Web as Corpus Workshop

Co-chairs:

Marco Baroni
Adam Kilgarriff
Sebastian Hoffman

"When you have tons of data and tons of computation you can make things work that don’t work on smaller systems"

- Google's VP-engineering, Urs Hölzle

History within CL

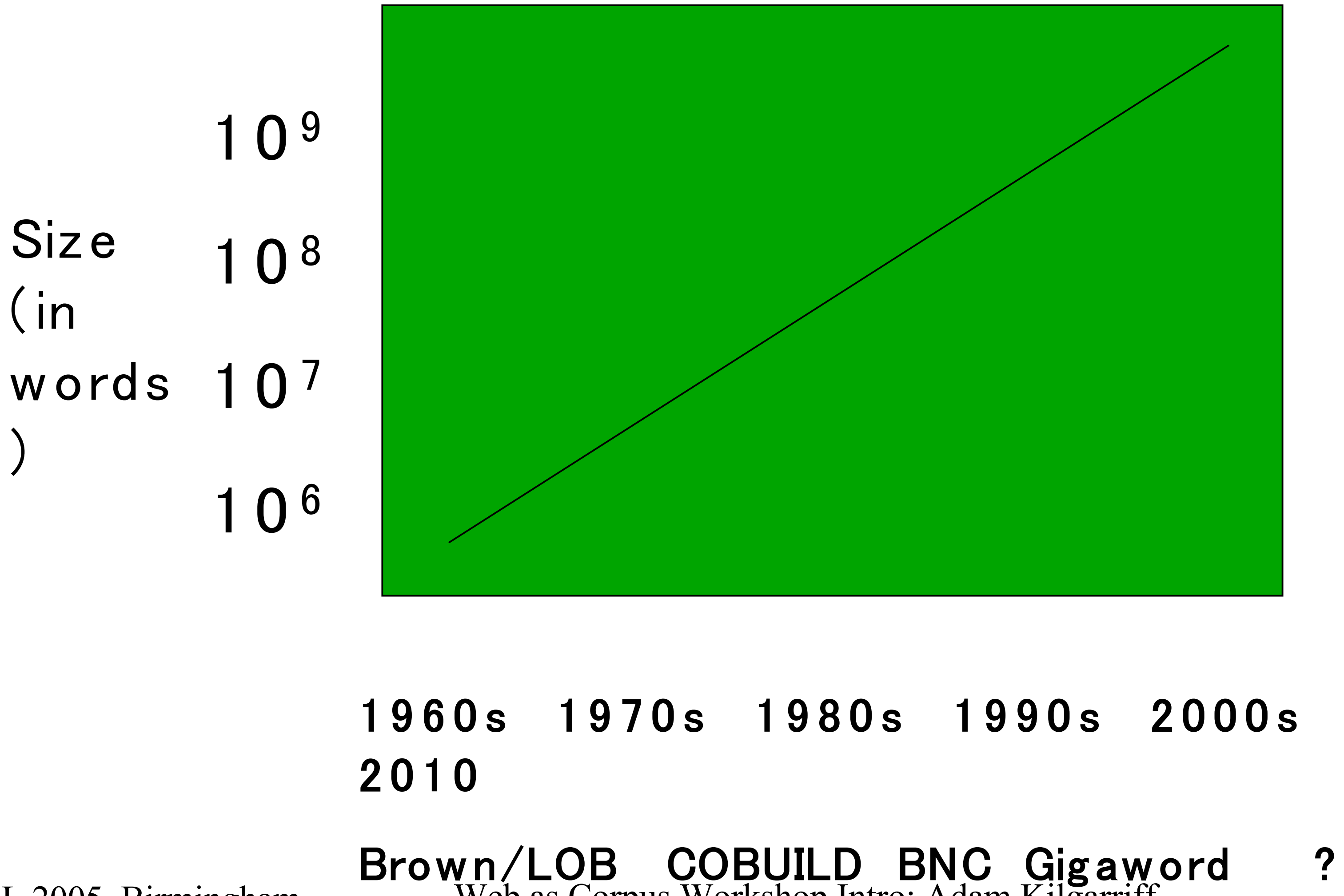
- 1989: corpora arrive on scene
- 1989-1993: “too dirty”: **battles**
- 1993: CL Special Issue: consummation

- …
- 1999: web arrives on scene
- 1999-2003: “too dirty”
- 2003: CL Special Issue
- .

History within CL

- 1989: corpora arrive on scene
- 1989-1993: “too dirty”: **battles**
- 1993: CL Special Issue: consummation
- 1993: WVLC workshop series starts
- …
- 1999: web arrives on scene
- 1999-2003: “too dirty”
- 2003: CL Special Issue
- 2005: WAC workshop series starts

History


Size (in words)
10^9
10^8
10^7
10^6
1960s
1970s
1980s
1990s
2000s
2010
Brown/LOB
COBUILD
BNC
Gigaword
?

Approaches

- Use Google hit counts
- Use snippets
- Use google, then download pages
- Spider from relevant starting sites

(Marco Baroni's analysis)

The Trouble with Google

- not enough instances (max 1000)
- not enough context
  - ca 10-word snippet around search term
- ridiculous sort order
  - search term in titles and headings
- linguistically dumb
  - not lemmatised
    - think/thinks/thinking/thought: four searches
  - not POS-tagged
    - mixes up beat (n) and beat (v)
  - and why not parsed

DIY

- do it ourselves
  - this community
- Wacky

Components

- web crawler
- filters/classifiers
  - language id, non-text, boilerplate, genre
- linguistic processor (optional)
- database/indexing
- statistical summariser (optional)
- user interface.

Programme

9.30	Welcome, goals *Adam Kilgarriff*
10.00	Crawling *Marco Baroni*
10.30	***coffee***
11.00	Creating specialized and general corpora using automated search engine querying *Marco Baroni and Serge Sharoff*
12.00	Small groups: what we have all been doing
1.00	**lunch**
2.30	Processing web-derived text *Sebastian Hoffman*
3.15	Indexing and interfaces *Stefan Evert and Adam Kilgarriff*
4.00	**coffee**
4.30	Representing genre-specific websites *Alexander Mehler and Rüdiger Gleim*
5.00	Small groups: “what are critical next steps for WaC activity?”
5.30	Plenary: where next?
6.10	**end**

Small groups (proposal)

Around topics:

- wac for theoretical linguistics
- wac for applied linguistics
  - language teaching, translation, terminology
- wac for nlp
- wac for lexicography
- wac for ontology engineering

Around problems:

- large crawls
- text processing, boilerplate removal, etc.
- indexing and interfaces