Diachronic syntax with and without parsed corpora

Rob Truswell, University of Edinburgh
rob.truswell@ed.ac.uk

CliS, Edinburgh, 2/12/16

Roadmap

1. Diachronic syntax
2. Parsed historical corpora
3. Getting more out of corpora

Section 1

Diachronic syntax

Types of change

- Language change occurs at a range of levels.
  - Lexical
  - Phonological
  - Morphological
  - Syntactic
  - Semantic
  - ...
- Research into different types of change raises different challenges.
- One commonality: reliance on distributional evidence.

Challenges for diachronic syntax

1. **Size**: syntactic units are bigger than lexical/phonological/morphological units. Reliable results require more data.
2. **Abstractness**: syntactic change typically involves changes in structural representations, only indirectly reflected in the attested forms.
3. **Speed**: syntactic change can happen very gradually, and very slowly.

- These challenges apply at least equally acutely to diachronic semantics.
- This is unsurprising, as most successful diachronic semantic research is grounded in distributional evidence.

Size

- *Ormulum*: late 12th century verse homilies (c.20k lines).
- Significant as one of very few lengthy early Middle English texts.
- Sample in Penn–Helsinki Parsed Corpus of Middle English (2nd edition, Kroch & Taylor 2000) contains:
  - 238106 characters (for the phonologists).
  - 52593 words (for the morphologists).
  - 2421 sentences (for the syntacticians).
- To make matters worse, there are arguably more degrees of freedom in syntactic structure than phonological or (at least inflectional) morphological structure.

Abstractness

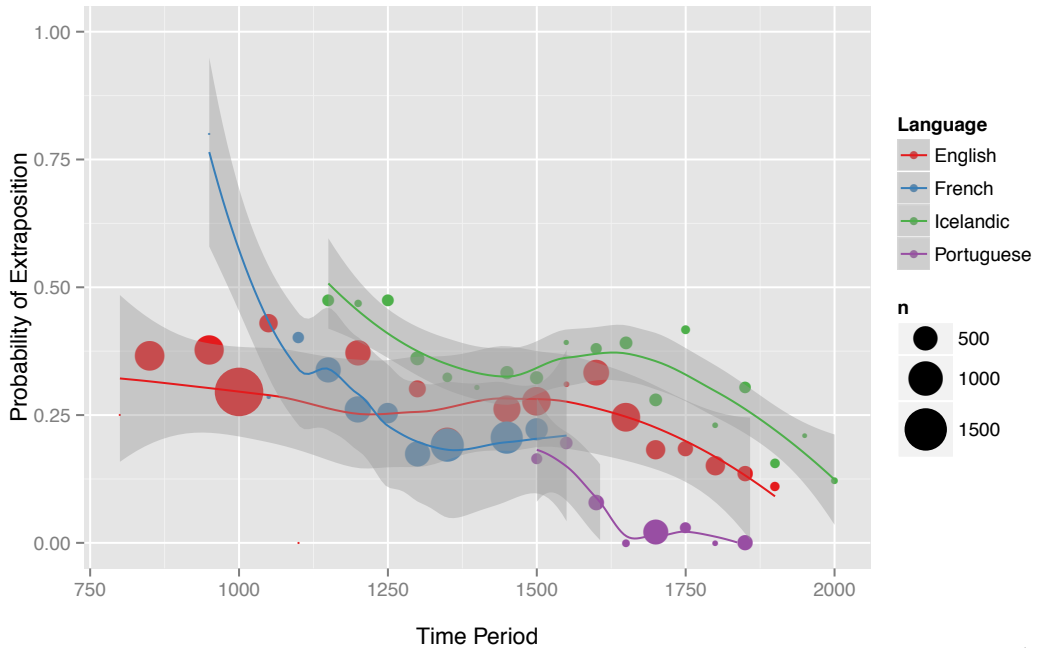
(1) womans mylke þat hase a knaue childe (cmthorn,9.166)

Liber de diversis medicinis, Thornton ms., c.1440

- The string is still grammatical today.
- Today, the meaning is unusual (milk that has a child).
- The Middle English interpretation (milk of a woman that has a knave child) is unattested since the early 19th century.
- Moral of the tale: certain syntactic properties cannot be reduced to properties of strings.

Speed

Four Languages (Subj Ex), over time



(Wallenberg 2015)

- Properties that are apparently stable over centuries may be part of a broader change.
- The reality of the broader change requires a thorough quantitative investigation.

Section 2

Parsed historical corpora

Layers of annotation

- A corpus is just a collection of texts.
- A good corpus will have chosen those texts in some principled way.
- An ideal corpus would:
  - Be enormous.
  - Be error-free.
  - Contain explicit annotation of:
    - Morphology (lemmatization);
    - Part of speech (tagging);
    - Syntactic structure (parsing);
    - Semantic relations (semantic role labelling).
- The ideal doesn't exist.
- Trade-off between corpus size and quality of annotation.
- All the above processes can be automated.
- The automated syntactic/semantic annotation processes are very imperfect.
- Hand-parsing is time-consuming (but see below).

A corpus with almost everything

- IcePaHC: the Icelandic Parsed Historical Corpus (Wallenberg et al. 2011)
- Step 1: find text.

Hans faðir Jón var Magnússon

A corpus with almost everything

- Step 2: lemmatize.

Hans-hann faðir-faðir Jón-jón var-vera Magnússon-magnússon

A corpus with almost everything

- Step 3: add POS info, including case, tense, etc.

(PRO-G Hans-hann)
(N-N faðir-faðir)
(NPR-N Jón-jón)
(BEDI var-vera)
(NPR-N MagnÃžsson-magnÃžsson)

A corpus with almost everything

- Step 4: add phrase-level bracketing.

```
( (IP-MAT (NP-SBJ (NP-POS (PRO-G Hans-hann))
                  (N-N faðir-faðir)
                  (NP-PRN (NPR-N Jón-jón)))
          (BEDI var-vera)
          (NP-PRD (NPR-N Magnússon-magnússon))
          (. ;-;)) (ID 1675.MAGNUS.BIO-OTH,.3))
```

- Not shown: annotating nonlocal syntactic dependencies, annotating semantic information (e.g. referential dependencies).
- This format typically includes nonlocal dependencies, but no semantic information.

What's already available?

Penn format

- The above format is the Penn format.
- Major historical corpora:
  - English (c.6m words, –1914, PPCHE, YCOE, Kroch & Taylor 2000, Taylor et al. 2003, Kroch et al. 2004, 2010).
  - English correspondence (c.2.2m words, c.1450–1700, PCEEC, Taylor et al. 2006).
  - French (c.1m words, –18th century, MCVF, Martineau et al. 2010).
  - Portuguese (c.900k words, c.1500–1900, Tycho Brahe, Galves & Faria 2010).
  - Icelandic (c.1m words, –20th century, IcePaHC, Wallenberg et al. 2011).
  - German (c.100k words, Luther, *Septembertestament*, Light 2011).
  - English poetry (c.100k words, 1100–1500, PCMEP, Zimmermann 2015).

What's already available?

Dependency treebanks

- The other major format annotates dependencies rather than constituency.
- Major example: PROIEL / TOROT / ISWOC.
  - PROIEL: Ancient Greek, Latin, Gothic, Classical Armenian, Old Church Slavonic (Haug & Jøhndal 2008).
  - TOROT: Old Church Slavonic, Old Russian, Middle Russian (Eckhoff & Berdicevskis 2015).
  - ISWOC: Old English, Old French, Portuguese, Spanish (Bech & Eide 2014).
- Total c.900k sentences.
- No principled divide between the two formats.
- Neither format is better.

So where are we?

- 20 years on, we have almost 20m words, from a range of IE languages.
- There's a clear acceleration in production of parsed historical corpora, with many more in the works (Low German, Early Middle English, more French, dialectal English, parallel Bible translations incl. non-IE European languages).
- That's simultaneously great, and depressing.
- **Size:** very few languages have sufficient parsed resources. All are IE.
- **Speed:** very few languages have parsed resources across sufficient time depth to track very slow change. All are IE.
- **Abstractness:** trade-off with coverage.

Mother's milk that has a knave child

```
(NP (NP-POS (N$ womans)
            (CP-REL *ICH*-1))
    (N mylke)
    (CP-REL-1 (WNP-2 0)
              (C +tat)
              (IP-SUB (NP-SBJ *T*-2)
                      (HVP hase)
                      (NP-ACC (D a) (N knaue) (N childe))))))
```

Section 3

Getting more out of corpora

Making the most of bad data

- Despite all the frustrations, diachronic syntax has never had it so good.
- Digital resources make it possible to answer research questions, accurately, with a fraction of the resources (and expertise) that would previously have been required.
- Corpora can't do everything, especially not given current limitations. But you can get a lot out of them.
- You can get more out of them through creative use of available resources:
  1. Work with unparsed corpora when possible.
  2. Look for hidden distributional evidence.
  3. Make your own parsed resources.

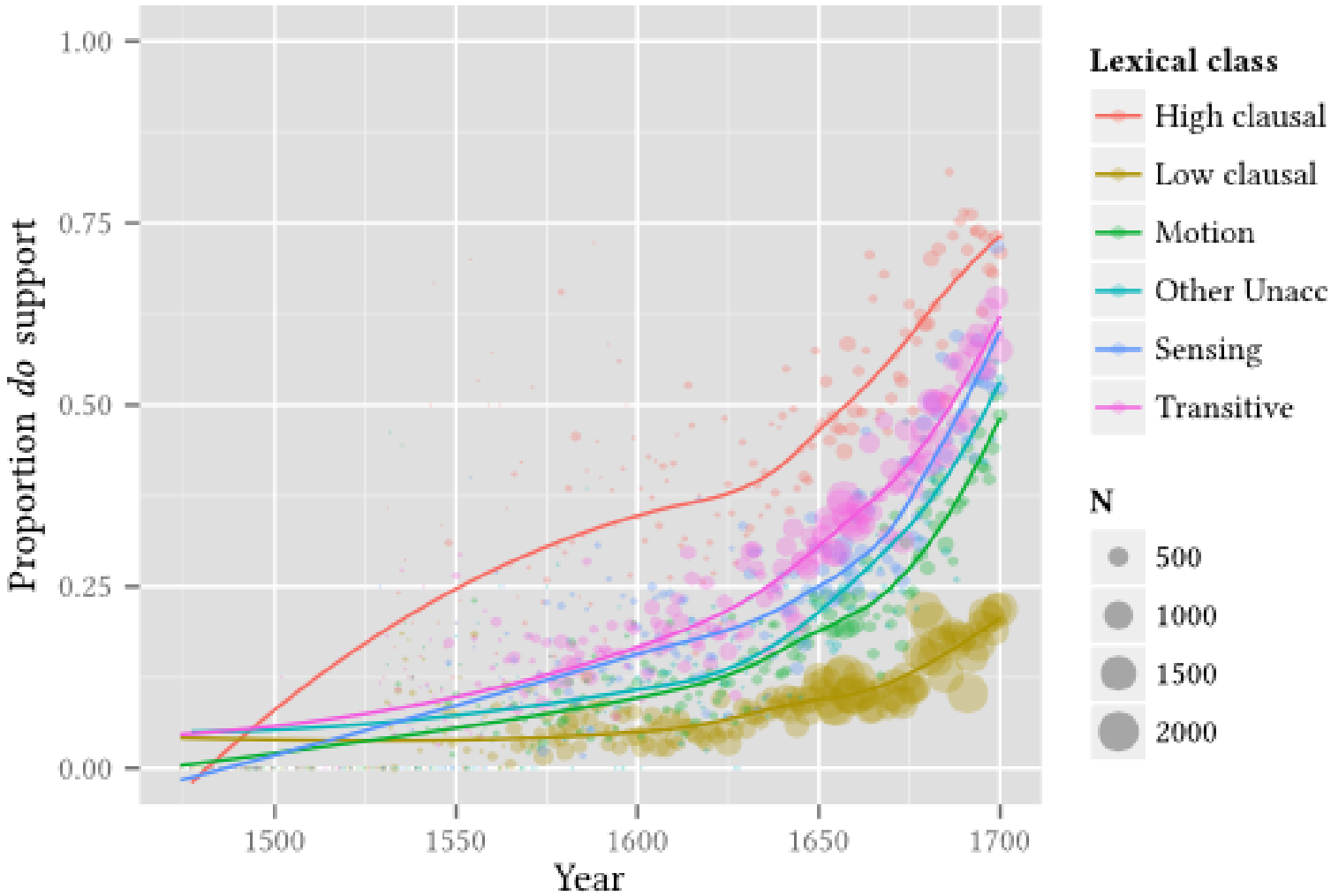
Subsection 1

Work with unparsed corpora

Why and how

- Unparsed corpora (still typically lemmatized and POS-tagged) are an order of magnitude bigger than parsed corpora.
  - PYCCLE (Ecay 2015b): c.100m words (not all public).
  - COHA (Davies 2012): c.400m words.
  - Google Books corpus: c.100bn words?
- Nuances in the diachrony, even for English, can be obscured by lack of data.
- Many syntactic questions can be operationalized as questions about word order.

Do-support: lexical class effects



(Ecay 2015a)

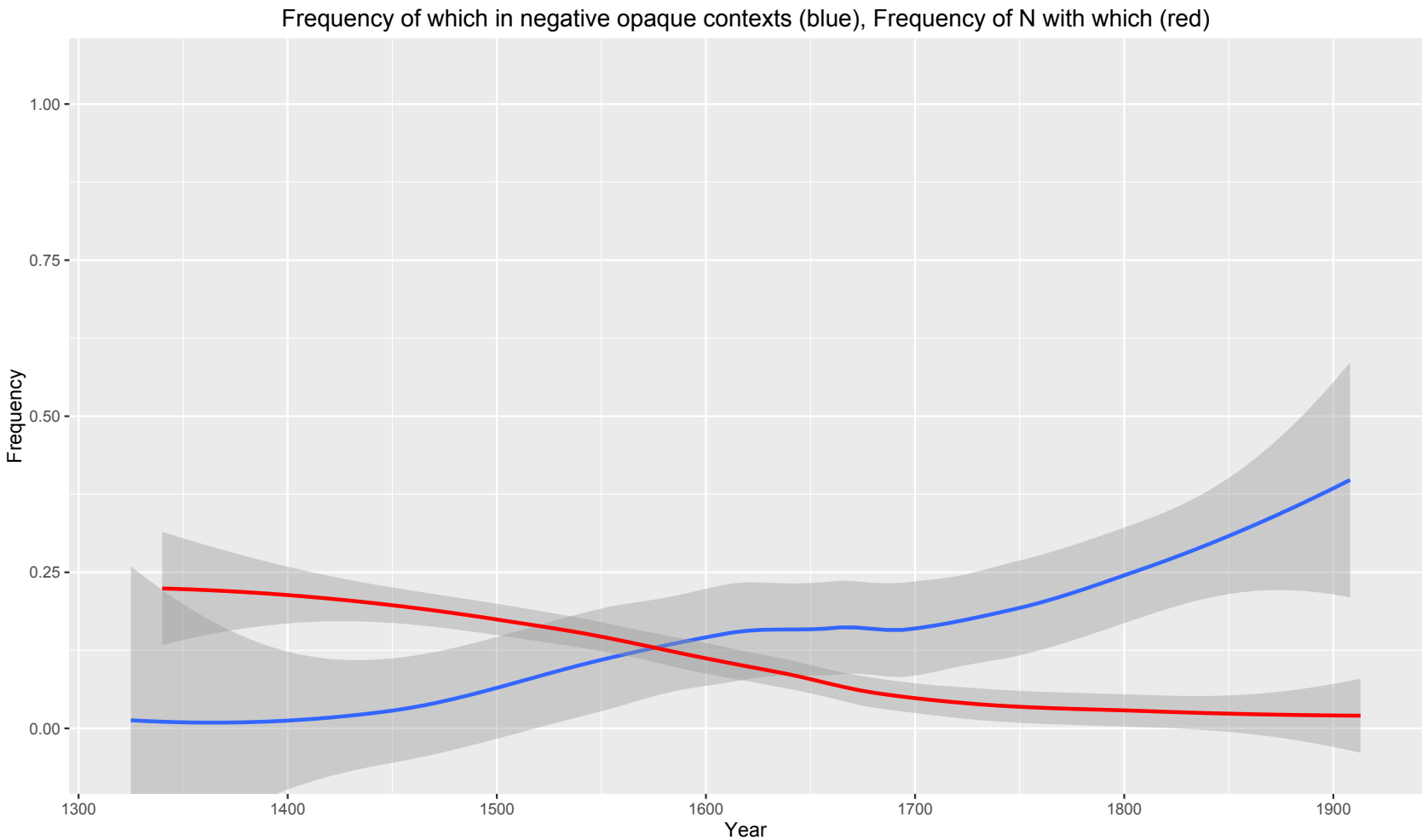
Subsection 2

Look for hidden distributional evidence

Restrictiveness

- A question addressed in unpublished work with Nik Gisborne: were the earliest English *wh*-relatives restrictive or nonrestrictive?
- It can feel that way, but direct evidence for restrictiveness is limited.
- Simple distributional test: nonrestrictive relatives cannot modify *no/few*, etc. Generates insufficient data for real confidence.
- More complex distributional test: it turns out that relatives of the form *which N* are always nonrestrictives. So we can track restrictiveness indirectly by tracking *which N*.

Correlation of simple and obscure evidence



Subsection 3

Make your own parsed resources

The process

- Parsing text is time-consuming but quite possible, and every little helps.
- Small-scale resources can be constructed as part of doctoral research or similar.
- Steps:
  1. Automated tokenization, POS-tagging, (lemmatization).
  2. Automated provisional parsing.
  3. Manual correction of provisional parse.
- A smart way to proceed is to look for high-quality resources where Step 1 has already been completed.
- PLAEME (work in progress with Rhona Alcorn, Jim Donaldson, Joel Wallenberg): building on high-quality annotated texts in LAEME (Laing 2013).
  - Bespoke rule-driven parsing process, building on syntactic information implicit in LAEME tags.
  - Hand-correction at c.4–500 words/hour.
  - Aiming for a 200k-word corpus on a £10k British Academy small research grant.

Summary

- Diachronic syntactic research is particularly challenging because syntactic objects are very big, very abstract, and often change slowly.
- Parsed corpora are slowly revolutionizing diachronic syntactic research by providing sufficient searchable data, in an appropriate format, to overcome those challenges.
- But high-quality parsed corpora are time-consuming and labour-intensive to produce.
- So corpus-based diachronic syntax can usefully:
  - operationalize research questions in terms of strings, not structures;
  - look for links between ideal data and available data;
  - make more resources!

References

Bech, K. & Eide, K. (2014). The ISWOC corpus. Department of Literature, Area Studies, and European Languages, Oslo. http://github.com.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora*, *7*, 121–157.

Ecay, A. (2015a). *A Multi-step Analysis of the Evolution of English Do-support*. PhD thesis, University of Pennsylvania.

Ecay, A. (2015b). The Penn–York Computer-annotated Corpus of a large amount of English based on the TCP (PYCCLE-TCP). https://github.com/uoy-linguistics/pyccle.

Eckhoff, H. M. & Berdicevskis, A. (2015). Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta*, *14-15*, 9–25.

Galves, C. & Faria, P. (2010). Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

Haug, D. & Jøhndal, M. (2008). Creating a parallel treebank of the Old Indo-European Bible translations. In Sporleder, C. & Ribarov, K. (Eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, (pp. 27–34).

Kroch, A., Santorini, B., & Delfs, L. (2004). Penn-Helsinki parsed corpus of Early Modern English.

Kroch, A., Santorini, B., & Diertani, A. (2010). The Penn–Helsinki Parsed Corpus of Modern British English (PPCMBE). Department of Linguistics, University of Pennsylvania.

Kroch, A. & Taylor, A. (2000). Penn-Helsinki parsed corpus of Middle English (2nd edition).

Laing, M. (2013). A Linguistic Atlas of Early Middle English, 1150–1325. Version 3.2, http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html.

Light, C. (2011). Parsed Corpus of Early New High German. https://enhgcorpus.wikispaces.com/home.

Martineau, F., Hirschbühler, P., Kroch, A., & Morin, Y. C. (2010). Corpus MCVF annoté syntaxiquement. Université d'Ottawa.

Taylor, A., Nurmi, A., Warner, A., Pintzuk, S., & Nevalainen, T. (2006). Parsed Corpus of Early English Correspondence. University of York and University of Helsinki. Distributed through the Oxford Text Archive.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). The York–Toronto–Helsinki Parsed Corpus of Old English prose (YCOE). Department of Language and Linguistic Science, University of York.

Wallenberg, J. (2015). Science of the experimentally possible: Very slow change and language technology. Paper presented at Research Links workshop, Campinas.

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus. http://www.linguist.is/icelandic_treebank.

Zimmermann, R. (2015). The Parsed Corpus of Middle English Poetry. University of Geneva.