Studies in Quantitative Linguistics

Ioan-Iovitz Popescu
Ján Mačutek
Emmerich Kelih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

Vectors and Codes of Text

RAM - Verlag
Vectors and codes of text

by

Ioan-Iovitz Popescu

in cooperation with

Ján Mačutek
Emmerich Kélih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

2010
RAM-Verlag
Studies in quantitative linguistics

Editors
Fengxiang Fan  (fanfengxiang@yahoo.com)
Emmerich Kelih  (emmerich.kelih@uni-graz.at)
Reinhard Köhler  (koehler@uni-trier.de)
Ján Mačutek  (jmacutek@yahoo.com)
Eric S. Wheeler  (wheeler@ericwheeler.ca)


Gedruckt mit Unterstützung der Karl-Franzens-Universität Graz

ISBN: 978-3-942303-02-6

© Copyright 2010 by RAM-Verlag, D-58515 Lüdenscheid

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

The present book is a continuation of our endeavour to introduce in textology new quantitative methods and evaluate some older ones (cf. Popescu et. al. 2009, Popescu, Mačutek, Altmann 2009; Tuzzi, Popescu, Altmann 2010). We illustrated the measurements and performed evaluations of texts from many languages. Needless to say, all results ensuing from the use of vectors, codes and chains must be tested on further languages and texts. Since this is ongoing empirical research, some modifications and adaptations of the methods presented may be necessary.

Nevertheless, the more we advanced the clearer we saw the abysmal playground hidden in texts. With some progress in using and elaborating quantitative methods in text analysis some new problems and formerly unrecognized phenomena appeared, thus we were confronted not only with methodological challenges, but also with new questions and problems in linguistic text theory in general.

We restricted ourselves to formal features (frequencies, codes and chains) accessible to the cooperating linguists and avoided sociolinguistic, psycholinguistic and other problems. We nevertheless hope that the methods presented could be made useful for other investigations, too.

The book consists of nine chapters. In Chapter 1 we introduce briefly the extensive domain of possible problems concerning comparisons and research strategies. In Chapters 2 to 6 we examine different vectors of texts, show their behaviour, compare texts and languages and take a step towards capturing the text dynamics looking at it from different points of view.

In Chapters 7 and 8 we goedelize the text in one special way, show breaks in the syntactic continuity in the text and ascribe to it its binary code which can be compared and tested.

The last chapter, Chapter 9, is devoted to chaining phenomena restricted here to Belza-Skorochdoľko chains, revealing many new vistas touching behaviour, perseveration, psycholinguistic and other aspects. As a matter of fact, each topic could be developed infinitely but we strived for presenting simple methods, developed tests and showed a way of ternary plotting.

We hope that other scholars will adopt the methods for different purposes and for analyzing other languages, in order to get stronger corroboration of the procedures presented.

In this place we want to express our gratitude to Claudiu Vasilescu, who patiently wrote for us dilettantes all the Excel programmes and discreetly concealed his amazement about our naivety. We had suffered much more without his kind help. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach).

I.-I. Popescu
## Contents

Preface

1. Introduction 1

2. The adjusted modulus 3
   2.1. German data 4
   2.2. Italian data 12
   2.3. Slavic data 15
   2.4. General data 22

3. The vector $T$ 26
   3.1. Stepwise and retrospective dissimilarity 26
   3.2. Dispersion 38
   3.3. Randomness 40
   3.4. Prospective dissimilarity 42

4. Vectorial method of text comparison 53
   4.1. Comparisons of texts 53
   4.2. Cross-linguistic comparison 55
   4.3. Vector distance 69

5. The ternary plot 76

6. Further simple methods for measuring the dynamics 95

7. The binary code of sentence 100
   7.1. Goedelization 100
   2.2. Breaks in the sequence 120

8. The binary code of text 124
   8.1. The classical method 124
   8.2. Other methods 127
   8.3. Using the binary code 130

9. Belza-Skorochod'ko chaining 135

10. Conclusions 146

Appendix I. Texts used 147
References 156
Author index 159
Subject index 160