Constructing a syntactically annotated corpus for grammatical research

Beatrice Santorini
University of Pennsylvania

NINJAL Colloquium - December 8, 2017

Roadmap

- Why parsed corpora?
- From POS-tagged to parsed corpus
- Issues in syntactic annotation
- Trade-offs in using POS-tagged vs. parsed corpora

What is a corpus?

- A corpus is simply a collection of texts.
- The texts may be exhaustive samples (texts in their entirety) or partial samples.
- The selection of the samples raises important questions to which there are no pat answers.
- Rather, sample selection is closely tied to the aims of a particular project.

How big should a corpus be?

- All other things equal, as big as possible!
- However, all other things are never equal.
- More on this later.
- No matter how big your corpus, you should know the size. For most purposes, the best measure is number of words (or characters for non-alphabetic writing systems).

Why construct corpora?

- There are many reasons.
  - Research in linguistics
  - Research and developing applications in computational linguistics
  - Lexicography
- So there are synchronic corpora, corpora of written language, speech corpora, corpora of particular genres, corpora of signed languages, historical corpora, diachronic corpora, and so on.

Why construct corpora?

- Constructing corpora is time-consuming, but once the work is done, the result can be used for many different projects.
- Corpora can be searched quickly and reliably.
- Hypotheses are easily put to the test and refined.
- Results are replicable.

Why construct corpora?

- Corpora can yield unexpected discoveries.
- Over time, they can be corrected, revised, augmented, and otherwise improved.
- With increasing corpus size, new kinds of scientific results become possible.

Our reasons for constructing parsed corpora

- Our particular interest is in diachronic syntax.
- The data are historical, so we need corpus data, since native speaker judgments are unavailable.

Our reasons for constructing parsed corpora

- Note: Diachronic ≠ historical.

- We are interested not just in comparing two invariant stages of a language (old vs. new).

- Rather, we often wish to propose mathematical models for the time course of linguistic changes.

Our reasons for constructing parsed corpora

- During a change, we observe synchronic variation.

- Even if native speaker intuitions concerning variable usage were available, they would not be detailed enough for our purposes.

Digitization

- Nowadays, when we speak of corpora, we mean ones that are digitized (rather than printed or composed of page images).
- Digitized corpora make it possible to:
  - perform searches and statistical analysis
  - make corrections and revisions on a large-scale and consistent basis
  - add annotation (= further linguistic information)

Digitization

- Digitized corpora go back to the 1960s.
  - Brown Corpus (Francis and Kučera 1967, 1M words)
- The parsed corpora of historical English that we and others have built at Penn and York are based on the diachronic portion of an early digitized corpus.
  - Helsinki Corpus of English Texts (Matti Rissanen et al., 1984-1991, over 1.5M words)
- Note: The most recent versions of the historical English parsed corpora now also include other texts.

Raw text

- In principle, corpora can consist of raw text.

Here is an example of raw text .

Normalization

- What we call raw text is not completely raw.
- Punctuation is split off.
- Contractions, portmanteau words, and the like are usefully separated into separate orthographic words.
  - English *can't* > *ca@ @n't*
  - French *du* 'of the' > *d@ @u*
- Words can also be joined.
  - *him_self*
  - *two_thousand*

Uses for raw text corpora

- Raw text corpora have been used in:
  - tracking word frequencies (Zipf's law)
  - lexicography (American Heritage Dictionary 1969 and many others since then)

Limitations of raw text

- One might think that (normalized) raw text is sufficient to study phenomena concerning individual lexical items.

- For instance, in the history of the English pronominal system, *you, ye* replaced *thou, thee.*

Limitations of raw text

- But orthographic words are often ambiguous, especially in older stages of a language.
- The pronoun *thee* can be spelled *the.*
- The article *the* can be spelled *ye.*

Part-of-speech (POS) tagging

- Words can be tagged (= annotated) for morphosyntactic information, including
  - part of speech (= basic category), such as noun, verb, adjective, etc.
  - inflectional features (number, tense, aspect, grammatical case, etc.)
  - other features (use vs. mention, native vs. foreign, disfluency, etc.)

POS-tagged text

Here/ADV is/BEP an/D example/N of/P

tagged/VAN text/N ./.

POS tagsets

- The Brown Corpus was already POS-tagged using a combination of automatic tagging and human correction.

- Tags for English tend to be derived from the Brown Corpus tagset.

- Tags for morphologically richer languages typically include more morphological information.

What questions might we address with tagged corpora?

- POS tagging distinguishes:

<u>Canonical spelling</u>	<u>Variant spelling</u>
the/D	ye/D
thee/PRO	the/PRO
ye/PRO	

Some surface-y morphosyntactic issues

- Clitic placement in Romance
- Adverb placement in Germanic as a diagnostic of verbal syntax (V-to-T raising)
- The rise of *do* support in English (Zimmermann 2017) - more on this later

Limitations of POS tagging

- Surface-y morphosyntactic phenomena are ones that can be searched for in connection with overt lexical material or POS tags.

- But in general, syntax concerns constituent structure, which is independent of individual lexical items.

Irreducibly syntactic phenomena

- Phrasal headedness (notably, OV vs. VO)

  John pizza eat will. —> John will eat pizza.

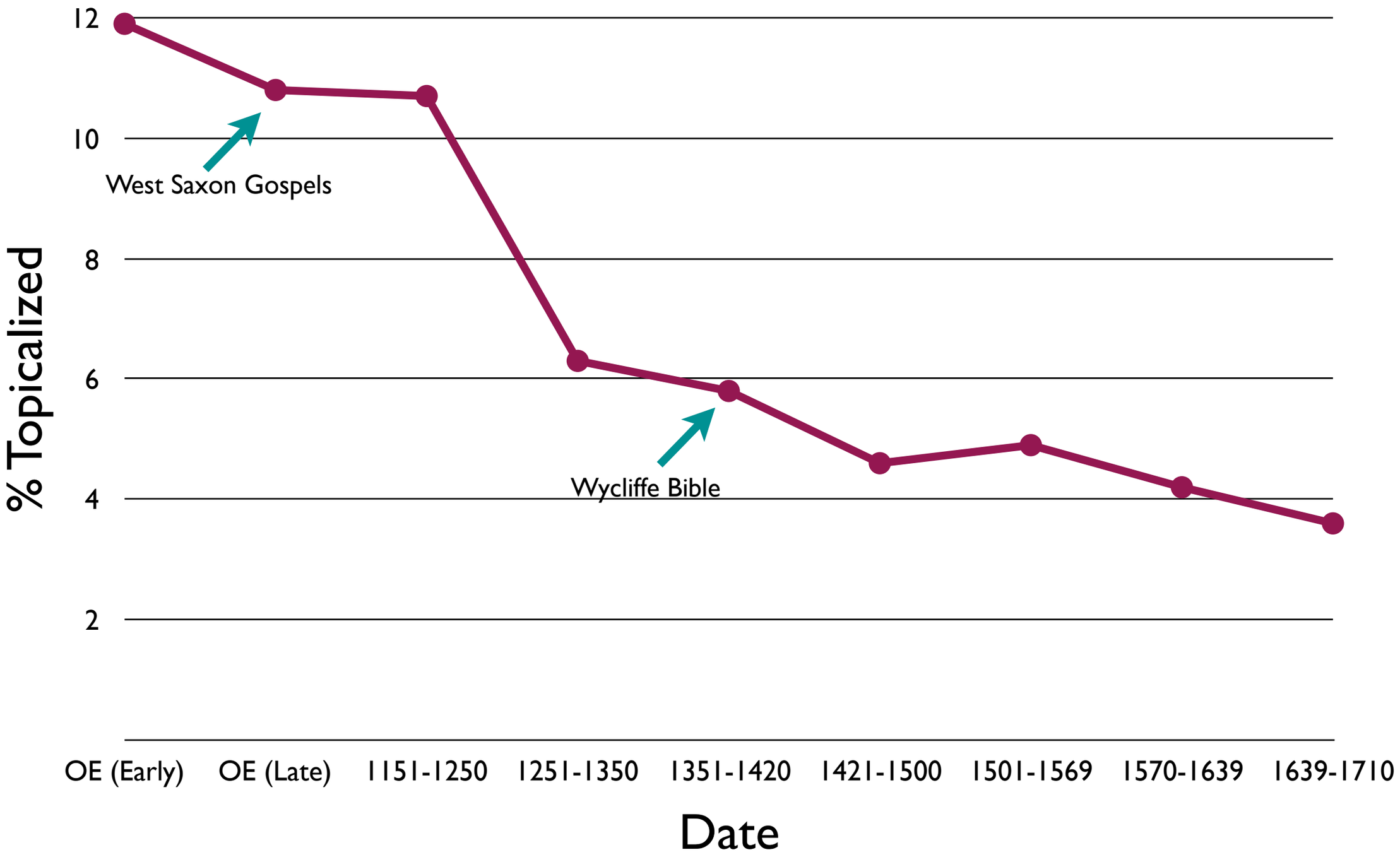
- Topicalization

  John will eat pizza. —> Pizza, John will eat.

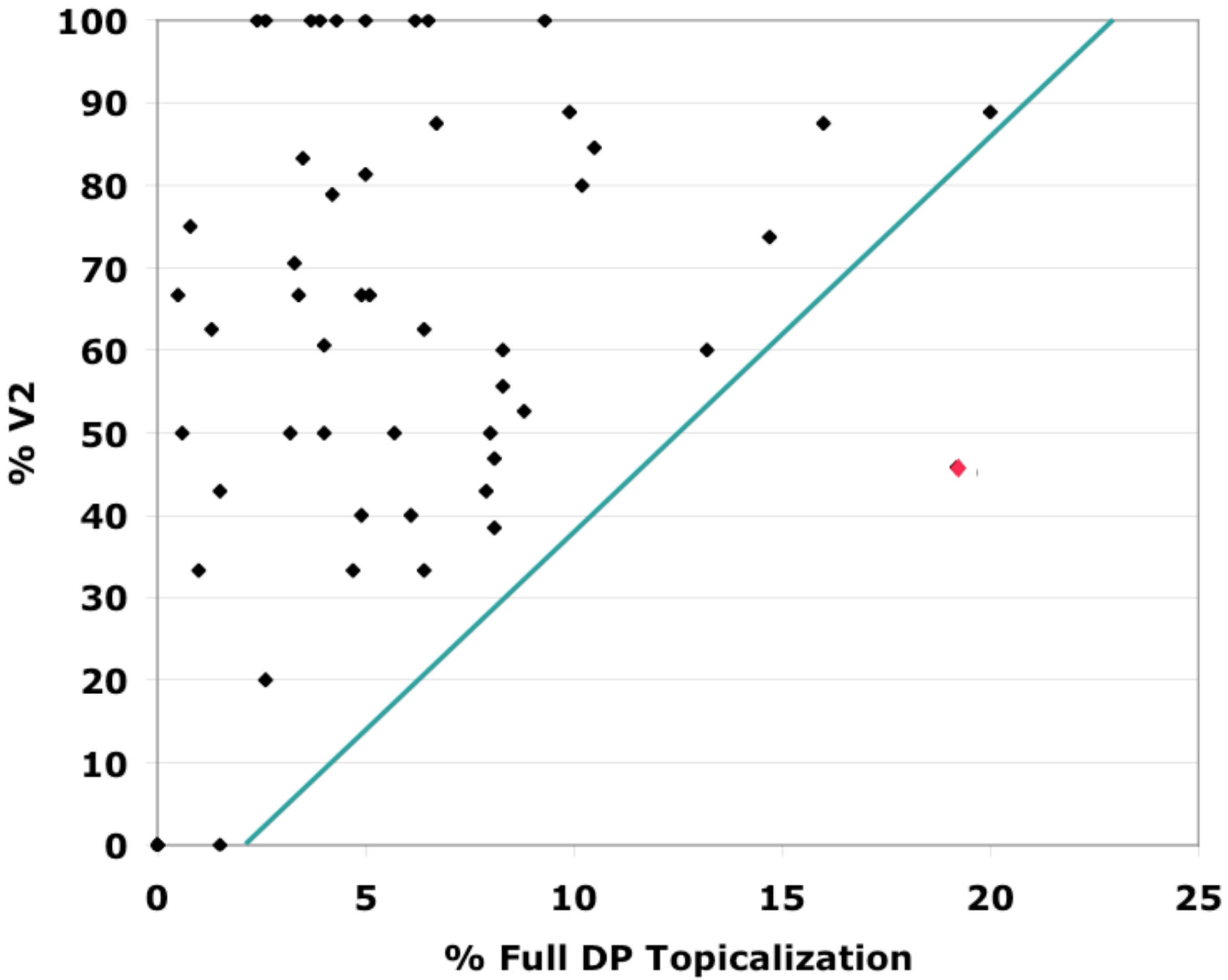
- Verb-second (V2)

  John will eat pizza. —> Pizza will John eat.

Decline of direct object topicalization in English



Correlation between frequencies of object topicalization and of V2 in Middle English texts (Wallenberg 2007)



More examples

Some handouts featuring research and results requiring parsed corpora can be found at:

http://www.ling.upenn.edu/~kroch/handouts/index.html

Parsed text

- Just as we added POS annotation to raw text, we can add further syntactic information to POS-tagged text.
  - Phrasal category (NP, VP, AdjP, etc.)
  - Phrasal subcategory (finite clause vs. infinitival clause vs. gerund / participial clause, etc.)
  - Grammatical function (subject, direct object, indirect object, etc.)
  - Semantic or other grammatical function (locative, temporal, measure phrase, secondary predicate, etc.)

Silent categories

- In addition to annotating the original text, we normalize the structures by adding silent categories, such as:
  - Silent subjects
  - Silent complementizers
  - Traces of movement (at least for some types of movement)

Parsed text

((IP-MAT (ADVP-LOC (ADV Here))

(BEP is)

(NP-SBJ (D an) (N example)

(PP (P of)

(NP (VAN parsed) (N text))))

(. .)))

Penn Treebank format

- The example on the previous slide is in so-called Penn Treebank format - the standard format for parsed corpora based on phrase structure grammars.

- The Penn Treebank is a parsed version of the Brown Corpus (Marcus, Santorini, and Marcinkiewicz 1993).

- The syntactic analysis is expressed in terms of labelled brackets, which are mathematically equivalent to syntactic trees.

- Individual sentences are enclosed in unlabeled wrapper parens (highlighted on the previous slide in bold red).

A note on file format

- At all stages of corpus creation, our files are plain text files (that is, they do not contain special formatting characters).

- Natural language processing algorithms (taggers, parsers, search programs) expect files in plain text format.

- Plain text files also give us a degree of independence from proprietary software.

What do we mean by “text”?

- Plain text refers to the format of a file for computational purposes (cf. “save as” options)

- Raw text refers to the information content of a file (no POS tags or syntactic structure).

- So there is no contradiction when we say that a POS-tagged or parsed file is a text file. What we mean is that the raw text and the annotation are encoded in the same way — as plain text.

A note on XML markup

- Corpora can be marked up (= annotated) in XML, a system similar to HTML (used to display content on the web).

- The markup generally concerns features of text organization (section headings, subheadings, etc.), typography (italic, bold, etc.).

A note on XML markup

- XML can be used to annotate corpora for linguistic features.
- As long as the information expressed in the markup is the same as in our notation, the choice of notation doesn't matter, because one notation can be automatically converted to the other.
- We don't use XML because:
  - The markup adds lots of material without lots of information.
  - The tools for XML corpora are too primitive for our purposes.
  - If it became necessary, the conversion from Penn Treebank format to an XML-compliant annotation would be trivial.

A note on stand-off annotation

- In stand-off annotation, the text and the annotation are stored in different files.

- The items in the various files obviously need to be linked, introducing computational complexity.

- So we stay away from stand-off annotation.

From tagged to parsed text with an automatic parser

- Get a parsed training corpus
- Train an automatic parser (the parser learns the rule-based or statistical patterns in the training corpus)
- Run the trained parser on your tagged corpus
- Correct the output

What if there is no training corpus?

- Training corpora are available for many languages.
- But not for all historical stages of a language or for recherché languages (like Sumerian or Old Florentine).

Revision queries to the rescue

- We can use a feature of the query language CorpusSearch called revision queries.

- http://corpussearch.sourceforge.net

- Revision queries allow us to specify structures and modify them in well-defined ways.

POS-tagged text, revisited

Here/ADV

is/BEP

an/D

example/N

of/P

tagged/VAN

text/N

./.

POS-tagged text, reformatted to conform to Penn Treebank format

```
( (IP-MAT (ADV Here)
          (BEP is)
          (D an)
          (N example)
          (P of)
          (VAN tagged)
          (N text)
          (. .)))
```

Corpus revision query 1.0

query: ({1}D hasSister {2}N)

AND (D iPrecedes N)

add_internal_node{1,2}: NP

Output of revision query 1.0

((IP-MAT (ADV Here)

(BEP is)

(NP (D an) (N example))

(P of)

(VAN tagged)

(N text)

(. .)))

Corpus revision query 1.1

query: ({1} D | ADJ | ADJR | ADJS | Q | QR | QS | VAN hasSister {2}N | NS)

AND (D | ADJ | ADJR | ADJS | Q | QR | QS | VAN iPrecedes N | NS)

add_internal_node{1,2}: NP

Sample definitions file

noun: N | NS

adjective: ADJ | ADJR | ADJS | VAN

quantifier: Q | Q[RS]

pre_noun: D | $adjective | $quantifier

Corpus revision query 1.2

query: ({1}pre_noun hasSister {2}noun)

AND (pre_noun iPrecedes noun)

add_internal_node{1,2}: NP

Output of revision query 1.2

((IP-MAT (ADV Here)

(BEP is)

(NP (D an) (N example))

(P of)

(NP (VAN tagged) (N text))

(. .)))

Corpus revision query 2

query: ({1}P hasSister {2}NP)

AND (P iPrecedes NP)

add_internal_node{1,2}: PP

Output of revision query 2

((IP-MAT (ADV Here)

(BEP is)

(NP (D an) (N example))

(PP (P of)

(NP (VAN tagged) (N text)))

(. .)))

Chunked text

- Delimiting non-recursive NPs and PPs is relatively easy (whatever parsing method is used).

- Recursive structure is the tough nut.

- Some cases are easy, though (like PPs headed by *of*).

Corpus revision query 3

query: ({1}NP hasSister {2}PP)

AND (NP iPrecedes PP)

AND (PP iDoms P)

AND (P iDoms [oO]f)

move_to{2,1}:

Output of revision query 3

((IP-MAT (ADV Here)

(BEP is)

(NP (D an) (N example)

(PP (P of)

(NP (VAN tagged) (N text)))))

(. .)))

Remaining corrections

((IP-MAT (ADVP-LOC (ADV Here))

(BEP is)

(NP-SBJ (D an) (N example)

(PP (P of)

(NP (VAN tagged) (N text))))

(. .)))

Annotald - a tool for human correction of parsed corpora

- http://annotald.github.io

- Developed in connection with IcePaHC, a 1M-word corpus of historical Icelandic

  http://www.linguist.is/icelandic_treebank/Main_Page

- Browser-based (Google Chrome)

- Allows correction of POS and syntactic annotation

Annotation = God's truth, not !

- Annotation is intended to facilitate retrieval of examples.
- It does not necessarily provide the correct analysis of the examples.

Annotation guidelines for the Penn Parsed Corpora of Historical English

- http://www.ling.upenn.edu/~beatrice/annotation/
- This annotation system, first developed for Middle English by Tony Kroch and Ann Taylor, is also used for later stages (Early Modern English, Modern British English).
- With relatively minor modifications, it has been used for Old English, Icelandic, German, historical Romance, Ancient Greek, …

Annotation guidelines for the Penn Parsed Corpora of Historical English

- With auxiliary guidelines for annotating disfluencies, the guidelines are suitable for annotating corpora of speech transcripts (Appalachian, New York City, African-American Vernacular English, Russian child and child-directed speech, …)

Why not give the correct structure?

- The correct structure can
  - be unknowable in principle (OV vs. VO during phrase structure change in progress)
  - be unknowable in a particular case (synchronic ambiguity)
  - be knowable in principle, but not to us now
  - be too difficult or time-consuming to determine
  - involve details that are irrelevant for retrieving examples of interest (Kaynean representation of head-final structures)

A superficially simple example

((IP-MAT (NP-SBJ (PRO He))

(MD will)

(VB tell)

(NP-OB1 (D the) (N story))

(. .)))

VO — or OV with extraposition?

((IP-MAT (NP-SBJ (PRO He))

(MD will)

(NP-OB1 *T*-i)

(VB tell)

(NP-i (D the) (N story))

(. .)))

Synchronic ambiguity

- Even in a given context, it can be impossible to decide whether
  - a PP should attach high or low
  - a participial form is verbal or adjectival

    The duchess was <u>entertaining</u> last night.
  - Romance *que* is a wh- pronoun or a complementizer

Sample default rules

- PPs attach high.
- Participial forms are verbal.
- *Que* is a complementizer in comparative constructions and a wh- pronoun elsewhere.

Where does the trace go?

((CP-QUE-MAT (WNP-i (WD Which) (N story))

(IP-SUB (MD will)

(NP-SBJ (PRO he))

(NP-OB1 *T*-i) ← here? (OV)

(VB tell)

(NP-OB1 *T*-i)) ← or here? (VO)

(. ?)))

A crazy but useful default rule

((CP-QUE-MAT (WNP-i (WD Which) (N story))

(IP-SUB (NP-OB1 *T*-i)

(MD will)

(NP-SBJ (PRO he))

(VB tell))

(. ?)))

Another position that isn't true

((CP-QUE-MAT (WNP-i (WD Which) (N story))

(IP-SUB (NP-OB1 *T*-i)

(MD will)

(NP-SBJ (PRO he))

(VB tell))

(. ?)))

Further considerations

- Consistency is more important than absolute correctness.
  - Correct structures that are implemented inconsistently are not as useful as consistently implemented approximations.
- Use simple rules.
  - For instance, treat a word a foreign or not depending on whether it’s in the OED.

Notational variants

- It is very important to focus on the information content of annotation alternatives rather than on the superficial form.

- Alternatives are notational variants if one alternative can be automatically be converted to the other, and vice versa.

Notational variants

- One variant might still be preferable to the other for some purposes.

- For instance, the published corpus features the grammatical function tags OB1 and OB2.

- But when building the corpus, we use the notational variants ACC and DTV (even for Modern English) to minimize typos.

Notational variants

- Revision queries allow us to convert one notational variant to another.

- Users can adapt parsed corpora to suit their own preferences.

Sentences with modals

((IP-MAT (NP-SBJ (PRO They))

(MD will)

(VB come)

(ADVP-TMP (ADVR later))))

((IP-MAT (NP-SBJ (PRO They))

(MD will)

(IP-INF (VB come)

(ADVP-TMP (ADVR later)))))

Perception verb complements

((IP-MAT (NP-SBJ (PRO They))

(VBD saw)

(IP-INF (NP-SBJ (PRO him))

(VB arrive))))

((IP-MAT (NP-SBJ (PRO They))

(VBD saw)

(NP-OB1 (PRO him))

(IP-INF (VB arrive))))

Simplicity

- In the past, the tools for constructing parsed corpora were more limited than we have now.

- So simple representations (essentially, ones with fewer nodes) were attractive because they involved less effort for the annotator (easier to read, faster to correct).

- With the advent of more powerful tools, what is becoming increasingly important are simple annotation principles.

Branching conjuncts

(NP (NP (ADJ delicious) (N food))

(CONJP (CONJ and)

(NP (ADJ excellent) (NS drinks)))

Non-branching conjuncts: Simple representation

(NP (N food) (CONJ and) (NS drinks))

- Complicates annotation guidelines by adding special cases
- Complicates search queries
- Drives computational linguists nuts

Non-branching conjuncts: Simple principle

(NP (NP (N food))

(CONJP (CONJ and)

(NP (NS drinks)))

- No additional special guidelines
- Simpler search queries are less error-prone
- No more need for computational linguists to normalize

Future releases of the PPCHE

- In future releases of the PPCHE, we intend to streamline the annotation guidelines and the annotation itself along the lines just noted.

- In later corpora (historical French, Old Italian), we have already begun this process (notably in connection with the annotation of degree and comparative constructions).

Trade-offs and tensions in automatic annotation

- Now that you have a sense both of the usefulness of parsed corpora and of the effort involved in constructing them, we would like to make note of some trade-offs and tensions that are characteristic of the state of the art in natural language processing.

The importance of corpus size

- As we mentioned earlier, the bigger a corpus the better.

- Why?

- Many phenomena of interest are rare and might occur only a handful of times in even a 1M-word corpus.

The importance of corpus size

- Corpora are samples of language use. They yield estimates concerning the parameters of the underlying object of study. The bigger the sample, the more reliable the estimates.

- So we would like to base our research on very large corpora (hundreds of millions of words or even bigger).

Very large corpora — tagged

- Given current technology, very large corpora are increasingly available (at least for some languages).
- Automatic POS tagging is relatively accurate, so very large tagged corpora are increasingly available.

Very large corpora — parsed

- Automatic parsing is still not accurate enough to be useful for linguistic research.
- So we need humans to correct automatically parsed corpora.

Very large corpora — parsed

- But human correction of parsed texts is a real bottleneck.
- It takes talent.
- Even with talent, it is time-consuming.
- Hence, it is expensive and essentially unavailable for very large corpora.

Partially automating human correction

- Accuracy and consistency are noticeably improved.
- Parsing correction speeds are roughly quadrupled.
- Still far from good enough for very large corpora, though.

Some ideas, 1

- Can we triage the output of automatic parsing?

- If automatic parsing is almost perfect for short sentences, can we use only those as a representative subset of all sentences?

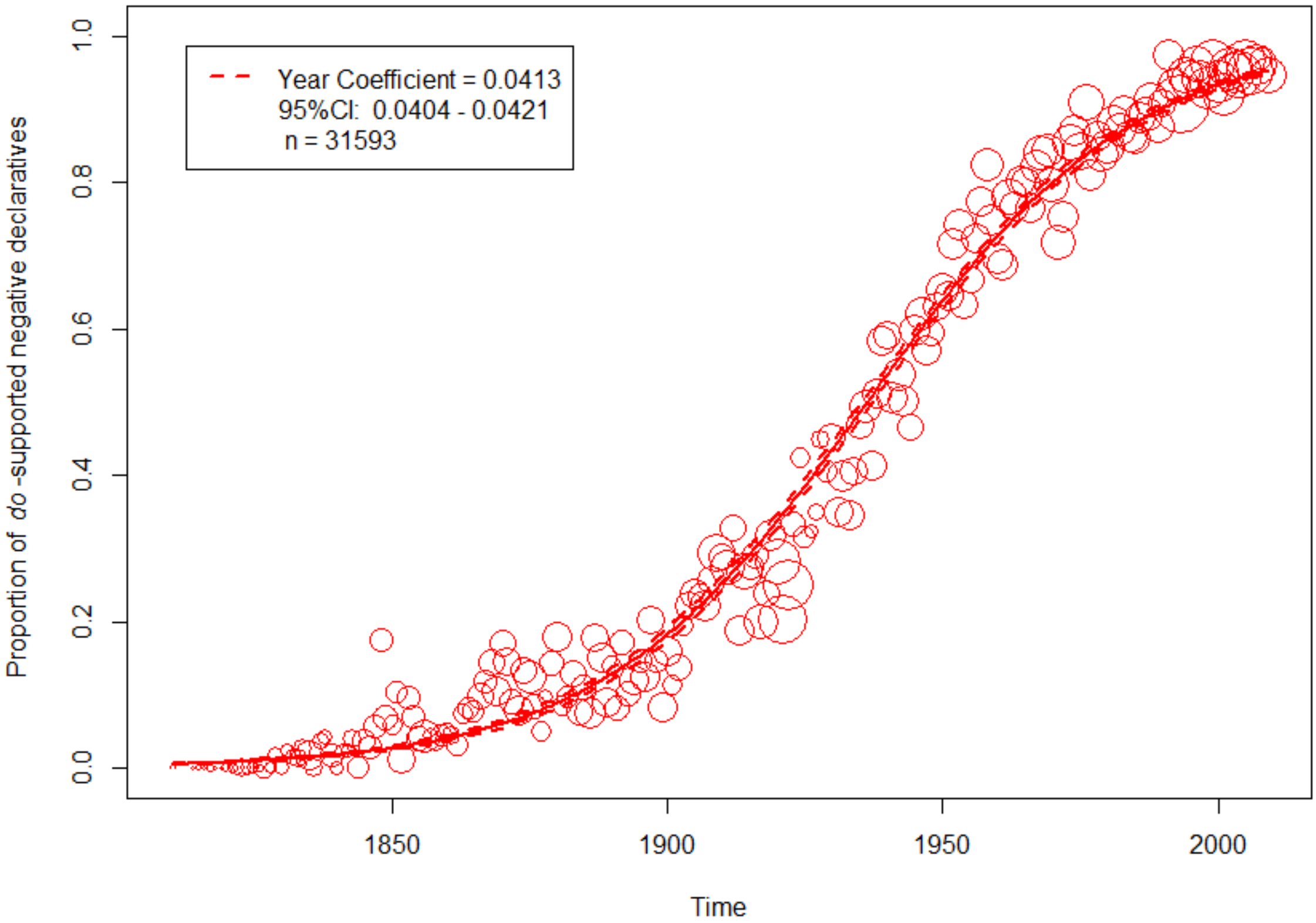
Some ideas, 2

- Can we live with parsing errors?
- Ubiquitous errors, as with PP attachment and conjunction structures, might be irrelevant for certain research questions.

Some ideas, 3

- Can we make a silk purse out of a sow's ear?
- Can we use POS-tagged corpora to extract representative subsets of data?

The development of *do*-support with possessive *have* in negative declaratives



Zimmermann 2017

- https://archive-ouverte.unige.ch/pages/about_thesis (search for “Zimmermann, Richard”)

- Results on previous slide are based on Corpus of Historical American English (COHA)

COHA

- Built by Mark Davies at Brigham Young University
- https://corpus.byu.edu/coha/
- Approx. 385M words, annotated by lemma and POS
  - Trees / tree / nn2
  - were / be / vbdr
- 115K texts from 1810-2009
- Each decade balanced by 4 genres (fiction, magazine, newspaper, non-fiction book)

Massaging COHA

- Zimmermann 2017 wrote a series of scripts to extract relevant tokens.

- The output contained various types of errors (see Chapter 2.3 for discussion).

- Correcting the output was “tedious and strenuous” (p. 74) and took about 2 years.

Irreducible syntax again

- The strategies that Zimmermann used for studying *do* support would <u>not</u> work for irreducibly syntactic phenomena like topicalization or V2.

- This is because they cannot be searched for with reference to individual lexical items or POS tags.

```
( (IP-MAT (NP-SBJ *pro*)
                (VBP thank)
                (NP-OB2 (PRO you))
          (PP (P for)
              (NP (PRO$ your)
                  (N attention)))
          (. .)))
```