IQLA-GIAT Summer School in
Quantitative Analysis of Textual Data
University of Padua, 21-25 September 2015

ABSTRACTS

➢ Andrei Beliankou (University of Trier, Germany)

Introduction in Programming with R
The open sourced R development environment gained on popularity not among statisticians from the old school but also in the Data Science and Business Analytics people. The main advantages of R is its vibrant community and the enormous amount of libraries for every task one can imagine. The course will focus on the very basics such as development environments (RStudio), installation of libraries, basic data structures and language constructs. The participants should get familiar with some advanced topics like graphical data exploration and windowing functions. After the course all participants will be able to continue their work with R on the own hand and educate themselves in advanced topics of R programming.

Quantitative Models of Linguistic Data
Based on the Introduction in Programming with R this course is aimed at modelling linguistic data. During the course all participants will master R facilities to work with discrete and continuous distributions, employ this knowledge for testing simple hypotheses. A further part of the course contains regression modelling, model evaluation and prediction steps. An application of these facts will be a two step exercise on the models of the Lexical Richness.

➢ Rocco Coronato (University of Padova, Italy)

(Under)mining Metaphors in Shakespeare with a little help from Quintilian
Both in real life and in literary analysis, metaphors are notoriously difficult to grasp. I would like first to analyze the classical and Elizabethan notions of metaphor, heavily based on the transition between animate and inanimate in correspondence to the still popularly believed Neoplatonic order of the cosmos (a point often eluded in our modern theories on metaphor). By using Google Big Query and Gephi, I will compare the metaphors in Twelfth Night with larger corpora (the whole text of Twelfth Night, Shakespeare's works and Elizabethan corpora). Does this analysis unveil the presence of cues and other rhetorical and theatrical forms? Is metaphor in Shakespeare always accompanied by three concomitant factors (usage of common words, rare or unique occurrence of these words in the text, shift between 'animate' and 'inanimate')? Does the visualization of metaphors show deeper clusters of meaning, thus favouring new interpretations of the text?

➢ Maciej Eder (University of Kraków, Poland)

Stylometry with the package 'Stylo': Explanatory methods
Stylometry with the package 'Stylo': Supervised methods
The workshop, split into two 1.5 hour lab sessions, will address varius types of stylometric analyses using the R package 'Stylo'. This library of R functions, and at the same time a suite of stylometric tools, is an open-source available from the repository CRAN. The first session will acquaint the participants with 'Stylo' and focus on explanatory distance-based methods (cluster analysis, multidimensional scaling, principal component analysis), while the second block will discuss supervised machine-learning approaches to

classification (delta, support vector machines, nearest shrunken centroids). Both blocks will focus on tools supplemented with user-friendly interfaces, so no expert knowledge of R in particular, or of programming in general is required. The texts used for the workshops will be provided by the instructor and the participants are encouraged to bring their own; if necessary, the participants' individual corpora will be expanded as needed and as available.

- **Reinhard Köhler (University of Trier, Germany)**

General Methodology in Empirical Linguistics. Evaluation of Data and Hypothesis Testing
What is the difference between fact, phenomena, and data? What is measurement? Which problems occur with measuring linguistic phenomena and how can they be overcome?
General methodological issues are discussed, which are the same in linguistics and in other empirical sciences. First, the nature of data as the results of (qualitative or quantitative) descriptions of facts and phenomena is in focus. Quantitative descriptions of text and language phenomena are called measurements. Purpose and methods of measurements are presented, which includes a consideration of measurement problems. Furthermore, Scale Levels (levels of measurement) must be taken into account when data are evaluated and mathematically processed.
Another topic is the role of composed properties and index formation. Important hints are given as to comparison of objects (such as texts) by means of individual values or indexes.
Then, the procedure of statistically testing linguistic hypotheses is analysed and some common tests are characterised, combined with warnings against the application of individual tests without checking their applicability to the given linguistic data.
Finally, remarks on the status and concepts of language, corpus, text, statistical sample, and population are made, which very often are confused.

A Crash Course in the Central Terms and Concepts of Science
In linguistics, computational and corpus linguistics, a dedicated education in the philosophy of science is rare. Therefore, terms and concepts of science are confused and misused in many cases, which may lead to severely false conclusions.
The course attempts to present the participants a thorough overview over the preconditions of scientific work and over the logics of scientific activities in general.
The course considers
- the concepts of facts, phenomena, and data
- the concepts and types of scientific problems and their logical structure
- the concept and the types of scientific concepts
- the concept and the types of scientific hypotheses and models
- the concept of laws
- the concept of scientific explanation and prediction
- the helix of scientific work from antecedens knowledge to new knowledge

- **George Mikros (National and Kapodistrian University of Athens, Greece)**

Cross-linguistic authorship attribution: Detecting stylomes in translated documents and texts written by bilingual authors
Authorship identification techniques have been extensively used for the attribution of texts in specific authors as long as these texts are produced originally from one of them in his/her mother-tongue. However, there is little experience in testing authorship identification methods in cross-linguistic areas where the translator is the aim of the identification or the source (training) documents and the disputed samples have been written in different languages. Stylometric theory assumes that each author possess a distinct, unique “writeprint” which is expressed quantitatively through the idiosyncratic occurrence variation of its most frequent linguistic structures and various indices of unconscious linguistic behavior such as lexical “richness” formulas, word and sentence lengths etc. Translations and cross-linguistic text productions test the theory of “writeprint” in its extreme. If the identity of the author survives through the process of translation or the L2 writing and can

be traced in a text that originally was written in another language then stylometric authorship attribution would increase its methodological robustness and reliability.
This talk will present a brief overview of the evolution of stylometry the last decades and will explain its basic methodological tools for extracting "writeprints" from texts (stylometric features, machine learning algorithms). Then it will focus in two different application areas:

- Translations of classic Russian literature will be used as a test-bed to investigate whether computational stylistic methods can be successfully applied to uncover not only the original author but also the translator.
- Tweets in English and Spanish written by bilingual English-Spanish users will be used to test whether we can find quantitative evidence of stylometric features that correlate between the two languages. The reported research results will be discussed under the general theoretical framework of quantitative linguistics.

Machine learning methods in computational stylistics using the caret R package
The aim of this lab session is to provide a hands-on experience using state-of-the-art machine learning methods using a specialized R package named caret. The caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

- data splitting
- pre-processing
- feature selection
- model tuning using resampling
- variable importance estimation
- as well as other functionality.

In the lab session we will explore its basic data preprocessing capabilities and perform various analyses using real research datasets performing predictive classifications in authorship attribution and author profiling tasks.

➢ **Paolo Nadalutti (IT expert)**

Visualizing content analysis information using web based tools
When dealing with big amounts of data - just like when analyzing text data – visualizing the information extracted by analysis is a key feature for both data exploration and results presentation. Recently a big number of new instruments have become available to transform data into meaningful pictures, particularly web-based and open source instruments that can help users to get free from the standard software plotting tools. In this course, starting from data coming from content analysis procedures and using open source and web based instruments, we will create a number of meaningful plots to highlight relations between texts or relevant variables. Further than explaining what each type of plot is and how it can be used, we will suggest what the best visual solutions are in order to plot words counts and frequencies, text-level summary variables, vocabulary growth rates, factorial analysis scatterplots, distance measures analysis.

➢ **Sven Naumann (University of Trier, Germany)**

Data-driven approaches to syntactic processing
In recent years, modern (computational and corpus) linguistics has developed a number of extremely powerful tools for a range of non-trivial language processing tasks /e.g. tokenization, tagging, chunking, ...). Looking back at generative linguistics, it seems that now problems are actually solved while before linguists wasted most of their precious time by inventing grammar formalisms nobody really needed. But a closer look reveals that things are not that simple.
In the first part of the class, we look at a number of modern parsing systems (Stanford parser, Berkeley Parser, …) and the probability models they use. Then, we focus on the grammar engineering approach many of them employ. I will argue that it leads to a representation of syntactic knowledge that can be considered as problematic from a linguistic and cognitive point of view.

Syntactic complexity: quantitative models and measures
Syntactic complexity is one of those linguistic terms that are used quite frequently though it is not entirely clear what exactly they refer to. Does syntactic complexity, for example, address specific aspects of the language system (structural complexity) or is it more closely related to the processing of linguistic units (cognitive complexity)?
Nevertheless, a large number of competing measures have been proposed as more or less plausible operationalizations for this concept. The class will (a) discuss a number of measures for syntactic complexity found in literature (b) give an overview of related studies in quantitative and synergetic linguistics.

- **Stefano Ondelli (University of Trieste, Italy)**

Compiling Corpora: Representativeness, Balance and Sampling
Corpus design is undeniably paramount for successful research; however no generally valid approach is available for compiling corpora since corpus selection greatly depends on the researcher's objectives. According to the purpose for which the corpus is used, the principles of sociolinguistics are called upon to define the variety and balance of its text –components. This presentation provides an overview of the issues at stake when dealing with the pre-processing of language tokens as well corpus representativeness and balance. Moreover, examples will be illustrated of recent research in corpus linguistics using samples of language for special purposes.

- **Pierre Ratinaud (University of Toulouse II, France)**

IraMuTeQ: corpus indexation, manipulation and simple description
The Reinert method in IraMuTeQ
Iramuteq is a free (as in free speech) software (licence GNU GPL) for data and textual manning. It's based on the R statistical software. It can perform different types of text analysis and visualization on large text corpora (over hundreds of millions of occurrences). One of its particularities is to reproduce Reinert Analysis (1983, 1991). The first lab will focus on the indexation and manipulation of corpus. I will present different possibilities for extracting sub-corpus and some general aspect of the interface. The second lab will be an in depth presentation of the Reinert method and of the different tools implemented in IraMuTeQ to help its interpretation.

- **Jacques Savoy (University of Neuchâtel, Switzerland)**

Application of Quantitative Linguistics methods to US political speeches
In this lecture we will focus on applying three quantitative linguistics methods (specific vocabulary, intertextual distance, and clustering) to political speeches. We will show that these three approaches can be viewed as "distant reading" revealing the underlying main trends and evolution included in a corpus, mainly from a stylistic point of view.
During this lecture, we will use both governmental (State of the Union addresses from 1790 to 2015) or US electoral speeches. Based on a method detecting the specific vocabulary, we can define the style and expressions significantly related to a given US president or candidate (e.g., as we for Clinton or Obama, or she with Polk). Based on this finding, we will also show how we can extract the most specific sentences of a given president. Moreover, we can determine lexical leaders, presidents who were able to introduce an expression reuse significantly by his followers (e.g., as we with F.D. Roosevelt, or God with R. Reagan).
Using an intertextual distance between speeches delivered by US presidents, we can measure a distance between presidents. Using a clustering algorithm, we can then generate an overview depicting the stylistic similarities between the 44 US presidents. This view can then be interpreted to derive some general stylistic trends over two centuries (in which both T. and F.D. Roosevelt have a particular place).