Data Processing and Management for Quantitative Linguistics with Foxpro

by

Fan Fengxiang

2010
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)

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 + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.

ISBN: 978-3-942303-03-3

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

Preface

Imagine a researcher of Shakespearean plays is studying the Bud's stylistic characteristics with the quantitative approach. He has all the plays totalling about a million words stored in XML files. The immediate task before him is to remove all the XML codes from the files to get "pure" text. Next, he needs the following data: a wordlist with frequencies and word length both in letters and syllables, the vocabulary richness and frequency spectrum of each of the plays, lexical similarity and distance among the plays, the average word length in syllables and the average sentence length of each of the plays, collocations of certain words, number of rare words—hapax legomena, vocabulary growth rate, etc. However, life of a linguistic researcher is not as simple as that. To get a wordlist with word frequencies he'll need to lemmatize all the word tokens in those plays, and as the research progresses, some ad hoc research inspirations may pop up and new data are needed; he also has to constantly rearrange the data trying to find some patterns and retrieve some for a closer look, etc. These tasks would take ages to complete manually. The well known American scholar Ione Dodson Young used 25 years to make a concordance for the complete poetic works of Byron; she started the work in 1940 and didn't compete it until 1965!

With Foxpro, a powerful data processing and managing system, all the above can be done in a matter of a few minutes. This book, *Data Processing and Management for Quantitative Linguistics with Foxpro* gives detailed descriptions and instructions on how to gather, process and manage large amount of linguistic data with this data managing system. This book is aimed at literary and linguistic researchers, teachers and students at the undergraduate or postgraduate levels, EFL/ESL teachers and students, etc. It is also a very good book for corpus linguistics, text mining, information retrieval, and natural language processing. No previous computer programming experience is required of the reader except the ability to use the Windows Operating System.

All the examples for the commands and functions, as well as the demonstration programs in the book, are literary/linguistic oriented and of the author's own creation, and the majority of them are immediately useful for serious research, after changing only the input and output file names and their path. This book can be used as a course book that takes roughly 36-lab hours to complete; it can also be used for self-study. There is a CD-ROM attached to the book with all the Foxpro tables, examples, demonstration programs and non-copy-right textual materials for all the programs, exercises and model answers to these exercises.

There are different versions of Foxpro, and the latest version is Visual Foxpro 9. The Foxpro needed in this book is Foxpro 6 or higher. Foxpro can process any language in the world; however, in this book, it's used mainly to deal with English, occasionally Chinese. With some changes, the programs in the book can also be adapted to process other languages.

The following are some suggestions for tackling this book.

Firstly, this book is not for reading, but for careful reading plus repeated practice. That is, the reader should sit in front of the computer trying out each of the operators, commands, functions and examples many, many times while reading it. The operators, commands and functions in this book, totalling about 200, were carefully selected and are the most fundamental for linguistic computing. In some other computer languages there are fewer commands and functions; however, the users have to create commands and functions themselves when needed, and this makes these types of languages more difficult to learn and use for linguistic researchers and students. The reader of this book is not expected to remember all these operators, commands, functions, etc, by heart. He or she can always come back to this book to refresh his or her memory.

Secondly, as mentioned before, used as a course book, it'll take about a semester, roughly 36 lab hours to complete, and for each lab hour, the students need at least two more hours for home practice. For self-study, it'll take half a year. A person hurrying through the book in 10 days will probably learn nothing.

Thirdly, all the examples and exercises were carefully planned. The reader is not expected to solve all the problems in the exercises. One of the purposes of the exercises are for making the reader think about the possible applications of the operators, commands and functions etc learned; if the reader is unable to do the exercises, that's perfectly normal for a beginner; in such cases, go to the model answers, analyse them and then try them out. This is an important learning process.

Lastly, the author hopes that the above will not scare off potential readers. Please bear in mind that there are no magic books in the world from which a beginner can learn a computer language in 10 or 20 days. Learning a computer language from scratch is not like reading Shakespeare or Goethe for the first time; it's a long and sometimes painful process, and patience and perseverance are a must. But once learned, it'll be an open sesame for the learner to the wonderful linguistic and literary treasure trove that can last a life time.

The author is deeply indebted to Professor Gabriel Altmann for his insightful suggestions for this book and for his constant stimulating research ideas from which the author has benefited greatly; without his support this book wouldn't be possible. The author also wishes to thank Professor Reinhard Köhler for reading the manuscript and for his expert advice.

Fan Fengxiang

Table of Contents