5th Trier Symposium
on Quantitative Linguistics

ABSTRACTS

December 7 – 8, 2007

Quantitative Analysis of the Evolution of Mikhail Lermontov's Style*

Sergey Andreev
Smolensk State University

One of the directions in quantitative analysis of idiostyle is the study of its development during creative activity of an author. In general it consists in finding out the degree of variation of certain linguistic parameters in texts of the same author but written at different time. The count of linguistic characteristics of such literary texts with consequent comparison of the obtained data makes it possible to single out features which, remaining constant, form the basis of idiostyle and, on the other hand, characteristics which tend to vary in time, reflecting its evolution.

This paper deals with the problem of the evolution of style of Mikhail Lermontov, one of the most famous poets of Russia who lived in the first part of the 19th century. Lermontov's life and his creative activity were very short because of his early tragic death at the age of 27. The question arises if there was any evolution in his style which can be reflected by linguistic (not purely literary) parameters. Traditionally Lermontov's creative activity is divided into two periods: Period 1 (1828 – 1836) and Period 2 (1837 – 1841). The division is made due to his biographical data: in 1836 he was sent in exile.

One of the most important conditions for choosing texts for such analysis seems to be their homogeneity since differences in contents, size and genre of poems may determine to a great extent the difference in the choice of linguistic means, obscuring the main tendencies in the author's idiostyle.

Taking into consideration the above-mentioned conditions, we chose for the analysis poems of the same (on the whole) contents and structure: lyrical verses of 4-feet iambic meter, written in 4-line stanzas. Lyrics were chosen because this genre reflects individual manner of a poet best of all. The 4-feet iambic meter was one of the most popular among Russian poets of the first half of the 19th century. Stanza characteristics is still another restriction which allows to reach structural homogeneity of texts.

All texts, which satisfied the above-mentioned requirements, were grouped into two classes, corresponding to the indicated periods, and analyzed from the point of view of possessing the following characteristics: morphological (parts of speech of words in different positions in lines), syntactic (pause, enjambement, etc.), rhythmic (omission of stress on the first, second and third ictuses) and rhyme (masculine, feminine).

The data obtained was used for the comparison of these two groups of texts with the help of one of multivariate methods of analysis – discriminant analysis. Its results showed that despite the short time of Lermontov's creative activity there are certain differences in his style which are reflected by linguistic parameters. The following parameters were found to possess the biggest discriminating force: the omission of the stress on the third ictus, the number of nouns at the end of the line, the number of lines ending with exclamation or interrogatory marks.

* Address correspondence to: Sergey Andreev, Smolensk State University, Przhevalskogo str. 4, 214000 Smolensk, Russian Federation. E-mail: smoL.an@mail.ru.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Data Management and Linguistic Analysis: MDS applied to RODA*

Sheila Embleton, Dorin Uritescu and Eric S. Wheeler
York University, Toronto

Multidimensional Scaling (MDS) is a statistical technique useful for portraying a large number of independent measures as a two-dimensional map. It can be used for conveying an overview of the linguistic distances among locations with related dialects.

We have implemented an MDS function in our Romanian Online Dialect Atlas (RODA) and used the function to explore linguistic relationships in the important dialect region of North-Western Romania, known as Crişana. It reveals some interesting and novel results about the dialect situation in that area.

* Address correspondence to: Sheila Embleton, Vice-President Academic, South 939 Ross Building, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3. E-mail: embleton@yorku.ca.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Laws of Language Outside Human Language*

Ramon Ferrer i Cancho
Universitat Politecnica de Catalunya, Barcelona

The XX century witnessed the discovery of many laws of language such as Zipf-Mandelbrot's or Menzerath-Altmann's. At the same time, the anthropocentric focus of standard linguistics on human language has been questioned at least by comparative psychology studies and a new branch within standard linguistics, i.e. biolinguistics. Here we show new evidence that some well-known quantitative linguistics laws of language are not unique to human language. Our findings suggest that there are abstract universal principles of information coding.

* Address correspondence to: Ramon Ferrer i Cancho, Departament de Llenguates i Sistemes Informatics, Universitat Politecnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona, Spain. E-mail: ramon.ferrericancho@gmail.com.

Systematic and System-based Studies of Grapheme Frequencies*

Peter Grzybek and Emmerich Kelih
Karl-Franzens-Universität, Graz

This presentation concentrates on letter frequencies and their theoretical modeling, the approach being assumed to be relevant for grapheme, phoneme, or sound frequencies alike. The results thus far obtained clearly indicate that the frequency of letters is regularly organized. Based on recent findings proving the negative hypergeometric function to be an adequate model, this study concentrates on a detailed examination of parameter behavior. Additional evidence is provided that the parameter behavior follows clear rules, too, implying both intra-lingual (language-specific) and inter-lingual processes: (a) As to language-specific tendencies, it is shown that all parameters of this distribution behave regularly, as long as the analysis is based on the system's inventory size, rather than on the class of items occurring in the given sample. (b) As to interlingual tendencies, which can be derived from the comparative analysis of various languages, it can be shown that the parameter behavior characterizing the individual languages can be traced back to a common regression model. As to an interpretation of this parameter behavior, it turns out that three factors must be taken into account: (1) the relative frequency of the most frequent class (P_1), (2) the mean value (m_1) of the distribution, and (3) the system's inventory size (I).

* Address correspondence to: Peter Grzybek, Karl-Franzens-Universität Graz, Institut für Slawistik, Merangasse 70, A – 8010 Graz. E-mail: peter.grzybek@uni-graz.at.

The Complexity of Learning a Polysynthetic Language*

Tarik Hadzibeganovic[1] and Sergio A. Cannas[2]
[1] University of Graz; [2] National University of Córdoba

We study the dynamics of Western-Greenlandic word learning in two orthographically different populations (shallow and deep orthography). In simple memorization tasks, subjects monitor a total of six 9-letter Western-Greenlandic words arranged in a letter matrix for a fixed number of seconds. Learning was measured following each of the 10 stimulus exposures. The visual stimuli were learned to a criterion of two consecutive perfect recalls. We obtained learning curves for both shallow and deep orthography language speakers and compared the results to their performance on the equivalent tasks with Finnish words and *n*-letter nonsense strings (non-words). We further analyzed the complexity of novel word learning in a polysynthetic language by using numerical simulations of a Langevin equation based neural network model with non-extensive cost functions. The resulting learning algorithm with a non-local learning rule was able to replicate the population-specific learning behavior to a high degree. The model further allows for the analysis of learning efficiency given the number of bits of random information an agent consumes as it proceeds in a learning task.

* Address correspondence to: Tarik Hadzibeganovic, Language Development & Cognitive Science Unit, University of Graz, A-8010/Austria. E-mail: ta.hadzibeganovic@uni-graz.at;

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Thematic Concentration of Text in Indian Languages*

B. D. Jayaram, K. S. Rajyashree and M. N. Vidya
Central Institute of Indian languages, Mysore

Thematic concentration is used as a tool to predict the topic and the genre of the given text. It is based on the h-point proposed by Hirsch and developed by Popescu, Best, Altmann for various other application. The present paper investigates thematic concentration in two Indian languages namely Marathi and Kannada and in four genres namely Aesthetics, Commerce, Natural Physical and Professional Sciences and Social Sciences. It is observed that the order of thematic concentration across the genres remains the same in both the languages.

* Address correspondence to: Sri B.D. Jayaram, Research Officer, Central Institute of Indian Languages, Mansagangotri, Hunsur Road, Mysore -- 570006, Karnataka, India.
E-mail: Jayaram//ciil@CIIL.STPMY.Soft.Net.

Sentence Length – Word Length. A Systematic Revision of the Arens Law*

Emmerich Kelih and Peter Grzybek
Karl-Franzens-Universität, Graz

In quantitative linguistics, the relationship between the units of two different linguistic levels has usually been studied with reference to the Arens Law, the latter being considered to be a special case of the Menzerath-Altmann Law. In recent studies on the relationship between word length (WL) and sentence length (SL), emphasis has been laid on the distinction between intra-textual and inter-textual approaches. Paying attention to this distinction, it turns out that, for the inter-textual perspective, there seems to be only weak evidence in coincidence with well-known linguistic regularities. Furthermore, with regard to the intra-textual level, a number of factors come into play:

(a) *Minimal SL*: for very short sentences (≤ 4), the Menzerathian tendency does not seem to play a crucial role;
(b) *Maximal SL*: for very long sentences (≥ 30), the Menzerathian tendency does not seem to play a crucial role;
(c) *Minimal SL*: if there are not enough *SL* data points as a basis of average *WL*, variance is too large to result in some kind of general tendency.

As has recently been shown, these factors represent a necessary, though not sufficient condition for the Arens Law to be efficient, resulting in constant WL-SL relations for given text types, rather than in that kind of non-linear regression to be expected according to Menzerathzian principles. As a reason for this, it has been suspected that *textual heterogeneity* might be another, necessary factor: as long as the data material consists of homogenous texts (i.e., from a specific text type), *WL* seems to be regulated by the text type's specific *WL* organization. Only in case data from different text types are combined, the necessary textual heterogeneity is provided for the Menzerathian principle to come into play. Literary texts are likely to be characterized by this intrinsic heterogeneity, being composed of diverging text elements such as dialogues, descriptive and narrative sequences, auctorial comments, etc. The present study is a first test of this hypothesis.

* Address correspondence to: Emmerich Kelih, Institut für Slawistik, Merangasse 70, A-8010 Graz.
E-Mail: emmerich.kelih@uni-graz.at.

Sequences of Sentence Length*

Reinhard Köhler, Sven Naumann
Trier University

Reasonable amounts of data for quantitative analyses have become available without serious problems from large linguistic text corpora since corresponding computer techniques can be used to extract the needed material. However, on higher levels of linguistic analysis, such as the levels of sentences, clauses, and phrases, the situation is much more complicated than on lower ones. Therefore, the development of algorithms and software which can be used to automatically extract quantitative data on these levels is highly desirable. We will concentrate here on the sentence and clause levels.

Though the concept *clause* is a quite common one in linguistics, there is no general consent about the (necessary and sufficient) criteria which would help to identify entities of this type in an unambiguous way. In our work, we used a multi-pass, iterative algorithm which is heuristic in nature. It takes a tokenized and tagged text as input and outputs a sequence of clauses for each of the sentences. The text is scanned sentence by sentence. Each sentence is decomposed into a sequence of potential clauses (*clause hypothesis*). Based on positional information, the distribution of finite and non-finite verb forms and conjunctions, this list is transformed into the final list of clauses of the sentence considered. Preliminary tests using newspaper articles indicate that the success rate is about 95%.

The present paper will introduce the heuristic principles our program is based on and report on first experiences with its output as data for studies in quantitative phenomena on the sentence and clause levels with emphasis on Menzerath-Altmann's Law and on the study of sequences of properties such as the recently introduced L-segments (cf. Köhler & Naumann, to appear).

* Address correspondence to: Reinhard Köhler, FB II – Linguistische Datenverarbeitung, Universität Trier, D – 54286 Trier. E-mail: koehler@uni-trier.de.

Contemplations on Corpus Infinity*

Jan Králík
The Czech language Institute, Prague

It seems to be obvious that direct comparison or confrontation of quantitative data gathered from different texts on one side and from corpora on the other side can not form strong bases for meaningfull interpretation. However, the refusal of such comparison or confrontation can not be general. At least, it should stay on stronger argumentation, than on a mere feeling. Contemplations on Corpus Infinity try to show that not all comparisons must be necessarily refused, and, which seems yet more important, contemplations will try to underline reasons why such comparison cannot be accepted for all cases generally. Two ways of argumentation will be taken in question: 1. construction based on the Menzerath-Altmann Law and its consequences, 2. construction based on axioms of the Probability Theory and their presumptions. The conception of infinity and its understanding will be discussed from the point of view of quantitative lingvistics, mathematics and corpora. Some examples will be given as to possible and impossible confrontations.

* Address correspondence to: Jan Králik, The Czech Language Institute, AV ČR, v.v.i. Letenská 4, 11851 Praha 1, CZ. E-mail: kralik@ujc.cas.cz.

Models of Graphemes Frequencies*

Ján Mačutek
Comenius University, Bratislava

Two new approaches to modeling grapheme frequencies will be presented. The first of them is a generalization of the geometric distribution resulting in a distribution which has not been described so far, namely

$$P_x = cp^{x-1}\left(1+\frac{a}{n-x+1}\right),\ x=1,2,\ldots,n,$$

where c is a normalization constant, $p \geq 0$, $a \geq -1$. The distribution belongs to the Wimmer-Altmann family, hence it is a special case of a very general model of linguistic laws. Its goodness-of-fit is roughly the same as for the negative hypergeometric distribution. On the other hand, the new distribution has no direct relation to binary urn schemes, which can possibly mean interpretability of the parameters.

The other approach does not follow the well known unified derivation of linguistic laws introduced by Wimmer and Altmann. In almost all cases observed grapheme rank-frequency distributions do not decrease smoothly and at least one "jump" occurs. The idea is to divide the frequencies into two parts and to model them separately. The new models, e.g. the "piecewise geometric" distribution, give some promising goodness-of-fit. Of course a qualitative interpretation of obtained results will be necessary.

* Address correspondence to: Jan Macutek, Dept. of Applied Mathematics and Statistics, Comenius University, Mlynska dolina, 84248 Bratislava, Solvakia. E-mail: jmacutek@yahoo.com.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Towards Diachronic Comparison of Morphological Profiles*

Alfonso Medina Urrea
Universidad Nacional Autónoma de Mexico, Mexico

The set of most prominent affixes and sequences of them of a wide variety of languages seems to be more intimate to those languages than any set of basic lexical items including cognates such as body parts, heavenly phenomena, personal pronouns, very basic numerals, etc. As it is known, measuring similarity among those basic sets of lexical cognates permits, among other things, estimation of how far back in time two or more languages were in fact the same language; usually in terms of millennia. Measuring similarity among sets of prominent affixes and sequences of them allows for comparison of diachronic stages of one language alone within much shorter periods of time.

In this presentation, quantitative data for three centuries of the Spanish language spoken in Mexico will be presented (XVIth, XVIIIth and XXth centuries) with the intent of corroborating (or not) intuitions put forward by philologists.

* Address correspondence to: Alfonso Medina Urrea, Ingenieria Lingüistica, Coordinación de Sistemas, Instituto de Ingenieria, Universidad Nacional Autónoma de México, Circuito Escolar S/N, Ciudad Universitaria, 04510 Covoacán DF, Mexico. E-mail: AMedinaU@iingen.unam.mx.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

A Model of the Distribution of Lexical Chains*

Alexander Mehler, Ulli Waltinger and Rüdiger Gleim
Bielefeld University, Bielefeld

Lexical chaining is the task of tracking semantically related tokens in texts where semantic relatedness is modelled by means of lexical reference systems as WordNet. Path-based chaining algorithms, e.g., judge tokens to be related subject to the shortest path between their WordNet types. Obviously, the goodness of chaining depends on the chaining resource. Chaining may ignore, e.g., words not covered by the operative reference system even if being central to the meaning of a text. On the other hand, co-occurrence networks may cover all types of a corpus by modelling corpus-specific word usages, but induce underspecified similarity judgements as they lack the type system of lexical reference systems. Thus, a framework for chaining is needed which integrates divergent resources and balances their deficits. In this paper we present such a framework for evaluating the impact of the operative chaining resource. This is done by means of a task-*independen*t evaluation of lexical chaining. It evaluates the impact of a chaining parameter (e.g. the chaining resource) subject to all other parameters (e.g. the chaining algorithm) being constant. Our starting points are texts dealing with a single topic (e.g. a newspaper article). For a corpus of such texts we estimate the expected value of chaining in terms of the size, coverage and structure of the chains being generated. The idea behind this approach is that a text is the better chained the more of its content units are covered and the more plausible the structure of the chains. We distinguish three aspects of this structural plausibility: chain topology, size and location. By evaluating these plausibility constraints we provide a model of the distribution of lexical chains subject to the variation of the operative chaining resource.

* Address correspondence to: Alexander Mehler, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Universitätsstr. 25, D-33615 Bielefeld. E-mail: Alexander.Mehler@uni-bielefeld.de.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Social Ontologies: Representation & Classification*

Alexander Mehler and Armin Wegner
Bielefeld University, Bielefeld

We investigate the principles of intertextual structure formation within large document networks which manifest the self-organization of web communities. More specifically, we focus on social ontologies by example of wikis. This is done by means of a comparative study which explores the differences of wiki-based social tagging. The paper tackles the following questions:

- *What is the least complex graph model that adequately describes wiki document networks in formal terms*?
- *What are topological characteristics of these networks*?
- *Can we reliably classify wikis in terms of their quantitative characteristics*?

To clarify the role of the communication function on the organization of social tagging we examine four different areas of communication. We argue that the investigation of intertextual structures by means of a quantitative approach is indispensable in order to discover the hidden order of social tagging. Following this line of research, we expand the research object in quantitative linguistics which traditionally relies on the analysis of units on the text level or on lower levels.

* Address correspondence to: Alexander Mehler, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Universitätsstr. 25, D-33615 Bielefeld. E-mail: Alexander.Mehler@uni-bielefeld.de.

Towards a Quantitative Model of Alignment in Communication*

Alexander Mehler, Petra Weiss, Olga Pustylnikov and Sara Maria Hellmann
Bielefeld University, Bielefeld

The approach to interactive alignment in communication (Pickering & Garrod, 2004) postulates two mechanisms of alignment:

- *priming* as a short-term mechanism of information percolation within the same or between different levels of representation
- and *routinization* as a long-term mechanism of expectation driven control of dialogue unfolding.

In this talk we present work in progress about building a quantitative model of alignment based on *lexical priming*. That is, we outline a framework of quantifying intrapersonal and interpersonal alignment of lexical units in dialogue. This framework is based on the theory of repetitions which postulates that texts can be characterized and reliably classified by the repetition structure of their constituents. We outline how this approach can be extended to model dialogical communication with its structuring by means of turns and multimodal signs (Rieser, 2007). More specifically, we present a model which focuses on the following questions:

- *How does a model of the repetition of alike elements in dialogue looks like?*
- *What does it mean to observe alignment in this setting?*
- *How to deal with multimodality and cross-modal signs?*

* Address correspondence to: Alexander Mehler, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Universitätsstr. 25, D-33615 Bielefeld. E-mail: Alexander.Mehler@uni-bielefeld.de.

Authorship Attribution Using Discriminant Function Analysis: Exploring Literary Stylistic Variation in Five Modern Greek Novels*

George K. Mikros
University of Athens, Athens

The aim of this paper is to explore quantitatively the literary style of Modern Greek novels and to conduct an authorship attribution experiment. The main research questions posed are:

- Is an arbitrary small portion of a novel, carrier of authorship information?
- Can the extraction of author-specific words (ASW) be used effectively in authorship attribution?

In order to answer the above research question we created a corpus of five Modern Greek novels written by four authors. Each novel was split in equal 200-word chunks and for each one we calculated a number of stylometric variables, which are briefly described below:

1) Lexical "richness" variables (Yule's K, TTR, Lexical Density, Ratio of hapax and dis-legomena, Lexical Frequency Profile).
2) Character level measures (Frequency of the letters and punctuation)
3) Word level measures (Average word length per text (in letters), Word length distribution, Part of Speech frequency and ratios)
4) Sentence level measures (Average length of sentences (in words), Standard deviation of sentence length per text)
5) Most frequent function words (FFW)
6) Most distinctive author-specific words (ASW)

The first four sets of variables (lexical "richness", word and sentence level measures and most frequent function words) have been extensively used in stylometric studies and their discriminatory power has been well documented. In addition, we will use the most frequent function words (FFW) of the corpus as well as the most distinctive author-specific words (ASW).

The latter method (ASW) has been applied previously in automatic text categorization (Mikros, 2003) and authorship attribution (Mikros, 2006) and has been proven superior to any other lexical selection method. It is based on frequency profiling and has already been used in English for different research purposes (Hofland & Johansson, 1982; Rayson et al., 1997; Granger & Rayson, 1998).

The resulting data will be analyzed using the multivariate statistical analysis Discriminant Function Analysis (DFA).

* Address correspondence to: George Mikros, Dept. of Italian and Spanish Language and Literature, Univ. of Athens, Panepistimioupoli Zografou, 15784, Athens, Greece. E-mail: gmikros@isll.uoa.gr.

Polish Flag-Words and Collective Symbols in Text Corpora. Quantitative Measures of Word Proximity*

Adam Pawłowski[1] and Maciej Piasecki[2]
[1]University of Wrocław; [2] Technical University of Wrocław

Flag-words and collective symbols are defined as words or expressions, which denotate or connotate high positive or negative values that can be put on flags or banners in a society (Pisarek, 2002, p. 7). They correspond to meanings and values particularly important for a given culture or community. It was hypothesised that f.-w. and c.s. could be extracted from text corpora by means of statistical methods. Two sets of f.-w. and c.s., created by Pisarek and Fleischer (2003), were applied as testing material. They were generated by surveys carried out on the representative groups of Polish population. From the anthropological viewpoint, these lexemes indicate the most fundamental values and categories existing in the human consciousness. Linguistically they can be associated with some "privileged" elements of the mental lexicon. The question arises how these internal elements are represented in the external language manifestations: we check whether high positions of lexemes in the m.l. of a representative group of respondents are correlated with some observable quantitative or systemic properties of words in text corpora.

The study follows a practical goal too. Surveys have always been laborious, expensive, and offered a limited insight into respondents' knowledge. Our objective is to investigate, whether the methods of the automatic extraction of semantic knowledge from text corpora lead to the same conclusions as surveys. If they do, some limitations of traditional methods, such as a real access to the population of respondents and high costs, would be overcome.

As the basis of analyses the IPI PAN Corpus of Polish was chosen (Przepiórkowski, 2004), as well as some other smaller corpora. The IPI PAN Corpus contains about one hundred million tokens and is now the largest annotated corpus of Polish. The following methods will be applied: the extraction of a *semantic relatedness measure* (SRM) and clustering methods. SRM assigns to every pair of lexical units a real number expressing a kind of semantic proximity. The extraction of SRM is based on the assumption of the so called *distributional hypothesis* (Harris, 1968), which says that lexical units occurring in similar contexts express similar meaning. In our approach, the contexts are described by lexico-morphosyntactic constraints and the distribution across contexts is represented by a coincidence matrix. An SRM is calculated on the basis of a transformed matrix according to some assumed distance measure. Clustering methods are used to identify sets of lexical units internally consistent in relation to the used SRM. The constructed SRMs are applied to the automatic identification of semantic profiles for flag-words. Among large number of lexical units which are semantically related to the given one, a coherent group is identified by clustering. We propose also a measure of semantic proximity of two lists of lexical units, which is based on the constructed SRM. The list proximity measure is used in the comparison of our results with the survey analysis presented in literature.

* Address correspondence to: Adam Pawłowski, Uniwersytet Wrocławski, Instytut Informacji Naukoweji Bibliotekoznawstwa, pl. Universytecki 9/13, 50-137 Wrocław, Polska. E-mail: apawlow@pwr.wroc.pl.

On Two Simplifications of the Japanese Writing System*

Haruko Sanada[1] and Gabriel Altmann[2]
[1]Saitama Gakuen University, Tokyo; [2]Lüdenscheid

It is well known that in Japanese there are three kinds of script, the ideographic kanji and the syllabic hiragana and katakana. Kanji has been introduced from China in the 5th century or earlier, and hiragana and katakana have been developed based on the cursive script of kanji, or made from a part of kanji or by reducing the strokes.

Script symbols have their properties which change in the course of time influenced by human requirements, especially that of the least effort (cf. Zipf, 1949), sufficient distinctiveness yielding redundancy, etc. Sanada and Yokoyama (2007) already mentioned that laws of the Synergetic Linguistics can be applied to Japanese kanji which have meanings and can behave like words. Our aim in this paper is to study the intricate ways of simplification and express them quantitatively. There are four possible ways of simplification and our aim is to find out which way has been chosen in Japanese:

(1) All signs are reduced to an approximately same complexity without regard to the original sign.
(2) Sticking to the complexity of the prototype: the more complex the prototype, the more complex the simplification.
(3) Simplification with a turning point: up to a certain complexity of the prototype the simplification follows way (2) but from a certain point, the more complex the prototype, the simpler the simplified forms. This third way of simplification has been observed in the change of hieroglyphs into hieratic script (cf. Hegenbarth-Reichardt & Altmann, 2007).
(4) Simplification with several turning points.

In this study we employ Altmann's method of measuring script complexity (2004). The measurement of complexity in Japanese is made difficult by the great number of different writing styles. Each font has its ornamentality (cf. Best & Altmann, 2007), a well developed creative activity in East-Asian calligraphy, and a slight movement of the brush (imitated in printing) can change the complexity of a sign.

The developments of complexity from kanji to hiragana pre-form and those from pre-forms to printed forms of hiragana accord with hypothesis (3), and those from printed to written form of hiragana accord with hypothesis (2). In the case from kanji to katakana the developments of complexity accord with hypothesis (3)

Though hiragana and katakana arose by different procedures, the general trend of script simplification seems to work. In both cases we obtain a concave curve with turning point signalizing that procedure (3) is not restricted to Egyptian hieroglyphs but has a more general validity. In any case we have shown that simplification is a procedure abiding by some subconscious control leading to analogous results in two quite different cultures.

* Address correspondence to: Haruko Sanada, Saitama Gakuen University, Saitama, Japan. E-mail: hsanada@iea.att.ne.jp

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Diversification in Icelandic and German Noun Inflection*

Petra Steiner
Universität Erfurt

Diversification has become established as a well-explained phenomenon in quantitative linguistics (Altmann, 1985; Rothe, 1991). Especially in the field of semantics, many investigations have been made (e.g. Altmann, 1985; Beőthy & Altmann, 1991; Nemcová, 2007). Diversification of inflectional morphology, however, is rarely taken into consideration.

This investigation deals with the different patterns of the inflectional paradigms for Icelandic and German nouns. Based on the paradigm lists of Kvaran (2005) and Steiner and Prün (2007), the values for the inflectional complexity of noun paradigms are calculated. Inflectional complexity is defined as the sum of all different inflectional affix types and the number of umlauts, ablauts or other allomorphs of noun stems. The number of inflectional paradigms of Icelandic nouns is higher than for German nouns, due to a variety of stem alternations produced by epenthesis, elision and ablaut-changes within some noun paradigms. This is connected with a higher inflectional complexity on the average of the former language.

A hypothesis concerning the distribution of this measure is derived from assumptions on diversification processes and tested on the frequency counts. Despite the differences in the frequencies and ranges, the two data samples can both be fitted to typical frequency distributions, showing the characteristics of diversification.

* Address correspondence to: Petra Steiner, Anglistische Sprachwissenschaft, Universität Erfurt, Nordhäuser Str. 63 99089 Erfurt. E-mail: petra.steiner@uni-erfurt.de

Power Laws and Other Heavy-Tailed Distributions and Associated Codes Related to Zipf's Law*

Flemming Topsøe
University of Copenhagen

The rank/frequency regularity in natural languages as expressed in 'Zipf's law" has puzzled the linguists and other scientists for many years. The attempt to model the law by considering an ideal person, the "Zipfean", who has an infinite vocabularium is bound to fail if the law is taken literally since that would involve a distribution with tails so heavy asto prevent normalization.

The suggestion to modify the modelling by considering power laws or their simple modifications is not convincing either as it renders no explanation of basic linguistic characteristica.

In a joint study with Peter Harremoës focus was on stability and flexibility of natural languages and it was argued that these two key features match a modelling with so called hyperbolic distributions (distributions not dominated by any power law). The good sense of this view will be explained in the talk and then an attempt to go further by modelling using special codes of the Zipfean's vocabularium will be initiated. Though primitive, it is attempted to model the learning process by describing a natural hierarchy of codes (using codes introduced previously by Elias and, independently, by Levenshtein).

* Address correspondence to: Flemming Topsøe, Dept. of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark. E-mail: topsoe@math.ku.dk.

5th Trier Symposium on Quantitative Linguistics, December 7-8, 2007

Efficiency of Flexible Parts-of-speech Systems*

Relja Vulanović
Kent State University, Ohio

Following Hengeveld, Rijkhoff and Siewierska (2005) and Vulanović (to appear), in this paper I continue my investigation of flexible parts-of-speech (PoS) systems. Speaking of languages that use four syntactic slots for the head and the modifier of each predicate and referential (i.e. noun) phrases, a PoS system is of type m (m = 1, 2, 3, 4) if it uses m lexeme classes to occupy those four slots. Flexible PoS systems are PoS systems of type 1-3. Since they have fewer than four lexeme classes, such PoS systems are prone to functional ambiguity which can be resolved by regulating word order and/or by relying on morphological markers. It is shown in (Vulanović, to appear) that the fixed word order suffices as a disambiguation device only in type 3 PoS systems; types 1 and 2 have to use morphology. This is now approached from the point of view of grammar efficiency (Vulanović, 2003). Efficiency of the three types of flexible systems is calculated under various assumptions regarding word-order rules and morphological markers. Maximally efficient structures are found and compared to 14 flexible languages in the linguistic sample used in Hengeveld, Rijkhoff and Siewierska (2005). It can be observed that grammar efficiency of natural languages is well below the theoretically possible maximum.

* Relja Vulanović, Dept. of Mathematical Sciences, Kent State University Stark Campus, 6000 Frank Ave. NW, North Canton, Ohio 44720, U.S.A. E-mail: rvulanov@kent.edu.

Logistic Regression Model of Preference and Familiarity in Letter Perception*

Shoichi Yokoyama
The National Institute for Japanese Language, Tokyo

Key words: letter perception, preference, familiarity, mere exposure effect, logistic regression model, Fechner's law, generalized matching law, corpus, maximum likelihood estimation

In Japanese, pairs of Kanji letters share the same meaning and pronunciation but exhibit varieties in their visual forms, for example "檜– 桧', they are called variants. Variants are commonly found in Japanese, they can often be characterized by a pairs of a "traditional" and a "simplified" forms. For example, a letter representing "cypress" can be transcribed with the traditional form "檜' and the simplified form "桧', both of which are pronounced the same.

The study discussed a kind of regression model to account for and to predict preference from familiarity in Kanji. We proposed that probability of which alternative of Kanji variants is preferred in 2-alternative forced-choice task can be decided with a linear function in mind expressed as follows:

$$Z = a\,(\mathrm{FamTrad} - \mathrm{FamSimp}) + b \qquad (1)$$

where FamTrad stands for familiarity of traditional variants and FamSimp for familiarity of the counterpart simplified variants. The study should link this linear function with logistic regression model. It is considered as an efficient way to explain phenomena that can be represented by 2-alternative, such as positive vs. negative. The logistic regression model is expressed as follows:

$$\log\{p1 / (1 - p1)\} = Z \qquad (2)$$

where Z is a linear function, the term p1 refers to the probability to choose Alternative 1, and 1–p1 describes the probability to choose Alternative 2 in forced-choice tasks.

$$\log\{p1 / (1 - p1)\} = a\,(\mathrm{FamTrad} - \mathrm{FamSimp}) + b, \qquad (3)$$

This study estimated the parameters by applying the method of maximum likelihood estimation to equation (3). In order to also examined whether regional differences are observed in the preference judgment task with character variants having been conducted in the two distinct regional areas, i.e. Tokyo and Osaka-Kyoto regions.

* Address correspondence to: Shoichi Yokoyama, The National Institute for Japanese Languages, Sociolinguistic Survey Group, 10-2 Midoricho, Tachikawa Tokyo 190-8561, Japan. E-mail: yokoyama@kokken.go.jp.

Multiple Logistic Regression Analysis for Formulating a Change in Language*

Yokoyama Shoichi[1], Sanada Haruko[2]
[1]The National Institute for Japanese Language; [2]Saitama Gakuen University

Many studies show that a process of the language change follows the S shape curve. There are few theoretical studies on the S shape curve of the language change whereas many empirical studies have been published.

This study proposes a new method to apply a multiple logistic regression model for dialectological data, so called glottogram, showing the process of analysis with virtual cases.

A multiple logistic regression model is given as

$$\log\{ p/(1-p) \} = Z, \tag{1}$$

where p is probability, Z is a linear combination shown in the form $Z = a1 \cdot X1 + a2 \cdot X2 + a3 \cdot X3 + \cdots + b$, and log is the logarithm to base e. The equation (1) can be transformed into equation (2) as

$$p = 1/\{1 + \exp(-Z)\}. \tag{2}$$

The multiple logistic regression model can be applied for an analysis of dialectological data considering the factors like age, the case of situations, and other factors denoting them as the variables $X1$, $X2$, $X3$, etc.

Employing this model, it is possible to analyze two or more factors of dialectological data including those with a nominal scale within one equation, to analyze data excluding disruptors, and to estimate future trend with a high precision even if observed data are not complete, a situation which often faces us. For investigations of written materials it is also possible to analyze a change of vocabulary size, a change of ratios of word types, or diverse language change with factors such as genre of the text or a gender of the author of the text.

We discuss a relation between the logistic regression model and a psychophysical model, and try to explain the fact that the language change follows an S shape curve by means of the "exposure relativity theory" (Yokoyama, 2006). This theory is to consider a change from an older word to a new word with the probability of success or failure which corresponds to the integration of the normal cumulative curve. It is well known that a logistic curve is very similar to the normal cumulative curve. The present study concludes that the ratio of the change from an older word to a new word theoretically behaves according to the logistic regression model given as (1), and the ratio of the change is also a function of the mere exposure effect. We can conclude that this psychophysical model includes human memory theory and it may be a powerful hypothesis to explain the background of the change in language which follows an S shape curve.

* Address correspondence to: Shoichi Yokoyama, The National Institute for Japanese Languages, Sociolinguistic Survey Group, 10-2 Midoricho, Tachikawa Tokyo 190-8561, Japan. E-mail: yokoyama@kokken.go.jp.

Quantitative Methods in Computational Dialectometry*

Thomas Zastrow
University of Tübingen

Computational dialectometry attempts to identify dialect regions with the help of mathematical, statistical or information theory related methods. The project Buldialects applies these methods to the Bulgarian language. This presentation will show two new methods which were developed at the University Tübingen: a geometrical approach in form of vector analysis and an adoption of the principal of entropy to dialectometry. In accordance to the results of traditional dialectology, these methods are viable to identify the main dialect regions in Bulgaria.The project Buldialects is a cooperation between the Rijksuniversiteit Groningen, the University Tübingen, the Bulgarian Academy of Science and the University Sofia. It is sponsored by the Volkswagen Stiftung.

Homepage:
http://www.sfs.uni-tuebingen.de/dialectometry/index.shtml

* Address correspondence to: post@thomas-zastrow.de