Methods and Applications of Quantitative Linguistics

Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012

Editors:
Ivan Obradović, Emmerich Kelih, Reinhard Köhler

Editorial assistance:
Sheila Embleton, Hermann Moisl

Technical editor:
Milan Radonjić

Printed with the support of the Government of the Republic of Serbia Ministry of Education, Science and Technological Development
QUALICO April 26-29, 2012, Belgrade, Serbia

Programme Committee:
Peter Grzybek (Graz), Emmerich Kelih (Graz), Reinhard Köhler (Trier),
Cvetana Krstev (Belgrade), Ján Mačutek (Vienna), George Mikros (Athens),
Ivan Obradović (Belgrade), Duško Vitas (Belgrade)

Organising Committee:
Cvetana Krstev, Ivan Obradović, Duško Vitas

Publishers
UNIVERSITY OF BELGRADE
ACADEMIC MIND, Belgrade, Serbia

Printed in Serbia by
ACADEMIC MIND, Belgrade

Circulation
80 copies

ISBN 978-86-7466-465-0
Previous QUALICO Proceedings:


Preface

The contributions to this volume are selected and peer-reviewed papers on the basis of talks which were delivered on the occasion of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia. The first QUALICO was held in Trier, Germany in 1991, the second one was organised in Moscow in the year 1994 by colleagues from the Faculty of Humanities of the Lomonosov University in Moscow and the department of Computational Linguistics at Trier University, during which the International Quantitative Linguistics Association (IQLA) was founded. Follow-up conferences have been organised regularly since 1997 by IQLA (QUALICO 1997 in Helsinki, Finland, QUALICO 2000 in Prague, Czech Republic; QUALICO 2003 Athens, Georgia USA, 5th Trier Symposium on QL 2007 and QUALICO 2009 in Graz, Austria).

QUALICO 2012 was, as usual, open to original papers without any limitations as to thematic or methodological aspects. Therefore, a large variety of topics has been addressed and numerous fields of linguistics and related fields have been touched. A number of contributions to the conference can be subsumed under the general problem of stylometrics, in particular authorship attribution or text classification with respect to stylistic features unveiling the authors’ membership in social or other groups, among them gender, age, or literary tradition. In this volume, George Mikros presents a stylometric approach to the analysis of the language of blogs, where authorship attribution and gender determination are the main topics. Another paper in this group is D. Benedetto’s, M. Degli Esposti’s, and G. Maspero’s report on a specific philological problem of disputed authorship. The approach presented is a mathematical method using an estimation of entropy.

Fundamental linguistic research is focussed on by another group of contributions. Beliankou, Köhler, and Naumann introduce us into a new field of quantitative-linguistic research by giving an example of a study on the analysis of the discourse-pragmatic problem of argumentation structures. Ján Mačutek and Radek Čech investigate the diversification of the cases of Czech nouns; their results support earlier investigations into comparable matters. Furthermore, interesting facts are found and reported which shed new light on frequency effects in morphology. On the background of the Menzerath-Altmann Law, Peter Grzybek offers results of his research on the relations between sentence length and the lengths of other linguistic levels, which are not directly adjacent to the sentence level. A new mathematical approach to Polikarpov’s and Krylov’s “word life cycle” hypothesis is presented by Vasilii Poddubnyy and Anatoliy Polikarpov. They describe a dynamic model of the evolution of linguistic signs based on dissipative processes and provide the empirical data on which the model was tested. Matilde Trevisiani’s and Arjuna Tuzzi’s paper
presents a mathematical masterpiece as an attempt at unveiling chronological patterns in the frequency of words in a text corpus. The specificity of this study is that the mathematical model tries to capture not only general tendencies, which could be motivated by linguistic or philological considerations, but at the same time the natural fluctuations and statistical deviations as a part of the deterministic model. Makoto Yamazaki delivers one of the so far very few quantitative works on text cohesion. His method is based on finding lexically identical elements in sequences of text passages.

The following papers can be grouped under the label application-oriented: Gordana Đurđaš, Ernst Stadlober and Emmerich Keliš focus their study on the analysis of the distribution of word length in Slovenian and Russian texts of different type. Within a text classificatory context, they propose generalisations of the Poisson distribution and compare several commonly used estimators. Ivan Obradović presents a method for extracting semantically related word pairs from aligned texts using frequency based-ranking. The results are promising with respect to computational-linguistic applications and may be of importance to automatic translation systems. Japanese texts are analysed by Haruko Sanada with respect to their thematic concentration using Altmann’s and Popescu’s TC measure, which is based on the H index as introduced into linguistics by Ioan-Iovitz Popescu. The results suggest that thematic concentration is heavily influenced by text type.

Another group consists of contributions with methodological topics. Łukasz Dębowski tests Hilberg’s conjecture for a selection of English prose using the Lempel-Ziv algorithm and finds an upper bound for the exponent. This contribution is interesting mainly from a mathematical point of view and may have implications in informatics. String similarity and the comparison of methods to find cognates is dealt with by Antonella Delmestri and Liviu P. Dinu. They suggest that orthographic learning methods may accurately detect traces of sound changes left in the orthography and outperform static phonetic systems. Jiří Milička delivers a paper on the mathematical relation between TTR and rank-frequency distribution. He describes a method to predict the development of the TTR curve for large corpora. Non-linearity is the focus of Hermann Moisl’s paper. He demonstrates the nature of data nonlinearity, reviews existing methods for detection of nonlinearity and proposes a way of measuring nonlinear relationships between data objects. Relja Vulanović extends the Piotrowski-Altmann Law to several dimensions by generalizing the logistic differential equation to a system of partial differential equations. The application of multi-dimensional variants of the law is exemplified on data from cross-linguistic studies of parts-of-speech systems.

All in all the selected papers provide a broad overview on current research topics in quantitative linguistics and related areas.

The editors, July 2013
CONTENTS

PREFACE 4

I STYLOMETRICS 9

Dario Benedetto, Mirko Degli Esposti, Giulio Maspero
Authorship Attribution and Small Scales Analysis Applied to a Real Philological Problem in Greek Patristics 11

George Mikros
Authorship Attribution and Gender Identification in Greek Blogs 21

II FUNDAMENTAL LINGUISTIC RESEARCH 33

Andrei Beliankou, Reinhard Köhler, Sven Naumann
Quantitative Properties of Argumentation Motifs 35

Peter Grzybek
Close and Distant Relatives of the Sentence: Some Results from Russian 44

Ján Mačutek, Radek Čech
Frequency and Declensional Morphology of Czech Nouns 59

Vasiliy Poddubnyy, Anatoliy Polikarpov
Stochastic Dynamic Model of Evolution of Language Signs Ensembles 69

Matilde Trevisani, Arjuna Tuzzi
Shaping the history of words 84

Makoto Yamazaki
Quantitative Analysis of Text Structure Using Co-occurrence of Words 96

III APPLICATION-ORIENTED RESEARCH 105

Gordana Duraš, Ernst Stadlober, Emmerich Kelih
The Generalized Poisson Distributions as Models of Word Length Frequencies 107

Ivan Obradović
A Method for Extracting Translational Equivalents from Aligned Texts 119

Haruko Sanada
Thematic concentration in Japanese prose 130

IV METHODOLOGICAL ISSUES 141

Łukasz Dębowski
Empirical Evidence for Hilberg’s Conjecture in Single-Author Texts 143
Contents

Antonella Delmestri, Liviu P. Dinu
An Assessment of String Similarity Methods for Cognate Identification 152

Jiří Milička
Rank-frequency Relation & Type-token Relation: Two Sides of the Same Coin 163

Hermann Moisl
Measurement of nonlinear distance in data derived from linguistic corpora 172

Relja Vulanović
A Multidimensional Generalization of the Piotrowski-Altmann Law 184