Aspects of Word Frequencies

by

Ioan-Iovitz Popescu
Ján Mačutek
Gabriel Altmann

2009
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

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.

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

Preface

During the preparation, layouting and printing the book „Word frequency studies“ (2009)[1] a great number of new ideas about texts arose which could not be inserted any more in the above book. They appeared in form of articles dispersed in different journals and omnibus volumes and touched a very variegated palette of problems. We try to collect them and show the connections between them if there are any. Besides, we shall try to develop some of the ideas a step further. Frequently we shall take recourse to the above mentioned book whose knowledge is, however, not presupposed. If necessary, the pertinent object will be explained.

The booklet can be used as a collection of lectures in textology for a seminary and can be managed in one semester, even without a teacher. At the same time, the methods presented in both books can be used for text mining.

Since the individual chapters are heterogeneous developments of different issues, there is sometimes no logical nexus between the subsequent chapters. This is caused also by the fact that textology is no closed discipline and develops very quickly in different directions. It extends especially to the study of modern forms of texts, namely SMS, SPAM, E-mail and Internet pages, all of which display some divergent properties brought about by the conditions of the medium and the purpose. We restrict ourselves to literary texts but the methods can be applied to these special texts mutatis mutandis.

In many chapters we try to show the way from text to language typology, text being the surface where one can find the reflections of language structure. Needless to say, this is only the beginning of an enterprise which can be developed more extensively. Combining the properties and processes in the deep layers of language and on the surface represented by texts one will perhaps be able to construct some time a theory encompassing both. It will not have an algebraic structure, it will not concern grammatical rules and it will not be deterministic. It can turn out to become anything else but it will not be able to avoid probability, the basis of communication.

We want to express our gratitude to all those who took part in the sampling of texts for the above mentioned book and whose results are used in this book, too, namely P. Grzybek, B.D. Jayaram, R. Köhler, V. Krupa, R. Pustet, L. Uhlířová, and M.N. Vidya.

In the first place we want to thank Fengxiang Fan for his thorough reading of the book and correcting our Middle-European English. All remaining errors were made after his reading the book.

I.-I.P., J.M., G.A.

---

[1] Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Contents

Preface

1. On text theory 1
2. On sampling and homogeneity 8
3. A new view of Zipf´s law 13
4. The h-point 24
5. Arc length 49
   - 5.1. Arc length and associated typological indicators 49
   - 5.2. Arc development 64
   - 5.3. Arc length as a function of text indicators 68
   - 5.4. Analysis of language levels 70
   - 5.5. Conclusions on language levels 90
6. Hapax legomena 99
7. Further typological considerations 111
8. Diversity of word frequency and typology 157
9. Nominal style 171
   - 9.1. Static approach 171
   - 9.2. Dynamic approach 174
   - 9.3. Prospects 177

Appendix 179

References 186

Author index 192

Subject index 194