Studies
in Quantitative Linguistics
5

Issues

in

Quantitative Linguistics

edited by

Reinhard Köhler

RAM - Verlag

Issues in Quantitative Linguistics

edited by

Reinhard Köhler

2009
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

The present volume is a collection of recent papers on diverse linguistic and textological topics but with a common epistemological and methodological background. They contribute to the field of quantitative linguistics either by results of the application of quantitative approaches to interesting problems or by presenting new ideas and methods. A number of these contributions are versions of papers presented on occasion of the 5th Symposium on Quantitative Linguistics in Trier, Germany, December 2007.

Two of the papers are devoted to research in the field of stylistics. *Sergey Andreev* presents an investigation of an author's (Lermontov's) style with the emphasis on its development over the years during his short life (1814-1841). 35 texts including 25 poems were selected as empirical material; characteristics from several levels of linguistic analysis (morphology, syntax, rhythm, rhyme) serve as style indicators. Andreev arrives at the conclusion that two main periods in Lermontov's life can be determined, and central phases can be differentiated from peripheral ones when the individual texts are attributed to the periods. On data from five Modern Greek novels written by four authors, *George Mikros* conducts an authorship attribution experiment comparing different sets of stylometric characteristics. Besides common indicators, Mikros uses the most frequent functions words and the most distinctive author-specific words. His results yield convincing superiority assessments.

The application of Multidimensional Scaling (MDS) to geolinguistic data, as presented and illustrated by *Sheila Embleton, Dorin Uritescu, and Eric Wheeler* allows, when integrated into a software package and a corresponding data-base to select, search, count, view, edit, and analyse the data according to the researcher's interest. MDS, one of the statistical methods to reduce the number of dimensions of multidimensional data (in this case to just two dimensions), was implemented by the authors in their Romanian Online Dialect Atlas. Their presentation of their MDS function, which can be used for conveying an overview of the linguistic distances among locations with related dialects, gives the reader an impression of the explorative power of the approach. Another paper on a geolinguistic topic is the one presented by *Thomas Zastrow* and *Erhard Hinrichs*. They compare two approaches to computational dialectometry, which they characterize as an information theoretic approach and a vector-based one, on a Bulgarian data set. They, too, illustrate their work and show that both methods yield the same results, thus corroborating the approaches in an impressive way.

Slavic letter frequencies form the topic of *Peter Grzybek's, Emmerich Kelih's, and Ernst Stadlober's* research, which systematically corroborates the hypothesis that these frequencies are distributed according to the negative hypergeometric distribution (NHG). A surprising result of the comparative studies on data from five Slavic languages is the dependency of the NHG parameters on

language-specific factors as well as on interlingual ones. The authors are able to single out individual factors and to show their influence on parameter behavior.

Quantitative studies in linguistics are almost exclusively based on a ”bag-of-words” model, i.e. they disregard the syntagmatic dimension, the arrangement of units in higher units or on higher levels and in the course of the given text. The paper contributed by *Reinhard Köhler and Sven Naumann* shows how motifs, the recently introduced sequences of linguistic features, can be used for the analysis of texts also on the basis of clause properties. A second aim of this paper is the development of an algorithm for the automatic identification and segmentation of clauses in German sentences as a prerequisite for the study of linguistic mass data on this level. Another study on linguistics motifs is contributed by *Ján Mačutek*. He devotes his paper to the aspect of motif richness in analogy to vocabulary richness, a very popular problem in some branches of QL, based on word lengths motifs with length measured in terms of the number of syllables. The data have been taken from two Slovak texts.

An experiment is reported by *Adam Pawłowski, Maciej Piasecki, and Bartosz Broda*. They compared Michael Fleischer’s word profiles – collective symbols distilled from surveys – to profiles generated by automatic extraction from a corpus. The project explores in how much the results of a distributional extraction from text data match with semantic information given by human subjects as obtained in surveys and word priming experiments.

Two papers are devoted to research in the area of morphology. *Olga Pustylnikov and Karin Schneider-Wiejowski* address the phenomenon of productivity in derivational morphology from the point of view of its quantification. They evaluate three quantitative approaches proposed in the literature to measure productivity of German noun suffixes. In addition, they apply a decomposition algorithm used in a multi-agent simulation to identify productive suffixes. As opposed to most other studies on morphological productivity, the authors enclose in their empirical material written texts as well as oral speech. *Petra Steiner* scrutinizes an aspect of inflectional morphology. She deduced, in analogy to models of semantic diversification known from G. Altmann’s works, hypotheses for the distribution of the complexity of inflectional paradigms and tests them with four different measures on data from the Icelandic language. *Relja Vulanović* investigates another aspect of grammar, viz. properties of parts-of-speech systems. Flexible parts-of-speech systems are analyzed from the point of view of grammar efficiency. Seventeen linguistic structures are considered, most of them corresponding to natural languages described in typological samples. Vulanović shows that grammar efficiency of natural languages is well below the theoretically possible maximum.

The diachronic perspective is reflected in *Shoichi Yokoyama’s* and *Haruko Sanada’s* paper on language change. They introduce the models of language change known from QL research (Altmann’s Piotrowski Law) and illustrate them on hypothetical data. Their specific point of view as presented in the paper is a psychological view on the mechanisms behind the process, i.e. they assume an

intra-personal variable as a critical factor which determines the dynamics of the phenomenon.

Jan Králík's "contemplation" discusses the concept of infinity from different points of view. This discussion forms the background of his methodological and epistemological argumentation around the question as to if, when and in how far text and corpus studies can be compared to each other. Arguments from the theory of probability as well as theoretical and empirical findings in quantitative linguistics are taken into account.

I would like to thank the contributors for their co-operation; special thanks are due to Gabriel Altmann for his invaluable support and critical reviews.

Trier, December 2009 RK

Contents

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 1-9.*

Lermontov: Dynamics of Style

Sergey Andreev

Introduction

One of the possible approaches to the study of the individual style of an author is the study of its evolution. Data about whether there are any statistically relevant changes in the characteristics of texts, written by the same author, but at different time, is not purely a theoretical question, but may also have practical applications in the sphere of text attribution, in forensic linguistics, for literary criticism, etc.

The subject of the study in this paper is the search for linguistic markers of the evolution of style of the Russian poet Mikhail Lermontov. Lermontov is one of the most famous authors in Russia. Though Lermontov lived a very short life (1814–1841), he created masterpieces in poetry and also in prose and drama. The poet's lyrics are almost unanimously divided by literary critics into two main groups, depending on the time when the poems were written: (1) during early or (2) during 'versatile' periods of Lermontov's creative activity. The year of division of Lermontov's creative life into these periods is 1837 when he wrote the poem *On the death of the poet*. This poem masterly depicted the public feelings of sorrow and indignation aroused by the death of the famous Russian poet A.S.Pushkin after a duel. The poem was copied again and again, passed from hand to hand. Lermontov whose name had been known only in a narrow circle of the poet's friends, became famous.

Though the poem caused serious difficulties in the life of Lermontov – he was arrested and then sent in exile to the Army acting in the Caucasus – it marked the start of a highly successful period of creative activity of the author. Literary critics stress the fact that starting from this time his manner and style had changed: earlier verses had many instances of imitation of other poets, and his later works are characterized by specific features peculiar to Lermontov only.

Even if literary peculiarities of lyrics written during both periods have been studied, linguistic properties of his texts, differentiating the periods, still need a more systematic analysis. This paper describes a possible approach to finding out the markers of such differences.

Material

The study of the dynamics of style of an author, to some extent, can be compared to longitudinal studies in psychology where the same individuals are tested dur-

ing their life at certain intervals to see if there are any variations in individual psychological properties.

If longitudinal experiments are rather difficult to organize because of life circumstances (illnesses, removal from the place of residence, other events in the life of both the experimental group members and the researchers), the 'diachronic' analysis of an individual style seems easier to conduct. It is enough to find proper texts, belonging to the same author, written by him at different periods of his literary career, and then compare the texts, using quantitative and qualitative analysis.

However, such a "longitudinal" analysis of individual style conceals certain difficulties, too. They consist in the first place in the selection of the material for the analysis. The problem often is what is to be considered as the same text and what as a new one. Thus in 1831 Lermontov wrote a poem in the album of E.A.Sushkova (Khvostova) *«Ya ne lublu tebya, strastej…» (I do not love you, passions…).* Later in 1834, when she recited this poem in his presence, he persuaded her to give him the text back for improvement. In 1837 he changed several lines in it and later in 1840 published the poem in his collection of works, but now with a different title – *«Rasstalis' mi, no tvoj portret…» (We parted, but your portrait…).*

In the present study we consider different editions of a poem as the same text if these variants coincide in more than 90 per cent of the lines. The change of the title of the text in such cases is not taken into account.

Another problem consists in choosing the date of the origin of poems – the actual date when the poem was written or the date of its first publication. There is even a third option – to accept the date when it became generally known. The latter can be traced by publications in anthologies and, of course, is less accurate. In other words, the question, which the investigation is faced with, is whether we take into account actual changes in the creative work of an author or the evaluation of the dynamics of his style as seen by the reading public. In case of the second alternative the study would reveal a virtual picture of the evolution – the effect that the poems exerted on the reading public and literary norms of the time rather than the change of the style itself.

In this research we accept the first option of the above-mentioned: we take into account the actual date when the poem was written. Besides the theoretical basis in favor of this decision, there exists a practical reason to accept this approach. Many of the poems by Lermontov, belonging to the first period of his creative activity, became known from manuscripts and albums, but were published much later, often after the death of the poet. Hence the time of the publications in this case would be to some extent misleading.

Chronological data was taken from the commentaries by Iraklij Andronnikov (1975).

Lermontov used different meters, especially in the second period. In order to achieve a common basis for the analysis and have homogeneous material we

chose lyrics written in 4-feet iambic meter, not exceeding 60 lines with the same type of stanzas (4-line stanzas or bigger stanzas divisible by 4). As a result, 35 texts which corresponded to the criteria were chosen for the analysis. They include 25 poems, belonging to the first period and 10 to the second.

All 25 lyrics of the first period were not published by Lermontov himself who did not include them into his single author's collection of works in 1840, though later mostly after his death they were published and many of them are now considered as masterpieces.

Characteristics

The parameters to be used during the analysis should be linguistically and poetically relevant, expressing important properties of verse texts. Verse is characterized by a very complicated organization of elements and strong interaction of different linguistic levels, forming integrity both vertically (between different lines within the same stanza and between different stanzas) and horizontally (between the initial and final elements of the line). That is why it seems reasonable to use characteristics, reflecting the features of different linguistic levels. This is in line with the studies aimed at solving problems, similar to ours, of automatic text classifications and text attribution (Gurney and Gurney 1998; Mikros 2006; Mikros and Carayannis 2000; Rudman 1998; Holmes 1994; Stamatatos, Fakotakis and Kokkinakis 2001; Kelih, Antić, Grzybek and Stadlober 2005). The difference of our approach from the above mentioned ones is that in our research we are interested in finding out not only characteristics which differentiate the groups of texts, but also pay attention to those which are integral for different stages of the author's creative activity.

Choosing characteristics we were guided by the following criteria. Characteristics must be well distinguished formally; they should not be correlated with one another (or, at least, the correlation should not be very high); their number must be less than the number of objects (texts). Besides, of course, they must reflect important features of lyrical poems, be poetically relevant.

The list of characteristics, which we used in this study, is based on the results of a number of previous studies, conducted in the sphere of style evolution analysis (Andreev 2003; Andreev 2002, 2008; Baevskij 1993).

Rhythmic features. We take into account the absence of stress on the so-called strong positions. Strong positions, or ictuses, in syllabo-tonic verse system are syllables on which according to the metric scheme a stress should fall. For iamb strong positions are even syllables.

In the following examples the stress is omitted on, correspondingly, the first, the second and the third ictuses. Examples are taken not from the poems by Lermontov but from American poetry to avoid the problem of Cyrillic orthography and translation of verses.

Wild was the day; the wintry sea <...>
(Bryant *The Twenty-second of December*).

The prosperous and beautiful <...>
(Emerson *The Park*).

They brought me rubies from the mine <...>
(Emerson *Rubies*).

Morphological parameters are represented in terms of traditional morphological classes (noun, verb, adjective, adverb, participle and pronoun). We counted the number of words belonging to these classes, occupying the last strong position in the line.

Syntactic properties are revealed by the following characteristics: the number of lines, which end in exclamation marks or question marks ("emotional ending"), the number of enjambements and syntactic pauses.

Enjambement takes place when a clause is not finished at the end of the line and continues on the other line:

Rhodora! If the sages ask thee why
This charm is wasted on the earth and sky <...>
(Emerson *The Rhodora*)

Pause is a break in the line, caused by a subordinate clause or another sentence:

True – time will seam and blanch my brow –
Well – I shall with aged men <...>
(Bryant *The Lapse of Time*)

Characteristics of **rhyme** reflect the number of masculine and feminine rhymes.

All 35 texts were analyzed with the help of this list of characteristics. The values showing the occurrence of every characteristic in texts were normalised over the size of these texts in lines. These data were used in discriminant analysis.

Discriminant analysis is a procedure whose purpose is to find characteristics, discriminating between naturally occurring (or a priori formed) classes, or to classify into these classes separate (unique) cases which are often doubtful and "borderline" (Klecka 1989; Warner 2008: 650-701). This method has been successfully used by linguists who are solving attribution problems (Mikros 2006; Mikros and Carayannis 2000; Tambouratzis et al. 2004).

At the first stage of our analysis the aim was to see if there are any variables which differentiate between two groups of texts and how these variables contribute to the discrimination.

As a result of the discriminant analysis the following characteristics from the above-mentioned list were included into the discriminant model: the number of nouns in the last strong position (in rhyme); the number of verbs in the last strong position (in rhyme); the omission of the stress on the third ictus; emotional ending.

Characteristics of texts are variables in the obtained discriminant function. In order to find out and compare the contributions of these variables, their standardized coefficients were calculated. Standardized coefficients of characteristics, relevant for the discrimination between the periods, are given in Table 1.

Table 1
Standardized Coefficients for Canonical Variables

Characteristics	Root 1
noun in the last strong position	-0.93
omission of stress on ictus 3	-0.58
emotional ending	-0.37
verb in the last strong position	0.31

The magnitude and sign of the coefficients of characteristics in Table 1 show us the association of each variable with the value of canonical discriminant function. The biggest input is made by the number of nouns in the last strong position (rhyme). It is followed in force by omission of stress on the 3rd ictus, emphatic ending and the number of verbs in the last strong position (rhyme).

In Table 2 means of canonical variables for two classes are given.

Table 2
Means of Canonical Variables

	Root 1
Group 1	0.54
Group 2	-1.34

Judging by the values and sign of the coefficients in Table 1 and the data in Table 2 we can state how the parameters are associated with group membership of the texts (Klecka 1989: 100–104; Warner 2008: 651–660). Three first variables in Table 1 are more characteristic of the second period of Lermontov's creative activity, the last characteristic (verb in the last strong position) of the first period.

It is interesting to note that enjambement appears to be stable during the whole creative activity of Lermontov, whereas this characteristic as a rule has a marked discriminant force when the styles of different poets are compared (Andreev 2002). On the other hand, the number of lines with emotional endings increased. In general this corresponds to the opinion that Lermontov's poetry is highly emotional (at the expense of semantic transparency). But the fact that this is a factor, differentiating two periods, is rather unexpected. It was possible to suppose that expressive syntax would be at least as important in the first period as in the second.

The opposition of verbs and nouns in the last strong position (rhymed) is very indicative. Verbs are certainly easier to rhyme. There exist a limited number of grammatical verbal suffixes and inflexions in Russian, which helps an author to make the so-called grammatical or morphological rhymes. The fact that Lermontov used many verbal rhymes during Period 1 may be accounted for by his literary inexperience at the beginning of his career as a poet. Still, another explanation is also possible. The increase of the number of nouns in rhyme and correspondingly the decrease of the number of verbs may be interpreted as transition from dynamic vision to a more static conception of the world.

Speaking about omission of the stress on the third ictus, it is possible to suggest that Lermontov started to prefer longer words at the end. This again, in our opinion, reflects the shift in his poetry to contents in rhyme at the expense of rhythmic formal regularity.

Nucleus and periphery

It is also possible to find out the poems that form the nucleus in both periods and those forming periphery. This is seen from Table 3, which demonstrates squared Mahalanobis distances from group centroids to the texts.

The squared Mahalanobis distance is a measure of distance between two points in multidimensional space. By a group centroid is understood the "mean point" representing the means for all independent variables in the multidimensional space in which each observation (text) was plotted. In our case the space is defined by four characteristics, included into the discriminant model. The smaller is the distance to the group centroid, the closer to the nucleus is the text.

Table 3
Squared Mahalanobis Distances from Group Centroids

Nucleus of Period 1	
Odinitchestvo (Loneliness)	1.50
Nistchij (Beggar)	1.40
"Ya ne dlya angelov i raya…" "Not for the angels and heaven I…"	1.11

"Ti molod. Tsvet tvoih kudrej…" "You are young. The colour of your locks…"	0.55
Gusar (Hussar)	0.87

Periphery of Period 1	
"Neredko ludi i branili…" ("Not rarely people scolded…")	7.87
Glupoj krasavitse (To a silly beauty)	9.91
"Metel' shumit, i sneg valit…" ("The snowfall chides and it snows heavily")	7.17
Parus (The Sail)	11.45

Nucleus of Period 2	
"Rasstalis' mi, no tvoj portret…" ("We parted but your portrait…")	1.45
"Spesha na sever iz daleka…" ("Hurrying to the North from afar…")	0.84
Opravdaniye (Justification)	0.98

Periphery of Period 2	
Vetka Palestini (The Branch from Palestine)	5.61
"Nad bezdnoj adskoyu bluzhdaya..." (Wandering over the abyss of hell…")	7.39
"Proschaj nemitaya Rossiya…" ("Farewell, unwashed Russia…")	4.83

The nucleus of the first period is formed of the texts in which the relations between the poet and the world can be characterized in terms of opposition, what is rather typical of romanticism. The narrator is alone and this is considered as the best choice in the cruel and shallow world. Love and friendship are causes of future suffering. The lyrical hero is described completely alone with no one to grieve at his death.

Lermontov's texts forming the periphery of the first period contain a variety of themes, presenting both a pessimistic picture of the world and optimistic vision of events when people are shown quite happy and joyful, though hostile to the author.

In the texts written during the second period and belonging to the nucleus we observe radical changes in the motives and themes as compared to the nucleus of Period 1. In *"Rasstalis' mi, no tvoj portret…" ("We parted but your portrait…")* the lyrical hero declares that love never ends. In *Opravdaniye (Justification)* love is described as the feeling of which people should be proud. The hero of *"Spesha na sever iz daleka…" ("Hurrying to the North from afar…")* is not willing to live if he does not see his friends. Loneliness so praised by Lermontov during the first period of his creative activity is no longer desired now.

In the texts of the periphery of the second period, like in the periphery of the first part of Lermontov's creative activity, a variety of themes are observed. They range from typical romantic motive of exile in *"Proschaj nemitaya Ros-*

siya…" ("Farewell, unwashed Russia") to praising love, which brings freedom even to prison in *"Nad bezdnoj adskoyu bluzhdaya..." (Wandering over the abyss of hell…")* and to description of victory of faith over time and dangers in *Vetka Palestini (The Branch from Palestine).*

Conclusion

Though the study of changes in the creative activity of authors is certainly not uncommon among literary critics, systematic analysis of *linguistic* differences, occurring in texts, written at different periods of an author's life, has so far attracted much less attention.

The results of this study, conducted on the material of iambic four-line stanza lyrics by Lermontov showed that despite the brevity of the period of the poet's creative work there are certain linguistic characteristics which possess discriminating force and differentiate two main periods of his life.

The analysis allowed us to compare nuclei and peripheries of the texts of both periods. The nucleus of every period is more homogeneous in thematic aspect than the periphery. From the first to the second periods in nuclei texts a change in motives and themes is observed. This change is correlated with changes in the morphological and syntactic structure of verse texts. Nouns forming conceptual basis of the verse text are shifted to the rhymed position and this becomes an important marker, reflecting the development of the author's style. The tendency to omit stresses on the third ictus is another important feature of Lermontov's style evolution.

References

Andreev, S. (2003). Estimation of similarity between poetic texts and their translations by means of discriminant analysis. *Journal of Quantitative Linguistics 10(2), 159-176.*

Andreev, V. (2002). Classification of poetic texts by means of the multivariate analysis. *Minsk state linguistic university bulletin 10, 141-146.*

Andreev, V. (2008). Variation of Style: Diachronic Aspect. *Digital Humanities 2008. Book of Abstracts (University of Oulu, June 24-29): 42-43*. Oulu, Finland.

Andronnikov, I. (1975). Primechanija [Comments]. In: *M.U.Lermontov. Collection of works: 507-618.* Moskva: Hudozhestvennaja Literatura.

Baevskij, V.S. (1993). *Pasternak - Lirik.* [Pasternak the lyric poet]. Smolensk: Trust-Imakom.

Gurney, P.J., Gurney, L.W. (1998). Authorship attribution of the Scriptores Historiae Augustae. *Literary and Linguistic Computing 13(3), 119-31.*

Kelih, E., Antić, G., Grzybek, P., Stadlober, E. (2005). Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs, W. Gaul. (eds.), *Classification – The Ubiquitous Challenge: 498-505.* Heidelberg: Springer.

Klecka, W.R. (1989). *Faktornyj, diskriminantnyj i klasternyj analiz*. [Factor, discriminant and cluster analysis]. Moscow: Finansi i statistika.

Mikros, G. (2006). Authorship Attribution in Modern Greek Newswire Corpora. In: O. Uzuner, Sh. Argamon, J. Karlgren (eds.), *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval*, Seattle, Washington, USA August 10: *43-47.*

Mikros, G., Carayannis, G. (2000). Modern Greek corpus taxonomy. *Proceedings of the 2nd International Conference on Language Resources and Evaluations, Athens, Greece, 31 May-2 June, Vol. 3, 129-134.*

Rudman, J. (1998) The State of Authorship Attribution Studies: Some Problems and Solutions. *Computer and the Humanities 31*, 351–365.

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. *Literary and Linguistic Computing 19(2), 221 – 242.*

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities 28(1), 87-106.*

Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities* 35 (2), 193-214.

Warner R.M. (2008). *Applied Statistics*. Los Angeles – London: Sage Publications.

R. Köhler (ed.) (2009), *Issues in quantitative linguistics: 10-16.*

Data Management and Linguistic Analysis:

Multidimensional scaling applied to Romanian Online Dialect Atlas

Sheila Embleton
Dorin Uritescu
Eric S. Wheeler

Context

In various papers (Embleton, Wheeler 1997a, 1997b, 2000; Embleton, Uritescu, Wheeler 2004, 2006a, 2006b, 2007a, 2008a), we have presented our projects for obtaining digitalized dialect data (from variously English, Finnish and Romanian) and applying the multidimensional scaling (MDS) technique to show the overall dialect picture. The Romanian Online Dialect Atlas (RODA) is now available online (Embleton, Uritescu, Wheeler 2007b) for others to use.

Here, we will briefly reiterate the context for developing RODA, and then focus on the implementation of a built-in MDS function, which allows this technique to be used readily with the Romanian data.

Romanian and Romance

The Romance languages, descendants of Vulgar Latin, divided early into a western and eastern branch. Romanian is the principal language remaining on the eastern branch and shows some interesting differences from the western languages such as French, Spanish or Italian (for example, the Romanian definite article follows the noun instead of preceding it). As such, Romanian is of interest not only to scholars of Romanian but to Romance language scholars in general.

The north west region of Romania, called Crişana, is a conservative area of the Romanian speech community, and it preserves dialect features that are not necessarily found in standard Romanian or other dialect regions.

Romanian Online Dialect Atlas (RODA)

Starting with a hard-copy dialect atlas, currently in two volumes (Stan, Uritescu 1996, 2003) with others to follow, we digitized the basic data. It consists of the several responses to over 400 indirect questions, at 120 locations, and represents

many years of detailed field work in the north west area of Romania. The notation used for recording the responses is rich, with subtle variations in pronunciation recorded, as well as notations on usage (e.g. "used by old people") and respondent's performance (e.g. "hesitation"). However, this rich notation necessitated a more sophisticated data representation, and custom software for presenting, storing and searching the underlying data, as we have discussed in Embleton, Uritescu, Wheeler (2007a).

The objective of the Romanian project was to provide greater scholarly access to the data. With RODA, one is able to:

- Select all or a subset of the data files (each file corresponding to a dialect map in the hardcopy atlas)
- Search for a string pattern in the selected data, specifying context (e.g. all word-final occurrences of /u/ that are syllabic, and not marked as special usage)
    - Count the occurrences of the pattern by location
    - View the results as a list, and revise the results manually
- Present the results as a map
- Create an interpretive map
- Store maps as images for further use
- Hear selected samples of the field data.

Most important for us, it is possible to apply computerized analysis methods to the data. To this end, we have built in a multidimensional scaling function that allows us to summarize a large amount of data as a single picture.

Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) was proposed in Embleton, Wheeler (1997a) as an effective way of showing the dialect relationships among many locations, based on many pieces of data. The technique can be used otherwise, but it is very effective with large data sets. Generally, it is very difficult to visualize the relationship among many dialect locations that differ from one another on many points (i.e. responses to many questions). MDS provides a means of doing this that is visual and easy to interpret. In one picture, every location is represented, according to its dialect distance from every other location.

Technique

The MDS technique begins by calculating the dialect distance between each pair of locations. While there are many ways this calculation could be done, one straight-forward way is to give each point of comparison (i.e. each selected file, representing a question given in the field) equal weight. If two locations agree on

their response, the distance for that point of comparison is zero, and if they disagree the distance is one. The distances are summed over all the points of comparison, giving an overall distance for that pair of locations. In our case, that distance can vary from 0 to over 400, if all the data files are used. See Figure 1.

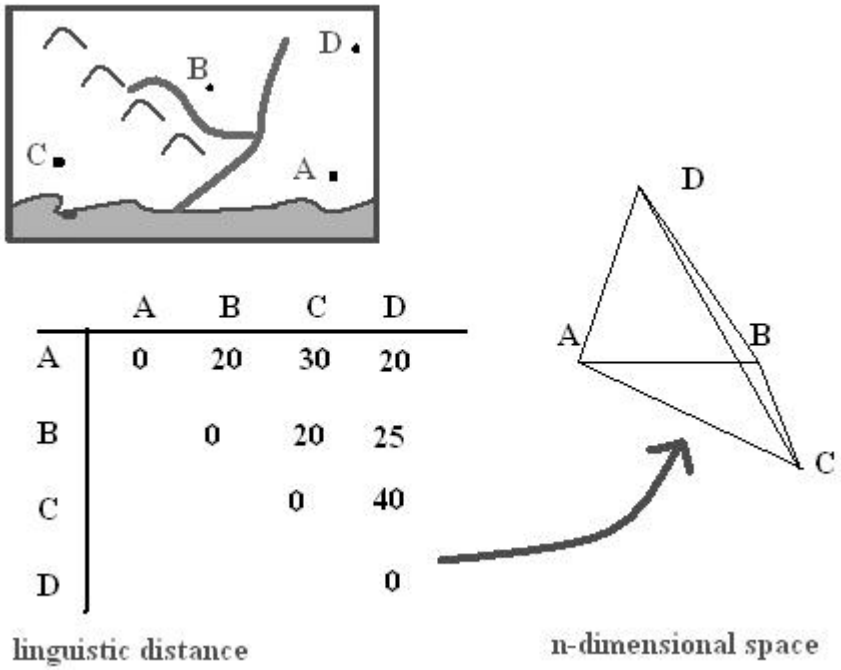


Figure 1. Linguistic distances among n+1 locations requires an n-dimensional space

The heart of the MDS technique comes when we try to place each of the 120 locations at an appropriate distance from every other locations. In general, *n*+1 locations require an n-dimensional space to do this placement exactly. MDS projects this large dimensional space down to 2 dimensions (i.e. a flat map) in a way that minimizes distortion of the original higher space. See Figure 2.

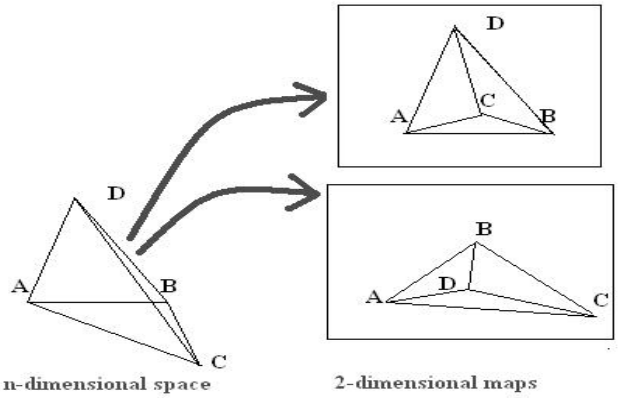


Figure 2. MDS projects an n-dimensional space to 2 dimensions in an optimal way

In the resulting map, generally, distant places are distant in the original data, and close places are either close because they are close in the data or (occasionally) as an artifact of the projection. Think of projecting the shadow of your hand on the wall: if you turn your hand appropriately, the shadow is a good map of the hand, although some parts may seem closer together than they really are.

Our implementation of MDS uses a basic unit of distance that is 1 or 0. If two locations have exactly the same response (in the primary data field, and ignoring any annotations to the data), the distance is zero; otherwise it is one. That means (using some invented data) that "cat" and "ca:t" are different, and just as different as "cat" and "feline". In both cases, the distance is one. One could imagine refining the measure so that more fundamental differences contributed more to the distance measure.

However, we think that over a large set of data (in our case, there is potentially more than 400 comparisons for each location), such subtle differences are not necessary. In a related study (Embleton, Uritescu and Wheeler forthcoming), we found that random selections of the underlying data files had a limited impact on the resulting MDS picture. The MDS picture of all the data was a good representative of the picture produced from (say) 80% or 90% of the data. For practical purposes, the MDS technique seems to be very tolerant of small changes to the underlying data.

Application

When we construct an MDS picture based on a large set of linguistic data, the result may or may not correspond to the original geography. In the case of the English and Finnish data that we analyzed, it was substantially similar to the geography. That is, English northern counties were at the opposite end of the MDS map from the southern counties, and the counties in-between were more or less in geographic order. There were some obvious exceptions: the extreme southwest of England was closer to London linguistically than geographically, but this was a well-recognized aspect of English dialects. There were also counties (e.g. Cambridgeshire) which were widely spread out, and yet most counties were very compact indicating a uniformity within the county as well as a separateness from other counties. In Finland, we found a similar situation, with dialect areas being very compact, and spreading from east to north to west, with a few notable but explainable exceptions.

In north west Romania, however, the situation was not so clear cut. Here, areas that were traditionally considered to be separate dialect areas were not necessarily compact, and most were interleaved with one another. See Figure 3. We discuss this situation and its implications for Dialectology in Embleton, Uritescu and Wheeler (2008b). Briefly, we think that dialect distinctions are present when one considers certain features selectively, but that in the aggregate, there is a unity to the region that is seen in the MDS map.

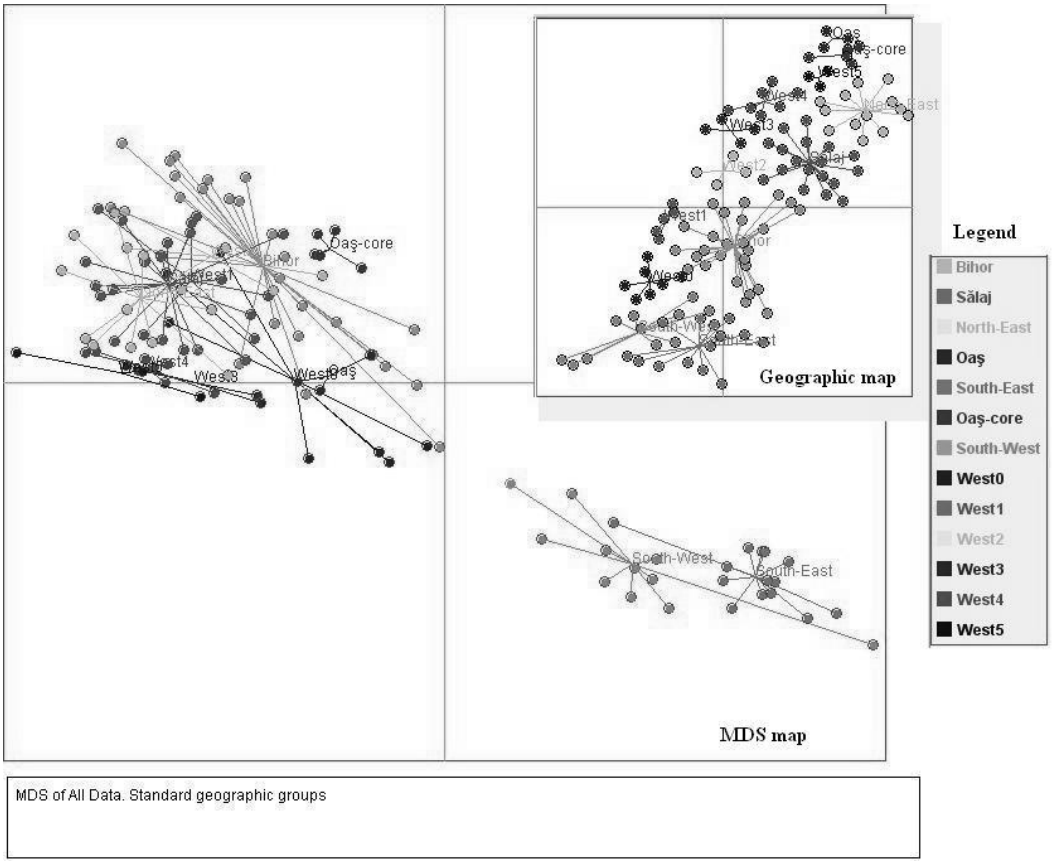


Figure 3. An MDS picture showing most dialect regions overlapping

The importance of the MDS picture, however, is not in its relationship to the original geography. Rather, the MDS picture is one that captures a large amount of data and expresses all the data relationships in a single picture. It is as if one could summarize an entire dialect atlas (or any selected subset of it) in one image.

This top-level summary is one that is inconceivable to create without a technique like MDS and an incredible amount of labour. With the MDS function built into RODA, and the access to digitalized data provide by RODA, it is not only easy to create this picture, but equally easy to experiment with different selections of data, and different geographical groups. As such, one can test out different theories of what defines a dialect.

Summary

RODA provides a large set of digital data on a dialect area of considerable importance to Romance linguistics. RODA makes the data much more accessible than it would be in hard-copy because users can select, search, count, view, edit,

display and analyze data to fit their needs. We have been able to do a number of interesting linguistic studies using the basic functions of RODA, including:

- An analysis of certain Latin words ending in *-um* and *-o*, which have reflexes here with word-final syllabic or non-syllabic *-u*.
- An analysis of non-palatalized dentals before front vowels. They occur widely in our region, although the common assumption that they occur only in one part of the region is reflected in their much more frequent occurrence there.
- The analysis of certain vowel raising and weakening phenomena, which we were able to show happened independently of one another because the patterns overlapped but did not match one another.

More details can be found in Embleton, Uritescu and Wheeler (2006a, 2008a).

Further, the built-in analysis tool for doing MDS gives us an easy way to provide a top-level view of the whole data set. This has given us greater insight into the data and the dialect situation in Romania.

References

Embleton, Sheila, Wheeler, Eric (1997a). Multidimensional Scaling and the SED Data. In: Wolfgang Viereck, Heinrich Ramisch (eds.), *The Computer Developed Linguistic Atlas of England 2*: *5-11*. Tübingen: Max Niemeyer.

Embleton, Sheila, Wheeler, Eric (1997b). Finnish Dialect Atlas for Quantitative Studies. *Journal of Quantitative Linguistics.4, 99-102.*

Embleton, Sheila, Wheeler, Eric (2000). Computerized Dialect Atlas of Finnish: Dealing with Ambiguity. *Journal of Quantitative Linguistics 7. 227-231.*

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2004). Romanian Online Dialect Atlas. An exploration into the management of high volumes of complex knowledge in the social sciences and humanities. *Journal of Quantitative Linguistics 11(3), 183-192.*

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2006a). *Seeing Words Change using the Romanian Online Dialect Atlas*. Presentation to International Linguistics Association. Annual Meeting. Toronto. April 2006.

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2006b). Defining User Access to the Romanian Online Dialect Atlas. Presentation to the 5th Congress of Société Internationale de Dialectologie et Géolinguistique (International Society for Dialectology and Geolinguistics). Braga, Portugal. August 2006. Published in *Dialectologia et Geolinguistica 16, 27-33.*

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2007a). Romanian Online Dialect Atlas: Data Capture and Presentation. In: Peter Grzybek, Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text*. *(Quan-*

titative Linguistics, 62), G. Altmann Festschrift: 87-96. Berlin and New York: Mouton de Gruyter**.**

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2007b). Online Romanian Dialect Atlas. http://vpacademic.yorku.ca/romanian

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2008a). *Digitalized Dialect Studies: North-Western Romanian.* Bucharest: Editura Academiei Române (Romanian Academy)

Embleton, Sheila, Uritescu, Dorin, Wheeler, Eric (2008b). Identifying Dialect Regions: Specific features vs. overall measures using the Romanian Online Dialect Atlas and Multidimensional Scaling. Leeds, UK: *Methods XIII Conference. August 2008.* publication forthcoming.

Embleton, Sheila, Uritescu, Dorin Wheeler, Eric (2009). The Stability of Multidimensional Scaling over Large Data Sets: Evidence from the Digitized Atlas of Finnish. In: Eva Havu, Mervi Helkkula, Ulla Tuomarla (2009), ‚*Mélanges en l'honneur de Juhani Härmä*; 207-214. Mémoires de la Société Néophilologique de Helsinki, ed. May 2009.

Nerbonne, John, Kleiweg, Peter, Manni, Franz, Heeringa, Wilbert (2007). Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. http://www.let.rug.nl/~nerbonne/papers/Nerbonne-Kleiweg-Manni.pdf

Stan, Ionel, Uritescu, Dorin (1996). *Noul Atlas lingvistic român. Crişana.* Vol. I. Bucharest: Academic Press.

Stan, Ionel, Uritescu, Dorin (2003). *Noul Atlas lingvistic român. Crişana.* Vol. II. Bucharest: Academic Press.

Uritescu, Dorin (1984a). Subdialectul crisean. In: V. Rusu (ed.), *Tratat de dialectologie românească: 284-320.* Craiova: Scrisul Românesc.

Uritescu, Dorin (1984b). Graiul din Tara Oasului. In: V. Rusu (ed.), *Tratat de dialectologie românească:* 390-399. Craiova: Scrisul Românesc.

Wheeler, Eric S. (2005). Multidimensional Scaling for Linguistics. In: Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook: 548-553.* Berlin: Walter de Gruyter.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 17-33.*

Slavic Letter Frequencies:
A Common Discrete Model and Regular Parameter Behavior?

Peter Grzybek
Emmerich Kelih
Ernst Stadlober

Letter Frequencies and Frequency Models in the Context of Dynamic and Synergetic Linguistics

In the framework of quantitative approaches to language, so-called "low-level" units of language – e.g. letters, phones, phonemes, etc. – have always played a major role, from the early beginnings of this discipline on. Whereas earlier attempts in this field, which were mainly mere letter or sound statistics and the like, were related not only to linguistic problems, but also to concrete practical or technical issues of different kinds (cf. Grzybek 2006, Grzybek & Kelih 2003), recent studies are much more theory-based and, in fact, theory-oriented. A major reason for this development can be seen in the rise of synergetic linguistics (cf. Köhler 2005); in this context, letters (and other "low-level" units) can be seen as linguistic entities which form, or rather are part of systems, the characteristics and needs of which seem to be quite easy to survey as compared to more complex systems, where one is concerned with multi-faceted needs and multi-level influences. Therefore, it seems likely and reasonable, that any understanding of these allegedly less complex systems will yield deep insight into the dynamic mechanism of linguistic systems, in general; seen from this perspective, the study of letter frequencies clearly goes beyond simple analyses on something like a linguistic playground, and it represents much more than a methodological test case, but is an important and valuable scholarly object in its own right, contributing to a deeper understanding of the dynamics of linguistic systems.

Since general characteristics of letter inventories and frequency organization are of primary relevance here, the specific frequency of individual letters fades into the background. Instead, the question in how far the system-bound organization of a letter frequency distribution underlies general regularities comes to the fore. To this end, the frequency distribution is transformed into a (descending) order, where the frequency of the most frequent letter is assigned to the first rank, and the most infrequent letter to the last rank. The crucial question then concentrates on the point whether the frequencies exhibit a particular relation, or proportion and how these relations can best be described by a theoretical model.

The theoretical background of this procedure has repeatedly been described in recent years (Grzybek & Kelih, 2003; Grzybek, Kelih & Altmann 2004); a redundant description of the method can be abandoned here. The usual assumption in this context is that the probability of a given class with value *x* or rank *r* is proportional to the next lower class (i.e., *x–1* or *r–1*, respectively). Based on this general assumption (cf. Altmann , Köhler 1996) one may establish the difference equation

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

the concrete solution of which depends on the function *g*(*x*). In the past, even relatively simple functions *g*(*x*), usually rational functions, have yielded concincing results for the frequency analysis of linguistic units from different levels. More recently, Wimmer & Altmann (2005, 2006) have generalized this approach, and within this generalization, many distributions relevant for linguistic modeling may be sub-summarized under a common "linguistic roof". Without going into details here, let it suffice to say that this approach has also been successfully applied in systematic analyses of letter frequencies for various Slavic languages.[1]

One major objective of all these studies has been to systematically test previously discussed frequency distribution models with consistent material across different languages. Taking into account different "philosophies" of writing systems, our intention is not to find an overall valid, "universal" model for letter frequencies. Rather, the concentration on different Slavic languages offers the chance to study typologically similar languages from one and the same linguistic family, i.e. languages which share some general common straits, but still display some variation; this might shed light on some factors influencing the system-related behavior of this linguistic level, and one should expect that, given an adequate model common to these languages, relevant changes might result in some interpretable parameter behavior yielding deep insight into the synergetic organization of this level.[2]

Thus far, only selected languages have been analyzed, and the results obtained should be taken with a pinch of salt. Anyway, as a first result, it turned out that most of the models discussed in the past turned out to be inadequate; only

[1] For Russian see Grzybek & Kelih (2003), Grzybek, Kelih & Altmann (2004), Grzybek, Kelih & Altmann (2005a) and Kelih (2007); for Slovak see Grzybek, Kelih & Altmann (2005b) and Grzybek, Kelih, Altmann (2006); for Ukrainian see Grzybek & Kelih (2005a), and for Slovene see Grzybek, Kelih, Stadlober (2006).

[2] In detail, these models are the zeta distribution, the Zipf-Mandelbrot distribution, the geometric distribution, the Good distribution, the Whitworth distribution, the negative hypergeometric distribution.

one model, the negative hypergeometric (NHG) distribution, could be shown to suitable for letter frequencies of the languages studied thus far.[3]

As compared to all other distribution models, the NHG distribution – which shall be presented in detail below – has the most parameters; it goes without saying that the more parameter a distribution model has, the more flexible it is. The parameters of a given distribution have to be estimated such that the model yields the best fit to the data under study. In former times, this estimation has been done by different estimation methods, where estimated values were determined with regard to theoretical characteristics of the given model. Today, this process of parameter estimation is increasingly, if not exclusively, done by special software tools: parameters are optimized via iterative procedures to obtain minimal deviations between theoretical and observed values.[4] As a matter of fact, parameter estimation is first and foremost a method to find the optimal parameter values. Then the corresponding model values have to be tested statistically for goodness of fit. Yet, fitting of the distribution model is of course not the ultimate aim; rather, this is one important step in the course of a quantitative linguistic study, which should, at the end, lead to some qualitative interpretation of the results obtained. At this phase the crucial transition from the *discovery* and *description* of particular regularities to their *interpretation* and eventual *explanation* should take place. This clearly defined step is not self-evident in qualitative linguistics, what sufficiently characterizes the latter's scientific status… One important step in this transition would be, of course, the availability of some interpretation of the parameter behavior. However, also in quantitative linguistics, ultimately striving at theoretical explanations, parameter interpretations have hardly ever been achieved and remain one of the crucial objectives of research.

This intention is the starting point of the present study: Based on the observation that obviously, for the description of Slavic letter frequencies, a complex model such as the NHG distribution with its three parameters *K*, *M*, and *n* to be estimated (for details see below) is needed, an attempt shall be made to approach at least some partial explanation of parameter behavior across the languages studied. This endeavor might then be considered to be successful if the

[3] Interestingly enough, the NHG distribution has been proven to be adequate not only for Slavic languages, but for German, as well, cf. Best (2004/05, Best 2005, Grzybek 2007a,b); further details must remain unconsidered, here.

[4] For reasons discussed elsewhere in detail, we do not work with continuous models and curves, here (as to this line of research, cf. the recent work by Kelih 2009), but with discrete frequency models, only. In the studies reported here all relevant approaches thus far pursued in studies on letter frequencies, have been tested for their adequacy. The goodness of fit has been tested with statistical procedures; first and foremost, this has been done by the chi-square test. Since the latter increases linearly with increasing sample size (resulting in significant deviations even in case of good fits), it is more reasonable to use the discrepancy coefficient $C = \chi^2/N$. Values of $C < 0.02$ are then interpreted as an indication of a good fit, values of $C < 0.01$ of a very good fit.

systematic of rank frequency behavior of Slavic letters might be grasped not only within each of the individual languages, but also in comparison across languages. In case this attempt should turn out to be successful, this would be an important step in explaining the necessity of such a complex model.

With these perspectives in mind, it seems reasonable to shortly summarize the state of the art and to present the languages and material analyzed, before delving into further details.

0. Previous Studies on Slavic Languages

Systematic studies on grapheme frequencies have been reported with regard to four Slavic languages: Russian, Ukrainian, Slovak, and Slovene, thus covering inventory sizes (I) in the interval of $25 \leq I \leq 46$, the minimum of 25 representing Slovene, the maximum of 46 representing Slovak (including diagraphs):

1. **The Slovenian data** are taken from Grzybek, Kelih & Stadlober (2006). The experimental framework of this paper may be summarized as follows: 30 individual texts from different text sorts (masters theses, journalistic comments, sermons, private letters, literary prose and scholarly articles) were analyzed. Across all 30 samples, the discrepancy coefficient for the NHG-Distribution was in the interval $0.022 \geq C \geq 0.0055$; for 26 of the 30 individual analyses, we had $C < 0.02$; for 6 of them even $C < 0.01$. Thus the NHG distribution seems to be an adequate model for the Slovenian grapheme frequencies analysed.
2. **Russian grapheme frequencies** were examined in Grzybek, Kelih, Altmann (2005a), involving 30 complete texts. Again in six different text sorts (drama, stories, poems, private letters, novel chapters and novel in verse) the grapheme frequencies were studied under two different conditions: (a) with the letter ‚ë' as a letter in its own right and without it (represented as 'e' instead) – inventory size thus changing from $I = 32$ to $I = 33$. This specific design did not primarily intend to make a „political" statement as to the status of letter; rather, it was meant to be a detailed analysis of the relevance of inventory size for grapheme studies. As a result it turned out that, apart from a systematic displacement (and in fact no significant differences) of entropy and repeat rate values, again the NHG distribution was a good model under both conditions (with $C < 0.02$ in 59 of 60 samples). With condition $I = 32$ – used for our re-analysis of the parameters below – 21 texts showed $C > 0.001$ and for the remaining rest $C < 0.002$ was obtained. In general, once again the NHG distribution, turned out to be a suitable model for grapheme frequencies in Russian.
3. **Slovak** grapheme frequencies were studied on the basis of 30 texts (literary prose, diploma theses, journalistic comments, fairy tales and "technical" texts) where again the NHG distribution turned out to be the

only adequate model (Grzybek, Kelih & Altmann 2005b und 2006); for Slovak, too, this holds true for two conditions, differing with regard to inventories: taking the three digraphs „dz“, „dž“ and „ch“ as separate units in their own right, inventory size is $I = 46$, otherwise $I = 43$. With $I = 46$, 25 of the 30 analyses yielded a fit of $C < 0.02$. The fitting results under condition $I = 43$ are as follows: 28 of 30 texts had $C > 0.02$; 10 texts had even $C > 0.01$. Two outliers ($C > 0.02$) can be explained due to the relatively small sample sizes of 562 and respectively 445 graphemes. Nevertheless, the NHG distribution fits well for Slovak grapheme frequencies too.

4. Finally, **grapheme frequencies of Ukrainian** were studied by Grzybek & Kelih (2005a). The study included 30 texts (drama, journalistic texts, poems, literary prose and scientific texts); inventory size here amounts to $I = 33$ (not counting the inverted comma as a separate grapheme). Again, the NHG distribution was shown to be an adequate model, with $C < 0.02$ in all 30 texts and even $C < 0.01$ for twenty of them.

1.1. Results: Details

Summarizing, one can say that the grapheme ranking behavior can be grasped by one type of model across the four languages studied: This model is the NHG distribution, in its 1-displaced form (since ranking usually starts with rank 1):

$$(2) \qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \quad x = 1,2,\ldots,n+1; \quad K > M > 0, \quad n \in \{1,2,\ldots\}$$

Taking n as one of three parameters of the NHG distribution fixed at $n = I-1$ (since the support of the NHG distribution is limited by $n+1$), only K and M remain as two free parameters to be estimated. With regard to the individual analyses, both parameters may differ for two reasons: first, they may differ within a given language (due to a “natural” variance of frequencies), and second, they may differ across languages (obviously due to some specific ranking behavior). To systematically analyze the parameter behavior of K and M, and to find possible general tendencies, it seems reasonable to calculate mean values of K and M within each of the given languages, along with 95% confidence intervals. Table 1 represents the corresponding values: in addition to the number of samples analyzed (n), information is given as to inventory size (I), as well as to mean values, upper and lower limits of the confidence intervals for K and M.

Table 1
Parameter Behavior of *K* and *M* in the Languages Analyzed

	n	*I*	$\overline{K}$	K_u	K_o	$\overline{M}$	M_u	M_o
Slovene	30	25	2.96	2.91	3.00	0.8351	0.8263	0.8439
Russian	30	32	3.14	3.10	3.18	0.8096	0.7990	0.8202
Ukrainian	30	33	2.96	2.92	3.01	0.8203	0.8082	0.8324
Slovak	30	43	4.07	4.00	4.14	0.8546	0.8389	0.8703

Based on these findings, first attempts have been undertaken to check the behavior of parameters *K* and *M* for regularities and look for possible interpretations (cf. Grzybek, Kelih 2005b; Grzybek, Kelih, Altmann 2005a). In these attempts, it has first been argued on a direct dependence of parameter *K* on inventory size *I*; as a consequence, the interpretation of *K* would be possible across languages. As compared to this, it has been argued in favor of a relation between parameters *K* and *M* in form of a linear relationship, though not across languages, but within each of the given languages. In other contexts (Grzybek 2007a,b) additional interpretations have been considered e.g. the possibility that parameter *M* may be related either to the first frequency (P_1) of a given distribution, or to its mean rank (m_1).

For the time being, these far-reaching perspectives will not be further pursued, here; instead, the observed dependence of *M* and *K* is analyzed in more detail, and a statistical test is presented which may be useful in the given situation.

Figure 1 demonstrates the relation between the two parameters for the four samples described above.

There is only a weak dependence of *M* on *K* across languages, but a significant linear relationship ($p < 0.001$) within each of the languages, *M* increasing with an increase of *K*, and with correlation coefficients *r* ranging from 0.79 to 0.89 and in all cases.

A comparison of the parameter behavior between the different languages shows that the overall tendency seems to be almost identical, displaying approximately parallel slopes of the four regression lines.

A closer inspection of Figure 1, however, displays two remarkable deviations from expectance:

1. for Ukrainian, parameter *K* seems to be smaller than expected (i.e., not in line with the inventory size interpretation);
2. the regression line for Slovene deviates from the scheme, despite its overall accordance with the general parallel tendency, being characterized by an intersection with the regression line of the Ukrainian data.

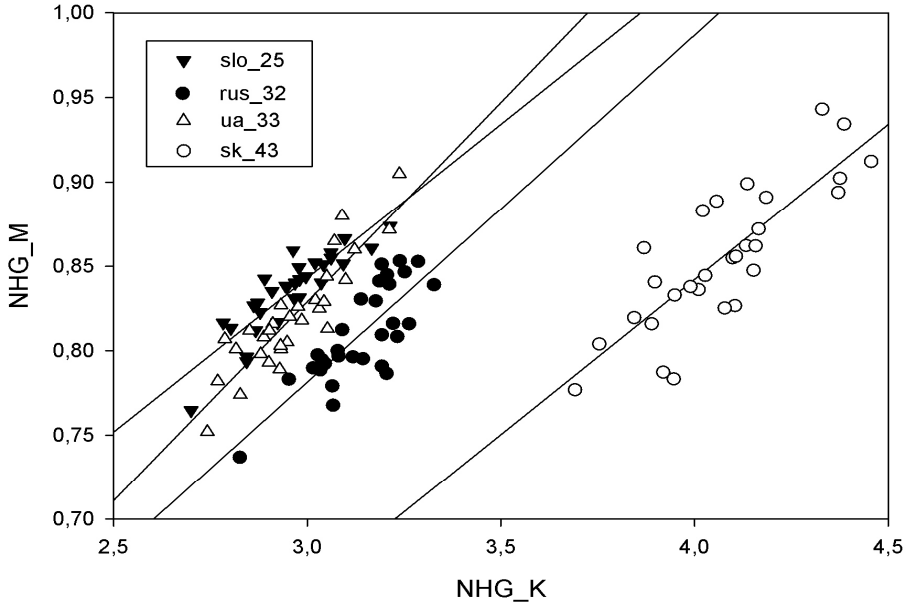


Figure 1. Dependence of parameter *M* on *K* for Slovenian, Russian, Ukrainian, and Slovak grapheme data

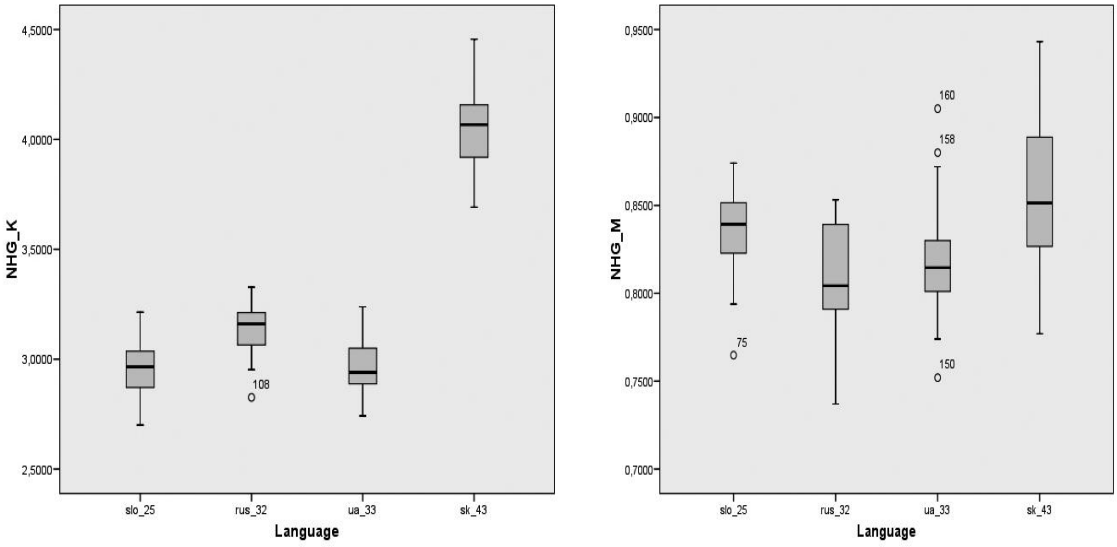
Focusing on the relation between *M* and *K*, only, we neglect the first problem in the given context and concentrate on the second issue. Thus, first stating the overall aptness of the NHG distribution for modeling Slavic letter frequencies, and second observing a general tendency in the behavior of parameter *M*, we can turn to the more specific question as to the observed deviations from the established rule.

1.2. Outliers and Extreme Values

A common first step in explaining the observed deviations from general parameter behavior can be seen involves an analysis of possible outliers and extreme values, which are eliminated from the analysis. This is usually be done be reference to the so-called interquartile range (IQR), which comprises the central 50% of all observations. Outliers and extreme values are defined as cases, for which the difference to the upper and lower limit of the IQR is more than 1.5 times (or 3 times, respectively) as large as the IQR.

The analysis can be illustrated by box plots, in which outliers can easily be detected and identified: they are located beyond or below the upper or lower line, which is defined by a concrete value of the given data set, and is maximally 1.5

times as large as the IQR – if there are no outliers, they are represented by the maximal and minimal values of the given sample. Figures 2a and 2b represent the four box plots for the parameter values of *K* and *M*.



(a) Parameter *K* **(b)** Parameter *M*

Figure 2. Box plot series

Indeed there are some outliers which can easily be identified. In case of parameter *K*, this is only one of the Russian private letters (# 260). In case of parameter *M*, we are concerned with four outliers, one from the Slovene data (# 22), and three from the Ukrainian data (# 332, 340, 342). Eliminating these outliers from the analysis and submitting the data again to a study of parameter behavior results in an only slightly changed picture of the regression lines, as illustrated by Figure 3.

As can be seen, the regression line for the Slovene data is still characterized by an intersection with the regression line of the Ukrainian data, but now this intersection is far away from all observed data points. Table 2 represents the regression equation for all four languages (after elimination of the outliers); these regression lines follow the equation $y = a + bx$ (in our case we have $M = a + bK$). Inventory size is denoted by *I*, sample size by *n* (after elimination of outliers);

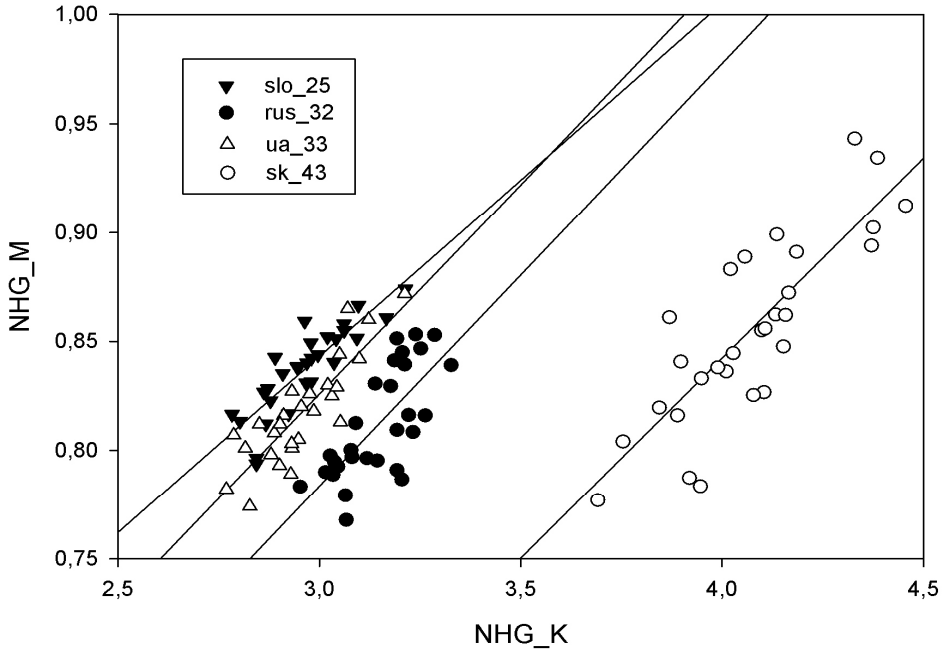


Figure 3. Relation between parameters *K* and *M* (after elimination of outliers)

Table 2
Regression coefficients: $M_i = a_i + b_i \cdot K_i$ and correlation coefficients *r*

	I	*n*	*b*	*a*	*r*
Slovene	25	29	0.1620	0.3571	0.86
Russian	32	29	0.1941	0.2013	0.72
Ukrainian	33	27	0.1921	0.2494	0.85
Slovak	43	30	0.1840	0.1067	0.83

Since it is the regression line for Ukrainian, which intersects with the Slovenian one, it is reasonable to test the difference between the two regression coefficients (slopes b_1 and b_2) for significance. For linear relations, this can be done by reference to a *t*-test statistic

(3) $$t = \frac{|b_1 - b_2|}{\sqrt{\frac{s^2_{y1.x1} \cdot (n_1 - 2) + s^2_{y2.x2} \cdot (n_2 - 2)}{n_1 + n_2 - 4} \cdot \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}} \right)}}$$

with $DF = n_1 + n_2 - 4$ degrees of freedom and $Q_x = \sum (x - \bar{x})^2$.

As a result, the comparison of the two regression coefficients b_1 for Slovene and b_2 Ukrainian shows the difference to be not significant, with a value of $t = 1.01$ and $DF = 52$ degrees of freedom ($p = 0.32$). This result naturally leads to a simultanous comparison of all four regression lines, rather than only two of them. Given there is no significant deviation from parallelism, this would yield a regression model common to all four samples studied.

1.3. A common regression model for Slavic letter frequencies?

With regard to a possible uniformity of the tendencies and, as a consequence, a common regression model, one may ask the specific question if the dependence of parameter *M* on *K* shows an identical trend for the four languages under study. This leads to the question of testing the differences between the regression coefficients and the parallelism of the regression lines for significance. An adequate procedure to test this is the multiple partial *F*-test; usually, this test is applied with regard to multiple linear regressions (cf. Kleinbaum et al. 1998) when the question of possible additional contributions of independent variables, which are not (yet) included in a given model, is at stake. The *F*-test thus tests the effect of expansion of a given model by the simultaneous addition of two or more variables. In its complete form, such a multiple model has the following form:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \beta_1^* X_1^* + ... + \beta_k^* X_k^* + \varepsilon \quad (4)$$

Here, *Y* is the dependent variable, α is the regression constant, and ε is a random error; *X*i and X_i^* are the independent variables, β_i and β_i^* the regression coefficients. The null hypothesis (H_0) to be tested includes the assumption that $X_1^*, X_2^*, ..., X_k^*$ do not significantly contribute to the prediction of Y, when X_1, $X_{2...}$ X_k are already included in the model; thus, for the complete model we have $H_0 : \beta_1^* = \beta_2^* = ... = \beta_k^* = 0$.

From this (second) formulation the reduced model under H_0 is:

$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \quad (5)$$

Thus, the variance (SSq_{reg}) explained by the model becomes larger by the addition of X_i^*; now, the following F statistics can be calculated:

$$F = \frac{\left[SSq_{reg}(complete\ model) - SSq_{reg}(reduced\ model) \right] / k}{SSq_{res}(complete\ model) / \left(n - p - k - 1 \right)} \quad (6)$$

In (6), SSq_{reg} denotes the variance (i.e. the sum of the squared deviations: sum of squared effects) of the complete or the reduced regression models, and SSq_{res} denotes the squared sum of the residuals of the complete model (sum of squared errors); n is the sample size, p is the number of regression coefficients in the reduced model, and k is the number of those regression coefficients which equal zero under the assumption of the null hypothesis (H_0).

In our case, the sample size is n = 115 texts (without outliers) from four sub-samples from four different languages (each with its own inventory size). Since the attribution to one of the languages is a nominal category, and since nominally scaled predictors cannot be directly introduced into a regression model, the relevant information has to be (re)-coded in a different manner. To this end, one introduces dummy coding: in this case, a variable is split into sub-variables (which are termed 'indicators') and coded dichotomously; each category is thus classified 'present' (1) or 'absent' (0). Membership of a given case within a given (sub)sample can thus be regarded as a dummy variable with the coding 0 and 1. The advantage of such a binary (0 vs. 1) coding is that dummy variables can be statistically treated like interval-scaled variables. A categorical variable with *k+1* values is thus transformed into *k* dummy variables each with two values 0 and 1. Since our variable "LANGUAGE" (with a given inventory size) has four categories, three dichotomous variables (D_1 to D_3) can be constructed which contain the same information as one categorical variable. In our case, we thus obtain the scheme represented in Table 3.

Table 3
Coding schema and dummy coding

	I		D_1	D_2	D_3
Slovene	*25*	*8*	0	0	0
Russian	*32*	*2*	1	0	0
Ukrainian	*33*	*5*	0	1	0
Slovak	*43*	*6*	0	0	1

Within this framework, our question as to the parallelism of regression lines turns out to be a special case of a more general situation: considering the regression lines to be parallel to each other if the predictive value of *Y* is not significantly changed by the addition of the additional variables, the described procedure as apt to be applied to this special case. In this case, the reduced model for *M* can be written as:

(7) $$M = \alpha + \beta_1 \cdot K + \beta_2 \cdot D_1 + \beta_3 \cdot D_2 + \beta_4 \cdot D_3 + \varepsilon \ .$$

This means that the regression lines of all four groups are parallel with identical slope β_1 (p = 4). The pre-conditions are thus fulfilled to make a comparison between the complete and the reduced model, with regard to variable 'LANGUAGE' in its re-coded (dummy-coded) form, by addition of the products of X_i and K as variables X_1^*, X_2^*, X_3^* to the dummy variables $X_1^* = KD_1$, $X_2^* = KD_2$, $X_3^* = KD_3$. Hence the full model can be written as

(8) $$M = \alpha + \beta_1 \cdot K + \beta_2 \cdot D_1 + \beta_3 \cdot D_2 + \beta_4 \cdot D_3 + \beta_1^* \cdot KD_1 + \beta_2^* \cdot KD_2 + \beta_3^* \cdot KD_3 + \varepsilon$$

In our case we are concerned with 7 parameters for the complete model (K, 3 dummy-coded variables, and 3 dummy products), for which we obtain the values SSq_{reg} = 0.0726 and SSq_{res} = 0.0229.

Interestingly enough, one obtains for the reduced model (7), which contains four variables, namely, the three dummy-coded variables in addition to K, a nearly identical value of SSq_{reg} = 0.0724. The error sum of squares (i.e., the sum of the squared deviations of the residuals) of the reduced model, too, is almost the same with SSq_{res} = 0.0231.

Thus, in this particular case, the assumption of parallelism seems likely to be confirmed by a statistical test, the F-test.

For the calculation of the F value we need the value of k (cf. (6)), which is obtained by the difference between the number of variables of the complete model (8) and that of the reduced model (5), in our case, $k = 7 - 4 = 3$, equivalent to the number of degrees of freedom for the numerator in (6). We also need the mean of the squared sum of residuals, which is represented by the quotient of the sum of the squared residuals (SSq_{res}) and the number of the degrees of freedom of the denominator, being calculated as $m = n - p - k - 1$; in our case, we thus have $m = 115 - 7 - 1 = 107$.

We now can calculate the F value as

$$F_{(FG_3=5, FG_2=107)} = \frac{(0.0726 - 0.0724)/3}{0.0229/107} = 0.31$$

With the given degrees of freedom, this F-value corresponds to a probability of p = 0.82, which is far from any statistical significance; as a consequence, we have to retain the null hypothesis ($H_0 : \beta_1^* = \beta_2^* = \beta_3^* = 0$), according to which the regression lines are parallel.

We can thus summarize that the dependence of parameter M on parameter K of the negative hypergeometric distribution behaves identically across the four Slavic languages studied which can be expressed as a common regression model. Within this model, the common regression coefficient (slope) is $\hat{\beta} = b \approx 0.1811$; accordingly, for the four languages under study parameter M can be estimated as

$\widehat{M} \approx \widehat{\alpha} + 0.18 \cdot K$; differences between the languages are a result of differences in the intercept $a = \widehat{\alpha}$ From this general model, the individual groups (i.e., the four languages each with their given inventory sizes) can be derived as special cases. Given the fact that the null hypothesis is to be retained, it is sufficient to do this with reference to the reduced model. Ignoring the error of estimation (ε), we obtain

Group 1:	$a + b_1 \cdot K$
Goup 2:	$(a + b_2) + b_1 \cdot K$
Group 3:	$(a + b_3) + b_1 \cdot K$
Group 4:	$(a + b_4) + b_1 \cdot K$

For our four languages we thus obtain the following special regression models

Slovene	25	$M = 0.18K + 0.3005$
Russian	32	$M = 0.18K + 0.2470$
Ukrainian	33	$M = 0.18K + 0.2826$
Slovak	43	$M = 0.18K + 0.1181$

Now, interpreting the intercepts as response variables of a regression model with inventory size *I* as the independent variable exhibits a clear tendency as illustrated in Figure 4. This tendency, based on four data points only, is not significant ($p = 0.07$; $r = 0.93$), however, obviously due to the deviating structure of the Ukrainian data; the analysis of more data from further Slavic languages will shed more light on this highly intriguing issue.

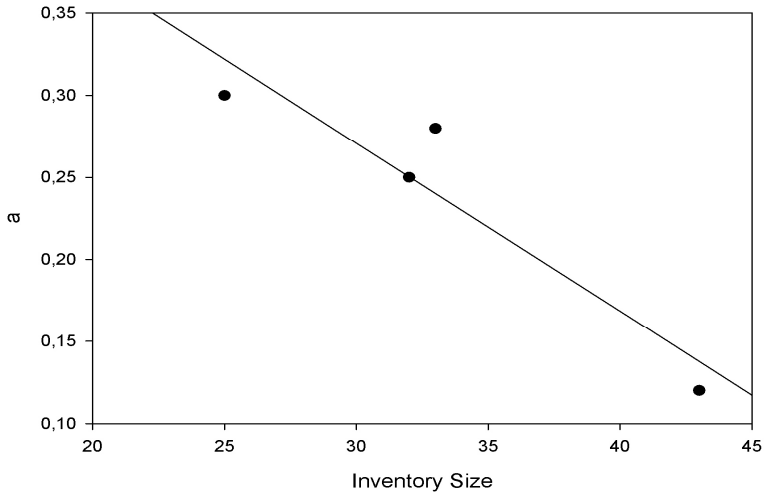


Figure 4. Relation between the intercepts of the regression model and inventory size *I*

With these findings, an important step seems to be made towards the intended analysis of the systematic parameter behavior, at least as far as parameter *M* of the NHG distribution is concerned: it turns out to be plausible that parameter *M* is closely related to parameter *K*, not across languages, but rather within a given language, only.

Yet, it is interesting to delve even deeper into the matter; a next logical step in this direction would be an answer to the question under what circumstances the outlined regression model is less effective than expected. To provide possible answers to this question, a next step might involve specific analyses of residuals, but this would clearly go beyond the scope of this paper.

2. Summary of Results and Future Prospectives

With regard to the findings reported above, we can summarize the most import results:

1. In systematic analyses of letter frequencies from four Slavic languages it could be shown that all can be adequately modeled by the NHG distribution, other models failed to be likely successful.
2. Parameter behavior of the NHG distribution seems to be highly regular; this regularity seems to be related to both language-specific and interlingual factors:
    a. the relevance of interlingual factors has already been discussed else where (cf., e.g., Grzybek & Kelih 2005); inventory size has been identified as a crucial factor influencing the distinction between languages; in this article, further arguments in favor of this notion have been brought forth, by showing that, at least for various Slavic languages, the relation between the parameters *M* and *K* of the NHG distribution follows a common linear regression model from which the individual languages may be derived as special cases;
    b. language-specific tendencies are expressed in parameter values, which lend themselves to discriminant analyses; also the specific relation between parameters *K* and *M* of the NHG distribution seems to be specific for individual languages.

    The study of additional languages, Slavic and non-Slavic, is necessary to gain more information on this specific situation. It seems plausible that, in addition to the above-mentioned language-specific and interlingual factors, also "local" factors may come into play, as can be seen in case of Ukrainian; here, additional analyses turn out to be necessary to grasp more exactly the boundary conditions of letter behavior. In this context, it has to be checked if the deviation of individual languages (as, e.g., Ukrainian in our case) may be caused by mere computational aspects of parameter estimation; to give an answer to this question, a qualitative parameter

interpretation is in order. We are still relatively far from this stage, but first analyses in this direction point at the importance of particular "first-order" characteristics such as inventory size (I), or first frequency (P_1), as well as of "second-order" characteristics, such as mean rank (m_1), entropy (H), repeat rate (RR), etc. – research in this direction is in progress now.

3. As this study shows, single corpus analyses of a given language cannot, as "representative" as they may be considered to be, uncover all mechanisms and processes at work in a language's dynamic system – in the given case we see that, within a language, there seems to be an intrinsic mechanism which synergetically regulates the dynamic balance of possibly contradictory forces, and which regulate the overall frequency behavior.
4. In order to identify trends, any sample must be checked for possible outliers and extreme values which eventually has to be eliminated from the analysis in order the trend to be uphold; in this respect, sample size, too, must be controlled to guarantee the statistical stability of tendencies (cf. Grzybek et al. 2009).
5. It is obvious that modifications of the model described above may be necessary when further languages (particularly from other than the Slavic family) will be taken into consideration; it may well turn out that the NHG model then soon turns out to be a special model relevant only for particular languages, or specific writing systems, etc.
6. In order to arrive at an explanation why the NHG is a suitable model for grapheme frequencies, its theoretical derivation must be carefully taken into account: since, in this case, it does not seem to make sense to interpret it in terms of an urn model, it might be reasonable, by way of an alternative, to derive the NHG rather as beta-binomial distribution, i.e., as a binomial distribution with its parameter p being variable and following a beta distribution: as Grzybek (in print) shows, this results in a new interpretation of the whole generating process. Seen from this perspective, a specific inventory size is not a "given" fixed starting point, but rather emerges as the diachronically motivated outcome of the dynamic speaker-hearer communicative interaction.

References

Altmann, G., Köhler, R. (1996). 'Language Forces' and synergetic modelling of language phenomena. In: Schmidt, P. (ed.), *Glottometrika 15, 62–76.* Trier

Best, K.-H. (2004/05). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21–32*.

Best, K.-H. (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Науковий вісник Чернівеского університета, віп. 231, 119–127.*

Grzybek, P. (2006). A very early Slavic letter statistic in the Czech journal *Krok* (1831): Jan Svatopluk Presl (1791-1849). *Glottometrics 13*, 88–91.

Grzybek, P. (2007a). What a difference an ‚E' makes. Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen. In: Deutschmann, P. unter Mitarbeit von P. Grzybek, L. Karničar, H. Pfandl (eds.), *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag: 105–128.* Wien: Praesens.

Grzybek, P. (2007b). On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies. *Glottometrics 15,* 82–91.

Grzybek, P. (in print). Graphem- und Phonemstatistik: Inventare – Modelle – Zusammenhänge – Typologie. In: Kempgen, S., Berger, T., Gutschmidt, K., Kosta, P. (eds.), *Die Slavischen Sprachen. Ein internationales Handbuch zu ihrer Struktur, ihrer Geschichte und ihrer Erforschung. Bd. 2.*

Grzybek, P., Kelih, E. (2003). Graphemhäufigkeiten (am Beispiel des Russischen) Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie 31, 131–162.*

Grzybek, P., Kelih, E. (2005a). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph." In: Altmann, G., Levickij, V., Perebejnis, V. (eds.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics: 159–179.* Černovci: Ruta.

Grzybek, P., Kelih, E. (2005b). Towards a general model of grapheme frequencies in Slavic languages. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages*: *73–87.* Bratislava: Veda.

Grzybek, P., Kelih, E., Altmann, G. (2004). Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25–54*.

Grzybek, P., Kelih, E., Altmann, G. (2005a). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das ‘ë'. *Anzeiger für Slavische Philologie 33, 117–140.*

Grzybek, P., Kelih, E., Altmann, G. (2005b). Graphemhäufigkeiten im Slowakischen. (Teil I: Ohne Digraphen). In: Nemcová, E. (ed.), *Philologia actualis slovaca.* [im Druck]

Grzybek, P., Kelih, E., Altmann, G. (2006). Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen. In: Kozmová, R. (ed.), *Sprache und Sprachen im mitteleuropäischen Raum: 661–684*. Trnava: GeSuS.

Grzybek, P., Kelih, E., Stadlober, E. (2006). Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik. *Anzeiger für Slavische Philologie 34, 41-74.*

Grzybek, P., Mačutek, J., Stadlober, E., Wimmer, G. (2009). Sample size estimation in linguistics – A new approach. In prep.

Kelih, E. (2007). Häufigkeiten von Graphemen und Lauten: Zwei Ebenen – ein Modell? (Re-Analyse einer Untersuchung von A.M. Peškovskij). In: Grzybek, P., Köhler, R. (2006) (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann On the Occasion of His 75th Birthday: 267–277*. Mouton de Gruyter: Berlin – New York [= Quantitative Linguistics, 62].

Kelih, E. (2009). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle. *Glottometrics 18, 53–69*.

Kleinbaum, D.G.; Kupper, L.L.; Muller, K.E. ([3]1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press. 3rd ed., rev.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch – An International Handbook*: *760–774*. Berlin u.a.: Walter de Gruyter [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Wimmer, G.; Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch – An International Handbook: 791–807*. Berlin u.a.: Walter de Gruyter..

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 329–335*. Dordrecht, NL: Springer.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 34-45.*

A contribution to quantitative studies on the sentence level

Reinhard Köhler
Sven Naumann

I. Clause detection

Why are clauses promising candidates for quantitative studies of syntactic properties of natural languages? Besides the fact that clauses have been used successfully in a number of different applications such as text-to-speech conversion, text-alignment, machine translation and discourse parsing, they are from a semantic point of view interesting units as they are the smallest units with a propositional content (i.e. containing a predication expressing a "complete thought").

But what kind of entities are clauses exactly? Like with many other basic concepts in linguistics (such as *word* and *sentence* e.g.), it is hard to come up with a precise definition of the term *clause*, especially if this definition is supposed to reflect the common usage of this term. The basic idea is that clauses are the sentential units full-fledged sentences are made of, leading to the conclusion that any sentence can be segmented into one or more clauses. Following this line of argument, it is often added that *clauses are word sequences which contain a subject and a predicate*. Leaving aside the fact that without further qualification this statement does not allow to draw a sharp line between sentences on the one hand and clauses on the other, it either has the consequence that expressions lacking an appropriate overt nominal or verbal element (like imperatives or infinitive constructions in many languages) cannot be classified as clauses, or clause detection presupposes some kind of advanced syntactic processing of the sentence which helps to identify implicitly realized subjects and predicates.

Another point is, that it is usually assumed that clauses can be nested (i.e. a clause can contain other clauses), but have to be nested properly (i.e. there are no crossing brackets or edges). Cross-serial dependencies as found in Dutch and other languages indicate that subjects and predicates can pattern in a way which may cause severe descriptive problems.

Finally it has to be noted that as there is no generally accepted operational definition of the concept *clause*, comparing the performance of different clause detection systems often is not an easy task.

First systems for automatically identifying clauses in the late 80th and the 90th used hand-crafted rules (Ejerhed 1988; Papagergiou 1997 and Leffa 1998) and proved to be quite successful. In recent years, a number of systems have

been designed and tested which use machine learning techniques such as decision graphs, neural networks, HMMs, etc. (cf. the CoNLL-2001 shared task, Sang/ Déjean 2001), thereby dispense with the tedious and time-consuming task to write appropriate language specific rules.

We have designed a clause detection system for German, which maps each sentence of a document onto a (possibly empty) sequence of non-overlapping clauses, thereby undoing any nesting of clauses found in the original data. In the sequence computed a clause B that is embedded in a clause A is positioned after clause A. The reason for this kind of processing is that it makes it easier to compute certain values and measures (e.g. clause sequences) needed for the quantitative studies presented in this paper. It is basically a rule-based system which uses a small number of general heuristics, which can easily be adapted to different languages.

The system accepts XML-documents in which sentence and paragraph boundaries are marked and morpho-syntactic tags have been assigned to all tokens. No further linguistic information (like chunk or phrase boundaries) are included.

Clause detection is considered as a three step process driven by a small number of heuristic rules: First, a set of clause candidates is computed by exploiting morpho-syntactic information and punctuation marks. As we try to maximize recall in this step, this list often contains candidates which do not represent a full clause but a part of a clause. In a second step, all the clause candidates are classified. Finally, complete clauses are removed from the list and new clause candidates are built by adjoining appropriate clause candidates already on the list. This last step is repeated until either the candidate list is empty or contains only candidates which cannot be adjoined.

Clause types

We distinguish three types of clauses:

(1) Finite clauses - clauses containing a finite verb form:
Haile rennt. (*Haile runs*.) / *... der beste Langstreckenläufer, den es jemals gab*./ (… *the best long-distance runner which ever lived*)

(2) Non-finite clauses – clauses containing no finite verb, but one or more non-finite verb forms:
Er scheint fest entschlossen, es mit jedem aufzunehmen. (*He seems to be determined to take everyone on*)

(3) Verbless clauses:
Ja! (*Yes*!) / *Hilfe*! (*Help*!)

Dealing with Insertions

Sentential insertions typically marked by pairs of brackets or hyphens or in case of reported speech by quotation marks, are extracted from the sentence and dealt with separately. This step is necessary in order to ensure that the heuristics for adjoining clause candidates work properly.

Generating Clause Candidates

The most important indicators of clause boundaries are punctuation marks and subordinating and coordinating conjunctions. Each punctuation mark is considered as signalling a (potential) clause boundary. The overgeneration introduced by this step is removed later on by the rules for adjoining clause candidates. Coordinating conjunctions like *und* (and), *oder* (or) pose a greater challenge. A colon precedes a coordinating conjunction only if the constituents combined are complete sentences, each of which could be used in isolation:

- Coordination of sentences:
  Man müsse von den deutschen Athleten lernen, denn die deutsche Disziplin sei eisern.
- Coordination of clauses:
- *In der Umkleidekabine des Olympiastadions in Atlanta schnürt Haile sich die blauen Adidasturnschuhe | und schließt die Augen zu einem Gebet. Als er auf dem Treppchen steht, die äthiopische Nationalhymne ertönt | und ihm die Tränen kommen, wird auch im Kinosaal bereits geschluchzt.*
- Coordination of other constituents:
- *Und liest seiner Mutter und den Geschwistern beim Schein einer Öllampe Gleichnisse aus der Bibel vor. Deshalb muss es auch Wettkämpfe in den Dörfern geben und nicht nur hier in Addis.*

In order to detect clause boundaries in case of coordinated clauses, a simple heuristic analyzing verbal elements is used to introduce extra clause boundaries before conjunctions combining clauses.

Classifying Clause Candidates

Word sequences marked as clause candidates in the first step are assigned to one of the following classes:

(A) subordinate clauses starting with a subordinating conjunction or a relative phrase and ending with a finite verb;
(B) main clauses and non-finite clauses;

(C) word sequences beginning with a subordinating conjunction or relative phrase but missing a finite verb, which are considered as first part of a complete clause;
(D) word sequences containing a finite verb but lacking structural properties which would allow to classify them as complete clauses without any doubt and
(E) word sequences which could not be assigned to one of the classes men tioned above.

Example (1)
Auf der Leinwand erscheint Haile Gebrselassie, [B_1]
Olympiasieger über 10.000 Meter, [E_2]
Weltsportler des Jahres 1998 und einer der besten, [E_3]
wenn nicht der beste Langstreckenläufer, [C_4]
den es jemals gab. [A_5]
Haile Gebrselassie ist Äthiopiens unumstrittener Nationalheld, [B_6]
dessen Lebensgeschichte nun von Leslie Woodhead in "Endurance", [C_7]
Ausdauer, [E_8]
verfilmt wurde. [D_9]

Removing and Adjoining Clause Candidates

The candidate list is processed by a simple iterative scheme: If a candidate is considered to represent a complete clause, it is removed from the list. Otherwise the list is searched for another candidate it can be combined with. Elements remaining in the list after the procedure has terminated are interpreted as (complete) clauses.

Process candidate list
Repeat until the clause candidate list is empty or no further operations are applicable:
(1) Remove all type-A clauses.
(2) If a type-B or type-C clause is followed by one or more type-E clauses, adjoin these clauses and reclassify the new clause.
(3) A type-C clause is followed by a type-D clause. If there is a second type-D Clause following, adjoin the type-C clause with the second type-D Clause; otherwise adjoin it with the first type-D clause. Reclassify the new clause.

Example (2)
Consider the first sentence from example (1) above: At the start, the candidate list contains the clause candidates [B_1], [E_2], [E_3], [C_4] and [A_5]. [A_5] is considered to be a complete clause and removed from the list. Now, the first three clause candidates are adjoined, generating a new clause candidate [$A_{1\text{-}3}$]. Finally, this new candidate is combined with [C_4] and after the resulting clause [$A_{1\text{-}4}$] is removed, the candidate list is empty and the process stops.

Data and first evaluation

We used our clause detection system to identify the clauses in 64 newspaper articles taken from the Trier TAZ corpus. Article length varied from 67 to 399 sentences (7070 sentences altogether), resulting in more than 10.000 clauses.

For a first evaluation of the performance of our system, we randomly selected 5 articles and hand-checked the automatically generated clause boundaries. The results of this first, tentative test indicate a success rate between 90 and 95%. Many clause detection errors were due to missing or wrongly placed punctuation marks and false morpho-syntactic tags.

II. Syntagmatic units on the sentence level: Length motifs

With these data, several studies became possible:

(1) the distribution of sentence lengths in terms of clauses;
(2) patterns of length motifs on the sentence level
  (a) the motif type rank-frequency distribution
  (b) the distribution of motif lengths;
(3) the Menzerath-Altmann-Law (sentence/clause/word).

Sentence length in terms of the number of clauses

Sentence length studies are usually conducted in terms of the number of words a sentence consists of – although this kind of investigation suffers from several shortcomings; among them the following are the most severe ones:

(1) Words are not the immediate constituents of sentences and do, therefore, not form units of appropriate granularity;
(2) it is very unlikely to get enough data for each length class as the range of sentence lengths in terms of words varies between unity and several

dozens. For this reason, the data are usually pooled but do not form smooth distributions nevertheless.

On the data generated by our program, we studied sentence lengths in terms of the number of clauses. Although the number of sentences per text in our corpus is limited it was sufficiently large to form satisfying samples. Fitting theoretical frequency distributions yielded a set of nicely fitting distributions around the Poisson distribution. Deviating texts followed distributions such as the Bissinger-Poisson, the Positive Poisson, and the Poisson distribution with exponential limit. The fits were very good (cf. Fig.1).

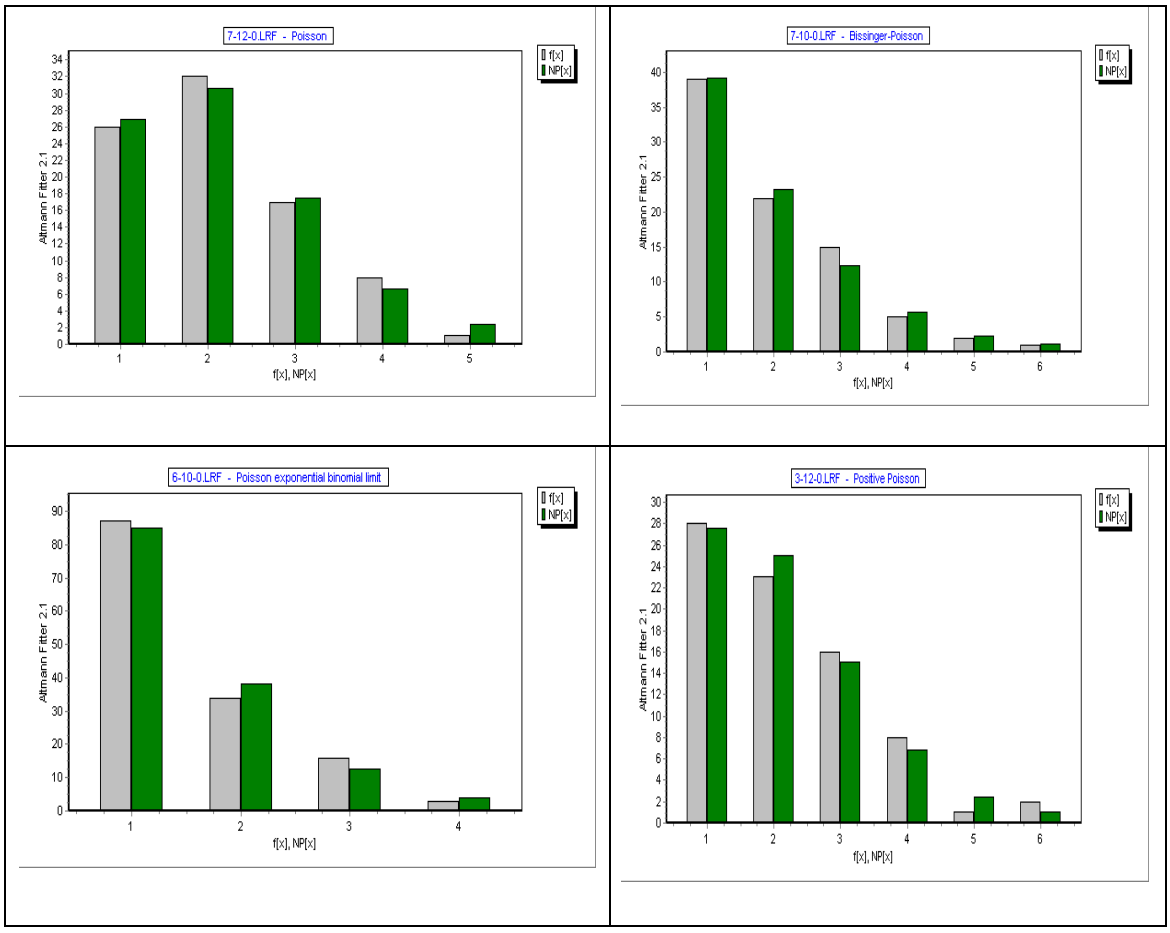


Figure 1: Fitting distributions from the Poisson family to the sentence length data

The length motif as a new unit

Another part of our studies concentrates on patterns of sequences of the values of several linguistic properties. Such sequences were introduced into linguistics as

new units for several reasons (later called „motifs“). In particular, quantitative properties of linguistic units and of texts have been studied almost exclusively with respect to their distributions whereas only very few investigations have been devoted to syntagmatic properties (cf. Köhler 2006a, 2006b). The first aspect was called “language in the mass” by Gustav Herdan as opposed to “language in the line”; a modern expression, in particular common in corpus linguistics, emphasises the first one as “bag-of-words” model. Distributional analyses reflect certain significant characteristics of linguistic units and texts but fail to unveil properties of the organisation of these units in their syntagmatic dimension. For the first aspect, appropriate units have not been available so far. Following Köhler (2006a, 2006b; Köhler/Naumann 2007), such new units can be defined in the following way:

> *A X-motif is a continuous series of equal or increasing values of the variable X in a linguistically organised structure.*

Here, the variable *X* represents any measurable property of a linguistic unit, such as frequency, length, polysemy, polytextuality, homonymy, synonymy etc. of morphs, words, phrases, clauses, sentences etc. (clearly, not every property is applicable to all units). The linguistic structure is, in most cases, a full text. An illustrative example is the definition of the L-motif as a monotonically increasing series of, say word length, measured in terms of the number of their syllables (or morphs). The sentence

> *“Word length studies are almost exclusively devoted to the problem of distributions.”*

corresponds to the five motifs

> (1-1-2) (1-2-4) (3) (1-1-2) (1-4)

if word length is measured in syllables. The new unit has some interesting properties: Motifs can always be determined in an objective, unambiguous and exhaustive way, they allow for an appropriate granularity on all levels of linguistic analysis, and are scalable because the procedure can be iteratively applied.

Previous studies on the basis of motifs were conducted on several properties on the word level. In our present paper, we study length motifs on the sentence level, measuring sentence length in terms of the number of clauses a sentence consists of.

The motif obtained in the described way form an inventory which can be investigated in analogy to a vocabulary. Hence, all kinds of analysis which are common for other linguistic units can also be conducted on the basis of our motifs.

One of the self-suggesting analyses is the determination of an appropriate probability distribution. The motif types from our data were best represented by the Zipf-Mandelbrot distribution with very good Chi-square values (close to unity, cf. Fig. 2).

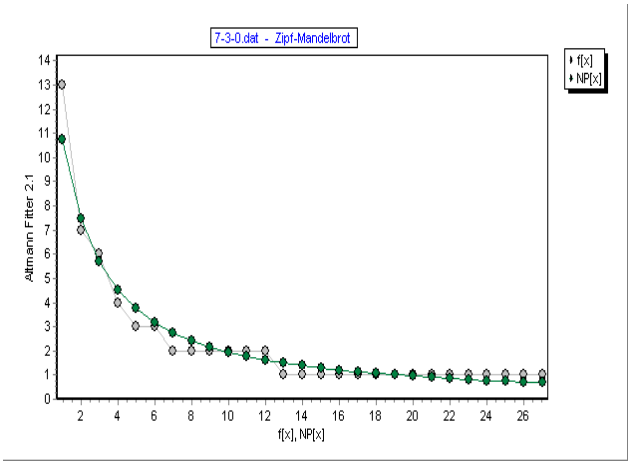


Figure 2: The Zipf-Mandelbrot distribution as fitted to the motif type data

The length of length motifs

Another obvious question concerns the lengths of the length motifs themselves. A theoretical model of the length distribution of L- motifs of sentences is obtained by the following considerations:

[1] In a given text, the mean sentence length, the estimation of the mathematical expectation of sentence length, can be interpreted as the sentence length intended by the text expedient (speaker/writer).

[2] Shorter sentences are formed in order to decrease decoding/processing effort (the requirement *minD* in synergetic linguistics) within the sentence. This tendency will be represented by the quantity *D*.

[3] Longer sentences are formed where they help to compactify what otherwise would be expressed by two or more sentences and where the more compact form decreases processing effort with respect to the next higher (inter-sentence) level. This will be represented by *H*.

[2] and [3] are causes of deviations from the mean length value while they, at the same time, compete with each other. We express this interdependence in form of Altmann's approach (Köhler/Altmann 1966): The probability of a sentence length x is proportional to the probability of sentence length x-1, where the proportionality is a linear function:

$$P_x = \frac{D}{x + H - 1} P_{x-1}.$$

D has an increasing influence on this relation whereas *H* has a decreasing one. The probability class *x* itself has also a decreasing influence, which reflects the fact that the probability of long sentences decreases with the length. This equation leads to the hyper-Poisson distribution (Wimmer/Altmann 1999, 281):

$$P_x = \frac{a^x}{{}_1F_1(1;b;a)\,b^{(x)}}, \quad x = 0,1,2,..., \quad a \geq 0,\ b > 0,$$

Where ${}_1F_1(1;b;a)$ is the confluent hypergeometric function

$${}_1F_1(1;b;a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}}.$$

and $b^{(x)} = b(b+1)\ldots(b+x-1)$. According to this derivation, the hyper-Poisson distribution, which plays a basic role with word length distributions (Best 1997b), should therefore also be a good model of L- motif length on the sentence level although motifs on the word level, regardless of the property considered (length, polytextuality, frequency), follow the hyper-Pascal distribution (Köhler 2006a; Köhler/Naumann 2007). In fact, many texts from our corpus follow this distribution (cf. Fig. 3). Others, however, are best modelled by other distributions such as the hyper-Pascal or the extended logarithmic distributions. Nevertheless, all the texts seem to be oriented along a straight line in the I/S plane of Ord's criterion (cf. Fig. 4).

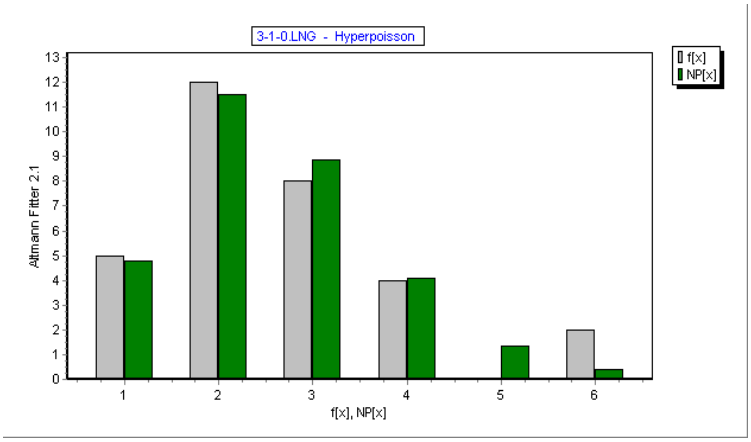


Figure 3: Fitting the hyper-Poisson distribution to the frequency distribution of the lengths of L- motifs on the sentence level.

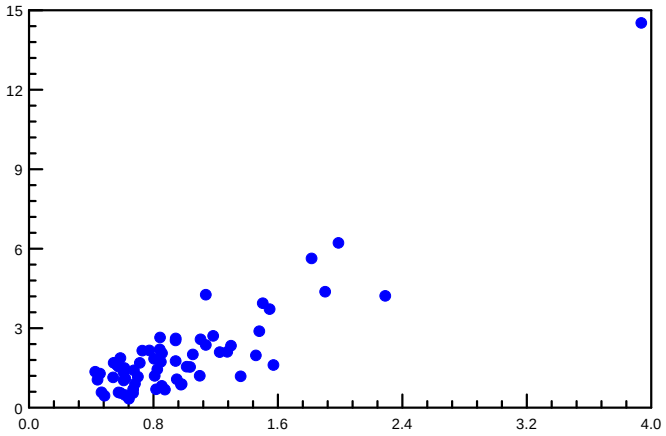


Figure 4: Ord's criterion of the frequency distribution of the lengths of the L- motifs.

Menzerath-Altmann's Law

As our algorithm provides data on the sentence/clause level in a rather easy way, large quantities of material can be generated which can be used for various investigations on this level of analysis. In order to check whether the quality of our data is sufficient we used it with the well-confirmed Menzerath-Altmann law. This law has been tested on data from many languages and on various levels of linguistic investigation. On the sentence level, however, not too many studies have been done for obvious reasons. Moreover, the existing results are not always comparable because there are no accepted standards and researchers often apply ad-hoc criteria.

As Fig. 5 shows, the results correspond closely to those studies which are based on manual identification and segmentation of clauses (However, we have not yet compared the parameter values of the different approaches).

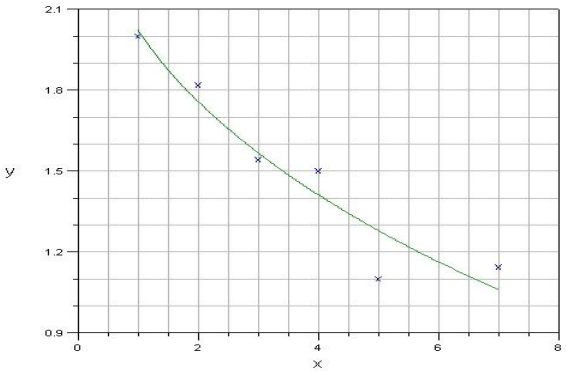


Figure 5: The Menzerath-Altmann Law on the sentence/clause/word level

Conclusion

Our algorithm can be used as a tool for the segmentation of German sentences into clauses. So far, we could not find remaining mistakes in the data which would disturb statistical studies. Now it seems worthwhile to work out criteria and algorithms for other languages.

In the course of our research we could show that the L- motifs succeed as units for the investigation of the syntagmatic dimension also on the sentence level. Further studies should be devoted to motifs of other properties on this level.

References

Köhler, Reinhard, Altmann, Gabriel (1996), "Language Forces" and synergetic modeling of language phenomena. In: *Glottometrika 15: 63-76.* Trier: WVT.

Köhler, Reinhard (2006a), Word length in text. A study in the syntagmatic dimension. In: Myslovičová, Sibyla (ed.), *Jazyk a jazykoveda v pohybe. Festschrift for S. Ondrejovič: 416-421*. Bratislava: Veda.

Köhler, Reinhard (2006b), The frequency distribution of the lengths of length sequences. In: J. Genzor, M. Bucková (eds.), *Favete linguis. Studies in honour of Victor Krupa. Bratislava: 145-152.* Slovak Academic Press.

Köhler, Reinhard, Naumann, Sven (2007), Quantitative text analysis using L-, F- and T-segments. In: Preisach, Burkhardt, Schmidt-Thieme, Decker (eds.), *Data Analysis, Machine Learning and Applications: 637-646*. Berlin, Heidelberg: Springer.

Wimmer, Gejza, Altmann, Gabriel (1999), *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Best, Karl-Heinz (1997b), Zum Stand der Untersuchungen zu Wort- und Satzlängen. In: *Third International Conference on Quantitative Linguistics. Helsinki, 172-176.*

Ejerhed, Eva (1988), Finding clauses in unrestricted text by finitary and stochastic methods. In: *Proceedings of the 2nd conference on applied natural language Processing, 219–227,* Austin, Texas.

Leffa, Vilson J. (1998), Clause processing in complex sentences. In: *Proceedings of the First International Conference on Language Resource & Evaluation, vol. 1, 937 – 943, May.*

Puscasu, Georgiana (2004), A Multilingual Method for Clause Splitting. In: *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. Birmingham, UK; http://clg.wlv.ac.uk/papers/puscasu-04a.pdf (Oct. 31, 2009).

Papageorgiou, Harris V. (1997), Clause recognition in the framework of alignment. In: Ruslan Mitkov and Nicolas Nicolov (ed.), *Recent Advances in Natural Language Processing: 417-425*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Sang, Erik F., Tjong, Kim, Déjean, Hervé (2001), Introduction to the CoNLL-2001 Shared Task: Clause Identification. In: Walter Daelemans & Rémi Zajac (ed.), *Proceedings of CoNLL-2001, 53–57*. Toulouse, France.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 46-50.*

Contemplations on Corpus Infinity

Jan Králík

Large corpora of texts brought unusual light into traditional quantitative methods of investigation. This light influenced also questions, which should be discussed by mathematicians. New questions concern already the basic axioms and the law of large numbers. The easiness of getting an amount of new statistical data from large corpora could underrate the theoretical base. But easiness itself could also be overestimated. In these contemplations I should like to show that doubts, which guarantee correct further steps, lead to careful re-valuation of some used terms and of their way of application.

[a]

The law of large numbers in its formulation by Bernoulli says over 300 years already, that – roughly expressed – relative frequencies possess a *limit value*. Such limit value can be identified with the conception of *probability*, if the axioms of the probability theory, as they have been introduced by Kolmogorov (1936) in the 30s of the 20th century, are used. Long satisfying experience with this identification, with its mathematical formulation as well as with exact proofs, did not lead to any doubts, nor to any discussion about its applicability, even in linguistics (Rényi 1972).

Mathematical research, independently, lead to the very important central limit theorem, roughly paraphrased: the sum of all distributions tends to one common law.

In a metaphor we could say: empirical distribution functions (and empiric distributions of random variables which are bound with them) possess a limit form. After some generalising steps, we could say in other words that common imaginations are fully right, when expecting – in simplest expression – more accurate information from larger samples.

This holds under some *conditions*, however. The essential condition is that the whole sample space would develop further – namely when going to infinity – in the same way as it did up to now. In the terms of the central limit theorem we should say: even when the sets of the observed random variables are distributed very differently, if their variances show "reasonable" behaviour, the corresponding limit distribution of standardised sums will be similar to the normal distribution as described by Gauss. The mentioned essential condition inconspicuously presupposes that we are staying in the space with homogeneous events, that we are staying in the space without extreme changes, where the "reasonable be-

haviour" of variance means that variance will not extremely exceed its presently known values.

Another one, not only inconspicuous, but even unexpressed idea is that infinity cannot be reached in any empirical way, nor any limit distribution comes to its real existence. A parallely deduced unexpressed idea would say, naturally, that *probability* must remain inscrutable, too.

The common conception of probability, however, is accepted without its close touch with the mathematical definition, even when applied in such a precisely measurable field as corpus linguistics. Here, the dangerous weakness of every simplification is multiplied by the extreme rise of the number of observed events. That is why the application of algorithms previously used in quantitative linguistics and interpretation of conclusions in corpus linguistics must be done with extreme care. Every detail in the mathematical formulation of the starting presumptions should be discussed anew, before probability theory would be applied. And similarly, conclusions based on empirical data from large corpora should be formulated most carefully (Králík 2001).

In linguistics, the presumptions enabling to apply the law of large numbers seem to be easily fulfilled because limited text sizes and limited inventories of elements or events are investigated. That is why the variance of relative frequencies around their mean values can be supposed as limited too, according to mathematical presumptions. There was no need to think about wider frames or further steps before large corpora became available.

The phenomenon of extremely large corpora introduced a dramatic change into this calm field. It was nearly forgotten that the law of large numbers has been postulated with respect to relative frequencies of events which are fully *repeatable*, or, in other words, with respect to events belonging to the same "equivalence class" in which such events appear, but they do not appear necessarily if the circumstances defining their "equivalence class" realise. In linguistics, such circumstances enabling the occurrence or non-occurrence of repeatable events may not be internal only, as, e.g. topical or contextual or generally bound to an entire text. In large corpora, the mentioned circumstances can be also external, as, e.g., bound to authors, genre, territories, time etc. Already because of this, some incomparability between the quantitative data computed from texts and from large corpora must exist (Králík 2004, 2006).

A brief contemplation concludes further that it is necessary to differ between two cases: (α) investigation of events, the inventory (repertoire) of which is *fixed* (such as phonemes, graphemes, syllables, grammatical categories, paradigms etc.) and (β) investigation of events, the inventory of which is generally *unlimited* (such as lexicon, word forms, phrase types, sentence graphs etc.). Such differentiation, in spite of its natural simplicity, opens another view on the principal idea on which the mathematical formulation of the limit behaviour is constructed.

This difference is well known, but it was mentioned occasionally only when e.g. the potentionality in language was discussed or when the idea of the theoretical category *nulax* was discussed (in an analogy to *hapax*) as theoretical number of elements which were not used or which were not present in the sample although they belonged to the set of possible eventualities.

It urges us to come back to the conception of probability as to the *limit value* to which the observed relative frequencies stream. We must accept that any text size, although it may be most extreme, remains limited in every respect, so that it cannot supply any "touch with infinity". Extreme text size can lead to an extremely fine but strict differentiation between the conceptions of relative frequency and probability. An extreme increase of the number of observed events need not necessarily bring extremely exact results because in large corpora, from the point of view of the investigated events, the multiplied sample space cannot be and is not homogenous. From the axiomatic definition it can be deduced that for every situation (as, e.g., time, author, text etc.) the corresponding probabilities are different. Generally, in linguistics, probabilities are not combinatorially countable numbers, nor are they exact relative values, but they must be seen as untouchable limit values in the mathematical sense of the law of large numbers, or – in the expression of a metaphor – as trends with strong influence.

Thus, if we accept that the limit values of probabilities differ for every text, for every author and for every time interval, a natural question appears: how should we understand relative data computed from large corpora? Numerical sequences and limits can be generally added and new mean values can be computed, but the same process cannot (must never) be done with their interpretations. Large corpora can be seen as a very useful sample space for events with a fixed inventory. For such cases, large corpora can offer a sufficiently accurate estimation for that part of language which they represent. Large corpora can be and they are most useful for many other directions in linguistic research even for events with unlimited repertoire. But quantitatively, for such events, large corpora can be seen as *set of texts* only. For such cases, large corpora represent the sum of many different representations. Therefore, a general interpretation of any relative frequencies remains impossible.

In a metaphor we could say that a single corpus offers one of the objective projections of a real situation of involved texts onto the grey horizon of infinity. Any new data from large corpora produce a new projection, drawing a sharp line, which fits well with the fuzzy silhouette we suspect at the infinite horizon. This silhouette, which itself does not possess a sharp contour – if we may go on in such metaphor – follows the top line of Gaussian mountains and hills.

In this situation, it may seem that for quantitative studies in large corpora, all the models of quantitative linguistics used up to now become insure, that all the up-to-now insignificant influences come to foreground, and we seem to be back at the beginning of all applications and interpretations.

Such an illusion may be alarming, but not hopeless. If the two principles, namely *axioms* and *laws* are taken in question, and if the presumptions are analysed again, new starting points may appear.

(1) It must be accepted that large corpora, in spite of their extreme size, are samples only. Because with respect to infinity, the sum of the texts remains fixed and limited when compared with all possible texts of the given language as a whole.

(2) A second fixed starting point is offered by answers to some questions around quantifications of usage and tolerance of deviations. Does the "extreme text-size" really mean the same as "towards infinity"? Is not the tolerance of 1 or 5 per cent of accuracy merely the tolerance over the extension of two decadic exponential steps? Is not the relation between the frequency of an event and the representativity of the sample mutually bound in the sense, that if the sample size rises extremely, it represents a different level of relative frequencies so that it already represents also different events?

Answers and solutions seem to be near. The law of large numbers does not model infinity, but it comments the finite cases from the point of view of infinity. It does not mean the challenge to go to infinity, but it shows and comments how the repeated observations would stream in a limit case. This law does not show, nor does it comment any single case, fixed situation or finite example. Similarly, the axiomatic definition of probability introduces a measure into a set of events. This measure is different within different circumstances, for different texts, among different authors, in different times.

Another metaphor could be offered: axioms and the law of large numbers offer the wisest conception of probability resembling the matchless limit mystery, which realises as far as in the hypothetic infinity. Large corpora, although they deal with extreme sample size, stream to these distant horizons much ambitiously than any previous quantitative studies. But they can never touch the limit mystery in fact.

[b]

To close these contemplations, yet another view should be mentioned briefly. The above quoted single cases, fixed situations or finite examples, when applied to different language levels, could be identified with levels of constructs in the sense of the Menzerath-Altmann law, as generalised by Hřebíček (Altmann 1980; Hřebíček 1997). In a comparable parallel, corpus could be seen as a *construct*, which stays very high over the level of texts. Natural texts could be taken as *constituents* of higher *sets of texts*, and such sets of texts could constitute – hypothetically in several further steps – a corpus. But such sets and further steps may not exist in reality, because they may not be intentional or because they may not be internally bound in any sense. This naturally means, however, that we

cannot be able to interpret quantitative characteristics of such sets in the traditional way, and the less we can be able to interpret the corresponding quantitative characteristics from large, inhomogeneous corpora.

Going towards higher text size, if texts *do not* form intended or natural higher sets, the majority of interpretations in traditional terms must *fail*. No sum or mean value (average) of quantitative characteristics computed from large corpora can be generally compared with traditional quantitative data from single texts or from *intended* sets of texts (intended corpora). Quantitative data from large, inhomogeneous corpora represent something completely different, that is not yet fully known.

Thus, also the Bochum-Trier School, as based on Menzerath-Altmann and Hřebíček, supports ideas which can be deduced from probability theory axioms and from the law of large numbers. These ideas correspond with intuitively quoted recommendation not to base quantitative studies blindly on disparate large corpora, but on more homogeneous data of one type. Smaller homogeneous corpora defined ad hoc could be a useful solution.

Meaningful interpretation of quantitative data from large corpora will come not before data from comparable large corpora will appear.

References

Kolmogorov, A. N. (1936). *Osnovnyje ponjatija teorii verojatnostej*. GITI, Moskva.

Rényi, A. (1962). *Wahrscheinlichkeitsrechnung*. Berlin: Deutscher Verlag für Wissenschaft.

Altmann, G. (1980). Prolegomena to Menzerath´s law. *Glottometrika* 2, *1-10*.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Králík, J. (2001). On quantitative characteristics of corpora approaching infinite size. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 149-152*. Trier: WVT.

Králík, J. (2004). Statistické úvahy nad rozsahem korpusu. *Jazyky a jazykověda: 267-271*. Praha: FF UK – ÚČNK.

Králík, J. (2006). Zamyšlení nad velkými výběry. In: Čermák, F., Blatná, R. (eds.), *Korpusová lingvistika: stav a modelové přístupy: 205-209.* Praha: NLN.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 51-60.*

Motif richness

Ján Mačutek

1. Introduction

A length motif (cf. Köhler 2006; Köhler, Naumann 2008) is a secondary unit consisting of a non-decreasing sequence[1] of lengths of the primary unit, e.g. a sequence of word lengths. For example, the sequence *"Length motifs are linguistic units behaving like other units"*, if word length is measured in terms of syllable numbers, consists of the motifs: 1-2, 1-3, 2-3, 1-2-2. In this paper we use word length motifs, but the results are applicable also to other units.

The main goal of this paper is to introduce and study motif richness as the ratio of the number of observed motifs to the number of all possible motifs, which is derived in Section 2. The problem was presented by Strauss, Fan and Altmann (2008). Theoretically derived results are applied to two Slovak texts, namely, one poem and one short story.

2. Maximum number of motifs

Denote L the length of the longest unit and M the length of the longest motif. For example, if $L = 3$ and $M = 3$ (which, even though not realistic, is short enough to list all possibilities), we can have the following 3-term non-decreasing sequences (we remind that we consider non-decreasing sequences in general here; see below for a discussion on the relationship between the number of non-decreasing sequences and the number of motifs):

1, 2, 3,
1-1, 1-2, 1-3, 2-2, 2-3, 3-3,
1-1-1, 1-1-2, 1-1-3, 1-2-2, 1-2-3, 1-3-3, 2-2-2 , 2-2-3, 2-3-3, 3-3-3.

Denote $N(sis,k,u)$ the number of strictly increasing k-term sequences (SIS) consisting of the numbers 1,2,..,u and $N(nds,k,v)$ the number of non-decreasing k-

[1] Exactly speaking, length motifs are the *longest possible* non-decreasing sequences of unit lengths. For example, the word length motifs in the first two verses of the Slovak national anthem (*Nad Tatrou sa blýska, hromy divo bijú*) are 1-2, 1-2-2-2-2. Without the condition of the *maximum* possible length of non-decreasing sequences we would face ambiguities, like, e.g., 1-2, 1-2-2, 2-2, etc.

term sequences (NDS) consisting of the numbers 1,2,…,*v* (in a NDS a number may occur more than once, e.g., for k = 2 we allow a sequence 2,2, etc). It holds

$$N(sis,k,n+k-1) = N(nds,k,n),$$

which can be proved as follows (it is to be reminded that the result is not new and it has long been known in combinatorial analysis).

Let $a_1, a_2, \ldots, a_k$ be a *k*-term NDS consisting of 1,2,...,*n*. Then $a_1, a_2+1, a_3+2, \ldots, a_k+k-1$ is a *k*-term SIS. Hence for every *k*-term NDS consisting of 1,2,..,*n* there is at least one *k*-term SIS consisting of 1,2,..,*n*+*k*-1, i.e., $N(nds,k,n) \le N(sis,k,n+k-1)$.

Now, let $a_1, a_2, , \ldots, a_k$ be a *k*-term SIS consisting of 1,2,..,*n*+*k*-1 (note that as it is a SIS, it holds $a_k \le n+k-1, \ldots, a_2 \le n+1, a_1 \le n$ and $a_j - a_{j-1} \ge 1$ for j = 2,3,…,*k*). Then $a_1, a_2-1, a_3-2, \ldots, a_k-(k-1)$ is a *k*-term NDS consisting of 1,2,..,*n*. Hence, for every *k*-term SIS consisting of 1,2,..,*n*+*k*-1 there is at least one *k*-term NDS consisting of 1,2,..,*n*, i.e., $N(sis,k,n+k-1) \le N(nds,k,n)$, which completes the proof.

Obviously, $N(sis,k,n+k-1) = \binom{n+k-1}{k}$ (we choose *k* numbers out of *n* and no permutations are allowed, as the sequence must be strictly increasing). As shown above, the number of all non-decreasing *k*-term sequences is the same, i.e.,

$$N(nds,k,n) = \binom{n+k-1}{k}.$$

As can be seen also from the example at the beginning of this section, if we consider non-decreasing sequences consisting of the numbers 1, 2 and 3 (i.e., *n*=3), there are $\binom{3+1-1}{1} = 3$ NDS's with the length 1, $\binom{3+2-1}{2} = 6$ NDS's with the length 2 and $\binom{3+3-1}{3} = 10$ NDS's with the length 3.

In our case, *n* is the length of longest unit (i.e., the length of the longest word in our examples), hence *n* = *L*. As the number *k* is the length of a sequence and *M* is the length of the longest motif, *k* attains the values 1,2,…,*M*. Hence the number of all possible non-decreasing sequences consisting of the numbers 1,2,…,*L* with not more than *M* terms - using the well known combinatorial identities $\binom{n}{k} = \binom{n}{n-k}$, $\sum_{j=k}^{n}\binom{j}{k} = \binom{n+1}{k+1}$ and supposing *L*>1) – is

$$\sum_{k=1}^{M} N(nds,k,L) = \sum_{k=1}^{M}\binom{k+L-1}{k} = \sum_{k=1}^{M}\binom{k+L-1}{L-1} = \sum_{j=L-1}^{L+M-1}\binom{j}{L-1} - \binom{L-1}{L-1} = \binom{L+M}{L} - 1.$$

Returning again to the example from the beginning of this section, if $L = 3$ and $M = 3$, we have $\binom{L+M}{L} - 1 = \binom{3+3}{3} - 1 = 19$ (i.e., there are 19 NDS's consisting of the numbers 1, 2 and 3 with the length not higher than 3).

The formula is not true if $L = 1$, but if a text is long enough, the probability that it contains only units with the length 1 is negligible.

The result obtained thus far (i.e., the number of all possible non-decreasing sequences consisting of numbers 1,2,…,*L* with not more than *M* terms) is not equal to the number of all possible motifs, as the motifs are restricted by further two conditions.

First, for every *k* there is one NDS ending with 1, namely the constant sequence containing only 1's. As *k* attains values from 1 to *M*, there are *M* NDS's consisting exclusively of 1's; however, only one of them can occur in a text (and, in addition, only as the last motif – with a possible exception of the last one, all motifs end with a number higher than 1). As a consequence, we must subtract *M*-1 from the number of NDS's.

Second, motifs beginning with *L* can occur only at the beginning of a text (a new motif means that a decrease was observed, i.e., the first number in the motif is preceded by a greater number in the previous motif, e.g., 1-1-2, 1-3, 2-3), hence only one of them can be observed. Again, there are *M* NDS's beginning with the number *L*, all of them being constant sequences consisting exclusively of *L*'s, which means that *M*-1 must be subtracted from the number of NDS's once more.

Hence, if the length of the longest unit is *L* and the length of the longest motif is *M* (and if $L > 1$), the number of all possible length motifs is

$$\binom{L+M}{L} - 1 - 2(M-1) = \binom{L+M}{L} - 2M + 1.$$

For example, if $L = 3$ and $M = 3$, there are $\binom{3+3}{3} - 2x3 + 1 = 15$ possible length motifs (only one of the sequences 1, 1-1, 1-1-1 can occur as the last motif and only one of the sequences 3, 3-3, 3-3-3 as the first motif).

3. Applications

We analyze word length motifs in two Slovak texts, namely, in a poem *Po ceste meteora* by Rudolf Dilong and in a short story *Z teplého hniezda* by Martin Kukučín. Word length is measured in syllables, zero-syllable words are not considered (cf. Antić, Kelih and Grzybek 2006). Motif length is measured in words and in syllables. Abbreviations were re-written using full (i.e., non-abbreviated) words. Obvious typo errors in the texts were corrected.

Our data support Köhler's hypothesis that length motifs and the units they consist of behave according to the same laws (cf. Best 2005 for an overview of word length models). Köhler (2006) theoretically derived and successfully fitted the hyperpascal distribution as a model for motif length measured in words and the negative binomial distribution (slightly modified by its left truncation) as a model for motif length measured in syllables (cf. Wimmer and Altmann 1999 for both distributions). We follow his approach here. As can be seen in the following tables, the short story data provide another corroboration of the theory; on the other hand, the analysis of the poem gives not so convincing results. While the goodness-of-fit for the hyperpascal distribution fitted to motif length measured in words is somewhere on the edge (C = 0.0208), the modified negative binomial distribution fitted to motif length measured in syllables yields C = 0.0288. From among possible explanations we mention the sample size (which is quite modest when compared with that of Köhler 2006) and the genre – rhythm in poetry may have a strong impact. E.g., it can be reasonable to respect stanzas and to split motifs which would otherwise overreach from a stanza to the next one. In any case it is necessary to analyze more texts from different languages and genres before a conclusion can be formulated.

Köhler and Naumann (2008) presented (and successfully tested on 66 German texts) the hypothesis that the rank-frequency distribution of length (and other) motifs behaves similarly to the rank-frequency distribution of words. The two Slovak texts analyzed in this paper are no exception, the Zipf-Mandelbrot distribution (cf. Baayen 2005 for the distribution as a model of word frequency, and also Wimmer and Altmann 1999) fits the data excellently.

We define the motif richness as the ratio

$$MR = \frac{number\ of\ observed\ motives}{number\ of\ all\ \ possible\ motives},$$

where the number of all possible motifs is $\binom{L+M}{L} - 2M + 1$ (with L being the length of the longest unit and M the length of the longest motif, cf. Section 2 for the derivation). For Dilong's poem *Po ceste meteora* we have $L = 5, M = 10$ and

90 observed length motifs, resulting in $MR = \frac{90}{2982} = 0.0302$. Similarly, for Kukučín's short story *Z teplého hniezda* with $L = 7, M = 12$ and 155 observed motifs we obtain $MR = \frac{155}{50365} = 0.0031$.

Table 1
Word length and motif length in a Slovak poem (2782 words, 859 motifs)

Rudolf Dilong: *Po ceste meteora*						
	word length (syllables) hyperpascal distribution		motif length (words) hyperpascal distribution		motif length (syllables) doubly truncated negative binomial distribution	
x	*f(x)*	*NP(x)*	*f(x)*	*NP(x)*	*f(x)*	*NP(x)*
1	1102	1097.28	96	99.67		
2	1214	1212.34	248	254.55	86	83.62
3	353	353.80	196	198.18	120	111.06
4	112	90.15	140	128.74	127	123.15
5	1	28.44	91	77.54	133	120.38
6			55	44.80	79	107.11
7			18	25.21	86	88.60
8			6	13.93	52	69.13
9			7	7.60	56	51.41
10			2	8.77	52	36.73
11					32	25.37
12					15	17.01
13					12	11.12
14					4	7.11
15					3	4.46
16					2	2.75
	k = 0.4923 m = 0.0955 q = 0.2142 χ^2 = 31.80, DF = 1 P(χ^2) ≈ 0 C = 0.0114		k = 0.7997 m = 0.1567 q = 0.5004 χ^2 = 17.83, DF = 6 P(χ^2) = 0.0067 C = 0.0208		k = 6.8243 p = 0.5485 L = 2, R = 16 χ^2 = 24.76, DF = 9 P(χ^2) = 0.0032, C = 0.0288	

Table 2
Word length and motif length in a Slovak short story (4315 words, 1440 motifs)

Martin Kukučín: *Z teplého hniezda*						
	word length (syllables) hyperpascal distribution		motif length (words) hyperpascal distribution		motif length (syllables) left truncated negative binomial distribution	
x	*f(x)*	*NP(x)*	*f(x)*	*NP(x)*	*f(x)*	*NP(x)*
1	1767	1809.17	178	189.03		
2	1572	1541.58	480	482.34	145	151.62
3	698	646.79	351	340.89	210	200.80
4	248	221.25	207	201.01	252	219.81
5	26	68.52	118	110.01	214	210.34
6	3	20.00	66	57.80	162	181.98
7	1	7.70	20	29.59	120	145.50
8			15	14.88	108	109.18
9			2	7.39	71	77.75
10			2	3.63	52	52.99
11			0	1.77	45	34.78
12			1	1.65	24	22.11
13					16	13.66
14					10	8.24
15					3	4.86
16					1	2.81
17					0	1.60
18					6	0.89
19					1	1.06
k = 1.6117 m = 0.4365 q = 0.2308 χ^2 = 55.5360 DF = 3 P(χ^2) ≈ 0 C = 0.0129			k = 0.7571 m = 0.1355 q = 0.4567 χ^2 = 12.2550 DF = 6 P(χ^2) = 0.0565 C = 0.0085		k = 7.7822 p = 0.5939 L = 2 χ^2 = 16.7547 DF = 10 P(χ^2) = 0.0800 C = 0.0116	

Table 3
Fitting the Zipf-Mandelbrot distribution to rank-frequency distributions of length motifs

Dilong: *Po ceste meteora*		Kukučín: *Z teplého hniezda*	
x	*f(x)*	*x*	*f(x)*
1	110	1	179
2	86	2	145
3	71	3	132
4	40	4	65
5	37	5	54
6	29	6	53
7	27	7	53
8	26	8	51
9	25	9	43
10	24	10	42
11	22	11	31
12	22	12	28
13	21	13	23
14	21	14	22
15	18	15	22
16	16	16	20
17	14	17	19
18	13	18	19
19-21	12	19	17
22	11	20	16
23-24	10	21	16
25	9	22	14
26-28	8	23	13
29-30	7	24	12
31	6	25-28	10
32-34	5	29-33	9
35-40	4	34-37	8
41-48	3	38-40	7
49-60	2	41-43	6
61-90	1	44	5
		45-51	4
		52-68	3
		69-92	2
		93-155	1
χ^2 = 26.6945 DF = 86 P(χ^2) = 1	a = 1.6772 b = 5.0834 n = 90	χ^2 = 31.1597 DF = 134 P(χ^2) = 1	a = 1.7120 b = 5.3058 n = 155

The maximum possible number of motifs can grow very quickly with increasing *L* and *M* (depending on their difference), hence another characteristic of motif richness, namely a sequence which will be called *partial motif richness* sequence (*PMR* in the following) can be more appropriate to compare two texts. The *PMR* is an *M*-term sequence defined as follows:

$$PMR_j = \frac{number\ of\ observed\ motives\ with\ the\ length\ j}{number\ of\ all\ possible\ motives\ with\ the\ length\ j},\ j = 1,2,\ldots,M$$

.

We remind that for a fixed *L* the number of all possible motifs with the length *j* (*j*=1,2,…, *M*) is $\binom{L+j-1}{j}$, cf. Section 2. Although a *PMR* consists of proportions, which are numbers between 0 and 1, it is not a probability distribution, namely, its sum is not normalized.

Obviously, also the frequency spectrum (i.e., numbers of observed motifs with particular lengths) behaves very regularly. Having only two samples, we postpone any research in this direction until more data are available.

The *PMR*'s for Dilong (a poem) and Kukučín (a short story) are compared on the following figure. One can easily see quite different curve shapes. At the first sight it seems that the sequence can be used for classification purposes, discrimination analyses, etc. Of course, also in this case more data are needed to judge its appropriateness and to find a general model (i.e., to fit a theoretical curve).

Table 4
Partial motif richness

	Dilong: *Po ceste meteora*			Kukučín: *Z teplého hniezda*		
Motif length	maximum	observed	PMR_j	maximum	observed	PMR_j
1	5	2	0.4000	7	4	0.5714
2	15	7	0.4667	28	9	0.3214
3	35	11	0.3143	84	21	0.2500
4	70	15	0.2143	210	27	0.1286
5	126	16	0.1270	462	30	0.0649
6	210	15	0.0714	924	32	0.0346
7	330	11	0.0333	1716	17	0.0099
8	495	6	0.0121	3003	11	0.0037
9	715	5	0.0070	5005	2	0.0004
10	1001	2	0.0020	8008	2	0.0002
11				12376	0	0
12				18564	1	0.0001

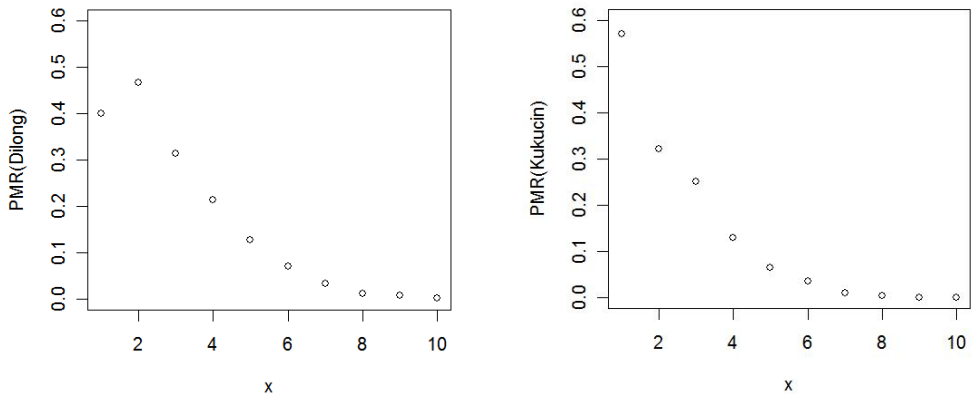


Figure 1. Partial motif richness

Acknowledgement
Supported by the research grant VEGA 1/3016/06.

Texts

Dilong, R. Po ceste meteora. http://zlatyfond.sme.sk/dielo/257/Dilong_Spod-Juzneho-kriza/12 (accessed 22-07-2008).

Kukučín M. Z teplého hniezda. http://zlatyfond.sme.sk/dielo/109/Kukucin_Z-tepleho-hniezda/1 (accessed 22-07-2008).

References

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 117-156.* Dordrecht: Springer.

Baayen, R.H. (2005). Word frequency distributions. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 397-409.* Berlin: de Gruyter.

Best, K.-H. (2005). Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 260-273.* Berlin: de Gruyter.

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: 145-152.* Bratislava: Slovak Academic Press.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Ch., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-645.* Berlin: Springer.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.

Wimmer, G., **Altmann, G.** (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 61-75.*

Content words in authorship attribution: An evaluation of stylometric features in a literary corpus

George K. Mikros

1 Introduction

Quantitative authorship attribution has regained interest the last years mainly due to progress made in machine learning methods and advancements in the computational analysis of corpora. Research made in the framework of information retrieval has produced a wide range of techniques which have been transplanted with great success in Natural Language Processing (NLP) classification problems such as automatic topic categorization, automatic genre classification, spam detection etc. Authorship attribution viewed as NLP classification problem has been a productive research field with impressive results in classification accuracy as long as the training corpus has sufficient size and is limited to only a handful of authors.

Although stylometric methods were traditionally applied to literary and religious texts in order to identify the real author behind a work of disputed or unknown authorship, nowadays they have been used in all kind of text genres, including news (Mikros, 2006), emails (de Vel et al., 2001), blogs (Schler et al., 2006), USENET news (Argamon et al., 2003) etc. The analysis of different text genres requires different sets of stylometric variables and for this reason all kinds of textual characteristics[1] have been used in order to maximize correct authorship attribution. Furthermore, a wide variety of statistical methods[2] have been applied in order to achieve the maximum discriminatory power between different authors in each study.

[1] Rudman (1997, p. 360) states that approximately 1,000 style markers have already been isolated and used in authorship attribution studies.

[2] The palette of statistical methods used in authorship attribution studies is enormous and contains among others, multinomial logistic regression (Madigan et al., 2005), discriminant function analysis (Mikros, 2006), principal component analysis (Burrows, 1992), neural networks (Tweedie et al., 1996), support vector machines (Diederich et al., 2003), bayes classifier (Clement & Sharp, 2003), decision trees (Koppel & Schler, 2003), genetic algorithms (Holmes & Forsyth, 1995).

2 Corpus used in the study

2.1 Corpus description

The field however, despite the advancements in the quantitative methodology described above, has not overcome some fundamental problems regarding the reliability of the results. One apparent deficiency is the lack of a well organized, standardized corpus for authorship attribution benchmarking similar to the corpora existing for topic categorization and information retrieval[3]. Some efforts (Juola et al., 2006) have been put forward in order to create such test corpora, but until now they have not gained wide recognition.

Another problem related with the corpora used in the authorship attribution studies is the lack of their homogeneity. As Rudman (1997) states the most striking deficiencies are:

- The improper selection, unavailability or fragmentation of the texts.
- The text normalization that often applies from the editor or the publisher causing serious distortion in the writer's style.
- The cross-validated texts should be controlled for genre, topic, date and medium when comparing to the training texts.

For this reason we decided to compile a corpus which would contain complete versions of literary texts from the same time period and would be strictly controlled for editorial normalization. We selected five novels from four widely known modern Greek writers published by the same publishing house (Kastaniotis Publishing House). All the novels were best-sellers in the Greek market and belong to the "classics" of the Modern Greek literature. More specifically, the corpus consists of:

1. *The mother of the dog*, 1990, by Matesis [47,929 words].
2. *Murders*, 1991, by Michailidis [72,831 words].
3. From the other side of the time, 1988, by Milliex [78,077 words].
4. *The dead liqueur*, 1992, by Xanthoulis [28,602 words].
5. *Dreams*, 1991, by Milliex [9,796 words] - Test novel.

The first four novels formed our main training corpus while the fifth novel (*Dreams*) has been kept out as a hold-out sample.

2.2 Experimental methodology

The main research question posed in this study is whether a specific set of lexical features named Author-Specific Words (ASW) (Mikros, 2006, 2007) perform better in authorship attribution than traditional function words and other kind of stylometric variables. In order to fully evaluate the performance of this feature

[3] Topic categorization research has widely used the Reuters Corpus (Rose et al., 2002) and information retrieval community the TREC text collections.

set we measured authorship attribution accuracy in our test corpus, as well as in a number of subcorpora which were created in order to measure the effect of two independent variables:

- **Text size**: We want to measure whether the ASW feature set performs better compared to the others feature sets in varying text sizes.
- **Corpus fragmentation**: We want to estimate the effect of the corpus fragmentation in authorship attribution accuracy and compare the performance of the ASW to the other feature sets as the percentage of fragmentation increases. We define corpus fragmentation as the percentage of the texts existing in the original corpus after removing randomly *n* percentage of texts, where *n* gradually increases in steps of 20%.

In order to experimentally verify the usefulness of the proposed feature set we created a number of subcorpora which were systematically controlled for the above mentioned two variables. The steps for the creation of our subcorpora are the following:

1. Slice each novel in text segments of varying size (50 – 100 – 200 – 500 words).
2. Create 4 different corpora for each size (50 words corpus, 100 words corpus etc.).
3. Subdivide each corpus further using random sampling in the text segments and create 4 extra subcorpora (20%, 40%, 60%, and 80% of text segments compared to the original corpus).
4. Calculate in each of the 20 resulting subcorpora 3 different feature groups:
    a. Common stylometric variables (ST)
    b. Frequent function words (FFW)
    c. Author-specific words (ASW)
5. Use Discriminant Function Analysis (DFA) in order to obtain authorship classification for each text segment in each subcorpus.

The size (number of text segments) of the resulting subcorpora is displayed in Table 1.

Table 1
Size of subcorpora (number of text segments) produced by controlling text segment size and percentage of corpus fragmentation

	Corpus fragmentation % of the original sample remained after random deletion of fixed percentage of text segments				
Text size	*20%*	*40%*	*60%*	*80%*	*100%*
50 words	979	1,901	2,788	3,773	4,736
100 words	498	929	1,453	1,896	2,367
200 words	209	491	718	955	1,181
500 words	87	191	309	380	471

The different subcorpora produced vary considerably in size ranging from 87 texts in the 500 words segments and 20% texts of the original corpus to 4,736 texts in the 50 words segments and 100% of the texts. Furthermore, all our text sizes are below 1,000 words[4] and we have included in the examined subcorpora 50 words and 100 words text segments, untypical small sizes for texts in literary authorship attribution pushing our methodological tools to their limits. Thus we have created a diverse and demanding experimental environment for adequately comparing the performance of the different feature sets in authorship attribution.

3 Feature sets

In this study we used three broad sets of stylometric features which contain both lexical and sublexical units. The first set, called Common Stylometric Variables (ST), includes some of the most frequent stylometric features utilized in authorship attribution research, such as Yule's K index of vocabulary richness and various word length measures. The second set, called Frequent Function Words (FFW) contains lexical variables that belong to the vocabulary class of the function words and have also been used extensively in modern authorship attribution research. The third set, called Author-Specific Words (ASW) contains content words that appear to be author-specific and their extraction will be described in section 4.2 of this paper. All the features used in this study are the following:

1) Common Stylometric Variables (ST)

i) Lexical "richness"

(a) Yule's K: Vocabulary richness index that exhibits stability in different text sizes (Tweedie & Baayen, 1998).

(b) Lexical Density: The ratio of functional to content words frequencies in the text, also known as Functional Density (Miranda & Calle, 2007).

(c) % of Hapax- and Dis- legomena: The percentage of words with frequency 1 and 2 in the text segment.

(d) Dis-/Hapax- legomena: The ratio of dis-legomena to hapax-legomena in the text segment, indicative of authorship style (Hoover, 2003).

(e) Relative entropy: Is defined as the quotient between the entropy of the text and its maximum entropy multiplied by 100. Maximum entropy for a text is calculated if we assume that every word appears with frequency 1 (Oakes, 1998: 62).

ii) Character level measures

[4] Research has shown that most stylometric features exhibit stability in text sizes over the 1,000 words (Juola et al., 2006, p. 172; Stamatatos et al., 1999, p. 208; Zhao & Zobel, 2007)

iii) Frequency of characters: The frequency of each letter in the text segment normalized in 1,000 words sample. We measured in total 31 letters (we calculated separately the frequencies of the stressed and the unstressed vowels since in Modern Greek spelling the stressed vowels have stress marked orthographically, thus representing different grapheme).

iv) Word level measures

(a) Average word length (per text) measured in letters.

(b) Word length distribution: The frequency of words of 1, 2, 3 … 14 letters long normalized in 1,000 words sample.

2) Frequent Function Words (FFW)

The frequency of the 80 most frequent function words, normalized in 1,000 words sample.

3) Author Specific Words (ASW)

The frequency of the 80 most distinctive content words (20 per author), normalized in 1,000 words sample.

4 Lexical variables in authorship attribution studies

4.1 Function words

There is a general consensus among researchers that function words are among the best discriminators of authorship style. The major advantages of using them are:

- They lack of prominence and for this reason their usage cannot be imitated.
- They belong to a universal closed category of vocabulary that have specific syntactic role and can be easily enlisted.
- Most of them belong to the most frequent words in all languages. For this reason they can be counted even in small portions of text avoiding the sparse matrices problems.
- Because of their syntactic role and their high frequency of occurrence they can reveal subconscious mechanisms of language variation, which are unique to each author.

Mosteller & Wallace (1984) were among the first to search for text attributes that were systematically topic-neutral. They ended up using specific function words, which had high frequency of occurrence and at the same time remained corpus-independent. However, function words were widely recognized as the most reliable feature set for authorship attribution, through the studies of Burrows and his coworkers (Burrows, 1989, 1992; Burrows & Craig, 1994; Burrows & Hassal, 1988). Their basic method consists in using the frequency of occurrence of the most frequent words of a corpus as variables in a multivariate analysis such

as Principal Component Analysis (PCA) or Cluster Analysis. PCA for example can be used to position texts in a two-dimensional map using the loadings of the first two principal components of the analysis. In most cases the texts grouped together in this graph have been written by the same author. The results of these studies were particularly successful in a wide spectrum of disputed authorship problems and were quickly adopted by others including Holmes & Forsyth (1995), Holmes (1992), Hoover (2001), Tweedie et al., (1998). A recent study (Koppel et al., 2003), using different experimental methodology from Burrows, concluded that function words are indeed the best candidates for a universal, corpus-independent feature set for authorship attribution. The researchers used the measure of "stability", which represents quantitatively the degree of available synonymy of a specific linguistic item. Function words are unstable, in the sense that they can be substituted easily in a passage, without affecting the meaning of the text. Argamon & Levitan (2005) arrive at the same conclusion by comparing the discriminatory power of the function words to other two sets of stylometric variables (frequent pairs of words and frequent collocations).

Despite the wide acceptance of function words as authorship indicators, there are studies which demonstrate some serious drawbacks in using them in authorship attribution research. The oldest and most influential criticism was posed by Damerau (1975) who summarized the main deficiencies of the function words in the following arguments:

- The notion of function words is insufficiently precise to permit determination of what words to study.
- Not all function words are context insensitive.
- The distribution analysis of the function words revealed that many of them are not randomly distributed.

Furthermore, there are studies that correlate function word usage mainly to genre and less to authorship. Binongo (1994) analyzing the style of the Filipino writer Joaquin found that the PCA analysis of the function words he performed revealed mainly genre differences and to a lesser extent authorial one. The same conclusion has been reached by Baayen et al. (1996) who observed that using function words as stylometric variables facilitates documents to cluster more successfully in means of text genre than authorship. They conclude that the most frequent syntactic structures are more effective authorship discriminators than the lexical ones. In a recent study Mikros & Argiri (2007) examined the topic and the authorship information carried in the 10 most frequent function words of Modern Greek. Statistical analysis revealed that half of these words do not have any discriminatory power over author or topic. From the remaining five, only the most frequent one discriminates exclusively authorship, while the others distinguish both author and topic. These results show that, although function words are indeed semantically free, they do however contribute indirectly to the meaning of the text. This is happening probably through syntax and discourse level, since

many function words construct phrase complexity and build cohesion patterns, which can indirectly be linked with topic information.

On the other hand many recent studies have found evidence that content words carry stylistic information suitable for authorship attribution, among others Baayen et al. (2002), Hoover (2004), Lancashire (1999).

The selection of content words in authorship attribution has been implemented using various methods:

- Word distinctiveness ratio (Ellegård, 1962)
- Mutual information (Luyckx & Daelemans, 2005)
- Information gain on classification categories (Schler et al., 2006)

In the next section (4.2) we will present a method for selecting automatically content words that provide the maximum discriminatory power between the authors under study.

4.2 Author-Specific Words (ASW)

In Mikros (2003) we introduced a frequency profiling method in order to select content words for using them as discriminating variables in text categorization tasks. In later studies (Mikros, 2006, 2007) we extended this method in order to detect author-specific words. The procedure we proposed is analyzed in seven discrete steps, explained briefly as follows:

1) Selection of the training corpus.
2) Formation of homogeneous subcorpora regarding the author of the included texts.
3) Creation of frequency wordlists (FWL) for each of the subcorpora (for example Author A FWL, Author B FWL, Author C FWL, and Author D FWL).
4) Comparison of each FWL with the unified FWL of the remaining authors, i.e., comparison of Author A FWL with the FWL which has been created joining Author B, Author C, and Author D FWLs.
5) Extraction of the *k* most author-characteristic words that exhibit maximum discriminating power between his vocabulary and the vocabulary of the other authors. The extraction is performed using Log-Likelihood measure.
6) Repetition of the procedure (stages 4 & 5) by deploying the remaining combinations of the available FWL comparisons.
7) Extraction of *n* words (in the previous example 4 X *k*) which can be used as Author-Specific lexical variables in an authorship attribution training set.

For the needs of our study we performed this methodology and we extracted 80 ASW (20 words per author). For every one of these words we calculated its frequency in each text of the corpus.

The resulting wordlists contain words that characterize the vocabulary of each author and at the same time discriminate it mostly from the vocabulary of the other authors. However, it should be noted that this methodology is designed

to address closed authorship problems, which the real author is one of a specific list of candidates.

5 Classification method

In order to explore the discriminating power of the selected variables in authorship attribution we used Discriminant Function Analysis (DFA). DFA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a priori defined groups. Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance (Hair Jr et al., 1995, p. 244). If the dependent variables have more than two categories, DFA will calculate *k*-1 discriminant functions, where *k* is the number of categories. Each function allows us to compute discriminant scores for each case for each category, by applying the formula:

$$D_{jk} = \alpha + w_1 x_{1k} + w_2 x_{2k} + ... + w_i x_{ik} ,$$

where

D_{jk} = Discriminant score of discriminant function *j* for object *k*
a = intercept
w_i = Discriminant weight for the independent variable *i*
x_{ik} = Independent variable *i* for object *k*

In order to validate the DFA results we used 2 different methods:

***U*-method**: it is based on the "leave-one-out" principle (Huberty et al., 1987). Using this method, the discriminant function is fitted to repeatedly drawn samples of the original sample. Estimates *k*-1 samples, eliminating one observation at a time from a sample of *k* cases.

Test novel: For one author (Milliex) we used a second novel, not included in the training data, in order to evaluate the classifier's accuracy in unforeseen data from the same author. The specific test resembles more closely to a real life authorship attribution problem.

6 Results

The best classification results were obtained by the ASW method using the lengthiest text segments (500 words). The overall authorship attribution accuracy was 97.8% for the *U*-method and 94.7% for the test novel. The detailed confusion matrix for the *U*-method is shown in Table 2.

Table 2
Confusion matrix of authorship attribution using ASW feature set and 500 words text segments

		Matesis	**Michailidis**	**Milliex**	**Xanthoulis**	**Total**
Count	**Matesis**	**93**	0	2	0	95
	Michailidis	0	**138**	7	0	145
	Milliex	0	1	**154**	0	155
	Xanthoulis	0	0	0	**57**	57
%	**Matesis**	**97.9**	0	2.1	0	100
	Michailidis	0	**95.2**	4.8	0	100
	Milliex	0	0.6	**99.4**	0	100
	Xanthoulis	0	0	0	**100**	100

In order to have a clear view of the performance of the three feature sets we used all the experimental datasets (c.f. Table 1) and for each one we measured authorship attribution accuracy separate for each feature set and each cross-validation method. The averaged results from the two validation methods can be seen in Chart 1 below.

The chart clearly shows that ASW method outperforms all other feature sets in all text sizes. In order to confirm this, we performed a two-way ANOVA with dependent variable the authorship attribution accuracy and factors, text size and feature set. The main effect of the feature set was $F(2, 24) = 26.08$, $p < 0.001$, such that the average authorship attribution accuracy across all text sizes was significantly higher for ASW, (M = 86.7%, SD = 8.2) than for FFW (M = 70.6%, SD = 13.3) and ST (M = 64.9%, SD = 8.4). The main effect for the text size was also found statistically significant, $F(3, 24) = 13.9$, $p < 0.001$. In order to analyze further the effect of the text size on the authorship attribution accuracy we performed the multiple comparisons HSD Tukey test. The results showed that 50 and 100 words texts display homogenous behavior and contrast to the 200 and 500 words texts, meaning that text segments below 100 words conceal authorship information and need to be modeled in a different way. The interaction effect was not significant, $F(6, 24) = 22.5$, $p > 0.05$.

Since ASW displays better results in authorship attribution related to the other feature sets we decided to analyze its performance more systematically. Therefore, we plotted the ASW authorship attribution accuracy over the different text sizes as can be seen in the Chart 2 below.

In addition we performed curve fitting and the best fit ($R^2 = 0.96$) was obtained using the following quadratic equation:

$$ASW\ accuracy = 68.9 + 0.159*(\text{Text Size}) - 0.000212*(\text{Text Size})^2.$$

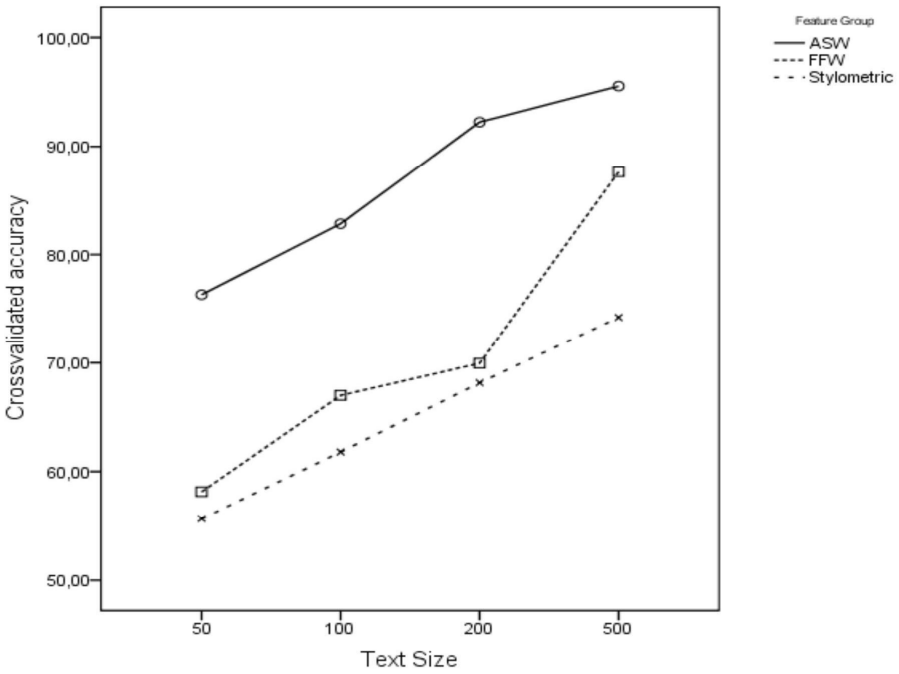


Chart 1. The effect of text size on classification results for different feature sets

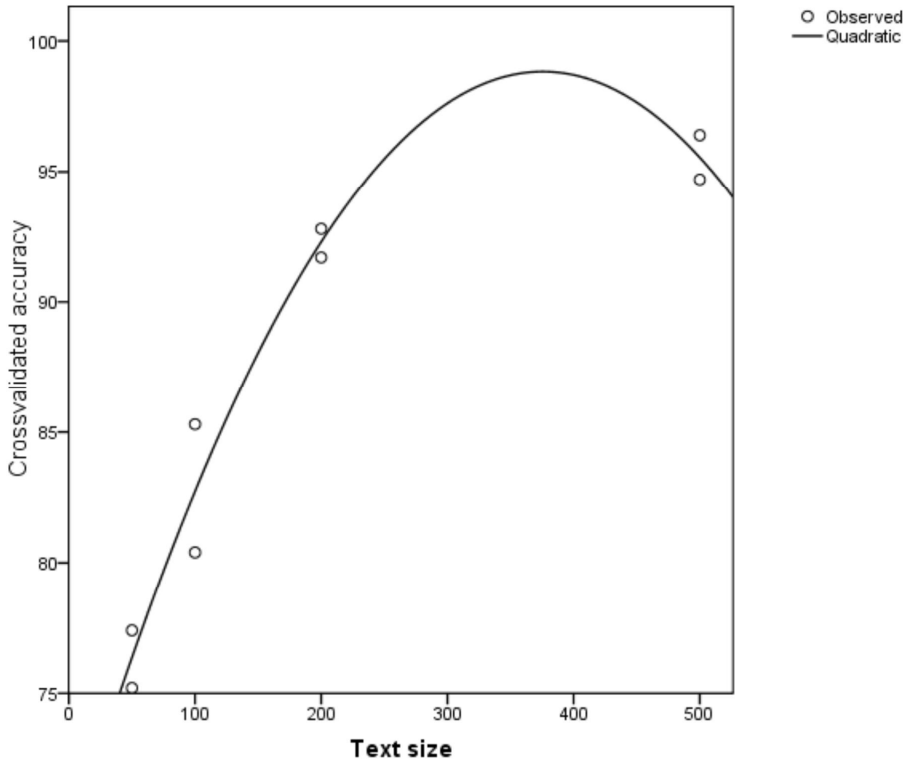


Chart 2. The quadratic fit of text size on ASW authorship attribution accuracy

The relationship of text size and ASW method performance is obviously non-linear. In fact the ASW performance boosts in the 50, 100 and 200 words text and flattens out in the 500 words texts still remaining better than the other two feature sets. This is a highly desirable characteristic since many critical applications of authorship attribution (e.g. forensic linguistics) use corpora with extremely short texts.

The same steps were followed in the study of the effect of corpus fragmentation in the performance of the three feature sets. The results can be seen graphically in Chart 3 below.

The chart confirms that ASW method outperforms all other feature sets in all samples sizes. However, the general effect of corpus fragmentation in authorship attribution accuracy is negligible. This impression was further corroborated by a two-way ANOVA with dependent variable authorship attribution accuracy and factors, corpus fragmentation and feature set. The main effect of the feature set was $F(2, 60) = 18.015$, $p < 0.001$, such that the average authorship attribution accuracy across all sample sizes was significantly higher for ASW (M = 86.2%, SD = 8.5) than for FFW (M = 74.2%, SD = 10.9) and ST (M = 64.8%, SD = 10.8). The main effect of corpus fragmentation was not found significant, $F(4, 60) = 0.34$, $p > 0.05$, meaning that corpus fragmentation does not influence authorship attribution accuracy even when the training sample is 80% reduced related to the original. The interaction effect was also not found significant, $F(8, 60) = 0.009$, $p > 0.05$.

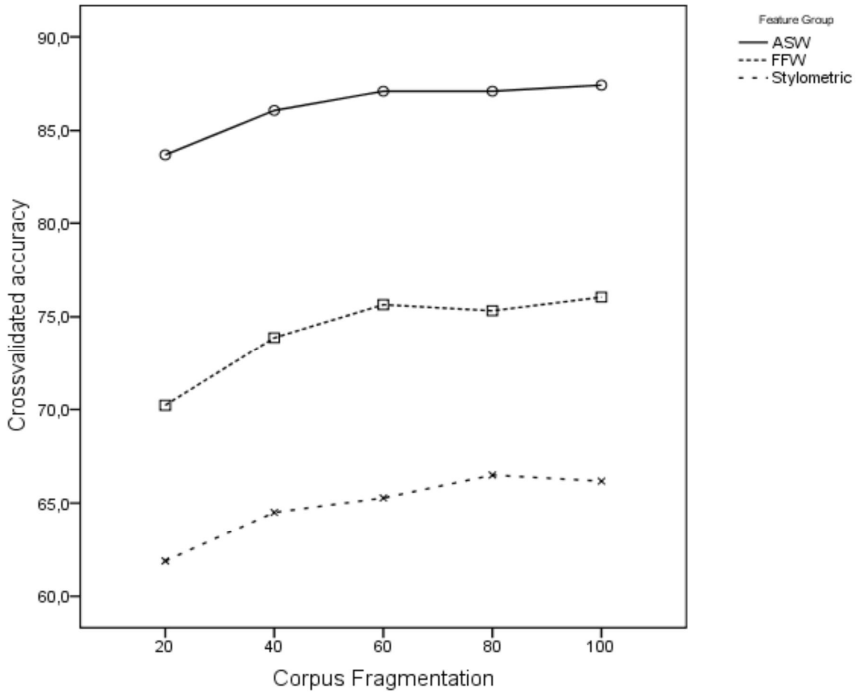


Chart 3. The effect of corpus fragmentation on classification results for different feature sets

7 Conclusions

This study compared three different feature sets and measured their performance in authorship attribution accuracy using literary corpus. The results showed that feature sets that contain content words should not be excluded from authorship attribution research since they can offer significant improvements on the attribution accuracy. More specifically ASW feature set can be used effectively in authorship attribution as long as there is a closed set of candidate authors. Our most important findings are summarized below:

- Authorship "genome" exists even in small text segments of 100 and 50 words especially if we take into consideration the frequency patterns of content words.
- The size of a text segment exhibits linear correlation with the precision of authorship attribution. However, if we examine only the ASW features the best fit is obtained by a quadratic function.
- Variations in sample size (from 20% to 100%) did not have an effect on the authorship attribution accuracy in a statistically significant way.
- ASW method outperforms the FFW and ST methods in all experimental conditions and performs particularly well in small text sizes.

The above results do not undermine the role of the function words in authorship attribution, since their value has been tested in many relevant studies. They remind us however, that authorship style is a mixture of unconscious and intentional language behavior. In order to capture fully this complex construct we need to correlate successfully both unconscious linguistic choices and purposeful selection of linguistic units.

Future research will be oriented to test ASW in authorship attribution problems with larger number of author candidates (> 20) and multilingual corpora.

Acknowledgments

The present study has been funded and published in the framework of the research programme **PYTHAGORAS I** which is co-funded by the European Social Fund (75%) and National Resources (25%) - Operational Program for Educational and Vocational Training II (EPEAEK II).

References

Argamon, S., Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC conference*, Victoria, BC, Canada.

Argamon, S., Šaric, M., Stein, S.S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 475-480.

Baayen, H., van Halteren, H., Neijt, A., Tweedie, F. (2002). An experiment in authorship attribution. *Actes des 6 Journées internationales d'analyse statistiques des donnés textuelles (JADT 02)*.

Baayen, H.R., van Halteren, H., Tweedie, F.J. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing 11*(3), *121-132*.

Binongo, J.N.G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: A statistical expression of a Filipino writer's style. *Literary and Linguistic Computing 9*(4), *267-279*.

Burrows, J.F. (1989). 'A vision' as a revision. *Eighteenth Century Studies 22, 551-565*.

Burrows, J.F. (1992). Computers and the study of literature. In: *Computers and Written Texts: 167–204*. Oxford: Blackwell.

Burrows, J.F., Craig, D.H. (1994). Lyrical drama and the "turbid mountebanks": Styles of dialogue in romantic and renaissance tragedy. *Computers and the Humanities 28*(2), *63-86*.

Burrows, J.F., Hassal, A.J. (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives. *Eighteenth Century Studies 21, 427-453*.

Clement, R., Sharp, D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing 18*(4), *423-447*.

Damerau, F.J. (1975). The use of function words frequencies as indicators of style. *Computers and the Humanities 9, 271-280*.

de Vel, O., Anderson, A., Corney, M., Mohay, G. (2001). Multi-Topic E-mail Authorship Attribution Forensics. *Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*, Philadelphia, PA, USA.

Diederich, J., Kindermann, J., Leopold, E., Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence 19*(1), *109-123*.

Ellegård, A. (1962). *A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772*: Acta Universitatis Gothoburgensis.

Hair Jr, J.F., Anderson, R.E., Tatham, R.L., Black, W.C. (1995). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ, USA: Prentice-Hall.

Holmes, D.I., Forsyth, R.S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing 10*(2), *111-127*.

Holmes, D.I. (1992). A Stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society, Series A, 155*(1), *91-120*.

Hoover, D. (2003). Another perspective on vocabulary richness. *Computers and the Humanities 37*, 151-178.

Hoover, D. (2004). Testing Burrow's Delta. *Literary and Linguistic Computing 19*(4), *453-475.*

Hoover, D.L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing 16*(4), *421-444*.

Huberty, C.J., Wisenbaker, J.M., Smith, J.C. (1987). Assessing Predictive Accuracy in Discriminant Analysis. *Multivariate Behavioral Research 22*(3), *307-329*.

Juola, P., Sofko, J., Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing 21*(2), *169-178.*

Koppel, M., Akiva, N., Dagan, I. (2003). A corpus-independent feature set for style-based text categorization. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, *1263-1276.*

Koppel, M., Schler, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, *69-72.*

Lancashire, I. (1999). Probing Shakespeare's idiolect in Troilus and Cressida. *University of Toronto Quarterly 68*(3), *728-767.*

Luyckx, K., Daelemans, W. (2005). Shallow Text Analysis and Machine Learning for Authorship Attribution. *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN 2004), 149-160.*

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., Ye, L. (2005). Author identification on the large scale. *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification*, Washington University School of Medicine, St. Louis, Missouri.

Mikros, G.K. (2003). Statistical approaches to automatic text categorization in Modern Greek: A pilot study for evaluating stylistic markers and statistical methods. *Proceedings of the 6th International Conference on Greek Linguistics*, Rethimno, Greece.

Mikros, G.K. (2006). Authorship attribution in Modern Greek newswire corpora. *Proceedings of the SIGIR 2006 International Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval*, Seattle, Washington, USA, *43-47*.

Mikros, G.K. (2007). Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts. In: R. Köhler, P. Grzybek (Eds.), *Exact methods in the study of language and text: 445-456.* Berlin / New York: Mouton de Gruyter.

Mikros, G.K., Argiri, E.K. (2007). Investigating topic influence in authorship attribution. *Proceedings of the SIGIR 2007 International Workshop on*

Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, Amsterdam, Netherlands, *29-35*.

Miranda, G.A., Calle, M.J. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing 22*(1), *49-66*.

Mosteller, F., Wallace, D.L. (1984). *Applied bayesian and classical inference. The case of The Federalist Papers* (2nd ed.). New York: Springer-Verlag.

Oakes, M.P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Rose, T.G., Stevenson, M., Whitehead, M. (2002). The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, *827-833.*

Rudman, J. (1997). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities 31*(4), *351-365*.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2006). Effects of age and gender on blogging. *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.*

Stamatatos, E., Fakotakis, N., Kokkinakis, G. (1999). Automatic authorship attribution. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Bergen, Norway, *158-164.*

Tweedie, F.J., Singh, S., Holmes, D.I. (1996). Neural network applications in stylometry: the Federalist Papers. *Computers and the Humanities 30, 1-10.*

Tweedie, F.J., Baayen, H.R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32*(5), *323-352.*

Tweedie, F.J., Holmes, D.I., Corns, T.N. (1998). The provenance of De Doctrina Christiana, attributed to John Milton: A statistical investigation. *Literary and Linguistic Computing 13*(2), *77-87.*

Zhao, Y., Zobel, J. (2007). Search with style: authorship attribution in classic literature. *Proceedings of the 30th Australasian Computer Science Conference (ACSC2007)*, Ballarat, Australia.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 76-105*.

Automatic extraction of word-profiles from text corpora
On the example of Polish collective symbols[1]

Adam Pawłowski
Maciej Piasecki
Bartosz Broda

1. Overview of the project

The scope and methodology of language sciences has changed over centuries. Rationalistic approach was characteristic for the Age of Enlightenment, historical and comparative studies dominated the 19th century, the mainstream paradigm of the 20th century were structuralism and generativism. One of the principal factors of change in recent developments of linguistics is computer engineering, which led to the progress of empirical studies, automation and acceleration of traditionally manual linguistic tasks, such as data excerption (word frequencies, collocations, etc.), as well as surveys. However, while all sorts of data excerption are the most evident outcome of corpus linguistics, the automation of elicitation tasks realised in surveys seems to be of moderate interest. This is rather surprising, as surveys are a costly, but a powerful tool of linguistic inquiry, which offers a direct insight into the mental lexicon of individual language bearers. On a more general level, theses particular idiosyncratic characteristics sum up to universal descriptions of collective symbolism in the whole language community. Below three cases of such traditional applications of surveys are discussed, followed by an attempt of automatic extraction of word profiles from text corpora, simulating real elicitation tasks.

The first attempt to reconstruct the kernel of Polish collective consciousness was realised by Pisarek (2002). He introduced the notion of a *flag-word* defined as a word or expression which denotates or connotates high positive or negative values that can be put on flags or banners in a given society (*ibid*. 7). Operationally a flag-word can be defined as a unit *x* that is socially accepted in the expressions: *Long live x!* or *Down with x!* Pisarek managed to select empirically a set of 54 flag-words which determined the kernel of Polish axiological system at the end of the second millennium (cf. Appendix).

[1] This work was partially financed by the Polish Ministry of Education and Science, project No. 3 T11C 018 29.

Semantic profiles of one hundred selected notions, including choronyms, ethnonyms, toponyms and professionyms, have been reconstructed in a twofold (1990 and 2000) comparative study by Bartmiński and his collaborators from the Lublin school of anthropological linguistics (Bartmiński 2006). Bartmiński and his collaborators used in their research the method of surveys. Although no convincing argument was presented, saying why precisely these notions were chosen, the result seems relevant to the automatic approach in knowledge extraction. Below an example of the empirical word profile of the noun/adjective GERMAN (Niemiec), as it appears in the second study (year 2000), is presented. Terms in Table 1 represent full responses of respondents to open questions and have been introduced by the researchers.

Table 1
Word profile GERMAN, empirical profile (Bartmiński 2006: 388–389).

GERMAN (N(n)iemiec)	language 5.23% (16), hard work (Fleißigkeit) 5.23% (16), exactness 4.9% (15), history 3.59% (11), living in Germany 3.27% (10), beer 3.27% (10), culture 2.61% (8), order 2.29% (7), German citizenship 1.96% (6), patriotism 1.96% (6), no answer 1.63% (5), blond hair 1.63% (5), cleanness 1.63% (5), sentiment of superiority 1.63% (5), richness 1.31% (4), nationalism 1.31% (4), nationality 1.31% (4), pedantry 1.31% (4), punctuality 1.31% (4), racism 1.31% (4), cars 1.31% (4), solidity 1.31% (4), wars 1.31% (4)

Finally, the system of Polish collective symbols was investigated by Fleischer (2003). Fleischer carried out his study on the empirical basis in accordance with the assumptions of his systemic theory of culture (Fleischer 2002). Collective symbols are defined as signs which connotate meanings and values particularly important for a given culture or community. They may be of any nature or structure: simple or complex, verbal or pictorial etc. (cf. Appendix). Below an example of the empirical profile of the symbol *communism* is presented. Also in this case short terms in Table 2 characterising the collective symbol have been derived by the researcher from longer responses of respondents.

Table 2
Collective symbol *communism*, empirical profile (Fleischer 2003: 127)

COMMUNISM (komunizm)	ideology 52, political system [ustrój] 46, utopia 33, system 25, evil 19, USSR 18, totalitarism 16, Lenin 13, past 13

The data presented above bring two important kinds of information. One is qualitative: word profiles are a sort of extensive, empirical and quasi pragmatic definitions of these basic items of collective consciousness. The other one is quantitative: word profiles have numerical dimension, they show which features dominate in the profile. The latter characteristic is particularly important for quantitative studies of vocabulary and some techniques of natural language processing which focus on the relationships of word senses and lead to the creation of semantic maps or word spaces (cf. Lund, Burgess 1996; Schütze 1998; Lin 1998; Lin, Pantel 2002; Widdows 2004).

It should be emphasized here that word profile is not equivalent to the so called semantic differential (Snider, Osgood 1969). While Ch. Osgood's semantic differential is a method of measurement of meaning where the same network of semantic features characterises different terms or notions, word profiles are specific for every lexeme.

2. Hypothesis

In this experiment it was hypothesised that the word profiles of collective symbols selected in traditional surveys could be efficiently extracted from text corpora by means of statistical methods. Henceforth, such profiles are called *automatic word profiles*. Efficiency of extraction has two possible meanings here. First, automatic word profiles could be compared with the natural ones. The inference is in this case quite simple: the greater the number of identical or synonymous words in the corresponding profiles, the more efficient and adequate the method of extraction. But at the same time one should assume that the weak intersection of profiles or even the absence of common words need not be regarded as a negative result. It is quite imaginable that automatic word profiles, when compared with the natural ones, simply display a different, but still cognitively salient network of lexical or notional relationships.

3. Goals of the study

The study follows several goals which can be roughly classified as cognitive, pragmatic, methodological and, last but not least, financial. As far as cognitive aspects are concerned, the study is expected to shed new light on the relationship between human knowledge representation and word networks in corpora. Seen from the pragmatic perspective, this study may extend the limits of linguistic inquiry into the

structures of mental lexicon. Actually surveys require one absolutely crucial prerequisite, namely the presence of language users. But linguistics on its part deals extensively with external language manifestations coming from various historical periods, extinct languages, and from distant regions of the world. If an automatic method of corpus mining could produce the results compatible to those obtained in surveys, i.e. reliable word profiles, the scope of linguistic research concerning mental vocabulary would enormously enlarge – it would encompass all the cultures or societies, provided they had left to the posteriority sufficient amounts of exploitable written records. Methodological perspective seems very promising too. Survey and elicitation task data refer to the process of introspection in data collection. Automatic method would thus complete (certainly not replace) the introspective approach in the creation of word profiles. Finally the financial outcomes of the research should not be disregarded neither. Surveys have always been laborious and expensive. Even though creating a corpus is definitely more costly than making one survey, corpora are reusable and allow multiple studies, which makes them financially viable or even profit-making, long-term enterprise.

4. Method

When choosing the method of extraction one should bear in mind that while introspective approach offers direct access to the mental lexicon, understood as a complex network of notional relationships represented in human memory, automatic techniques can rely only on structural relationships of linguistic units in text. Assuming that one knows the context of use of a given piece of text, including both its intended purpose, as conceived by the author, and the real usages, as made by the readers, one could perform an in depth analysis of lexical units and their relationships on the basis of syntactic and semantic information derived from the text in which they co-occur. However, there is no automatic technique that has such broad and exhaustive access to the linguistic information present in text. Automatic analysis of syntax is either limited to the identification of selected main constituents or dependencies (*shallow parsing*) or has selective coverage and low accuracy (*deep parsing*). Automatic semantic analysis on the level of phrases and sentences does not go beyond relatively small subsets of the natural language, while lexical semantic analysis lacks exhaustive semantic lexicons and suffers from the phenomenon of polysemy. An automatic method of knowledge extraction must thus construct description of the lexical semantics on the basis of the fragmentary and distributed information. Actually, the existing methods utilise mostly lexico-syntactic relations, which only partially reveal semantic content of lexemes.

At first glance the survey method and the corpus based automatic method seem to differ significantly in the range of knowledge they have access to, which is smal-

ler in the latter case. However, the proportions are opposite, when quantitative aspects are considered. Great corpora and engineering tools of information processing offer large possibilities of knowledge extraction, which should be compared with the knowledge of human informants. There are two main paradigms of the acquisition of lexical-semantic information (Pantel, Pennacchiotti 2006): pattern-based and clustering-based.

Pattern-based approaches originate from the observation that there are some repeated lexico-syntactic structures which involve two lexical units and mark them as an instance of some lexical-semantic relation, e.g. the pattern from the seminal work of Hearst (1998):

NP_0 ..... such as {NP_1, NP_2 . . . . (and | or)} NP_n

It implicates that each noun phrase NP_i is a hyponym of the noun phrase NP_0, or, more precisely, the hyperonymy relation holds between lexical units represented in the text by the noun phrases. The implicit assumption here is that one can construct patterns which are enough accurate enough to make correct implications on the basis of single occurrences of pairs of lexical units. However, it seems barely possible in intentional metaphoric uses. Such uses are hard to distinguish from more regular ones without deeper semantic and pragmatic analysis.

The clustering-based approaches are free, to some extent, from the drawback mentioned above. They originate directly from the *distributional hypothesis* of Harris (1968), which says that lexical units occurring in similar contexts convey similar meanings. The stronger the distributional similarity is, the closer the lexical units are in their meaning. The name of the clustering-based paradigm originates from the observation that analysis of the similarity of distributions of lexical units always results in a form of grouping of highly semantically related lexical units. The meaning relation among lexical units is described by a *Measure of Semantic Relatedness* (henceforth MSR), sometimes called *semantic similarity* (cf. Edmonds, Hirst 2002). The former term (MSR) seems, however, more adequate, as it expresses several different types of semantic relations. MSR is a function such that:

$$MSR: LU \times LU \rightarrow R \tag{1}$$

where LU is the set of lexical units and R the set of real numbers, i.e. MSR assigns a real value to a pair of lexical units.

A typical blueprint for the MSR extraction, which is presented below, is common for most approaches:

- Construction of the co-incidence matrix.

- Transformation of the initial, raw frequencies.
- Weighting.
- Similarity computation.

A co-incidence matrix $\boldsymbol{M}$ is a matrix of the size: $|LU| \times |C|$, where $|LU|$ is the size of the set of lexical units, and C is the set of contexts. A matrix cell $\boldsymbol{M}[l,c]$ stores the number of occurrences of the lexical unit l in the context c. A context of the given lexical unit a can be defined in different ways on different levels of granularity[2] and details, so it can be:

- *document* – a whole document in which a occurs (cf. Landauer, Dumais 1997);
- *text window* – a consecutive sequence of words in text such that l is located in the centre (cf. Landauer, Dumais 1997);
- *co-occurrence in a text window* with a lexical unit b – a lexical unit b which co-occurs with a in the text window describes a (cf. Schütze 1998; Lund, Burgess 1996);
- *co-occurrence with the related lexical unit b* (*lexico-syntactic relations*) – b is related to a by some specific relation (mostly syntactic), there are two elements that constitute the description: b and the relation (cf. Ruge 1992, Lin 1998).

Using documents and text windows lexical units are described by assigning them to the topics of the contexts. However, documents should be semantically coherent and, moreover, text windows should be large enough. Landauer and Dumais (1997) achieved good results while applying their *Latent Semantic Analysis* method (referred to as LSA) to encyclopaedia entries as documents. However, in the case of documents taken from a general corpus results are quite bad, i.e. the extracted meaning associations of lexical units are vague and partially accidental[3].

Representation of the semantics of a lexical unit a on the basis of its collocations with the b units can be also considered as a kind of description by the co-occurrence in a text window. But in that case the window is limited to the nearest consecutive words (cf. Biemann *et. al.* 2004) and no other types of relations linking more distant units in a sentence can be explored. More precise information can be delivered by analysing co-occurrence in a text window of the lexical unit being described (a) on the basis of the structural associations with other lexical units used as elements of the description (b). In the previous experiments (Piasecki, Broda 2007), it was observed that the exploration of this information brings improvement

[2] Granularity is a measure of the size of components and the extent to which the whole is divided to subparts. Granularity also refers to the status of parts in relation to the overall structure of the whole.

[3] The re-implementation of LSA for Polish was presented in Piasecki and Broda (2007).

in the accuracy of description. Formal representation of co-occurrences displays information of two kinds:

- structural – mostly syntactic, based on relationships with other lexical units in natural language expressions (phrases, clauses, sentences, etc.);
- lexical – meanings of the lexical units *a* is related to.

For example, in the case of the following lexico-syntactic relation:

modified_by(*inteligentny*) (*intelligent*)

the composition of a lexical unit and relations describes the lexical unit *a* as being in the *modify syntactic relation* to the adjective *inteligentny* (*intelligent*) which posses some *lexical meaning*. In the case of modification by an adjective, the meaning of this relation is quite clear, but it is not always the case of the lexico-syntactic relations, e.g. modification by a noun in genitive may be very ambiguous, including, inter alia, possessiveness and/or meronymy.

This kind of description is more general than the one based on collocations, as it is not limited to the nearest, direct context, but takes into account deeper syntactic and semantic relations in a sentence. In Piasecki, Broda (2007) it was shown that the approach based on exploration of lexico-syntactic relations, when applied to Polish for the task of the extraction of MSR, is superior to the methods based on co-occurrence in a text window.

Identification of the relation instances in text requires an automatic tool capable of performing this task efficiently and with reasonably high accuracy. According to the first condition, it is necessary to process a corpus of the size of 250-700 millions of words in a reasonable time of one to a few day, e.g. in the case of 250 millions of words and one day of the processing time, a tool must process about 176 thousands of words per minute. Fortunately, a very high accuracy, which is measured as the percentage of correctly recognised relation instances, is not required, but it is desirable, as some errors, especially less systematic usages, can be hidden by the statistical mass of the collected data.

Recognition of lexico-syntactic relations is typically based on subsequent shallow parsing of the corpus (cf. Lin 1998; Weeds, Weir 2005). However, the lack of a robust, shallow parser for Polish required the use of a simpler tool for the corpus pre-processing. Polish is an inflective language with the relaxed word order but a lot of structural information is encoded in the morphosyntactic properties of word forms. Some form of the agreement of morphosyntactic categories is the basis for most syntactic structures. The morphosyntactic tagger called TaKIPI (Piasecki 2007) was applied to disambiguate Parts of Speech (PoSs), values of grammatical catego-

ries (like number, gender, case, etc.) and lemmas. The accuracy of TaKIPI is 93.44% when calculated in relation to PoSs and grammatical categories for all tokens in text. The accuracy of lemmatisation is about 98%. TaKIPI is a rule based tagger and includes an implementation of a language called JOSKIPI, which is a language of morphosyntactic constraints allowing for description of morphosyntactic dependencies e.g. the necessary agreement in the form of logical constraints. For example contexts, in which the above-mentioned modification of *a* by the adjective *b* occurs, can be described by a JOSKIPI constrain of the following scheme:

1.1 look for *b*, such that *b* agrees with *a* on number, gender, case, to the left from the position of *a* until the beginning of the sentence
1.2 if no b is found then return false
1.3 check tokens between *a* and *b* for the presence of only words that do not contradict the found agreement
1.4 if no contradiction was found then return true
1.5 perform a symmetric procedure for the right context of *a* up till the end of the sentence.

If the constraint is fulfilled for the given context of *a* – the true value was returned – then an instance of the corresponding lexico-morphosyntactic relation was identified and the corresponding matrix cell $\mathbf{M}[a, r_{i,b}]$ is increased.

The accuracy of the identification of relation instances on the basis of constraints is worse than the accuracy typically obtained with the use of shallow parsers. The constraints express only partial conditions, however the influence of the error of recognising the relation instances on the extracted MSR appeared to be small, while the accuracy of the MSR was at least comparable with the accuracy of MSRs extracted for English with the help of shallow parsers (cf. Piasecki *et. al.* 2007a). The constraints are not limited to the nearest context of *a*, and their application can result in recognition of the associations over the whole sentence. The advantage of the method is the utilisation of the morphosyntactic tagger, which is a relatively simple language tool available for many natural languages.

The following lexico-morphosyntactic relations were used in the experiments discussed in the empirical section (*a* represents a noun being described and *b* is the lexical element of the relation instance):

- *a* modified by *b*, where *b* is an adjective or adjectival verb participle;
- *a* modified by *b*, where *b* is a noun or gerund in genitive;
- *a* is co-ordinated with *b* (a noun or gerund) by one of the selected conjunctions, i.e. *ani* (*or, either, neither*), *albo* (*either … or*), *czy* (*or*), *i* (*and*), *lub* (*or*), and *oraz* (*and, as well as*);
- *a* is possible subject of *b* which is a verb.

Each occurrence of an instance of the relation *R*, written r_b=*R*(□,*b*) found in corpus (recognised on the basis of the appropriate constraint value) is recorded in the coincidence matrix:

$$\mathbf{M}[a, r_b] = \mathbf{M}[a, r_b] + 1 \quad (2)$$

As lexico-morphosyntactic relations, by their co-occurrences with the given lexical unit *a*, deliver some information concerning the meaning of *a*, we will further call such relations *features* of *a*.

The raw frequencies acquired from the corpus are biased and are very rarely used unmodified for the MSR extraction. A process of transformation, presented in Fig. 1, was proposed, which combines the variety of different techniques of filtering, selection and weighting. Many elements of the proposed process can be found in literature, e.g., filtering (Lin 1998; Geffet, Dagan 2004) or implicit local selection (Weeds, Weir 2005). However, such a complex picture of the subsequent stages of processing is rarely presented in literature.

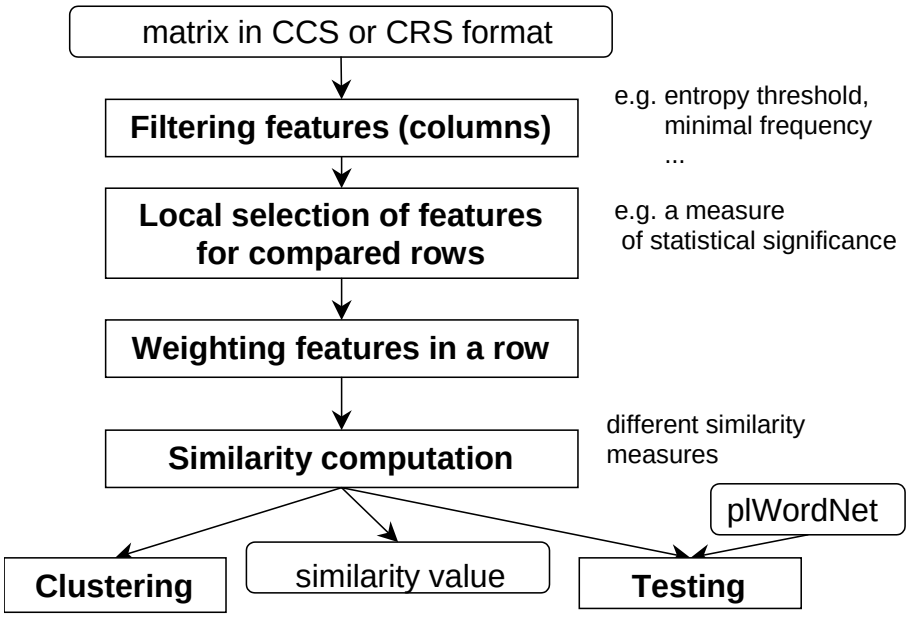


Figure 1. General scheme of the raw co-incidence matrix transformation

The general aim of the process of transformation is the selection of informative features and their modification which would provide a measure of the content they deliver. One should emphasise here the distinction between a *global selection*, i.e. *filtering* (see Fig. 1) done on the basis of the properties of the whole matrix, and a

local selection, i.e. selection of some features facilitating comparison of a pair of lexical units. The local selection is used in an implicit way in (Weeds, Weir 2005), while different techniques of filtering are often applied during the extraction of an MSR. Filtering results in the permanent eliminations of some features from further processing. In the approach described here, filtering is based on the combination of the *maximal entropy threshold* and two heuristic criteria: the *minimal noun frequency* and the *minimal number of informative features*.

By the application of the entropy threshold, one should eliminate features uniformly occurring with the large number of nouns, as well as those which do not deliver enough information discriminating among different nouns. Features were sorted in a descending way and 1% of the top features with the highest entropy were eliminated.

Lexical units occurring smaller number of times in the corpus than the minimal frequency threshold are filtered out, too. For such lexical units it is difficult to extract any reliable description of their meaning on the basis of the data acquired from corpora. The minimal frequency was experimentally set to the value 5, which is often used in literature (cf. Mohammad, Hirst 2006).

The threshold of the minimal number of informative features defines how many different features a given lexical unit must be associated with in order to achieve some minimal level of description. If the number of features describing a lexical unit is too small, it is not possible to
compare this lexical unit in a conclusive way with other units on the basis of its description. Only features passing the entropy threshold are counted and the threshold was set to 20 features during the experiments.

After filtering, only those nouns and those features are kept that seem to be described in enough detail and/or deliver information allowing for meaning relatedness description. However, even after the filtering phase, the feature values remain raw frequencies strongly biased by the applied corpus. In order to eliminate this bias, some feature weighting procedure is necessary. Many weighting functions have been proposed in literature, to name few examples: the Lin's measure of similarity (1998), the CRM scheme and several its instantiations in (Weeds, Weir 2005), or a recursive weighting function proposed in (Geffet, Dagan 2004). Nevertheless, when applied to the available corpus of Polish, any of the several tested weighting functions did not bring satisfying results, see (Broda, Piasecki 2007) and (Piasecki *et. al.* 2007a). The main reason seemed to be direct dependency of the feature values on the collected frequencies. It was assumed that a set of features ordered by the ranking of their importance was closer to the way in which a human describes an object than a representation consisting of a vector of real values, in which the subtle differences in value are outside the conscious understanding. On the basis of this assumption, a method of weighting features by ranking, called *Rank Weight Function*

(RWF) was proposed (Piasecki *et. al.* 2007a). RWF covers the two middle steps in the process presented in Fig. 1, namely: the local selection and weighting:

For each lexical unit *a*

2.1 Local selection:
  2.1.1 for each feature r_b: $\mathbf{M}[a, r_b] = f_{sig}(a, r_b, \mathbf{M}[a, r_b])$
    where $f_{sig}()$ is some pre-defined measure of significance of the feature r_b for the lexical unit *a*
  2.1.2 if *significant*($\mathbf{M}[a, r_b]$)=*false* then $\mathbf{M}[a, r_b]=0$
2.2 Weighting
  2.2.1 sorting of $\mathbf{M}[a, r_b]$ in descending order
  2.2.2 for each $\mathbf{M}[a, r_b]$ □ 0
    $\mathbf{M}[a, r_b] = No\_of\_significant(\mathbf{M}[a, r_b])+1 - rank(\mathbf{M}[a, r_b])$

As the measure of significance $f_{sig}()$ any function can be used that calculates a numerical value, expressing how significant is the occurrence of *a* in the context supporting the feature r_b. In the case of our experiments, the best results were achieved with the application of the *z-score* measure of statistical significance as $f_{sig}()$:

$$zscore(a,r_b) = \frac{M[a,r_b] - \frac{TF_a TF_{r_b}}{W}}{\sqrt{\frac{TF_a TF_{r_b}}{W}}} \quad (3)$$

where TF_a, $TF_{r\mathrm{b}}$ are the total frequencies of the noun *a* and the feature r_b, and W is the number of words in the processed corpora. In the case of too small values of TF_a, $TF_{r\mathrm{b}}$ the approximation used in (3) would not be reliable. However, we filtered out very infrequent lexical units (the threshold TF_a>10, discussed earlier) and lexical units with too few features, i.e. less than 5. Among the lexical units left in the matrix only a very small number of lexical units is of lower frequency (less than 30). Thus we assumed that in the vast majority of cases the equation (3) can be a reliable basis for the estimation of the strength of association between *a* and r_b.

In the step 0 one eliminates those features that after the transformation by $f_{sig}()$ appear to be insignificantly associated with *a* and *significant*(z) returns *true* if *z* is statistically significant with the threshold of 0.9999 probability of significance. Matrix cells set to zero in this step are not taken into account in the next step.

Features are then ordered by their significance (step 1.1.2). The significant ones receive as their new values the reversed ranks (positions) in the order, i.e. the

most significant features is assigned the number of all significant features for *a* and the least significant gets 1 as its value.

The number of the significant features can vary from a few to several thousands for more frequent and polysemous words. Thus each lexical unit is finally characterised by a vector of natural numbers, in which the values represent the relative importance of the corresponding features for the given lexical unit.

As the representation abstracts from the exact frequencies, there is no possibility to compare the vectors by probabilistic measures, and only heuristic and geometric measures can be applied. The best results were achieved with the application of the *cosine measure* for calculation of the similarity of row vectors. The cosine measure was used in all experiments presented here, and all the presented values of semantic relatedness are cosines of the two row vectors representing the lexical units being compared. All row vectors have been previously transformed by the RWF method, as earlier defined. Henceforth, the symbol MSR(RWF) is used for an MSR based on the RWF function, specified in the equation (1), which is based on the RWF function defined above.

5. Automatic identification of word profiles

MSR(RWF) produces higher values for lexical units being more closely semantically related, e.g. a program using MSR(RWF) achieved 90.92% of accuracy in a task of selecting a synonym of the question word from the four possible answers (Piasecki *et. al.* 2007b); the test was generated automatically from an electronic thesaurus of Polish called plWordNet (Derwojedowa *et. al.* 2008). It was assumed that on the basis of the values produced by the MSR(RWF) groups of lexical units, expressing meanings similar to the considered collective symbols, can be identified and labelled. Henceforth, collective symbols will be referred to as *focus words*, where the distinction will not be important in the context of automatic extraction of word profiles.

Due to its construction, MSR(RWF) reflects semantic relations of lexical units in the corpus. As the goal of the study was general, collective understanding of the analysed collective symbols, several Polish disjoint corpora were collected in order increase representativeness of the whole:

- the *IPI PAN Corpus* (IPIC), (www.korpus.pl), 250 millions of words, including: newspaper, literature, parliamentary documents and scientific works (Przepiórkowski 2004),
- the corpus of electronic edition of *Rzeczypospolita* (Polish newspaper), 113 millions of words, full editions from the years: January 1993 to March 2002,

- a corpus of *Large Text Documents* (LTD) in Polish collected from Internet, 214 millions of words, mostly literary works; documents already present in one of the other two corpora and all documents including larger number of words non-recognised by the Polish morphological analyser *Morfeusz* (Woliński 2006) were removed manually.

The entire joint corpus included 578 millions of words.

As focus words are rather general terms and no relevant associations with more specific lexical units were expected to appear, a kind of filtering was applied to eliminate less general lexical units as potential members of automatic word profiles. Consequently, only 13 285 one word and two word nominal lexical units were selected for the construction of the coincidence matrix. These lexical units come from:

- plWordNet,
- the Polish-English dictionary of (Piotrowski, Saloni 1999),
- two word lexical units from the general dictionary of Polish (SJP 2007),
- and the lexical units occurring more than 1000 times in IPIC.

The collected list was next compared with the *stop list* consisting of functional words and Proper Names. All lexical units included in the stop list were removed. On the basis of the joint corpus pre-processed by TaKIPI, a huge matrix of coincidence (13 285 lexical units × 271 563 features having non-zero total number of co-occurrences) was constructed, where selected lexical units were described by lexico-morphosyntactic features based on 61 064 adjectives and adjectival participles and 210 313 nouns and the four types of constraints presented in Section 0.

It was assumed that lexical units from a focus word profile are strongly related to the given focus word on the basis of meaning associations present in the corpus. Profile of each focus word w was simulated by extracting a list of $k = 20$ lexical units which express the highest semantic relatedness to w on the basis of the constructed MSR(RWF). Each automatically generated focus word profile was ordered in the descending order by the value of MSR(RWF). There is no absolute interpretation of MSR(RWF) values, as it depends, to some extent, on the number of significant features describing a lexical unit being analysed (here a focus word). So, it is hard to find any threshold for the MSR(RWF) values in general (a kind of confidence interval), and the criterion of taking the $k = 20$ most related ones turned out to be the most reasonable solution. An example of the automatic profile of the collective symbol *tolerancja* (*tolerance*), which was extracted from the joint corpus, is presented below in Table 3 – complete results are available on the web page of results (Pawłowski et. al. 2008):

Table 3
Automatic word profile constructed for the collective symbol *tolerancja* (*tolerance*)

MSR(RWF)	Profile members	Translations
0.258777	*wyrozumiałość*	*forbearance*
0.245173	*życzliwość*	*friendliness*
0.240451	*otwartość*	*openness*
0.215774	*pluralizm*	*pluralism*
0.20611	*wrażliwość*	*sensitivity*
0.196616	*poszanowanie*	*respect*
0.193327	*współczucie*	*sympathy*
0.192132	*zrozumienie*	*understanding*
0.191465	*uczciwość*	*honesty*
0.18945	*dobroć*	*goodness*
0.18698	*miłosierdzie*	*mercy*
0.183296	*cierpliwość*	*patience*
0.18323	*humanitaryzm*	*humanity*
0.179528	*równość*	*equality*
0.178427	*przyzwoitość*	*decency*
0.175568	*serdeczność*	*cordiality*
0.175561	*humanizm*	*humanism*
0.173376	*szacunek*	*respect*
0.170978	*bezinteresowność*	*disinterestedness*
0.170641	*braterstwo*	*brotherhood*

The extracted profiles were compared with the profiles of collective symbols created by human informants. The method of comparison consisted in calculating the size of the intersection. The results are presented and discussed in the Section 0.

A method of multidimensional scaling was then applied to identify sets of lexical units internally consistent in relation to the extracted MSR(RWF). Clustering helped identify, among a larger number of lexical units, a small and semantically coherent group of lexemes. However, as the number of units in the matrix is very large (13 285), clustering was limited to focus words and lexical units present in the profiles of collective symbols of Fleischer (2002). Examples of clusters of focus words are presented in Fig. 2 and 3. Automatic and survey-driven word profiles, in a form of lists, are presented in Tables 6 and 7.

6. Experiments and results

In the first series of experiments different MSR algorithms were tested in order to inspect the shape of automatically extracted word profiles. Among many possible MSR algorithms, three algorithms were selected on the basis of the previous applications to synonymy detection (cf. Piasecki et. al. 2007a, 2007b):

- MSR(RWF) – discussed in Section 0;
- MSR(LogEnt) – follows the transformation proposed in LSA (Landauer, Dumais 1997), i.e. logarithmic scaling of matrix cells and entropy normalisation of matrix rows[4];
- MSR(PMI) – a measure proposed in (Lin, Pantel 2002), discussed below.

In MSR(PMI) the matrix transformation is based on *pointwise mutual information* measure (Lin, Pantel 2002):

$$pmi_{a,r_b} = \log \frac{P(a,r_b)}{P(a)\,P(r_b)} \tag{4}$$

where, as in Section 0, a is a lexical unit being described, r_b is a feature, $P(a,r_b)$ is the probability of a and r_b co-occurring as associated, i.e. a occurring in the relation r_b, $P(a)$, $P(r_b)$ are unconditional probabilities of a and r_b occurrences, respectively. All three MSRs apply the cosine measure over transformed matrix row vectors to compute the final value of semantic relatedness.

Structure of the extracted automatic word profiles can be evaluated in two ways: by comparison to some existing pattern, and by manual inspection. For the needs of comparison lists resulting from Fleischer's research were used as a gold standard. Some minor modifications had to be introduced into the words profiles: all lexical units were change to their morphological base forms (as some lexical units were originally presented in plural) and some adjectives occurring in word profiles were removed, as automatic word profiles were constructed exclusively on the basis of nominal lexical units.

Because automatic profiles have constant length and in many cases it is greater than the length of the Fleisher's profile, we used two measures to describe the quality of automatic profiles: *precision* and *recall*. Let S_a denotes the set of lexical units

[4] In LogEnt transformation, every matrix cell is scaled by the natural logarithm and next divided by entropy of the corresponding matrix row.

from the word profile of the lexical unit *a*, which was obtained from surveys and A_a is a lists of lexical units from the automatic word profile of *a*. Precision of the results extracted for *a* is defined as following:

$$Precision_a = \frac{card(S_a \cap A_a)}{card(A_a)} \tag{5}$$

Precision is the percentage of lexical units from the automatic word profile of *a* that occur also in the word profile of *a*, build on the basis of a survey.

Recall is defined in a similar way:

$$Recall_a = \frac{card(S_a \cap A_a)}{card(S_a)} \tag{6}$$

Recall is equal to the percentage of lexical units from the manually acquired word profile of *a* that are covered also by the automatic word profile of *a*. It measures the extent to which it is possible to reconstruct the word profile of *a*, while the precision gives the error of the reconstruction obtained.

Results of average precision and recall calculated for all words and obtained by the application of different MSRs constructed on the basis of the joint corpus are presented in Table 4.

Table 4
Comparison of MSRs trained on the daily "Rzeczpospolita"

MSR	**Precision**	**Recall**
MSR(LogEnt)	4.26%	6.05%
MSR(PMI)	9.40%	13.41%
MSR(RWF)	18.40%	27.84%

The result obtained with MSR(RWF) is significantly better then the one obtained in the case of the other two methods[5]. Both indices (precision and recall) are based on the percentage of co-occurrence and have linear character. However it is very hard to set any threshold for them on the basis of deductive reasoning. Definitely, we did not

[5] It correlates with the results of comparative evaluation presented in (Piasecki *et. al.*, 2007a).

even come close to the results of thresholds, but as the automatic construction focuses more on the linguistic properties of lexical units and semantic relations expressed in the domain of "human" language use and knowledge representation, it is hard to predict what kind of agreement between both methods we should expect in the case of an "ideal automatic method".

Next the influence of different types of corpora on the result was also tested. The results are presented in Table 5.

Table 5
Results of MSR(RWF) applied to different corpora[6]

Corpus	**Precision**	**Recall**
"Rzeczpospolita"	18.40%	27.84%
LTD	20.56%	30.96%
IPIC	19.73%	29.43%
Joined corpora	20.79%	31.48%

The results obtained when exploring different corpora show that the use of a larger corpus (e.g. the joined corpora) gives results matching closer the word profiles acquired manually. Moreover, slightly better matching was obtained with LTD corpus which contains high number of literary works and richer, more descriptive language, when compared with the corpus of relatively short, electronic news (daily newspaper "Rzeczpospolita").

7. Clustering

Automatic word profiles show semantic relatedness of different lexical units to the given focus word. However, such a one dimensional presentation does exhaust opportunities created by the automatic methods. Network of semantic relations between lexical units can be visually revealed by clustering methods that extract groups of lexical units in such a way that units in one group are more semantically related to themselves than to the units from the outside of the group. Such clusters express structure of the semantic space.

[6] Corpora characteristics can be found in section 5.

In the next series of experiment a ready to use package called CLUTO, which implements several clustering methods, was used (Karypis 2002). As the input a similarity matrix computed on the basis of MSR(RWF) for selected lexical units was used. Only lexical units from the set of focus words and Fleisher's word profiles were included in the input. An extended list would result in clusters created according to less clear meaning dimensions, as clustering algorithms explore all dependencies in data. Among several clustering algorithms implemented in CLUTO, graph based and agglomerative methods were tested. According to the manual inspection, the best results were obtained with the agglomerative algorithm[7] of strict, hierarchical grouping, in which each lexical unit belongs to exactly one cluster and the clusters are organised into a binary tree. Each node of the tree joins two clusters that express the highest similarity among existing ones in the given phase of the tree construction.

Two selected fragments of the overall tree of clusters are presented in Figure 2 and Figure 3. The complete tree of clusters is available on the web page of results (Pawłowski et. al. 2008).

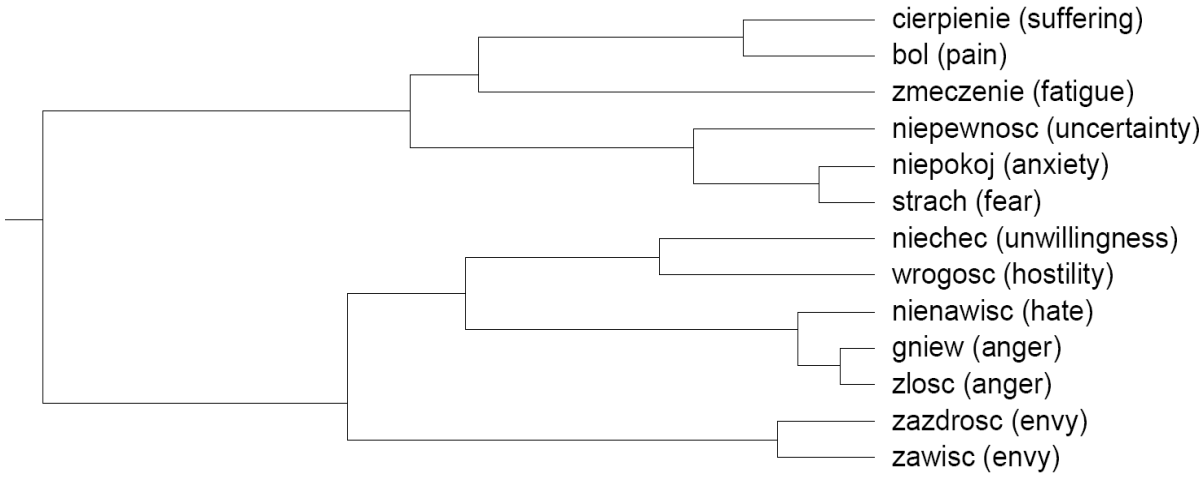


Figure 2. Fragment of clustering tree containing words with negative polarity

[7] Using upgma criterion function.

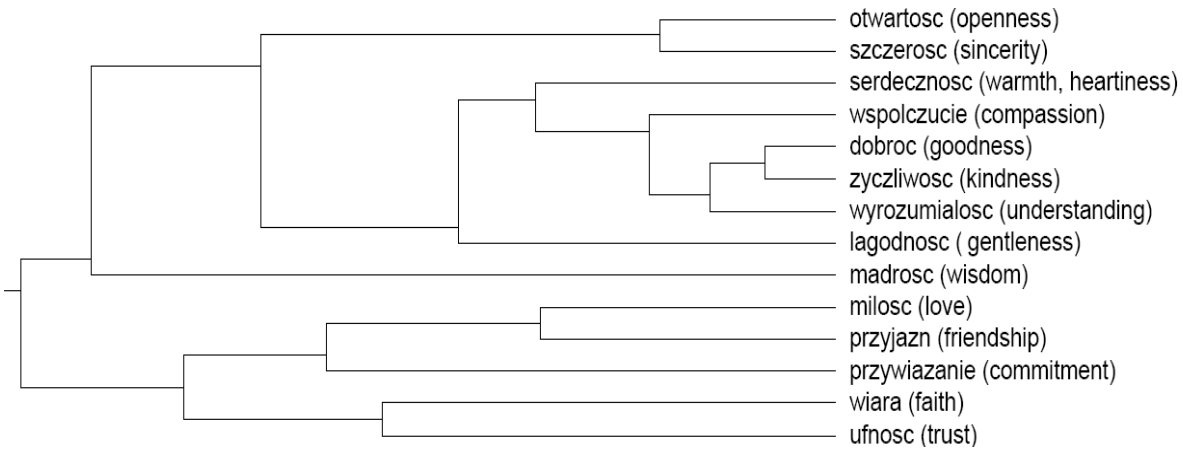


Figure 3. Fragment of clustering tree containing words with positive polarity

8. Examples of automatic word profiles

An example of resemblance of two word profiles, one extracted from the corpus and the other one acquired in a survey from human informants, is presented in Tab.6. Lexemes which appear in both profiles of the collective symbol *wolność* (freedom, liberty) are bold-faced.

Table 6

Comparison of two profiles for the collective symbol *wolność* (*freedom, liberty*)

Automatic word profile			**Profile acquired from human informants**		
MSR(RWF)	**Members**	**Translations**	***F***	**Members**	**Translati-ons**
0.326826	***swoboda***	***liberty***	65	***niezależność***	***indepen-dence***
0.253527	***suwerenność***	***sovereignty***	62	***swoboda***	***liberty***
0.251555	***równość***	***equality***	20	***demokracja***	***democracy***
0.222793	*pluralizm*	*pluralism*	16	*radość*	*joy*
0.207782	***niezależność***	***independence***	15	***równość***	***equality***
0.202406	*poszanowanie*	*respect*	14	***niepodległość***	***sovereignty***
0.192328	*prawo*	*law*	14	*pokój*	*peace*
0.190235	*prywatność*	*privacy*	12	*miłość*	*love*
0.18658	***demokracja***	***democracy***	12	***samodzielność***	***self-reliance***
0.172485	*autonomia*	*autonomy*	10	*szczęście*	*happiness*

Automatic word profile			**Profile acquired from human informants**		
MSR(RWF)	**Members**	**Translations**	***F***	**Members**	**Translati-ons**
0.16991	*tolerancja*	*tolerance*			
0.16405	*godność*	*dignity*			
0.161999	*bezpieczeństwo*	*safety*			
0.160893	***samodzielność***	***self-reliance***			
0.160312	*dobrobyt*	*prosperity*			
0.155195	***niepodległość***	***sovereignty***			
0.154079	*sprawiedliwość*	*justice*			
0.151656	***niezawisłość***	***independence***			
0.150943	*pomyślność*	*~well-being*			
0.148648	***równouprawnie-***	***equality of***			

An example of a less successful comparison of two profiles for the collective symbol *praca* (*work, labour, occupation*), probably caused by its different senses in Polish, is presented in Table 7.

Table 7
Comparison of two profiles for the collective symbol *praca*
(*work, labour, occupation*)

Automatic word profile			**Profile acquired by the survey**		
MSR(RWF)	**Members**	**Translations**	***F***	**Members**	**Translations**
0.284441	***robota***	***labour***	69	*pieniądze*	*money*
0.255903	*działalność*	*activity*	34	***zajęcie***	***occupation***
0.22462	***zajęcie***	***occupation***	31	*satysfakcja*	*satisfaction*
0.224472	*działanie*	*action*	30	*zarobek*	*earnings*
0.203023	***wysiłek***	***effort***	28	***wysiłek***	***effort***
0.202735	*zadanie*	*task*	25	*zawód*	*profession*
0.198546	***czynność***	***action***	21	*konieczność*	*necessity*
0.195763	*przedsięwzięcie*	*undertaking*	20	*obowiązek*	*duty*
0.1927	*szkolenie*	*training*	17	***robota***	***labour***
0.187332	*dyskusja*	*discussion*	15	*przyjemność*	*pleasure*
0.183773	*zabieg*	*procedure, endeavour*	15	*zadowolenie*	*satisfaction*
0.17763	*negocjacja*	*negotiation*	13	*zmęczenie*	*tiredness*

Automatic word profile			**Profile acquired by the survey**		
MSR(RWF)	**Members**	**Translations**	***F***	**Members**	**Translations**
0.175914	*debata*	*debate*	11	***czynność***	*action*
0.174264	*śledztwo*	*investigation*	11	*utrzymanie*	*~ livelihood*
0.166727	*współpraca*	*cooperation*	10	*dobro*	*good*
0.165385	*badanie*	*examination*			
0.165262	*przygotowanie*	*preparation*			
0.162105	*inwestycja*	*investment*			
0.159772	*dorobek*	*achievements*			
0.156046	*staż*	*apprentices-*			

9. Evaluation of the results

Semantics is by far the most complex part of language sciences. This is why quantitative measures of accuracy of an algorithm or of a method of a meaning-based word linkage are rather difficult to implement. They involve important human assistance (surveys, evaluation tests, often requiring some basic semantic knowledge) and should take into consideration pragmatic aspects (communicative context). One solution to this problem is a domain-sensitive approach, which consists in selecting more or less monothematic texts and thus eliminating some part of the polysemy and/or homonymy due to semantic incoherence of the "general language". This, however, was not the idea here – great corpora were applied in order to approximate possibly stable and constant semantic word connections. Instead, the solution applied was expert evaluation, which helped characterize at least some of the semantic relations discovered and give them a quantitative dimension.

However biased and "unequal" the results may seem, they are definitely noteworthy from the linguistic point of view. In almost all the cases, the program perfectly generated lists of synonyms (the highest values of the MSR measure). In this sense the distributional method implemented here is a powerful tool for lexicography, e.g. in the creation of dictionaries. The process of generation is fully automatic and requires only verification by a human. Further developments should lead to the creation of a graphic online interface visualizing "flat" word lists.

When looked up from a different perspective, the results might seem biased. It turned out that the distributional approach does not perfectly simulate human semantic associations (comparison with the list of Fleischer's collective symbols). Automatic extractions produce in the first place straightforward and close synonyms and

only then meronyms, which convey rich and interesting information[8]. In reality, human cognitive skills allow more complex associations, which can be classified as a broad and selective meronymy. Probably in the case of so called word priming experiments, especially with collective symbols, the axiological system of human values determines the selection of associated notions and respective lexical units.

It is worth mentioning here that the distributional approach produces quite different results, when compared with collocation measures. Collocations give word's characteristics, they describe the features most often associated with the given lexical unit (and notion). All the parts of speech may appear as attributes, but the preferred ones are, for nouns at least, adjectives. The algorithm based on the MSR produces the list of synonyms and meronyms which do not characterise a given lexical unit, but might replace it many contexts. Despite ignoring collocational information, it is more powerful, at least for Polish and other inflectional languages, because it takes take into account syntactic relations of disjoint lexical units in a sentence (a very frequent phenomenon in non-positional languages).

When analyzed quantitatively, the automatic method applied to the daily "Rzeczpospolita" corpus produced on the average only 15% of "bad" associations, i.e. contradictory to human linguistic intuition. This results would be even better, but a few automatic word profiles were entirely wrong. A good example of a misshapen profile is *bezrobocie* (*unemployment*). The results do not prove that the method is wrong, but that the collective symbol *unemployment* (negative or neutral value, common to many language users) has no precise and unique semantic profile. Actually, it has different meanings in various official discourses: economical, social, leftist, rightist etc. Probably a well targeted corpus would reveal one of the multiple facets of this complex notion. The results of the automatic and "natural" profiling are presented in the Table 8.

Table 8
Automatic word profile constructed for the flag-word *bezrobocie* (*unemployment*)

MSR(RWF)	Profile members	Translations
0.288122	*inflacja*	*inflation*
0.270549	*deficyt*	*deficit*
0.227205	*frekwencja*	*frequency*
0.227177	*popyt*	*Demand*
0.214060	*rentowność*	*rentability*

[8] Meronymy (holonymy) in a narrow sense is defined as a part-whole relation. Here meronymy is understood in a broad sense as a greater number of various relations, which include inter alia causality, content-container, product-producer relations, etc.

0.208806	*zapotrzebowanie*	*demand*
0.20666	*śmiertelność*	*mortality*
0.202789	*obrót*	*turn over*
0.197665	*zadłużenie*	*indebtedment, liabilities*
0.196904	*przestępczość*	*criminality*
0.172604	*spożycie*	*consumption*
0.166415	*nakład*	*investment*
0.161406	*wydajność*	*effectiveness*
0.160582	*dynamika*	*dynamics*
0.158016	*dysproporcja*	*disproportion*
0.156589	*podaż*	*offer*
0.155969	*zatrudnienie*	*employment*
0.155837	*liczba*	*number*
0.15239	*zapas*	*stock*
0.150035	*oprocentowanie*	*rate*

10. Conclusions

The study has proven that the access to the Mental Lexicon of Polish language bearers through text corpora is possible, but rather limited. The degree of similarity between natural and automatic word-profiles in the case of raw and unprocessed list of lexical units was relatively small. The intersection covered only 18.4% of the automatic profiles (*precision*) and 27.84% of the "natural" ones (*recall*). However, the result depends on the type of measure (MSR) applied. In the case of a simple method which ignores structural properties of text, i.e. MSR(LogEnt), the obtained result was much worse: 4.26% of precision and 6.05% of recall.

Application of more sensitive measures, such as MSR(RWF) and MSR(PMI), produced better results. In the case of the MSR(PMI), applied to the corpus of the daily "Rzeczpospolita", precision rose to 9.4% and recall to 13.41%. The best results were obtained with the MSR(RWF): respectively 18.4% and 27.8%.

Another way of improving the result of the study seemed fine-tuning of the corpora used. The initial, common sense assumption was that word profiles would be very content-sensitive. This would suggest that collective symbols and/or flagwords will be better characterised by the social and/or humanistic parts of the corpus, while scientific or administrative texts should produce worse results. Testing this hypothesis was, however, hardly possible, as discourses and genres are almost inseparably blended in the corpora dominated by press texts. Sorting them out of the masses of texts without automatic selection procedures resembles Cinderella's work,

which consisted in separating good and bad lentils – without the help of friendly pigeons. Further tests proved that the best data source are literary texts (LTD corpus, 214 millions of words). Surprisingly enough, it turned out that even better result was produced by the joined corpus (578 millions words). As the difference was not great in this case, a sound compromise between size and quality of corpora used in the automatic extraction of collective symbols and/or flag-words is contemporary literature, where axiological content is present and principles of good writing are respected.

An attempt was also made to point out advantages and disadvantages of the automatic text-mining method, when compared with the traditional techniques of data collection, such as surveys, elicitation tasks and/or word priming. Among the advantages one should emphasize:

- relatively easy targeting of the inquired population (despite the above-mentioned difficulties, thematically, socially or otherwise focussed corpora can be created);
- low price due to reusability of corpora;
- unlimited variety of exploration tasks (any set of lexemes can be profiled);
- limited access to the elements of the Mental Lexicon of the "absent" language bearers, provided a decent stock of texts be available.

Some disadvantages of the method of automatic word extraction also merit attention here. One should not forget that units in a corpus-driven word profile, based on the distributional method, are related by synonymy or antonymy rather then meronymy and/or free associations, which appear less frequently or on distant places of the list. And these are the latter, which are the most revealing and informative in discovering hidden structures of Mental Lexicon. In this respect "natural" data elicitation is superior to the automatic data extraction. From the perspective of a human informant, an unexpected feature of the automatic extraction of word profiles is the presence of antonyms. This is however quite justified in an automatic system considering lexico-morphosyntactic relations, because both synonyms and antonyms display very similar distributional patterns. As far as financial arguments are concerned, the price of context-sensitive Internet surveys is decreasing and this data source might be recalled as an alternative data source in research. Without denying the utility of this technique in general, one should bear in mind that the results of such surveys are heavily biased by the autoselection of respondents who enter a given WWW page, thus violating good practices in empirical sciences. At this stage of the research one should finally admit that there are still shortages of the lemmatizer of Polish, which produces a small percentage of wrongly attributed basic forms.

To conclude, the distributional method, based on the Measure of Semantic Relatedness (MSR) and data from great text corpora, should be regarded as a highly autonomous and efficient explorative tool, which helps reveal significant relations-

hips and regularities in the Mental Lexicon of members of a communicative community. At the present stage of research, it can complete and even partly replace introspective methods, such as surveys, elicitation tasks and word priming experiments.

References

Bartmiński J. (ed.) (2006). *Język, wartości, polityka* [Language, values and politics]. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Biemann, C., Bordag, S., Quasthoff, U. (2004). Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences. *Proceedings of LREC2004*. Lisboa, Portugal, ELRA.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., Broda, B. (2008), Words, Concepts and Relations in the Construction of Polish WordNet. In: A. Tanács et al. (ed.), *Proceedings of the Global WordNet Conference*. Seged, Hungary January 22–25 2008, University of Szeged, *162-177*.

Drabik, L., Sobol, E. (eds.). *Słownik języka polskiego* [Dictionary of Polish] (referred to as SJP 2007). Warszawa: Polskie Wydawnictwo Naukowe.

Fleischer, M. (2003). *Polska symbolika kolektywna* [Polish collective symbols]. Wrocław: Dolnośląska Szkoła Wyższa Edukacji TWP.

Edmonds, P., Hirst, G. (2002). Near-Synonymy and Lexical Choice. *Computational Linguistics 28(2), 105-144.*

Geffet, M., Dagan, I. (2004). Vector Quality and Distributional Similarity. Proceedings of the 20th international conference on Computational Linguistics, COLING2004, *247-254*.

Harris, Z.S. (1968). *Mathematical Structures of Language*. New York: Interscience Publishers.

Hearst, M. A. (1998). *Automated Discovery of WordNet Relations*. In: C. Fellbaum (ed.), *WordNet – An Electronic Lexical Database*. Massachusetts: The MIT Press.

Mohammad, S., Hirst, G. (2006). Distributional Measures of Concept-distance: A Task-oriented Evaluation Proceedings. *Conference on Empirical Methods in Natural Language Processing*, (EMNLP 2006), Sydney, Australia.

Karypis, G. (2002).*CLUTO – a clustering toolkit*. Technical Report 02-017, University of Minnesota, Department of Computer Science, August.

Landauer, T., Dumais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition. *Psychological Review 104(2), 211-240.*

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. *COLING 1998, ACL, 768-774.*

Lin, D., Pantel, P. (2002), Concept discovery from text. *Proceedings of COLING-02, 2002, 577-583.*

Lund, K., Burgess, C. (1996). Producing High-dimensional Semantic Spaces from Lexical Co-occurence. *Behavior Research Methods, Instrumentation, and Computers 28, 203-208.*

Pantel, P., Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, The Association for Computer Linguistics, 113-120.*

Pawłowski, A., Piasecki, M., Broda, B. (2008). *Automatic extraction of word-profiles from text corpora – results of experiments.* The web page containing the reserach results: http://plwordnet.pwr.wroc.pl/g419/WordProfiles/.

Piasecki, M. (2007), Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *TASK Quarterly 11, 151-167*.

Piasecki, M., Broda, B. (2007). Semantic Similarity Measure of Polish Nouns Based on Linguistic Features. In: W. Abramowicz (ed.), *Business Information Systems, 10th International Conference, BIS 2007, Poznan, Poland, April 25-27, 2007, Proceedings.* Springer, LNCS 4439.

Piasecki, M., Szpakowicz, S., Broda, B. (2007a). Automatic Selection of Heterogeneous Syntactic Features in Semantic Similarity of Polish Nouns. *Proceedings of the Text, Speech and Dialog 2007 Conference,* Springer, LNAI 4629.

Piasecki, M., Szpakowicz, S., Broda, B. (2007b). Extended Similarity Test for the Evaluation of Semantic Similarity Functions. In: Z. Vetulani (ed.), *Proceedings of the 3rd Language and Technology Conference, October 5-7, 2007, 104-108.* Poznań, Poland: Wydawnictwo Poznańskie.

Piotrowski, T., Saloni, Z. (1999). *Kieszonkowy słownik angielsko-polski i polsko-angielski* [Pocket English-Polish and Polish-English dictionary]. Warszawa: Wilga.

Pisarek, W. (2002), *Polskie słowa sztandarowe i ich publiczność* [Polish flag-words and their audience]. Kraków: Universitas.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version.* Warszawa: Institute of Computer Science PAS.

Ruge, G. (1992), Experiments on Linguistically-based Term Associations. *Information Processing and Management 28, 317-332.*

Schütze, H. (1998). Automatic Word Sense Discrimination Computational Linguistics. *Computational Linguistics 24, 97-123.*

Snider, J.G., Osgood, Ch.E. (eds). *Semantic Differentia Technique*. Chicago: Aldine Publishing Company.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics 24, 97-123.*

Weeds, J., Weir, D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics 31, 439-475*.

Widdows, D. (2004). *Geometry and Meaning*. Stanford: CSLI Publications.

Woliński, M. (2006). Morfeusz – a practical tool for the morphological analysis of Polish. In: M. A. Kłopotek et al. (ed.*), Intelligent Information Processing and Web Mining – Proceedings of the International IIS: IIPWM'06 Conference held in Wisła, Poland, June, 2006 Springer, 511-520.*

Appendix

Flag Words (Pisarek 2002)

	FLAG WORD	**(+)**	**(–)**
1.	MIŁOŚĆ (love)	69,1%	0,2%
2.	RODZINA (family)	55,8%	0,3%
3.	ZGODA (agreement)	50,4%	0,5%
4.	WOLNOŚĆ (freedom, liberty)	49%	0,2%
5.	SPRAWIEDLIWOŚĆ (justice)	46,7%	0,5%
6.	TOLERANCJA (tolerance)	43,4%	0,4%
7.	ZDROWIE (health)	43,1%	0,2%
8.	PRACA (work)	41,3	0,5%
9.	UCZCIWOŚĆ (honesty)	40,8%	0,2%
10.	WIARA (faith)	40,5%	1,2%
11.	OJCZYZNA (homeland)	36,7%	0,1%
12.	NAUKA (science)	36,2%	0,5%
13.	OPIEKA (care)	35,7%	1,0%
14.	PRAWDA (truth)	33,4%	0,2%
15.	BEZPIECZEŃSTWO (security)	31,6%	0,3%
16.	RÓWNOŚĆ (equality)	30,8%	1,0%
17.	DOBRO INNYCH (others' good)	21,8%	0,9%
18.	GODNOŚĆ (honour)	19,6%	0,4%
19.	PIĘKNO (beauty)	18,9%	0,4%
20.	NARÓD (nation)	17,8%	0,4%
21.	WARTOŚCI CHRZEŚCIJAŃSKIE (Christian values)	10,8%	2,5%
22.	SUKCES (success)	10,7%	0,6%

23. PATRIOTYZM (patriotism)	10,5%	1,2%
24. REFORMY PAŃSTWA (state reforms)	9,7%	11,7%
25. SOLIDARNOŚĆ (solidarity)	9,1%	5,4%
26. TRADYCJA (tradition)	7,9%	0,3%
27. PAŃSTWO (state)	7,6%	0,7%
28. DOBRO WŁASNE (own good)	6,9%	1,9%
29. NOWOCZESNOŚĆ (modernity)	6,1%	1,4%
30. LUKSUS (luxury)	5,3%	5,5%
31. PRZEDSIĘBIORCZOŚĆ (enterprise)	4,5%	0,6%
32. REKLAMA (publicity)	4,4%	15,1%
33. EUROPA (Europe)	2,4%	0,9%
34. EROTYKA (eroticism)	2,1%	18,7%
35. SAMORZĄD (self-government)	2,0%	2,1%
36. LEWICA (left)	1,9%	13,7%
37. SOCJALIZM (socialism)	1,9%	18,6%
38. WALKA (struggle)	1,8%	23,5%
39. PRYWATYZACJA (privatisation)	1,6%	11,3%
40. KAPITALIZM (capitalism)	1,4%	10,8%
41. ABORCJA (abortion)	1,2%	43,5%
42. PRAWICA (right)	1,2%	3,8%
43. CENZURA (censorship)	0,9%	34,0%
44. LUSTRACJA (lustration)	0,8%	20,2%
45. OBCY KAPITAŁ (foreign capital)	0,7%	19,4%
46. KLERYKALIZM (clericalism)	0,6%	21,2%
47. ANARCHIA (anarchy)	0,5%	46,7%
48. ELITA (elite)	0,5%	18,9%
49. ZAZDROŚĆ (envy)	0,4%	55,3%
50. GLOBALIZM (globalism)	0,3%	6,2%
51. KORUPCJA (corruption)	0,3%	66,6%
52. BRZYDOTA (ugliness)	0,1%	36,5%
53. ZAKŁAMANIE (lie)	0,1%	68,6%
54. DYKTATURA (dictatorship)	0,0%	55,9%

Collective Symbols (Fleischer 2002)

	Collective Symbol	**R**
1.	WOLNOŚĆ (freedom, liberty)	92
2.	MIŁOŚĆ (love)	90
3.	POKÓJ (peace)	90

4. RODZINA (family) 89
5. PRZYJAŹŃ (friendship) 88
6. SPRAWIEDLIWOŚĆ (justice) 88
7. DOBRO (good, wealth) 86
8. UCZCIWOŚĆ (honesty) 86
9. DOBROĆ (being good) 85
10. DOM (house, home) 85
11. TOLERANCJA (tolerance) 85
12. PRAWDA (truth) 84
13. SAMODZIELNOŚĆ (independence) 80
14. OCHRONA ŚRODOWISKA (environmental protection) 78
15. KULTURA (culture) 75
16. PRACA (work) 75
17. WIERNOŚĆ (faithfulness) 75
18. HONOR (honour) 74
19. GODNOŚĆ (dignity) 73
20. NIEZALEŻNOŚĆ (independence) 69
21. OJCZYZNA (motherland) 66
22. WIARA (faith) 64
23. PATRIOTYZM (patriotism) 57
24. TRADYCJA (tradition) 56
25. DEMOKRATYCZNY (democratic).....53
26. DEMOKRACJA (democracy).....52
27. NARÓD (nation) 51
28. PAŃSTWO (state) 49
29. PLURALIZM (pluralism) 37
30. KOŚCIÓŁ (church) 22
31. IDEOLOGIA (ideology).....13
32. PRAWICA (political right) 13
33. LEWICA (political left).....-5
34. NOMENKLATURA (nomenclature) .....-30
35. NACJONALIZM (nationalism) .....-35
36. LENISTWO (laziness) .....-38
37. KOMUNA (communism).....-46
38. KOMUNIŚCI (communists).....-48
39. KOMUNIZM (communism).....-49
40. EGOIZM (egoism).....-51
41. BEZROBOCIE (unemployment).....-55
42. KŁAMSTWO (lie).....-57
43. TOTALITARYZM (totalitarianism) .....-58

R. Köhler (ed.) (2009), *Issues in quantitative linguistics: 106-125.*

Measuring Morphological Productivity

Olga Pustylnikov
Karina Schneider-Wiejowski

1 Introduction

Predominant approaches in the area of language simulations (e.g., Steels, 2005; Vogt, 2003) deal with random vocabularies, while focusing on the emergence of meaning-form relations or compositionality (Kirby, 2007). Other use game theoretic assumptions modeling human communication strategies (Jäger, 2008), but still with random language input. This is done in order to simulate the emergence of language controlling different parameters. However, we might be interested in evolution of language that already has achieved a particular state in development. Pustylnikov (2009) presented a simulation model that analyzes natural language input (e.g., German) rather than random words, and takes this language as the base for communication between the agents. The simulation model was initially designed to examine the evolution of suffixes in word formation. That is, the use of a particular suffix when it comes to create a new word is assumed to depend on the language use in general (Tomasello, 2005), and on the lexicon a person has acquired in particular (Bybee, 1988: Ch. 7, 119–141.). In this paper, we use the language decomposition algorithm of this model to study productivity of suffixes.

Suffixes that have the same function (e.g., to derive an adjective from noun) are supposed to compete during the evolution of language. For example, a suffix that is preferred to derive an adjective from verb is likely to be reused in future word formations (e.g., *ease* > *eas-y*). These affixes are called productive affixes (Baayen, 1991).

A lot of work in respect to morphological productivity was done synchronically, so that productivity of several affixes was compared within a single period of time (see e.g., Prell, (1991), Habermann (1994) and Stricker (2000)). There was some effort to combine synchronic and diachronic aspects of productivity (Munske, 2002). Bauer (2001) emphasizes:

> "A second problem for word-formation with the distinction between synchrony and diachrony is that it is frequently the case that a diachronic event is the evidence for a synchronic state" (2001:27).

Thus, it is reasonable to consider both perspectives when we understand productivity as a gradual process, and aim to measure its development. Cowie et al. (2002:418) also emphasize the dynamics in change of word formation as the main aspect, and speak of *diachronic productivity*.

In this article, we present a combined approach to exploring morphological productivity in German. We examine productivity *diachronically* within the register of newspaper texts comparing corpora of 17th-19th century German to a 20th newspaper corpus. Further, we consider a spoken corpus of German in order to test the divergence in productivity values for spoken and written data. Since new word formations are more expected to happen spontaneously in speech, rather than in written language we evaluate this difference here. The evaluation for synchronic and diachronic productivity is made using four quantitative measures: two introduced by Baayen (1991, 1992), one proposed by Kreyer (2009), and one based on the simulation model discussed here.

The paper is organized as follows: Section 2 discusses the literature on morphological productivity. Section 3 describes the multi-agent simulation model allowing to study the productivity of a particular language. The corpora used are presented in Section 4. In Section 5, we describe quantitative measures applied here to measure productivity. The obtained results are discussed in Section 6. Finally, the conclusions are drawn in Section 7

2 Related Work

The phenomenon of morphological productivity has long been discussed theoretically in linguistic literature (see e.g., Schultink (1961), Plag (1999). And it has long been discussed as an insolvable mystery of derivational morphology (see Aronoff (1976:36)). But, the idea of morphological productivity is older than Aronoffs work, and the first remark on an aspect that is important to describe this phenomenon is made by Willmanns (1899) who describes derivation types by *vitality* and *persistency*. The same idea is expressed by Kruisinga (1932:22) who speaks of *living suffixes* and so called *dead suffixes*. An important step and often quoted statement comes from Schultink (1961):

> "We see productivity as a morphological phenomenon as the possibility for language users to coin unintentionally an in principle unlimited number of new formations, by using the morphological procedure that lies behind the form-meaning correspondence of some known words." (In: Evert & Luedeling (2001:167))

Three important aspects are mentioned in this statement: unintentionality, unlimitedness and regularity. All criteria given by Schultink (1961) can be seen as interdependent criteria. Unintentionality is in opposition to creativity. Words constructed by creativity are easily recognized as new words, this cannot be said at all for words produced unintentionally. The aspect of unlimitedness is a general property of natural language, and the productivity of a language is supported by the use of some regular affixes. In contemporary German, for example, it is possible to create many new adjectives with the suffix *-mäßig*: bananen-mäßig (banana like), kaffee-mäßig (coffee like), and in principle almost all can be *-mäßig*. Thus, the aspect of regularity is a very important one for derivation and word formation.

Aronoff (1976:36) makes an interesting statement in respect to morphological productivity. He points out that it is necessary to develop a method that allows estimating the number of all *possible words* formed by a word formation rule (WFR), and not just the ones given in sampled text:

> "There is a simple way to take such restrictions into account: we count up the number of words which we feel could occur as the output of a given WFR (which we can do by counting the number of possible bases for that rule), count up the number of actually occurring words formed by that rule, take a ratio of the two, and compare this with the same ratio for another WFR. In fact, by this method we could arrive at a simple index of productivity for every WFR: the ratio of possible to actual words."

This is the point where quantitative approaches come into play. Mostly all ideas and definitions made by the mentioned authors above are aspects that can be summarized as qualitative aspects of morphological productivity except for the idea of Aronoff that has not been formalized yet.

For Baayen (1992), there are a couple of aspects that can activate but also weaken a word formation process. The aim of Baayens research is to calculate the **probability** of finding new words in a sampled text that are formed by morphological process. Therefore, Baayen (1992) develops methods to measure morphological productivity in a quantitative way. Some of these measurements will be discussed in Section 5, and applied in our study.

Some other new findings of morphological productivity in a qualitative way are made by Plag (1999), Hay (2001) and Bauer (2001). All this work is based on the English language.

For German there is still less work done for productivity in derivational morphology. There are some single studies that try to detect consolidated findings. Scherer (2005) makes a corpus study for 400 years (1600-2000), and tries to find out whether the noun building suffix *-er*, which should be productive in contemporary German, has changed during time. She finds out that word formation processes are subject to diachronic change and that they can be measured in terms of productivity.

Another quantitative study is made by Luedeling & Evert. (2004) who investigate a special suffix from German that is normally used for medical description of words, but can also be used in other word formations: *-itis*. Another study comes from Schneider-Wiejowski (2009) who investigates four German suffixes (*-nis*, *-heit/-keit/-igkeit*[1], *-ung* and *-sal*) in a diachronic way in a German variety spoken in Switzerland. It can be shown that there is a morphological change during the 20th century in this variety. Some of these suffixes become more productive whereas other suffixes lose their efficiency of building new words.

The present study tests some of the productive suffixes from Schneider-Wiejowski (2009) in a German newspaper genre *diachronically* (17-20th century), and *synchronically* in a 20th century spoken language corpus.

3 Simulation Framework

3.1 Game Architecture

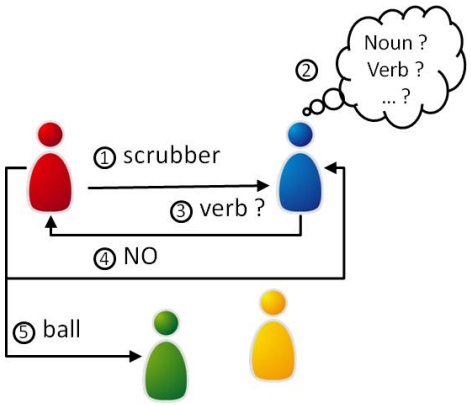


Figure 1. A possible game round in the first stage: An adult utters a word to the child (1), the child makes a guess (2-3), the adult values the answer of the child (4). If the answer was wrong, he picks the next agent and speaks to him (5), otherwise the child speaks to his neighbor. In the second stage, the answer of the adult is always YES, so that children speak more on their own.

For the purpose of the present work, we evaluate the decomposition module from the Morphological Derivation Game (MDG) introduced in Pustylnikov (2009). This model grounds on the theoretical assumptions of morphological processing Frauenfelder & Schreuder (1992), and was developed to test morphological processing mechanisms in a multi-agent simulation.

[1]The suffixes *-heit/-keit/-igkeit* are allomorphs, and therefore combined to the same suffix category.

The inter- and intra-generational model of MDG consists of two types of agents adults: *A* and children *C*. In a dialogue situation (i.e., a round of the game), there is one adult who "moderates" the course of the game selecting a *C* at random and talking to it. The adult brings new words into play asking the child to guess the part-of-speech (POS) of a particular word. It might resemble the following communication situation:

a) Did you see the *scrubber*?
b) *Scrubber*?

Obviously the child (b) does not know the meaning of the word *scrubber* yet, and has to guess it. Literature on first language acquisition states that children use all the material present in the speech of the adult (Bybee, 1988) to uncover the meaning. In MDG, the agents are endowed with a 'human-like' decomposition module that segments the words, similar to what children are supposed to do Dressler & Karpf (1995), and identifies suffixes characteristic for a part-of-speech. The decomposition module operates on the acquired words, and filters out the most probable derivation suffixes of the input language. Pustylnikov (2009) could show that the decomposition module succeeds to identify real derivation suffixes in English and German. In case of the random language[2], children negotiate new derivation suffixes that are different from those identified by the adult (since no significant suffixes are present in the input data). In this paper, we investigate the suffixes segmented from German input data for their productivity (see Sec. 5). That is, we ask: do the suffixes identified by the decomposition procedure correlate with empirical findings on productive suffixes for German? Does material from 17th-19th century German differ from the 20th century with respect to productive suffixes?

3.2 Decomposition Module

The decomposition module is based on two sources: *lexicon* and *word-family* file. Here, we concentrate on the suffix induction from lexicon, which is the main mechanism in MDG to induce suffixes from raw data.[3]

[2] By a random language, we mean here randomly generated strings from the Latin alphabet without any internal word structure.

[3] In the case of word families, derivational suffixes are induced decomposing the words of a word family (e.g., *Geograph* > *geography* > *geographical*) into stem and possible suffixes. See Pustylnikov (2009) for details.

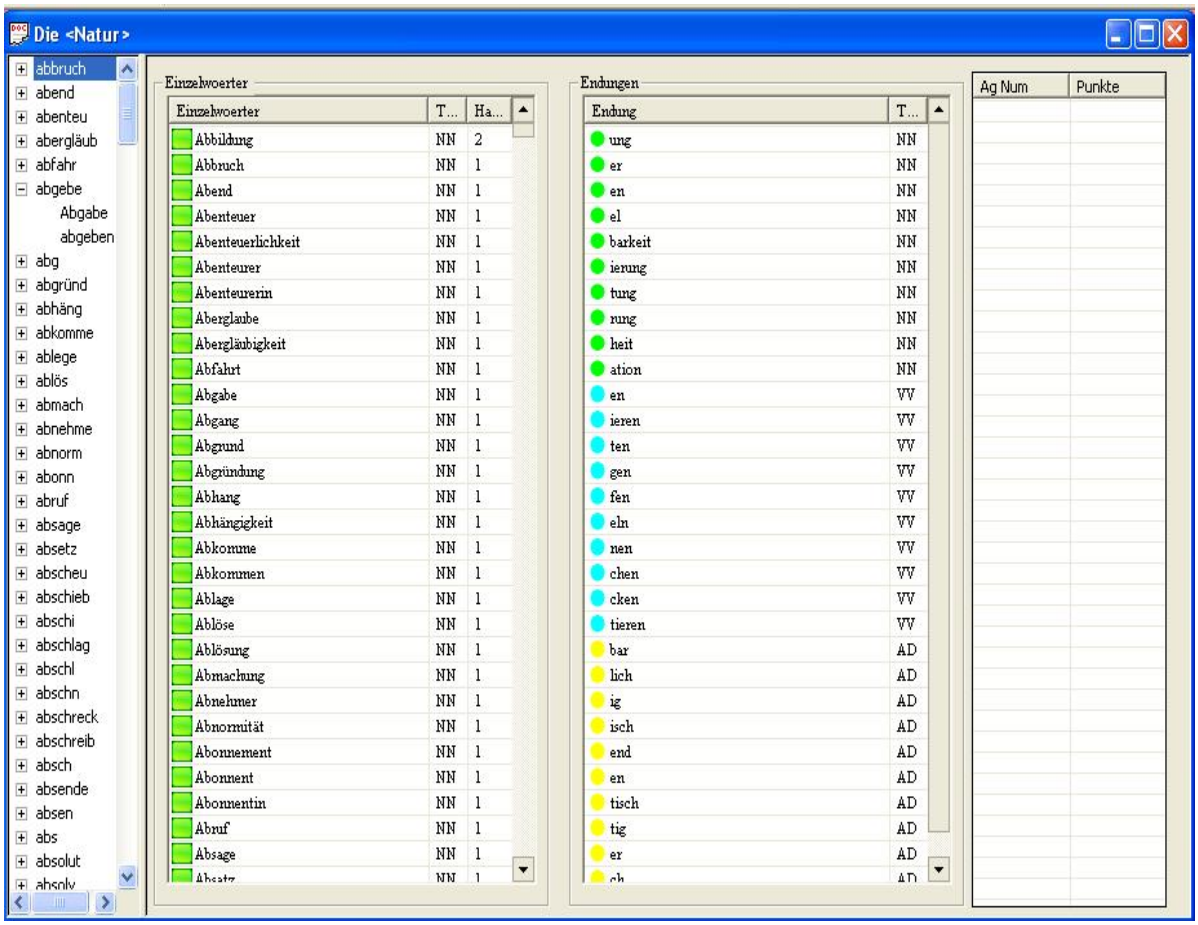


Figure 2. A snapshot of MDG after parsing a German newspaper corpus.

Figure 2 shows a screenshot of the MDG after parsing the German newspaper corpus *Süddeutsche Zeitung* (see Section 4). The first tree view displays the word families, the second the single words, the third shows suffixes of nouns, verbs and adjectives identified by decomposition, and ranked according to their relevance (estimated by the below algorithm). The last column contains the agent scores, which is empty at the beginning. Our hypothesis is that productive suffixes can be identified in language solely observing the structure of the words. The below algorithm, the same for adults and children, tries to find those suffixes that are best suited for word formation analyzing words from the lexicon. The adults' lexicon represents a corpus of texts (e.g., newspaper corpus) tagged with POS information. Adults process the whole lexicon, and have a greater chance to find the regularities of suffix use (Figure 2 shows the knowledge of the adult). Child agents acquire this

knowledge partially, analyzing the words uttered by the adult during the game. Both types of agents use the following algorithm:

1. For each word class *c* and each word *w* encountered by the agent all possible suffixes $\{s_1,...,s_n\} \in w$ are extracted and stored together with their frequencies F_i (for $F_i > 20\%$). [4]
2. If $F_1 \sim F_2$ according to a similarity threshold *sim* (.1 in this step) the shorter suffix is discarded. For example, if '*-isch*' and '*-ch*' have the same (or similar) frequency, then '*-ch*' is discarded.[5]
3. Step 2 is repeated comparing suffixes of *w* to all suffixes of all other words with a more sensitive value of *sim* (i.e., increasing the similarity threshold to .0000001 so that only suffixes with nearly the same frequencies are compared).
4. For each remaining suffix *s* a list of suffixes that have no substrings in common escape for *s* is constructed. That is, a list for the suffix *-ung* contains *{-tung, -lung, -kung, ...}*.

In sum, a suffix is said to be productive according to the above algorithm if it is frequent (frequency above the threshold), and also preceded by a large number of different substrings (i.e., the length of the suffix-list in Step 4).

4 Corpora

Table 1
Sizes of 3 German Corpora: 17th-19th, 20th century newspaper corpus *Süddeutsche Zeitung* (2004) and a spontaneous speech corpus (German) Hinrichs & Kuebler (2005).

Corpus	**GerManC**	**SZ**	**SpokenC**
Size	110,448	975,526	362,795

In the present study, we use three different types of data: two corpora from the newspaper genre, and one corpus of spontaneous speech. The *GerManC* project[6]

[4] 20% are determined heuristically in order to filter out inappropriate suffixes.
[5] This is done in order to avoid duplicates. Since all possible suffixes are analyzed separately, the suffix *-ch* needs not to be considered twice - as a separate suffix, and as a suffix within *-isch*.
[6] http://www.llc.manchester.ac.uk/research/projects/germanc/

represents a historical corpus of German from 1650 to 1800. The aim of this project is to compile a representative corpus of written German for these time steps. This period is a period of language change because, on the one hand, the modern standard was formed during it, and on the other hand, competing regional norms were finally eliminated. Currently only newspaper samples are available, but in future the project plans to add other genres too. The newspaper corpus consists of 2000 word samples from five regions (North German, West Central German, East Central German, West Upper German, East Upper German) within three periods of fifty years (1650-1700, 1701-1750 and 1751-1800). For each region three samples were taken for each period so that 110,448 words are available to work with. We test the total amount of data available for the period from 17th-19th century to compare with the newspaper corpus from the 20th century.

The 20th century newspaper corpus is extracted from the 10 years release of the German newspaper *Süddeutsche Zeitung* (2004). We consider a sample of the total corpus that comprises 975,526 tokens (henceforth abbreviated with SZ). The *Süddeutsche Zeitung* corpus was used in recent studies to evaluate text classification techniques (see, e.g., Mehler et al. (2007)).

Finally, we measure the degree of productivity in the *Tübinger* corpus of spontaneous speech Hinrichs & Kuebler (2005). This corpus comprises 362,795 tokens of spoken dialogue, and is used here as a baseline to evaluate the productivity measures. Productivity is assumed to be the probability of using an affix when it comes to generate a new word (Baayen 1991). Although, neologisms are expected to appear in spoken language, the quantitative measures proposed in the literature are widely applied to written data. In this paper, we aim to test them for a spoken corpus, too. The reason is that large divergence in results between spoken and written data could indicate a weakness of a particular productivity measure.

5 Productivity Measures

Some German suffixes like -*heit* (Schön-*heit* 'beauty', Gesund-*heit* 'health') or -*ung* (Digitalisier-*ung* 'digitalization', Computerisier-*ung* 'computerization') are very productive because it is possible to create many (new) words using them. But there are also suffixes that do not account for word formation at all. Although, there are existing lexicalized words like Ereig-*nis* or Hinder-*nis* composed with the suffix -*nis*, this suffix would not produce any new word because it is completely unproductive. Derivational productivity is gradual, and some attempts to measure productivity were made in the past. In this section, we test some of these approaches.

5.1 Productivity *P* in the narrow sense

One very popular measure is the one proposed by Baayen (1991):

(1) $$P = \frac{n_1}{N}.$$

This measure calculates the proportion of single tokens (i.e., hapax legomena) of words ending with, for example, *-ung* divided by the number of all tokens ending with *-ung* (*N*) in the corpus. By calculating the ratio of hapaxes, the types which only have one token, Baayen gets an estimate of the probability of becoming novel forms with a given affix. As noted by Baayen, and also remarked elsewhere (Bauer, 2001) this measure is dependent on the size of the corpus. This property does not allow for direct comparisons of the values obtained from corpora of different size. *P* is according to Baayen best suited as an index of productive vs. unproductive affixes.

Table 2 shows the values of *P* for the three corpora. Obviously, the values cannot be compared directly due to different corpus sizes. However, we can compare the productivity of the suffixes within a corpus ranking them according to the values of *P*, and then examine whether this ranking changes among the corpora (i.e. we can perform a rank test). In the 17th-19th century corpus *GerManC -nis* is most productive followed by *-heit/-keit*, *-ung* and *-er*, which have similar values.

Table 2
Noun suffixes in GerManC, SZ and SpokenC according to *P*. The values are obtained for the total number of tokens.

	GerManC			SZ			SpokenC		
suffixes	n_1	*N*	*P*	n_1	*N*	*P*	n_1	*N*	*P*
-nis	4	10	.4	55	1357	.04	55	1357	.04
-ung	156	864	.18	1016	20415	.049	1016	20415	.049
-er	129	803	.162	1061	22980	.046	1061	22980	.046
-heit/-keit	79	281	.281	234	2480	.094	234	2480	.094
token	110,448			975,526			362,795		

Table 3
Noun suffixes in GerManC and SZsmall measured with *P* for corpus samples of equal size.

	GerManC			SZsmall		
	n_1	*N*	*P*	n_1	*N*	*P*
-nis	4	10	.4	25	121	.02
-ung	156	864	.18	357	2086	.171
-er	129	803	.162	361	2607	.138
-heit/-keit	79	281	.281	83	262	.316
token	110,448			110,448		

According to *P*, *-nis* seems to be more productive but when we look at the number of tokens (10) ending with *-nis*, and the number of hapaxes (4) the high value of .4 is explained by a small difference between the two, rather than, by a high productivity of *-nis*. Considering the other two corpora, SZ and SpokenC, *-nis* appears to be less productive. The sample of SZ of the size of GerManC (Table 3) shows also smaller values for *P*.

We can interpret this result as a shift in productivity with respect to *-nis*. However, obviously the values of *P* are biased by a small number of tokens, and cannot be accepted as a true indicator of a productivity shift for *-nis* (although, such a shift might exist in language).

5.2 Hapax-conditioned Productivity *P**

Table 4
Noun suffixes in GerManC, SZ and SpokenC according to *P**.

	GerManC	*P**	SZ	*P**	SpokenC	*P**
hapaxes	1354		11352		730	
-er	129	.09	1061	.09	52	.07
-ung	156	.11	1016	.09	64	.08
-heit/-keit	52	.04	234	.02	9	.01
-nis	4	.003	55	.004	7	.009

Table 5
Noun suffixes in GerManC and SZ measured with *P** in two corpus samples of equal size

	GerManC	*P**	SZ	*P**
hapaxes	1354		3789	
-er	129	.09	361	.09
-ung	156	.11	357	.09
-heit/-keit	52	.04	83	.02
-nis	4	.003	25	.006

To overcome the drawbacks of *P*, Baayen (1992) proposes another measure *P** that allows to compare affixes - within and across corpora. This measure compares hapaxes of a particular morphological category (e.g., suffix *-ung*) to all hapaxes of a part of speech (POS) (e.g., nouns) in the corpus, and asks about the contribution of this affix to all singular word formations.

(2) $$P^* = \frac{n_1}{h_t}$$

In the above formula, n_1 denotes the number of hapaxes formed by a particular word formation rule for a part of speech *t*. The quantity h_t denotes the total number of hapaxes in the corpus. Bauer (2001) doubts whether this index really measures the right thing, asking: "What proportion of new coinages use affix X?" That is, *P** assumes that productive formations are within the total amount of hapax legomena in the corpus. Bauer (2001) contrasts *P** with *P* stating that *P* asks: "What proportion of words using affix X are new coinages?" However, the assumption that hapax legomena are really new formations is also implied by *P*, and can be doubted, especially when we deal with written corpora. That is, words might occur only once in a corpus indicating an unproductive and lexicalized formation process. We are not in the position to rule out the problem of appropriate corpus selection, but we consider a spoken corpus to additionally verify the expressiveness of the measures.

Finally, Baayen assigns different functions to both measures and recommends applying them as complementary. This is done in the present paper.

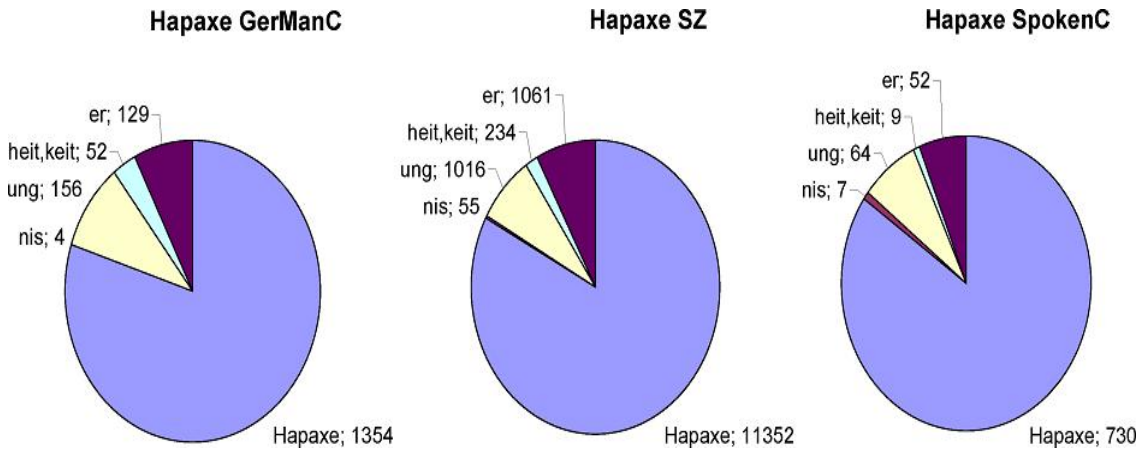


Figure 3. Noun hapaxes in GerManC, SZ and SpokenC.

The values of *P** (Tables 4-5, Figure 3) look much different from what is shown in Table 2 for *P*. According to *P**, *-ung* and *-er* have almost the same productivity, followed by *-heit/-keit* and *-nis*. The picture is consistent for different periods of time (GerManC and SZ), as well as for the spoken corpus.

5.3 Productivity based on Type Frequency

The first two measures presented in this section assume that the number of hapaxes in the corpus is an indicator of productivity. Kreyer (2009) argues that it is reasonable to investigate not only hapax legomena but the total distribution of types formed by a particular word formation rule. He proposes a measure to account for this task:

(3) $$P' = \frac{n_1}{\sqrt{1}} + \frac{n_2}{\sqrt{2}} + \ldots + \frac{n_x}{\sqrt{X}} = \sum_{i=1}^{X} \frac{n_i}{\sqrt{i}},$$

with n_i being the number of types that occur i times in the corpus. X is the total number of different frequencies of types i formed by the word formation rule. *P'* takes its maximal value if $n_1 = X$, that is, if there are only hapaxes in the corpus. Its minimal value is achieved for $n_X = 1$ if $X > 1$. That is, for example, if there is one type in the corpus ($n_X = 1$), and this type occurs 1000 times ($X = 1000$). Thus, *P'* takes the total frequency distribution of types into account giving less weight to highly-frequent types. A high value of *P'* indicates high productivity.

Table 6
Values of *P*' measured for GerManC, SZsmall, SZ and SpokenC.

suffixes	GerManC	SZsmall	SZ	SpokenC
-nis	.8309	.6959	.4939	.8189
-ung	.7660	.7397	.5725	.6446
-er	.7360	.7330	.5886	.6649
-heit/-keit	.7774	.8525	.7446	.7030

Results for our data are shown in Table 6. With respect to *-nis* the results are not plausible, which can be explained by a small number of different types in the GerManC and SpokenC. Figure 4 illustrates the frequency rank of *-nis* in different corpora on a log-log plot. There are only two different type frequencies (hapaxes and dis legomena) in the two corpora. While *P*' weights upper ranks higher, a small frequency spectrum of types results in higher values of *P*'. In contrast, the other two corpora (SZsmall and SZ) have a larger frequency spectrum of types, which lowers the overall result of *P'*. So, at least for *-nis* we cannot make any judgments about productivity according to *P'*.

5.4 Simulation Rank

Simulation Rank (SR) is an index that results from empirical ranking of suffixes identified by the decomposition module of MDG. According to the filtering algorithm presented in Section 3 the SR ranks the suffixes with respect to their probability to be selected for word production. Ten suffixes[7] are selected at most, and the three best ranked are those used for production. On the one hand, we evaluate the MDG's potential in identifying productive suffixes, on the other hand, we compare the ranking of SR to the results of productivity measures described above.

Table 7
Noun suffixes in GerManC, SZ and SpokenC according to *P*.

SR	GerManC	SZ	SpokenC
1	-er	-er	-er
2	-ung	-ung	-en
3	-en	-en	-ung

[7] "Ten" is a filtering parameter, which can be varied to include less relevant suffixes.

The decomposition module of MDG identifies only *-er* and *-ung* as appropriate for word production. Suffixes *-nis* and *-heit/-keit* are not present in the best-of-ranking. This finding is in line with the results obtained by *P**, and in line with the literature about these suffixes. Suffix *-en* is additionally ranked within the three best suffixes identified by the algorithm. The results are consistent for different time periods (17th-19th vs. 20th century) and genres (newspaper vs. spoken language).

6 Discussion

In this paper we have evaluated language decomposition techniques applied to measure morphological productivity. The simulation based productivity measurement represents a frequency based approach to productivity. This approach identifies productive suffixes in a language, and can be used to test productivity assumptions for particular suffixes.

Measuring the productivity of certain German suffixes with established methods for measuring productivity could show that at first glance it looks like *-nis* should be very productive from 1650 to 1800, whereas the other three noun forming suffixes *-er*, *-ung* and *-heit*/*-keit* seem to be less productive. The suffix *-nis* generates the highest value of *P* in the corpus of written speech.

For the SZ corpus one can say that all values except for the value of *P* for *-heit*/*-keit* are similar to each other. The allomorphs of *-heit* show a very high value of *P* (0.094). For the spoken corpus we observed according to *P* that *-ung* is the most productive suffix followed by *-nis*, whereby *-er* and *-heit*/*-keit* are not productive at all.

This finding does not confirm the assumptions that are made in the linguistic literature. Fleischer & Barz (1995) as well as Lohde (2006) say that all suffixes explored here should be very productive for forming new words in contemporary German except for *-nis*. The suffix-*nis* is declared as unproductive. Therefore, we have to question our findings and find explanations to state this. First, we can think about the value of *P* itself and what it should express. All words that occur only once in a given sample should be hapax legomena in the sense of Baayen. This means, hapax legomena are put on the same level with newly created words. But there is no guarantee for making this implication. There are many reasons why one word only occurs once in a sampled text:

- The word is new and was created by word formation process.
- The word is very old and rarely used.
- The word is not new, and it is still used in a language but it only occurs once by chance.

The other thing one has to consider is that *P* should be interpreted as a constrained value dependent on corpus size. And in our first experiment the total number of tokens is not the same for all corpora. The corpus from the 17th century is a very small one, and it consists only of approximately 100,000 tokens whereas the SZ corpus is nine times larger.

Therefore, we decided to select a sample of the SZ corpus of about 100,000 tokens and to look at *P* again. After shortening the available token size, all values change. The new calculation shows that *-heit/-keit* should be the most productive one of the four suffixes followed by *-ung* and *-er*. *-nis* is unproductive. This conclusion comes closer to our intuition about the productivity of *-nis*, however, *-ung* and *-er* are more productive than *-heit/-keit*, which is not confirmed by the results.

If we look at the results of the second experiment in which hapax-conditioned productivity was examined we can come to another conclusion. Table 4 show that *-nis* is less productive than the other three suffixes independent of time period or genre. Suffixes *-er* and *-ung* should be very productive, and *-heit* can still be interpreted as a productive suffix. And these values do not change after shortening the number of examined tokens from the SZ corpus (Tab. 5). Amazingly there is the same structure for all ratings and these values do not depend on the size of the corpus. That means, according to *P** there is no shift in productivity in the period of time examined here. On the one hand, this result shows the stability of *P** in contrast to *P*. On the other hand, it is probably not sufficient to consider hapaxes only, and we are better off to look at the full frequency distribution of words formed by an affix.

To account for this, we used the measure of Kreyer (2009) that combines all type frequencies in a single measure giving to less frequent types (i.e., hapaxes etc.) higher weights. The results (Tab. 6) for *-nis* are biased by the small number of types, and cannot be treated as representative. The overall results for *-ung* and *-er* seem to be plausible, since both suffixes have similar values in all the corpora, which is in line with our expectations. The suffix *-heit/-keit*, however, has better values, than *-ung* and *-er*, which is understandable, too. These allomorphs are declared as productive in German, and it is possible to create new words using them. In sum, it seems plausible to look at the type frequency spectrum as Kreyer does, however the significance of the results with respect to productivity needs further testing.

Other factors like phonological and semantic constraints for selecting a suffix might also influence productivity of an affix but a quantitative model capturing these kinds of information is still missing. Therefore, there is a research gap, and in future this question should be investigated more in depth, especially for German.

As the theoretical background it is possible to describe the phenomenon of productivity with the grammaticalization theory, which is discussed to a great extent

in the literature (see e.g., Diewald (1997), Heine (2003), Hopper & Traugott (2003), Lehmann (1995)). Grammaticalization is a cycling process. That is, existing lexical items are worn down, but at the same time new grammatical affixes are created. Although, there is a controversial discussion about the question whether the grammaticalization theory can be used to explain the emergence of affixes and not - in the narrow sense - just for changes from syntax to clitics (see Nuebling (2006:72)), we assume that it can be used for derivational morphology. If grammaticalization is a cycling process, it should be also explained what happens with affixes that become unproductive. We assume that the German languages (and other languages too) have a *balanced derivation system*. If one affix becomes unproductive, another one will get more productive. If one suffix dies, another one is born. There are a few examples of present day German for affixoids, emerging affixes that will maybe convert to affixes once. One of such an affixoid is *höllen-* (*hell-*) that is very productive now, especially in spoken language: there can be a *Höllenhitze* ('a hell heat') or one can have a *Höllentag* ('a hell day').

7 Conclusion

In this article, we analyzed morphological productivity of four German suffixes using four different approaches. We evaluated three quantitative measures proposed by (Baayen, 1991; Baayen, 1992; Kreyer, 2009), and one frequency based approach used for language decomposition. The simulation based ranking allows us to identify productive suffixes in language and rank them according to their productivity. The results of the ranking are consistent with the literature on productivity of these suffixes. Results of SR and P^* when applied to different kinds of corpora show almost the same degrees of productivity, either for different time spans, or for different corpus genres. This finding indicates that either the degree of productivity for these suffixes has not changed much in this time span, or this variation is not captured by the measures. To answer this question, additional experiments with speakers' intuition about the use of suffixes, as well as more quantitative tests are needed. The measure P shows a variation among the different kinds of corpora, however, this variation is not confirmed by findings from the literature. In sum, although, quantitative approaches like those examined here succeed to distinguish productive suffixes from the unproductive ones, there is still much work to be done to grasp the dynamics driving productivity from synchronic and diachronic perspectives.

Acknowledgments

This research is supported by the German Research Foundation *Deutsche Forschungsgemeinschaft* (DFG) in the Collaborative Research Center 673 'Alignment in Communication'. We would also like to thank Alexander Mehler for fruitful discussions and comments.

References

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, Mass.: MIT Press

Baayen, H. (1991). Quantitative aspects of morphological productivity. In: J. M. Geert Booij (ed.), *Yearbook of Morphology: 109–149*. Dordrecht: Kluwer.

Baayen, H. (1992). On frequency, transparency, and productivity. *Yearbook of Morphology 1992*: *181–208*.

Bauer, L. (2001). *Morphological productivity*. Cambridge: Cambridge University Press.

Bybee, J.L. (1988). *Morphology as lexical organization*. New York: Academic Press.

Cowie, C., Dalton-Puffer, C. (2002). Diachronic word-formation over time: Theoretical and methodological considerations. In: Javier Días Vera. *A Changing World of Words. Studies in English Historical Lexicography, Lexicology and Semantics*: *410–437*. Rodopi.

Diewald, G. (1997). *Grammatikalisierung: Eine Einführung in Sein und Werden grammatischer Formen*. Tübingen: Niemeyer.

Dressler, W.U., Karpf, A. (1995). The theoretical relevance of pre- and protomorpholgy in language acquisition. *Yearbook of Morphology 1994: 99–122*. Dordrecht: Kluwer.

Evert, S., Lüdeling, A. (1991). Constraining psycholinguistic models of morphological processing and representation: the role of productivity. In: G. Booij, Marle, J.v. (eds.), *Yearbook of morphology 1991: 165–183.* Dordrecht: Kluwer.

Evert, S., Lüdeling, A. (2001). Measuring morphological productivity: is automatic preprocessing sufficient? In: P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*: *167–175*. Lancaster: Peter Lang.

Fleischer, W., Barz, I. (1995). *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Max Niemeyer Verlag.

Habermann, M. (1994). *Verbale Wortbildung um 1500. Eine historisch-synchrone Untersuchung anhand von Texten Albrecht Dürers, Heinrich Deichlers und Veit Dietrichs*. Berlin: de Gruyter.

Hay. J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics 39(4), 1041–1070.*

Heine, B. (2003). Grammaticalization. In: Joseph, B. D. & Janda, R. D. (Eds.), *The Handbook of Historical Linguistics: 575–601*. Oxford: Blackwell.

Hinrichs, E.W., Kübler, S. (2005). Treebank profiling of spoken and written German. In: *Proc. of the Fourth Workshop on Treebanks and Linguistic Theories (TLT)*, December 2005. Barcelona, Kluwer Academic Publishers.

Hopper, P.J., Traugott, E.C. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.

Jäger, G. (2008). Evolutionary stability conditions for signaling games with costly signals. *Journal of Theoretical Biology 253, 131–141.*

Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In: C. Lyon, C. Nehaniv, A. Cangelosi (eds.), *Emergence of Communication and Languag*e: 253–268. Berlin: Springer Verlag..

Kruisinga, E. (1932). *A Handbook of Present-Day English, volume 5*. Groningen: Noordhoff.

Lehmann, C. (1995). *Thoughts on grammaticalization*. LINCOM studies in theoretical linguistics 1. LINCOM Europa,. München/Newcastle: Lincom (= Lincom Studies in Theoretical Linguistics 01).

Lohde, M. (2006). *Wortbildung des modernen Deutschen. Ein Lehr- und Übungsbuch*. Tübingen: Narr.

Lüdeling, A., Evert, S. (2004). The emergence of productive non-medical *-itis*: corpus evidence and qualitative analysis. In: *Proceedings of the First International Conference on Linguistic Evidence: 351–370.* Berlin/New York: Mouton de Gruyter.

Mehler, A., Geibel, P., Pustylnikov, O. (2007). Structural classifiers of text types: Towards a novel model of text representation. *Journal for Language Technology and Computational Linguistics (JLCL) 22(2), 51–66.*

Munske, H. (2002). Wortbildungswandel. In: Habermann et al. (ed.), *Historische Wortbildung des Deutschen: 23–40*. Tübingen: Niemeyer.

Nübling, D., Dammel, A., Duke, J., Szczepaniak, R. (2006). *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen: Gunter Narr Verlag.

Plag, I. (1999). *Morphological productivity. Structural constraints in English derivation*. Berlin-New York: Mouton de Gruyter.

Prell, H.-P. (1991). *Die Ableitung von Verben aus Substantiven in biblischen und nichtbiblischen Texten des Frühneuhochdeutschen*. Frankfurt am Main: Lang.

Pustylnikov, O. (2009). Modeling learning of derivation morphology in a multi-agent simulation. In: *Proceedings of IEEE Africon 2009*. IEEE, September 2009. AFRICON 2009, Nairobi, IEEE.

Scherer, C. (2005). *Wortbildungswandel und Produktivität. Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Schneider-Wiejowski, K. (2009) Sprachwandel anhand von Produktivitätsverschiebungen in der schweizerdeutschen Derivationsmorphologie. *Linguistik online 38, 2009.*

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren 2, 110–125.*

Steels, L. (2005). The role of construction grammar in language grounding. http://www.isrl.uiuc.edu/ amag/langev/paper/steels_constructionGrammar.html, (Retrieved November 2005).

Stricker, S. (2000). *Substantivbildung durch Suffixableitung um 1800: untersucht an Personenbezeichnungen in der Sprache Goethes.* Heidelberg: Universitätsverlag Winter.

Süddeutscher Verlag. (2004). *Süddeutsche Zeitung 1994-2003.* 10 Jahre auf DVD. Süddeutscher Verlag. (München).

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, March 2005.

Vogt, P. (2003). Iterated learning and grounding: From holistic to compositional languages. In: S. Kirby (ed.), *Proceedings of Language Evolution and Computation Workshop/Course at ESSLLI*, pages 76–86, Vienna, 2003.

Willmanns, W. (1899).. *Deutsche Grammatik. 2. Abteilung. Wortbildung*. Verlag von Karl J. Trübner, Strassburg, 2 edition.

Appendix

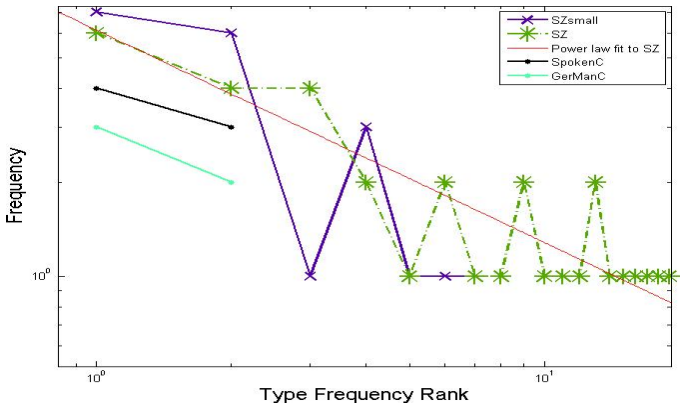


Figure 4: Frequency distributions of types formed with *-nis* in GerManC, SZsmall, SZ and SpokenC on a log-log plot (values from Table 8).

Table 8
Frequencies of hapax, dis, ..., *n*-legomena formed with *-nis* in GerManC, SZsmall, SZ and SpokenC.

Type Frequency	GerManC	SZsmall	SZ	SPOK
1	3	7	6	3
2	2	6	4	4
3		1	4	
4		3	2	
5		1	1	
6		1	2	
7		1	1	
8			1	
9			2	
10			1	
11			1	
12			1	
13			2	
14			1	
15			1	
16			1	
17			1	
18			1	
19			1	

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 126-154.*

Diversification in Icelandic Inflectional Paradigms[1]

Petra Steiner

1 The Implications of Zipf's Forces on Morphosyntax

Zipf's (1949) well-known principle of least effort comprises two forces:

(a) Diversification in language occurs whenever one linguistic unit of one level is associated with more than one unit of another level. Considering lexical meaning, one form (e.g. *field*) can be associated with more than one meaning (e.g. the senses of *grassland*, *runners in a race*, *physical field* etc.). On the level of inflectional morphology one form, for example one inflectional suffix, can indicate various different (combinations) of grammatical properties, sometimes even within the same inflectional paradigm. This is usually referred to as *case syncretism* (see Jakobson 1971: 67) and is defined as "homonymy of inflection markers in some grammatical domain" (Müller, Gunkel & Zifonun 2004: 6).

(b) Unification as the antagonistic process leads to a one-to-one relationship between form and meaning or form and function. Inflectional morphemes in agglutinative languages are typical instances of this.

Usually, diversification (of meaning) is described as caused by speaker's requirements, because thus the inventory is kept small, and effort for differentiating meaning can be minimized. At the hypothetical extreme point of such a process, there would be only one form left with many meanings, uses, or functions. On the opposite side, unification meets the requirements of the hearer, who prefers minimizing the effort of disambiguation. In the hypothetical extreme case, every form would be associated with exactly one meaning, one use or one function. Every function would be expressed by one form. However, as both forces are operating, the speaker's original tendencies are blocked because the hearer's requirements urge the speaker to use a few meanings often and the rest of them rarely. This must lead to characteristic distributions of frequencies. However, before these are derived, two widely known approaches to the explanation for inflection classes are to be discussed in the following section.

[1] I like to thank my colleagues Birgit Hellwig, Claudia Prün and especially Ásta Svavarsdóttir for their assistance. Many thanks to Gabriel Altmann, who carefully supervised the statistical part of this paper.

2 Principles in Inflection

In modern psycholinguistics, the concept of Zipf's force of unification is mirrored in the "Principle of Contrast": "Speakers assume that any difference in form signals a difference in meaning." (Clark 2003: 144). By this assumption, language acquisition is facilitated if forms have a minimal range of meaning or functions. This principle clearly contradicts the reality of inflectional languages, where one morphosyntactic function can be expressed by more than one allomorph and the same form can function as different morphs (see Carstairs-McCarthy 1994: 738). As an alternative, Carstairs-McCarthy (1994: 741) discusses a wider interpretation of 'meaning'. In that sense, an inflectional affix could indicate a unique membership of an inflectional paradigm and thus the Principle of Contrast (or of unification) would still hold. This is also formulated as the assumption of *lexical economy*, where "a single form predicts the full paradigm – and thus determines the inflection class – of any (non-suppletive) lexeme" (Blevins 2004: 52f). However, for Icelandic, this pattern does not apply, as the same form (e.g. *–ar*) can belong to several paradigms and different paradigm cells (e.g. genitive singular, nominative plural). Besides this, this assumption would result in paradigms with completely distinct inflectional affixes. Carstairs-McCarthy tries to avert the failure of the Principle of Contrast by postulating the constraint of the "No Blur Principle":

> Within any set of competing inflectional affixal realizations for the same paradigmatic cell, no more than one can fail to identify inflection class unambiguously (1994: 742).

He introduces the following (very condensed) system of inflectional affixes for Icelandic monosyllabic feminine nouns (Carstairs-McCarthy 1994: 740).

	CLASS A	CLASS B	CLASS C	CLASS D
GEN sg:	-ar	-ar	-ar	-(u)r
NOM pl:	-ir	-ar	-(u)r	-(u)r

For this small selection of inflectional forms, the principle applies correctly. However, in contrast to Carstairs-McCarthy's claim that this principle holds in general for Icelandic noun paradigms, it can be shown that more than one allomorph occurs in more than one paradigm. For instance, *-ur*, zero morph and *-i* can each function in more than one paradigm as indicators for masculine nominative singular. The same holds for *-ur* as well as the zero morph of feminine nominative singular forms. Masculine genitive singular can be signified by *-a*, *-s* and *-ur* in more than one class. Regarding the No Blur Principle for genitive singular feminine nouns, which is the second example in Carstair-McCarthy (1994: 740), it can be shown that this neglects the fact that *-ur* does not only mark the genitive singular feminine for the inflectional paradigms of *vík* and *nótt*,

but also for those of *móðir*, *dóttir* and *systir* (see Table 9, Appendix). The constraint of monosyllabicity evades these exceptions.

Although the No Blur Principle in its absolute can be refuted straightforwardly, there is certainly a statistical connection between allomorphs and their inflectional paradigm membership. Considering this as a statistical hypothesis would diminish the need of explanations for 'exceptions' (cf. Carstairs-McCarthy 1994: 747) and the construction of macroclasses to avoid ambiguities that are inconsistent with the principle (cf. McCarthy 1994: 751).

While the 'Principle of Contrast' needs a lot of support from other principles, as well as constraints and reclassifications into macroclasses to be upheld as valid, syncretism is often 'explained' by underspecification (e.g. Müller, Gunkel & Zifonun 2004: 6f):

> According to these analyses, standard case features like [+nom], [+acc], [+dat], and [+gen] are decomposed into the more primitive features [±gov(erned)], [±obl(ique)]. (Müller, Gunkel & Zifonun 2004: 7)

Those features are considered to be systematic and underlying some principles, e.g. the *Syncretism Principle*, which postulates "Identity of form implies identity of function (in a domain Σ, and unless there is evidence to the contrary)" (Müller 2004: 197). This is a variant of the Principle of Contrast. For its underpinning, cases and inflection class features are decomposed into primitive features including those which designate those properties which two or more inflectional paradigms have in common. Logically, this contributes to the 'predictability' of class membership. According to this approach, inflectional class features trigger inflection (Müller 2004: 211), but the question why such features should be necessary is left unanswered.

Current approaches concerning *syncretism* and *lexical economy* regard the astounding complexity of inflectional systems primarily from the hearer-learner's perspective and then find ways to demonstrate that the complexity is really simpler than it appears. However, considering speakers' requirements adds the following aspects to the picture:

(a) Speakers tend towards assimilation, epentheses, deletions, shortenings and everything that facilitates pronunciation. These simplifications on the phonetic and phonological level result in diversity of morphological forms.

(b) Speakers tend to keep inventories of linguistic units small (Köhler 2005: 766ff). This applies not only to free morphemes but also to inflectional forms. For Icelandic, only fourteen inflectional morphemes exist, including some complementary variants.

(c) Inflectional classes can be subject to sociolinguistic constraints. Iceland was never an "isolated speech community" (cf. Carstairs-McCarthy 1994: 784), but has always averted external linguistic influences: formerly from Danish and German and nowadays from English. There are even well-

attested cases of paradigm shifts, which were reversed due to social constraints. For instance, the noun *læknir* (physician) was shifted from a new paradigm of five different morphs back to the former one with six inflectional affixes. Nowadays the first paradigm is considered "wrong" or non-standard (Kress 1982: 59; Kvaran 2005b: 1746f).

3 Hypotheses

Diversification processes do not only operate on the inventory of inflectional affixes, but also within (a) the inflectional paradigms and (b) their organization.

(a) Within inflectional paradigms, unification in its extreme form leads to a 1:1 relation of form and grammatical meaning, resulting in a different affix for each cell within the inflectional paradigm. On the other hand, the speaker's requirements demand in their extreme form that one form marks all cases. But if there is just one form used, no distinctive features are needed and the form could be completely omitted. As both tendencies work within Icelandic inflection, there are a few (and by tendency short) inflectional affixes used with high frequency, while some (and by tendency longer) inflectional affixes are used with relatively low frequency. Within a range of eight inflected word-forms twenty-two different combinations of ordered distributions of word-form types are theoretically possible. However, it can be expected that the forces of unification and diversification "tug" at both ends of this distribution.

(b) The number of Icelandic inflectional forms is influenced by the combinatorial potential of eight word-form cells. However, as the Zipfian forces must have an effect on the distribution of inflectional paradigms itself, the number of inflectional forms among the declension classes must be distributed according to a typical linguistic diversification distribution. This has been shown for German inflectional noun classes, for which the left truncated negative binomial distribution yields a good fit (see Steiner & Prün 2007). Icelandic does not only use inflectional affixes and ablaut for case marking, but also epentheses, deletion and other means. Therefore, classes of inflectional paradigms can be defined in different ways. In the following, four different measures of paradigm complexity will be defined. This will lead to different classifications of inflectional paradigm classes. It can be assumed that each such inflectional paradigm will or will not change its number of different suffixes and alternation stems within a certain period of time with a certain probability. For the modeling of such a process, we start from the assumption that a certain complexity will be attained with probability p; then the probability that out of n possibilities exactly x will be realized, can be expressed by the binomial distribution as shown below in Chapter 5.2.

4 Data

Islandic is a highly inflectional language, possessing a large variance in allomorphy. Noun paradigms are inflected according to four cases, for gender (masculine, feminine, neuter) and number (singular, plural) (Árnason 2005: 1566ff, Barðdal 2001: 11f). The complete paradigms of Icelandic nouns comprise eight positions. Table 1 presents some examples. Note that the iconic principle (Haiman 1980) does neither hold for the first example (*hestur*$_{\text{NOM.SING}}$: *hest*$_{\text{ACC.SING}}$) nor for the second one (*hafnar*$_{\text{GEN.SING}}$: *hafna*$_{\text{GEN.PL}}$).

Table 1
Four inflectional paradigms of Icelandic nouns

	hestur (horse)		**höfn (harbor)**	
	SING	PLUR	SING	PLUR
NOM.	hestur	hestar	höfn	hafnir
ACC.	hest	hesta	höfn	hafnir
DAT.	hesti	hestum	höfn	höfnum
GEN.	hests	hesta	hafnar	hafna

	móðir (mother)		**björn (bear)**	
	SING	PLUR	SING	PLUR
NOM.	móðir	mæður	björn	birnir
ACC.	móður	mæður	björn	birni
DAT.	móður	mæðrum	birni	björnum
GEN.	móður	mæðra	bjarnar	bjarna

Within noun paradigms, syncretism exists for neuter (nominative and accusative, singular and plural), masculine (accusative and dative for many paradigms) and feminine, where there are three identical affix forms of genitive, dative and accusative for most singular forms. Svavarsdóttir (1993: 55; 58ff) provides a systematic overview of affixes and their inflectional function. Noun paradigms comprise up to three different alternation stems caused by assimilation processes (Árnason 2005: 1565f, Kress 1982: 43f). Besides umlaut/ablaut alternation, there are variant forms of synkopes (e.g. *himinn*$_{\text{NOM.SING.MASC}}$: *himni*$_{\text{DAT.SING.MASC}}$ 'sky'), contractions (e.g. *tré*$_{\text{NOM.SING.NEUT}}$: *trjám*$_{\text{DAT.SING.NEUT}}$ 'tree/wood') (Kress 1982: 47) and epentheses. Diachronically, shifts of inflectional classes depend upon morphophonological conditions, e.g. for dative singular *–i* is attached if the stem ends in a consonant group. As articles are suffixed, this epenthesis leads to the avoidance of big consonant clusters (see Kress 1982: 69). This results in a shortening of the average syllable length and can be considered as effect of Menzerath's law (Altmann & Schwibbe 1989).

However, this investigation treats the data from a mere synchronic point. Therefore epentheses and other phonological changes are not considered if they occur over the whole paradigm. Obsolete forms which are still present in idioms are not included, also suppletive forms (e.g. *eyrir* vs. *aurar* 'øre'), as well as pluralia and singularia tanta are left out. Table 9 (in the Appendix) shows an overview of the full paradigms for Icelandic nominal inflection. The table is based on the overviews of Kress (1982), Kvaran (2005a) and Scholten (2000) and lists the inflectional classes according to Kress (1982) and Scholten (2000). If the class is unlisted in one of those grammars, a reference to Kvaran (2005a) is given. Gender as well as stem alternation by ablaut, syncope, elision and epenthesis are considered as distinctive features. As complexity of inflection can be understood in different ways, four different operational measures are defined. c_4 is defined as the number of different affixes of a paradigm. c_3 is defined as the sum of c_4 and the number of different alternation stems by ablaut. c_2 equals c_3 and the number of alternation stems by syncope and elision. Finally, c_1 comprises all kinds of alternation including epentheses, which are often phonologically determined. Different complexity measures yield different paradigm classes. For instance, considering solely inflectional affixes as criterion creates less classes than taking ablaut alternation into consideration additionally. Therefore some paradigms have to be unified and lead to a smaller set of classes.

f is the number of different word-forms within a paradigm, with the theoretical maximum of 8. For hypothesis (a), the unordered (integer) partitions of word-form types and the partitions of inflectional affixes are listed in p_1 and p_2. For instance, the paradigm for *móðir* (see Appendix, Table 9) comprises three forms of *móður*, two of *mæður*, while the other forms are singletons, therefore p_1 is (3 + 2 + 1 + 1 + 1). Regarding solely the affixes, the partition p_2 is (5 + 1 + 1 + 1). As can be easily seen, *f* equals the number of partitions in p_1.

Note that the frequency of paradigms in lexis and text is disregarded here. Some of the listed inflectional patterns are limited to only a few lexemes. This certainly has an effect on long-term changes within the structure of the paradigms. However, this should be the subject of a different investigation.

5 Tests of the statistical hypothesis

5.1 Distribution of Partitions

The counts of all possible combinations of unordered partitions of word-form types lead to the following picture presented in Table 2.

Table 2
Combinations of word-form types and their frequencies

Partitions of word-forms	Frequency
8	0
7 + 1	0
6 + 2	0
6 + 1 + 1	1
5+ 3	0
5 + 2 + 1	0
5 + 1 +1 + 1	5
4 + 4	0
4 + 3 + 1	0
4 + 2 + 2	0
4 + 2 + 1 + 1	7
4 + 1 + 1 + 1 + 1	5
3 + 3 + 2	0
3 + 3 + 1 + 1	10
3 + 2 + 2 + 1	2
3 + 2 + 1 + 1 + 1	22
3 + 1 + 1 + 1 + 1 + 1	2
2 + 2 + 2 + 2	0
2+ 2 +2 + 1 + 1	0
2+ 2 + 1 + 1 +1 + 1	15
2 + 1 + 1 + 1 + 1 + 1 +1	22
1 + 1 + 1 + 1 + 1 + 1 + 1 +1	0
Σ	91

As can be seen in Table 2, only 10 of 22 possible combinations exist in Icelandic noun paradigms. Of the unique combinations, at least the paradigm for 6 + 1 + 1 applies only to a small set of lexemes (*herra*, *séra*). A closer look at the attested combinations shows a preference for distributions with a strong gradient in the first part and a long tail for the remaining values.

Table 3 presents the frequencies for the patterns of inflectional affixes. A comparison with Table 2 indicates that stem alternation results in a lower steepness within the frequency distributions of inflectional forms. The same can be examined for German umlaut (Steiner & Prün 2007: 631). Uniform frequencies do not exist and the picture is clearly one which is typical for diversification. Frequencies for *f* are presented in Table 4.

Table 3
Combinations of inflectional affixes and their frequencies

Partition of affixes	Frequency
8	0
7 + 1	0
6 + 2	0
6 + 1 + 1	1
5+ 3	0
5 + 2 + 1	0
5 + 1 +1 + 1	10
4 + 4	0
4 + 3 + 1	0
4 + 2 + 2	0
4 + 2 + 1 + 1	7
4 + 1 + 1 + 1 + 1	7
3 + 3 + 2	0
3 + 3 + 1 + 1	10
3 + 2 + 2 + 1	4
3 + 2 + 1 + 1 + 1	15
3 + 1 + 1 + 1 + 1 + 1	4
2 + 2 + 2 + 2	0
2+ 2 +2 + 1 + 1	0
2+ 2 + 1 + 1 +1 + 1	13
2 + 1 + 1 + 1 + 1 + 1 +1	20
1 + 1 + 1 + 1 + 1 + 1 + 1 +1	0
Σ	91

Table 4
Frequencies *n* of paradigms with *f* word-form types (from Table 2)

f	*n*
3	1
4	24
5	27
6	17
7	22

Obviously, partitions with four to seven word-form types are preferred and are uniformly distributed (χ^2 = 2.35, DF = 3, P = 0.50). The group *f* = 3 is an outlier comprising only the two loan words *herra* (‘lord’) and *séra* (‘reverend’).

5.2 Distributions of complexity of inflectional paradigms

For the testing of the hypothesis (b) concerning the complexity of noun paradigms, the Altmann-Fitter (2000) is used. The complexity is calculated according to the assumptions in section 3. As Table 5 to Table 8 and Figure 1 to Figure 3 show, the 3-displaced binomial distribution

$$(1) \qquad P_x = \binom{n}{x-3} p^{x-3} q^{n-x+3}, \qquad x = 3, 4, ..., n+3$$

yields an excellent fit to all data.

Table 5
Frequency classes of complexity c_1
in Icelandic noun inflection paradigms
(data from Table 9)

c_1	$f(c_1)$	$NP(c_1)$
3	1	1.42
4	9	8.06
5	25	19.64
6	23	26.59
7	14	21.59
8	15	10.51
9	3	2.85
10	1	0.33
	n = 7, p = 0.44, X^2 = 6.97, DF = 4, P = 0.14	

Table 6
Frequency classes of complexity c_2
in Icelandic noun inflection paradigms
(data from Table 10)

c_2	$f(c_2)$	$NP(c_2)$
3	2	1.62
4	12	8.49
5	20	19.07
6	19	23.79
7	13	17.82
8	11	8.00
9	3	2.00
10	1	0.21
	n = 7, p = 0.43, X^2 = 6.42, DF = 4, P = 0.17	

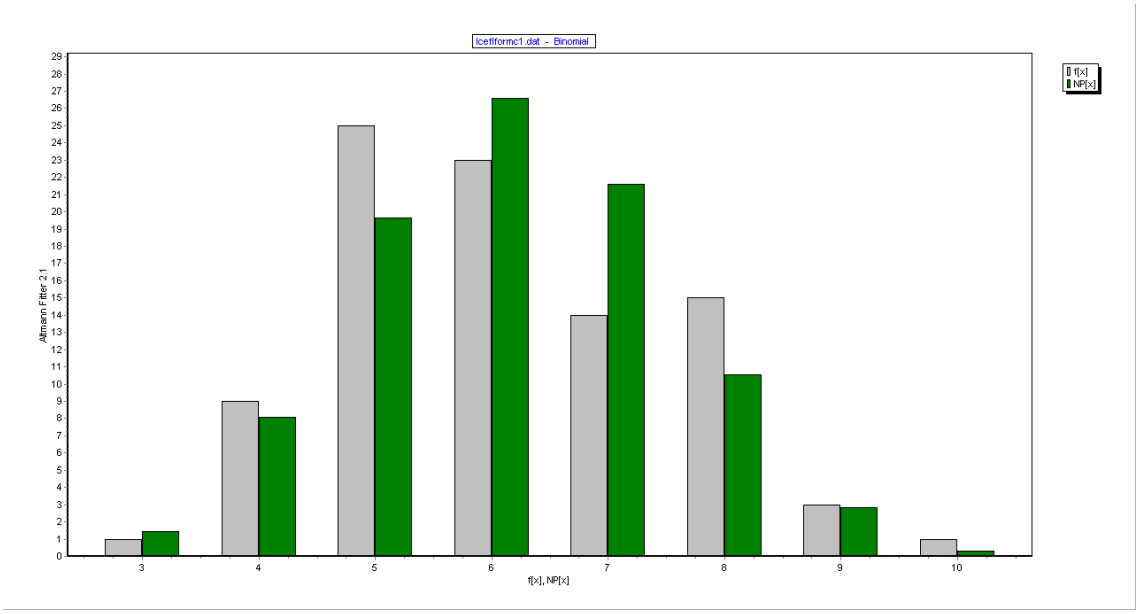


Figure 1. Frequency classes of different inflectional classes including all alternation stems, modeled by the binomial distribution (gray: empirical data; green: computed data)

Table 7
Frequency classes of complexity c_3
in Icelandic noun inflection paradigms
(data from Table 11)

c_3	$f(c_3)$	$NP(c_3)$
3	2	1.78
4	13	8.88
5	20	18.84
6	16	22.27
7	13	15.79
8	8	6.72
9	4	1.75
	$n = 7$, $p = 0.41$, $X^2 = 7.44$	
	$DF = 4$, $P = 0.11$	

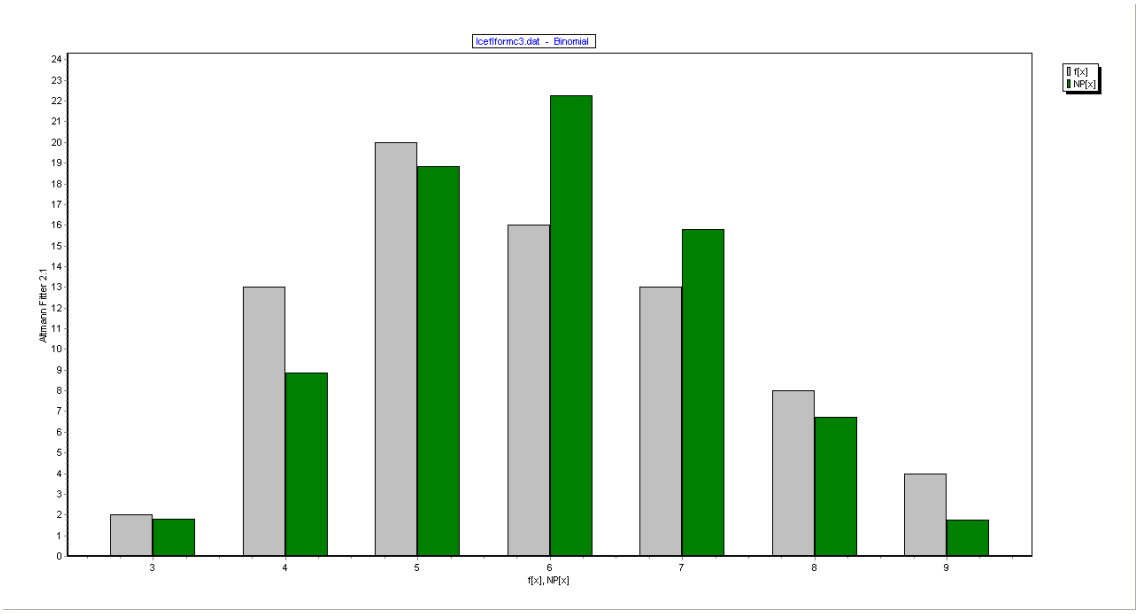


Figure 2. Frequency classes of different inflectional classes comprising ablaut, modeled by the binomial distribution

Table 8
Frequency classes of complexity c_4
in Icelandic noun inflection paradigms
(data from Table 12)

c_4	$f(c_4)$	$NP(c_4)$
3	2	3.45
4	17	11.83
5	12	16.91
6	12	12.89
7	9	6.91
	n = 6, p = 0.36, X^2 = 4.98 DF = 2, P = 0.10	

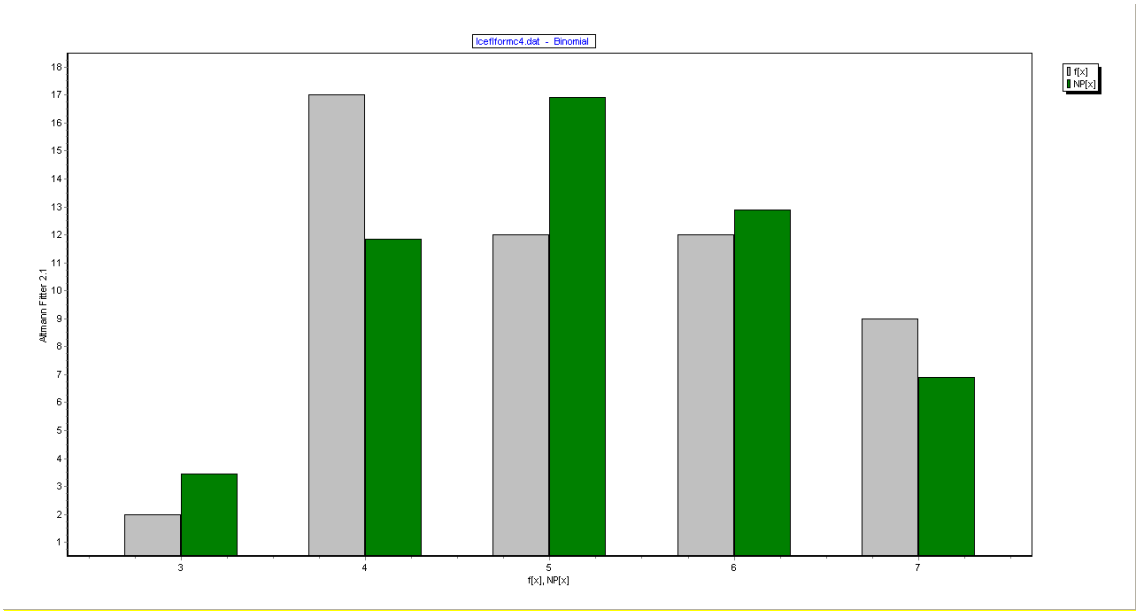


Figure 3. Frequency classes of different inflectional classes comprising affixes, modeled by the binomial distribution

According to the results of the chi-square test, the hypothesis that the frequencies of complexity show the effect of diversification can be accepted for all measures of complexity.

6 Discussion

While the left truncated negative binomial distribution yields a good fit for German inflectional noun classes (see Steiner & Prün 2007), the model of the binomial distribution matched better with the Icelandic data. At this point of the investigation, the question of why this is so has to be left open. An obvious advantage of the chosen model is certainly the limitation on a restricted class number, which corresponds with the fact that the number of complexity is determined by the number of paradigm cells combined with the features considered (ablaut, epenthesis etc.).

In a next step, it would be of interest to analyze how far the effects of diversification show up in dependence on the size of lexeme sets for which a certain paradigm applies. Counting the lexemes for each paradigm could lead to a smoother, unimodal distribution compared to the bimodal one in Table 4. The frequencies of the word-form types and affix partitions are clearly non-random: partitions with one or two different wordforms do not exist, though there are five possible combinations. Partitions of eight variants do not exist either. While this can be explained by the Zipfian forces, the peak of seven different forms remains astounding. Possibly, the bimodality gives hints towards changes in paradigm classes. The sociolinguistic constraints on Icelandic should receive a closer look under this aspect. In general, the theoretical distribution of paradigms is shifted towards the greater paradigm classes. Diversification can be assumed as the most important force for this.

Last but not least, looking at text corpora could answer such questions as: Are there relations between properties of inflectional class and productivity in word formation? Certainly, further research on these phenomena will be rewarding.

References

Altmann, Gabriel, Schwibbe, Michael (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim/Zürich/New York: Olms.

Altmann-Fitter 2.1 for Windows 95 / NT. Lüdenscheid: RAM-Verlag, 2000.

Árnason, Kristjan (2005). The standard languages and their systems in the 20th century I: Icelandic. In: Bandle, O. et al. (eds.), *The Nordic Languages*: *An International Handbook of the History of the North Germanic Languages*. *Vol. 2, 1560-1573.* Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science 22.2).

Barðdal, Jóhanna (2001). *Case in Icelandic: A Synchronic, Diachronic and Comparative Approach.* Lund: Department of Scandinavian Languages. (Lundastudier i Nordisk Språkvetenshap A 57).

Blevins, James P. (2004). Inflection Classes and Economy. In: G. Müller, L. Gunkel, G. Zifonun (eds.), *Explorations in Nominal Inflection: 41-85.* Berlin/New York: de Gruyter (Interface Explorations 10).

Beőthy, Erzsébet, Altmann, Gabriel (1991). The diversification of meaning of Hungarian verbal prefixes I. "meg-". In: Rothe, U. (ed.), *Diversification processes in language: grammar: 60-66.* Hagen: Rottmann.

Carstairs-McCarthy, Andrew (1994). Inflection classes, gender and the Principle of Contrast. *Language 70, 737-88.*

Clark, Eve V. (2003). *First Language Acquisition*. Cambridge: Cambridge University Press.

Haiman, John (1980). The Iconicity of Grammar: Isomorphism and Motivation. *Language 56, 515-540.*

Jakobson, Roman (1971). Beitrag zur allgemeinen Kasuslehre – Gesamtbedeutungen der russischen Kasus. In: Jakobson, Roman: *Selected Writings 2: Word and Language: 23-71*. The Hague: Mouton.

Köhler, Reinhard (2005) Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik/ Quantitative Linguistics: Ein internationales Handbuch/An international handbook: 760-774.* Berlin/New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft /Handbooks of Linguistics and Communicative Science 27).

Kress, Bruno (1982). *Isländische Grammatik*. Leipzig: VEB Verlag Enzyklopädie Leipzig.

Kvaran, Guðrún (2005a). *Íslensk tunga II. Orð. Handbók um beygingar- og orðmyndunarfræði*. Reykjavík: Almenna bókafélagið.

Kvaran, Guðrún (2005b). Written languages and forms of speech in Icelandic in the 20th century. In: Bandle, O. et al. (eds.), *The Nordic Languages*: *An International Handbook of the History of the North Germanic Languages*. *Vol. 2, 1742-1749*. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science 22.2).

Müller, Gereon (2004). On Decomposing Inflection Class Features: Syncretism in Russian Noun Inflection. In: Müller, G., Gunkel, L., Zifonun, G. (eds.), *Explorations in Nominal Inflection: 189-227*. Berlin/New York: de Gruyter (Interface Explorations 10).

Müller, Gereon, Gunkel, Lutz, Zifonun, Gisela (2004). Introduction. In: Müller, G., Gunkel, L., Zifonun, G. (eds.), *Explorations in Nominal Inflection: 1-20.* Berlin/New York: de Gruyter (Interface Explorations 10).

Rothe, Ursula (1991). Diversification of the Case in German: Genitive. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 140-156.* Hagen: Rottmann

Scholten, Daniel (2000). *Einführung in die isländische Grammatik*: *Ein Lehrbuch für Anfänger und Fortgeschrittene*. München: Philyra.

Svavarsdóttir, Ásta (1993). *Beygingakerfi nafnorða í nútímaíslensku.* [The

morphological system of nouns in Modern Icelandic.] Reykjavík: Málvísindastofnun Háskóla Íslands.

Steiner, Petra, Prün, Claudia (2007). The effects of diversification and unification on the inflectional paradigms of German noun. In: Grzybek, P., Köhler, R. (eds), *Exact Methods in the Study of Language and Text*. *Dedicated to Gabriel Altmann at the Occasion of his 75th Birthday: 623-632*. Berlin/New York: de Gruyter.

Zipf, George K. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley.

Appendix

Table 9
The inflectional paradigms of Icelandic nouns

Ns ... Gp: Nominative singular ... genitive plural; *abl etc*: ablaut, syncope, elision, epenthesis;
g: gender; *Kress, Sch, Kv*: references to Kress (1982), Scholten (2000), Kvaran (2005);
p_1: pattern of word-form type partitions; p_2: pattern of affix partitions
f: number of different word-forms (≤ 8) = number of partitions;

***measures of complexity*:**
c_1: number of affixes, ablaut, **epenthesis,** syncope, elision ...; c_2: number of affixes, ablaut, **syncope, elision**;
c_3: number of affixes, ablaut; c_4: number of affixes

r	Ns	As	Ds	Gs	Np	Ap	Dp	Gp	abl etc.	example	*g*	Kress	Sch/Kv	c_1	c_2	c_3	c_4	*f*	p_1	p_2
1	-	-	-	-ar	-	-	-um	-a	1x	mús	f	173	W7	5	5	5	4	5	3 2 1 1 1	5 1 1 1
2	-	-	-	-ar	-ar	-ar	-m	-a		stó	f	142	A15	4	4	4	4	4	3 3 1 1	3 3 1 1
3	-	-	-	-ar	-ar	-ar	-um	-a		nál, vél	f	140	A13	4	4	4	4	4	3 3 1 1	3 3 1 1
4	-	-	-	-ar	-ar	-ar	-um	-a	r-syn	lifur	f	140.1	A16	5	5	4	4	4	3 3 1 1	3 3 1 1
5	-	-	-	-ar	-ar	-ar	-um	-a	v-ep	stöð	f	143	VA2	5	4	4	4	4	3 3 1 1	3 3 1 1
6	-	-	-	-ar	-ar	-ar	-um	-a	i-elis	heiði, helgi	f	144	JA6	5	5	4	4	4	3 3 1 1	3 3 1 1
7	-	-	-	-ar	-ar	-ar	-um	-a	j-ep	skel	f	144	JA8	5	4	4	4	4	3 3 1 1	3 3 1 1
8	-	-	-	-ar	-ir	-ir	-um	-a		tíð, mynd, brún	f	157	I8, 20-38	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
9	-	-	-	-ar	-ir	-ir	-um	-a	1x	höfn, öxl	f	158	I8a	6	6	6	5	5	3 2 1 1 1	3 2 1 1 1
10	-	-	-	-ar	-ir	-ir	-um	-a	1x	verslun	f	159	I9	6	6	6	5	5	3 2 1 1 1	3 2 1 1 1
11	-	-	-	-ar	-r	-r	-m	-a		frú	f	142	A15	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
12	-	-	-	-ar	-r	-r	-m	-a	1x	fló, brú	f	170	W3	6	6	6	5	5	3 2 1 1 1	3 2 1 1 1
13	-	-	-	-ar	-ur	-ur	-um	-a	2x	önd, nögl	f	168	W5	7	7	7	5	5	3 2 1 1 1	3 2 1 1 1
14	-	-	-	-ar	-ur	-ur	-um	-a	1x	bók, gát	f	171	W4	6	6	6	5	5	3 2 1 1 1	3 2 1 1 1

15	-	-	-	-ar	-ur	-ur	-um	-a		kind, geit	f	172	W6a	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
16	-	-	-i	-ar	-ir	-i	-um	-a	2x	björn	m	162	U2	8	8	8	6	6	2 2 1 1 1 1	2 2 1 1 1 1
17	-	-	-i	-ar	-ur	-ur	-um	-a	2x	hönd	f	168	W5	8	8	8	6	6	2 2 1 1 1 1	2 2 1 1 1 1
18	-	-	-	-r	-r	-r	-m	-a		á	f	142	A15	4	4	4	4	4	3 3 1 1	3 3 1 1
19	-	-	-	-s	-	-	-m	-a	já-ep	tré	n	134	A11c	5	4	4	4	4	5 1 1 1	5 1 1 1
20	-	-	-	-s	-	-	-um	-a	i-eli	klæði, tæki	n	138	JA3(a)	5	5	4	4	4	5 1 1 1	5 1 1 1
21	-	-	-	-s	-ar	-a	-um	-a	v-ep	spör	m	121	20-16	6	5	5	5	5	3 2 1 1 1	3 2 1 1 1
22	-	-	-	-s	-ar	-a	-um	-a		bjór, ís, strætó	m	120	A9	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
23	-	-	-	-s	-ir	-i	-um	-a		hver	m	155	I7	6	6	6	6	6	3 1 1 1 1 1	3 1 1 1 1 1
24	-	-	-	-s	-ir	-i	-um	-a	1x	Às	m	163		7	7	7	6	6	3 1 1 1 1 1	3 1 1 1 1 1
25	-	-	-i	-	-ar	-a	-um	-a	1x	dans	m	124.3	A7	6	6	6	5	5	3 2 1 1 1	3 2 1 1 1
26	-	-	-i	-s	-	-	-um	-a	j-eph	ber, kyn	n	138	JA4	6	5	5	5	5	4 1 1 1 1	4 1 1 1 1
27	-	-	-i	-s	-	-	-um	-a		borð, hús, högg	n	133	A11	5	5	5	5	5	4 1 1 1 1	4 1 1 1 1
28	-	-	-i	-s	-	-	-um	-a	1x	land, barn	n	135	A11a	6	6	6	5	6	2 2 1 1 1 1	4 1 1 1 1
29	-	-	-i	-s	-	-	-um	-a	ð-/r-syn	hreiður, höfuð	n	136	A12	6	6	5	5	5	4 1 1 1 1	4 1 1 1 1
30	-	-	-i	-s	-	-	-um	-a	1x	meðal	n	135	A11b	6	6	6	5	6	2 2 1 1 1 1	4 1 1 1 1
31	-	-	-i	-s	-ar	-a	-um	-a		fugl	m	124.3	A6	6	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
32	-	-	-i	-s	-ar	-a	-um	-a	1x, r-syn	akur, hamar	m	120	A8	8	8	7	6	6	2 2 1 1 1 1	2 2 1 1 1 1
33	-	-	-	-ur	-ur	-ur	-um	-a		vík	f	169	W6	4	4	4	4	4	3 3 1 1	3 3 1 1
34	-	-	-	-ur	-ur	-ur	-um	-a	2x	mörk	f	168	W5a	6	6	6	4	4	3 3 1 1	3 3 1 1
35	-	-	-	-ur	-ur	-ur	-um	-a	1x	nótt	f	168	W4a	5	5	5	4	4	3 3 1 1	3 3 1 1
36	-	-u	-u	-ar	-ar	-ar	-um	-a		drottning	f	141	A14	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
37	-a	-a	-a	-a	-ar	-a	-um	-a		herra	m	176.3	20-27	3	3	3	3	3	6 1 1	6 1 1
38	-a	-a	-a	-a	-u	-u	-um	-a	n-ep	auga	n	184	N4a	4	3	3	3	4	4 2 1 1	5 1 1 1
39	-a	-a	-a	-a	-u	-u	-um	-a	1x/n-ep	hjarta	n	184	N4b	5	4	4	3	4	4 2 1 1	4 2 1 1
40	-a	-u	-u	-u	-ur	-ur	-um	-a	1x	fata, taska	f	180	N5b	5	5	5	4	4	3 2 2 1	3 2 2 1
41	-a	-u	-u	-u	-ur	-ur	-um	-a	n-ep	tunga	f	181.2	N6a	5	4	4	4	5	3 2 1 1 1	3 2 2 1
42	-a	-u	-u	-u	-ur	-ur	-um	-a	1x/n-ep	saga	f	182.2	N6b	6	5	5	4	5	3 2 1 1 1	3 2 2 1
43	-a	-u	-u	-u	-ur	-ur	-um	-a		lilja	f	180	N5a, 20-40	4	4	4	4	4	3 2 2 1	3 2 2 1
44	-i	-a	-a	-a	-ar	-a	-um	-a		penni, tími, foringi	m	176	20-27	4	4	4	4	4	5 1 1 1	5 1 1 1

45	-i	-a	-a	-a	-ar	-a	-um	-a	1x	hani, bakari	m	177	N1	5	5	5	4	4	5 1 1 1	5 1 1 1
46	-i	-a	-a	-a	-ir	-i	-um	-a	1x	Dani	m	148	N1a	5	5	5	4	4	4 2 1 1	4 2 1 1
47	-i	-a	-a	-a	-ir	-i	-um	-a	j-ep	Tyrki	m	150	N2a	5	4	4	4	4	4 2 1 1	4 2 1 1
48	-i	-a	-a	-a	-ur	-ur	-um	-a	1x	nemandi, bóndi	m	185	P1, P2	5	5	5	4	5	3 2 1 1 1	4 2 1 1
49	-i	-i	-i	-i	-ar	-ar	-um	-a		lygi	f	183	N8	4	4	4	4	4	4 2 1 1	4 2 1 1
50	-i	-i	-i	-i	-ir	-ir	-um	-a		ævi, beiðni	f	183	N7	4	4	4	4	4	4 2 1 1	4 2 1 1
51	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	2x, r-syn	faðir	m	174	R1a	7	7	6	4	5	3 2 1 1 1	5 1 1 1
52	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	móðir, dóttir	f	174	R2a,b	6	6	5	4	5	3 2 1 1 1	5 1 1 1
53	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	bróðir	m	174	R1	6	6	5	4	5	3 2 1 1 1	5 1 1 1
54	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	r-syn	systir	f	174	R2c	5	4	4	4	4	5 1 1 1	5 1 1 1
55	-l	-	-	-s	-ar	-a	-um	-a		hóll, makrill, stóll	m	124.2	A5	6	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
56	-l	-	-i	-s	-ar	-a	-um	-a	l-syn	lykill	m	119	A4	8	8	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
57	-l	-	-i	-s	-ar	-ar	-um	-a	2x, l-syn	ketill	m	119.2	20-5	10	10	9	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
58	-n	-	-i	-s	-ar	-a	-um	-a	n-syn	himinn	m	119	A2	8	8	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
59	-n	-	-i	-s	-ar	-a	-um	-a		steinn	m	124.2	20-16	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
60	-r	-	-	-ar	-ir	-i	-um	-a	j-ep	bær	m	151	I6	8	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
61	-r	-	-	-r	-r	-r	-m	-a	1x	kýr	f	167	W2	5	5	5	4	4	4 2 1 1	4 2 1 1
62	-r	-	-	-s	-r	-	-m	-a		skór	m	130	A10	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
63	-r	-	-	-s	-ar	-a	-um	-a		snjór	m	122	A10	6	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
64	-r	-	-	-s	-ar	-a	-um	-a	i-del	læknir, kíkir	m	123	JA1	7	7	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
65	-r	-	-	-s	-ar	-a	-um	-a	v-ep	már	m	121		7	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
66	-r	-	-	-s	-ar	-a	-um	-a	v-ep, 1x	týr	m	121		8	7	7	6	7	2 2 1 1 1 1	2 2 1 1 1 1
67	-ur	-	-	-ar	-ir	-i	-um	-a	j-ep	lækur, leggur	m	149	I5	8	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
68	-ur	-	-	-ar	-ir	-i	-um	-a		litur	m	147	I4	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
69	-ur	-	-	-ar	-ir	-i	-um	-a	1x	staður	m	147	20-24	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
70	-ur	-	-	-s	-ar	-a	-um	-a	j-ep	niður	m	123	JA2	7	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
71	-ur	-	-	-s	-ar	-a	-um	-a	v-ep	söngur	m	121	VA1	7	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
72	-ur	-	-	-s	-ir	-i	-um	-a		smiður	m	148	20-22	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
73	-ur	-	-	-s	-ir	-i	-um	-a	1x	dalur	m	148	I3	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
74	-ur	-	-	-s	-ir	-i	-um	-a	j-ep	belgur, leikur	m	150	I5a	8	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1

75	-ur	-	-i	-ar	-ar	-a	-um	-a		skógur	m	129.2	A1a	6	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
76	-ur	-	-i	-ar	-ir	-i	-um	-a		fundur	m	147	I1	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
77	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	mánaður	m	164	U5	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
78	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	háttur	m	163	U3	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
79	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	söfnuður	m	164	U5	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
80	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	sonur	m	163	U4	8	8	8	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
81	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	fjörður	m	162	U2	9	9	9	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
82	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	köttur, völlur	m	161	U1	9	9	9	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
83	-ur	-	-i	-ar	-ur	-ur	-um	-a	1x	fótur	m	165	W1b	7	7	7	6	7	2 1 1 1 1 1 1	3 1 1 1 1 1
84	-ur	-	-i	-s	-ir	-i	-um	-a		gestur	m	148	I2	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
85	-ur	-	-i	-s	-	-	-um	-a	3x	maður	m	165	W1d	9	9	9	6	7	2 1 1 1 1 1 1	3 1 1 1 1 1
86	-ur	-	-i	-s	-ar	-a	-um	-a		hestur	m	118	A1	7	7	7	7	7	2 1 1 1 1 1 1	2 1 1 1 1 1 1
87	-ur	-i	-i	-ar	-ar	-ar	-um	-a		hildur	f	145	JA5	5	5	5	5	5	3 2 1 1 1	3 2 1 1 1
88	-ur	-i	-i	-ar	-ar	-ar	-um	-a	j-ep	ylgur	f	145.1	JA7	6	5	5	5	5	3 2 1 1 1	3 2 1 1 1
89	-ur	-i	-i	-ar	-ir	-ir	-um	-a		bruður	f	145		6	6	6	6	6	2 2 1 1 1 1	2 2 1 1 1 1
90	-ur	-ur	-i	-ar	-ur	-ur	-um	-a	r-syn	vetur	m	165	W1A	6	6	5	5	5	4 1 1 1 1	4 1 1 1 1
91	-ur	-ur	-i	-s	-ur	-ur	-um	-a	r-syn	fingur	m	165	W1c	6	5	5	5	5	4 1 1 1 1	4 1 1 1 1

Table 10
Unification according to complexity c_2

0.	*Ns*	*As*	*Ds*	*Gs*	*Np*	*Ap*	*Dp*	*Gp*	*abl etc*	*example*	*g*	c_1	c_2
1.	-	-	-	-ar	-	-	-um	-a	1x	mús	f	5	5
2.	-	-	-	-ar	-ar	-ar	-m	-a		stó	f	4	4
3.	-	-	-	-ar	-ar	-ar	-um	-a		nál, vél, stöð, skel	f	5,4	4
4.	-	-	-	-ar	-ar	-ar	-um	-a	r-syn	lifur	f	5	5
5.	-	-	-	-ar	-ar	-ar	-um	-a	i-elis	heiði, helgi	f	5	5
6.	-	-	-	-ar	-ir	-ir	-um	-a		tíð, mynd, brún	f	5	5
7.	-	-	-	-ar	-ir	-ir	-um	-a	1x	höfn, öxl	f	6	6
8.	-	-	-	-ar	-ir	-ir	-um	-a	1x	verslun	f	6	6
9.	-	-	-	-ar	-r	-r	-m	-a		frú	f	5	5
10	-	-	-	-ar	-r	-r	-m	-a	1x	fló, brú	f	6	6
11	-	-	-	-ar	-ur	-ur	-um	-a	2x	önd, nögl	f	7	7
12	-	-	-	-ar	-ur	-ur	-um	-a	1x	bók, gát	f	6	6
13	-	-	-	-ar	-ur	-ur	-um	-a		kind, geit	f	5	5
14	-	-	-i	-ar	-ir	-i	-um	-a	2x	björn	m	8	8
15	-	-	-i	-ar	-ur	-ur	-um	-a	2x	hönd	f	8	8
16	-	-	-	-r	-r	-r	-m	-a		á	f	4	4
17	-	-	-	-s	-	-	-m	-a	já-ep	tré	n	5	4
18	-	-	-	-s	-	-	-um	-a	i-eli	klæði, tæki	n	5	5
19	-	-	-	-s	-ar	-a	-um	-a		bjór, ís, strætó, spör	m	6,5	5
20	-	-	-	-s	-ir	-i	-um	-a		hver	m	6	6
21	-	-	-	-s	-ir	-i	-um	-a	1x	Às	m	7	7
22	-	-	-i	-	-ar	-a	-um	-a	1x	dans	m	6	6
23	-	-	-i	-s	-	-	-um	-a		ber, kyn, borð, hús, högg	n	6,5	5
24	-	-	-i	-s	-	-	-um	-a	1x	land, barn	n	6	6

25	-	-	-i	-s	-	-	-um	-a	ð-/ r-syn	hreiður, höfuð	n	6	6
26	-	-	-i	-s	-	-	-um	-a	1x	meðal	n	6	6
27	-	-	-i	-s	-ar	-a	-um	-a		fugl	m	6	6
28	-	-	-i	-s	-ar	-a	-um	-a	1x, r-syn	akur, hamar	m	8	8
29	-	-	-	-ur	-ur	-ur	-um	-a		vík	f	4	4
30	-	-	-	-ur	-ur	-ur	-um	-a	2x	mörk	f	6	6
31	-	-	-	-ur	-ur	-ur	-um	-a	1x	nótt	f	5	5
32	-	-u	-u	-ar	-ar	-ar	-um	-a		drottning	f	5	5
33	-a	-a	-a	-a	-ar	-a	-um	-a		herra	m	3	3
34	-a	-a	-a	-a	-u	-u	-um	-a	n-ep	auga	n	4	3
35	-a	-a	-a	-a	-u	-u	-um	-a	1x/n-ep	hjarta	n	5	4
36	-a	-u	-u	-u	-ur	-ur	-um	-a	1x	fata, taska	f	5	5
37	-a	-u	-u	-u	-ur	-ur	-um	-a		tunga, lilja	f	5,4	4
38	-a	-u	-u	-u	-ur	-ur	-um	-a	1x/n-ep	saga	f	6	5
39	-i	-a	-a	-a	-ar	-a	-um	-a		penni, tími, foringi	m	4	4
40	-i	-a	-a	-a	-ar	-a	-um	-a	1x	hani, bakari	m	5	5
41	-i	-a	-a	-a	-ir	-i	-um	-a	1x	Dani	m	5	5
42	-i	-a	-a	-a	-ir	-i	-um	-a	j-ep	Tyrki	m	5	4
43	-i	-a	-a	-a	-ur	-ur	-um	-a	1x	nemandi, bóndi	m	5	5
44	-i	-i	-i	-i	-ar	-ar	-um	-a		lygi	f	4	4
45	-i	-i	-i	-i	-ir	-ir	-um	-a		ævi, beiðni	f	4	4
46	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	2x, r-syn	faðir	m	7	7
47	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	móðir, dóttir	f	6	6
48	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	bróðir	m	6	6
49	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	r-syn	systir	f	5	4
50	-l	-	-	-s	-ar	-a	-um	-a		hóll, makrill, stóll	m	6	6

51	-l	-	-i	-s	-ar	-a	-um	-a	l-syn	lykill	m	8	8
52	-l	-	-i	-s	-ar	-ar	-um	-a	2x, l-syn	ketill	m	10	10
53	-n	-	-i	-s	-ar	-a	-um	-a	n-syn	himinn	m	8	8
54	-n	-	-i	-s	-ar	-a	-um	-a		steinn	m	7	7
55	-r	-	-	-ar	-ir	-i	-um	-a	j-ep	bær	m	8	7
56	-r	-	-	-r	-r	-r	-m	-a	1x	kýr	f	5	5
57	-r	-	-	-s	-r	-	-m	-a		skór	m	5	5
58	-r	-	-	-s	-ar	-a	-um	-a		snjór, már	m	7,6	6
59	-r	-	-	-s	-ar	-a	-um	-a	i-del	læknir, kíkir	m	7	7
60	-r	-	-	-s	-ar	-a	-um	-a	v-ep, 1x	týr	m	8	7
61	-ur	-	-	-ar	-ir	-i	-um	-a	j-ep	lækur, leggur, litur	m	8,7	7
62	-ur	-	-	-ar	-ir	-i	-um	-a	1x	staður	m	8	8
63	-ur	-	-i	-ar	-ar	-a	-um	-a		skógur	m	6	6
64	-ur	-	-i	-ar	-ir	-i	-um	-a		fundur	m	7	7
65	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	mánaður	m	8	8
66	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	háttur	m	8	8
67	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	söfnuður	m	8	8
68	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	sonur	m	8	8
69	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	fjörður	m	9	9
70	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	köttur, völlur	m	9	9
71	-ur	-	-i	-ar	-ur	-ur	-um	-a	1x	fótur	m	7	7
72	-ur	-	-	-s	-ar	-a	-um	-a	j/v-ep	niður, söngur	m	7	6
73	-ur	-	-	-s	-ir	-i	-um	-a	1x	dalur	m	8	8
74	-ur	-	-	-s	-ir	-i	-um	-a		smiður, belgur, leikur	m	8,7	7
75	-ur	-	-i	-s	-ir	-i	-um	-a		gestur	m	7	7
76	-ur	-	-i	-s	-	-	-um	-a	3x	maður	m	9	9
77	-ur	-	-i	-s	-ar	-a	-um	-a		hestur	m	7	7
78	-ur	-i	-i	-ar	-ar	-ar	-um	-a		hildur, ylgur	f	6,5	5
79	-ur	-i	-i	-ar	-ir	-ir	-um	-a		bruður	f	6	6
80	-ur	-ur	-i	-ar	-ur	-ur	-um	-a	r-syn	vetur	m	6	6
81	-ur	-ur	-i	-s	-ur	-ur	-um	-a	r-syn	fingur	m	6	5

Table 11
Unification according to complexity c_3

0.	*Ns*	*As*	*Ds*	*Gs*	*Np*	*Ap*	*Dp*	*Gp*	*abl etc*	*example*	*g*	c_1	c_2	c_3
1.	-	-	-	-ar	-	-	-um	-a	1x	mús	f	5	5	5
2.	-	-	-	-ar	-ar	-ar	-m	-a		stó	f	4	4	4
3.	-	-	-	-ar	-ar	-ar	-um	-a		nál, vél, stöð, skel, lifur, hei-ði, helgi	f	5,4	5,4	4
4.	-	-	-	-ar	-ir	-ir	-um	-a		tíð, mynd, brún	f	5	5	5
5.	-	-	-	-ar	-ir	-ir	-um	-a	1x	höfn, öxl	f	6	6	6
6.	-	-	-	-ar	-ir	-ir	-um	-a	1x	verslun	f	6	6	6
7.	-	-	-	-ar	-r	-r	-m	-a		frú	f	5	5	5
8.	-	-	-	-ar	-r	-r	-m	-a	1x	fló, brú	f	6	6	6
9.	-	-	-	-ar	-ur	-ur	-um	-a	2x	önd, nögl	f	7	7	7
10.	-	-	-	-ar	-ur	-ur	-um	-a	1x	bók, gát	f	6	6	6
11.	-	-	-	-ar	-ur	-ur	-um	-a		kind, geit	f	5	5	5
12.	-	-	-i	-ar	-ir	-i	-um	-a	2x	björn	m	8	8	8
13.	-	-	-i	-ar	-ur	-ur	-um	-a	2x	hönd	f	8	8	8
14.	-	-	-	-r	-r	-r	-m	-a		á	f	4	4	4
15.	-	-	-	-s	-	-	-m	-a	já-ep	tré	n	5	4	4
16.	-	-	-	-s	-	-	-um	-a	i-eli	klæði, tæki	n	5	5	4
17.	-	-	-	-s	-ar	-a	-um	-a		bjór, ís, strætó, spör	m	6,5	5	5
18.	-	-	-	-s	-ir	-i	-um	-a		hver	m	6	6	6
19.	-	-	-	-s	-ir	-i	-um	-a	1x	Às	m	7	7	7
20.	-	-	-i	-	-ar	-a	-um	-a	1x	dans	m	6	6	6
21.	-	-	-i	-s	-	-	-um	-a		ber, kyn, borð, hús, högg, hreiður, höfuð	n	6,5	6,5	5

22.	-	-	-i	-s	-	-	-um	-a	1x	land, barn	n	6	6	6
23.	-	-	-i	-s	-	-	-um	-a	1x	meðal	n	6	6	6
24.	-	-	-i	-s	-ar	-a	-um	-a		fugl	m	6	6	6
25.	-	-	-i	-s	-ar	-a	-um	-a	1x, r-syn	akur, hamar	m	8	8	7
26.	-	-	-	-ur	-ur	-ur	-um	-a		vík	f	4	4	4
27.	-	-	-	-ur	-ur	-ur	-um	-a	2x	mörk	f	6	6	6
28.	-	-	-	-ur	-ur	-ur	-um	-a	1x	nótt	f	5	5	5
29.	-	-u	-u	-ar	-ar	-ar	-um	-a		drottning	f	5	5	5
30.	-a	-a	-a	-a	-ar	-a	-um	-a		herra	m	3	3	3
31.	-a	-a	-a	-a	-u	-u	-um	-a	n-ep	auga	n	4	3	3
32.	-a	-a	-a	-a	-u	-u	-um	-a	1x/n-ep	hjarta	n	5	4	4
33.	-a	-u	-u	-u	-ur	-ur	-um	-a	1x	fata, taska	f	5	5	5
34.	-a	-u	-u	-u	-ur	-ur	-um	-a		tunga, lilja	f	5,4	4	4
35.	-a	-u	-u	-u	-ur	-ur	-um	-a	1x/n-ep	saga	f	6	5	5
36.	-i	-a	-a	-a	-ar	-a	-um	-a		penni, tími, foringi	m	4	4	4
37.	-i	-a	-a	-a	-ar	-a	-um	-a	1x	hani, bakari	m	5	5	5
38.	-i	-a	-a	-a	-ir	-i	-um	-a	1x	Dani	m	5	5	5
39.	-i	-a	-a	-a	-ir	-i	-um	-a	j-ep	Tyrki	m	5	4	4
40.	-i	-a	-a	-a	-ur	-ur	-um	-a	1x	nemandi, bóndi	m	5	5	5
41.	-i	-i	-i	-i	-ar	-ar	-um	-a		lygi	f	4	4	4
42.	-i	-i	-i	-i	-ir	-ir	-um	-a		ævi, beiðni	f	4	4	4
43.	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	2x, r-syn	faðir	m	7	7	6
44.	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	móðir, dóttir	f	6	6	5
45.	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	bróðir	m	6	6	5
46.	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	r-syn	systir	f	5	4	4
47.	-l	-	-	-s	-ar	-a	-um	-a		hóll, makrill, stóll	m	6	6	6

48.	-l	-	-i	-s	-ar	-a	-um	-a	l-syn	lykill	m	8	8	7
49.	-l	-	-i	-s	-ar	-ar	-um	-a	2x, l-syn	ketill	m	10	10	9
50.	-n	-	-i	-s	-ar	-a	-um	-a	n-syn	himinn, steinn	m	8,7	8,7	7
51.	-r	-	-	-ar	-ir	-i	-um	-a	j-ep	bær	m	8	7	7
52.	-r	-	-	-r	-r	-r	-m	-a	1x	kýr	f	5	5	5
53.	-r	-	-	-s	-r	-	-m	-a		skór	m	5	5	5
54.	-r	-	-	-s	-ar	-a	-um	-a		snjór, kíkir, már, læknir	m	7,6	7,6	6
55.	-r	-	-	-s	-ar	-a	-um	-a	v-ep, 1x	týr	m	8	7	7
56.	-ur	-	-	-ar	-ir	-i	-um	-a	j-ep	lækur, leggur, litur	m	8,7	7	7
57.	-ur	-	-	-ar	-ir	-i	-um	-a	1x	staður	m	8	8	8
58.	-ur	-	-i	-ar	-ar	-a	-um	-a		skógur	m	6	6	6
59.	-ur	-	-i	-ar	-ir	-i	-um	-a		fundur	m	7	7	7
60.	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	mánaður	m	8	8	8
61.	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	háttur	m	8	8	8
62.	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	söfnuður	m	8	8	8
63.	-ur	-	-i	-ar	-ir	-i	-um	-a	1x	sonur	m	8	8	8
64.	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	fjörður	m	9	9	9
65.	-ur	-	-i	-ar	-ir	-i	-um	-a	2x	köttur, völlur	m	9	9	9
66.	-ur	-	-i	-ar	-ur	-ur	-um	-a	1x	fótur	m	7	7	7
67.	-ur	-	-	-s	-ar	-a	-um	-a	j/v-ep	niður, söngur	m	7	6	6
68.	-ur	-	-	-s	-ir	-i	-um	-a	1x	dalur	m	8	8	8
69.	-ur	-	-	-s	-ir	-i	-um	-a		smiður, belgur, leikur	m	8,7	7	7
70.	-ur	-	-i	-s	-ir	-i	-um	-a		gestur	m	7	7	7
71.	-ur	-	-i	-s	-	-	-um	-a	3x	maður	m	9	9	9
72.	-ur	-	-i	-s	-ar	-a	-um	-a		hestur	m	7	7	7
73.	-ur	-i	-i	-ar	-ar	-ar	-um	-a		hildur, ylgur	f	6,5	5	5
74.	-ur	-i	-i	-ar	-ir	-ir	-um	-a		bruður	f	6	6	6
75.	-ur	-ur	-i	-ar	-ur	-ur	-um	-a	r-syn	vetur	m	6	6	5
76.	-ur	-ur	-i	-s	-ur	-ur	-um	-a	r-syn	fingur	m	6	5	5

Table 12

Unification according to complexity c_4

r	Ns	As	Ds	Gs	Np	Ap	Dp	Gp	abl etc.	example	*g*	c_1	c_2	c_3	c_4
1	-	-	-	-ar	-	-	-um	-a	1x	mús	f	5	5	5	4
2	-	-	-	-ar	-ar	-ar	-m	-a		stó	f	4	4	4	4
3	-	-	-	-ar	-ar	-ar	-um	-a		nál, vél, stöð, skel, lifur, heiði, helgi	f	5,4	5,4	4	4
4	-	-	-	-ar	-ir	-ir	-um	-a		tíð, mynd, brún, höfn, öxl, verslun	f	6,5	6,5	6,5	5
5	-	-	-	-ar	-r	-r	-m	-a		frú, fló, brú	f	6,5	6,5	6,5	5
6	-	-	-	-ar	-ur	-ur	-um	-a		önd, nögl, bók, gát, kind, geit	f	5,6,7	5,6,7	5,6,7	5
7	-	-	-i	-ar	-ir	-i	-um	-a	2x	björn	m	8	8	8	6
8	-	-	-i	-ar	-ur	-ur	-um	-a	2x	hönd	f	8	8	8	6
9	-	-	-	-r	-r	-r	-m	-a		á	f	4	4	4	4
10	-	-	-	-s	-	-	-m	-a	já-ep	tré	n	5	4	4	4
11	-	-	-	-s	-	-	-um	-a	i-eli	klæði, tæki	n	5	5	4	4
12	-	-	-	-s	-ar	-a	-um	-a		bjór, ís, strætó, spör	m	6,5	5	5	5
13	-	-	-	-s	-ir	-i	-um	-a		hver, Às	m	7,6	7,6	7,6	6
14	-	-	-i	-	-ar	-a	-um	-a	1x	dans	m	6	6	6	5

15	-	-	-i	-s	-	-	-um	-a		ber, kyn, borð, hús, högg, hreiður, höfuð	n	6,5	6,5	6,5	5
16	-	-	-i	-s	-	-	-um	-a	1x	land, barn, meðal	n	6	6	6	5
17	-	-	-i	-s	-ar	-a	-um	-a		fugl, akur, hamar	m	8,6	8,6	7,6	6
18	-	-	-	-ur	-ur	-ur	-um	-a		vík, mörk, nótt	f	6,5,4	6,5,4	6,5,4	4
19	-	-u	-u	-ar	-ar	-ar	-um	-a		drottning	f	5	5	5	5
20	-a	-a	-a	-a	-ar	-a	-um	-a		herra	m	3	3	3	3
21	-a	-a	-a	-a	-u	-u	-um	-a		auga, hjarta	n	5,4	4,3	4,3	3
22	-a	-u	-u	-u	-ur	-ur	-um	-a	1x	fata, taska, tunga, lilja, saga	f	5,4	5,4	5,4	4
23	-i	-a	-a	-a	-ar	-a	-um	-a		penni, tími, foringi, hani, bakari	m	5,4	5,4	5,4	4
24	-i	-a	-a	-a	-ir	-i	-um	-a	1x	Dani, Tyrki	m	5	5,4	5,4	4
25	-i	-a	-a	-a	-ur	-ur	-um	-a	1x	nemandi, bóndi	m	5	5	5	4
26	-i	-i	-i	-i	-ar	-ar	-um	-a		lygi	f	4	4	4	4
27	-i	-i	-i	-i	-ir	-ir	-um	-a		ævi, beiðni	f	4	4	4	4
28	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a		faðir, bróðir	m	7,6	7,6	6,5	4
29	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	1x, r-syn	móðir, dóttir	f	6	6	5	4
39	-ir	-ur	-ur	-ur	-ur	-ur	-um	-a	r-syn	systir	f	5	4	4	4
31	-l	-	-	-s	-ar	-a	-um	-a		hóll, makrill, stóll	m	6	6	6	6

32	-l	-	-i	-s	-ar	-a	-um	-a	l-syn	lykill	m	8	8	7	7
33	-l	-	-i	-s	-ar	-ar	-um	-a	2x, l-syn	ketill	m	10	10	9	7
34	-n	-	-i	-s	-ar	-a	-um	-a	n-syn	himinn, steinn	m	8,7	8,7	7	7
35	-r	-	-	-ar	-ir	-i	-um	-a	j-ep	bær	m	8	7	7	7
36	-r	-	-	-r	-r	-r	-m	-a	1x	kýr	f	5	5	5	4
37	-r	-	-	-s	-r	-	-m	-a		skór	m	5	5	5	5
38	-r	-	-	-s	-ar	-a	-um	-a		snjór, kíkir, már, læknir	m	7,6	7,6	6	6
39	-r	-	-	-s	-ar	-a	-um	-a	v-ep, 1x	týr	m	8	7	7	6
40	-ur	-	-	-ar	-ir	-i	-um	-a	j-ep	lækur, leggur, litur, staður	m	8,7	8,7	8,7	7
41	-ur	-	-i	-ar	-ar	-a	-um	-a		skógur	m	6	6	6	6
42	-ur	-	-i	-ar	-ir	-i	-um	-a		fundur, fatnaður, mánaður, háttur, söfnuður,sonur, fjörður, köttur, völlur	m	9,8,7	9,8,7	9,8,7	7
43	-ur	-	-i	-ar	-ur	-ur	-um	-a	1x	fótur	m	7	7	7	6
44	-ur	-	-	-s	-ar	-a	-um	-a	j/v-ep	niður, söngur	m	7	6	6	6
45	-ur	-	-	-s	-ir	-i	-um	-a		smiður, belgur, leikur, dalur	m	8,7	8,7	8,7	7
46	-ur	-	-i	-s	-ir	-i	-um	-a		gestur	m	7	7	7	7
47	-ur	-	-i	-s	-	-	-um	-a	3x	maður	m	9	9	9	6

48	-ur	-	-i	-s	-ar	-a	-um	-a		hestur	m	7	7	7	7
49	-ur	-i	-i	-ar	-ar	-ar	-um	-a		hildur, ylgur	f	6,5	5	5	5
50	-ur	-i	-i	-ar	-ir	-ir	-um	-a		bruður	f	6	6	6	6
51	-ur	-ur	-i	-ar	-ur	-ur	-um	-a	r-syn	vetur	m	6	6	5	5
52	-ur	-ur	-i	-s	-ur	-ur	-um	-a	r-syn	fingur	m	6	5	5	5

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 155-175.*

Efficiency of Flexible Parts-of-Speech Systems

Relja Vulanović

1. INTRODUCTION

In this paper, I consider the classification of parts-of-speech (PoS) systems proposed in Hengeveld (1992) and further discussed in Hengeveld, Rijkhoff, & Siewerska (2004) – HRS from this point on, Hengeveld & Rijkhoff (2005), and Rijkhoff (2007). Parts of speech are approached in these papers from the functional point of view – they are defined according to the propositional functions they can have, that is, according to the syntactic slots they can occupy. Four propositional functions are considered: P – the head of the predicate phrase (PredPh), R – the head of the referential (noun) phrase (RefPh), p – the modifier of PredPh, and r – the modifier of RefPh. Verbs, nouns, adjectives, and manner adverbs are the specialized lexeme classes for these four propositional functions. The class of verbs consists of lexemes that can only be used as the head of PredPh. Nouns are lexemes that can function as the head of RefPh. Adjectives are used as modifiers of RefPh and manner adverbs as $_{\text{modifiers}}$ of PredPh. Manner adverbs are the only kind of adverbs considered since other kinds of adverbs often modify larger units within the sentence and not just the head of PredPh. This approach to lexeme classes can be applied to any other lexeme class which is then defined based on how it can normally function without using any $_{\text{special}}$ grammatical devices. The HRS sample of 50 languages shows that there is a great variation among languages when lexeme classes are viewed in this functional way.

Consider, for instance, Samoan and Warao, two languages from the HRS sample. All examples below are taken from HRS.

Samoan (Mosel & Hovdhaugen, 1992: 80, 80, 305, 394)

(a) ‘Ua mālosi le **lā**.
PERFECT strong ARTICLE sun
‘The sun is strong.’ (‘The sun strongs.’)

(b) ‘Ua **lā** le aso.
PERFECT sun ARTICLE day
‘The day is sunny.’ (‘The day suns.’)

(c) le fale **ta‘avale**
ARTICLE house car
‘the garage’ (‘the car house’)

(d) ‘Ua ma‘i **misela** le tama.
PERFECT sick measles ARTICLE boy
‘The boy has got the measles.’ (‘The boy sicks measelwise.’)

The four Samoan examples show that translational equivalents of English nouns can have any of the four propositional functions: *lā* in (a) functions as R, *lā* in (b) as P, *ta‘avale* in (c) as r, and *misela* in (d) as p. These are not just some special cases – the characteristic of the whole Samoan PoS system is that almost any word can be used to indicate any propositional function. This kind of large open class of lexemes is called in HRS the class of *contentives*. A contentive is a lexeme which can function as P, R, r, and p.

Warao (Romero-Figeroa, 1997: 49f, 49f, 119)

(a) **yakera**
beauty
‘beauty’

(b) Hiaka **yakera** auka saba tai nisa-n-a-e.
garment beauty daughter for she buy-SINGULAR-PUNCTUAL-PAST
‘She bought a beautiful dress for her daughter.’

(c) Oko kuana yaota-te arone **yakera** nahoro-te...
we hardness work-NON-PAST although beauty eat-NON-PAST
‘Although we work hard and eat well...’

The word *yakera* in these three Warao examples is used with three propositional functions: R in (a), r in (b), and p in (c). However, it cannot function as P. This illustrates that Warao has two lexeme classes, verbs, which are only used for P, and non-verbs, used for the remaining three propositional functions.

PoS systems of both Samoan and Warao are flexible in the sense that one lexeme class can have more than one propositional function. The single lexeme class of contentives in Samoan has to be used with more flexibility than the two lexeme classes in Warao. Ngiti (Kutsch Lojenga 1993) would be an example of a language belonging to the next type on the scale of decreasing flexibility of lexeme classes. It has verbs, nouns, and a lexeme class of modifiers which function as r and p. The scale ends in *rigid* languages, like English, in which there is a different lexeme class

for each propositional function. The PoS systems of the four languages mentioned above are shown in Table 1. Each of the four systems illustrates one of the types in the HRS classification. Note that the type number is equal to the number of lexeme classes.

Table 1
Partial typology of PoS systems based on what lexeme classes have what propositional functions (C = contentive, V = verb, Λ = non-verb, N = noun, M = modifier, a = adjective, m = manner adverb).

Language	PoS system type	P	R	r	p
Samoan	1	C			
Warao	2	V	Λ		
Ngiti	3	V	N	M	
English	4	V	N	a	m

The full HRS classification contains three more rigid PoS system types. These, type 5-7, systems have fewer than four propositional functions and are not discussed in this paper. Although type 4 is rigid, I will refer to it as *differentiated* (the term used in HRS) in order to distinguish it from the other rigid system types.
The seven HRS PoS system types mentioned thus far constitute the basic types of the classification, but there are also six intermediate types.

Table 2
Basic and intermediate PoS system types 1-4

Language	PoS system type	P	R	r	p
Samoan	1	C			
Mundari	1.5	C			
			Λ		
Warao	2	V	Λ		
Turkish	2.5	V	Λ		
				M	
Ngiti	3	V	N	M	
Lango	3.5	V	N	M	
					m
English	4	V	N	a	m

There is one intermediate PoS system type between each pair of consecutive basic types. The three flexible intermediate types are shown in Table 2 which is an extension of Table 1. I conveniently denote those intermediate types as 1.5, 2.5, and 3.5 (they are labeled 1/2, 2/3, and 3/4 in the HRS notation). The HRS sample contains one example of each of these intermediate types. The PoS system of Mundari is classified as 1.5, that is, between types 1 and 2 (see also Hengeveld & Rijkhoff (2005) for a detailed discussion of this language). Turkish is a type 2.5 language and Lango – type 3.5. An intermediate PoS system has lexeme classes that are compatible with both basic PoS types that the intermediate system is in between. For instance, Lango has a large open class of lexemes that can have both modifier functions, but also a large open class of manner adverbs. Note in Table 2 the particular way the lexeme classes of the intermediate types are used for the four propositional functions.

Because they have fewer lexeme classes than propositional functions, the PoS systems of type 1-3 are prone to ambiguity. Some ambiguity is even possible in the intermediate type 3.5 because of the way the four lexeme classes are used. In such PoS systems, there is a greater need to mark propositional functions syntactically or morphologically. This is analyzed in HRS through examination of the linguistic sample used there. A formal combinatorial analysis (Vulanović, 2008a) reveals that the flexible languages in the sample tolerate some amount of functional ambiguity. Syntactic or morphological markers are sometimes missing and even if they exist, they are not always used in such a way that ambiguity gets eliminated. Obviously, flexible languages do not use the available grammatical markers in the most efficient way. While the nature of this result in Vulanović (2008a) is only qualitative, the present paper provides its quantitative version. This is done by calculating grammar efficiency (Vulanović, 1991, 1993, 2003, 2007) for most of the flexible languages in the HRS sample. In order to enable further comparisons, some differentiated languages from the sample are also included, as well as a few hypothetical structures.

All this requires several levels of abstraction. First of all, the classification of PoS systems is itself an abstraction. The cut between PoS system types, as already witnessed by the intermediate types, is not so clear. To quote Rijkhoff (2007: 718), "the various types of PoS systems [...] should be regarded as reference points on a scale rather than distinct categories. Because languages are dynamic entities, they can only approximate the ideal types in this classification." Secondly, the analysis is kept on the level of simple intransitive sentences which are described formally in order to represent the PoS system and other grammatical features of each language considered. Four kinds of sentences are modeled together: those having both heads and no modifier, then those with both heads and just one modifier (r or p), and finally those with all four propositional functions. The formalization of linguistic structures is done in such a way that it is possible to apply the formula from

Vulanović (2007) for calculating grammar efficiency. Two new variations of this formula are also used. They are introduced in order to define two other types of grammar efficiency which are of interest in this paper. The calculation of grammar efficiency is the last step of the formal approach undertaken here. When this is done with all languages in the sample, a list of grammar-efficiency values is obtained. The numbers show that grammar efficiency of natural languages is much smaller than what is theoretically possible. This is so because, on the one hand, word order is not as free as possible, and, on the other hand, lexeme classes and grammatical markers are not assigned to the propositional functions in the most efficient way. The effect of these two factors is shown separately by the two new versions of grammar efficiency. The quantification carried out in this paper enables further analyses. One of them is presented here, viz. the correlation analysis between the PoS system type and the grammar-efficiency value.

Warao is used in section 2 to illustrate how the linguistic structures are represented formally. This is followed by a description of grammar efficiency in sections 3 and 4. The steps for calculating grammar efficiency are illustrated by providing details for Warao again. In section 5, the example of Samoan is presented in order to show the whole procedure once more. The results for the whole sample are given in section 6. They are analyzed in section 7, where some conclusions are made.

2. THE FORMAL REPRESENTATION OF LINGUISTIC STRUCTURES

As already mentioned in the introduction, only simple intransitive sentences are considered when describing linguistic structures. Since the heads are obligatory and the modifiers optional, there are 38 possible orders of the four propositional functions. They can be described as

$$\{PR, RP\} \cup P\{P, R, p\} \cup P\{P, R, r\} \cup P\{P, R, p, r\},$$

where P denotes the permutation operator – it produces all permutations of the set it is applied to. The count of 38 comes from $2 + 3! + 3! + 4! = 38$. However, discontinuous phrases, like in PRp or RprP, are not going to be considered in this paper. This leaves the following 18 possible orders:

PR, RP, PpR, pPR, RPp, RpP, PRr, PrR, RrP, rRP, (1a)

PpRr, PpRr, pPRr, pPrR, RrPp, RrpP, rRPp, rRpP. (1b)

Most of the languages in the sample have a fixed word order, so only a few of the above 18 orders appear in each language. For instance, Warao has a fixed basic word order which corresponds to RrpP. Therefore, only 4 orders of propositional functions are possible in Warao:

$$\text{RP, RrP, RpP, RrpP.} \tag{2}$$

From now on, I will use the phrase *word order* to mean also the order of propositional functions. This is because the two orders are closely related.

The structure of Warao is presented here following the description in HRS. The original source is Romero-Figeroa (1997). Warao has a postposition *tane* marking the manner adverb. There are no other markers relevant for the present discussion. If this marker is indicated symbolically as a '+', we can say that Warao uses three g*rammatical conveyors* (two lexeme classes, V and Λ, and a lexeme-class-marker combination, Λ+) to perform the four propositional functions. In this paper, a grammatical conveyor is a lexeme class, a grammatical marker, or a combination of the two, in other words, anything that is used in a language to indicate the propositional functions. If the number of grammatical conveyors is denoted by k and the number of lexeme classes by l, then $k \geq l$. In Warao, $k = 3$ and $l = 2$.

The following mapping, denoted by Φ, between grammatical conveyors and propositional functions exists in Warao:

$$\Phi: \quad \text{V} \rightarrow \text{P}, \quad \Lambda \rightarrow \text{R, r}, \quad \Lambda\text{+} \rightarrow \text{p}. \tag{3}$$

This simply shows what conveyors have what propositional functions. Because of the orders in (2) and the mapping in (3), there are four possible intransitive sentences in Warao, represented formally in (4),

$$\Lambda\text{V}, \quad \Lambda\Lambda\text{V}, \quad \Lambda\Lambda\text{+V}, \quad \Lambda\Lambda\Lambda\text{+V}. \tag{4}$$

Note that here lexeme-class symbols also denote members of each respective class. Thus, ΛV means a sentence consisting of non-verb followed by a verb. Again because of (2) and (3), the sentences in (4) can be parsed as follows:

$$\Lambda\text{V} \rightarrow \text{RP}, \quad \Lambda\Lambda\text{V} \rightarrow \text{RrP}, \quad \Lambda\Lambda\text{+V} \rightarrow \text{RpP}, \quad \Lambda\Lambda\Lambda\text{+V} \rightarrow \text{RrpP}. \tag{5}$$

We can see that no sentence has ambiguous interpretation. If ρ indicates the number of successful parses of all permitted sentences and ρ_0 is the number of ambiguous parses, then, according to (5), in Warao, $\rho = 4$ and $\rho_0 = 0$. These values are important for the discussion of grammar efficiency.

A natural question to ask is how efficient this described linguistic structure is. We can see that the Warao word order is more rigid than what the rule in (3) can permit without creating ambiguity. Sentence VΛ, for instance, would be analyzed as PR without any ambiguity. What is then the greatest possible number of unambiguous parses for the given assignments of the grammatical conveyors, that is, for the given mapping Φ? Let this number be denoted by ρ_A. In the case of assignments like in (3), $\rho_A = 12$ because the following parses can be considered:

ΛV → RP, VΛ → PR, ΛΛV → RrP|rRP, VΛΛ → PRr|PrR, (6a)

ΛΛ+V → RpP, Λ+VΛ → pPR, ΛVΛ+ → RPp, VΛ+Λ → PpR, (6b)

ΛΛΛ+V → RrpP|rRpP, Λ+VΛΛ → pPRr|pPrR, (6c)

ΛΛVΛ+ → RrPp|rRPp, VΛ+ΛΛ → PpRr|PprR. (6d)

The symbol '|' is used above in the sense of an exclusive 'or' – one of the two possible parses should be chosen. The choice is always about the relative order of R and r since P and p are unambiguously identified, the former because of the specialized lexeme class (V), and the latter because of the marker. When finding ρ_A, note that we should only consider such permutations of the grammatical conveyors, which can be interpreted as at least one of the 18 parses listed in (1). For instance, the permutation ΛVΛ is not included in (6) – it cannot be parsed successfully because the continuity of RefPh is assumed. When forming the above permutations, the order of the marker and the word it marks is kept fixed.

However, there are more parsing attempts than the successful parses in (6). Assume that parsing is done from left to right, one word at a time, and the number of words is not known in advance. Under these assumptions, there are three parsing attempts to parse ΛV, for instance. One of the attempts can be finished successfully, giving RP like in the first parse in (6a), but Λ can be initially also interpreted as either r or p. These two parsing attempts are abandoned when it is realized that the second word is neither Λ nor +. The parsing attempts, successful or not, are represented in this case as follows:

ΛV → RP/r-/p-,

where the symbol '-' indicates that the parse is abandoned. The order of parsing attempts is unimportant here. If one attempt is finished successfully, the remaining ones should still be done because some of them may be also successful, giving rise

to ambiguity. Therefore, all parsing attempts are always considered. For the remaining sentences in (6), they are listed below:

$V\Lambda \rightarrow$ PR/Pr-/Pp-, $\Lambda\Lambda V \rightarrow$ RrP/Rp-/rRP/p-, $V\Lambda\Lambda \rightarrow$ PRr/PrR/Pp-,

$\Lambda\Lambda+V \rightarrow$ Rr-/RpP/rR-/p-, $\Lambda+V\Lambda \rightarrow$ R-/r-/pPR/pPr-,

$\Lambda V\Lambda+ \rightarrow$ RPp/r-/p-, $V\Lambda+\Lambda \rightarrow$ PR-/Pr-/PpR/Ppr-,

$\Lambda\Lambda\Lambda+V \rightarrow$ RrpP/Rp-/rRpP/p-, $\Lambda+V\Lambda\Lambda \rightarrow$ R-/r-/pPRr/pPrR,

$\Lambda\Lambda V\Lambda+ \rightarrow$ RrPp/Rp-/rRPp/p-, $V\Lambda+\Lambda\Lambda \rightarrow$ PR-/Pr-/PpRr/PprR.

The total number of attempted pareses is 44. This count is denoted by ρ*. Thus, ρ* represents the number of parsing attempts, of which at least one is successful, applied to all permutations of all possible sentences. The more lexeme classes, the fewer parsing attempts. For instance, in a type 4 language, there is just one parsing attempt of sentence aNV and it is successful (aNV → rRP). In this sense, the value of ρ* depends on how complicated mapping Φ is (how much it is different from a one-to-one mapping). Smaller values of ρ* indicate more efficient structures, all other things being equal.

If there are three grammatical conveyors, (3) is not necessarily the most efficient way of assigning them to propositional functions. There may be some other mappings which will give less than 44 parsing attempts for the value of ρ*. This issue is addressed in general terms in the next section.

In section 6, the structure of intransitive sentences in each language of the sample is described using the above formalism and the following template exemplified on Warao.

Type 2: Warao

Basic order: fixed RrpP

$k = 3$ ($V \rightarrow P$, $\Lambda \rightarrow R, r$, $\Lambda+ \rightarrow p$)

$\Lambda V \rightarrow RP$, $\Lambda\Lambda V \rightarrow RrP$, $\Lambda\Lambda+V \rightarrow RpP$, $\Lambda\Lambda\Lambda+V \rightarrow RrpP$

$\rho = 4$, $\rho_0 = 0$, $\rho_A = 12$, $\rho^* = 44$, $Q_{WO} = 1/6$

The third line of the template shows the permitted sentences and how they are parsed. There is one quantity in the fourth line, Q_{WO}, which still needs to be explained. This is done in section 4.

3. GRAMMAR EFFICIENCY

The concept of grammar efficiency is introduced in Vulanović (1991) and further developed in Vulanović (1993, 2003, 2007). The latest version of the grammar efficiency formula is used in the present paper. This formula is a slight modification of the one presented in Vulanović (2003). Grammar efficiency, *Eff*, is defined as follows:

$$Eff = \gamma\, Q\, n/k. \tag{7}$$

In this formula, γ is a scaling coefficient ensuring that $Eff = 1$ for maximally efficient grammars. The quantity n is a measure of the amount of linguistic information that needs to be conveyed and k is the number of grammatical conveyors, as already defined in section 2. If the grammar conveys more information with fewer conveyors, it is more efficient. This is captured by formula (7). In this paper, n is the number of propositional functions (syntactic slots) and is therefore kept fixed, $n = 4$.

The quantity Q is here referred to as the *parsing ratio* and is defined by

$$Q = (\rho - \rho_0)/\rho^*, \tag{8}$$

with ρ, ρ_0, and ρ^* as in the previous section. The following points motivate the formula for Q:

- The grammar is more efficient if it has fewer word order rules, that is, if ρ is greater.
- The grammar is more efficient if it permits fewer ambiguous sentences, that is, if ρ_0 is less.
- The grammar is more efficient if its permitted sentences require fewer parsing attempts, that is, if ρ^* is less.

The only essential difference between the definition of *Eff* in Vulanović (2007), which is used here, and that in Vulanović (2003) is in ρ^*. The value of ρ^* in Vulanović (2003) is easier to calculate, but is less closely related to parsing.

Returning to the example of Warao, we can see from (8) that

$$Q = (4 - 0)/44 = 1/11, \tag{9}$$

thus the coefficient γ is the only missing component in the efficiency formula (7). This coefficient is related to maximally efficient grammars. A maximally efficient grammar has to satisfy certain conditions, see Vulanović (2003) for details. The

most important ones are that it cannot permit ambiguous sentences and that it has the greatest value of the parsing ratio within the class of grammars Γ_k with the given number of grammatical conveyors k (recall that n is not a variable here). For a given value of k, what can be varied within class Γ_k is the assignment rule Φ and word order rules. All possible grammars in Γ_k, which create no ambiguity, should be explored and the greatest parsing ratio, denoted by Q_k, should be found. Theoretically it may happen that Q_k does not exist, but this will not be an issue in the present paper. Then the coefficient γ can be expressed as

$$\gamma = k/(n\, Q_k). \tag{10}$$

Equations (10) and (7) guarantee that for maximally efficient grammars $Eff = 1$. On the other hand, when (10) is used in (7), it follows that

$$Eff = Q/Q_k, \tag{11}$$

which, form this point on, is the formula to be used instead of (7). Therefore, grammar efficiency in (11) means a measure of the parsing ratio relative to the greatest possible parsing ratio within the class of grammars with the same k.

In order to find efficiency of the Warao structure, we have to find Q_3. Some details of how this is done are given in the next section, where it is shown that $Q_3 = 5/8$. Therefore, using (9) and (11), we get for Warao:

$$Eff = (1/11)/(5/8) = 8/55.$$

Two other types of efficiency are also of interest. They are used here for the first time. One of them is the *efficiency for the given assignments of grammatical conveyors*, denoted by Eff_{A}. Assume that the mapping Φ, which shows how grammatical conveyors are assigned to propositional functions, is fixed. Word order is then the only thing that can be varied within the set of grammars in Γ_k which have the same mapping Φ. This gives $Q_{\mathrm{A}} = \rho_{\mathrm{A}}/\rho^*$ for the greatest possible parsing ratio among such grammars. The same way (11) is interpreted, so here, Eff_{A} is the parsing ratio measured relatively to Q_{A},

$$Eff_{\mathrm{A}} = Q/Q_{\mathrm{A}} = (\rho - \rho_0)/\rho_{\mathrm{A}}. \tag{12}$$

If $Eff_A = 1$, word order is as free as possible for the given assignments of grammatical conveyors. Using the values found for Warao, we get that for this language

$$Eff_{\mathrm{A}} = 4/12 = 1/3,$$

which is greater than the overall efficiency *Eff*.

The other efficiency is the *efficiency for the given word order*,

$$Eff_{\mathrm{WO}} = Q/Q_{\mathrm{WO}}, \tag{13}$$

where Q_{WO} is the greatest possible parsing ratio for the given word order, obtained by changing assignments of grammatical conveyors. To find Q_{WO}, all grammars with the same k and the same fixed word order, but with different mappings Φ, should be considered. If $Eff_{WO} = 1$, the assignment mapping used is the most efficient one for the given word order. It is shown in the following section that in Warao $Q_{\mathrm{WO}} = 1/6$, as already reported in the Warao template at the end of section 2. This value and (9), used in (13), give that for Warao

$$Eff_{\mathrm{WO}} = (1/11)/(1/6) = 6/11.$$

4. THE GREATEST PARSING RATIOS

The definition of maximally efficient grammars assumes that those grammars cannot permit ambiguous sentences. This is impossible to achieve with just one grammatical conveyor, thus the parsing ratio Q_1, in the sense of its supposed use in formulas (10) and (11), does not exist. However, Q_1 is not needed in the analysis of the HRS sample because this sample does not contain any type 1 language without markers. In other words, in the languages of the sample, $k \geq 2$.

The greatest parsing ratios can be found in the remaining cases $k = 2, 3, 4$. The results are

$$Q_2 = 4/11, \quad Q_3 = 5/8, \quad \text{and} \quad Q_4 = 1.$$

More details on calculations are provided below. The case $k = 3$ is considered first because this continues the discussion of the Warao example.

$k = 3$.

All possible ways of assigning 3 grammatical conveyors should be investigated and the parsing ratio should be calculated for each possibility. Word order should be maximally free. Therefore, $\rho = \rho_{\mathrm{A}}$ in every case considered and all the listed values of the parsing ratio are in fact equal to the corresponding Q_{A}. The greatest parsing ratio resulting from this is the desired Q_3. There are 6 possible

assignment mappings Φ. They are listed below together with the values of their corresponding parsing ratios. The first one is like in type 3 PoS systems, see Table 1.

1) Type 3: V → P, N → R, M → r, p, Q = 16/28 = 4/7

2) V → P, Λ → R, r, m → p, Q = 12/24 = 1/2

3) V → P, Λ → R, p, a → r, Q = 17/30

4) C → P, p, N → R, a → r, Q = 12/24 = 1/2

5) C → P, R, a → r, m → p, Q = 15/24 = 5/8

6) C → P, r, N → R, m → p, Q = 17/30

C and Λ are used above in a more general sense than in Table 1. C indicates any lexeme class which has P and at least one more propositional function. Similarly, Λ denotes any lexeme class functioning as R and one or both modifiers.

Mapping Φ in case 2) is like in (3), but m and Λ+ are parsed differently. Whenever a grammatical conveyor consists of a lexeme class combined with a marker, this causes more parsing attempts than a fully distinguishable lexeme class. This is why markers are not considered in the above cases. Cases 2) and 4) are analogous to each other. In both of them, one head is combined with the modifier of the same phrase. Similarly, cases 3) and 6) are also mathematically equivalent. It is interesting to mention that some natural languages have recently been reported (Hengeveld & van Lier, to appear) to fall outside the classification presented in Table 1; one of them is Hungarian which behaves like in 2).

It can be concluded from cases 1) – 6) that Q_3 = 5/8, which is the parsing ratio obtained in case 5). Some details for this case follow. Calculations of parsing ratios in other cases proceed along the same lines.

In case 5), there are two successful parses of sentence CC,

CC → PR/RP,

and 10 attempted parses of three-word sentences (keep in mind that both PredPh and RefPh have to be continuous),

CCa → PRr/RP-, CaC → PrR/RrP, aCC → rRP,

CCm → PR-/RPp, CmC → PpR/RpP, mCC → pPR.

Further, there are 12 attempted parses of four-word sentences,

CaCm → PrR-/RrPp, CamC → Pr-/RrpP,

CmCa → PpRr/RpP-, CmaC → PprR/Rp-,

aCCm → rRPp, aCmC → rRpP, mCCa → pPRr, mCaC → pPrR.

Therefore, ρ* = 24 in this case. The total number of intransitive sentences, which require these 24 parsing attempts, is 15, which is at the same time the count for ρ_A. This is because whenever there are several successful pareses of a sentence, a word order rule can be imposed to permit only one of those successful parses and thus to eliminate ambiguity. Then it follows from (8) that $Q = 15/24 = 5/8$ for this structure.

The smallest value of ρ* listed in cases 1) – 6) is 24. This enables us to determine the value of Q_{WO} for Warao. Since Warao permits 4 orders of propositional functions (RP, RrP, RpP, RrpP), it follows that $Q_{WO} = 4/24 = 1/6$.

$k = 2$.

In this case, 7 possible mappings Φ should be explored. Only the final results are listed below.

a) C → P, R, Λ → r, p, $Q = 8/22 = 4/11$

b) C → P, p, Λ → R, r, $Q = 8/28 = 2/7$

c) C → P, r, Λ → R, p, $Q = 12/34 = 6/17$

d) Type 2: V → P, Λ → R, r, p,
and the analogous N → R, C → P, r, p, $Q = 9/34$

e) a → r, C → P, R, p, and m → p, C → P, R, r, $Q = 9/37$

The greatest parsing ratio is therefore $Q_2 = 4/11$.

It is interesting to observe that the type 2 and 3 languages do not give the greatest parsing ratios for $k = 2$ and $k = 3$, respectively. In both classes of structures, the greatest value of Q is obtained when the same lexeme class functions as both heads (C → P, R). This is discussed further in Vulanović (2008b). The value $Q = 4/7$ for type 3 languages is the greatest parsing ratio among those languages with $k = 3$ which have the class of verbs assigned to P. And, in the class with $k = 2$, type 2

languages are the only ones that have V → P. In other words, if the assignment V → P is assumed obligatory, then the largest possible *Q* values in the type 2 and 3 languages are the greatest parsing ratios for their respective classes.

$k = 4$.
This is the differentiated PoS system, which is included here so that the flexible PoS systems can be compared to it. With 4 lexeme classes, each lexeme class has exactly one propositional function. This is why $\rho_A = \rho^* = 18$ (the 18 orders in (1)) and $Q_4 = 1$.

5. ANOTHER EXAMPLE

In order to once more illustrate the procedure for finding all the relevant quantities, details for Samoan are provided in this section. This also exemplifies the case when a marker is used to mark not a lexeme class but the whole referential phrase. Again, I use here the description from HRS which is itself based on Mosel & Hovdhaugen (1992).

Samoan is a type 1 language with two possible orders, PpRr and RrPp. This gives 8 possible word orders,

PR, PpR, PRr, PpRr, RP, RrP, RPp, RrPp.

In Samoan, the PpRr order is unmarked, whereas the RrPp order requires a marker, *o*, in front of RefPh. Therefore, there are two grammatical conveyors ($k = 2$), C and its marked form ‘C. The following mapping is used:

C → P, R, r, p, ‘C → R, r if no preceding C is interpreted as R or r. (14)

There are 6 intransitive sentences with 8 possible parses in all,

CC → PR, ‘CC → RP, CCC → PpR/PRr, ‘CCC → RPp/RrP, (15a)

CCCC → PpRr, ‘CCCC → RrPp, (15b)

which means that $\rho = 8$. PpR/PRr and RPp/RrP form two pairs of ambiguous parses, so the number of ambiguous parses is $\rho_0 = 4$.

To find ρ^*, we have to consider all permutations of the 6 sentences in (15), but to exclude those permutations which cannot be parsed successfully. Starting from sentence CC, we can see that it requires 8 parsing attempts,

CC → PR/Pp-/Pr-/RP/Rr-/Rp-/rR-/pP-.

This is so because of the mapping in (14) and the continuity of both PredPh and RefPh. Similarly,

‘CC → RP/Rr-/Rp-/rR- and C‘C → PR/R-/r-/p-.

There are therefore 16 attempted parses of two-word sentences. Three-word sentences require 40 parsing attempts:

CCC → PRr/PrR/PpR/Ppr-/RPp/RrP/Rrp-/RpP/rRP/rRp-/pPR/pPr-,

‘CCC → RPp/RrP/RpP/rRP/rRp-, C‘CC → PRr/PrR/R-/r-/p-,

CC‘C → PR-/Pr-/PpR/RP-/Rr-/Rp-/rR-/pP-.

Finally, there are 28 attempted parses of four-word sentences:

CCCC → PR-/PrR-/PpRr/PprR/RPp-/RrPp/RrpP/RpP-/
rRPp/rRpP/pPRr/pPrR,

‘CCCC → RPp-/RrPp/RrpP/RpP-/rRPp/rRpP,
CC‘CC → PR-/Pr-/PpRr/PprR/RP-/Rr-/Rp-/rR-/pPRr/pPrR.

Thus, the total count is ρ* = 16 + 40 + 28 = 74. Formula (8) then gives

$Q = (8 - 4)/74 = 2/37$

for Samoan.

Ten sentences had to be considered above in order to find ρ*, which means that $\rho_A = 10$. Then, formula (12) gives

$Eff_A = (8 - 4)/10 = 2/5$.

In order to determine Q_{WO} for Samoan, for which $k = 2$, we have to refer to mappings a) – e) in the previous section. The mapping under a) gives the smallest value of ρ* (ρ* = 22), but it also produces many ambiguous sentences in spite of the fact that word order is restricted to PpRr and RrPp. For instance, CC → PR/RP, or CΛC → PpR/RrP. This is why mapping a) does not give the greatest parsing ratio for the Samoan word order. However, mapping b) creates no ambiguous sentences

and provides the greatest possible parsing ratio with the Samoan word order, Q_{WO} = 8/28 = 2/7. Therefore, Samoan can be described using the following template:

Type 1: Samoan
Basic order: PpRr; variation: RrPp with RefPh marked
k = 2 (C, RefPh marker ‘)
CC → PR, ‘CC → RP, CCC → PpR/PRr, ‘CCC → RPp/RrP,
CCCC → PpRr, ‘CCCC → RrPp
ρ = 8, ρ_0 = 4, ρ_A = 10, ρ* = 74, Q_{WO} = 2/7

6. RESULTS FOR THE LANGUAGE SAMPLE

Three flexible languages from the HRS sample are excluded: Tagalog, because it has no p (and therefore, n = 3, not 4); Ngiti, because it has a discontinuous PredPh; and Lango, because it has 6 grammatical conveyors. Those languages can be discussed analogously to the rest of the sample, but their inclusion would complicate the exposition. Instead of Lango, a hypothetical intermediate 3.5 type language with 4 lexeme classes is included in the sample. Another hypothetical structure considered is a type 1 PoS system similar to Samoan. It shows a more efficient way of using one grammatical marker. Five type 4 languages are also included for comparison – those from the HRS sample which have a fixed order of PredPh and RefPh and no markers (so that k is kept equal to 4). Finally, Hurrian, one of the languages in the HRS sample, is represented here in two versions. Hurrian is described in HRS as a language without markers and this is one version modeled here. The other version, which I call Hurrian*, has a RefPh marker. This model is motivated by the presence of a morpheme *-n* in some texts. This morpheme “mediate[s] between subject and predicate linking the one with the other into a unified utterance’’ (Speiser, 1941: 172). This brings the total number of languages in the sample to 17. Warao and Samoan have been described above in more details. The same procedure is used for the remaining languages. Their structures are described here based on how they are presented in Table 3 in HRS. The original source for each language can be found in HRS. The templates containing the final results are listed below.

Type 1: Hypothetical (like Samoan, but with fixed WO and RefPh always marked)
Basic order: fixed PpRr
k = 2 (C, RefPh marker ‘)
C‘C → PR, C‘CC → PRr, CC‘C → PpR, CC‘CC → PpRr
ρ = 4, ρ_0 = 0, ρ_A = 7, ρ* = 29, Q_{WO} = 2/11

Type 1.5: Mundari
Basic order: fixed rRpP
k = 2 (C, Λ)
CC → RP, ΛC → RP
XXC → RpP/rRP, XXXC → rRpP, where X = C or X = Λ
ρ = 18, ρ_0 = 8, ρ_A = 25, ρ* = 199, Q_{WO} = 2/11

Type 2: Hurrian
Basic order: fixed rRpP
k = 2 (Λ, V)
ΛV → RP, ΛΛV → RpP/rRP, ΛΛΛV → rRpP
ρ = 4, ρ_0 = 2, ρ_A = 9, ρ* = 34, Q_{WO} = 2/11

Type 2: Hurrian*
Basic order: fixed rRpP
k = 3 (Λ, V, RefPh marker ’)
Λ’V → RP, ΛΛ’V → rRP, Λ’ΛV → RpP, ΛΛ’ΛV → rRpP
ρ = 4, ρ_0 = 0, ρ_A = 11, ρ* = 36, Q_{WO} = 1/6

Type 2: Imbabura Quechua
Basic order: rRpP; variation: pPrR with RefPh marked
k = 3 (Λ, V, RefPh marker ’)
ΛV → RP, VΛ’ → PR, ΛΛV → RpP/rRP, ΛVΛ’ → pPR,
VΛΛ’ → PrR, ΛΛΛV → rRpP, ΛVΛΛ’ → pPrR
ρ = 8, ρ_0 = 2, ρ_A = 20, ρ* = 59, Q_{WO} = 1/3

Type 2.5: Turkish
Basic order: fixed rRpP
k = 3 (Λ, V, M)
ΛV → RP, ΛΛV → rRP/RpP, ΛMV → RpP,
ΛΛΛV/ΛΛMV/MΛΛV/MΛMV → rRpP
ρ = 8, ρ_0 = 2, ρ_A = 35, ρ* = 112, Q_{WO} = 1/6

Type 3: Ket
Basic WO: fixed rRpP
k = 3 (N, V, M)
NV → RP, MNV → rRP, NMV → RpP, MNMV → rRpP
ρ = 4, ρ_0 = 0, ρ_A = 16, ρ* = 28, Q_{WO} = 1/6

Type 3: Miao and Tidore (like Ket, only their basic order is fixed RrPp)

Type 3.5: hypothetical (like Lango, only without markers)
Basic WO: fixed RrPp
k = 4 (N, V, M, m)
NV → RP, NMV → RrP, NVM → RPp, NVm → RPp,
NMVM → RrPp, NMVm → RrPp
$\rho = 6$, $\rho_0 = 0$, $\rho_A = 26$, $\rho^* = 40$, $Q_{WO} = 2/9$

Type 4: Babungo
Basic order: fixed RrPp
k = 4 (N, V, a, m)
NV → RP, NaV → RrP, NVm → RPp, NaVm → RrPp
$\rho = 4$, $\rho_0 = 0$, $\rho_A = 18$, $\rho^* = 18$, $Q_{WO} = 2/9$

Type 4: Japanese (like Babungo, only its basic order is fixed rRpP)

Type 4: Hittite and Itelmen
Basic order: RrpP and rRpP
k = 4 (N, V, a, m)
NV → RP, NaV → RrP, aNV → rRP, NmV → RpP,
NamV → RrpP, aNmV → rRpP
$\rho = 6$, $\rho_0 = 0$, $\rho_A = 18$, $\rho^* = 18$, $Q_{WO} = 1/3$

Type 4: Arapesh
Basic order: RrPp, RrpP, rRPp, and rRpP
k = 4 (N, V, a, m)
NV → RP, NaV → RrP, aNV → rRP, NVm → RPp, NmV → RpP,
NaVm → RrPp, NamV → RrpP, aNVm → rRPp, aNmV → rRpP
$\rho = 9$, $\rho_0 = 0$, $\rho_A = 18$, $\rho^* = 18$, $Q_{WO} = 1/2$

Table 3. Results for the language sample.
[m] language with a grammatical marker
[v] language with variable word order

Type	Language	k	Q	Eff_A	Eff_{WO}	Eff
1	Samoan[mv]	2	2/37	.400	.189	.149
1	Hypothetical[m]	2	4/29	.571	.759	.379
1.5	Mundari	2	10/199	.400	.276	.138
2	Hurrian	2	1/17	.222	.324	.162
2	Hurrian*[m]	3	1/9	.364	.667	.178
2	Imbabura Quechua[mv]	3	6/59	.300	.305	.163

2	Warao[m]	3	1/11	.333	.545	.145
2.5	Turkish	3	3/56	.171	.321	.086
3	Ket, Miao, Tidore	3	1/7	.250	.857	.229
3.5	Hypothetical	4	3/20	.231	.675	.150
4	Babungo, Japanese	4	2/9	.222	1.000	.222
4	Hittite[v], Itelmen[v]	4	1/3	.333	1.000	.333
4	Arapesh[v]	4	1/2	.500	1.000	.500

Table 3 shows the results for the three types of grammar efficiency across the sample. Some conclusions can be derived from these values. They are presented in the following section.

7. CONCLUSIONS

It is obvious that grammar efficiency of natural languages is well below the theoretically possible maximum. The values of *Eff* in Table 2 are between 8.6% and 50% of the maximum, the average being 22.6%. This is partly because word order is too rigid and partly because the existing grammatical conveyors are not assigned to the four propositional functions in the most efficient way. The values of Eff_{A} average .314 (31.4% of the maximum), which shows that word order is not used as freely as theoretically possible. The hypothetical type 1 language and Arapesh have the freest word order (in the relative sense of how much word order can be free in those languages without producing ambiguity) as indicated by their values of Eff_{A} which are the two greatest ones in the table. On the other hand, only type 4 languages use the maximally efficient assignment of lexeme classes ($Eff_{\mathrm{WO}} = 1$). This is by default since there is just one way of assigning four lexeme classes to four propositional functions. For the twelve flexible languages (type $l < 4$) in Table 2, the average of Eff_{WO} values is .553 (55.3% of the maximum). This average is .684 for the whole table.

Variable word order should, generally speaking, increase *Eff* and Eff_{A}. This is indeed so with Hittite, Itelman, and Arapesh when compared to Babungo and Japanese. However, the picture changes in the presence of grammatical markers. It all depends on how the marker is used. For instance, the marker is not used so efficiently in Imbabura Quechua and, in spite of some variability in word order, the values of *Eff* and Eff_{A} do not surpass those of Hurrian* where word order is fixed but where the marker is put to a more efficient use. The marker is also used more efficiently in Hurrian* than in Warao. It should also be mentioned that the use of the marker in Warao as a *preposition* would increase its *Eff* to .267. Samoan is another example of the fact that natural languages do not use grammatical markers in the

most efficient way. The structure of the hypothetical type 1 language is more efficient than the structure of Samoan even with its freer word order.

All this confirms the conclusions from Vulanović (2008a), but it should be pointed out that the qualitative reasoning of Vulanović (2008a) cannot enable comparisons of languages the way they are done here. This shows the importance of using grammar efficiency to represent linguistic structures quantitatively.

Another conclusion is that intermediate type languages are less efficient than either type they comprise (in terms of *Eff*, not in terms of Eff_A or Eff_{WO}). This can be compared to the situation found in grammar efficiency models of syntactic change, see Vulanović (1995, 2005) for instance. In intermediate stages of a syntactic change, grammar efficiency is less than in the beginning of the change or in its end.

It should be kept in mind that all the above conclusions hold true for the models of simple intransitive sentences, which are considered here. Some things may change if more complicated structures or combinations of different structures are analyzed.

Finally, it is interesting to see whether there is some correlation between PoS types (variable *l*) and each of the three kinds of efficiency. In order to analyze this more realistically, I excluded the two hypothetical languages from the sample. As for the two languages of intermediate types, I left them in the sample although the values 1.5 and 2.5 are chosen for convenience, in the absence of anything else. If a linguistic continuum is assumed between, say, PoS system types 1 and 2, then any real number from the interval (1,2) may be possible to assign to an intermediate type language between types 1 and 2, depending on how close or far this language is to the end-point PoS systems. Of course, this is hard to estimate, if not impossible, and is beyond the scope of the present paper.

The results for linear correlation show nothing significant as far as Eff_A is concerned, but the coefficient of determination for Eff_{WO} and *Eff* is $R^2 = .849$ and $R^2 = .527$, respectively. The high value R^2 for Eff_{WO} is remarkable. Word order is controlled when Eff_{WO} is evaluated and, on the other hand, it seems that Eff_A and *Eff* depend on word order in a less predictable way. The other factor, the mapping of grammatical conveyors onto the propositional functions, offers some explanation why there is a positive correlation between the PoS system type and Eff_{WO}. If the type is higher, there are fewer possibilities of assigning lexeme classes to the propositional functions. Therefore, there is a greater chance that Eff_{WO} is closer to the maximally possible value if the type is higher. The culmination of this is $Eff_{WO} = 1$ for type 4 languages. However, it is not easy to explain fully at this point why the correlation is so strong. The linear correlation between PoS types and Eff_{WO} certainly requires further investigation.

REFERENCES

Hengeveld, K. (1992). Parts of speech. In: Michael Fortescue, Peter Harder, Lars Kristoffersen (Eds.), *Layered Structure and Reference in Functional Perspective*: *29–55*. Amsterdam/Philadelphia: John Benjamins.

Hengeveld, K., Rijkhoff, J. (2005). Mundari as a flexible language. *Linguistic Typology 9, 406–431.*

Hengeveld, K., Rijkhoff, J., Siewierska, A. (2004). Parts-of-speech systems and word order. *Journal of Linguistics 40, 527–570.*

Hengeveld, K., Lier, E.v. (to appear). An implicational map of parts of speech. *Linguistic Discovery*. Preprint available at http://home.hum.uva.nl/oz/hengeveldp/publications/An_implicational_functional_Map_of_Parts_of_Speech_160109.pdf. Accessed 25 March 2009.

Kutsch Lojenga, C. (1993). *Ngiti (Nilo-Saharan 9).* Köln: Reimer.

Mosel, U., Hovdhaugen, E. (1992). *Samoan Reference Grammar (Instituttet for sammenlignende kulturforskning B85).* Oslo: Scandinavian University Press.

Rijkhoff, J. (2007). Word classes. *Language and Linguistics Compass 1, 709–726.*

Romero-Figeroa, A. (1997). *A Reference Grammar of Warao (LINCOM Studies in Native American Linguistics 6).* München: LINCOM.

Speiser, E.A. (1941). *Introduction to Hurian (The Annual of the American Schools of Oriental Research 20)*. New Haven: American Schools of Oriental Research.

Vulanović, R. (1991). On measuring grammar efficiency and redundancy. *Linguistic Analysis 21, 201–211.*

Vulanović, R. (1993). Word order and grammar efficiency. *Theoretical Linguistics 19, 201–222.*

Vulanović, R. (1995). Model-based measuring of syntactic change. *Journal of Quantitative Linguistics 2, 67–76.*

Vulanović, R. (2003). Grammar efficiency and complexity. *Grammars 6, 127–144.*

Vulanović, R. (2005). The rise and fall of periphrastic *do* in affirmative declaratives: A grammar efficiency model. *Journal of Quantitative Linguistics 12, 1–28.*

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics 20, 399-427.*

Vulanović, R. (2008a). The combinatorics of word order in flexible parts-of-speech systems. *Glottotheory 1, 74-84.*

Vulanović, R. (2008b). A mathematical analysis of parts-of-speech systems. *Glottometrics 17, 51-65*.

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 176-192.*

Logistic regression model for predicting language change

Shoichi Yokoyama

Haruko Sanada

1. Introduction

Scientific studies, including language-related research, are concerned with phenomena of changes and the mechanisms behind the phenomena. Research in other natural science has shown that phenomena, in which choices and changes are involved, can be efficiently represented by s-shape curves and by corresponding logistic regression models in biological and social research, such as studies on the increase of animals in population ecology and increasing popularity of new products in economics. In fact, research has said that language changes are often represented by S-shaped curves, which starts with little and slow changes at the beginning, fast and vast changes in the middle, and little and slow changes again at the end. For example, Chambers (2006) reported the shift of the past tense of "sneak" in Golden Horseshoe, Canada, the past tense of "sneak" is traditionally "sneaked", however "snuck" appeared and spread. Figure 1.1 shows the ratio of the people who use "snuck" according to the age groups plotted on the X axis.

Figure 1.1 clearly shows that the shift from "sneaked" to "snuck" is represented by an s-shape curve. S-shape curves are also known to resemble to and correspond to logistic regression functions as shown in Figure 1.2. In language studies, the X-axis corresponds to time, such as ages of the participants and the time of data collection. The degrees of language changes are plotted on the Y-axis, representing the replacement of an old form by a new form.

Inoue (2000), indeed, has investigated language changes over forty years in Japanese. In his study, time is defined as the physical time of data collection to be plotted on the X axis. The language changes to be plotted on the Y axis were represented by the standardization of a regional variety. The data were collected three times by the National Institute for Japanese Language and twice by Inoue (2000) in Yamagata, a northeast region of Japan, in which a distinct local dialect is often observed. His data demonstrated that the language shift was described by an S-shaped curve by plotting the ratio of the regional and the counterpart standardized forms observed in the given geographical area over the five times of data

collection.

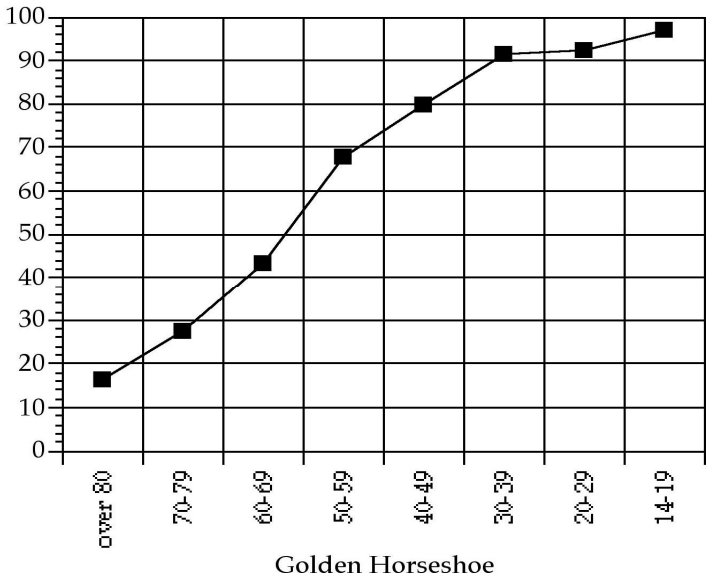


Figure 1.1. Change of past tense of “sneak”: “sneaked” vs. “snuck”

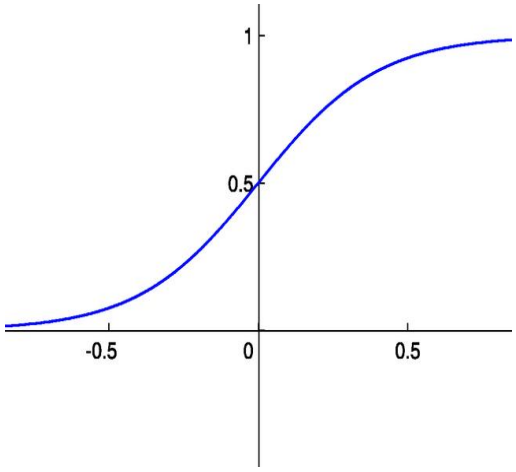


Figure 1.2. Logistic function

Studies, including but not limited to Inoue (2000), have shown empirical evidence as to the efficiency of the S-shaped curves to describe language changes, however, not much is available yet as to the theoretical and methodological research on the nature of the S-shaped curves and the mechanisms of language changes behind the observed phenomena. Thus, this paper provides a methodological suggestion for

analyses of language changes with multiple variables by applying a logistic regression model. It also introduces a theoretical explanation for the mechanism behind language change phenomena, which can often be described by S-shaped curves, by employing a psychophysical model.

2. Models of language changes 1: Linear regression model

Altmann, Buttlar, Rott, and Strauss (1983) provided an approximate curve, an s-shape curve, which describes language changes. The model is expressed as follows:

$$p = 1 / \{1+A\exp(- Kt) \}, \quad (1)$$

where p stands for observed ratio of responses and t represents time. K and A are constants, which are calculated the by least square method. It should be noted that Altmann et al. (1983) made a significant contribution, theoretically and methodologically, by explicitly describing language changes with a mathematical model with time as the variable. In fact, a number of studies have been conducted based on their model to describe language changes over time, particularly in Europe. For example, Sanada (2002) described the decrease of Sino-Japanese words in dictionaries of technical terms.

Equation (1) can be transformed in Equation (2) as follows:

$$\log\{(1 / p) - 1\} = \log A - Kt, \quad (2)$$

which is, apparently, a logistic function. Logistic models, in fact, represent the "density effect" model in population and biology studies. The density effect refers to the phenomena, in which increase of the population is interfered when the increase exceeds the capacity of the given environments. For example, excessive increase of population might trigger spread of diseases or lack of food. In short, the model of language changes by Altmann et al. (1983) as expressed by Equation (1) is theoretically identical to the density effect model as expressed by Equation (2). Indeed, a simplest differential equation of density effect can be expanded to Equation (1) when the density of population is proportionately allocated, indicating a theoretical comparability between the two models. In other words, the comparability among the models indicates that the principle mechanism behind the phenomena of changes may be commonly shared across areas of studies.

The model of language changes by Altmann et al. (1983) should be noted, however, their model leaves a few issues unsolved. First, the model as expressed by Equation (1) does not allow computation when the value of p is zero or 1.0. For

example, let us go back to the example of the past tense form "snuck" of "sneak". Suppose that the new form "snuck" is so popular that all the participants in their 20s and 30s use it. The probability of "snuck" is 100%, i.e. the *p* value equals to *p* = 1.0, with 0% of the counterpart "sneaked". Then, the left part of Equation (2) cannot be calculated because log 0 does not theoretically exist. Such a result consequently denies the existence of the given responses, i.e. "snuck", by the participants of the younger population in this example, which contradicts with the actually observed phenomena. In other words, the model is invalid for prediction of a language shift from an old to a new form.

The second issue is concerned with the candidate variable included in the model. As shown in Equations (1) and (2), time is the only variable to describe language changes, which is obviously counter-intuitive. Language changes presumably involve multiple variables, including physical, social, and psychological factor. In addition, "time" can be represented by multiple measures, such as "age of the participants" and "time of data collection". Thus, accurate description and prediction of language changes should consider multiple variables at the same time, including those related to time.

This paper proposes a logistic regression model by the maximum likelihood estimation as a solution for the two issues mentioned above. The model is expressed as follows:

$$\log\{p/(1 - p)\} = Z\,, \tag{3}$$

where *p* refers to the probability of response ratio in a binary-option task and *log* is the logarithm to the base *e*. *Z* stands for responses of a binary-option task, which can be expressed by a linear function with multiple variables as in $Z = a_1x_1 + a_2x_2 + b$. This paper refers to $p/(1 - p)$ as *odds* and $\log\{p/(1 - p)\}$ as *logit*. Logistic regression models, in other words, are expressed with logits on the left side and multiple regression models on the right side in the equations. Equation (3) can be further transformed to Equation (4) as follows:

$$p = 1/\{1 + \exp(-Z)\}, \tag{4}$$

where *exp* stands for exponential function. Equations (3) and (4) can be expanded to Equations (5) and (6), with X_1 being time and a_1 and *b* being the constants.

$$\log\{\, p/(1 - p)\,\} = a_1x_1 + b, \tag{5}$$

therefore

$$p = 1/\{1 + \exp[-(a_1x_1 + b)]\}, \tag{6}$$

which is a logistic regression model commonly applied to medical studies, exhibiting its efficiency to predict the ratio of occurrences in binary phenomena.

A logistic regression model as expressed by Equations (5) and (6), which is originated from Equation (3), is seemingly applicable to studies on language changes. In fact, Equation (3) can be deduced from Equation (2) by mathematical manipulation. Equation (2) can also be transformed in Equation (2.a.) as follows:

$$\log\{ p / (1 - p) \} = Kt - \log A , \qquad (2a.)$$

where t represents time and A and K are constants.

Applications of logistic regression analyses are, indeed, found in sociolinguistic studies as well. Labov (1972), for example, employed a logistic regression model to analyze empirical data. Studies with logistic regression analyses are found in Japanese as well (Hibiya, 1988; Matsuda, 1993, Yokoyama & Wada, 2006), all inductively showing empirical applicability of logistic regression models to describe language changes. Little research, however, theoretically explained the mechanism behind the S-shaped curves and logistic regression models. Thus, it should be of contribution to theoretically explain the S-shaped curves of language changes. Such research should also provide a method of analyses of language changes. Thus, this paper proposes a method to analyze language changes, which can be represented by an S-shaped curve with multiple variables, by applying a multiple logistic regression models.

3. Models of language changes 2: Logistic regression models

Sociolinguistic studies have shown that language changes involve multiple variables, such as contexts, environments, genders, and geographical regions. Thus, it should be methodologically necessary to consider multiple variables in a model to express language changes, so the critical factors are identified among various candidates. It should also be noted that models should be valid and applicable to a wide range of data possibly observable, including but not limited to, probability values of zero and 1.0, which previous models could not accommodate. The first purpose of this paper, therefore, is to propose a model, a logistic regression model with multiple variables, in specific, as a methodological solution.

This section simulates the applicability of the proposed logistic regression model using existing data by Chambers (2006). For example, let us consider the psychological condition of the participants in the data described in Table 1. Suppose that the participants of 50s and 60s were in a marked psychological con-

dition, due to the environment of data collection, for example, or to the incompetence of the interviewers. Also suppose that those participants in a marked psychological condition used the new form "snuck" 30% to 50% more often than in normal psychological conditions. Table 1 below summarizes such a hypothetical data, with the two independent variables, i.e. age and psychological condition, and with the independent variable, i.e. ratios of observed occurrences of "sneaked" vs. "snuck". The value 1 for the psychological condition denotes a marked psychological condition, whereas zero represents a normal condition.

Table 1
Hypothetical data modified from Chambers (2006)

Age	Psychological condition	Ratios of observed occurrences of "snuck"
85	0	18
75	0	28
65	1	95
55	1	95
45	0	80
35	0	91
25	0	92
15	0	98

The values of parameters in the model were estimated based on Equation (7) as follows:

$$Z = a_1x_1 + a_2x_2 + b, \tag{7}$$

where Z represents the ratio of "snuck" responses as opposed to "sneaked", x_1 and x_2 denote the age and the psychological condition, respectively. The values of the constants a_1 and a_2 were $a_1 = -0.076$ and $a_2 = 0.970$, respectively, and the b value was $b = 4.782$ by the method of maximum likelihood estimation.

The probability estimated by the logistic regression model is expressed by Equation (6) as follows:

$$p = 1 / \{ 1 + \exp [-(a_1x_1 + a_2x_2 + b)]\}, \tag{6}$$

as mentioned earlier. Thus, the rate of "snuck" occurrences against "sneaked" to

be estimated is obtained by applying the values of constants to Equation (6), which is represented as follows:

$$p = 1 / \{ 1 + \exp(0.076\, x_1 - 0.970\, x_2 - 4.782)\}, \qquad (6.a.)$$

which yielded the predicted values of probability *P* as shown in Figure 3.1 below. Figure 3.1 also shows the observed occurrence rate of "snuck". Likewise, Figure 3.2 shows the observed and predicted occurrence rates of "snuck" based on the regression model with single variable as expressed by Equation (1) by least square method. It is visually clear, based on Figures 3.1 and 3.2, that the multiple regression model as expressed by Equation (6) predicted the responses more precisely than the model with single-variable did.

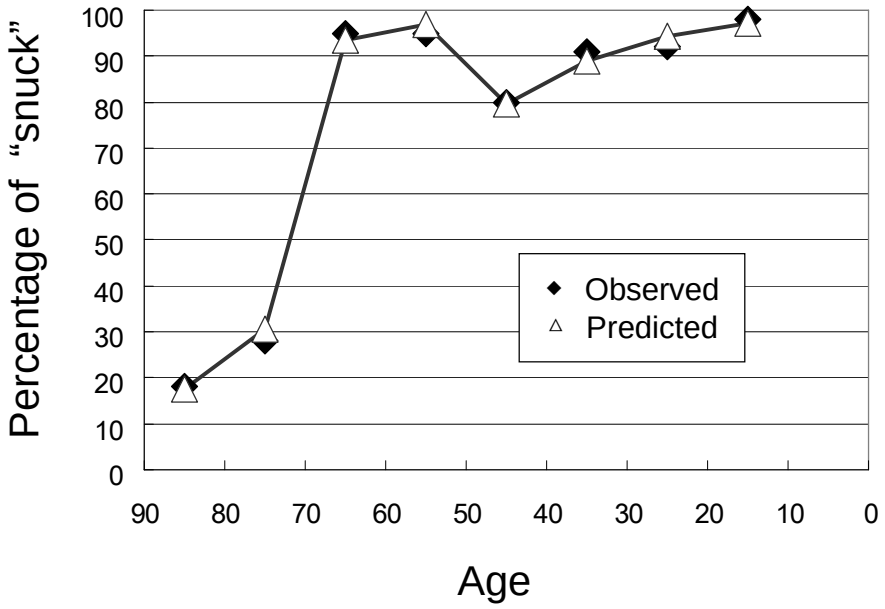


Figure 3.1. Prediction from multiple logistic regression model as expressed by Equation (6)

These results indicate that logistic regression models of the S-shaped curves allow analyses with multiple variables, which was not available with single-variable regression models without less significant variables. Analysis with multiple variables is apparently more powerful and accurate, as well as intuitively valid, and thus, more precise analyses and prediction should be available thorough multiple logistic regression models. For example, for sociolinguistic studies, gend-

ers, occupations, and locations of data collection can be included in the analysis together. For investigation of written languages, for example, writer-dependent variables, such as genders, and text-dependent variables, such as word types and genres, can be included in the analyses at the same time.

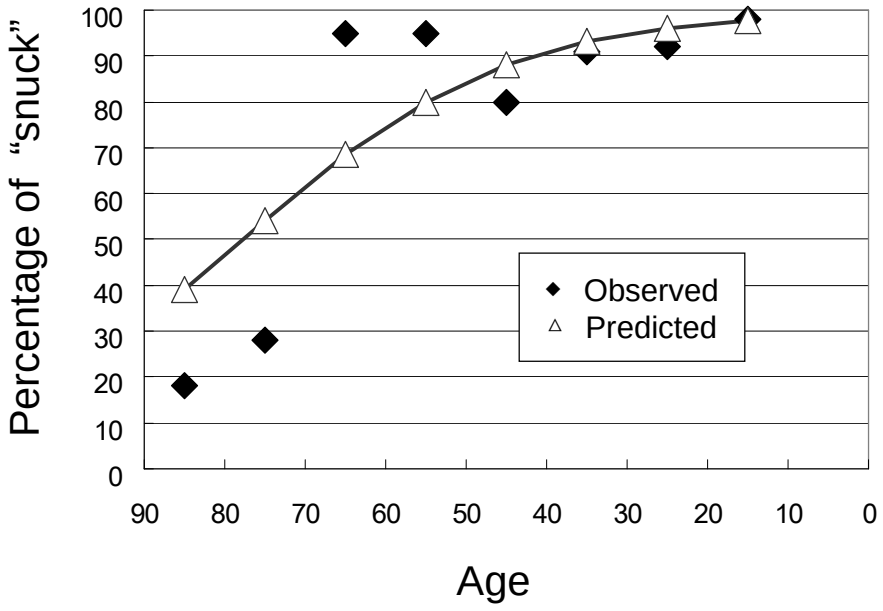


Figure 3.2. Prediction from single logistic regression model

4. Psychophysical model and the S-shaped curve of language changes

The second issue, which this paper addresses, is the theoretical explanation for a mechanism behind the phenomena of language changes. Previous research has addressed this issue, of course. For example, Altmann et al. (1983) employed a model to describe language changes, which is represented by S-shaped curves. Their model was, in principle, a differential function of the density effect. Thus, their model introduced that the density effect affects language changes, though implicitly, with the density being a factor external of language users.

This paper, in contrast, asserts that a psychological behavior, which is internal of language users, reflect languages changes, consequently yielding S-shaped curves. It is assumed that language changes are produced by binary options in human psychology, which is based on relative strength of the two

competing forms in the memory of language users (Yokoyama, 2006, 2007). It should be noted here, intuitively, empirically, and experientially that behavior and changes involving languages are attributable to both user-internal and external variables, possibly with interactions of the two as well. However, for the purpose of argument in this paper, a psychological factor, i.e. user-internal variable, is a major issue to be discussed. In specific, the discussion focuses on the role of memory strength, which is attributable to relative frequencies of the two competing forms of languages.

4.1. Logistic regression models and normal distribution

Regression models often assume normal distributions, and a part of the area is a critical measure used in statistical analyses. The cumulative sum of area, based on a normal distribution, is represented by an S-shaped curve as in Figure 1.2. as pointed out in testing research, an area of studies in psychological testing. The S-shaped curve can be expressed by a model as follows:

$$p = 1 \,/\, \{1 + \exp[\, -Da(\theta - b)] \}, \qquad (7.1)$$

where p is the accuracy rate of the test, b represents the degree of difficulty of the test, a does the distinctive power of the test, and θ is a constant. The b is an inflection point as well. Equation 7.1 above can also be expressed as follows:

$$p = 1 \,/\, \{1 + \exp(\, -DZ\,) \}, \qquad (7.2)$$

when $a = K$, $\theta = t$, $-ab = B$ and $Z = Kt + B$. Equation 7.2 is an approximation of the cumulative distribution function of a normal distribution if the value of D equals to $D = 1.7$, according to Lord and Novick (1968). Such comparability between Equation (7.2) and the cumulative distribution function of a normal distribution allows comparability between Equation (7.2) and a logistic regression model.

Equation (7.2), indeed, can be expanded to a logistic regression model if the value of D equals to $D = 1.0$ as expressed by Equation (4). Such comparability indicates that logistic regression models are an approximate representation of the cumulative distribution function of a normal distribution. In order to verify whether logistic regression models efficiently represent the cumulative distribution function, the differences were computed between the cumulative distribution function of a normal distribution with the standard deviation of 1.7 (SD = 1.7) and the values obtained from the logistic regression model, as shown in Figure 4.1.1. The errors fall in between ±0.01, indicating that logistic regression models can re-

present the cumulative distribution function with SD = 1.7 quite precisely, as well-known in testing research.

Cumulative sum of areas of a normal distribution can be mathematically obtained by integral calculation, however, the calculation is complex and demanding. On the contrary, logistic regression models can be computed without mathematical manipulations, and are thus an efficient method to approximately represent the cumulative sum of areas of a normal distribution.

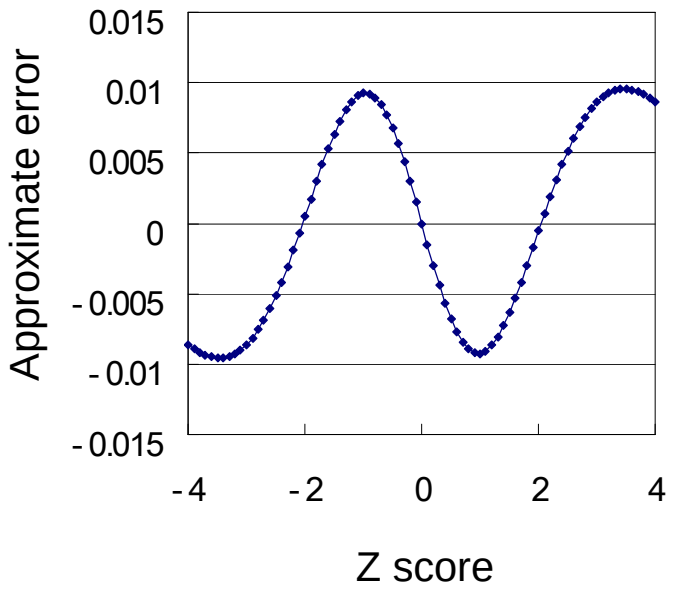
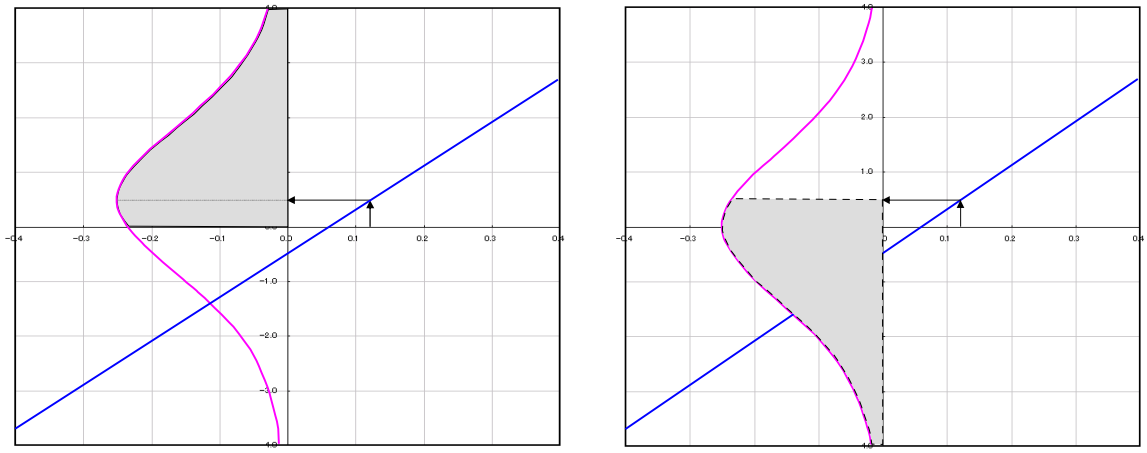


Figure 4.1.1 Differences between the cumulative distribution function of a normal distribution with the standard deviation of 1.7 (SD = 1.7) and the logistic regression model

4.2. Threshold model: Success, failure, and instability

Methodological efficiency of logistic regression models to represent a normal distribution also allows the computation of probability by a threshold model. For example, let us cite an example of language changes in Canada, in which the past tense form of "sneak" exhibits an old and a new forms, i.e. "sneaked" and "snuck". Suppose that the use of new form "snuck" is denoted by a "success" for the sake of methodological efficiency, following the convention which encodes a set of

binary options as either “success” or “failure”. Also suppose that variable *Z* represents the competency that enables “success” and that the variable Z is determined by variable *x*, i.e. age of the participants, in the “sneaked” vs. “snuck” example. Logistic regression assumes that the participants sharing the same *Z* values may produce either “success” or “failure”. In other words, “success” or “failure” is a probabilistic phenomenon, which allows both “success” and “failure”. The resultant phenomena, i.e. either “success” or failure”, depends on variable *Z*, which is dependent on variable *x*, but not a direct and linear function of *Z*. The probability for one of the either form to be observed is the highlighted area of the normal distribution represented by Figure 4.2.1. The *x* axis indicates age and the *Y* does the *Z* values, which can be obtained from the linear function. “Success” or “failure”, however, may vary unstably, depending on other variables besides *Z* values of competency, such as contexts, physical and psychological conditions, and luck. Such instability may vary, either positively or negatively, contributing to the “success”, however, its variability is assumed to exhibit regularity with a normal distribution with the *Z* value at the median in psychophysics. For example, a “success” reflects competencies as well as luck, and neither of the two alone guarantees a “success”. In other words, the normal distribution of the variability represents a combination of competency and instability.



(a) (b)

Figure 4.2.1 Sum of the area above the threshold equals to the probability value *p*

Let us suppose that the normal distribution curve in Figure 4.2.1(a) moves along the *Y* axis, i.e. the *Z* values. Suppose that the distribution curve stops at the point, at which the mean of the normal distribution coincides with the *Z* value corresponding to a given age, i.e. variable *x*. The standard deviation value of such

a distribution is SD = 1.7. The point at which $Z = 0$ is called a threshold, and the sum of the area above the threshold equals to the probability value p, i.e. rate of the new form "snuck" to be chosen against the old form "sneaked".

The probability value p equals to that obtained by a procedure as follows. First, imagine moving the normal distribution curve in Figure 4.2.1(b) along the Y axis, i.e. the Z values. Second, move the distribution curve until its mean of the normal distribution falls where Z value equals to zero. In other words, the mean, i.e. the peak of the distribution curve, falls at $Z = 0$ in Figure 4.2.1(b), whereas the peak is at any of the given values of Z in Figure 4.2.1(a). The value of the standard deviation equals between the two curves represented by Figures 4.2.1(a) and (b), with SD = 1.7. Third, define the threshold according to the given value of variable x, i.e. age. Fourthly, compute the sum of the area below the threshold. The value of the sum of areas computed as this is the probability value p for "success".

The method described above is applicable to categorical data as Table 3.1, allowing us to estimate that "snuck" will be observed if the probability value is $p > 0.50$. Likewise, "sneaked" is expected when the probability value is $p < 0.50$. Furthermore, this model can handle multiple variables, which is not explicitly treated in the statistical references known to the authors.

The threshold model is based on normal distributions, however, a logistic regression model is also applicable to computation of the probability values p, because logistic regression models can efficiently represent the cumulative sum of areas of a normal distribution with a standard deviation of 1.7. The logistic regression model to be used for such a purpose as probability computation can be expressed as follows:

$$p = 1 \,/\, \{1 + \exp(-Z)\}, \tag{4}$$

4.3. Exposure relativity theory: Memory strength and S-shaped curves

The purpose of this section is to present a model of Z values and to provide an explanation as to why the threshold model yields varying values for Z.

Let us assume, to start with, that a shift from the old form "sneaked" to the new form "snuck" reflects the relative frequency of exposure to the two competing forms. In other words, the more you are exposed to the new form, the more likely you choose the new form against the old counterpart. Such a role of relative frequency has been, indeed, introduced as the exposure relativity theory by Yokoyama (2006), based on the mere exposure effect and the generalized matching law.

Exposure relativity theory asserts as follows. As shown by the studies of the mere exposure effect, repeated exposure to unfamiliar forms of languages

increases favorability for the new forms. Increase of favorability contributes to more frequent use of the new form in the given community, which, in turn, relatively decreases the frequency of the old form. Increased use of the new form increases frequency of exposure to the new form even more, whereas less use of the old form decreases the relative frequency of exposure to the old form in the given community. Such changes of relative frequencies consequently reinforce the use of the new form even more. In other words, the exposure relativity theory asserts that use and exposure reciprocally reflect each other. The model of exposure relativity theory (Yokoyama, 2006) is expressed as follows:

$$\log \{p \,/\, (1 - p)\} = S \log (r_1 \,/\, r_2) + \log b \,, \qquad (8)$$

where p represents the frequency of the new form "snuck", r_1 the frequency of exposure to the new form "snuck", r_2 the frequency of exposure to the old form "sneaked", with S being sensitivity and log b being a response bias. Equation (8) is a logistic regression model, with the right side log (r_1 / r_2) representing the relative frequency of exposure called "exposure relativity" (Yokoyama, 2006) that contributes to the preference for the new form "snuck" against the old form "sneaked". Equation (8), in fact, can be transformed to (4), i.e. a logistic regression model with multiple variables as well. The discussion of this paper employs a logistic regression model as a practically efficient representation of the threshold model based on a probabilistic perspective, which assumes a normal distribution.

The model expressed by Equation (8) can explain a wide range of variations, such as regional and generation variations. For the sake of argument, let us take variations according to ages as an example. It is conceivable that older people are exposed to the old form "sneaked" more frequently than to the new form "snuck". Thus, let us assume that the relationship of exposure frequencies between the new form "snuck" and the old counterpart "sneaked" is expressed as $r_1 < r_2$, where r_1 denotes exposure frequency to the new form and r_2 refers to that of the counterpart old form. Let us also assume that the younger generation is exposed equally to the new and old forms, i.e. $r_1 = r_2$. These assumptions naturally lead to more frequent observations of the new form "snuck" among the younger generation than among the older generation, as represented by an S-shaped curve as in Figure 1.1, for $r_1 < r_2$ applies to the older generations whereas $r_1 = r_2$ does to the younger generations. The Z values can be obtained by Equation (9) as follows:

$$Z = S \log (r_1 \,/\, r_2) + \log b \,, \qquad (9)$$

by simultaneously equating Equations (3) and (8). In sum, Equation (9) is a proposed model of language changes from one form to the other, i.e. a binary

option, obtained by mathematical manipulations. Therefore, further validation with empirical evidences is necessary. Further theoretical research is also called for because the model is a deductive explanation of the given phenomena, i.e. *Z*, and therefore it is an expression of the mechanism of phenomena.

The theoretical explanation proposed here, which is based on the exposure relativity theory, expresses the effect of relative frequencies of exposure in human memory, in that more frequent items provide relatively more strengths in memory than less frequent items do. For example, old generations are more frequently exposed to the old form "sneaked" compared to the new form "snuck", therefore the relative strength of the old form "sneaked" is stronger in their memory compared to the counterpart new form "snuck". Such a relative gap of strength in memory consequently leads to more probability to choose the old form and less probability to choose the new form. On the other hand, the younger generations are, presumably, exposed to the old form "sneaked" less frequently than the older generations due to their limited experience with the old form in their environments. Because the younger generation has not got much exposure to the old form, the relative exposure frequency to the new form is higher among the younger generations than among older generations.

For example, let us suppose that the old form "sneaked" was used once yesterday. The older generation has seen it yesterday as well as many times over the past tens of years of their lives. Thus, the total number of times, for which they have seen it could be ten times, hundred times, etc. cumulatively. On the other hand, suppose that the younger generation has seen the old form "sneaked" only yesterday but never before for they haven't lived long enough to get many opportunities to see the old form. Thus, the younger generation has seen the old form "sneaked" only once in their lives. Suppose the counterpart new form "snuck" was also used once yesterday, which results in the same frequency of exposure for both the older and younger generations as for yesterday. Then, the relative frequency of exposure to the new and old forms, i.e. r_1 / r_2, for the older generation is $r_1 / r_2 = 1 / 10 = 0.1$, whereas that for the younger generation is $r_1 / r_2 = 1 / 1 = 1.0$. Therefore, the new form "snuck" leaves more room and strength in the memory of the younger generation, whereas it only takes up a very small impression, or strength, in the memory of the older generation in their entire lives. Such a relatively smaller strength of the new form leads to an only small probability to choose the given new form among the older generation. Attention to memory, i.e. a mechanisms that is internal of language users, and the assumption of frequency effect, i.e. user-external factor, is presumably critical to research on language changes, because language use is, in principle, a social behavior produced by human minds. By the same token, recent brain studies, which focus on user-internal mechanisms, should contribute to research of language changes and

behavior, provided that such studies consider social and inter-personal interactions that intra-personal factors experience.

A shift from an old form to a new one is comparable to the standardization of regional varieties. For example, Inoue (2000) asserts that the speed of standardization of a regional variety reflects the amount of inter-personal contacts, i.e. geographical neighborhood effect, and frequency of the observed phenomenon of the given language shift. He further proposes a model of standardization speed of regional varieties as follows:

$$V = f (C \cdot freq), \qquad \text{Modified from Inoue (2000)}$$

where V denotes the speed of language shift, C represents the amount of inter-personal contacts, and $freq$ does the frequency of the observed phenomenon of the given language change with f being a given function. The variables in his model are inter-personal contacts and frequency of language phenomenon, to which exposure frequency to language forms is attributable, in the given communities. His model thus consequently supports the basic principle/concept of the exposure relativity theory which considers exposure frequency as a major variable explaining language phenomena.

Little research is available so far, which considers probability of language change phenomena to be observed. Conceptual connection between the multiple logistic regression model and the exposure relativity theory should contribute to the advancement of the language studies, for the model expresses the S-shaped curve of language changes, as discussed in this paper, and the theory accounts for a basis which interweaves both psychology and observable phenomena, i.e. intra- and inter-personal variables, both of which are presumably critical to language changes.

5. Summary and discussion

This paper first demonstrated the efficiency of the S-shaped curves by analyzing a hypothetical data with a logistic regression model. In addition, estimation and probabilistic use of the models were not explicitly introduced in research on language changes.

The paper further proposed a theoretical view as to the mechanism behind the S-shaped curves of language changes. Altmann and his associates proposed the notion of S-shaped curves to describe language changes, however, their theoretical background is based on socio-linguistic perspective, focusing on inter-personal factors observable in the given communities. This paper, in contrast, employs a

psychological perspective, which considers intra-personal variables as a critical factor. In other words, this study and the series of studies by Yokoyama (2006, 2007) considers memory to be precise, i.e. an intra-personal variable, as a critical factor governing the mechanism of the S-shaped curve of language change phenomena. Such a psychological perspective has a long tradition in psychology, which focuses on intra-personal variables, however, the model proposed in this paper includes both social phenomena, i.e. inter-personal, as well as memory, i.e. an intra-personal psychological mechanism. In addition, it argues that both of the two factors reciprocally affect each other and play critical roles in the mechanism of the S-shaped curves to explain the language changes. This paper further applied the proposed model to the hypothetical data in this paper for the purpose of prediction. The proposed model should contributes to the research of language changes by allowing researchers to include and combine the two distinct variables, i.e. inter-personal and intra-personal ones, interwoven in a quite simple and efficient model.

The model should allow identification of the most critical factors that explain the language changes over a long period of time, when the nation has experienced vast economic and socio-cultural changes.

References

Aitchison, J. (1991) *Language change: progress or decay?* 2nd ed. Cambridge: Cambridge University Press.

Altmann, G., von Buttlar, H., Rott, W., Strauss, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical Linguistics: 104-115* (*=Quantitative Linguistics.* Vol. 18). Bochum: Brockmeyer.

Baum, W.M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior 22, 231-242*

Belke T.W., Belliveau J. (2001). The general matching law describes choice on concurrent variable-interval schedules of wheel-running reinforcement. *Journal of the Experimental Analysis of Behavior 75, 299 - 310.*

Chambers, J.K. (1998). Social embedding of changes in progress. *Journal of English Linguistics 26, 5-36.*

Chambers, J.K. (November, 2006). Paper presented at the meeting of National Institute for Japanese Language, Tokyo, Japan.

Collett, D. (2003). *Modelling Binary Data*. London:Chapman and Hall.

Elliot. R., Dolan. R. (1998). Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging

study. *The Journal of Neuroscience 18, 4697-4704.*

Fagen, R. (1987). A generalized habitat matching rule. *Evolutionary Ecology 1, 5-10*

Hibiya, J. (1988). *A quantitative study of Tokyo Japanese.* Doctoral Dissertation, University of Pennsylvania, 1988.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lord, F., Novick, M. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Matsuda, K. (1993). Dissecting analogical leveling quantitatively: The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change 5, 1-34.*

Moin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology 85, 1035-1048.*

Sanada, H. (2002). *Kindai nihon-go ni okeru gakujutsu-yougo no seiritsu to tei-chaku* [Emergence and establishment of academic terminologies in modern Japanese]. Tokyo, Japan: Junbunsha.

Woolverton, W.L, Alling, K. (1999). Choice under concurrent VI schedules: comparison of behavior maintained by cocaine or food. *Psychopharmacology 141, 47-56.*

Yokoyama, S. (2006). Mere exposure effect and generalized matching law for preference of Kanji form. *Mathematical Linguistics 25, 199 - 214.*

Yokoyama, S., Wada, Y. (2006). A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law. *Glottometrics 12, 63 -74.*

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology 9, 1-27.*

R. Köhler (ed.) (2009), *Issues in Quantitative Linguistics: 193-203.*

Quantitative Methods in Computational Dialectometry

Thomas Zastrow
Erhard Hinrichs

Introduction

The study of language variation as evidenced in regional dialects and dialect continua has a long tradition in linguistics. In dialectology, dialects are examined and analysed with the help of isoglosses. An isogloss is a boundary between two different realizations of a particular linguistic feature such as different pronunciations or different lexical items for a particular concept or object (Chambers 1980). Accordingly, dialect data are typically collected as: a) phonetic data which represent changes in the pronunciation of a specific word and b) lexical data which show the use of different words across dialects. The linguistic analysis of such data sets is traditionally recorded in language or dialect atlases that display isogloss maps of individual linguistic features or of bundles of them.

As more and more dialect data have become available in electronic form, the methodology of traditional dialectology has been extended by several quantitative methods. This has led to a new subfield of (computational) dialectometry. Apart from linguistics, computational dialectometry utilizes methods and concepts from other scientific disciplines, in particular from statistics, pattern recognition and machine learning. These methods include relational approaches such as "Relativer Identitätswert" (Goebl 1982), edit distance algorithms (Levenshtein algorithmus, Nerbonne 1999) and alignment algorithms (Kondrak 1999).

In this paper, two new approaches to computational dialectometry are introduced that extend previously developed and widely used approaches in dialectometry. Both methods focus on phonetic data and are applied here to an electronically available data set for regional variants of Bulgarian. As a point of comparison, the results obtained by these two methods are compared with traditional isogloss maps obtained for Bulgarian on the basis of lexical data.

The Data Set

In cooperation with the Bulgarian Academy of Science and the University Sofia[1], a large dialect data set of the Bulgarian language has been compiled (Zhobov 2006). This data set covers 197 geographical defined points (so called sites), spread over the whole territory of Bulgaria. The data set consists of two parts:

[1] http://www.sfs.uni-tuebingen.de/dialectometry

- Phonetic data which show the dialectal variations in the pronunciation of 118 commonly used words that have been collected for all 197 sites. This data is encoded in XSampa, an electronic form of the International Phonetic Alphabet (IPA).
- Lexical data. This data show the use of different lexical variations for 114 lemmata, common to all sites.

Both types of the data are encoded in XML, using DTDs for validation. The ClarK system (Simov 2004) was used to create the data. Data analysis was supported by the eXist XML database management system[2] and its seamless Java programming interface.

Workflow and Visualization of Phonetic Data

Dialect data has the form of a two dimensional matrix, which allows investigations in two directions (Figure 1):

- Single-Word-All-Sites: in the SWAS direction, one word in all sites is examined, for example the pronunciation of the word "red" in all of the 197 sites.
- Single-Site-All-Words: in the SWAS direction, all different words for one site are examined.

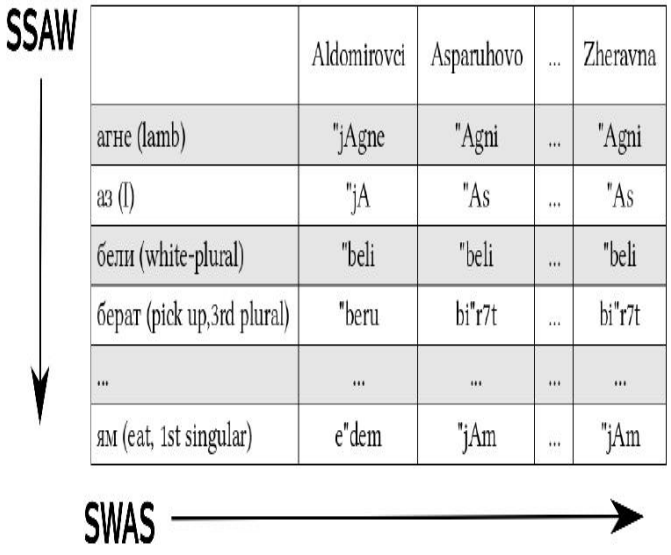


Figure 1. Investigations are possible in two directions

Depending on the dialectometric method which is applied, both or only one of these two directions are possible. The investigations in this paper are relying on the SSAW direction[3].

After a method is applied to a data set, a distance or similarity matrix can be build from the results. These matrices are always symmetric, so that the dis-

[2] http://exist.sourceforge.net

[3] Another common method in Computational Dialectometry is the so called Levenstein Distance. This method relies on the SWAS direction.

tance between a site S1 and a site S2 has the same value than the distance from S2 to S1.

The matrices can now be analyzed and visualized in several ways[4]. The analysis contains methods like a synopsis map or several algorithms of (hierarchical) clustering (Schulte im Walde 2003). The subsequent visualization can be in the form of dendrograms (only clustering) or in the form of Voronoi maps. In addition, maps which show isoglosses or rays between the dialects are possible. An isogloss map shows the borders between dialects, while a ray map does the opposite: it visualizes the communication between the dialects (Figure 2).

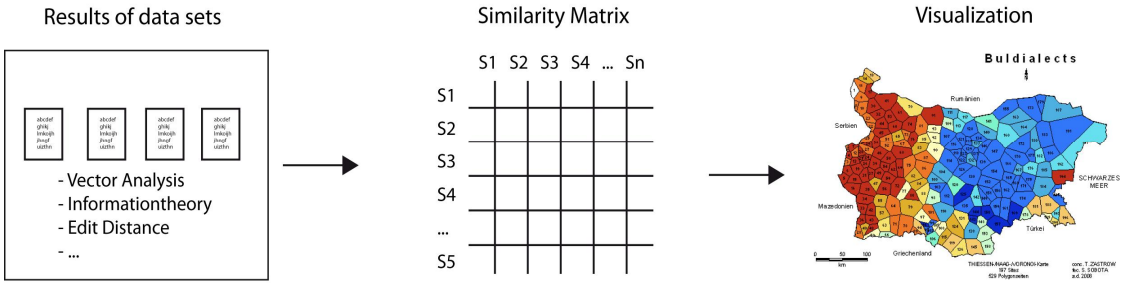


Figure 2. Workflow in Computational Dialectometry

Information Theory

In his seminal paper, *A Mathematical Theory of Communication*, Claude Elwood Shannon laid the foundation for the field of information theory and developed a measure for quantifying the information contained in a message (Shannon 1948). In deference to its founder, information theory measures the amount of information contained in a message in terms of Shannon units, which represent the minimum number of bits needed to encode a given message.[5]

More formally, the amount of information in Shannon of every atomic element z of a given data set can be calculated with the help of the probability *p(z)* of that element:

$$I(z) = -\log_2 p(z)$$

This formula expresses the information in terms of the base-2 number system. The equation assigns to an element with high probability a low amount of information and vice versa. Actually, the logarithmic character of the formula results in an overvaluation of elements with low probability.

---

[4] Two mature software packages for analyzing dialect data are available: the VDM software, written by Edgar Haimerl at the University Salzburg and the L04 package, written by Peter Kleiweg at the University Groningen. Where it was not possible to use one of these packages, analysis and visualization was performed with Java programming.

[5] In contrast to Bit, the unit Shannon is continous and not discrete (Cover 2006).

Summing up the information for all elements of a data set gives the information amount of the whole data set[6]:

$$I(DS) = -\sum_{i=1}^{n} \log_2 p(z_i)$$

This definition of information can now be adapted to computational dialectometry: for every site the amount of information is calculated in the SSAW direction[7]. The atomic elements are the XSampa codes[8]. The calculation can be performed in two ways:

- The information is calculated individually for every site on the basis of the probabilities the elements have only in the actual site.
- The probabilities of the elements can be calculated on the basis of the whole data set. Then, the amount of information is calculated for every site on the basis of these probabilities

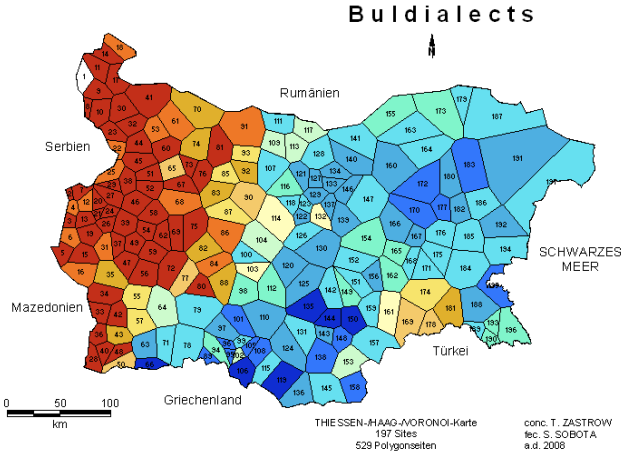


Figure 3. The amount of information in the sites, visualized with the VDM soft ware

[6] Readers familiar with information theory might wonder why Shannon is used as the measure of information instead of the more commonly used measure of entropy. The only difference between the two measures concerns the range of numbers used and the number of rounding operations that need to be performed. The calculation of entropy involves an additional multiplication step of the information of an element with its probability and thus an additional internal rounding operation of partial results. Since each additional rounding operation introduces an element of imprecision, the simpler calculation of Shannon values is to be preferred for purely technical grounds.

[7] Calculations in the SWAS direction are also possible. Calculations in this direction would show which words are carrying bigger and which are carrying smaller dialectal variations: the greater the amount of information, the higher the dialectal variations.

[8] Another possibility would be, to resolve the XSampa codes into phonetic features and then use the features as atomic elements.

The site specific amounts of information can be used as basis for creating a similarity matrix as described above.

The map in Figure 3 shows the results of this calculation and was created with the help of the VDM software. It shows a synopsis map with site number 1 as reference point. The amount of information was calculated individually for every site[9]. That means, that the colors of the polygones around the sites are calculated in relation to site number one. Red regions represent groups of sites that are most similar to the reference site, with color saturation corresponding to the relative degree of similarity to the reference site[10]. Blue regions represent groups of sites that are most dissimilar to the reference site, with color saturation corresponding to the relative degree of dissimilarity to the reference site.

Despite its analytic virtues, information theoretic approaches to dialectometry have one disadvantage: they take into account only the relative frequency of the atomic elements, but not its position within individual words that make up the data set. This means that word boundaries are ignored and the elements in the data set are treated like an unordered set.

Consider the following example: Data sets 1 and 2 are containing the element A one time, data set 3 contains the A two times. This results in identical information theoretic value for data set 1 and data set 2 (Figure 4).

	1	2	3	4	5
Data set 1	x	x	A	x	x
Data set 2	A	x	x	x	x
Data set 3	A	A	x	x	x

Figure 4. Data set 1 and 2 containing the same amount of information

The next section introduces a method that is fine grained enough to take into account individual words and to treat each word as an ordered set of elements.

Vector Analysis

Vector Analysis is a subfield of geometry (Schwartz 1960). Adapted to computational dialectometry, it has the advantage that it allows inspection of the whole data set in a single data structure. For a more detailed description of the vector

[9] If the probabilities are calculated on the basis of the whole data set as described above, the results are nearly the same. But other data sets could produce other results here.

[10] In the VDM software, synopsis maps are specified with some more parameters, for more information see http://ald.sbg.ac.at/dm/. All the maps in this paper are created with the same set of parameters, only the underlying similarity matrix changes.

analysis in computational dialectometry, see Hinrichs (2007). Other dialectometric methods such as edit distance or alignment algorithms are only able to compare the data set word by word, aggregating the partial results into a single value in a second analysis step. By contrast, vector analysis uses a single data structure that represents the trace of a focused element for an entire data set.

Vector analysis models vectors in a two- or higher dimensional space. A vector is defined as an array with a starting and an end point (Figure 5).

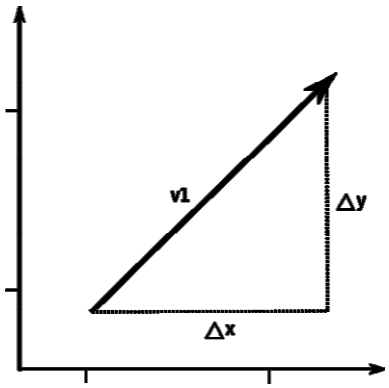


Figure 5. A vector in a two-dimensional space

The length of a vector can be calculated with the help of the Pythagorean theorem:

$$|v1| = \sqrt{\Delta x^2 + \Delta y^2}\,.$$

In computational dialectometry, chains of vectors can be used to trace the appearances and movement of an (phonetic) element through a data set. These vector chains are individual fingerprints for every data set, and their length[11] can be used to compare different data sets. Figure 6 shows on the left the first few single vectors of a vector chain. On the right side, six vector chains representing the vowel "e" are shown: the first three (Asparuhovo, Kozichino and Chernomorec) are in the eastern, while the other three sites (Rakovica, Stakevchi, Zelenigrad) are in the western part of Bulgaria. Just from the visual inspection, it is clear that the vector chains representing the western part of Bulgaria are longer than the ones in the east.

The examples in Figure 6 are showing vector chains in the SSAW direction: all the words of a site are taken into account to form the vector chain. In addition, vector analysis also allows comparisons in the SWAS direction. These comparisons show the degree of variation of individual words across data sets.

[11] Beside the length of a vector chain, the movement of the traced element from left to right in degrees could also be used as an individual feature.

A vector chain by itself has no interpretive value. But when two or more vector chains are compared to each other, the length of the vector chains represent[12]:

- the number of elements in the data set
- the position changes of the element within the data set

The map in Figure 7 was again produced with the help of the VDM software, visualizing the lengths of the vector chain for the vowel "e". The same parameters as for the map in the information theoretical approach were used. This map shows nearly the same structures as in the map above, but additionally there are some finer structures. For example on the eastern border, the anticipated transitional dialects to the Serbian language are visible.

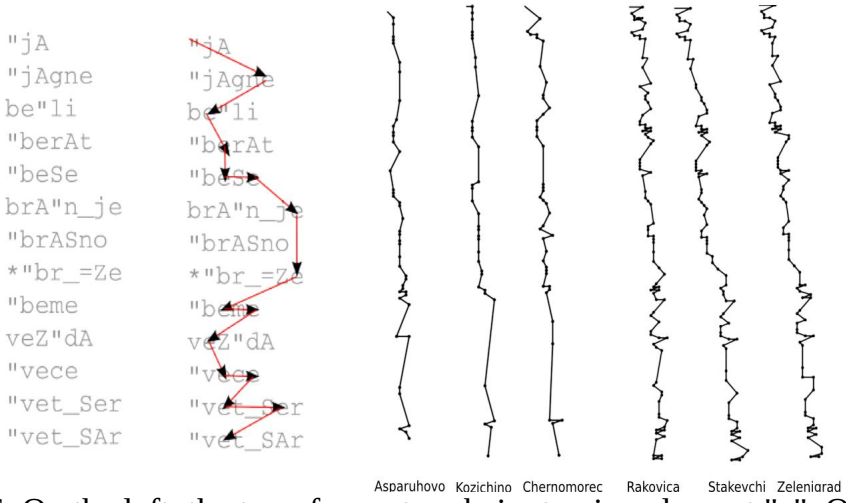


Figure 6. On the left, the top of a vector chain, tracing element "e". On the right, six complete vector chains for element "e", ordered from west to east

On the other hand, an analysis in terms of vector chains has one potential disadvantage: it takes into account only one focused element profiled in the vector chain. In principle, it is possible to aggregate two or more vector chains to a new one. For example, in this way the vector chains of all vowels could be aggregated. But dialect distinctions often manifest themselves in the variation of a small number of prominent differences so that dialect distinctions may not be clearly visible in aggregated vector chains.

[12] The order of words in the data sets under comparison has to be the same. If this is not the case, the vector chains are no longer comparable. John Nerbonne (personal communication) has pointed out to us that vector length is not invariant for all consistent permutations of words within data sets. For practical purposes this invariance does not have adverse effects if one simply takes the average length of vector chains across different permutations.

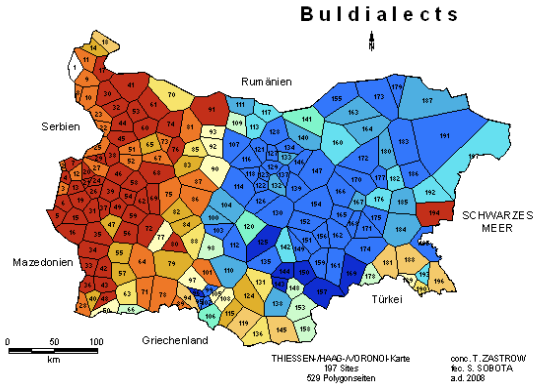


Figure 7. The lengths of the vector chains of element "e"

Lexical Data

In contrast to the computational methods shown above, this section presents a more conventional approach that is based on isoglosses of lexical data. Here the second part of the Bulgarian data set, the lexical data with 114 different lemmata, is used.

Instead of calculating individual values for every site and then aggregating them into similarity matrices, the lexical data is shown in the SWAS direction: for every lemma, one map is produced[13] which displays the lexical variation across sites. When such isogloss maps are compared for different words, the patterns of variation often do not coincide. Moreover, there is often no complete match with the patterns of variation obtained for the pronunciation data. However, the lexical and phonetic patterns of variation are also not in direct opposition to each other. Most of the time, the lexical isoglosses concern subparts of the dialect regions identified on the basis of phonetic data.

a) In some cases, lexical variation of a lemma is limited to a few sites. The map on the right shows this for the Bulgarian words for autumn which differ only in a few southern sites.

[13] Another possibility to analyse lexical data would be the use of the Relativer Identitätswert (RIW), see Goebl 1982.

b) Other lexical variation patterns of a lemma show a small number of geographically discontinuous subregions of variation (rendered here in green), with additional variation (rendered here in blue) clustering around such subregions. The area around the capital Sofia in the western part of Bulgaria and shown here in green often constitutes such a subregion of variation.

c) For some words, foreign influences along the borders (Serbia in the west, Greek and Turkey in the south and Romania in the north) can be seen. The purple variation in the south is located on the border to Greece, while the blue one points to Serbia.

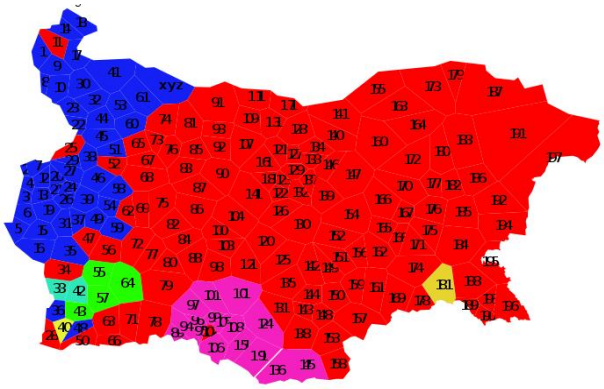


Figure 8-9. Different structures of lexical variations

Comparison to traditional dialectology

The Bulgarian dialects were already subject of some scientific investigations. Gutschmidt (2002) describes the main dialect distinctions of the Bulgarian language as follows:

- The so-called jat line named after the presence or absence of the semivowel j divides the eastern and western part of Bulgaria by a line running from the north in a southwestern direction.
- On the border to Serbia and north of Sofia, there is a belt of transitional dialects between Bulgarian and Serbian.
- In the south, the mountains of the Rhodopes form a separate and heterogeneous dialect region. In the middle of that segment, there is a break in the direction of Plovdiv.
- The sites at the border to Turkey form a region of their own.

These findings are clearly reflected in the dialectometric results for the phonetic data that have been presented in this paper. In contrast, the lexical data show more coarse-grained patterns of variation. It remains to be seen whether the use of lexical data can lead to further refinement of the clustering obtained exclusively on the basis of phonetic data.

Conclusions

In this paper, two novel approaches to computational dialectometry have been presented: an information theoretic approach and a vector-based approach. Both approaches lead to nearly identical results for the Bulgarian phonetic data set. These results conform to the observations in traditional dialectology. Future work on data sets for additional language will have to show whether the encouraging results obtained for Bulgarian will generalize.

An additional direction for future research concerns the length of segments under consideration. The results reported in the present paper are based on unigrams and thus do not reflect information about the immediate context to the left or to the right of a given segment. Since phonological processes are typically conditioned by local context, it seems appropriate to apply the same methods introduced here to bigrams or trigrams of segments, provided that this does not lead to severe problems of data sparseness.

Another open question concerns the utility of lexical data, which do not seem to yield enough "resolution" in comparison with the phonetic data. One reason for this lack of resolution could be that Bulgaria is a relative small country with a language that is in close language contact with languages spoken in surrounding countries. This language contact situation is substantiated by the fact that regions of variation typically occur near the national borders of Bulgaria. Thus, lexical variation may be the result of external linguistic influence to a much higher degree than phonetic variation.

Acknowledgements
This work was carried out as part of a collaborative research project "Measuring linguistic unity and diversity in Europe" funded by the Volkwagen-Stiftung. We are extremely grateful to Hans Goebl for making his VDM software available to us. We would also like to thank Hans Goebl and our collaborators at the University of Groningen (John Nerbonne, Jelena Prokic) and in Sofia (Petya Osenova, Kiril Simov, Georgi Kolev, Petar Shishkov and Vladimir Zhobov) for extended discussion on scientific matters related to the research reported here.

References

Chambers, J.K., Trudgill, P. (1980). *Dialectology*. Cambridge: Cambridge University Press.

Cover, T.M., Thomas, J.A. (2006). *Information Theory*. Hoboken: Wiley.

Goebl, H. (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien:Verlag der Österreichischen Akademie der Wissenschaften.

Gutschmidt, K. (2002), *Bulgarisch. Wieser Enzyklopädie des europäischen Ostens. 10: 219-234.* Klagenfurt: Wieser Verlag.

Hinrichs, E., Zastrow, T. (2007). A vector-based approach to dialectometry. *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands*.

Kondrak, G. (1999). *Alignment of Phonetic Sequences*. Toronto: Technical Report CSRG-402, University of Toronto.

Nerbonne, J., Heeringa, W., Kleiweg, P. (eds.) (1999). *Edit Distance and Dialect Proximity.* Stanford: CSLI Press.

Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes.* Stuttgart: AIMS Working Papers, Institut für Maschinelle Sprachverarbeitung.

Schwartz, M., Green, S., Rutledge, W.A. (1960). *Vector Analysis*. New York: Harper.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 1948, 27, 379-423* and *623-656*.

Simov, K., Simov E., Ganev, H., Ivanova, K., Grigorov, I. (2004). The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004, Lisbon, Portugal, 2004: 235-238.*

Zhobov, V. (2006). *Description of the Sources for the Pronunciation Data.* Sofia: Department of Slavic Philologies, University of Sofia.

Authors

Andreev, Sergey
Smolensk State University
Foreign Languge Department
ul. Prževalaskogo 4
RUS-214000 Smolensk, Russia
Email: smol.an@mail.ru

Broda, Bartosz
Wrocław University of Technology,
Institute of Informatics,
Poland
Email:maciej.piasecki@pwr.wroc.pl

Embleton, Sheila
York University
VP Academic & Provost
4700 Keele Street
M3J1P3 Toronto, Canada
Email: embleton@yorku.ca

Grzybek, Peter
Institut für Slawistik,
Karl-Franzens Universität,
Merangasse 70,
A-8010 Graz, Austria.
Email: grzybek@uni-graz.at

Hinrichs, Erhard
University of Tübingen
Institute for Linguistics
Wilhelmstr. 19
D-72074 Tübingen, Germany
Email: eh@sfs.uni-tuebingen.de

Kelih, Emmerich
Institut für Slawistik,
Karl-Franzens Universität,
Merangasse 70,
A-8010 Graz, Austria.
Email: emmerich.kelih@uni-graz.at

Köhler, Reinhard
LDV II, Universität Trier
Universitätsring 15
D-54296 Trier, Germany
Email:koehler@uni-trier.de

Králík, Jan
The Czech Language Institute, CAS
Letenská 4
CZ-11851-Praha 1
Czech Republic
Email: kralik@ujc.cas.cz

Mačutek, Ján
Institut für Slawistik,
Karl-Franzens Universität,
Merangasse 70,
8010 Graz, Austria.
Email: jmacutek@yahoo.com

Mikros, George K.
Dept. of Italian and Spanish
Language and Literature,
University of Athens,
Panepistimioupoli Zografou
15784 Athens, Greece
Email: gmikros@isll.uoa.gr

Naumann, Sven
LDV II, Universität Trier
Universitätsring 15
D-54296 Trier, Germany
Email: Sven.Naumann@uni-trier.de

Pawłowski, Adam
Wrocław University,
Inst.of Library and Information
Science, Poland
Email: apawlow@uni.wroc.pl

Piasecki, Maciej
Wrocław University of Technology,
Institute of Informatics, Poland
Email: maciej.piasecki@pwr.wroc.pl

Pustylnikov, Olga
Texttechnology Group,
SFB 673, University Bielefeld
Bielefeld, Germany
Email: Olga.pustylnikov@techfak.uni-bielefeld.de

Schneider-Wiejowski, Karina
Linguistics and Literature,
SFB 673, University Bielefeld
Bielefeld, Germany
Email:
karina.wiejowski@uni-bielefeld.de

Sanada, Haruko
Faculty of Humanities,
Saitama Gakuen University
1510, Kizoro, Kawaguchishi,
Saitama 333-0831
Japan
Email: h_sanada@nifty.com

Stadlober, Ernst
Institute of Statistics
Graz University of Technology
Münzgrabenstraße 11/III
A-8010 Graz, Austria
Email: e.stadlober@tugraz.at

Steiner, Petra C.
Dept. of Modern Foreign Languages,
University of Mary Washington
1301 College Avenue
Fredericksburg, VA
22401 USA
Email: psteiner@umw.edu

Uritescu, Dorin
York University, Glendon College
French Studies
2275 Bayview Avenue
Toronto, Ontario
Canada M4N3M6
Email: dorinu@yorku.ca

Vulanović, Relja
Dept. of Mathematical Sciences,
Kent State University Stark
Campus, 6000 Frank Ave NW,
North Canton, OH 44720, USA.
Email: rvulanov@kent.edu

Wheeler, Eric S.
York University
School of Information Technology
33 Peter Street
Markham, Ontario
Canada, L3P2A5
Email: wheeler@ericwheeler.ca

Yokoyama, Shoichi
National Institute for Japanese
Language, 10-2 Midori-cho,
Tachikawa City, Tokyo 190-8561
Japan.
Email: yokoyama@kokken.go.jp

Zastrow, Thomas
University of Tübingen
Institute for Linguistics
Wilhelmstr. 19
D-72074 Tübingen, Germany
Email:
thomas.zastrow@uni-tuebingen.de