Using the Penn Parsed Corpora of Historical English with CorpusSearch

Waseda Workshop on the PPCHE

Anthony Kroch & Beatrice Santorini
University of Pennsylvania
December 12, 2017

- Slides for this workshop

  www.ling.upenn.edu/~kroch/handouts/

- CorpusSearch user's guide

  corpussearch.sourceforge.net/CS-manual/Contents.html

- Annotation manual for PPCHE

  www.ling.upenn.edu/~beatrice/annotation/

What is a morphosyntactically annotated corpus?

- morphological tagging
  case, gender, number features on nouns
  tense, mood, aspect features on verbs, etc.

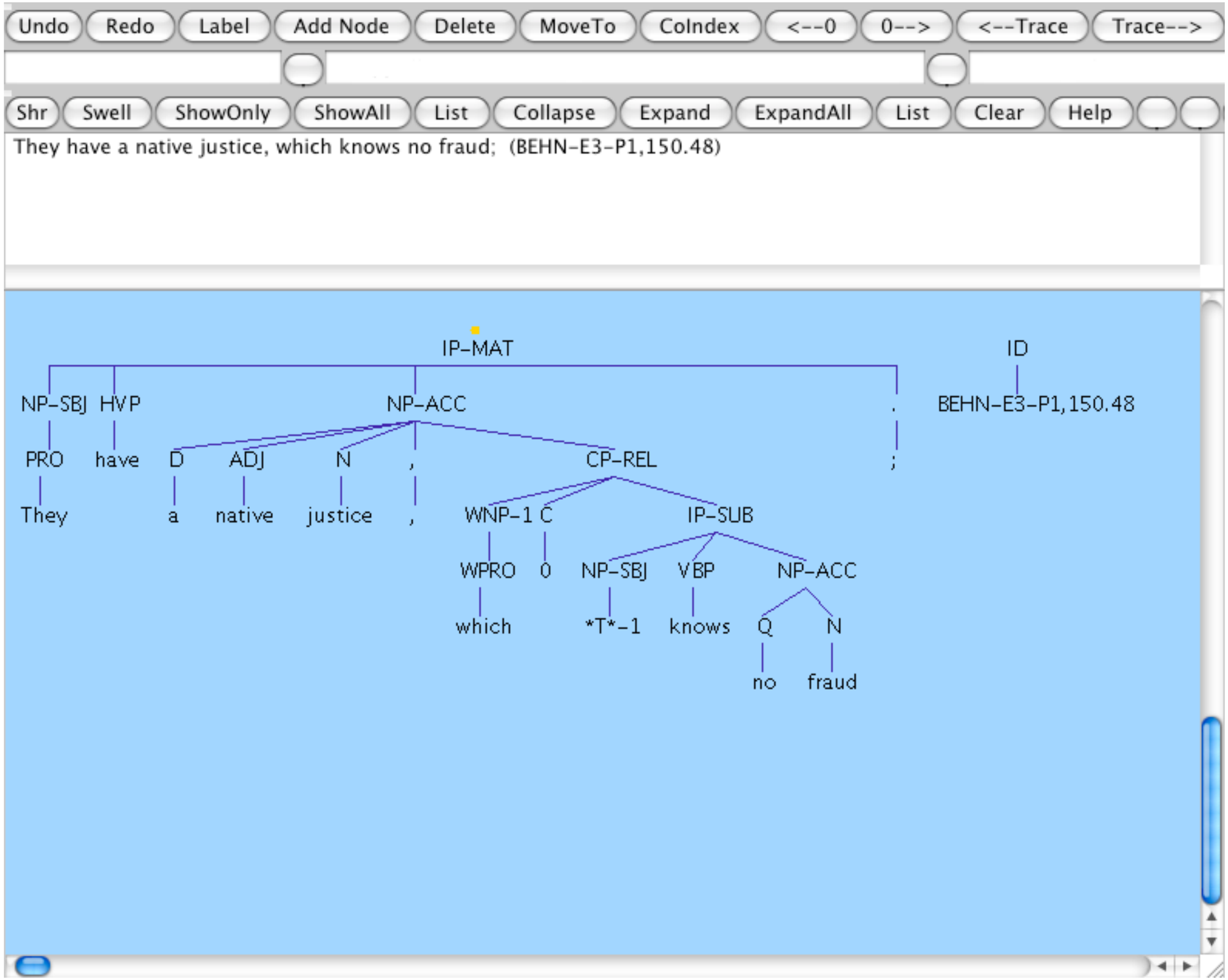
- lemmatization
  word sense disambiguation
  spelling normalization

- part of speech tagging
  elementary syntactic functions

- syntactic parsing
  hierarchical structure of phrases/clauses
  grammatical function of phrases/clauses

An example sentence

```
((IP-MAT (NP-SBJ  (PRO They))
         (HVP have)
         (NP-OB1  (D a)
                  (ADJ native)
                  (N justice)
                  (, ,)
                  (CP-REL (WNP-1 (WPRO which))
                          (C 0)
                          (IP-SUB  (NP-SBJ *T*-1)
                                   (VBP knows)
                                   (NP-OB1 (Q no)
                                           (N fraud)))))
 (. ;))
(ID BEHN-E3-P1,150.48))
```


Undo
Redo
Label
Add Node
Delete
MoveTo
CoIndex
<--0
0-->
<--Trace
Trace-->
Shr
Swell
ShowOnly
ShowAll
List
Collapse
Expand
ExpandAll
List
Clear
Help
They have a native justice, which knows no fraud; (BEHN-E3-P1,150.48)
IP-MAT
ID
NP-SBJ
HVP
NP-ACC
.
BEHN-E3-P1,150.48
PRO
have
D
ADJ
N
,
CP-REL
;
They
a
native
justice
,
WNP-1
C
IP-SUB
WPRO
0
NP-SBJ
VBP
NP-ACC
which
T-1
knows
Q
N
no
fraud

Reformated sentence for visualization

```
((IP-MAT        (NP-SBJ      (PRO They))
                             (HVP have)
                (NP-OB1      (D a)
                             (ADJ native)
                             (N justice)
                             (, ,)
 (CP-REL        (WNP-1       (WPRO which))
                             (C 0)
 (IP-SUB        (NP-SBJ      *T*-1)
                             (VBP knows)
                (NP-OB1      (Q no)
                             (N fraud)))))
                             (. ;))
                             (ID BEHN-E3-P1,150.48))
```

The example with lemmatization

```
((IP-MAT    (NP-SBJ    (PRO       (ORTHO They)
                                  (METAWORD
                                            (LEMMA (HEADWORD they)
                                                   (OEDID 200700)))))
                       (HVP       (ORTHO have)
                                  (METAWORD
                                            (LEMMA (HEADWORD have)
                                                   (OEDID 84705))))
            (NP-OB1    (D         (ORTHO a)
                                  (METAWORD
                                            (LEMMA (HEADWORD a)
                                                   (OEDID 4))))
                       (ADJ       (ORTHO native)
                                  (METAWORD
                                            (LEMMA (HEADWORD native)
                                                   (OEDID 125304))))
                       (N         (ORTHO justice)
                                  (METAWORD
                                            (LEMMA (HEADWORD justice)
                                                   (OEDID 102198)))))
                       (,         (ORTHO ,)
                                  (METAWORD
                                            (LEMMA (HEADWORD ,)
                                                   (OEDID NA))))
```

```
(CP-REL     (WNP-1      (WPRO     (ORTHO which)
                                  (METAWORD
                                            (LEMMA (HEADWORD which)
                                                   (OEDID 228284)))))
                        (C        (METAWORD (ALT-ORTHO 0)
                                            (LEMMA 0)))
(IP-SUB
            (NP-SBJ               (METAWORD (ALT-ORTHO *T*-1)
                                            (LEMMA 0)))
                        (VBP      (ORTHO knows)
                                  (METAWORD
                                            (LEMMA (HEADWORD know)
                                                   (OEDID 104157))))
            NP-ACC      (Q        (ORTHO no)
                                  (METAWORD
                                            (LEMMA (HEADWORD no)
                                                   (OEDID 127437))))
                        (N        (ORTHO fraud)
                                  (METAWORD
                                            (LEMMA (HEADWORD fraud)
                                                   (OEDID 74298)))))))

                        (.        (ORTHO ;)
                                  (METAWORD

                                            (LEMMA (HEADWORD .)
                                                   (OEDID NA)))))
                        (ID       BEHN-E3-P1,150.48))
```

- morphological tagging
  case, gender, number features on nouns
  tense, mood, aspect features on verbs, etc.

- lemmatization
  word sense disambiguation
  spelling normalization

- part of speech tagging
  elementary syntactic functions

- syntactic parsing
  hierarchical structure of phrases/clauses
  grammatical function of phrases/clauses

The annotation task

- Annotation is multilevel and complex, so that using human effort for the whole job is impractical.
- At the same time, accuracy is crucial and unattainable at present with fully automated methods.
- In consequence, parsed corpora are built by interleaving automated analysis with human correction of the output.

Annotation software

- Wide range of software for automatic part-of-speech tagging and other software for automatic parsing.
- Software for correcting the errors of automated taggers.
- Annotald software for the correction of the errors of automatic parsers (annotald.github.io).
- CorpusSearch revision queries for semi-automatic parsing and parsing correction.

Available parsed corpus resources for European languages using the Penn annotation scheme

English Parsed Corpora, I

- Anthony Kroch and Ann Taylor. *Penn-Helsinki Parsed Corpus of Middle English, second edition*. University of Pennsylvania, 2000. (http://www.ling.upenn.edu/hist-corpora)

  1.3 million words

- Anthony Kroch, Beatrice Santorini, and Ariel Diertani. *Penn-Helsinki Parsed Corpus of Early Modern English*. University of Pennsylvania, 2004.

  1.8 million words

- Anthony Kroch, Beatrice Santorini, and Ariel Diertani. *Penn Parsed Corpus of Modern British English*. University of Pennsylvania, 2010.

  3.0 million words

English Parsed Corpora, II

- Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. *York-Toronto-Helsinki Parsed Corpus of Old English Prose, first edition*. Oxford Text Archive, 2003.
  (http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm)

  1.5 million words

- Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. *Parsed Corpus of Early English Correspondence, first edition*. Oxford Text Archive, 2006.

  2.2 million words

A sample of other languages, I

- Eiríkur Rögnvaldsson et al. *Icelandic Parsed Historical Corpus (IcePaHC), version 0.9, 8/2011*. (http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC))

  ≈1 million words

- France Martineau et al. *MCVF Corpus of Historical French*. University of Ottawa, 2010. (http://www.arts.uottawa.ca/voies/)

  ≈1 million words

- Charlotte Galves et al. *Tycho Brahe Corpus of Historical Portuguese*,. University of Campinas, São Paulo, Brazil, 2010. (http://www.tycho.iel.unicamp.br/~tycho/corpus/en/)

  ≈2 million words, 0.8 million parsed to date

Other languages, II

- Prashant Pardeshi et al. NINJAL Parsed Corpus of Modern Japanese (NPCMJ) & Keyaki Treebank

  ≈500K words

- Christina Tortora et al. The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE)

  ≈1 million words

- Christina Tortora et al. A Corpus of New York City English (CUNY-CoNYCE).

  multi-million word corpus under construction

Coding queries

A canonical order sentence

```
( (IP-MAT (NP-SBJ (NPR John))
          (VBP likes)
          (NP-OB1 (N pizza))
          (PUNC .)))
```

A topicalized sentence

```
( (IP-MAT (NP-OB1 (N Pizza))
          (PUNC ,)
          (NP-SBJ (NPR John))
          (VBP likes)
          (PUNC .)))
```

A verb-second (V2) sentence

```
( (IP-MAT (NP-OB1 (N Pizza))
          (VBP likes)
          (NP-SBJ (NPR John))
          (PUNC .)))
```

A coding query example

```
node: IP-MAT*

ignore_nodes: PUNC | \**

coding_query:

// grammatical status of first constituent coded in column 1

1: {

    subj:  (IP-MAT* iDomsFirst NP-SBJ*)

    obj:    (IP-MAT* iDomsFirst NP-OB1*)

    temp: (IP-MAT* iDomsFirst *-TMP)

    -:  ELSE

  }
```

Coding query, column 2

```
// position of finite verb

2: {

    \1:  (IP-MAT* iDomsNum 1 finite_verb)

    \2:  (IP-MAT* iDomsNum 2 finite_verb)

    \3:  (IP-MAT* iDomsNum 3 finite_verb)

    -:  ELSE

  }
```

Coding query, column 3

```
// subject-verb inversion?

3: {

    subj-fin:       (IP-MAT* iDoms NP-SBJ*)

                AND (IP-MAT* iDoms finite_verb)

                AND (finite_verb precedes NP-SBJ*)

    fin-subj:       (IP-MAT* iDoms NP-SBJ*)

                AND (IP-MAT* iDoms finite_verb)

                AND (NP-SBJ* precedes finite_verb)

    -:  ELSE

  }
```

Coding query, column 4

// status of subject

4: {

pron: (IP-MAT* iDoms NP-SBJ*)

AND (NP-SBJ* iDomsOnly PRO)

np: (IP-MAT* iDoms NP-SBJ*)

-: ELSE

}

The canonical order sentence, coded

((IP-MAT (CODING-IP-MAT subj : 2 : subj-fin : np)

(NP-SBJ (NPR John))

(VBP likes)

(NP-OB1 (N pizza))

(PUNC .)))

The topicalized sentence, coded

```
( (IP-MAT (CODING-IP-MAT obj : 3 : subj-fin : np)

          (NP-OB1 (N Pizza))

          (PUNC ,)

          (NP-SBJ (NPR John))

          (VBP likes)

          (PUNC .)))
```

The V2 sentence, coded

```
( (IP-MAT (CODING-IP-MAT obj : 2 : fin-subj : np)

          (NP-OB1 (N Pizza))

          (VBP likes)

          (NP-SBJ (NPR John))

          (PUNC .)))
```

Extracting coding strings for quantitative analysis

Run a .q file with only the following single line:

```
print_only: CODING*
```

Coding at more than one node

Sometimes it is useful to combine coding strings that CS generates at more node, for example at IP and at NP. It is possible to concatenate the strings into a single string.

This possibility requires the use of a function, called *concat* in the revision query module of CS, which we describe later on in this presentation.

The output for our toy example

subj : 2 : subj-fin : np

obj : 3 : subj-fin : np

obj : 2 : fin-subj : np

Some more realistic output

…

subj:2:subj-fin:np

obj:3:subj-fin:np

obj:2:fin-subj:np

obj:2:fin-subj:pro

subj:2:subj-fin:np

subj:2:subj-fin:pro

subj:2:subj-fin:np

subj:2:subj-fin:np

subj:2:subj-fin:np

obj:3:subj-fin:pro

temp:3:subj-fin:np

…

Importing coding strings into an R dataframe

	date	text	genre	clause	sbj	fin	finord	nonfinord	fnonford	DO	IO	PPT	PPA	PPC	SOord	finDOord	nfDOord	DOquant	DOlen
9	0842	stras_p	p	rel	s_pro	main	s-fin	NONE	no-aux	o_tra	o_np	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
10	0842	stras_p	p	adv	s_Pro	mod-pouv	s-x-fin	s-non	non-fin	o_cli	NONE	NONE	NONE	NONE	s-o	o-fin	non-o	cli	1
11	0842	stras_p	p	rel	s_Dem	mod-pouv	s-x-fin	s-non	non-fin	o_tra	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
12	0900	eulal_p	p	Mat	s_NP	main-etre	fin-s	NONE	no-aux	p_np	NONE	NONE	NONE	NONE	o-s	o-fin	NONE	ind	2
13	0900	eulal_p	p	Mat	s_amb2	main-avoir	s-fin	NONE	no-aux	o_np	NONE	NONE	NONE	NONE	s-o	fin-o	NONE	conj	3
14	0900	eulal_p	p	Mat	s_NP	mod-voul	fin-s	non-s	fin-non	o_cli	NONE	NONE	NONE	NONE	o-s	fin-o	o-non	cli	1
15	0900	eulal_p	p	Mat	s_amb3	mod-voul	s-fin	s-non	fin-non	o_np	NONE	NONE	NONE	NONE	s-o	fin-o	non-o	def	1
16	0900	eulal_p	p	Mat	s_Pro	main	s-fin	NONE	no-aux	o_np	NONE	NONE	NONE	NONE	s-o	fin-o	NONE	def	3
17	0900	eulal_p	p	adv	s_Pro	main	s-x-fin	NONE	no-aux	o_np	NONE	NONE	NONE	NONE	s-o	o-fin	NONE	def	1
18	0900	eulal_p	p	rel	s_tra	main	s-fin	NONE	no-aux	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
19	0900	eulal_p	p	Mat	s_NP	mod-pouv	s-fin	s-non	fin-non	o_cli	NONE	NONE	NONE	NONE	s-o	o-fin	o-non	cli	1

A case study: the rise of recipient passives in English (Bacovcin 2012)

Theme passives and recipient passives in Modern English

(1) John gave the books to Mary.

(2) The books were given to Mary (by John).

(3) John gave Mary the books.

(4) Mary was given the books (by John).

(5) *The books were given Mary (by John).

Theme passives and recipient passives in Modern German

(1) Hans gab der Maria den Artikel.
DAT ACC

(2) Der Artikel wurde der Mary (von Hans) gegeben.

(3) *Die Maria wurde den Artikel (von Hans) gegeben.

(4) Der Maria wurde der Artikel (von Hans) gegeben.

Ditransitive sentences in Early Middle English

(1) John gave Mary the book.

(2) John gave the book to Mary.

(3) John gave to Mary the book.

(4) John gave the book Mary.

Theme passives and recipient passives in Early Middle English

(1) The books were given to Mary (by John).

(2) The books were given Mary (by John).

(3) *Mary was given the book (by John).

German double accusatives

(1) Hans hat die Kinder Geschichte gelehrt.
ACC ACC

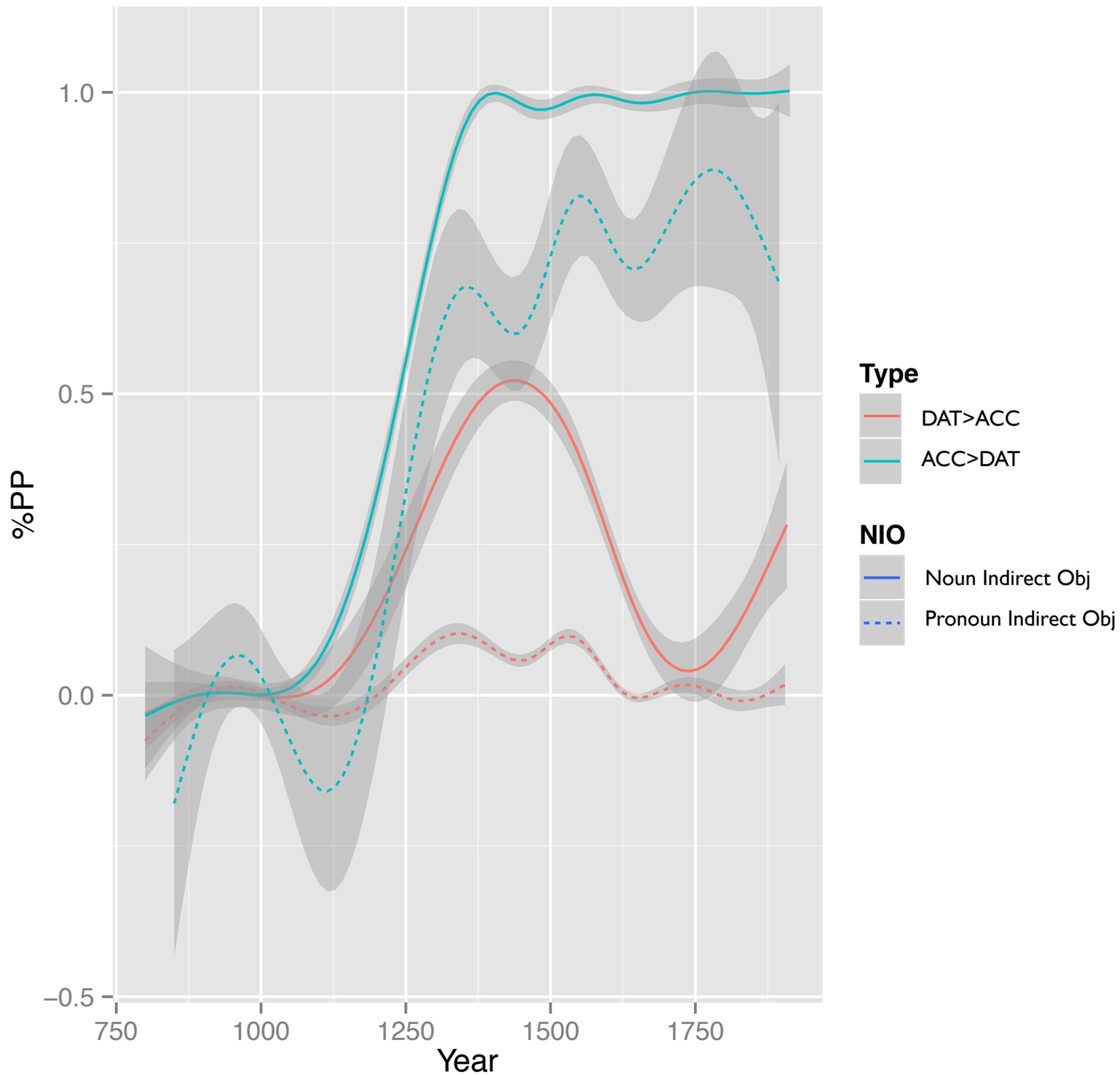
(2) ?Hans hat den Kindern Geschichte gelehrt.
DAT ACC

(3) *Geschichte wurde die Kinder gelehrt.

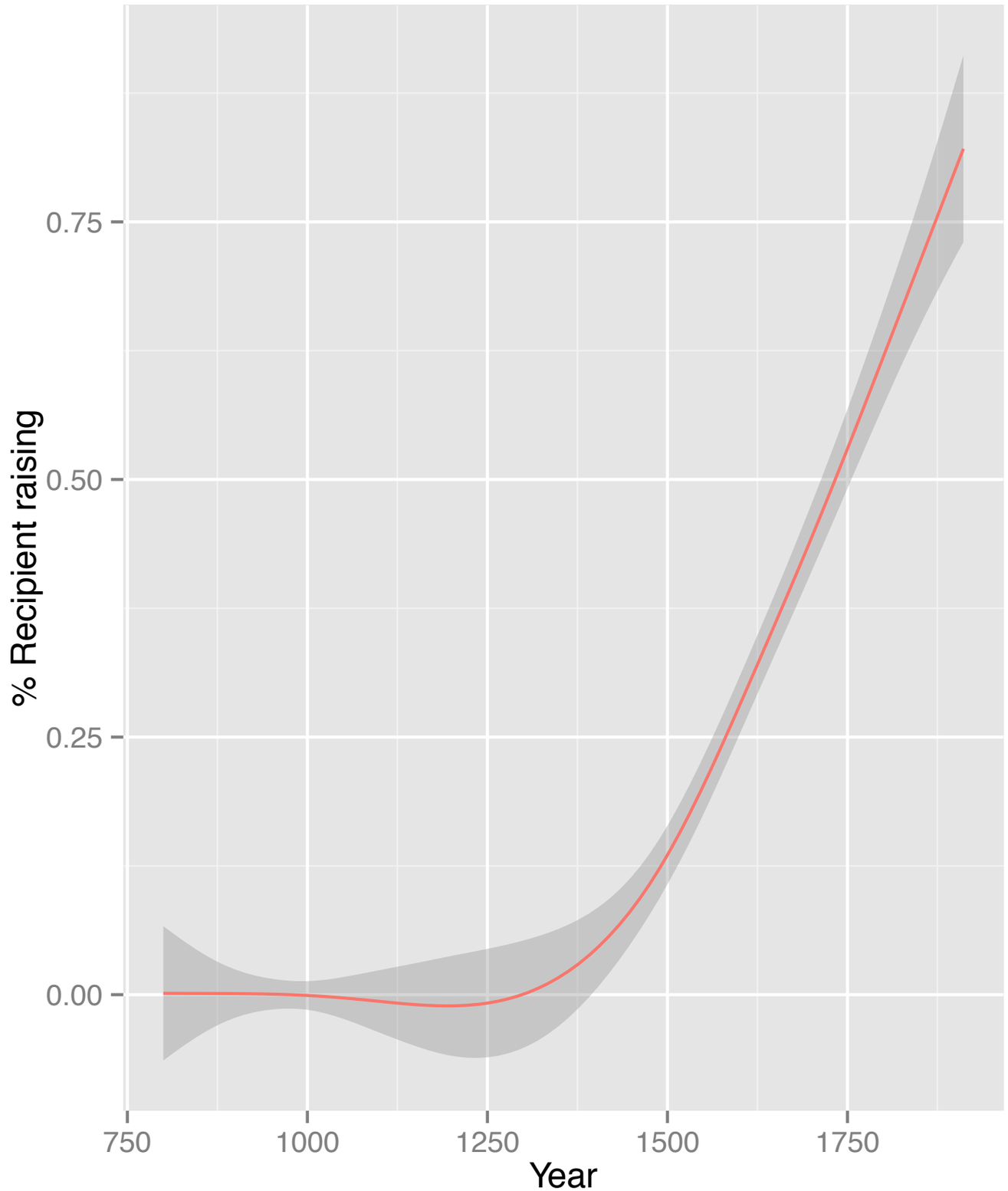
(4) Geschichte wurde den Kindern gelehrt.

(5) Die Kinder wurden Geschichte gelehrt.

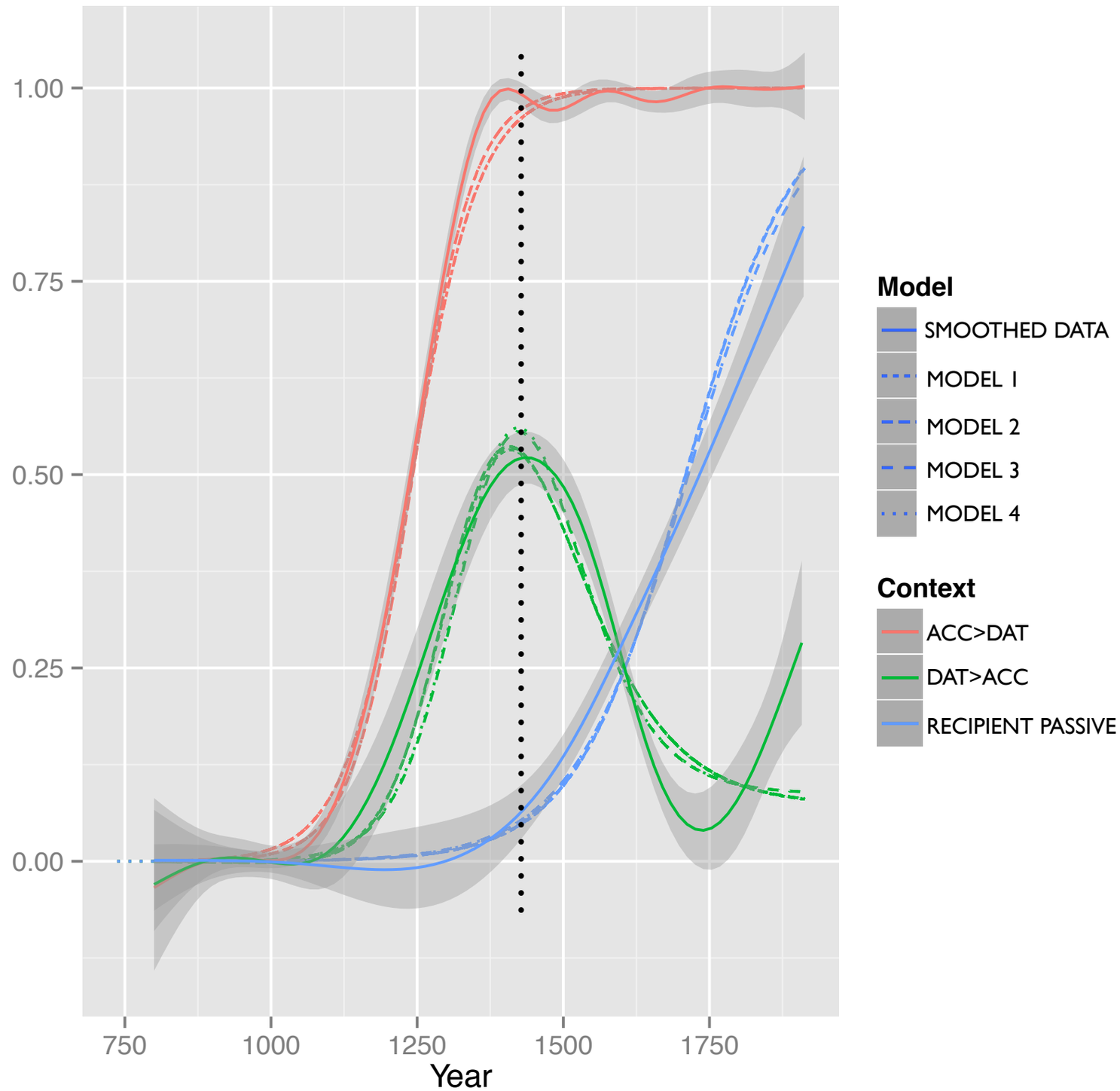
Rise in the use of prepositional indirect objects in English



Rise in recipient passives in English



Markov Chain Monte Carlo simulations of the change



Revision queries

Revision query 0.0: Concatenating coding strings

copy_corpus: t

query: (NP* iDoms CODING-NP*)

AND (CODING-NP* iDoms [1]{2}.*)

AND (NP* IDoms CP-REL*)

AND (CP-REL* iDoms CODING-CP-REL*)

AND (CODING-CP-REL* iDoms [2]{1}.*)

concat{2, 1}:

English annotation for modals: Monoclausal structure

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

(MD will)

(VB come)

(ADVP-TMP (ADVR later))))

Romance annotation for modals: Biclausal structure

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

(MD will)

(IP-INF (VB come)

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

Revision query 1.0:
From monoclausal to biclausal structure

node: $ROOT

query: (IP-* iDoms MD)

AND (IP-* iDoms [1]{1}.*)

AND (MD iPrecedes [1].*)

AND (IP-* iDomsLast [2]{2}.*)

add_internal_node{1,2}: IP-INF

But what about punctuation?

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

(MD will)

(VB come)

(ADVP-TMP (ADVR later))

(PUNC .)))

Revision query 1.1: Ignoring punctuation

node: $ROOT

ignore_nodes: PUNC

query: (IP-* iDoms MD)

AND (IP-* iDoms [1]{1}.*)

AND (MD iPrecedes [1].*)

AND (IP-* iDomsLast [2]{2}.*)

add_internal_node{1,2}: IP-INF

Revision query 2:
From biclausal to monoclausal structure

node: $ROOT

query: (IP-* iDoms MD)

AND (IP-* iDoms {1}IP-INF)

AND (MD iPrecedes IP-INF)

delete_node{1}:

ECM annotation

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

(VBD saw)

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

(VB arrive))))

Accusativus cum infinitivo annotation

```
( (IP-MAT (NP-SBJ (PRO They))
          (VBD saw)
          (NP-OB1 (PRO him))
          (IP-INF (VB arrive))))
```

Revision query 3.0:
From ECM to A.c.I.

node: $ROOT

query: (IP-* iDoms IP-INF)

AND (IP-INF iDoms {1}NP-SBJ)

move_up_node{1}:

replace_label{1}: NP-OB1

Revision query 4.1: From A.c.l. to ECM

node: $ROOT

query: (IP-* iDoms {1}NP-OB1)

AND (IP-* iDoms {2}IP-INF)

AND (NP-OB1 iPrecedes IP-INF)

move_to{1,2}:

replace_label{1}: NP-SBJ

But we <u>don't</u> want to revise cases of object control

```
( (IP-MAT (NP-SBJ (PRO They))
          (VBD persuaded)
          (NP-OB1 (PRO him))
          (IP-INF (TO to)
                  (VB come))))
```

Revision query 4.2: Restricting the revision to matrix “saw”

node: $ROOT

query: (IP-* iDoms {1}NP-OB1)

AND (IP-* iDoms V*) AND (V* iDoms saw)

AND (IP-* iDoms {2}IP-INF)

AND (NP-OB1 iPrecedes IP-INF)

move_to{1,2}:

replace_label{1}: NP-SBJ

Revision query 4.3: Using iDomsMod

node: $ROOT

query: (IP-* iDoms {1}NP-OB1)

AND (IP-* iDomsMod V* saw)

AND (IP-* iDoms {2}IP-INF)

AND (NP-OB1 iPrecedes IP-INF)

move_to{1,2}:

replace_label{1}: NP-SBJ

A trivial definitions file

```
see:     see* | saw
```

Revision query 4.4: Using the trivial definitions file

```
node: $ROOT

define: trivial.def

query:      (IP-* iDoms {1}NP-OB1)
            AND (IP-* iDomsMod V* see)
            AND (IP-* iDoms {2}IP-INF)
            AND (NP-OB1 iPrecedes IP-INF)

move_to{1,2}:

replace_label{1}: NP-SBJ
```

A less trivial definitions file

```
feel:      feel* | felt

hear:      hear*

let:       let*

see:       see* | saw

ECM-verb:      $feel | $hear | $let | $see
```

Revision query 4.5:
Using the less trivial definitions file

node: $ROOT

define: less-trivial.def

query: (IP-* iDoms {1}NP-OB1)

AND (IP-* iDomsMod V* ECM-verb)

AND (IP-* iDoms {2}IP-INF)

AND (NP-OB1 iPrecedes IP-INF)

move_to{1,2}:

replace_label{1}: NP-SBJ

End