Problems
in
Quantitative Linguistics
1

by

Udo Strauss
Fengxiang Fan
Gabriel Altmann

Second edition

2008
RAM-Verlag

Studies in quantitative linguistics

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, IX +132 pp. (2nd edition).

ISBN: 978-3-9802659-4-2

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, recheck 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