

QUITA

Quantitative Index Text Analyzer

Miroslav Kubát
Vladimír Matlach
Radek Čech

2014

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.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp.
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics Vol. 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol. 3*. 2013, IV + 403 pp.

14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014. VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in Language*. 2014, VIII+123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to ext and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014, VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, VII+106.

ISBN: 978-3-942303-28-6

© Copyright 2014 by RAM-Verlag, D-58515 Lüdenscheid, Germany

RAM-Verlag

Stüttinghauser Ringstr. 44

D-58515 Lüdenscheid

Germany

RAM-Verlag@t-online.de

<http://ram-verlag.de>

Contents

1	Introduction.....	1
2	System requirements.....	3
3	Install and Starting the Application	4
4	Creating a New Project.....	5
5	Creating and loading texts	7
5.1	Supported files and encodings	8
6	Indicators to compute	9
6.1	Frequency Structure Indicators	10
6.1.1	Type-Token Ratio (<i>TTR</i>)	10
6.1.2	<i>h</i> -point (<i>h</i>).....	11
6.1.3	R_1	14
6.1.4	Repeat Rate (<i>RR</i>).....	16
6.1.5	Relative Repeat Rate of McIntosh (RR_{mc}).....	19
6.1.6	Hapax Legomena Percentage (<i>HL</i>)	20
6.1.7	Lambda (<i>λ</i>).....	21
6.1.8	Gini Coefficient (<i>G</i>)	30
6.1.9	R_4	34
6.1.10	Curve length (<i>L</i>)	35
6.1.11	Curve length R Indicator (<i>R</i>).....	36
6.1.12	Entropy (<i>H</i>)	38
6.1.13	Adjusted Modulus (<i>A</i>)	40
6.2	Miscellaneous indicators.....	42
6.2.1	Verb Distances (<i>VD</i>).....	42
6.2.2	Activity (<i>Q</i>) & Descriptivity (<i>D</i>).....	43
6.2.3	Writer's View (<i>a</i>).....	44
6.2.4	Average Tokens length (<i>ATL</i>).....	47

6.2.5	Thematic Concentration (<i>TC</i>)	48
6.2.6	Secondary Thematic Concentration (<i>STC</i>)	52
6.2.7	Proportional Thematic Concentration (<i>PTC</i>).....	55
7	Pre-processing: Tokenization, Lemmatization, POS Tagging.....	59
7.1	QUITA and third party tools	60
7.2	Tokenizer.....	60
7.2.1	Available tokenizers	62
7.3	Lemmatizer	63
7.4	POS Tagger	65
8	Post-processing.....	66
8.1	N-grams.....	66
8.2	Text Length Reduction.....	67
9	Cache Settings.....	69
10	Starting Calculations	70
11	Results	71
11.1	Underlined Results alias “Alternative data”.....	71
11.1.1	Types	72
11.1.2	Tokens	73
11.1.3	Frequencies.....	73
11.1.4	POS Frequencies	74
11.1.5	Thematic concentration.....	74
11.1.6	Secondary Thematic Concentration	75
11.1.7	Activity.....	75
11.1.8	Descriptivity.....	76
11.1.9	Token Length Frequency Spectrum	76
11.2	Results tools.....	77
11.2.1	Text Selection.....	77
11.2.2	View Text	77

11.2.3	Comparison of Results	78
11.2.4	Comparison of Projects	80
11.2.5	Chart Wizard	83
11.2.6	Copy Results	87
12	Tools	89
12.1	Random Text Creator.....	89
12.2	Binary File Translator.....	91
13	Additional information	95
14	References	96
15	Appendix	98
15.1	Text 1	98
15.2	Text 2	98
15.3	Word list of Text 1 and Text 2	99
15.4	Text 3	105
15.5	Text 4	105

1 Introduction

Quantitative experimental methods have been increasingly used in the humanities in recent years. We can hardly imagine the disciplines of social science, such as psychology, sociology or economics, without a quantitative approach. On the other hand, the majority of linguists, historians and especially literary critics are still refusing to use quantitative methods. One of the reasons is the fact that those researchers consider quantitative methods, and especially statistical methods, too difficult to apply to their field. QUITA (Quantitative Indicator Text Analyzer) is a tool which aims to help all people who try to analyse texts by quantitative methods.

QUITA is a program to enable researchers from various disciplines (linguistics, criticism, history, sociology, psychology, politics, biology, etc.) to analyse texts using quantitative methods. There are many indicators which measure various characteristics of text. Although the authors of QUITA focused mainly on indicators connected to the frequency structure of a text, there are also functions for several other characteristics. Since QUITA is designed especially for researchers outside quantitative linguistics, it includes functions for the most basic and common indicators.

Given that the main purpose of QUITA is to provide a user-friendly tool of quantitative text analysis for researchers without a deeper knowledge of quantitative linguistics, statistics or programming, QUITA also provides simple statistical comparisons and the ability to create charts. There is no need to use any additional software such as spreadsheet applications or special statistical programs. QUITA is therefore the program that combines all the essential parts of any quantitative research effort: obtaining results, statistical testing and graphical visualization.

The QUITA manual is written with step-by-step instructions. All tools are concisely described and accompanied by screen-shots. Every indicator is briefly presented (complete with references), and mathematically defined. There are also examples of computation and statistical comparison.

Although the manual provides users with all the essential information about QUITA, it was not possible to cover most topics in deeper detail. For this purpose, we highly recommend the book *Word frequency studies* (Popescu et al. 2009) which is a comprehensive overview about quantitative analysis using indicators based on the frequency structure of a text. The book *Aspects of Word Frequencies* (Popescu et al. 2009) is also well worth reading. Detailed examples of computing most indicators used in QUITA can be found in *Metody kvantitativní analýzy (nejen) básnických textů* (Čech et al. 2014).

Since we aim to help as many researchers as possible, QUITA is distributed as freeware. Thus anyone can use QUITA without any restrictions. The latest version of the software is available on the website <https://code.google.com/p/oltk/>. In published work, acknowledgement of QUITA would be appropriate and appreciated.

2 System requirements

Supported Systems: Windows XP, Windows Vista, Windows 7, Windows 8.

System requirements: NET Framework 3.5

Optional system requirements: Python, Perl

NOTE:

Users are automatically notified that it is necessary to install the system requirements with the download links.