Computational semantics for the humanities

Diarmuid Ó Séaghdha

Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge
do242@cam.ac.uk

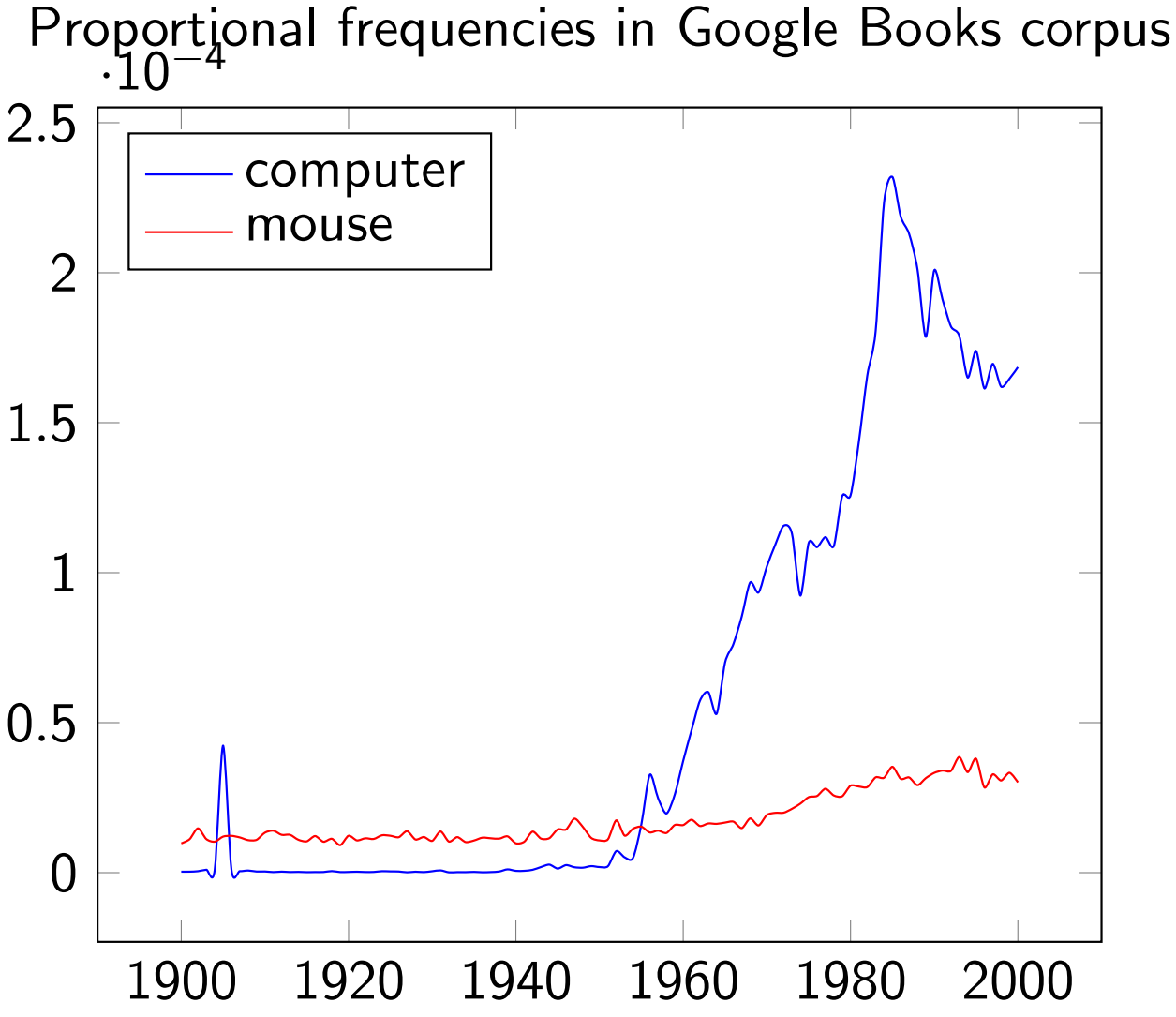
Translation and the Digital
25 April 2014

UNIVERSITY OF CAMBRIDGE

Introduction

- "Big Data" revolution:
  - We have access to more textual data than any human could ever read.
  - We can perform *some* kinds of automated analysis over large datasets.
- For humanities researchers:
  - Data mining is a tool that facilitates asking questions about language use.
  - Data mining is not a question or an answer.
- Natural Language Processing (NLP) research gives us computational methods for analysing and interpreting text.

Corpus frequency



Semantics: The distributional hypothesis

- Imagine that *tezgüino* is a rare English word, and you saw the word used in the following sentences:
  1. A bottle of *tezgüino* is on the table.
  2. Everyone likes *tezgüino*.
  3. *Tezgüino* makes you drunk.
  4. We make *tezgüino* out of corn.

(Lin, 1998)

- Can you guess what *tezgüino* means?
- What kind of things do you expect will be similar to *tezgüino*?
- **The Distributional Hypothesis:** Two words are expected to be semantically similar if they have similar patterns of co-occurrence in observed text.

Co-occurrences and similarity

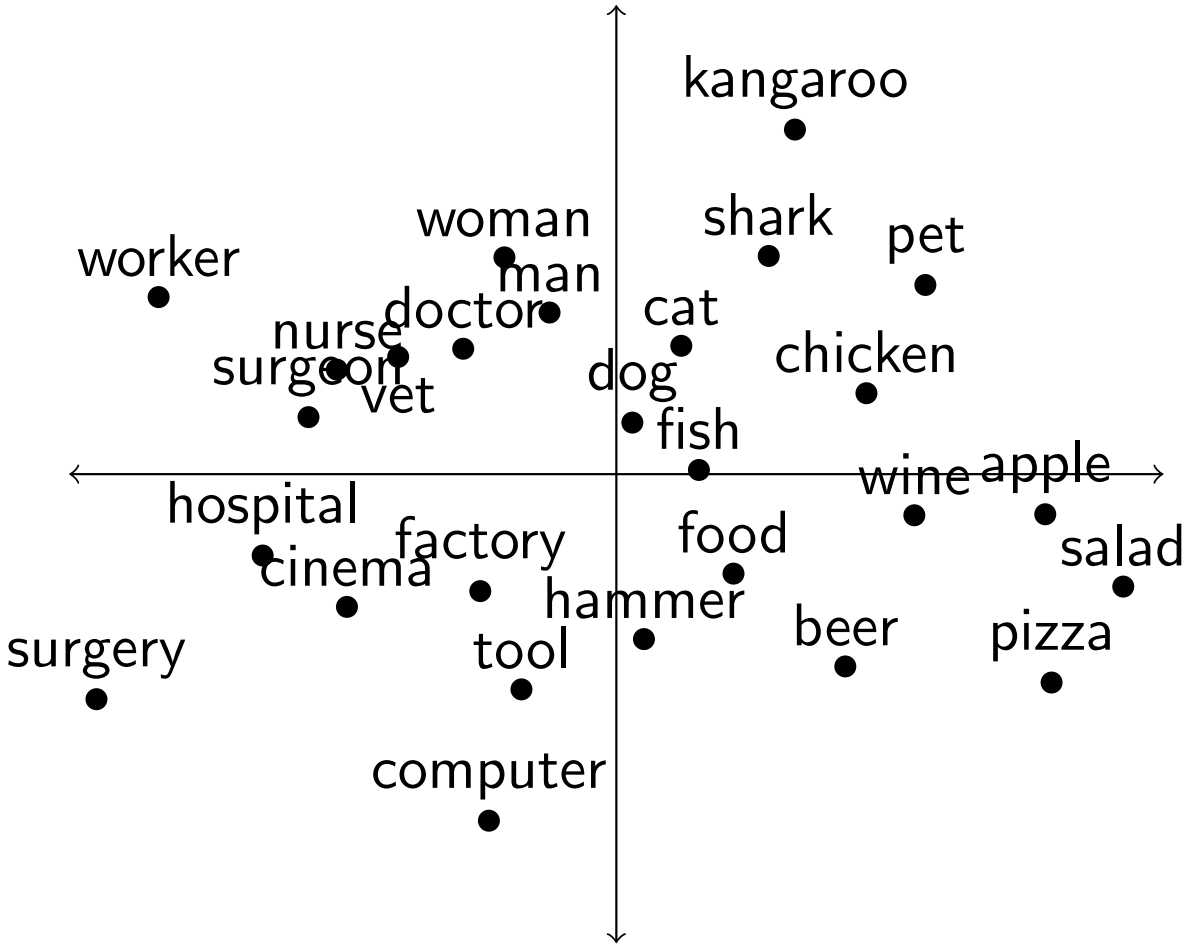
- We can produce a distributional "profile" of a word from a corpus:

  farmer: *part-time, sheep, peasant, tenant, wife, crop*, . . .
  doctor: *nurse, junior, prescribe, consult, patient, surgery*,. . .
  hospital: *psychiatric, memorial, discharge, admission, clinic*, . . .

- We can compute similarity between words by comparing their profiles.

Semantic space visualisation

British National Corpus, top 5000 dependencies



Discovering semantic classes

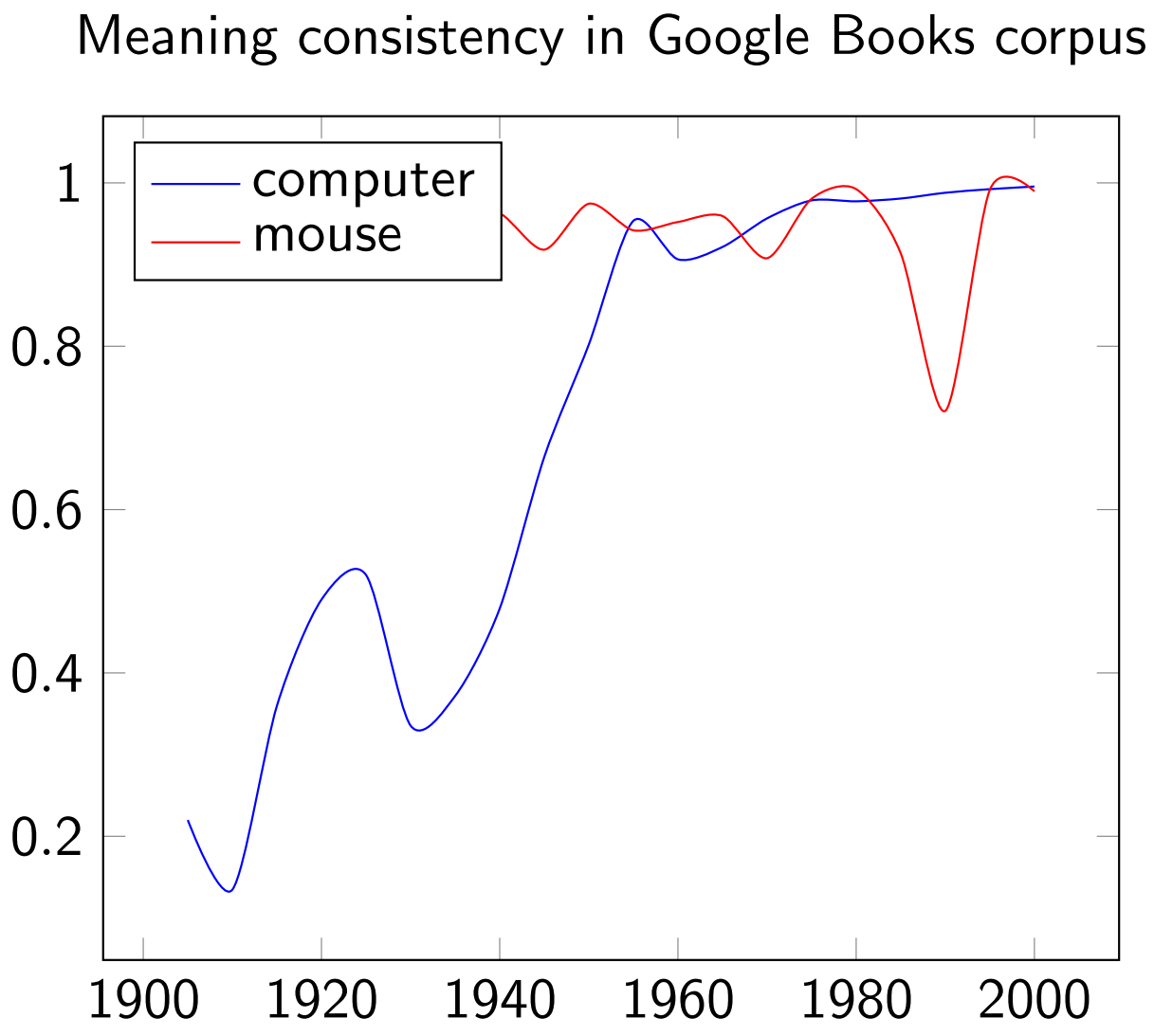
BNC nouns, method related to Latent Dirichlet Allocation (topic modelling)

Class 1	Class 2	Class 3	Class 4
attack	test	line	university
raid	examination	axis	college
assault	check	section	school
campaign	testing	circle	polytechnic
operation	exam	path	institute
incident	scan	track	institution
bombing	assessment	arrow	library
offensive	sample	curve	hospital

Tracking meaning over time

- Ongoing project (with Meng Zhang)
- We know that language changes over time.
- Words change their meaning by adding and losing senses and associations.
- Can we study this behaviour in a large corpus?
- Goal: "word biographies".
- A historian of ideas might be interested in what a word meant to people at different points in time.

Tracking meaning over time



Conclusion

- We have methods for extracting meaning from document collections:
  - Comparing words and texts
  - Clustering words/concepts
  - Identifying themes in a corpus
  - Identifying associations between words/concepts
- We need users in other fields to provide interesting questions.
- If you have ideas, say hi! Or send me an email at `do242@cam.ac.uk`.