Constructing a parsed corpus of Early Modern English

Beatrice Santorini
University of Pennsylvania
http://www.ling.upenn.edu/∼beatrice/corpus-ling/index.html

IV Encontro de Corpora - USP
24 August 2004

Overview of presentation

- Some useful URLs
- Motivation for constructing electronic parsed historical corpora
- Goals and principles of our annotation
- How we build a corpus - a flowchart
- CorpusSearch - a search engine for parsed corpora

Some useful URLs

- Annotation manual
  - For beginners
    http://www.ling.upenn.edu/~ataylor/ppcme-lite.htm
  - For advanced users
    http://www.ling.upenn.edu/~ataylor/ppcme2-man-toc.htm

- CorpusSearch manual
  http://www.ling.upenn.edu/mideng/csdocs/CSRefToc.htm

Why construct historical *corpora* ?

- Recourse to native speaker intuitions impossible
- Hence, we need representative historical corpora (= collections of texts)
- Corpora can be important even for synchronic studies
  - How do people *actually* speak/write (as opposed to how they *say* they do)?
  - Basis for statistical parsers

Why *parsed* historical corpora?

- The syntactic structure of sentences is not completely determined by the words and their linear order
- Sentences can be structurally ambiguous
- Sentences can be produced by distinct grammars
- Hence, we need corpora that are annotated with appropriate information

Synchronic structural ambiguity

- ```
  ( (IP-MAT (NP-SBJ (VAG Flying) (NS planes))
            (MD can)
            (BE be)
            (ADJP (ADJ dangerous))
            (. .)))
  ```

- ```
  ( (IP-MAT (IP-SBJ (NP-SBJ *arb*)
                    (VAG Flying)
                    (NP-OB1 (NS planes))
            (MD can)
            (BE be)
            (ADJP (ADJ dangerous))
            (. .)))
  ```

Variation between two grammars

Earlier forms of English showed variation between an old (OV) and a new (VO) grammar.

- ```
  ( (IP-MAT (NP-SBJ they)
            (MD will)
            (NP-OB1 the old house)   <--- old (OV)
            (VB buy)
            (. .)))
  ```

- ```
  ( (IP-MAT (NP-SBJ they)
            (MD will)
            (VB buy)                 <--- new (VO)
            (NP-OB1 the old house)
            (. .)))
  ```

Uncertainty between two grammars

- ```
  ( (CP-QUE (WNP-1 Which house)
            (IP (MD will)
                (NP-SBJ they)
                (NP-OB1 *T*-1)  <--- old (OV)?
                (VB buy)
            (. ?)))
  ```

- ```
  ( (CP-QUE (WNP-1 Which house)
            (IP (MD will)
                (NP-SBJ they)
                (VB buy)
                (NP-OB1 *T*-1)  <--- new (VO)?
            (. ?)))
  ```

Why *electronic* parsed historical corpora?

- To ensure representativity, we need large corpora
- Annotation by hand is slow, expensive, and error-prone
- The answer: automate annotation as much as possible
- Electronic corpora are (relatively) easy to correct and update
- Electronic corpora can be built in stages

Further advantages of electronic corpora

- Electronic corpora can be searched quickly and reliably
- Research hypotheses are more easily tested and refined
- Results become replicable across research groups
- Increased search speed makes possible different *kinds* of results

Goals and principles of our annotation

- Parsed corpus consists of straight-up ASCII
  - Structural information is represented as labeled bracketing
  - No hidden formatting codes
  - No dependence on obsolescent software
  - If necessary, we would use ISO-Latin-1, ISO-Latin-2, Unicode
- Annotated corpus = God's truth, not
  - The primary goal of our annotation is to facilitate searches for various constructions of interest.
  - The goal is not (!) to associate every sentence with a correct structural description.

Dealing with uncertainty and ambiguity

- As many syntactic categories as possible should have clear meanings so that the number of unclear cases is minimized.
- We try to avoid decisions that are controversial, very time-consuming, or otherwise difficult.
- To that end, we sometimes omit information.
  - Adjectival vs. verbal passive (*The door is shut*)
  - VP boundaries
- In other cases, we use default rules.
  - Location of wh- traces (= gaps)
  - PP attachment ("when in doubt, attach high")

OV, or VO + leftward pronoun movement?

- (PP (P until)
    (CP-ADV (C 0)
        (IP-SUB (NP-SBJ (N death))
            (DOP do)
            (VP (NP-OB1 (PRO us))
                (VB part)))))

- (PP (P until)
    (CP-ADV (C 0)
        (IP-SUB (NP-SBJ (N death))
            (DOP do)
            (NP-1 (PRO us))
            (VP (VB part)
                (NP-OB1 *T*-1)))))

Omitting undecidable information

Our solution: a 'flat' structure without a VP

```
(PP (P until)
    (CP-ADV (C 0)
            (IP-SUB (NP-SBJ (N death))
                    (DOP do)
                    (NP-OB1 (PRO us))
                    (VB part))))
```

Question movement revisited

- ```
  ( (CP-QUE (WNP-1 Which house)
            (IP (MD will)
                (NP-SBJ they)
                (NP-OB1 *T*-1)  <--- old (OV)?
                (VB buy)
            (. ?)))
  ```

- ```
  ( (CP-QUE (WNP-1 Which house)
            (IP (MD will)
                (NP-SBJ they)
                (VB buy)
                (NP-OB1 *T*-1)  <--- new (VO)?
            (. ?)))
  ```

An incorrect, yet useful, structure

Our solution: we consistently put the trace in a position that is linguistically unmotivated, but competely predictable and so exploitable for searches.

```
( (CP-QUE (WNP-1 Which house)
          (IP (NP-OB1 *T*-1)
              (MD did)
              (NP-SBJ they)
              (VB buy)
          (. ?)))
```

PP attachment - high or low?

- ```
((IP-MAT (NP-SBJ They)
          (VBD painted)
          (NP-OB1 (D the) (N man))
          (PP (P with)
              (NP (D a) (N brush)))
          (. .)))
```

- ```
( (IP-MAT (NP-SBJ They)
          (VBD painted)
          (NP-OB1 (D the) (N man)
                  (PP (P with)
                      (NP (D a) (N telescope))))
          (. .)))
```

Omitting undecidable information

A useful solution: undecidable or difficult cases are attached high by default.

```
( (IP-MAT (NP-SBJ They)
          (VBD saw)
          (NP-OB1 (D the) (N man))
          (PP (P with)
              (NP (D a) (N telescope)))
          (. .)))
```

Argument *se*

In (European) Portuguese, the clitic *se* can function either as a true argument or as a grammatical function–changing morpheme.

- ```
  ( (IP-MAT (NP-SBJ A Marta)
            (VB-D lavou)
            (NP-OB1 a roupa)
            (. .)))
  ```

- ```
  ( (IP-MAT (NP-SBJ A Marta)
            (NP-OB1 (CL se))        <--- argument  'se'
            (VB-D lavou)
            (. .)))
  ```

Passive *se*

- ```
  ( (IP-MAT (NP-SBJ os jarros)
            (NP-SE (CL se))          <--- passive 'se'
            (VB-D quebraram)
            (. .)))
  ```

- ```
  ( (IP-MAT (NP-SBJ os jarros)
            (SR-D foram)
            (VAN-P quebrados)
            (. .)))
  ```

Omitting undecidable information

```
( (IP-MAT (NP-SBJ las crian,cas)))
          (NP-??? (CL se))
          (VB-D lavaram)
          (. .)))
```

Did the children wash themselves? se = NP-OB1
Or were they washed by someone else? se = NP-SE

A useful solution: undecidable or difficult instances of *se* are labelled NP-SE by default

How we build a parsed corpus - a flowchart

- POS tagging
  - Automatic preprocessing (punctuation, contractions)
  - Automatic tagging (Brill 1995)
  - Human correction
- Parsing
  - Automatic parsing (Collins 1996, Bikel 2004)
  - Human editing (= correction + addition of information)
- Final editing (partially automated)

Correction software

- We use correction software developed in connection with the Penn Treebank (http://www.cis.upenn.edu/~treebank) and implemented in Emacs Lisp
- Incorrect tags are corrected by positioning cursor on item to be corrected and entering correct tag
- Proposed tag is checked to ensure that new tag is legal
- Incorrect structures can be corrected with mouse clicks and modifier keys
- All correction software leaves input text inviolate

POS tagging - Automatic stage

- Text is tokenized
  - Punctuation is split off from words
  - Contractions are decomposed into (possibly abstract) constituents
    we'll $\rightarrow$ $we/PRO $'ll/MD {TEXT:we'll}/CODE
    pelos $\rightarrow$ $por/P $os/D {TEXT:pelos}/CODE
- Text is run through tagger (in our case, Brill 1995)

The Brill tagger

- Step 1:
  Based on a training corpus (= a relatively large corpus of already tagged text), each word is tagged with its most frequent part of speech

  He/PRO opened/VBD a/D can/MD of/P soup/N

- Step 2:
  Tagger guesses at the tag for words that are not in the training corpus

  Wimple/? $\rightarrow$ Wimple/NPR
  wimple/? $\rightarrow$ wimple/N

The Brill tagger, 2

- Step 3:
  Tagger refines guesses from Step 2 on the basis of morphological clues

  wimpleless/N $\longrightarrow$ wimpleless/ADJ

- Step 4:
  Tagger adjusts tags from Step 1 in light of context

  . . . a/D can/MD of/P soup/N $\longrightarrow$ . . . can/N . . .

Sample raw text

# indicates continuation of line in source edition.

```
My Lord,
   I return my most humble thankes for y=e= honour of y=r=      #
Lord=ps= letter.
I have not yet bin any were, but at shopes and a veseting; but  #
I
believe shall be on Munday at a ball at St. Jeames, where, as   #
they
tell me, ther is a famose new danser to apere, which is to      #
charme
us all, but not make amends for y=e= loss of M=rs= Ibbings who  #
danced
at Lincolns Inn Feild and is lately dead.
```

Sample tokenized text

Punctuation has been split off.

```
My Lord , I return my most humble thankes for y=e= honour
of y=r= Lord=ps= letter .
I have not yet bin any were , but at shopes and a veseting ;
but I believe shall be on Munday at a ball at St. Jeames ,
where , as they tell me , ther is a famose new danser to
apere , which is to charme us all , but not make amends
for y=e= loss of M=rs= Ibbings who danced at Lincolns
Inn Feild and is lately dead .
```

Sample tagged text before correction

Tagger errors are highlighted in red. The narrow text formatting facilitates human correction.

My/PRO$ Lord/N ,/, I/PRO
return/VBP my/PRO$ most/QS
humble/ADJ thankes/NS for/P
y=e=/D honour/N of/P
y=r=/PRO$ Lord=ps=/N$
letter/N ./.
I/PRO have/HVP not/NEG
yet/ADV bin/BEN any/Q
were/BED ,/, but/P at/P
shopes/NS and/CONJ a/D
veseting/VAG ;/.
but/CONJ I/PRO believe/VBP
shall/MD be/BE on/P
Munday/NPR at/P a/D ball/N
at/P St./NPR Jeames/NPR ,/,

where/WADV ,/, as/P they/PRO
tell/VBP me/PRO ,/, ther/EX
is/BEP a/D famose/ADJ new/ADJ
danser/N to/TO apere/VB ,/,
which/WPRO is/BEP to/TO
charme/VB us/PRO all/Q ,/,
but/P not/NEG make/VB
amends/NS for/P y=e=/D loss/N
of/P M=rs=/NPR Ibbings/NPR
who/WPRO danced/VBD at/P
Lincolns/NPR Inn/NPR
Feild/NPR and/CONJ is/BEP
lately/ADV dead/ADJ ./.

Sample tagged text after correction

Tagger errors are highlighted in red; human corrections in green.

My/PRO$ Lord/N ,/, I/PRO return/VBP my/PRO$ most/QS humble/ADJ thankes/NS for/P y=e=/D honour/N of/P y=r=/PRO$ Lord=ps=/N$ letter/N ./. I/PRO have/HVP not/NEG yet/ADV bin/BEN any/Q were/BED*/WADV ,/, but/P at/P shopes/NS and/CONJ a/D*/P veseting/VAG*/N ;/. but/CONJ I/PRO believe/VBP shall/MD be/BE on/P Munday/NPR at/P a/D ball/N at/P St./NPR Jeames/NPR ,/, where/WADV ,/, as/P they/PRO tell/VBP me/PRO ,/, ther/EX is/BEP a/D famose/ADJ new/ADJ danser/N to/TO apere/VB ,/, which/WPRO is/BEP to/TO charme/VB us/PRO all/Q ,/, but/P*/CONJ not/NEG make/VB amends/NS for/P y=e=/D loss/N of/P M=rs=/NPR Ibbings/NPR who/WPRO danced/VBD at/P Lincolns/NPR*/NPR$ Inn/NPR Feild/NPR and/CONJ is/BEP lately/ADV dead/ADJ ./.

Parsing - Automatic stage

- POS-tagged text is stripped of all but correct tags
- Text is run though a parser (Collins 1996, Bikel 2004)
- As we have seen, output of parser is in the form of formatted labeled bracketing, in which depth of indenting corresponds to depth of structural embedding

The Collins parser

- Parses strings according to structures most frequently associated to input in a training corpus
- Chooses likely attachment on the basis of both POS tags and lexical items
  - paint the man with a brush (high attachment)
  - paint the man with a telescope (low attachment)
- Like the Brill tagger, the Collins parser can be trained

Parsing - Human editing stage

Editing operations include:

- Changing syntactic tags
- Adding subcategory information
  - ADVP $\rightarrow$ ADVP-TMP, ADVP-LOC, . . .
  - CP $\rightarrow$ CP-THT, CP-QUE, CP-CMP, . . .
  - NP $\rightarrow$ NP-SBJ, NP-OB1, NP-MSR, . . .
- Changing attachment level
- Breaking up run-on sentences or consolidating fragments

Parsing - Human editing stage, 2

- Adding empty categories (gaps, silent understood subjects, etc.)
- Adding matching indices to gaps and their antecedents
  - What did you drink _?
- Adding matching indices to expletives ('it', 'there') and their associates
  - It is clear that they are coming .
  - There is a unicorn in the garden.

Sample parsed text, before correction

```
( (IP-MAT (NP-SBJ (PRO I))
          (HVP have)
          (NEG not)
          (ADVP (ADV yet))          <--- missing -TMP label
          (BEN bin)
          (NP-ACC (Q any))
          (CP (WADVP (WADV were))   <--- parser misled by
          (, ,)                          unusual word boundary
          (C 0)                     <--- spurious complementizer
          (PP (P but) (P at)        <--- parser wrongly treats
              (CONJP (CONJ and)          'but at' like 'out of'
                     (PP (P a)
                         (NP (N veseting)))))))
          (. ;)))
```

Sample parsed text, after correction

```
( (IP-MAT (NP-SBJ (PRO I))
          (HVP have)
          (NEG not)
          (ADVP-TMP (ADV yet))
          (BEN bin)
          (ADVP-LOC (Q any) (WADV were)
                    (, ,)
                    (PP (P but)
                        (PP (PP (P at)
                                (NP (NS shopes))
                            (CONJP (CONJ and)
                                   (PP (P a)
                                       (NP (N veseting))))))))
          (. ;)) (ID ALHATTON,2,240.6))
```

Some recent advances in automation

- The Collins parser is now superseded by Bikel 2004
- Bikel parser based on similar principles as Collins parser
- Allows modification of linguistic parameters, allowing more cross-linguistic flexibility
- Outputs includes grammatical function tags (-SBJ, -OB1, -OB2)

Some recent advances in automation, 2

- Allows multiple passes through a corpus, each pass respecting the previous ones.
  - Multiple passes simplify editing task (divide and conquer)
  - Simplification means improvements in speed and consistency
  - Editing could be carried out by a mixture of more and less highly trained annotators.

- Advances in query language allow yet further automation of corpus construction.

Project management

- Mean editing speed (in language well-known to annotator):
  2,000 words/hours for POS-tagging
  1,000 words/hours for parsing

- Annotators can work approx. 4 hours/day or 20 hours/week

- Annotators are relatively easy to find and train for POS-tagging, but quite a bit harder to find and train for parsing (people are used to thinking about words, but not in terms of constituent structure)

So how long does it take to produce a parsed corpus of 1 M words?

- POS-tagging stage
  - 1,000,000 words / 2,000 words/hours = 500 hours
  - 500 hours / 20 hours/week = 25 weeks
- Parsing stage
  - 1,000,000 words / 1,000 words/hours = 1,000 hours
  - 1,000 hours / 20 hours/week = 50 weeks
- Total: 75 weeks

CorpusSearch, a search engine for parsed corpora

- A corpus without a search program is like the Internet without Google
- Enter CorpusSearch (Randall 2000), a dedicated search engine for parsed corpora
- Written in Java
- Runs under Linux, Mac, Unix, Windows

Properties of CorpusSearch

- Basic search functions are linguistically intuitive
  (immediately) precedes, (immediately) dominates
- End user can custom-define further linguistically relevant search expressions
- Searches can disregard material as necessary
- A key feature: The output of CorpusSearch is itself searchable

A key feature: Searchable output

- Complicated and error-prone monster queries can be implemented as a sequence of simpler queries.
- Sequences of queries are consistent with the way that corpus research proceeds, via a successive refinement of hypotheses.
- Generating searchable output slows CorpusSearch down somewhat (searches of 1-2M words can take 2-3 minutes)

A simple sample query

```
node: IP*
query: ((IP* iDomsNum1 NP-ACC)
        AND (IP* iDomsNum2 MD))
```

- IP* matches IP-MAT, IP-SUB, IP-INF, etc.
- CorpusSearch searches the corpus for constituents with the label(s) specified in node.
- Whenever it finds such a constituent, it checks whether the material in the constituent matches the condition(s) in query.
  No match: I will eat the pie.
  Match: The pie will I eat. (possible in older forms of English)
- Matching instances of node are recorded in an output file.

A possible query, but long-winded and error-prone

```
node: IP*
query: ((IP* iDomsNum1 NP-ACC | NP-DAT | NP-GEN)
        AND
        (IP* iDomsNum2 BE-PRES | BE-PAST |
                       DO-PRES | DO-PAST |
                       HAVE-PRES | HAVE-PAST |
                       MD | VB-PRES | VB-PAST))
```

A better way

```
define: v2.def

node: IP*

query: ((IP* iDomsNum1 OBJECT)
        AND (IP* iDomsNum2 FINITE-VERB))
```

Contents of the definition file v2.def:

```
OBJECT: NP-ACC | NP-DAT | NP-GEN
FINITE-VERB: *-PRES | *-PAST | MD
```

Ignoring material

- CorpusSearch ignores certain material by default.
  - punctuation
  - page numbers
  - editorial comments
- The default is overridable.
- In addition, other material can be ignored as convenient or necessary (gaps, interjections, parentheticals, vocatives, etc.).

Recent advances in CorpusSearch

- NOT and OR now function more intuitively
- Extraction of subcategorization frames
- "Search and replace" annotation support

Search and replace annotation support

- According to our annotation guidelines, all of the following sentences have parallel structures and include a (possibly silent) complementizer (= subordinating conjunction).
  - I know _ you are coming.
    I know that you are coming.
  - They wonder when _ you arrived.
    They wonder when that you arrived.
    (possible in older forms of English)

- In the past, silent complementizers had to be added by hand or with Perl scripts.

- Now, silent complementizers (and if necessary, traces) can be added automatically, saving days or even weeks of work

Before and after "search and replace"

- ```
  ( (IP-MAT (NP-SBJ (PRO They))
            (VBP wonder)
            (CP-QUE (WADVP (WADV when))
                    (IP-SUB (NP-SBJ you)
                            (VBD arrived))) (. .)))
  ```

- ```
  ( (IP-MAT (NP-SBJ (PRO They))
            (VBP wonder)
            (CP-QUE (WADVP (WADV when))
                    (C 0)                       <--- added
                    (IP-SUB (ADVP *T*)          <--- added
                            (NP-SBJ you)
                            (VBD arrived))) (. .)))
  ```

Automatic regularization of P+D combinations

- ```
  (PP (P+D-F-P pelas)
      (N-P meninas))
  ```

- ```
  (PP (P $por)
      (NP (D-F-P $as)
          (CODE {TEXT:pelas})
          (N-P meninas)))
  ```

An example from the EModEng corpus

Points of interest (see next slide)

- Expletive *there* is coindexed with logical subject
- Annotation indicates where (silent) relative pronoun is interpreted
- Tokens are identified by reference labels

```
ALHATTON  2,    241.  7
text ID   vol.  page  serial token number
```

Volume number is optional; serial token number is unique within text.

Example sentence 1

```
( (IP-MAT (NP-SBJ=1 (EX There))
          (BEP is)
          (NP-1 (ONE one) (NPR M=r=) (NPR Colson)
                (CP-REL (WNP-2 0)
                        (C 0)
                        (IP-SUB (NP-SBJ (PRO I))
                                (BEP am)
                                (ADJP (ADJ shure)
                                      (CP-THT (C 0)
                                              (IP-SUB (NP-ACC *T*-2)
                                                      (NP-SBJ (PRO$ my) (N Lady))
                                                      (HVP has)
                                                      (VBN seen)
                                                      (PP (P at)
                                                          (NP (N diner)
                                                              (PP (P w=th=)
                                                                  (NP (PRO$ my)
                                                                      (N Unckle)))))))))))
          (. .)) (ID ALHATTON,2,241.7))
```

A second example from the EModEng corpus

Points of interest (see next slide)

- Annotation indicates dependency between measure phrase (*so much*) and degree complement clause
- Locative (as well as directional and temporal) AdvPs are specially marked.

Example sentence 2

```
( (IP-MAT (NP-SBJ (PRO I))
          (HVP have)
          (NP-ACC (QP (ADVR so) (Q much)
                      (CP-DEG *ICH*-1))
                  (N business))
          (ADVP-LOC (ADV here))
          (CP-DEG-1 (C y=t=)
                    (IP-SUB (NP-SBJ (PRO I))
                            (VBP hope)
                            (CP-THT (C 0)
                                    (IP-SUB (NP-SBJ (PRO$ my) (N Lady))
                                            (MD will)
                                            (VB excuse)
                                            (NP-ACC (PRO me))
                                            (PP (P till)
                                                (NP (ADJS next)
                                                    (N post))))))))
          (. .)) (ID ALHATTON,2,245.46))
```