Corpora Past, Present and Future

Aaron Ecay and Susan Pintzuk
University of York

Before Electronic Corpora

Two methods of historical linguistic research:

- Research based on scholars’ impressions of the general patterns in the language (e.g. Canale 1978, Mitchell 1985, Roberts 1985)
  - Problem: are the results representative of the language as a whole?
- Manual searches of substantial text samples, keeping counts, tracking variation, creating small electronic corpora (e.g. Bean 1983, Kohonen 1978, Pintzuk 1991, Taylor 1994)
  - Problem: tedious, time consuming, error prone

Before Electronic Corpora

More serious problems:

- Results are generally not reproducible
- Difficult to determine whether results are representative of the language as a whole
- Necessitates considerable time and effort for quantitative results (for syntactic research, a minimum of one million words)
- Small corpora that are produced are inadequate for subsequent studies, because of the focus of the original study
- Low frequency data do not appear in the corpus

Electronic Corpora for Historical Texts

History of English

- *The Helsinki Corpus of English Texts* (1991, Matti Rissanen and his team)
- *Dictionary of Old English Corpus* (starting 1986, Antonette Healey and her team)

Advantages:

- electronic form
- searchable by word processing software, e.g. Word; electronic text viewers, e.g. WordCruncher; on-line web tools, e.g. created for DOE Corpus

Unannotated Electronic Corpora

Although searchable, they had disadvantages:

- low precision errors (getting too much data)
- low recall errors (getting too little data)
- searches are possible only for lexical items and their combinations

Electronic Corpora

1. Low precision error (too much data): searching for demonstratives in noun phrases
   a. I like *that* hat.
   b. I like *that*.
   c. I know *that* he bought a hat.
   d. I like the hat *that* he bought.

Electronic Corpora

2. Low recall error (too little data): searching for relative clauses

   a. the hat *that* he bought

   b. the hat *which* he bought

   c. *which* hat did he buy? (low precision)

   d. the hat he bought (low recall)

Electronic Corpora

One solution to low precision errors:
part-of-speech tagging

3. a. I like *that*/D hat/N.
   b. I like *that*/D.
   c. I know *that*/C he bought a hat.
   d. I like the hat *that*/C he bought.

Annotated (Parsed) Corpora

The real solution to both low recall and low precision errors is morpho-syntactic annotation: morpho-syntactically annotated corpora ('parsed' corpora) contain structural information not available from words and their strings.

Parsed corpora make possible the search for abstract constructions:

- relative clauses
- clauses that begin with a temporal NP
- clauses that contain both a direct and an indirect object

Parsed Corpora

4. Search for relative clauses:

   a. $[_{\text{NP}}$ the hat $[_{\text{CP-REL}}$ that/C he bought]]

   b. $[_{\text{NP}}$ the hat $[_{\text{CP-REL}}$ which/WH he bought]]

   c. $[_{\text{NP}}$ which/WH hat] did he buy?

   d. $[_{\text{NP}}$ the hat $[_{\text{CP-REL}}$ he bought]]

Parsed Corpora

5. Double object construction versus object + temporal phrase:

   a. I gave [$_{\text{NP-OB2}}$ John] [$_{\text{NP-OB1}}$ the book]

   b. I met [$_{\text{NP-OB1}}$ John] [$_{\text{NP-TMP}}$ last week]

Parsed Corpora

In the 1980’s and 1990’s, parsed corpora existed for Modern English, mainly through the Penn Treebank Project, but not for historical languages and texts.

Kroch and Taylor to the rescue!

Parsed Corpora of Historical English

- Old English prose (850-1150)
- Old English poetry (850-1150)
- Middle English prose (1150-1500)
- Early English correspondence (1400-1700)
- Early Modern English (1500-1700)
- Modern British English (1700-1900)

Parsed Corpora

Disadvantages:

- difficult and time-consuming to create
- annotators must have
  - ˜ knowledge of the language
  - ˜ strong syntactic background
  - ˜ strong general linguistic background

Parsed Corpora

Are parsed corpora good value for money?

YES !!!

Parsed Corpora

The feasibility of corpus-based linguistic research depends on

- the availability of sufficient data
- the straightforward retrieval of relevant information from the corpus

Parsed corpora give us both.

Parsed Corpora

Types of annotation

- part of speech
- organisation of phrases and clauses
- functional information (e.g. grammatical role, clause type)
- semantic relations (e.g. theme, goal, experiencer)
- information structure (e.g. topic, focus)

Parsed Corpora

```
/~*
he is in a great hurry to be married (ID AUSTEN-180X,162.53)
*~/
( (IP-MAT (NP-SBJ (PRO he))
          (BEP is)
          (PP (P in)
              (NP (D a)
                  (ADJ great)
                  (N hurry)
                  (IP-INF (TO to)
                          (BE be) (VAN married)))))
  (ID AUSTEN-180X,162.53))
```

Parsed Corpora of Other Languages

- Tycho Brahe Corpus, a parsed corpus of historical Portuguese - Charlotte Galves (University of Campinas, Brazil) and collaborators
- Modéliser le changement: les voies du français, a parsed corpus of historical French – France Martineau (University of Ottawa) and collaborators
- Icelandic Parsed Historical Corpus (IcePaHC) - Eiríkur Rögnvaldsson (University of Iceland) and collaborators
- The Parsed Old and Middle Irish Corpus (POMIC) - Elliot Lash (Dublin Institute for Advanced Studies: School of Celtic Studies)
- and more, from students at Penn and others (e.g. Old Saxon)
- and more using dependency parsing, e.g. The New Testament in Greek, Latin, Gothic, Armenian and Old Church Slavonic (PROIEL, Dag Haug)

The future of parsed corpora

- Manual parsing of parallel texts
- Convergence of different corpus traditions
- More capable annotation strategies
- Automated analysis

Manual parsing of parallel texts

- Collaborative project under the aegis of the Centre (collaborators at York, Newcastle, Campinas)
- Parsing Calvinist/early Protestant translations of the New Testament in 12 languages
  - 6 Germanic
  - 4 Romance
  - Welsh, Basque

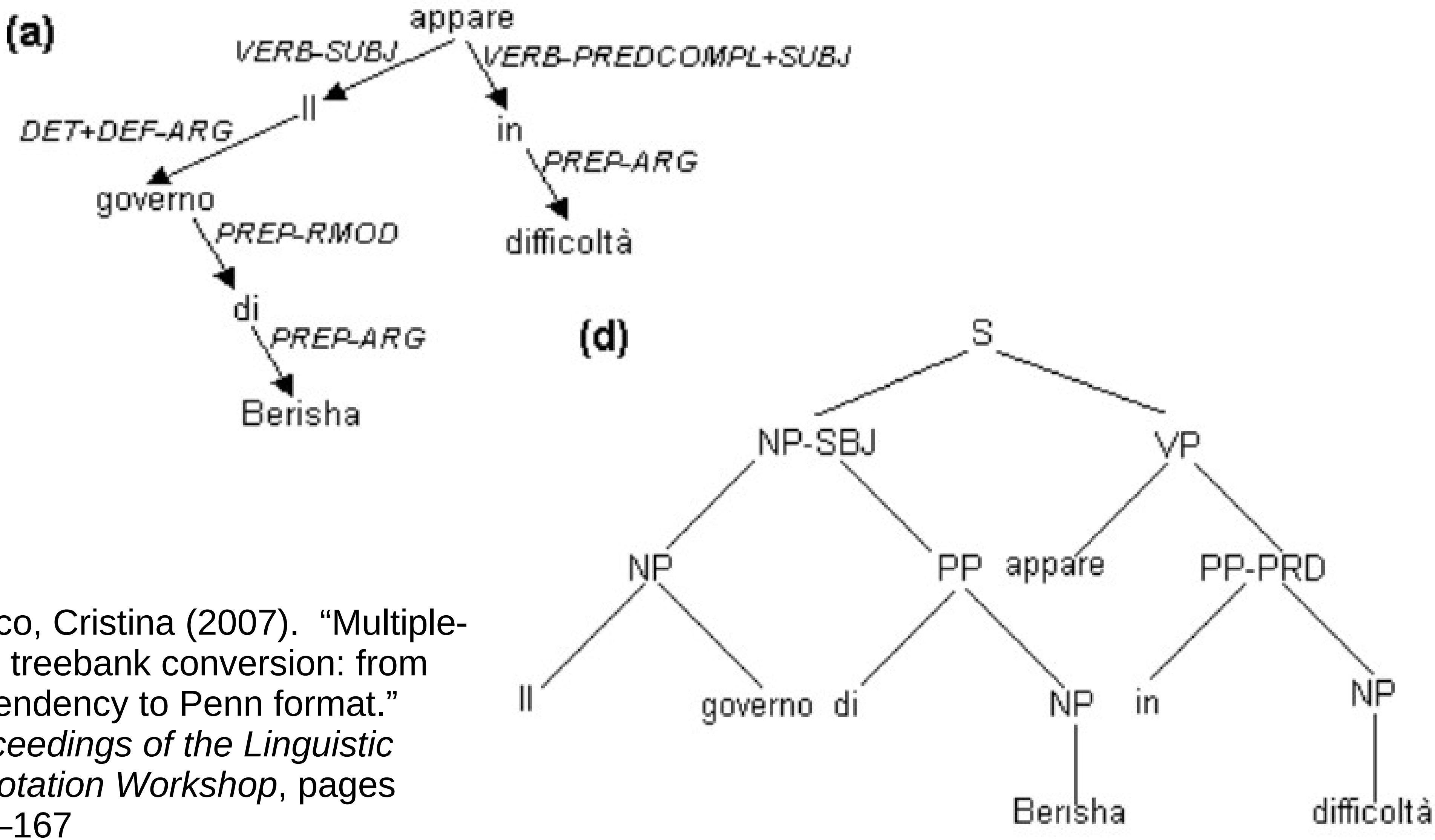
Parallel parsed corpora

- Empirical goals
  - Study information structure cross-linguistically
  - Gain insight into the effect of information structure and syntactic change within and across families
- Durable byproducts
  - “Seed” parsed corpora of ~8 new languages to enable futher research (training sample, annotation guidelines)
  - Techniques and computer programs for querying parallel parsed corpora

Convergence of corpus traditions

- Dependency parsing is winning (has won?) in Computer Science
  - Easier problem
  - More tractable algorithms
- Google has dependency-parsed their corpus
  - 345 billion words!
  - Tagset size = 12, unchecked OCR, occasionally flaky metadata, data isn't accessible, …
  - But they did it!

Dependency vs. phrase structure annotation



Bosco, Cristina (2007). “Multiple-step treebank conversion: from dependency to Penn format.” *Proceedings of the Linguistic Annotation Workshop*, pages 164–167

Dependency and phrase-structure annotation

- Phrase structure annotation is a strict superset of dependency annotation
  - But the differences are small
- It is possible to translate a dependency corpus into a phrase-structure one
  - Centre collaborators at Penn (Tony Kroch, Seth Kulick) are working on an automatic parser which uses dependency algorithms “under the hood,” but translates to phrase structure output

More capable annotation strategies

- There is lots of linguistic information that parsed corpora may not capture
  - Morphology: gender/number/case, inflection class
  - Information structure: old/new information, pronoun/anaphor reference
  - Lexical: lemma

Extending parsed corpora

- Work is ongoing to add these features to existing corpora
  - YCOE has case information
  - Icelandic corpus has case and lemma
- Ongoing project at Penn to add lemmata to the PPCHE
- Standardization of the annotation formats

Automated analysis of large corpora

- EEBO/ECCO: large database of scanned English texts from 1473-1800
  - EEBO is more complete, contains ~1/3-1/2 of all printed material up to 1700 => 48k texts
  - ECCO has “only” 2.5k texts
  - Together: 1B words
- TCP: OCR + manual correction of EEBO/ECCO
- Half of this corpus is publically available (as of Jan 1); other half only available to partner institutions

Using these corpora for research

- An unannotated corpus isn't much use...
- POS tagging is a “solved problem”
  - The best algorithms (under controlled conditions) do about as well as very good humans
- We have large training corpora (PPCEME, PPCMBE)

Challenges and opportunities

- Have local information (POS tags, adjacency); no long-distance information (wh-dependencies, subject/object, …)
- If you're lucky enough to study a phenomenon that's local: ☺
  - Do support in negative declaratives:
    - PRO DO NEG V
    - PRO V NEG
  - Make sure that: PRO=nominative
  - Exclude certain constructions: “not only,” “not to VB”
    - (Others will sneak by: “not X but Y”)

Semi-local approach

- If you're studying something that's a little non-local, generate and filter
  - Do-support in affirmative polarity questions
    - DO PRO[nom] V
    - V PRO[nom]
    - Collect all instances of the patterns and manually discard the non-questions

Truly non-local information

- If you need really long-distance information, this corpus won't help you very much
    - E.g. subject vs. object relatives

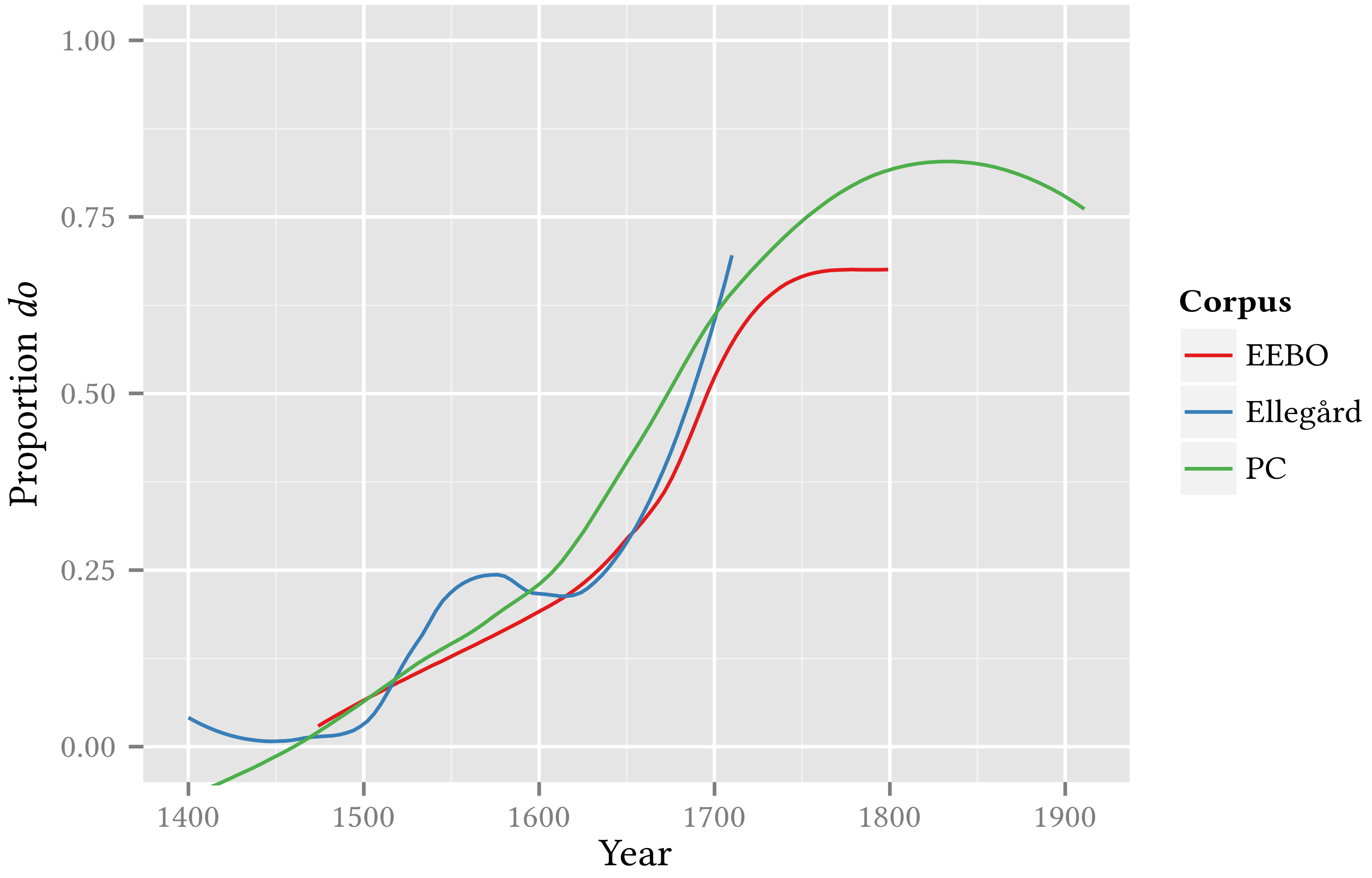
Plans for the development of the corpus

- Since half of the TCP corpus has recently been publically released, this tagged dataset can be made public
  - First public release
  - Clean up POS tagging
  - Generate some text metadata

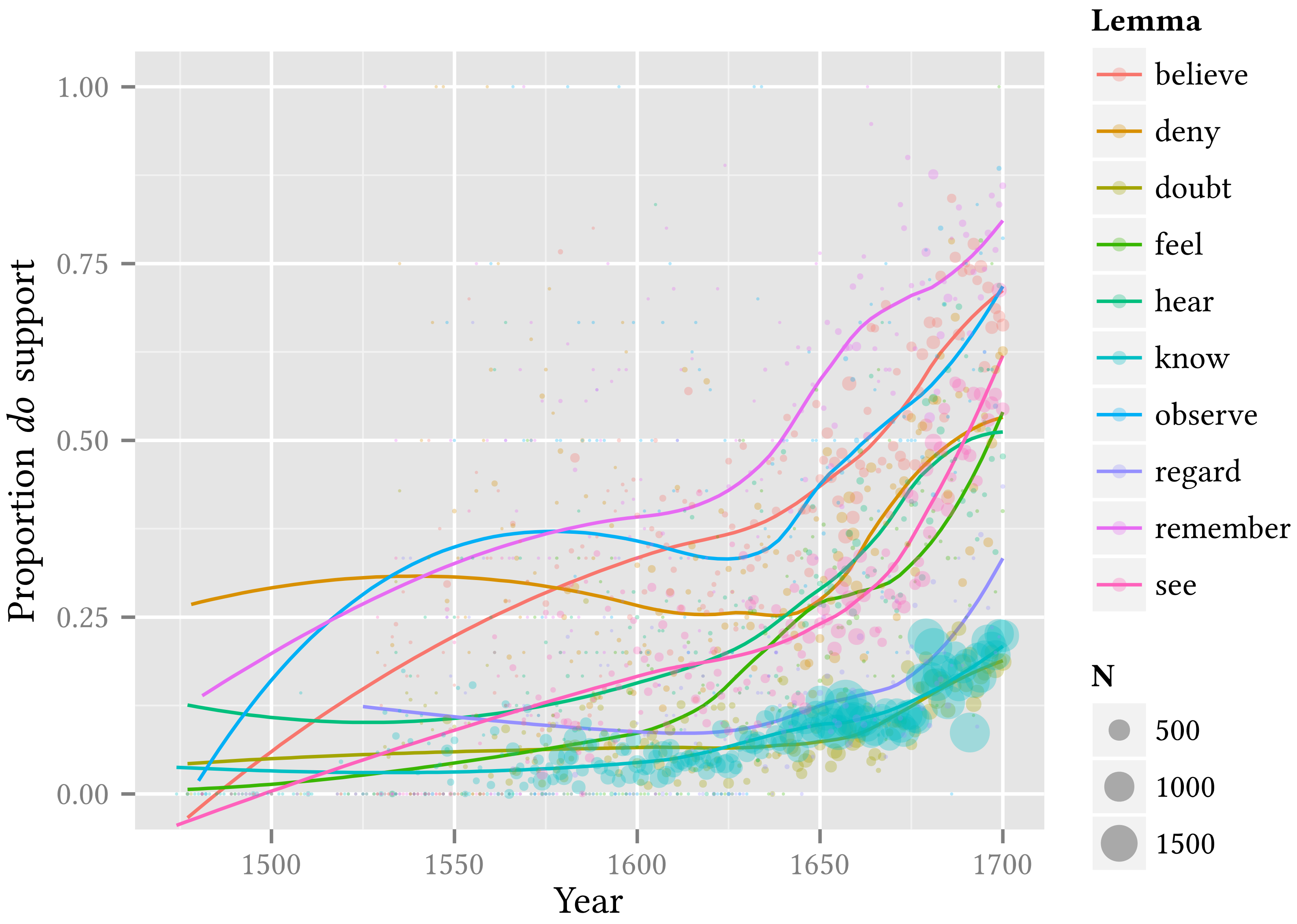
Longer-term plans

- Create more detailed text metadata
  - Author birthplaces/residences → geographic map of changes' spread
- Lemmatize the corpus
  - Tedious, orthography-based … but users shouldn't have to reinvent the wheel
- Automatically parse the corpus
  - At least with a dependency annotation scheme

Results from the corpus



Results at the lexical level



Conclusions

- There are multiple projects underway in the Centre which will enhance the breadth and depth of available corpora
- The success of these projects depends on an intellectually diverse and lively set of collaborators
- And on making the results accessible, to formal linguists as well as colleagues in other disciplines

Questions?