ALGODAN Algorithmic Data Analysis

Computational methods for complexity and burstiness in natural language

Jefrey Lijffijt

Joint work:

Panagiotis Papapetrou, Kai Puolamäki, Heikki Mannila

Aalto University, Dpt. of Information and Computer Science

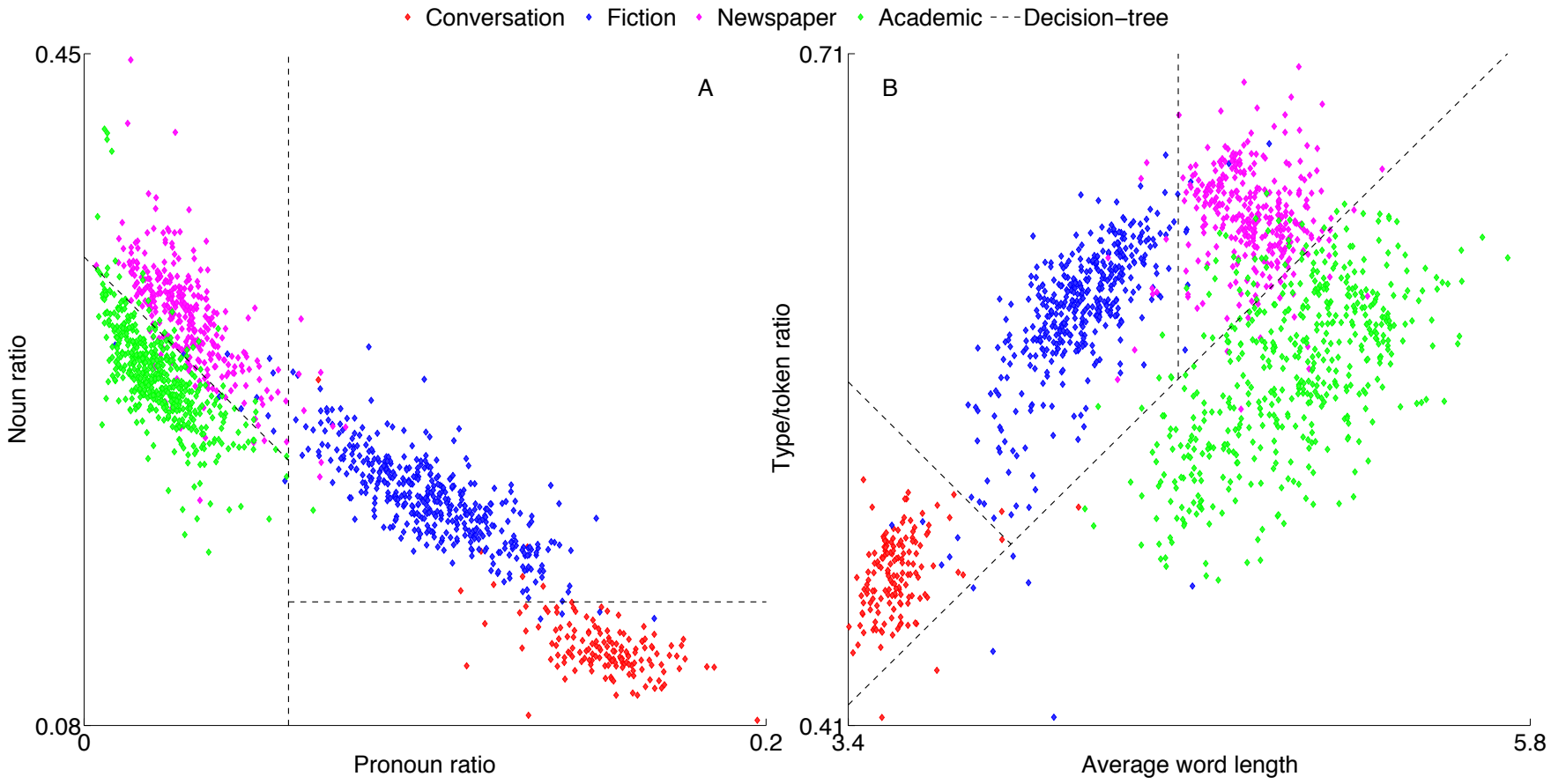
Tanja Säily, Turo Vartiainen, Terttu Nevalainen

University of Helsinki, Dpt. of Modern Languages

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Early work: Classification and clustering of genres



- Paper under rotation.

Early work: Word sense disambiguation

- Query words that are difficult to query directly
  - Premodifying –ing participles
- Use automatically tagged + parsed corpus
  - Noisy tags and poor parsing
- Goal: generate rules
  - [POS = AJ0 and SUBST_NEXT = True → Prem. -ing part.]
- 0-1 Classification problem

- Turo Vartiainen, Jefrey Lijffijt. **Premodifying -ing participles in the parsed BNC**. **To appear in *Corpus Linguistics and Variation in English: Theory and Description.* Amsterdam/New York: Rodopi.**

Recent work: Inter-arrival time distributions

- Count space between consecutive occurrences: **and**

Finnair believes that it will be able to resume its scheduled service to **and** from New York on Monday, after two days of cancellations caused by hurricane Irene. All three airports serving New York City have been closed because of the hurricane **and** Finnair was forced to cancel flights on Saturday **and** Sunday. The airline is not certain when its scheduled service can be resumed, but the assumption is that Monday afternoon's flight from Helsinki will depart. Some Finnair passengers whose final destination is not New York have been rerouted **and** some have delayed travel plans. The company has also offered ticket holders a refund. *YLE*

- $IA_{and} = \{29, 9, 39, 29\}$

- Hypothesis: this captures the behavior pattern of words

Recent work: Finding keywords (statistical tests)

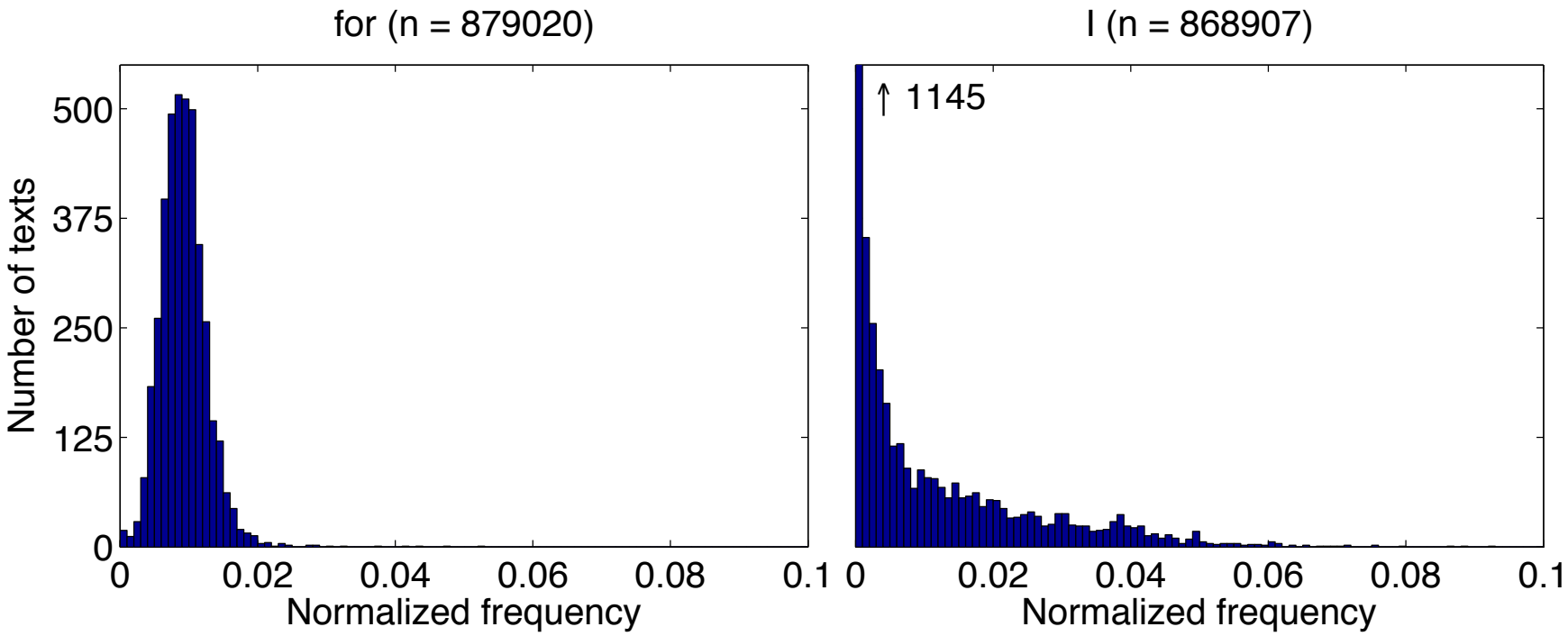
- Given two corpora (collections of texts) *S* and *T*
  - Corpus with fiction prose
  - *S* = male authored, *T* = female authored
- Find all words that are *significantly* more frequent in *S* than in *T*, or vice versa

Word	Freq in *S*	Freq in *T*
sergeant	57	32
Total	400.000	400.000

- Is this statistically significant?

Recent work: Finding keywords (statistical tests)

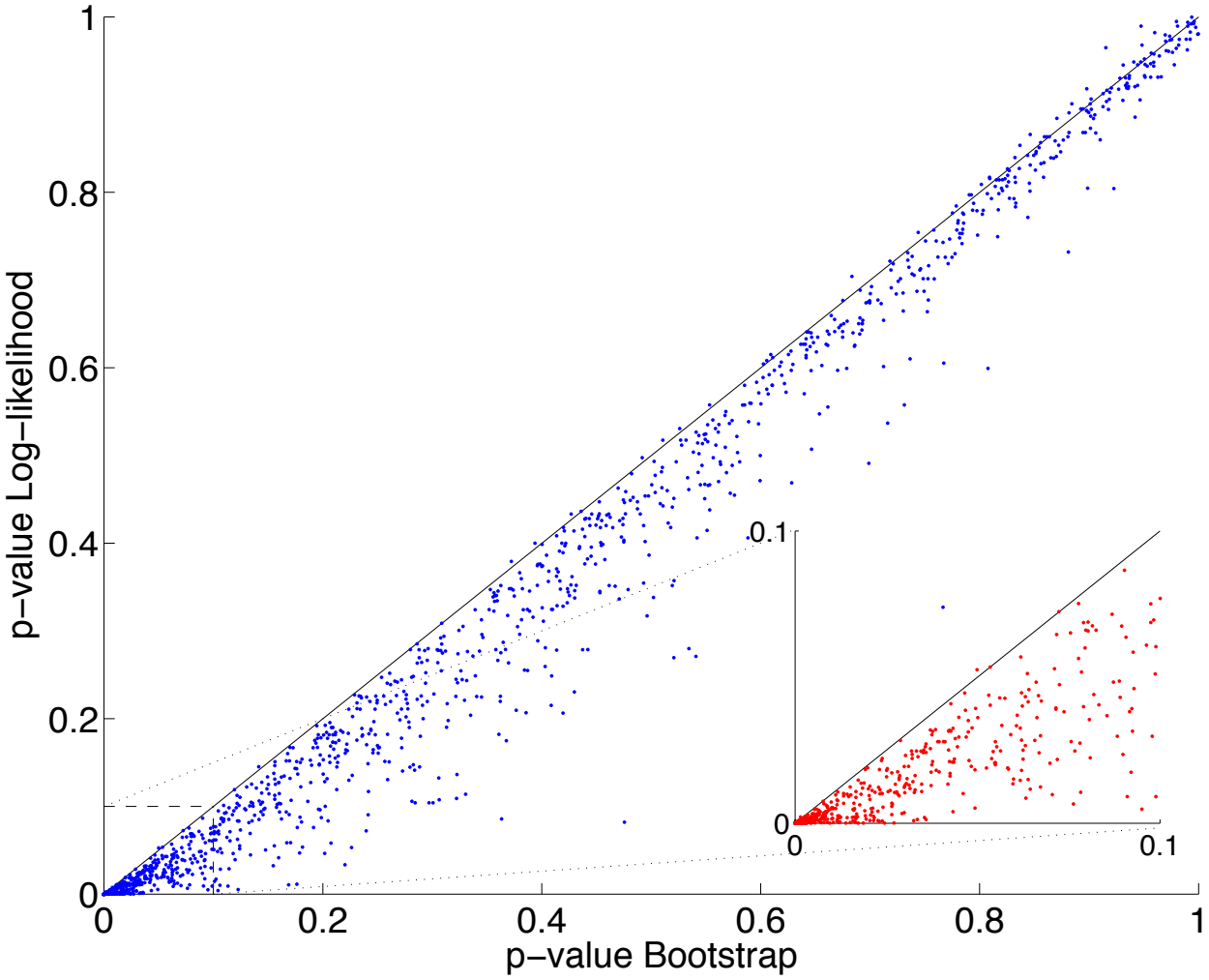
- Words are bursty → do not use bag-of-words model



Data: British National Corpus, 4049 texts

Recent work: Finding keywords (statistical tests)

- J. Lijffijt, P. Papapetrou, K. Puolamäki, H. Mannila. **Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping.** ***ECML-PKDD 2011.***
- Journal version about to be submitted.



Recent work: Finding keywords (statistical tests)

- $\alpha \leq 0.01$ in a text of 2000 words

Word	Freq in BNC ($x10^6$)	Weibull β	Binomial test	Bootstrap test
a	2.2	1.01	61	72
for	0.9	0.93	29	37
I	0.9	0.57	29	110

- β is the shape parameter of the Weibull fit
- Smaller β gives larger differences

Upcoming/now: Burstiness and inter-arrival times

- Segmentation of text
  - Find change-points in sequence of inter-arrival times
- Clustering of words
  - We know burstiness is related to parts-of-speech
  - Can we group words based on inter-arrival times?
- Stochastic model for inter-arrival times
  - Current best (Weibull) does not fit that well
  - How to take into account correlations

Upcoming/now: Burstiness and inter-arrival times

- Relating diseases, genes and tissues through text mining of scientific articles (PubMed abstracts)
  - Disease → key-words → genes
  - Collaboration with Hautaniemi Lab

Computational methods for complexity and burstiness in natural language

- Early work
  - Classification and clustering of text
  - Word sense disambiguation
- Later work
  - Burstiness and inter-arrival times of words
  - Finding key-words: statistical testing for linguistics
- Upcoming/now
  - Clustering words based on inter-arrival times
  - Segmentation of text based on inter-arrival times
  - Stochastic models for inter-arrival times
  - Relating genes and diseases through key-word analysis