The Lambda-structure of Texts

by

Ioan-Iovitz Popescu
Radek Čech
Gabriel Altmann

2011
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)

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.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp

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

Preface

The problem of frequency structuring of a text is not only very old but it has at present a great number of different aspects. The most popular ones are the studies of vocabulary richness, type-token ratio, rank-frequency distributions and the frequency spectrum, to mention only some of them. Vocabulary richness is a central property of the text useful in the study of language learning, in forensic linguistics, in style studies, in the literary development of a writer etc. Many researchers tried to find a relationship between the number of types and that of tokens (text size) but even if sometimes they succeeded to stabilize the relation, in the formula a variable remained whose sampling distribution was not known. What is the expectation of V (vocabulary) and the text size (N)? And even if text size can sometimes be determined in advance (e.g. for a press article), the vocabulary cannot. How can the standard deviation of the vocabulary be derived? The answer has never been given and nobody has tried to solve the problem, not even empirically.

The present study shows that if we descend a level deeper, viz. from the vocabulary as a whole to its components, i.e. words and their frequencies, a stable indicator (called lambda) of frequency structure (*cum grano salis* the basis of vocabulary richness) can be set up which does not depend on text size (N) and whose variance can be asymptotically derived. This fact enables us to set up tests for comparing individual texts, individual authors, genres, and languages, to follow the deployment of a text and the evolution of a writer through years. It allows us to study the jumps in the individual chapters/parts of a text and to express quantitatively different aspects of text dynamics.

Needless to say, even if we exemplify the study using 1185 texts in 35 languages, the research is not finished. On the contrary, many more texts must be analyzed, new aspect should be discovered and for every aspect test procedures must be devised. Further, frequency structure is not an isolated property. It is associated with other different properties, but first such a connection must be hypothesized and the other properties must be quantified, too, before we begin to set up hypotheses. It can be conjectured that frequency structure is also an element of Köhler´s control cycle but the way to show it will be very long.

Acknowledgements

We are very much obliged to Fazli Can, Fengxiang Fan, Emmerich Kelih, Viktor Krupa, Ján Mačutek, Haruko Sanada, Claudiu Vasilescu and Eric Wheeler who in different ways helped us to prepare this volume. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157.

Contents