Issues in Quantitative Linguistics

edited by

Reinhard Köhler

2009
RAM-Verlag

Studies in quantitative linguistics

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.

ISBN: 978-3-9802659-9-7

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