

Problems in Quantitative Linguistics

2

by

**Reinhard Köhler
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.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2*. 2009, VII +
.....

ISBN: 978-3-9802659-5-9

© Copyright 2009 by RAM-Verlag, D-58515 Lüdenscheid

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

Preface

*„It is not just that research begins with problems:
research consists in dealing with problems all the way long.”*
(Mario Bunge, *Philosophy of science. Vol. 1: From problem to theory.*
New Brunswick, London: Transaction Publishers, 2007, p. 187)

Finding a scientific problem is the first task of a young scientist. Solving it is the next one. A solution, however, does not finish a problem; on the contrary, every solution opens up a series of new problems. Thus, from time to time it would be useful for every scientific discipline to resume the topical problems, show some new ones and shed light on other aspects of old problems.

We present a collection of problems in the field of quantitative linguistics – as far as it is possible to find Ariadne’s thread in the jungle of its differently developed sub-disciplines. The whole field consists of *membra disiecta* and we try without too much violence to draw the reader’s attention to the way of unification, where theory building may begin. Today, it is not easy to imagine that in an empirical science a theory might arise without at least elementary quantification. Though in the problems presented here there is still a lot of qualitative work to be done, we try to convince the reader to form quantitative concepts, to strive for elementary quantitative solutions, to link some problems with some existing theories or to open a new field of research.

In the first volume of this series the authors presented problems concerning phonemics, script, grammar, lexicology, textology, semantics, synergetics, psycholinguistics, typology, different general problems and the relations of length and frequency to other properties. In the present volume, most of the above-mentioned domains are treated, too, but besides, a number of problems concerning pragmatics, proverbs, drama, philosophy of science, motifs, dialectology etc. are added.

If the reader decides to solve one of the problems, it is recommended to look first in “Problems Vol. 1” where a more elementary, preparatory problem concerning the same domain may be presented. If a problem has been successfully solved, one should always try to generalize it, to test the result on data from several languages or texts, to seek deviations, outliers, to enrich it with subsidiary conditions and to systematize it, i.e. to embed it in a more general framework from which it can be derived.

If one meets “hard” problems, the first step may be purely inductive, e.g. fitting a simple function to data mechanically, but in the next step, the tentatively tested function should be substantiated as to the question “why should this function be chosen?” which is nearer to a future explanation than a verbal description of the discovered phenomenon.

II

The problems presented here vary from classroom exercises in quantitative linguistics over take-off platforms for publications to themes for research projects.

Readers are invited to report on publications which departed from a problem in this collection or in the first volume to the editor of the *Journal of Quantitative Linguistics* (<http://www.ldv.uni-trier.de/index.php?koehler>) or the editor of *Glottometrics* (www.gabrielaltmann.de) Solutions to one of the problems may also be submitted for publication in one of these journals.

Readers are also invited to contribute more new problems by sending a description to one of the above-given addresses.

R.K., G.A

Contents

Preface	I
1. Phonology and script	1
1.1. Zipf's assimilation	1
1.2. Zipf's accent problem	1
1.3. Script distinctiveness	2
1.4. Entropy of script system distinctiveness	3
1.5. Script complexity 2	4
1.6. Canonical speech segments	5
1.7. Phonetic comparison of cognate languages	7
1.8. Phonetic word structure	8
1.9. Phonetic distortion of borrowings	9
2. Grammar	13
2.1. Fenk's hypothesis	13
2.2. Zipf's adverb hypothesis (1)	14
2.3. Zipf's adverb hypothesis (2)	14
2.4. Auxiliary words	15
2.5. Valency and text frequency	16
2.6. Valency and rank-order	16
2.7. Case diversification in Ugro-Finnic languages	17
2.8. Valency and compounding	18
2.9. Valency and derivation	19
2.10. Valency and synonymy	20
2.11. Valency and length	21
2.12. The control cycle of valency	21
2.13. Valency of nouns and adjectives	22
2.14. Valency: the distribution of variants	22
2.15. Valency and complementation patterns	23
2.16. Distribution of the semantic subcategories of arguments	24
2.17. Number of arguments and number of semantic subcategories	24
2.18. Frequency and allomorphy	25
2.19. Semantic relevance of affixes (1)	25
2.20. Semantic relevance of affixes (2)	26
2.21. Word order and topic assignment	27
2.22. Syntactic properties	28
2.23. Efficiency of the P-O-S system	28
2.24. Length and complexity of syntactic structures	29
2.25. Grammar, text, corpus, language	29

2.26. Functional dependences in syntax	30
2.27. Distribution of complexity	31
2.28. Information structure (1)	31
2.29. Information structure (2)	32
2.30. Diversification of the aspect	33
2.31. Case control	34
3. Semantics	36
3.1. Verb and noun polysemy	36
3.2. Polysemy of parts-of-speech	37
3.3. Synonymy and morphological productivity	38
3.4. Synonymy and postpositional phrases	38
3.5. Semantic partitioning of space	39
3.6. Synonymy and the morphological status of the word	39
3.7. Word senses (1)	40
3.8. Word senses (2)	41
3.9. Distribution of word synonymy	41
3.10. Synonymy and polysemy	42
3.11. Synonymy, length and frequency of words	43
4. Lexicology	44
4.1. Definition chains (verbs and adjectives)	44
4.2. Survival of word classes	45
4.3. Frequency and survival of words	46
4.4. Word class distributions 2	47
4.5. Vocabulary comparisons	50
4.6. Word commonness	51
4.7. Indicator of association	52
4.8. Word stability	53
4.9. Word length and meaning generality	55
5. Textology	57
5.1. Belza-Skorochod'ko's chaining coefficient	57
5.2. Crowding of autosemantics	59
5.3. Semantic reduction in texts	60
5.4. Rank-frequency distribution and arc length	61
5.5. Popescu's vocabulary richness	62
5.6. Alliteration	63
5.7. Alliteration structure	64
5.8. Autosemantic dissortativity	65
5.9. Superhreb	66

5.10. Golden section (1)	66
5.11. Strange attractor of writer's view	67
5.12. Aristotle's Categories	68
5.13. The Skinner effect	69
5.14. The <I,J> scheme	69
5.15. Text cohesion (1)	71
5.16. Text cohesion (2)	72
5.17. Text cohesion (3)	73
5.18. Hapax legomena and Markov chains	75
5.19. The frequency sequence of words	76
5.20. Golden section 2	76
6. Typology and universals	78
6.1. Arc length and typology	78
6.2. Length of morphs	78
6.3. Diversification constant	80
6.4. Synthetism – analytism	81
6.5. Methodological problems	83
6.6. Word order (1)	84
6.7. Word order (2)	85
6.8. Phoneme sequences	85
6.9. Saporta's consonant sequences	86
6.10. Word frequency and analytism	87
7. Synergetics	89
7.1. Frequency and polytextuality	89
7.2. Polysemy and polytextuality	90
7.3. Morph length and phoneme inventory	91
7.4. Frequency and polysemy	92
7.5. Diversification distribution	93
7.6. System boundaries and interactions	94
7.7. Language and text	95
7.8. Frequency and age	96
7.9. Word length and age	96
7.10. Valency and polysemy	97
7.11. Complement to synergetic problems	97
7.12. Phonotactics: exploitation of linguistic material	99
7.13. Word length and polysemy in Chinese	100
7.14. Length and frequency of affixes	101

8. Philosophy of science and general problems	102
8.1. Degree of constituency	102
8.2. Exercises in philosophy of science	103
8.2.1. Concept	103
8.2.2. Problem	104
8.3. Rank-frequency, a general approach	106
8.4. Universals, laws and theories	107
8.5. Observability	108
9. Different issues	109
9.1. Arc length and language evolution	109
9.2. Politeness	109
9.3. Word class distribution in proverbs	110
9.4. Köhler motives in proverbs	111
9.5. Semantic roles in proverbs	112
9.6. Number and length in proverbs	112
9.7. Sentence structures in proverbs	113
9.8. The recognition of variants in phraseological elements	113
9.9. Synonymy and impoliteness	114
9.10. Death process in dialectology	114
9.11. Length motives	115
9.12. Frequency and production effort (continuation)	116
9.13. Fourier analysis	117
10. Pragmatics	119
10.1. Frequency distribution of speech acts	119
10.2. Homogeneity, similarity and hierarchy of persons	121
10.3. Distances between equal acts	122
10.4. Scaling of speech acts	123
10.5. Distribution of scaled values of speech acts	124
10.6. Weight motives	124
10.7. Drama as a time series of speech acts	125
10.8. Some properties of speech acts sequences	126
10.9. Drama and comedy	126
10.10. The development of drama	127
10.11. Speech act herbs	127
10.12. Towards a theory of speech acts	128
10.13. Length of dialogue contributions	128
10.14. Discourse frequency (1)	129
10.15. Discourse frequency (2)	130
10.16. Discourse frequency (3)	131

10.17. Rhetorical structure (1)	132
10.18. Rhetorical structure (2)	132
10.19. Rhetorical structure (3)	133
10.20. Rhetorical structure (4)	133

Author index

Subject index