Towards interactive visual analysis of corpora

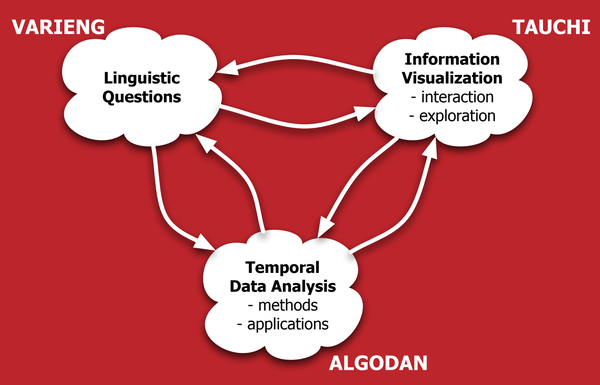
DAMMOC

"Data Mining Tools for Changing Modalities of Communication"

WHO?

Project funded by the Academy of Finland for 2009-2011, led by:

- Heikki Mannila, Aalto University
  - ALGODAN – Algorithmic Data Analysis (National Centre of Excellence)
- Terttu Nevalainen, University of Helsinki
  - VARIENG – Research Unit for Variation, Contacts and Change in English (National Centre of Excellence)
- Kari-Jouko Räihä, University of Tampere
  - TAUCHI – Tampere Unit for Computer-Human Interaction



WHAT?

Aim: to enable easier, faster and more powerful analysis of corpora

- Identify **methods in data analysis and visualization** that are useful for the study of language
  - Test cases: linguistic complexity, language variation and change
  - Pattern discovery, clustering, …
- Develop these methods into **software tools** accessible to corpus linguists
  - Open-source: R, Mondrian

WHY?

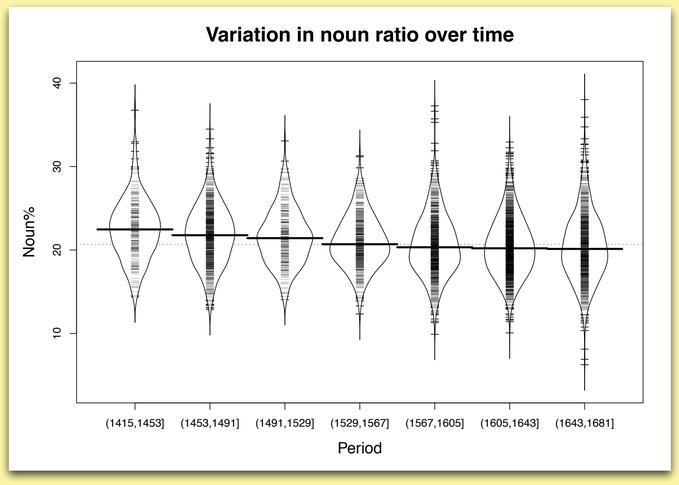
- More and more (annotated) corpora available
- Development of sophisticated but user-friendly tools for analysis lagging behind
  - Insight needed from recent advances in data mining and visualization
- To facilitate research in the development of computational methods

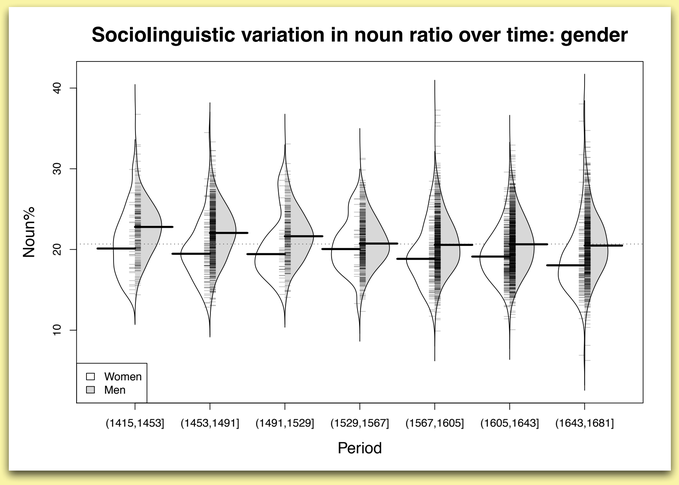
Project web page:
http://tauchi.cs.uta.fi/projects/dammoc

ICAME 2010 talks:

- Vartiainen, Turo and Lijffijt, Jefrey: *Pre-modifying -ing participles in the parsed BNC*
- Säily, Tanja, Nevalainen, Terttu and Siirtola, Harri: *Variation in noun and pronoun frequencies: Gendered drift or a corpus artefact?*

dammoc-project@helsinki.fi





Variation in noun and pronoun frequencies: Gendered drift or a corpus artefact?

Tanja Säily, Terttu Nevalainen, Harri Siirtola

Parsed Corpus of Early English Correspondence (PCEEC)

Results

- Proportion of nouns per letter decreases over the centuries
  - *English letter-writing becomes less focused on information over time?*
- Women use more pronouns and fewer nouns than men in every subperiod
  - *Gendered styles similar to Present-day English?*

Visualization: **beanplots** (Kampstra 2008, *Journal of Statistical Software* 28)

Clustering of genres

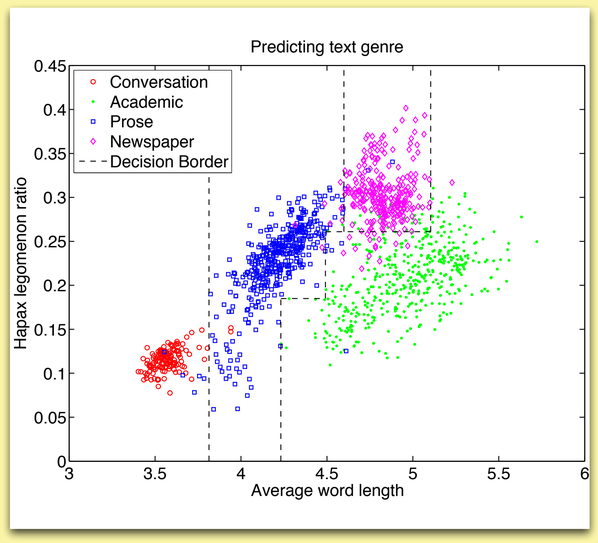
Clustering texts according to genre

Jefrey Lijffijt, Heikki Mannila, Terttu Nevalainen

Four pre-annotated genres from the British National Corpus (BNC-XML)

Results

- 93 % prediction accuracy using decision tree with two simple text measures
  - *Percentage of nouns* and *percentage of pronouns*
  - *Hapax legomena* and *average word length*
- Nearly equal prediction accuracy using unsupervised hierarchical clustering

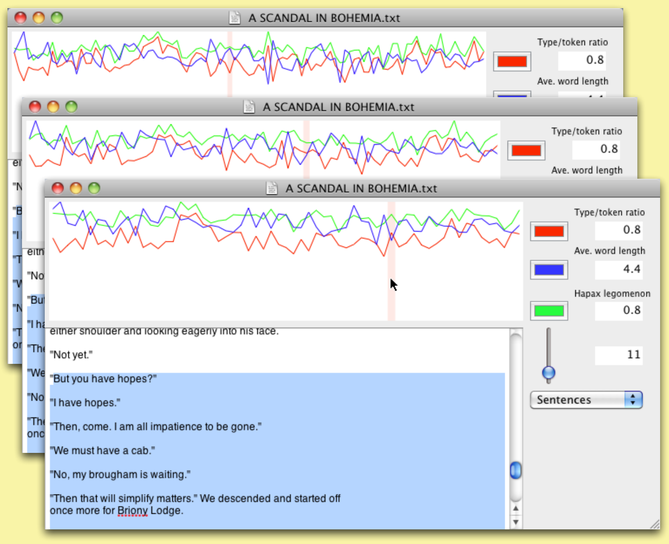


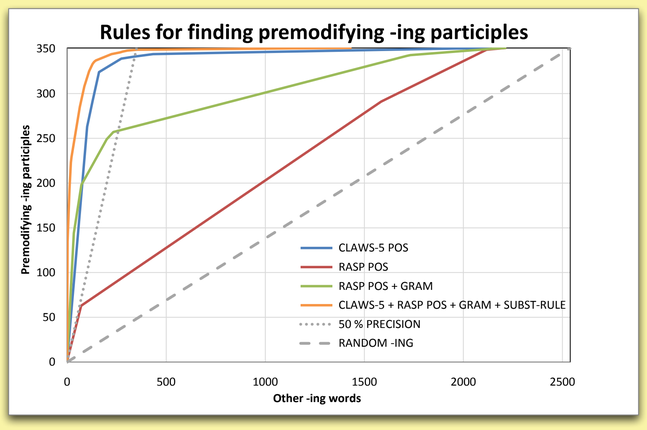
Interactive adjustment of window size

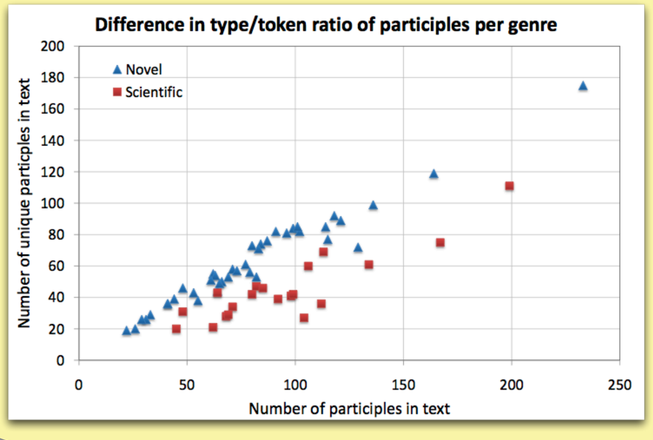
Harri Siirtola, Tanja Säily, Terttu Nevalainen

Type/token ratio, average word length, and proportion of hapax legomena are sensitive to window size

- Interactive tool for exploring **text complexity**, finding optimal window size for measures







Finding premodifying -ing participles in the parsed BNC

Turo Vartiainen, Jefrey Lijffijt

The parsed British National Corpus (Briscoe et al. 2006; Andersen et al. 2008)

Results

- Combining the information from the BNC-XML and the parsed BNC yields the most accurate results

Pilot study

- The participles have distinct functions in different genres
  - The type/token ratio of participles can be used to identify e.g. scientific texts from novels in the corpus