) studied the distribution of co-references and noticed a dependency between the number of co-references and the number of sentences with a parameter of text cohesion. Let z be the number of co-references, k the number of sentences, v the number of unique words and n the total number of word forms. Hřebíček supposes that:

$$dz = a \cdot k \quad dw, \tag{1}$$

where w is a text cohesion parameter defined as the relative mean frequency of words ($w = v/n$). He postulates also that

$$dz = a \cdot w \quad dk\,. \tag{2}$$

Experimentally we observed in his sample that z is almost linearly dependant on k as in (2) and the solution should be:

$$Z = 3.4 \cdot k\,. \tag{3}$$

Using synergetic linguistics, Köhler (1999) observed that component dependency upon position does not appear to be a classical power law such as between other parameters (frequency, length, lexicon size, minimization of production effort, minimization of decoding effort, ...). Synergetic linguistics defines complexity as the number of immediate constituents of a construction (syntactic, association, ...). A corpus called Suzanne contains 4621 different types and 90821 occurrences. Complexity is assessed at the word level for a word occurring at a given position (1, 2, ...) inside a syntactic construction; Köhler used a hyper-Pascal distribution to fit the data.

3 Evolution of "content-words" in corpora of biology

In this section, we present a law describing the evolution of single named entities over time and from different corpora of biology.

3.1 Corpora and named-entities extraction

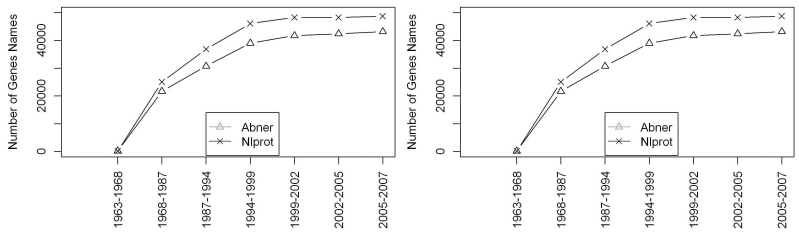
We have created two corpora on molecular biology. This chapter presents the tools we used, dedicated to a named-entity extraction for this kind of domain corpora. Both corpora consist of documents grabbed from the Medline free electronic document base, each document having a title and an abstract (`http://www.ncbi.nlm.nih.gov/pubmed/`). The first corpus, CorpusM, is focused on species of mice and their embryo development. CorpusM contains 34529 documents (titles+abstracts). The second corpus, CorpusH, is focused on the human species with regard to embryos, placenta and cancer. CorpusH contains 77333 documents. The corpora are prepared for the next step (i.e. named-entity extraction) by splitting them into different files containing only sentences, using the tool, lingpipe (Carpenter 2004). After processing CorpusH contains 515500 sentences distributed over 12 time intervals between 1963 and 2007. CorpusM contains 276100 sentences distributed over 7 time intervals.

A main issue to be addressed by text processing in molecular biology is protein and gene name extraction. This is a first step towards more general issues widespread in biology concerning the understanding of gene function and the networks required to make a cell. We used two distinct tools to achieve this task. The first tool is Abner (Settles 2005) which uses machine learning with a training corpus. Its method is based on conditional random-field models. It uses regular-expression formalism but without syntactic and semantic rules. It found 60611 noun phrases in CorpusM, and 82903 noun phrases in CorpusH. The second tool is Nlprot (Mika and Rost 2004) which also uses machine learning with a training corpus. Its method is based on syntactic rules and support vector machine classifiers. It also uses biological dictionaries but no explicit semantic rules. It found 42427 noun phrases in CorpusM, and 48086 noun phrases in CorpusH.

3.2 Experimental results and modelling

We plot by time the number of gene/protein names found (Figure 1). The graphs show a growth in the number of noun phrases extracted for each period, but if we take into account the percentage of abstracts for each period, it becomes constant: 95% of the documents have an abstract after the third time point for CorpusM, and 88% after the fifth time point for CorpusH. The titles

are not explicit enough to contain all protein and gene names described in each paper and induce a bias for content analysis when they are used on their own.



(a) CorpusM *(b)* CorpusH

Figure 1: Number of protein/gene names

We observe a linearity between the number of named entities and the number of sentences. This linearity occurs when abstracts are well indexed in the database; hence, this occurs after 1994 because gene/protein names occur mainly in abstracts. As we can see in Figure 1 for CorpusM for the same number of sentences we obtain a constant number of gene/protein names extracted. Let S be the number of sentences. Hence, we formulate the relationship:

$$\hat{N}_{\text{CorpusM}} = k_{\text{CorpusM}} \cdot S\,, \tag{4}$$

where $\hat{N}_{\text{CorpusM}}$ is the mean number of named entities for CorpusM, and

$$k_{\text{CorpusM}} = 0.98 \pm 0.07\,.$$

Values for k vary from one extractor to another expressing a dispersion (0.91 for Abner, 1.05 for nlprot). For CorpusH, we can formulate the relationship as follows:

$$\hat{N}_{\text{CorpusM}} = k_{\text{CorpusH}} \cdot S\,, \tag{5}$$

where $\hat{N}_{\text{CorpusH}}$ is the mean number of named entities for CorpusH, and

$$k_{\text{CorpusH}} = 0.81 \pm 0.08\,.$$

Values for k vary from one extractor to another expressing variation (0.73 for Abner, 0.90 for nlprot). We note first that the relationship between $\hat{N}_{\text{CorpusH}}$ or $\hat{N}_{\text{CorpusM}}$ and S is linear. This linearity is not far from what was observed with co-references (3). Second, they are very similar but they differ in their parameter values k_{CorpusH} and k_{CorpusH}.

In Section 4, we present a law describing the evolution of "content-word" associations over time using a corpus of biology.

4 Model of "content-word" dependencies over time

4.1 Corpus and "content-words"

First, we used a corpus of biology based on a type of protein called a "prion" (or prp), which is responsible for mad-cow disease. We call this corpus CorpusP and it consists of 6658 documents. A second corpus of biology is widely focused on plant epidemiology. We call it CorpusE and it consists of 5545 documents. A third corpus, not linked to biology, is linked to recent clustering techniques. We call it CorpusC and it consists of 982 documents. Data (title, abstracts, keywords) have been collected from the Science Citation Index database (`http://isiwebofknowledge.com/`).

We used a tool called Beluga (Turenne and Barbier 2004) to extract what has been defined in Section 2 as complexity. Complexity has been adapted at word level without any syntactic assumption through the clustering idea widely applied to lexical organization. The clustering algorithm is based on a sequential pattern extraction algorithm (Agrawal and Srikant 1994) which catches the co-occurrence of n words s times: s is called support, n words is called a sequential pattern (i.e. a word cluster); a context is a document. The algorithm is not sophisticated enough to catch clusters with a specific distance or similarity but is able to extract all co-occurring combinations of specific strings in a set of contexts. Hence if a pattern of length 5 is found, it means than its subpatterns of length 2, 3 or 4 are also extracted. We exploit this property to extract "content-word" dependencies. As a preprocessing step, Beluga extracts noun phrases and author names from the corpus saving also their positions and the document's ID number in which they occur. For our purposes the noun phrases will serve as "content-words". Beluga can define a set of time points to extract associations for each time point. The granularity of the time scale can be more than one year. As a maximum, it is possible to define 50 intervals, from 1960 to 2010. Sometimes, to take into account the evolution of science, studies prefer to take two years as an elementary time step. For CorpusP this is the set of time intervals, for author names and noun phrases there are nine intervals from 1985 to 2002. For CorpusE we defined a step of one year from 1995 to 2005. For CorpusC 5 intervals were defined from 1991 to 2003. Support for both has been set to 3 for CorpusP, and 2 for CorpusC and CorpusE. Such time scales were defined to get an equi-distribution for documents. For instance, for CorpusC and CorpusE each interval at the beginning contained an average of 200 documents per interval. The association frequency threshold is ideally set at 2 but can be higher to solve computational limits of memory storage. In the next part of this paper we call $T = 1$ the most recent interval, $T = 2$ the immediate successor, etc.

4.2 Experimental results and modelling with "content-words"

Extraction experiments for "content-words" are similar to those presented in the previous subsection. In distributions of patterns over time intervals for a given size of pattern we observe the same results as for the author names distribution:

1. the distribution is irregular,
2. it becomes tight when the size of the pattern increases,
3. the distribution follows the profile of $N = 1$ when the size increases.

Table 1 represents the relevant data.

Table 1

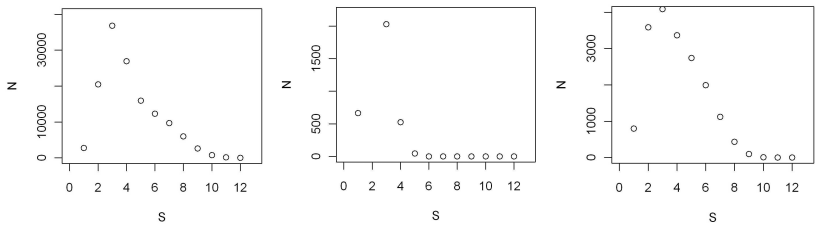
	# Patterns			# Maximal patterns	
Size of patterns	Period 1	Period 2	Period 4	Period 1	Period 4
1	2737	667	796	52	48
2	20506	2444	3586	98	100
3	36807	2031	4089	90	90
4	26960	524	3364	32	68
5	15945	45	2745	20	30
6	12288	0	1992	6	12
7	9688	0	1120	21	28
8	6008	0	432	0	8
9	2655	0	99	0	0
10	790	0	10	0	0
11	143	0	0	0	0
12	12	0	0	0	0

Figure 2 shows the distribution of patterns per size given an interval T. We see that the distribution has local maxima in the same way as observed for author names but the shape is not exactly the same if we look at intervals $T = 1$ and $T = 4$. The distribution has a "bump" towards the tail.

We modelled the asymmetric distribution with a mixed distribution composed of two weighted beta distributions, with property $p+q = 1$, and A being a normalization constant:

$$y = A\left[p \cdot \frac{1}{B(\alpha_1,\beta_1)} x^{\alpha_1-1}(1-x)^{\beta_1-1} + q \cdot \frac{1}{B(\alpha_1,\beta_1)} x^{\alpha_2-1}(1-x)^{\beta_2-1}\right] . \quad (6)$$

Our proposed model fits well the data for both periods $T = 1$ and $T = 4$. The datasets represented in Table 1 and illustrated in Figure 2,) had were transformed as follows: $X = x/12, Y = (y - min(y) + 0.001)/(max(y) - min(y) + 0.002)$. Using the package MASS (function fitdistr) (*R*-Project, 2004) for learning parameters we find the following values: for $T = 1$; $p = 0.78$, $A = 0.32$;



(a) $T = 1$ *(b)* $T = 2$ *(c)* $T = 4$

Figure 2: Distribution of patterns per size given a period T for CorpusP

$\alpha_1 = 2.6$, $\beta_1 = 10.5$, $\alpha_2 = 8.5$, $\beta_2 = 9$. For $T = 4$ the parameters are $p = 0.86$, $A = 0.29$; $\alpha_1 = -1.6$, $\beta_1 = 0.5$, $\alpha_2 = -0.7$, $\beta_2 = 0.25$; and for $T = 2$, $p = 1$, $A = 0.18$; $\alpha_1 = 3$, $\beta_1 = 18$.

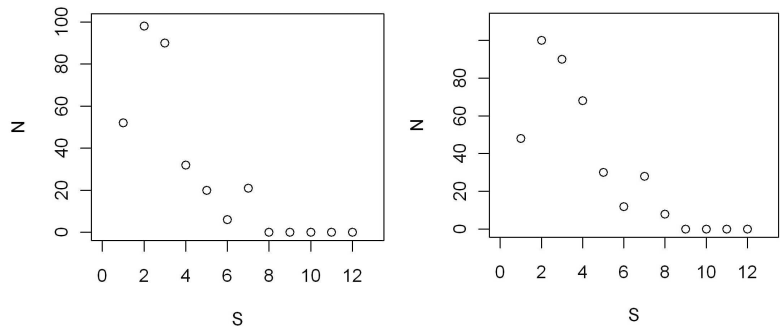
When q is zero, the data are only modelled by one beta distribution. It seems that the choice of one of the values is unpredictable and comes typically from the domain described by the corpus. Some theoretical assumption could justify the choice of a beta distribution to fit our empirical data better than a normal or a log-normal shape. If the dependent variable y (in our case the number of patterns) is a dynamic function of the independent one in the way that x (in our case the size of a pattern) causes a (relative) change in y, and if this change is inversely proportional to some constant α on the one hand and is at the same time limited by another inverse proportion determined by two constants β and c, then we obtain

$$\frac{dy}{y} = \frac{\alpha}{x} - \frac{\beta}{c-x},$$

where dy/y is the first derivative of y divided by the current value of y (hence the relative change). β can be interpreted as a "social diversity" factor, and α as a "topic attractivity" factor. α activates y and β inhibits y. We observed that $\beta > \alpha$. Solving this differential equation yields $y = x^{\alpha} \cdot (c-x)^{\beta}$, which is a function, and by norming it per division by $B(\alpha+1, \beta+1)$ and setting $c = 1$ one gets a beta distribution. This relation varies over time. We observe over time a merging process that can lead to the fact that groups of size n will merge to produce "content-word" groups of size $2n$. A feature of the mad-cow crisis was the Nobel Prize in biology awarded to Prusiner in 1996, period $T = 2$ (1996–1997). Unfortunately this domain event occurred between $T = 1$ (1998–2002) and $T = 4$ (1989–1994) where local maxima occur and does not make an impact on the size of the patterns. An external factor, e.g. a change to terminology by different authors, could explain such phenomena. We observe that for $T = 5$ (1993–1994) for author names, the average size of the pattern becomes large; it corresponds to the period of $T = 4$ for "content-words". But

this observation is only a partial explanation because it is not reproduced for $T = 1$ for author names.

As a pattern of length 3 is composed of three patterns of length 2, we could imagine that for a given prolific period of noun-phrase association and new emerging concepts, patterns of length 5 or 6 produce lots of smaller patterns. Let us call a *maximal pattern* a pattern that does not belong to a pattern of larger length. Figure 3 shows the same distributions as in figure 2 with $T = 1$ and $T = 4$. We still can observe a double-headed distribution which seems typical of the domain. We made complementary experiments with other text collections. It seems that a double-headed distribution does not appear systematically (even when making variations of the time interval with a time step of one or two years). For CorpusE the distribution of patterns by length lets us understand that the average can vary from one period to another but the distribution can be modelled easily with α_1 =1. If we process a small corpus from a field different from biology, CorpusC, the result, in this case, is the same.



(a) $T = 1$ *(b)* $T = 4$

Figure 3: Distribution of maximal patterns by size given a period T for CorpusP

5 Conclusion

Processing domain corpora, here molecular biology, and analyzing the distribution of "content-words" provides evidence of a specific distribution which does not correspond to already known laws discovered about diachronic phenomena. For *Single "Content-Words"* we found that a linear-shape law explains the regular presence of noun-phrases per sentence. But single noun-phrases do not provide real semantic information about the content of a corpus. The most representative factor is the association of "content-words" with their usage contexts. For the *association of "Content-Words"* we found that the distribution of

association size over time can be, sometimes but not always, a mixed distribution of small and double-size associations. Stable distributions require a hypothetical framework to achieve the status of being laws, and further studies need to be conducted on the contexts of distributions to embed them within more theoretical hypotheses such as in a synergetic framework. As complementary validation, the normalization of the data (ratio per number of documents) could improve the stability of proposed models.

References

Agrawal, R.; Srikant, R.
1994 "Fast Algorithms for Mining Association Rules". In: Bocca, J. (ed.), *Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile*. Hove, East Sussex: Morgan Kaufmann, 487–499.

Baayen, R.H.; Renouf, A.
1996 "Chronicling the Times: Productive Lexical Innovations in an English Newspaper", in: *Language*, 72; 69–96.

Baum, L.E.; Petrie, T.
1966 "Statistical Inference for Probabilistic Functions of Finite State Markov Chains", in: *The Annals of Mathematical Statistics*, 37/6; 1554–1563.

Beőthy, E.; Altmann G.
1984 "Semantic diversification of Hungarian verbal prefixes". In: Rothe, U. (ed.), *Glottometrika 7*. Bochum: Brockmeyer, 45–56.

Box, G.E.P.; Jenkins, G.M.
1970 *Time series analysis: forecasting and control.* San Francisco: Holden-Day.

Carpenter, R.
2004 "Phrasal Queries with LingPipe and Lucene." In: Voorhees, E.M.; Buckland, L.P. (eds.), *Proceedings of the 13th Meeting of the Text Retrieval Conference (TREC).* Gaithersburg, Maryland.

Dubois, J.
1962 *Étude sur la dérivation suffixale en Français moderne et contemporain.* Paris: Larousse.

Embleton, S.
1986 *Statistics in Historical Linguistics.* Bochum: Brockmeyer.

Hřebíček, L.
2005 "Text Laws." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 348–361.

Jespersen, O.
1929 *Growth and structure of the English Language.* New York.

Köhler, R.
1999 "Syntactic structures. Properties and interrelations", in: *Journal of Quantitative Linguistics*, 6; 46–57.

Kroeber, A.L.; Chrétien, C.D.
1937 "Quantitative Classification of Indo-European Languages", in: *Language*, 13/2; 83–103.

Labov, W.
1980 *Locating Language in Time and Space.* New York: Academic Press.

Markov, A.A.
1913 "Primer statističeskago izsledovanija nad tekstom 'Evgenija Onegina'

illjustrirujuščij svjaz' ispytanij v cep"', in: *Izvestija Imperatorskoj Akademii Nauk / Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg*, ser. VI/7/3; 153–162.

Mika, S.; Rost, B.
2004 "NLProt: extracting protein names and sequences from papers", in: *Nucleic Acids Research*, 32; W634–W637.

Neuhaus, H.J.
1973 "Zur Theorie der Produktivitat von Wortbildungssystemen". In: Cate, A.P.; Jordens, P. (eds.), *Linguistische Perspektiven. Referate des VII Linguistischen Kolloquiums Nijmegen 1972.* Tubingen: Niemeyer, 305–317.

Nurius, P.S.
1983 "Methodological Observations on Applied Behavioral Science", in: *The Journal of Applied Behavioral Science*, 19/3; 215–228.

Oono, S.
1956 "Kihon-goi ni kansuru ni-san no kenkyuu [= Studies on the basic vocabulary of Japanese: In the Japanese classical literature]", in: *Kokugogaku*, 24; 34–46.

Pawłowski, A.
1997 "Time-Series Analysis in Linguistics. Application of the ARIMA Method to Some Cases of spoken Polish", in: *Journal of Quantitative Linguistics*, 4; 203–221.

Pawłowski, A.
1999 "Language in the line vs language in the mass: On the efficiency of sequential modelling in the analysis of rhythm", in: *Journal of Quantitative Linguistics*, 6/1; 70–77.

Piotrowskij, R.G.
1968 *Informacionnye izmerenija jazyka.* Leningrad: Nauka.

Polikarpov, A.
1993 "A Model of the Word Life Cycle." In: Köhler, R.; Rieger, B.B. (eds.), *Contributions to Quantitative Linguistics.* Dordrecht: Kluwer, 53–63.

R development core team
2004 "*R*: A Language and Environment for Statistical Computing". Vienna: R Foundation for Statistical Computing.

Saussure, F. de
1959 *Course in General Linguistics.* Glasgow: Fontana/Collins.

Settles, B.
2005 "ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text", in: *Bioinformatics*, 21/14; 3191–3192.

Shannon, C.
1948 "The mathematical theory of communication", in: *Bell System Technical Journal*, 27; 379–423.

Swadesh, M.
1952 "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos", in: *Proceedings of the American Philosophical Society*, 96; 452–463.

Swadesh, M.
1955 "Towards greater accuracy in lexicostatistic dating", in: *International Journal of American Linguistics*, 21; 121–137.
Turenne, N.; Barbier, M.
2004 "BELUGA: un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine. Première application au cas des maladies à prions." In: Hébrail, G.; Lebart, L. (eds.), *Proceedings of extraction et gestion de connaissances.* Clermont-Ferrand, 423–428.
Zipf, G.K.
1929 "Relative frequency as a determinant of phonetic change", in: *Harvard studies in classical philology*, 40; 1–95.
Zipf, G.K.
1932 *Selected studies of the principle of relative frequency in language.* Cambridge/Mass.: Harvard University Press.
Zipf, G.K.
1935 *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin.
1945 "The meaning-frequency relaltionship of words", in: *The Journal of General Psychology*, 33; 251–256.

Subject index

Author index

T

U

V

W

Y

Z

Authors' addresses

Andreev, Sergej
Smolensk State University
Foreign Languages Department
RUS–214000 Smolensk, ul. Prževalskogo 4, Russian Federation
email: smol.an@mail.ru

Buk, Solomija
Ivan Franko National University of Lviv
Department for General Linguistics
UA–79000 Lviv, 1 Universytetska St., Ukraine
email: solomija@gmail.com

Čech, Radek
University of Ostrava
Department of Czech language
CZ–701 03 Ostrava, Reální 5, Czech Republic
email: radek.cech@osu.cz

Dębowski, Łukasz
Polish Academy of Sciences
Institute of Computer Sciences
PL–01237 Warszawa, ul. J.K. Ordona 21, Poland
email: ldebowski@ipipan.waw.pl

Đuraš, Gordana
Joanneum Research
Institut für Angewandte Statistik und Systemanalyse
A–8010 Graz, Steyrergasse 17-19, Austria
email: gordana.djuras@joanneum.at

Embleton, Sheila
York University
VP Academic & Provost
M3J 1P3 Toronto, 4700 Keele Street, Canada
email: embleton@yorku.ca

Grzybek, Peter
Universität Graz
Institut für Slawistik
A–8010 Graz, Merangasse 70, Austria
email: peter.grzybek@uni-graz.at

Humenchyk, Olha
Ivan Franko National University of Lviv
Department for General Linguistics
UA–79000 Lviv, 1 Universytetska St., Ukraine
email: see: Buk, Solomija

Kelih, Emmerich
Universität Graz
Institut für Slawistik
A–8010 Graz, Merangasse 70, Austria
email: emmerich.kelih@uni-graz.at

Köhler, Reinhard
Universität Trier
Linguistische Datenverarbeitung
D–54296 Trier, Universitätsring 15, Germany
email: koehler@uni-trier.de

Kravcova, Anastasija
Tomsk State University
Computer Science Department
RUS–636037 Seversk, ul. Kalinina 133, kv. 159, Russian Federation
email: askravtsova@gmail.com

Králík, Jan
Czech Academy of Sciences
Czech Language Institute
CZ–11851 Praha, Letenská 4, Czech Republic
email: kralik@ujc.cas.cz

Krstev, Cvetana
University of Belgrade
Faculty of Philology
RS–11000 Belgrade, Studentski trg 3, Serbia
email: cvetana@matf.bg.ac.rs

Leal, Jerónimo
Pontifical University of the Holy Cross
Faculty of Theology
I–00186 Roma, Via dei Farnesi 82, Italy
email: jleal@pusc.it

Loiseau, Sylvain
Université Paris Ouest
Modyco Laboratory
CNRS
F–92001 Nanterre Cedex, 200 avenue de la République, France
email: sloiseau@u-paris10.fr

Mačutek, Ján
Universität Graz
Institut für Slawistik
A–8010 Graz, Merangasse 70, Austria
email: jmacutek@yahoo.com

Mal'tseva, Lilija
Ivan Franko National University of Lviv
Department for General Linguistics
UA–79000 Lviv, 1 Universytetska St., Ukraine
email: see: Buk, Solomija

Martynenko, Gregory
St. Petersburg State University
Philological faculty, Department of Mathematical Linguistics
RUS–199034 St. Petersburg, Universitetskaya nab. 11, Russia
email: g.martynenko@gmail.com

Maspero, Giulio
Pontifical University of the Holy Cross
Faculty of Theology
I–00186 Roma, Via dei Farnesi 82, Italy
email: maspero@pusc.it

Naumann, Sven
Universität Trier
Linguistische Datenverarbeitung
D–54296 Trier, Universitätsring 15, Germany
email: naumsven@uni-trier.de

Obradović, Ivan
University of Belgrade
Faculty of Mining and Geology
RS–11000 Belgrade, Đušina 7, Serbia
email: ivano@rgf.bg.ac.rs

Obuljen, Aljoša
University of Belgrade
Faculty of Mathematics
RS–11000 Belgrade, Studentski trg 16, Serbia
email: aobuljen@matf.bg.ac.rs

Poddubnyj, Vasilij
Tomsk State University
Computer Science Department
RUS–634 012 Tomsk, Kievskaya Street 86-B, Apt. 16, Russian Federation
email: pvv@inet.tsu.ru

Radulović, Vanja
University of Belgrade
Faculty of Philology
RS–11000 Belgrade, Studentski trg 3, Serbia
email: vanja.radulovic@gmail.com

Rovenchak, Andrij
Ivan Franko National University of Lviv
Department for Theoretical Physics
UA–79005 Lviv, 12 Drahomanova Street, Ukraine
email: andrij@ktf.franko.lviv.ua

Sanada, Haruko
Saitama Gakuen University
Faculty of Humanity
Department of Human Cultures
333-0831 Saitama, 1510 Kizoro, Kawaguchi-shi, Japan
email: h_sanada@nifty.com

Sherstinova, Tatiana
St. Petersburg State University
Philological Faculty
Laboratory of Experimental Phonetics
RUS–199 034 St. Petersburg, Universitetskaya nab. 11, Russian Federation
email: sherstinova@gmail.com

Shevelyov, Oleg
Tomsk State University
1’, 2”, Barcelona, Carrer de Muntaner 250, Spain
email: oshevelyov@gmail.com

Stadlober, Ernst
Graz University of Technology
Institute for Statistics
A–8010 Graz, Münzgrabenstrasse 1, Austria
email: e.stadlober@tugraz.at

Tanomura, Tadaharu
Osaka University
Graduate School of Letters
560-8532 Toyonaka, Machikaneyamacho 1-5, Japan
email: tanomura@let.osaka-u.ac.jp

Turenne, Nicolas
INRA
Unité Mathématique Informatique et Génome UR 1077
F–78350 Jouy-en-Josas, Domaine de Vilvert, France
email: nicolas.turenne@jouy.inra.fr

Uritescu, Dorin
York University
Glendon College
French Studies / Linguistics and Language Studies Programme
M4N 3M6 Toronto, 2275 Bayview Avenue, Canada
email: dorinu@yorku.ca

Vitas, Duško
University of Belgrade
Faculty of Mathematics
RS–11000 Belgrade, Studentski trg 16, Serbia
email: vitas@matf.bg.ac.rs

Vydrin, Valentin
Museum of Anthropology and Ethnography
Russian Academy of Sciences, Saint Petersburg
RUS–199 034 St. Petersburg, Universitetskaya naberezhnaja 3, Russian Federation
email: vydrine@gmail.com

Wheeler, Eric
York University
School of Information Technology
L3P 2A5 Markham, Ontario, 33 Peter Street, Canada
email: wheeler@ericwheeler.ca