Building and searching large diachronic parsed corpora

Beatrice Santorini
Department of Linguistics
University of Pennsylvania

Deutsch Diachron Digital
Berlin - December 2003

1 A useful URL

http://www.ling.upenn.edu/∼beatrice/corpus-ling.html

- Slides (expanded version of this presentation)
- Handout
- Links to corpus homepages
- Links to manuals

2 Some tokens from the EModEng corpus

2.1 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))
```

See end of handout for list of POS tags and syntactic tags.

2.2 Sentence 2

```
( (IP-MAT (NP-SBJ (PRO I))
          (HVP have)
          (NP-ACC (NP-MSR (QP (ADVR so) (Q much)
                          (CP-DEG *ICH*-1)))
                  (N buisness)                      <------ not a typo!
                  (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))
```

2.3 Sentence 3

```
( (IP-MAT (CONJ And)
          (ADVP (ADV so))
          (, ,)
          (NP-SBJ (PRO he)
                  (CP-REL (WNP-1 0)
                          (C that)
                          (IP-SUB (NP-SBJ *T*-1)
                                  (VBD vsed)
                                  (IP-INF (TO to)
                                          (VB teache)))))
          (, ,)
          (DOD did)
          (NEG not)
          (ADVP (ADV commonlie))
          (VB vse)
          (IP-INF (TO to)
                  (VB beate))
          (. ,)) (ID ASCH,12R.29))
```

3 Sample output of a CorpusSearch query

3.1 Header

```
/*
PREFACE:  regular output file.
CorpusSearch copyright Beth Randall 2000.
Date:  Thu Dec 04 18:38:30 EST 2003

command file:     ddd.q
input files:      /pkg/ling/MIDENG/EMODENG/ref/penn2/ambass-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/anhatton-p2-e3.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/apoole-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/armin-p2-e2.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/asch-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/aungier-p2-e3.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/authnew-p2-e2.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/authold-p2-e2.ref
output file:      /home/beatrice/corpus-linguistics/ddd.out

remark:  finds instances of do-support

definition file:  ddd03.def
shorthand:  ((finite_do precedes NEG)
  AND (NEG precedes VB))
node:   IP-MAT*|IP-SUB*
query:  ((DOD|DOP precedes NEG)
  AND (NEG precedes VB))
*/
```

3.2 Sample token with CorpusSearch report

```
--------------------------------- SNIP ---------------------------------
/*
HEADER:
source file:  ASCH
*/
/~*
And so, he that vsed to teache, did not commonlie vse to beate,
(ASCH,12R.29)
*~/
/*
1 IP-MAT: 18 DOD did, 19 NEG not, 22 VB vse
*/

(0 (1 IP-MAT (2 CONJ And)
             (3 ADVP (4 ADV so))
             (5 , ,)
             (6 NP-SBJ (7 PRO he)
                       (8 CP-REL (9 WNP-1 0)
                                 (10 C that)
                                 (11 IP-SUB (12 NP-SBJ *T*-1)
                                            (13 VBD vsed)
                                            (14 IP-INF (15 TO to) (16 VB teache)))))
             (17 , ,)
             (18 DOD did)
             (19 NEG not)
             (20 ADVP (21 ADV commonlie))
             (22 VB vse)
             (23 IP-INF (24 TO to) (25 VB beate))
             (26 . ,))
      (ID ASCH,12R.29))
--------------------------------- SNIP ---------------------------------
/*
FOOTER
source file:  ASCH
hits found:  4
tokens containing the hits:  4
total tokens searched:  191
*/
--------------------------------- SNIP ---------------------------------
```

3.3 Summary

```
/*
SUMMARY:  regular output file.

command file:     ddd.q
input files:      /pkg/ling/MIDENG/EMODENG/ref/penn2/ambass-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/anhatton-p2-e3.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/apoole-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/armin-p2-e2.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/asch-p2-e1.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/aungier-p2-e3.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/authnew-p2-e2.ref
              /pkg/ling/MIDENG/EMODENG/ref/penn2/authold-p2-e2.ref
output file:      /home/beatrice/corpus-linguistics/ddd.out

source files, hits/tokens/total
  AMBASS                0/0/29
  ANHATTON              0/0/41
  APOOLE                0/0/13
  ARMIN                 1/1/399
  ASCH                  4/4/191
  AUNGIER               1/1/45
  AUTHNEW               0/0/793
  AUTHOLD               3/3/574
grand total hits :  9
grand total tokens containing hits:  9
grand total tokens searched:  2085
*/
```

4 Sample output of a coding query

4.1 Header

```
/*
    PREFACE: coding file.
    CorpusSearch copyright Beth Randall 1999.

    Date:  Thu Dec 04 18:54:11 EST 2003

command file:     ddd.c
input file:       /pkg/ling/MIDENG/EMODENG/ref/penn2/asch-p2-e1.ref
output file:      /home/beatrice/corpus-linguistics/ddd.cod

    remark:
        1: variant: o)ld, n)ew
        2: time period
        3: sex of author: f)emale, m)ale
        4: type of text: b)ible, n)onvernacular, v)ernacular

    definition file:  ddd03.def
```

4.2 Header, continued

```
1:  {
    o: (finite_verb precedes NEG)
    n: ((finite_do precedes NEG)
      AND (NEG precedes VB))
    }

2:  {
    1: (*AMBASS* inID)
    3: (*ANHATTON* inID)
    1: (*APOOLE* inID)
    2: (*ARMIN* inID)
    1: (*ASCH* inID)
    3: (*AUNGIER* inID)
    2: (*AUTHNEW* inID)
    2: (*AUTHOLD* inID)
    }

3:  {
    f: (*ANHATTON* inID)
    f: (*APOOLE* inID)
    m:  ELSE
    }

4:  {
    v: (*ANHATTON* inID)
    v: (*APOOLE* inID)
    v: (*ARMIN* inID)
    b: (*AUTHNEW* inID)
    b: (*AUTHOLD* inID)
    n:  ELSE
    }

*/
```

4.3 Sample token with CorpusSearch report and coding string

```
--------------------------------- SNIP ---------------------------------
/~*
And so, he that vsed to teache, did not commonlie vse to beate,
(ASCH,12R.29)
*~/

(0  (0 CODING n:1:m:n)
    (1 IP-MAT (2 CONJ And)
              (3 ADVP (4 ADV so))
              (5 , ,)
              (6 NP-SBJ (7 PRO he)
                        (8 CP-REL (9 WNP-1 0)
                                  (10 C that)
                                  (11 IP-SUB (12 NP-SBJ *T*-1)
                                             (13 VBD vsed)
                                             (14 IP-INF (15 TO to) (16 VB teache)))))
              (17 , ,)
              (18 DOD did)
              (19 NEG not)
              (20 ADVP (21 ADV commonlie))
              (22 VB vse)
              (23 IP-INF (24 TO to) (25 VB beate))
              (26 . ,))
    (27 ID ASCH,12R.29))
--------------------------------- SNIP ---------------------------------
```

4.4 Summary

```
/*
    SUMMARY:  coding file.

command file:     ddd.c
input file:       /pkg/ling/MIDENG/EMODENG/ref/penn2/asch-p2-e1.ref
output file:      /home/beatrice/corpus-linguistics/ddd.cod

    source files, total:
        ASCH           191
    grand total sentences searched:  191
*/
```

5 POS tagging

5.1 List of POS tags

ADJ	adjective
ADJR	adjective, comparative
ADJS	adjective, superlative
ADV	adverb
ADVR	adverb, comparative
ADVS	adverb, superlative
ALSO	The words ALSO (except when = AS) and EKE
BAG	present participle of BE
BE	infinitive of BE
BED	past of BE
BEI	imperative of BE
BEN	perfect participle of BE
BEP	present of BE
C	complementizer
CODE	nonlinguistic material
CONJ	coordinating conjunction
D	determiner
DAG	present participle of DO
DAN	passive participle of DO
DO	infinitive of DO
DOD	past of DO
DOI	imperative of DO
DON	perfect participle of DO
DOP	present of DO
ELSE	the word ELSE (in the collocation OR ELSE)
EX	existential THERE
FOR	the complementizer FOR
FP	focus particle
FW	foreign word

HAG	present participle of HAVE
HAN	passive participle of HAVE
HV	infinitive of HAVE
HVD	past of HAVE
HVI	imperative of HAVE
HVN	perfect participle of HAVE
HVP	present of HAVE
ID	token identifier
INTJ	interjection
MD	modal verb
MD0	untensed modal verb (possible in Middle English)
N	common noun, singular or mass
NEG	negation
NPR	proper noun, singular
NPRS	proper noun, plural
NS	common noun, plural
NUM	cardinal number
ONE	the word ONE (except as focus particle)
OTHER	the word OTHER (except as conjunction)
OTHERS	plural nominal use of OTHER
P	preposition or subordinating conjunction
PRO	personal pronoun
Q	quantifier
QR	quantifier, comparative (MORE, LESS, FEWER)
QS	quantifier, superlative (MOST, LEAST, FEWEST)
RP	adverbial particle
SUCH	the word SUCH
TO	infinitival AT (in northern Middle English) and TO

VAG	present participle of ordinary verb
VAN	passive participle of ordinary verb
VB	infinitive of ordinary verb
VBD	past of ordinary verb
VBI	imperative of ordinary verb
VBN	perfect participle of ordinary verb
VBP	present of ordinary verb
WADV	wh- adverb
WARD	the morpheme WARD
WD	wh- determiner
WPRO	wh- pronoun
WQ	WHETHER introducing indirect questions
X	completely unclear POS
$	possessive marker (as in "John his book")

5.2 Comments

- $ can also be appended to a tag to indicate possessive.

  the/D applicant's/N$
  King/NPR Henry's/NPR$
  my/PRO$ mother/N
  whose/WPRO$ books/NS

- Present and past tags include subjunctive uses.

  if it be/BEP so
  I wish that it were/BED so

- Compound tags are possible.

  everlasting/ADV+VAG
  whosoever/WPRO+ADV+ADV

6 Syntactic annotation

6.1 List of syntactic tags

ADJP	adjective phrase
ADJP-LOC	locative ADJP
ADJP-SPR	ADJP secondary predicate
ADVP	ordinary adverb phrase
ADVP-DIR	directional ADVP
ADVP-LOC	locative ADVP
ADVP-TMP	temporal ADVP
CONJP	conjunction phrase
CP-ADV	adverbial complement clause
CP-CAR	clause-adjoined relative
CP-CLF	it-cleft
CP-CMP	comparative clause
CP-DEG	degree complement clause
CP-FRL	free relative clause
CP-QUE	question
CP-REL	relative clause
CP-THT	ordinary complement clause
CP-TMC	tough-movement complement
FRAG	fragment
IP-ABS	absolute clause
IP-INF	complement infinitive
IP-INF-ADT	adjunct infintive
IP-INF-DEG	degree infinitive
IP-INF-PRP	purpose infinitive
IP-MAT	matrix clause
IP-PPL	participial clause
IP-SMC	small clause

NP	noun phrase
NP-ADT	adjunct NP
NP-ADV	NP adverb
NP-COM	NP complements of nouns (as in “this side Eden”)
NP-DOP	dative of possession (in Middle English)
NP-LFD	left-dislocated NP
NP-MSR	measure NP
NP-OB1	first object
NP-OB2	second object
NP-POS	possessive NP
NP-RFL	reflexive NP
NP-SBJ	subject
NP-SPR	NP secondary predicate
NUMP	number phrase
PP	prepositional phrase
QP	quantifier phrase
QTP	quotation phrase
REF	reference (if part of source text)
RRC	reduced relative clause
WADJP	wh- ADJP
WADVP	wh- ADVP
WNP	wh- NP
WPP	wh- PP
WQP	wh- QP
XX	completely unclear constituent

6.2 List of empty categories

(NP-SBJ *con*)	subject elided under conjunction
(NP-SBJ *exp*)	empty expletive subject
(NP-SBJ *pro*)	"small pro" subject
(NP-SBJ *arb*)	arbitrary silent subject in ECM infinitives (as in "I hear tell")
(XP *)	elided or understood constituent (say, a modal or a verb)
(XP T)	wh- trace
(XP ICH)	non-wh trace (ICH = Interpret Constituent Here)

6.3 Comments

- Unclear or undecidable hierarchical level of constituents are indicated by X

ADJX	not clear whether ADJ or ADJP
NX	not clear whether N or NP

  ll

- Syntactic tags can be modified by dash tags

-PRN	parenthetical or appositive
-RSP	resumptive element
-SPE	direct speech

- Multiword FW sequences are labelled appropriately (FRENCH, LATIN, SPANISH, WELSH, etc.)

- Numerical indices encode dependencies between T or ICH constituents and their antecedents.