Problems in Quantitative Linguistics 1

by

Udo Strauss
Fengxiang Fan
Gabriel Altmann

Second edition

2008
RAM-Verlag
Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Peter Grzybek (grzybek@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)


ISBN: 978-3-9802659-4-2

© Copyright 2008 by RAM-Verlag, D-58515 Lüdenscheid

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

This book is the first volume in the series “Problems in Quantitative Linguistics”, which presents selected proposals for research, problems, questions, hypotheses, and exercises taken from various quantitative-linguistic fields. Only very few of the issues presented here have been studied in previous investigations; each of them is of serious scientific interest and can lead to findings, which may contribute to the construction of a complex linguistic theory.

The problems are of different degree of difficulty and cause different effort if tackled. Many of them can help students in the choice of themes for theses, academic teachers in finding appropriate exercises and examples for their courses, or researchers looking for new enterprises. Most of the hypotheses afford an opportunity to form an original contribution to one of the QL fields by finding a first answer to a given question, a solution to problem, a new method or approach, or an application of existing ones to new linguistic data.

The great majority of the problems concern interrelations between two or more linguistic entities. The reader is asked to set up exact definitions, quantifications and measurement methods, to collect data, perform tests, find an empirical function or derive a function from theoretical assumptions; a complete solution, however, is not always required. In the few cases where a solution or method can be found in the references the reader should feel encouraged to test it on data from other languages, text types, dictionaries etc. or to find an alternative solution.

The individual problems are presented in a unified form throughout the book as follows: (1) A hypothesis or a problem is given together with sources that should be read. These sources often provide preliminary analyses of the problem and further references. (2) A procedure is proposed with suggestions for the appropriate steps in the analysis. Sometimes, an in-depth analysis of the presented problem is given. (3) References are provided where the interested reader can find the first mention or a deeper analysis of the problem. A corresponding remark indicates if a reference is mandatory before a problem can be approached.

The instructions given with the problems do not always contain ready-made formulas; in these cases, the reader is referred to the references or to statistics text books.

The following general recommendations may help with a successful work:

1. Linguistic examples cannot be considered as evidence of a phenomenon, pattern, trend or law. The only appropriate empirical basis consists of data from complete objects (e.g. texts) or random samples.
2. A correlation analysis is not acceptable as a result; the same is true of a simple test of differences between objects. You should rather find at least an empirical function.

3. English or German are fine but we recommend enriching your study by at least one other language.

4. Empirical findings are often prematurely generalised. Corresponding empirical statements should be tested on several languages, text types, authors etc. depending on the kind of hypothesis.

5. Concepts, quantifications and measurements must be defined in an absolutely explicit and unequivocal way. Avoid concepts you cannot operationalise with sufficient exactness.

6. Always try a derivation of the function or distribution you assume for your data from reasonable theoretical assumptions. Often, proportionality considerations may be successful as a number of hypotheses in synergetic linguistics have shown.

7. If a function or distribution seems inadequate with respect to your data, re-check your data (sources, pre-processing, amount, artificial factors etc.), calculation, computational procedures – and your assumptions. Change or correct whatever turns out to be wrong and try once more.

8. If your mathematical model fails again: sometimes, there are some boundary conditions which affect a relation (although we think that the law of gravitation is valid we observe that some objects, e.g. birds, do not drop). Find such boundary conditions in your case and consider them as independent variables. Re-formulate your hypothesis correspondingly and start again.

9. No hypothesis should be definitively rejected or definitively accepted. Corroboration is a matter of degree.

10. To clarify your thoughts, work out a diagram of the relationship including parameters and requirements (cf. the notation in synergetic linguistics).

11. Keep in mind that data are constructs, i.e. to some extent artificial. Data collection consists in transforming facts via hypotheses (or a weaker form of assumptions or expectations) into statements. Hence, one should first set up an explicit and plausible hypothesis – then search for data.

12. If it is difficult to determine which variable is dependent and which is independent, try to integrate both variants in a larger control cycle or at least test both directions.

13. After solving several problems try to integrate all of them in a control cycle. Fill the missing vertices and edges by hypothetical ones and try to find them empirically.

14. Never give basic data in the form of percentages; always present absolute numbers.

15. When a problem is solved, do not consider it the final solution; see it as part of a greater perspective and try to describe this perspective.
16. If you think you need a classification do not just classify mechanically using a method at hand. Instead, try to set up a theory and deduce an appropriate classification from this theory.

17. Do not use functions with many parameters (e.g. polynomials) because later on these parameters will have to be interpreted (i.e. adhere to “Occam’s razor”).

18. If possible, as linguist, cooperate with a programmer and a mathematician. If you are mathematician you should seek an experienced linguist, otherwise a good mathematical model may be developed – however without linguistic interpretation and hence without use.

19. Try to apply solved problems introduced in this book using new data (from other languages) so that existing theories can be corroborated or rejected.

20. Do not consider linguistic units as given a priori. Define units operationally in such a way that they can be used in hypotheses, even if their segmentation might seem somewhat artificial. Keep in mind that those linguistic units are theoretically prolific which can be used in formulating laws (not in grammatical rules).

21. Always prefer functions or distributions with a good theoretical foundation to ones which possibly displays a better fit but have no linguistic background, i.e. use empirical functions only at the beginning of a research.

22. There are nine chapters in this book. The contents of the individual chapters are not strictly homogeneous but furnish a relatively broad view of possible problems that can be solved using quantitative methods. Within each chapter, the problems are arranged alphabetically. Some problems have been analysed in more detail. Neither the chapters nor the problems need to be read successively; one can choose a problem according to one’s own preference and specialisation.

Acknowledment
We are very obliged to Reinhard Köhler who patiently read the whole book, corrected our style and our Pidgin, improved some argumentations and gave us a number of useful hints. The rest consists of our besetting sins.
Contents

Introduction 1

Chapter 1. Phonemics and script 2
  Accent and frequency 1
  Canonical word structure 1
  Consonants and clusters 2
  Distribution of canonical forms 2
  Distributional calculus 3
  Distributional gaps 4
  Evolution of script complexity 4
  Exploitation of canonical forms 5
  Letter frequency 5
  Measurement of distinctness 6
  Measurement of ornamentality 6
  Phoneme frequency and word frequency 7
  Phoneme inventory and word length 7
  Power law 8
  Ranking syllable types 9
  Script complexity 9
  Script simplification 10
  Syllable frequency 10
  Syllable structure 11
  Tendency towards vowel harmony 13
  Two-dimensional syllable structure 14
  Word length and supra-segmentals 15

Chapter 2. Grammar 16
  Behagel’s “law” 16
  Co-occurrence and cohesion 17
  Cotextuality and variation 18
  Frequency and case 18
  Frequency and cohesion 19
  Frequency and derivation 19
  Frequency and irregularity 20
  Frequency and valency 21
  Frequency of sentence patterns 22
  Grammaticalisation 22
  Morph frequency 22
  Morpheme polysemy and morpheme frequency 23
  Morphological productivity of stems 23
VI

Sequential word class frequency 24
Verb classification 25
Verbs and persons 25
Word class distributions 27

**Chapter 3. Compounds and lexicology** 28

Age and compounding propensity 28
Collocations 28
Compound length and component length 29
Compound length and compound cotextuality 30
Compound length and polysemy 30
Compound length and semantic correspondence 31
Compounds and semantic correspondence 31
Compound forming and associations 32
Compound forming and emotionality 32
Cotextuality and compounding propensity 33
Dissortativity of compounding 33
Distribution of compound length 34
Distribution of synonyms 34
Increase of loan words 35
Lexical chains 36
Lexical networks 37
Stem length and compounding propensity 38
Word length and synonymy 38

**Chapter 4. Textology** 40

The association graph of a text 40
Autosemantic pace filling 40
Carroll’s vector 41
Constraint measure for text 42
Cotextuality and frequency 42
Distances between equally long sentences 43
Distances between lexemes 43
Euphony 44
Hirsch-Popescu-point problems 45
Hrebs 46
Hurst’s exponent 47
Köhler’s word length motives 1 48
Köhler’s word length motives 2 49
Köhler’s word length motives 3 50
Köhler’s sentence length motives 51
Lorenz curve 51
<table>
<thead>
<tr>
<th>Chapter 5. Frequency and length</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of word length 1</td>
<td>67</td>
</tr>
<tr>
<td>Distribution of word length 2</td>
<td>67</td>
</tr>
<tr>
<td>Distribution of word length and Ord´s criterion</td>
<td>68</td>
</tr>
<tr>
<td>Frequency and compounding propensity</td>
<td>68</td>
</tr>
<tr>
<td>Frequency and irregularity</td>
<td>69</td>
</tr>
<tr>
<td>Frequency and letter utility</td>
<td>70</td>
</tr>
<tr>
<td>Frequency and markedness/complexity</td>
<td>70</td>
</tr>
<tr>
<td>Frequency and order in freezes</td>
<td>72</td>
</tr>
<tr>
<td>Frequency and phoneme complexity</td>
<td>73</td>
</tr>
<tr>
<td>Frequency and phoneme form</td>
<td>74</td>
</tr>
<tr>
<td>Frequency and production effort</td>
<td>74</td>
</tr>
<tr>
<td>Frequency and productivity</td>
<td>75</td>
</tr>
<tr>
<td>Frequency and reduction</td>
<td>75</td>
</tr>
<tr>
<td>Frequency and variety</td>
<td>77</td>
</tr>
<tr>
<td>Length and frequency</td>
<td>78</td>
</tr>
<tr>
<td>Length and polysemy</td>
<td>80</td>
</tr>
<tr>
<td>Length and word classes 1</td>
<td>81</td>
</tr>
<tr>
<td>Length and word classes 2</td>
<td>82</td>
</tr>
<tr>
<td>Sentence length and clause length</td>
<td>82</td>
</tr>
<tr>
<td>Word length and polytextuality</td>
<td>83</td>
</tr>
<tr>
<td>Word length and position in sentence</td>
<td>84</td>
</tr>
<tr>
<td>Word/morph length and composition</td>
<td>85</td>
</tr>
</tbody>
</table>
### Chapter 6. Semantics, synergetics, psycholinguistics

- Abstractness
- Distribution of polysemy
- Familiarity and frequency
- Familiarity of slang words
- Kanji frequency
- Learning and complexity
- Learning with children
- Meaning and frequency
- Morpheme inventory and morpheme polysemy
- Morphology vs. phonemics
- Phoneme inventory vs. morpheme length
- Polysemy and compounding
- Semantic classes
- Semantic diversification

### Chapter 7. Typology

- Entropy and synthetism
- Homonymy and synonymy of affixes 1
- Homonymy and synonymy of affixes 2
- Inflection in general
- Morph length
- Popescu’s typological indicator $a$
- Root length and extent of derivation
- Synthetism in language
- Vocalic language
- Word length and agreement
- Word order and inflection

### Chapter 8. General problems

- Distributions
- Entropy and inventory size
- Fitting a distribution
- Setting up hypotheses by means of factor analysis
- Iconicity
- Index formation
- Menzerath’s law
- Naranan-Balasubrahmmanyan distribution
- Ord’s criterion
- Repeat rate and entropy
- Sample size
- The problem of infinity
Tightness/Cohesion
Zipf’s and Zipf-Mandelbrot’s law

Chapter 9. Research projects
Frumkina’s law (Word occurrence in passages)
Skalička’s typological system
Synonymy
Word frequency and collateral properties

Author index
Subject index