- What counts as a dataset?
  - Includes maps, census, English, French, Latin text
  - How big do they need to be?
    - “Half a billion words is nice” (for syntactic analysis)
  - What would be a good reference corpus?
- What do we want to do with historical texts?
  - Can they be turned into a corpus?
    - Need to be digitised and transcribed (text mining feedback loop)
    - Which texts to select to form a representative corpus
      - Select from EEBO
      - to add to transcribed part of EEBO?
  - Find out what Maggie’s words mean? E.g. ‘tantarbawde’
- Have you POS tagged your text? Did it work?
- Where are we with semantic tagging?
- Requirements from historians collaborating with software developers
  - Are tools scalable for large corpora?
  - Cross language support

Disciplines

- Linguistics
  - 17
- Literary and linguistic computing
  - 3
- Computer Science
  - 8
- Computational linguistics
  - 4
- Information science
  - 1
- Humanities
  - 1
- History
  - 1
- Undecided/other/all of the above?
  - 3
- Free lunch
  - 0

End of day one

- How can we get the communities to talk to each other?
- What do we/they need?
  - Simple user interface – going the way of Google
  - Advanced user interface – for those who want to do it the hard way

Size

- Function of intended use
- Product of biases of corpus creators
- CED, CEEC, etc. perfectly fine for historical pragmatics
- Bottom line: not make size a limitation

Accessibility

- Via Web
- Cost (free?? to end-users)
- User interface (may be a much larger and diverse audience than we had first intended)

Searchability

- Spelling, morphology: wildcards
- Syntax: POS tagging
- Semantics: Collocates, comparing collocates, synonyms, user-defined lists
- [Pragmatics]
- Q: Is what we study a function of the corpus, or what we really want to study?

Morning Day Two

- Size limitation: 200M – 700M words?
- Accessibility
  - By machine, web services, Google API
- Access to full text, copyright problems!!
  - Fair use
  - Content providers need to understand what KWIC is
  - Who to talk to? AHRC? JISC?
  - “Money, lawyers and guns”
  - Who can negotiate for us? Academic associations? ALLC, ACH, ACL?

Software requirements

- Language support
  - Multilingual, historical, endangered, those without a written form, future proof
- Standards
  - (fixed) Unicode, XML
  - (fluid) Tools, API, Transcription, Annotation
  - (Backward compatibility) Storage
- Tools
  - Static corpora and researcher's data
  - Not just text
  - Toolbox
  - Ease of use (programming interface, graphical interface or ? for those who don't use email)
  - Accessibility (Google API, web services)
  - Don't talk about the tools when presenting them, do it on paper first
- Scalable from 100K to 1 billion words
  - To be used by computational AND corpus people
- Email Lisa Lena!!!!!

Where to next?

- From the workshop?
- For this community?
- For other communities?
  - History (of science, social), archaeology, politics
- Possibilities:
  - Edited collection
  - Journal special issue
  - Slides online
  - Training materials / curriculum
  - Mailing list / online discussion
  - The next workshop?