Local and global lexicon: a novel approach to quantifying persistence

Jefrey Lijffijt

Department of Information and Computer Science
Aalto University

Introduction

- Development of data mining tools for study of language change
  - Joint work with Tanja Säily, Terttu Nevalainen

- Today: Repetition
  - General measure: Type/token ratio (over time)
  - Localization: Significance testing
  - Understanding: Unexpected repetition

Motivation

- Persistence of linguistic structures
  - Both conscious and subconscious (priming)
- Ample evidence for lexical and syntactic persistence (Bock 1986, Pickering & Branigan 1999), also from corpus linguistics (Gries 2005, Szmrecsanyi 2005, Dubey et al. 2008)
  - Always specific questions, never bottom up
- Frequency(word, text) = Author + Topic + Priming
  - Too hard → Deep understanding of word frequency needed first

Repetition at text level

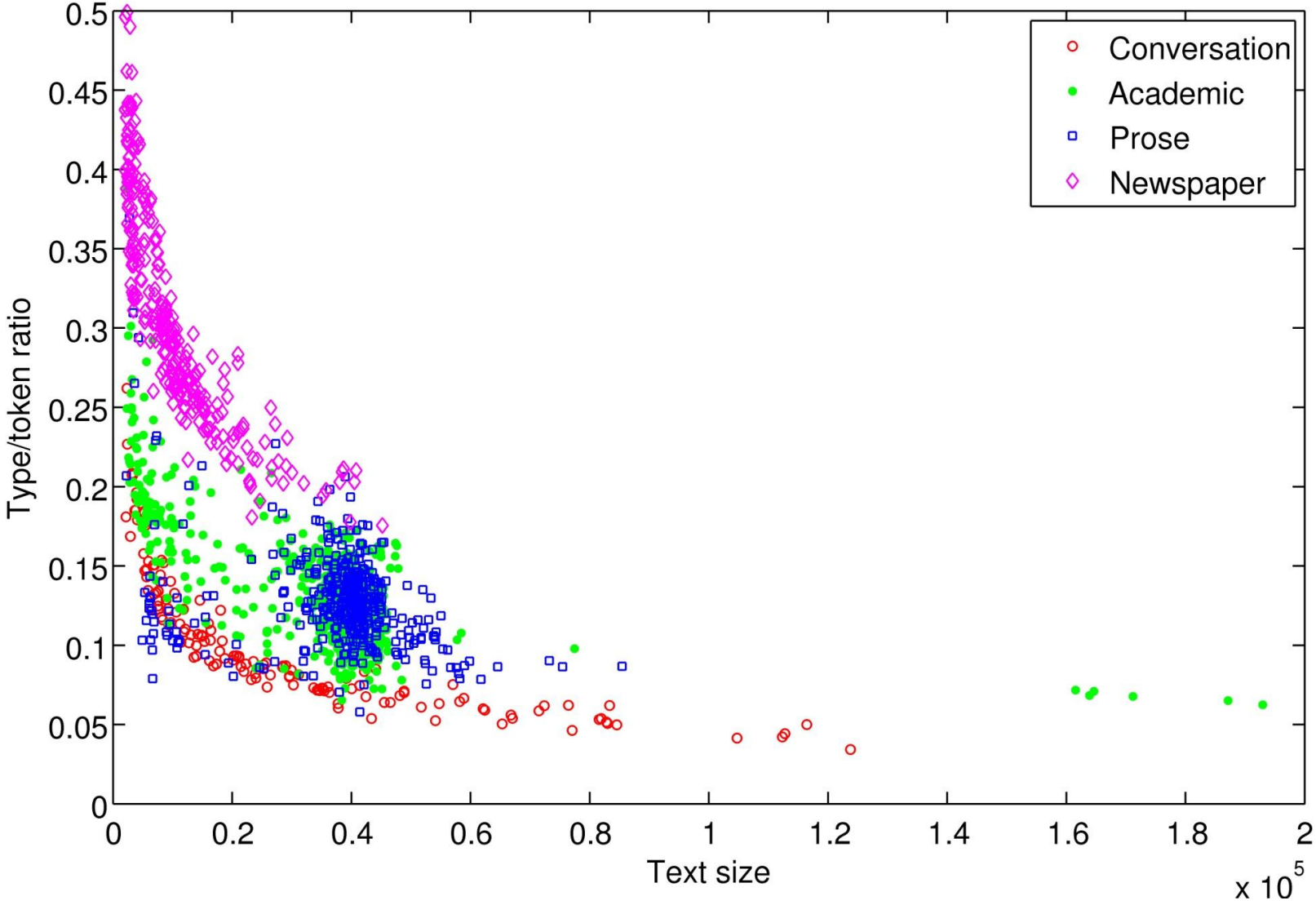
- Type/token ratio (TTR) =

  # unique words / # words

- General measure for amount of repetition in a text

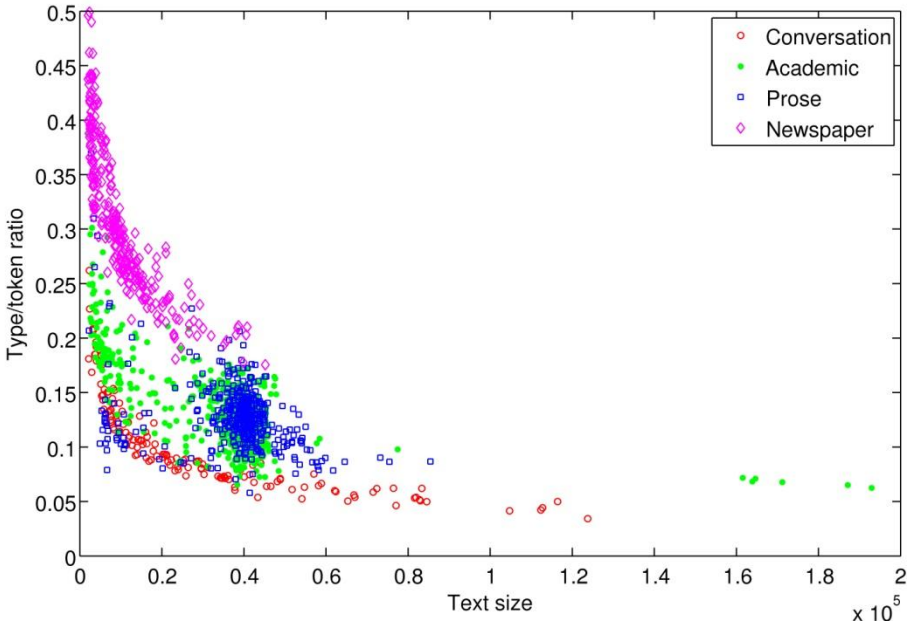
- British National Corpus (BNC-XML)
  - Plain words
  - Ignore punctuation, capitalization

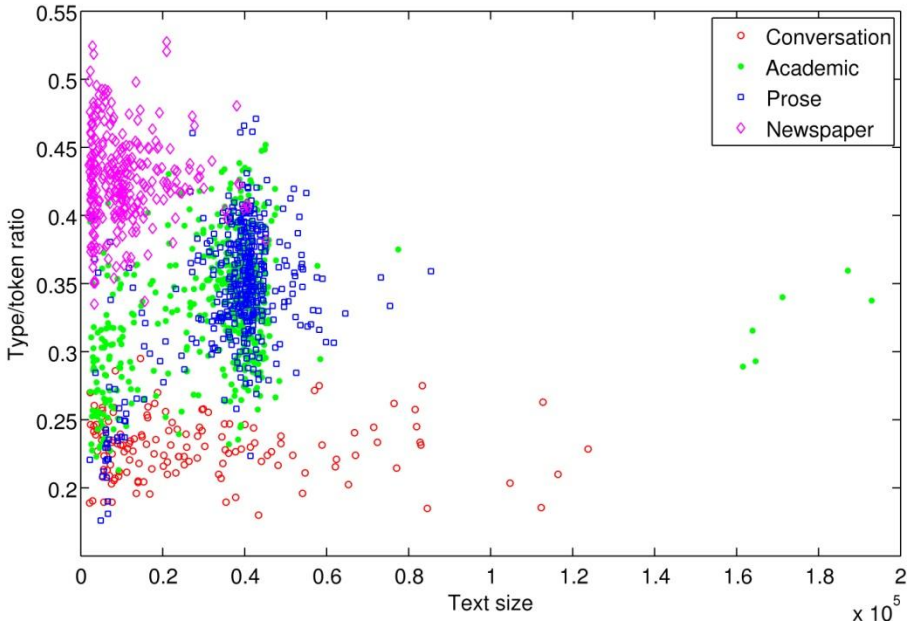
Repetition depends on context



Sampling

- Solve text length bias by using sampling
  - For example 100 samples of 2,000 words





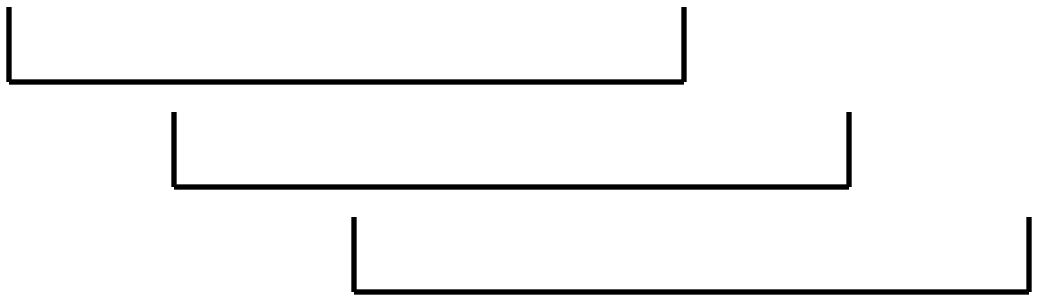
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

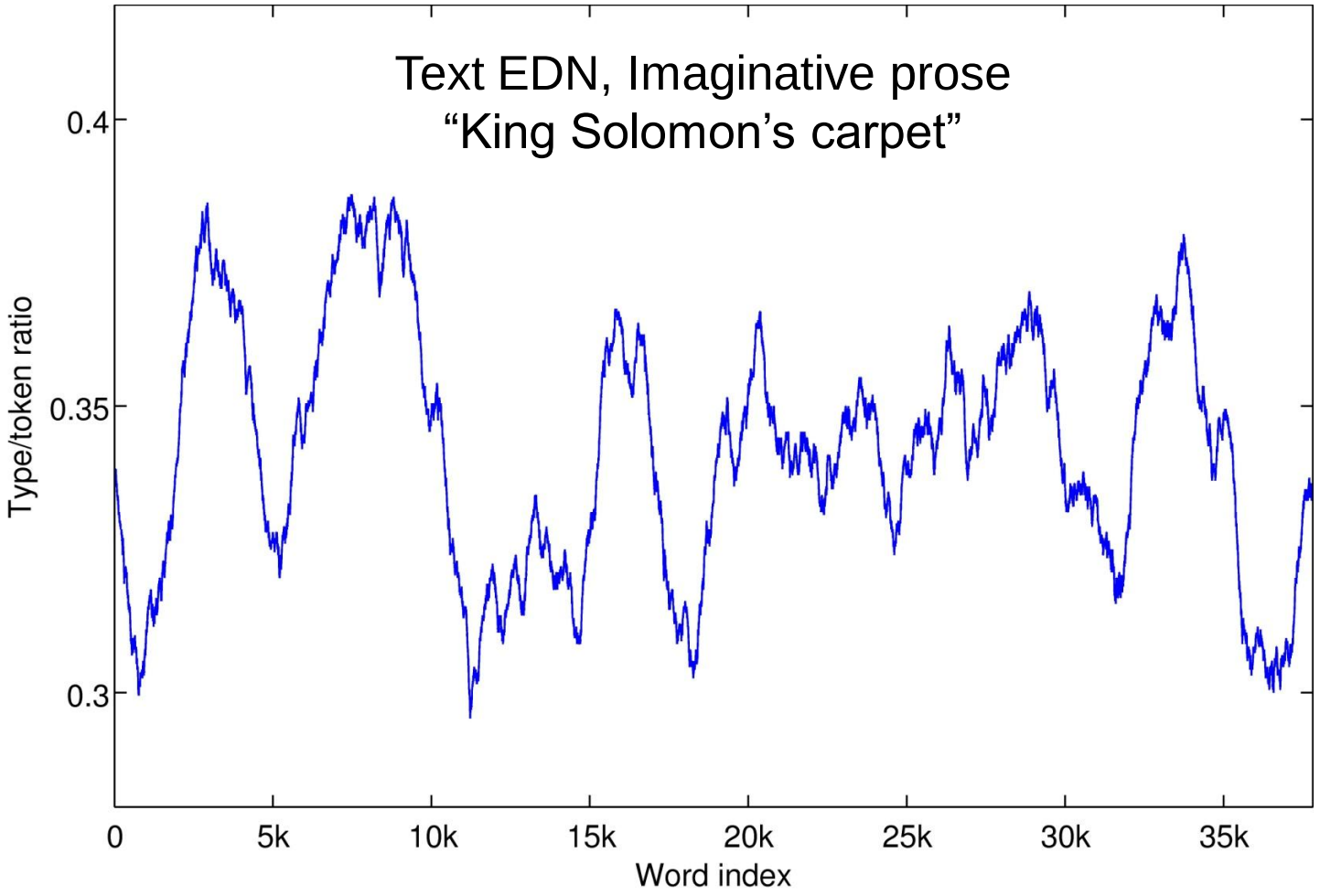
TTR over time

- Text as sequence of words
- Samples of 2000 words
  - Incremental sliding window

- $(w_1, w_2, w_3, \ldots, w_{2000}, w_{2001}, w_{2002}, \ldots)$



TTR over time



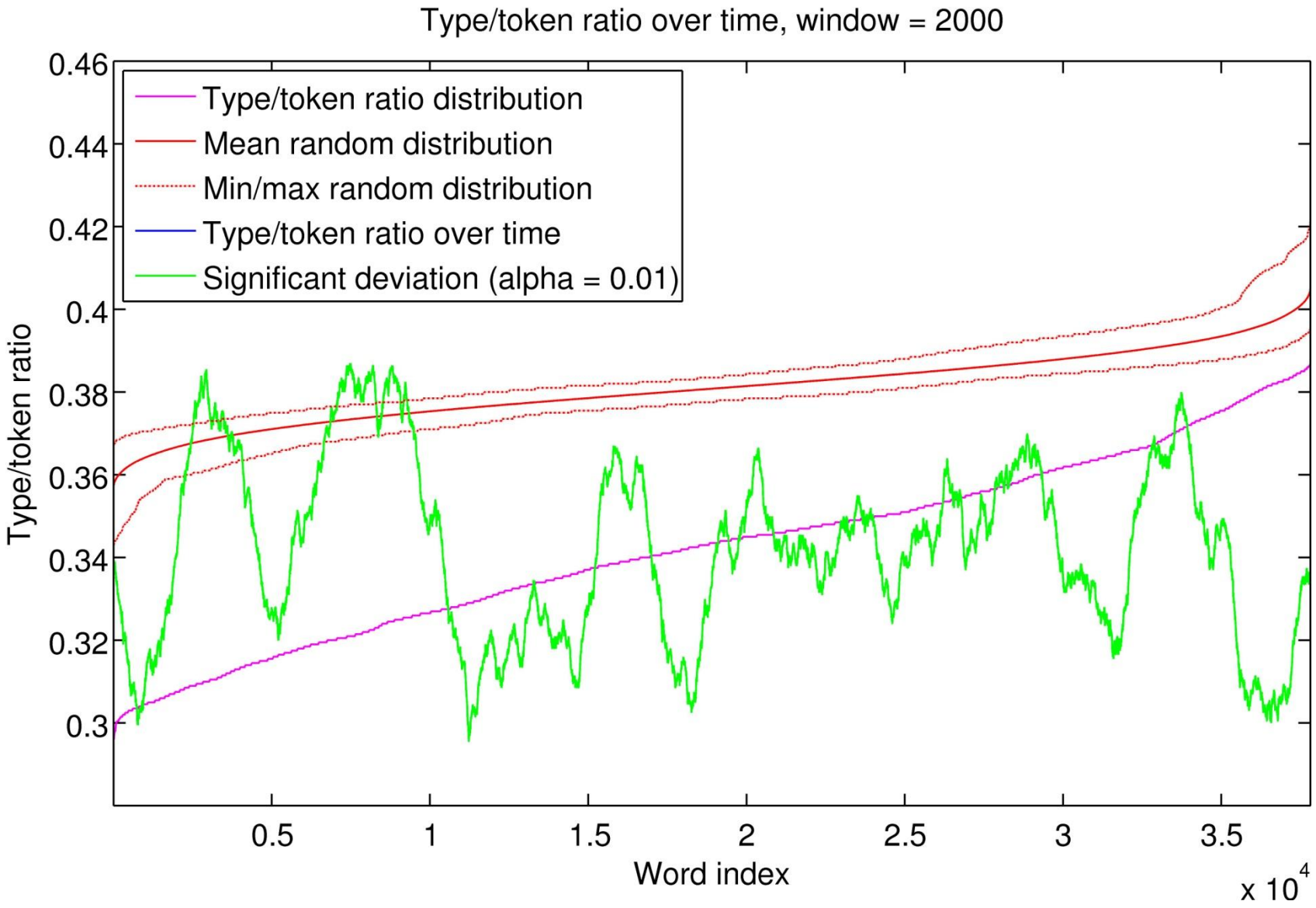
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

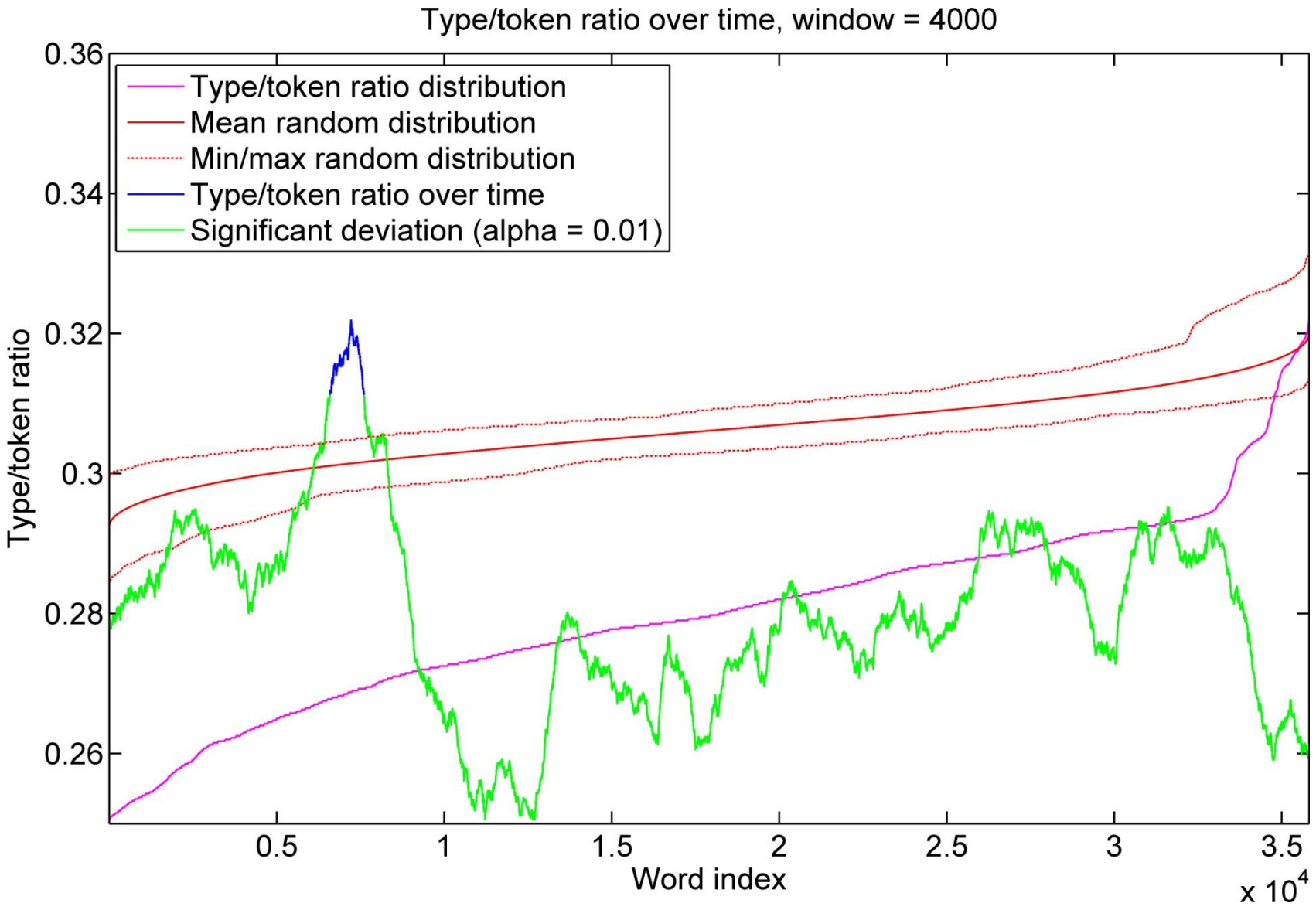
Significance testing

- Randomization approach
  - Produce 500 graphs based on random permutation
    - Equal frequency for every word
    - 1 sample = 1 graph
  - Dependency between windows
    - Compare equal ranks (Lowest vs. lowest points, highest vs. highest points)
  - Multiple hypothesis: Benjamini-Hochberg

Significant deviation



Significant deviation



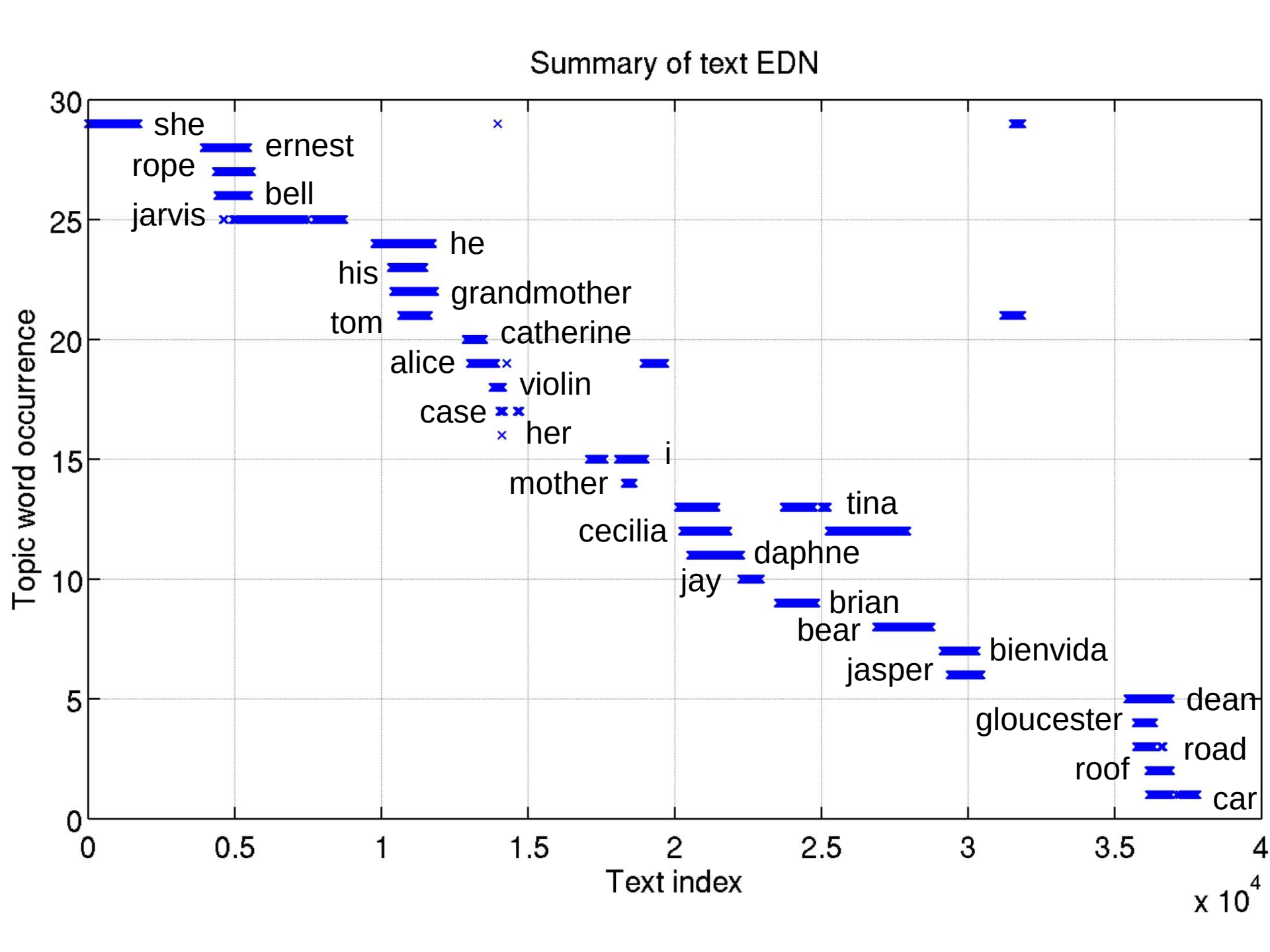
What does this mean?

- Word frequency distribution inside a text changes over time!

- Can we explain the significant repetition?
- Most frequent words account for most repetition
  - This is also expected in randomized text
  - Significant drop → unexpected repetition

Finding unexpected repetition

- Test probability for each word for each window
  - Likelihood function (Kleinberg 2004)
  - Likelihood ratio (Dunning 1993)
  - Chi-square inaccurate for very large or very small samples (Rayson et al. 2004)
- Likelihood ratio seems good choice
  - P(freq | prob-global) / P(freq | prob-local)


Summary of text EDN
Topic word occurrence
Text index
x 10^4
0
5
10
15
20
25
30
0
0.5
1
1.5
2
2.5
3
3.5
4
she
ernest
rope
bell
jarvis
he
his
grandmother
tom
catherine
alice
violin
case
her
i
mother
tina
cecilia
daphne
jay
brian
bear
bienvida
jasper
dean
gloucester
road
roof
car

Conclusion

- Significant repetition can be found using randomization approach
- Local explanation can be given using likelihood ratio
- We can construct a timeline of unexpectedly frequent words to summarize a text
- Open questions:
  - What is a good cut-off value for likelihood ratio
  - Can we detect topic shift in a text?