Studies
in Quantitative Linguistics
10

Ioan-Iovitz Popescu
Radek Čech
Gabriel Altmann

The Lambda-structure of Texts

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)


© Copyright 2011 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 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, forensic linguistics, style studies, the literary development of a writer etc. Many researchers tried to find a relationship between the number of types and that of tokens (text length), but even if sometimes they succeeded to determine the relation, in the formula a variable remained whose sampling distribution was not known. What is the expected value of $V$ (vocabulary size) and the text length $(N)$? And even if text length can sometimes be determined in advance (e.g. for a press article), the vocabulary size cannot. How can the standard deviation of the vocabulary size 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 $N$ 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. 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

Preface

1. Introduction 1

2. Data 10

3. Comparison of texts 16
   3.1. Individual comparisons 16
   3.2. Group comparisons 20
   3.3. Characterizing groups 26

4. Comparison of authors 32

5. Comparison of genres 42

6. Comparison of languages 49

7. Text development 58
   7.1. Change of lambda 58
     7.1.1. Difference in the height 60
     7.1.2. Difference in the profile 62
   7.2. Mean sequential difference 63
   7.3. Runs 68
   7.4. Length of phases 72
   7.5. Length of runs 73
   7.6. Uniformity 74

8. Historical development 78
   8.1. Development in a language 78
   8.2. Development of a writer 85

9. Child language development 93

10. Pathological texts 98

11. Conclusions 102

References 106
Appendix 110
Author index 179
Subject index 180