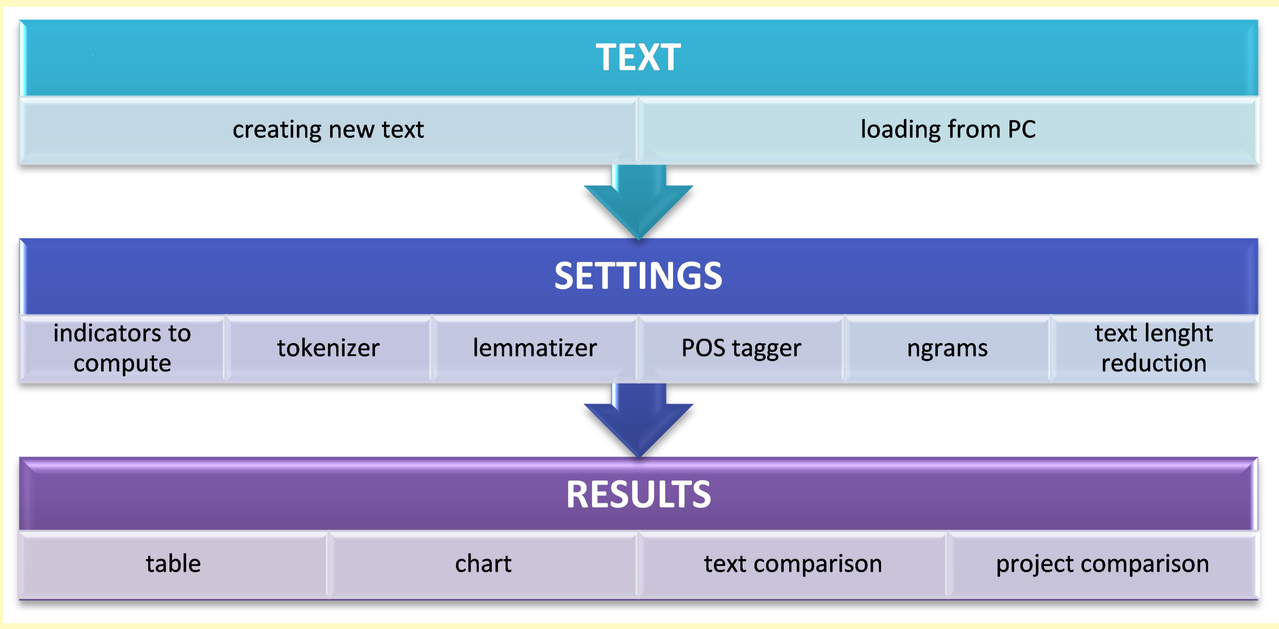
QUITA

Quantitative Index Text Analyzer

Miroslav Kubát
Vladimír Matlach

Department of General Linguistic, Palacký University, Czech Republic

Our aim is to provide a user-friendly tool of quantitative text analysis for researchers from various disciplines (linguistics, criticism, history, sociology, psychology, politics, biology, etc.). QUITA combines all important parts of any quantitative research: obtaining results, statistical testing and graphical visualization. There is no need to use any additional software such as spreadsheet applications or special statistical programs.



TEXT-PROCESSING

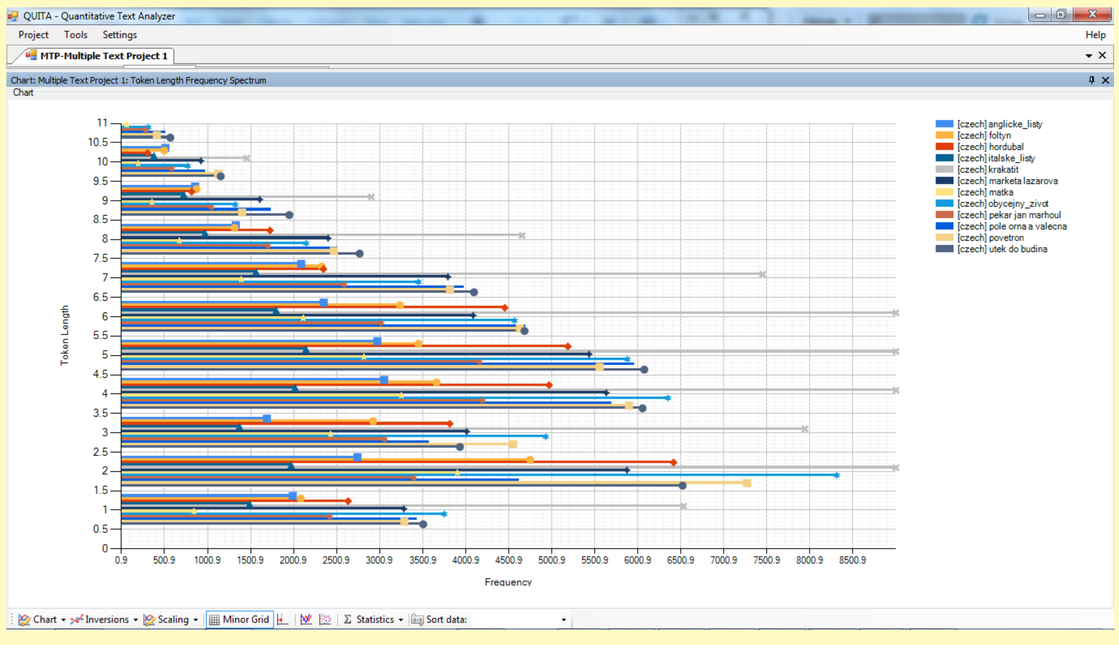
Pre-processing

- Tokenizer (word, line, char, DNA Triplet, DNA Nucleotide)
- Multilingual lemmatizer (AR, CZ, DE, DK, EN, ES, FI, FR, IT, NL, PT, RO, RU, SE)
- POS Tagger (It distinguishes parts of speech in a text)

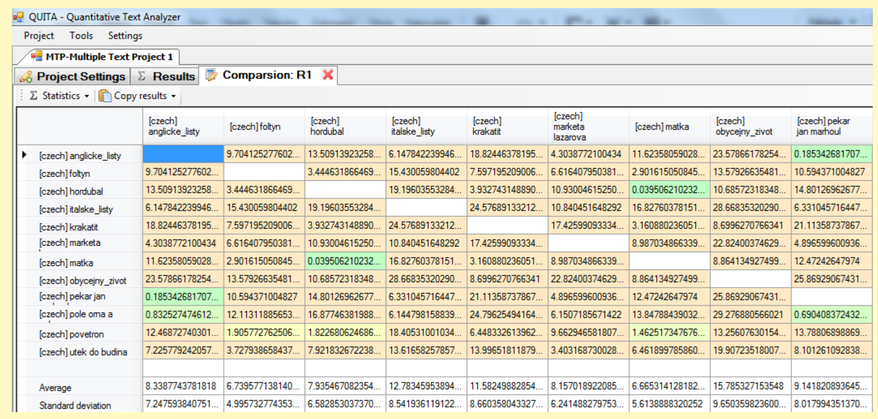
Post-processing

- N-grams (QUITA enables creating char, word or whatever n-grams)
- Text length reduction

CREATING CHARTS



STATISTICAL COMPARISON



	[czech] anglicke_listy	[czech] foltyn	[czech] hordubal	[czech] italske_listy	[czech] krakatit	[czech] marketa lazarova	[czech] matka	[czech] obycejny_zivot	[czech] pekar jan marhoul
[czech] anglicke_listy		9.704125277602...	13.50913923258...	6.147842239946...	18.8246378195...	4.3038772100434	11.62358059028...	23.57866178254...	0.185342681707...
[czech] foltyn	9.704125277602...		3.444631866469...	15.430059804402	7.597195209006...	6.616407950381...	2.901615050845...	13.57926635481...	10.594371004827
[czech] hordubal	13.50913923258...	3.444631866469...		19.19603553284...	3.932743148890...	10.93004615250...	0.039506210232...	10.68572318348...	14.80126962677...
[czech] italske_listy	6.147842239946...	15.430059804402	19.19603553284...		24.57689133212...	10.840451648292	16.82760378151...	28.66835320290...	6.331045716447...
[czech] krakatit	18.82446378195...	7.597195209006...	3.932743148890...	24.57689133212...		17.42599093334...	3.160880236051...	8.6996270766341	21.11358737867...
[czech] marketa	4.3038772100434	6.616407950381...	10.93004615250...	10.840451648292	17.42599093334...		8.987034866339...	22.82400374629...	4.896599600936...
[czech] matka	11.62358059028...	2.901615050845...	0.039506210232...	16.82760378151...	3.160880236051...	8.987034866339...		8.864134927499...	12.47242647974
[czech] obycejny_zivot	23.57866178254...	13.57926635481...	10.68572318348...	28.66835320290...	8.6996270766341	22.82400374629...	8.864134927499...		25.86929067431...
[czech] pekar jan marhoul	0.185342681707...	10.594371004827	14.80126962677...	6.331045716447...	21.11358737867...	4.896599600936...	12.47242647974	25.86929067431...	
[czech] pole orna a valecna	0.832527474612...	12.11311885653...	16.87746381988...	6.144798158839...	24.79625494164...	6.1507185671422	13.84788439032...	29.276880566021	0.690408372432...
[czech] povetron	12.46872740301...	1.905772762506...	1.822680624686...	18.40531001034...	6.448332613962...	9.662946581807...	1.462517347676...	13.25607630154...	13.78806898869...
[czech] utek do budina	7.225779242057...	3.727938658437...	7.921832672238...	13.61658257857...	13.99651811879...	3.403168730028...	6.461899785860...	19.90723518007...	8.101261092838...
Average	8.3387743781818	6.739577138140...	7.935467082354...	12.78345953894...	11.58249882854...	8.1570189220850...	6.665314128182...	15.785327153548	9.141820893645...
Standard deviation	7.247593840751...	4.995732774353...	6.582853037370...	8.541936119122...	8.660358043327...	6.241488279753...	5.613888320252	9.650359823600...	8.017994351370...

INDICATORS TO COMPUTE

Frequency Structure indicators

- Type-Token Ratio (*TTR*)
- *h*-point (*h*)
- Vocabulary Richness (R_1)
- Repeat Rate (*RR*)
- Relative Repeat Rate of McIntosh (RR_{mc})
- Hapax Legomenon Percentage (*HL*)
- Lambda (Λ)
- Gini Coefficient (*G*)
- Vocabulary Richness (R_4)
- Curve length (*L*)
- Curve length Indicator (*R*)
- Entropy (*H*)
- Adjusted Modulus (*A*)

Miscellaneous indicators

- Verb Distances (*VD*)
- Activity (*Q*) & Descriptivity (*D*)
- Writer's View (α)
- Average Tokens length (*ATL*)
- Thematic Concentration (*TC*)
- Secondary Thematic Concentration (*STC*)

oltk.upol.cz/software

Acknowledgement: QUITA was supported by the student project IGA (no. FF_2013_031) of the Palacký University, Olomouc.