The Geneva Corpus of Middle English Poetry: its construction and possible applications

Richard Zimmermann

Université de Genève

November 16, 2013

Outline

GeCMEP - overview

- The Geneva Corpus of Middle English Poetry (GeCMEP) is a fully annotated and syntactically parsed corpus.

GeCMEP - overview

- The Geneva Corpus of Middle English Poetry (GeCMEP) is a fully annotated and syntactically parsed corpus.
- Time: 1150-1420 (Helsinki periods M1, M2, M3)

GeCMEP - overview

- The Geneva Corpus of Middle English Poetry (GeCMEP) is a fully annotated and syntactically parsed corpus.
- Time: 1150-1420 (Helsinki periods M1, M2, M3)
- Size: goal is 100,000 words before end of PhD, but in principle open-ended

GeCMEP - overview

- The Geneva Corpus of Middle English Poetry (GeCMEP) is a fully annotated and syntactically parsed corpus.
- Time: 1150-1420 (Helsinki periods M1, M2, M3)
- Size: goal is 100,000 words before end of PhD, but in principle open-ended
- Parsed according to the rules of the Penn Parsed Corpus of Middle English (Kroch and Taylor 2000)

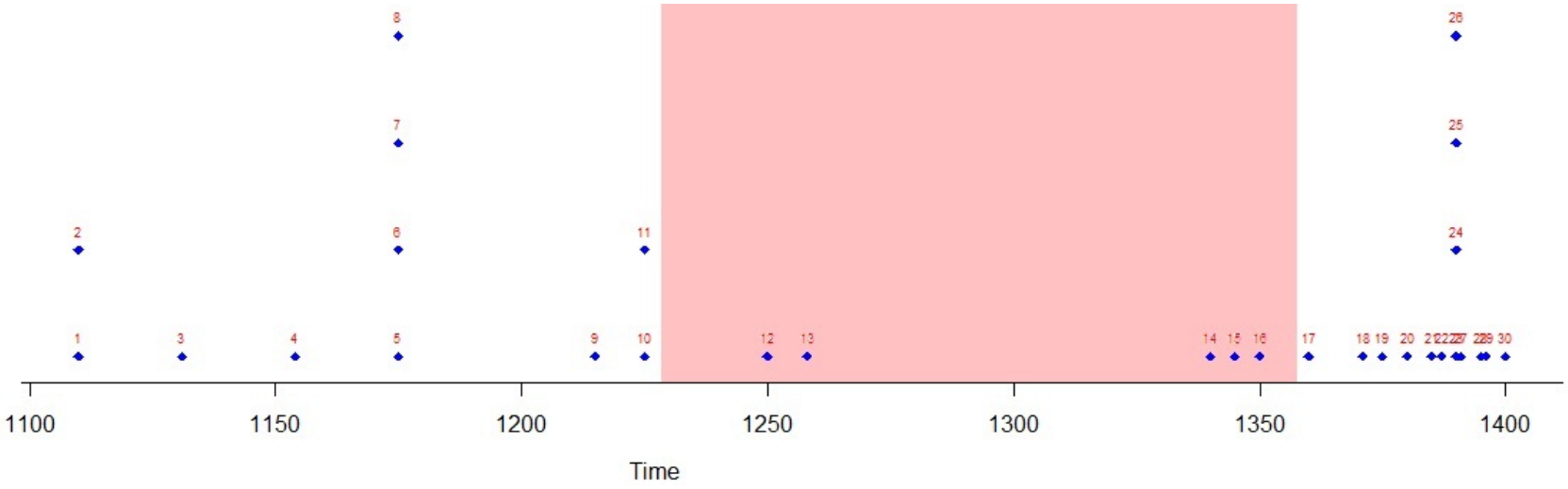
Go to the PPCM2 Manual

Example parse

(1) hwa swa ne forȝefeð heore hating, ne god ne forȝeueð him na þing
who so not forgives their hating, no God not forgives them no thing
'Whoever doesn't forgive their hate, God will not forgive them anything'
(PatNost,111.67.220)

```
(IP-MAT (NP-LFD (CP-FRL (WNP-1 (WPRO hwa) (ADV swa))
            (C 0)
            (IP-SUB (NP-SBJ *T*-1)
                (NEG ne)
                (VBP for+gefe+d)
                (NP-OB1 (PRO$ heore) (N hating)))))
    (, .)
    (NP-SBJ (Q ne) (NPR god))
    (NEG ne)
    (VBP for+geue+d)
    (NP-OB2-RSP (PRO him))
    (NP-OB1 (Q na) (N +ting))
    (. .))
(ID PatNost,111.67.220))
```

Major Middle English Prose Texts 1100-1400



1 In Festis Sancti Marie
2 Eluciderium
3 Anglo-Saxon Chronicle Continuation I
4 Anglo-Saxon Chronicle Continuation II
5 Lambeth Homilies
6 Trinity Homilies
7 Vespasian Homilies
8 Vices and Virtues
9 Ancrene Riwle
10 Katherine Group
11 Wooing Group
12 Kentish Sermons
13 Proclamation of Henry III
14 Ayenbite of Inwit
15 Richard Rolle
16 Earliest Prose Psalter
17 Mirror of Edmund
18 Travels of Sir Mandeville
19 Rievaulx' De Institutione
20 Wycliffe's works
21 Chaucer's Translation of Boethius
22 Trevisia's Polcychronicon
23 Julian of Norwich
24 Chaucer's Parson's Tale
25 Chaucer's Tale of Melibee
26 Texts from the Vernon Ms.
27 Chaucer's Treatise on the Astrolabe
28 The Cloud of Unknowing - author
29 Walter Hilton
30 The Chronicles of England

→ ME verse can help to close the prose gap c. 1250-1350

Creation of a basic text file I

- Find an electronic version of the text you want to parse

Some online resources...

Creation of a basic text file I

- Find an electronic version of the text you want to parse
- Example: Morris (1972) *An Old English Miscellany Containing a Bestiary*

Go to Archive.org

Some online resources...

Creation of a basic text file I

- Find an electronic version of the text you want to parse
- Example: Morris (1972) *An Old English Miscellany Containing a Bestiary*

  Go to Archive.org

- Copy and paste into a .txt file with ANSI formatting so that special characters like thorn, yogh etc. can be read

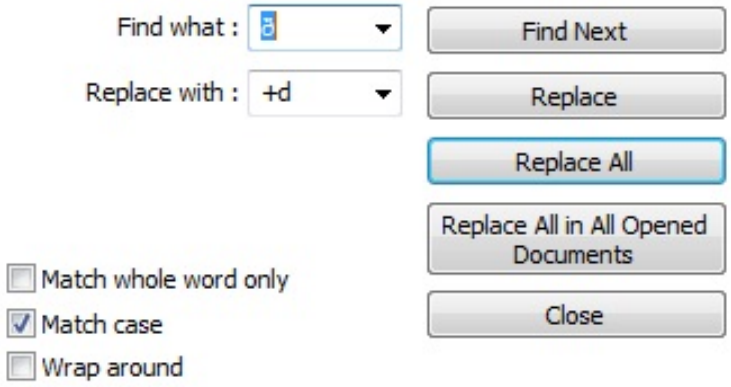
  Open Bestiary.txt

Some online resources...

Creation of a basic text file II

- Remove mark-ups, comments, footnotes, page & folio numbers, translations, critical apparatus etc. (UTF-8)

Replacement of ð in Notepad++

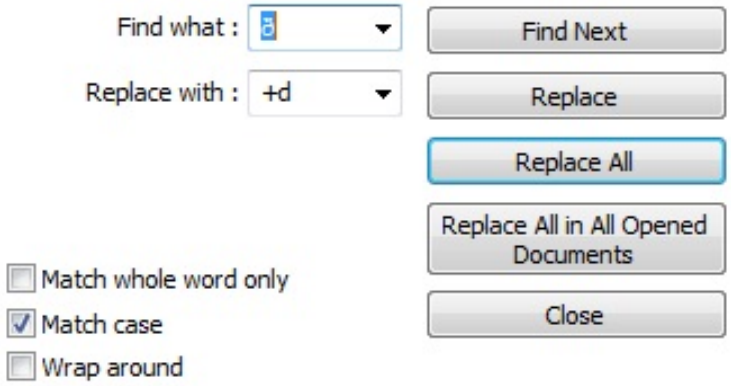


Creation of a basic text file II

- Remove mark-ups, comments, footnotes, page & folio numbers, translations, critical apparatus etc. (UTF-8)
- Replace special characters, e.g. *þ* → *+t*, *ȝ* → *+g* etc.

  Open Bestiary2.txt

Replacement of ð in Notepad++



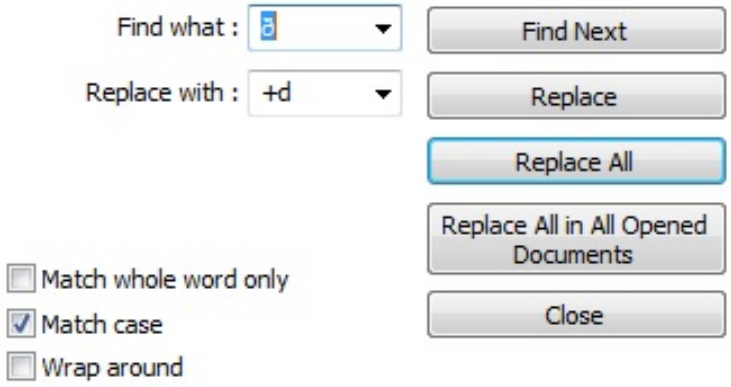
Creation of a basic text file II

- Remove mark-ups, comments, footnotes, page & folio numbers, translations, critical apparatus etc. (UTF-8)
- Replace special characters, e.g. *þ* → *+t*, *ȝ* → *+g* etc.

  Open Bestiary2.txt
- Reformat the file such that there is only one word per line

  Open Bestiary3.txt

Replacement of ð in Notepad++



Part-of-Speech Annotation

- POS-annotation with *Tree Tagger* (Schmid 1995)

Example output of TreeTagger

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Part-of-Speech Annotation

- POS-annotation with *Tree Tagger* (Schmid 1995)
- takes a word and assigns to it a POS-tag and a lemma

Example output of TreeTagger

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Part-of-Speech Annotation

- POS-annotation with *Tree Tagger* (Schmid 1995)
- takes a word and assigns to it a POS-tag and a lemma
- GeCMEP does not include lemmas; lemmas not used

Example output of TreeTagger

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Part-of-Speech Annotation

- POS-annotation with *Tree Tagger* (Schmid 1995)
- takes a word and assigns to it a POS-tag and a lemma
- GeCMEP does not include lemmas; lemmas not used
- supervised tagger; training lexicon and training input taken from PPCME2 prose texts

Example output of TreeTagger

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Part-of-Speech Annotation

- POS-annotation with *Tree Tagger* (Schmid 1995)
- takes a word and assigns to it a POS-tag and a lemma
- GeCMEP does not include lemmas; lemmas not used
- supervised tagger; training lexicon and training input taken from PPCME2 prose texts
- accuracy: c. 85% of all tags are assigned correctly

Example output of TreeTagger

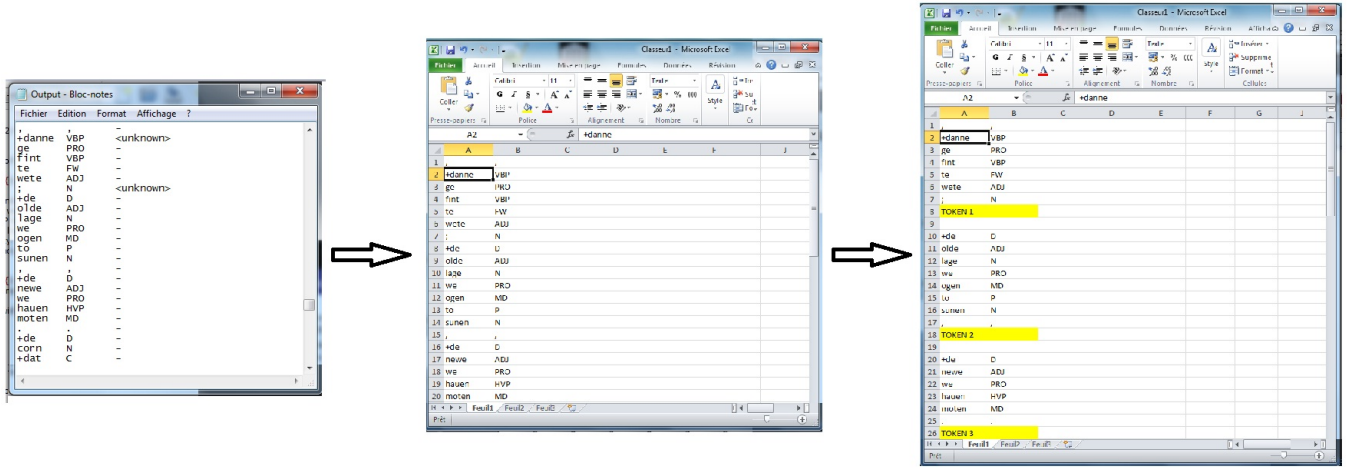
word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Part-of-Speech Annotation

Demonstration

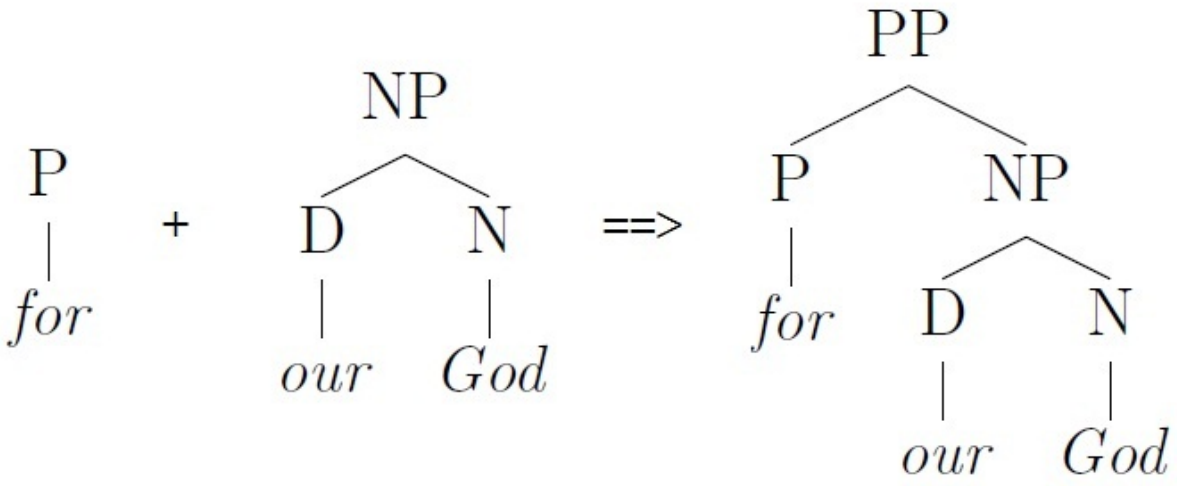
Tokenization

- sentence boundaries and IDs inserted manually in a spreadsheet



Shallow parsing procedure

- shallow parsing builds simple syntactic structures with regular expressions (Abney 1991)
- e.g. prepositional phrases can be build with an instruction like "whenever there is a P immediately before an NP, then bracket them together into a PP"



CorpusSearch revision queries

- tokens are chunked with revision queries of CorpusSearch 2 (Randall 2004)

Example revision query

```
define: cs.def
node: IP*|CP*
copy_corpus: t
query: (IP*|CP* idoms {1}P)
AND (IP*|CP* idoms {2}NP)
AND (P iprecedes NP)

add_internal_node{1,2}: PP
```

- windows bat file runs very many of such revision queries; output of one query feeds into the next
- simple python script converts tokens into right input format
- then manual correction until all tokens correspond to PPCME2 guidelines

CorpusSearch revision queries

Demonstration

Online Documentation for the GeCMEP

- Currently the database includes information on 15 parsed text files and a general bibliography

Online Documentation for the GeCMEP

- Currently the database includes information on 15 parsed text files and a general bibliography
- Text information specifies factors that might determine syntactic variation: date of composition, date of manuscript, dialect, versification, literary subjects

Online Documentation for the GeCMEP

- Currently the database includes information on 15 parsed text files and a general bibliography
- Text information specifies factors that might determine syntactic variation: date of composition, date of manuscript, dialect, versification, literary subjects
- In addition, cross-references to the three standard catalogues of ME (verse) texts: IMEV, Wells and MEC.

Online Documentation for the GeCMEP

- Currently the database includes information on 15 parsed text files and a general bibliography
- Text information specifies factors that might determine syntactic variation: date of composition, date of manuscript, dialect, versification, literary subjects
- In addition, cross-references to the three standard catalogues of ME (verse) texts: IMEV, Wells and MEC.

Go to the GeCMEP Database

Searching the corpus

- the corpus can be searched with query files of CorpusSearch 2 (Randall 2004)

Searching the corpus

- the corpus can be searched with query files of CorpusSearch 2 (Randall 2004)
- all functional labels, POS-tags and specific spellings can be searched for

Searching the corpus

- the corpus can be searched with query files of CorpusSearch 2 (Randall 2004)
- all functional labels, POS-tags and specific spellings can be searched for
- complex relations between elements can be specified (relative order of elements, number of words, identical indices, ...)

Realization of *verb-pronoun* structures

(2) a. þe þurst him dede more wo
the thirst him did more woe

þen hevede raþer his hounger do.
than had earlier his hunger done

'The thirst caused him more misery than his hounger had done before' (FoxWolf,56.273.68)

b. þe wox bicharde him, mid iwisse,
the fox deceived him, with certainity,

For he ne fond nones kunnes blisse
for he not found none kind's bliss

'The fox had deceived him indeed, because he didn't find any kind of bliss' (FoxWolf,224.278.295)

Search queries for *verb-pronoun* structures

- simple CorpusSearch query files to find such constructions:

opro-V.q

```
define: cs.def
node: IP*
query: (IP* idoms NP-OB*)
AND (NP-OB* idomsonly PRO)
AND (IP* idoms
VBP|VBD|DOP|DOD|HVP|HVD)
AND (NP-OB* precedes
VBP|VBD|DOP|DOD|HVP|HVD)
```

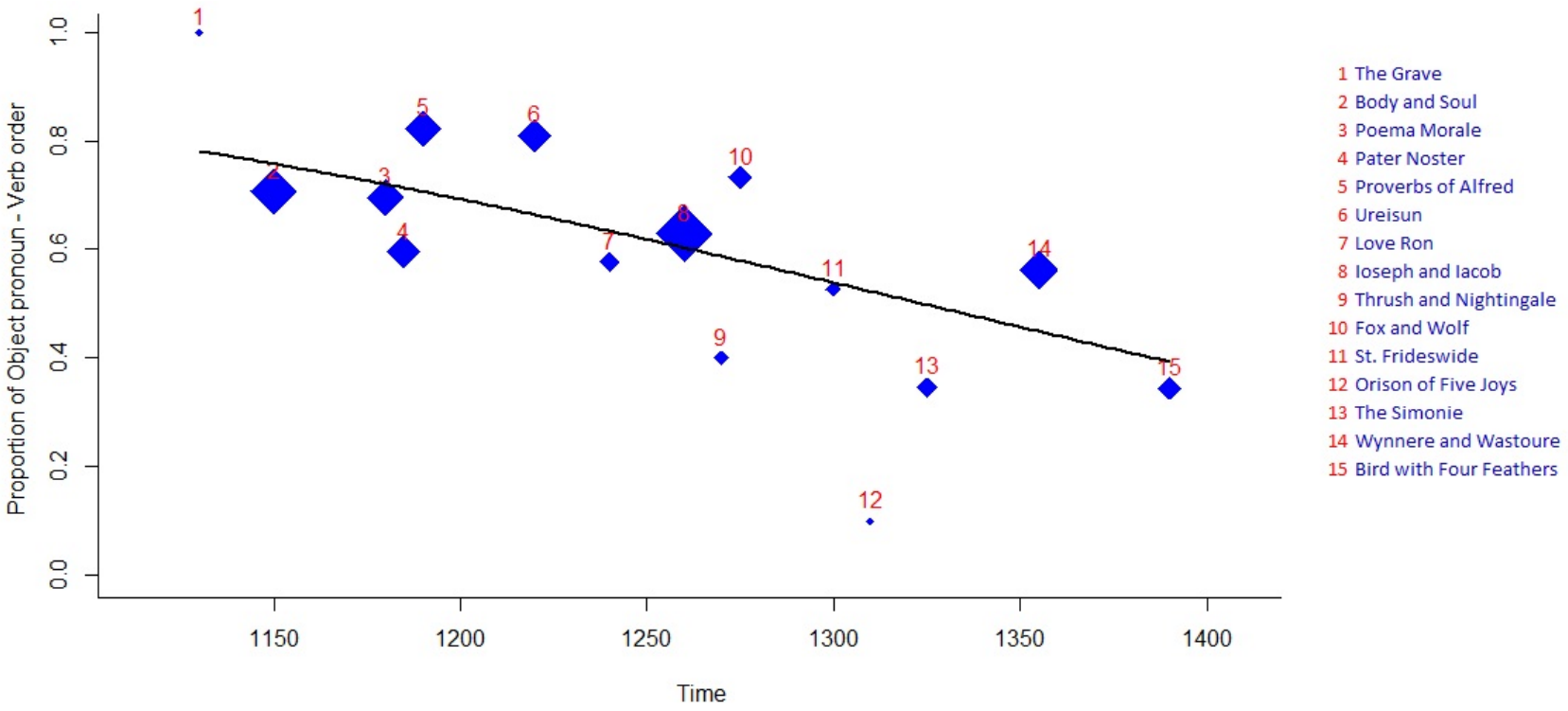
V-opro.q

```
define: cs.def
node: IP*
query: (IP* idoms NP-OB*)
AND (NP-OB* idomsonly PRO)
AND (IP* idoms
VBP|VBD|DOP|DOD|HVP|HVD)
AND
(VBP|VBD|DOP|DOD|HVP|HVD
precedes NP-OB*)
```

Search queries for *verb-pronoun* structures

Demonstration

Development of *verb-pronoun* structures



→ Significant decline of *object pronoun* - *verb* orders (c. 80% to c. 40%) measurable in Middle English verse 1150-1400

Benefits

- Collecting data from parsed corpora with automated search queries ...
  - ... saves a lot of time. Going through tens of thousands of words manually takes weeks or even months - with parsed corpora it's a matter of seconds.

Benefits

- Collecting data from parsed corpora with automated search queries ...
  - ... saves a lot of time. Going through tens of thousands of words manually takes weeks or even months - with parsed corpora it's a matter of seconds.
  - ... causes fewer mistakes. Humans can easily overlook examples or tally them up wrong - computers count correctly.

Benefits

- Collecting data from parsed corpora with automated search queries ...
  - ... saves a lot of time. Going through tens of thousands of words manually takes weeks or even months - with parsed corpora it's a matter of seconds.
  - ... causes fewer mistakes. Humans can easily overlook examples or tally them up wrong - computers count correctly.
  - ... assures replicability. Researchers can quickly double-check quantitative claims in the literature.

Benefits

- Collecting data from parsed corpora with automated search queries ...
  - ... saves a lot of time. Going through tens of thousands of words manually takes weeks or even months - with parsed corpora it's a matter of seconds.
  - ... causes fewer mistakes. Humans can easily overlook examples or tally them up wrong - computers count correctly.
  - ... assures replicability. Researchers can quickly double-check quantitative claims in the literature.
  - ... increases objectivity. Search criteria must be made explicit in the query files hence the ex-/inclusion of a particular sentence is independent of the individual researcher.

Benefits

- Collecting data from parsed corpora with automated search queries ...
  - ... saves a lot of time. Going through tens of thousands of words manually takes weeks or even months - with parsed corpora it's a matter of seconds.
  - ... causes fewer mistakes. Humans can easily overlook examples or tally them up wrong - computers count correctly.
  - ... assures replicability. Researchers can quickly double-check quantitative claims in the literature.
  - ... increases objectivity. Search criteria must be made explicit in the query files hence the ex-/inclusion of a particular sentence is independent of the individual researcher.

$\rightarrow$ increased scientificity

Locative relatives: *there* → *where* ...

(3) a. his halie nome we nomen and beren.
his holy name we took and bore

In þe font þer we iclensed weren.
in the font where we cleansed were

'We took and bore his name in the font where we were cleansed.' (PatNost,36.59.70)

b. To Oxenford is messageris he sende, that hi soghte
to Oxford his messengers he sent, that they sought

This maide ware heo were ifounde and to him broghte.
this maid where she were found and to him brought

He sent his messengers to Oxford so that they might seek and find this maiden, where she was, and bring her to him.' (Fridesw,47.64)

and temporal subordinators: *then* → *when*

(4) a. Ac ure drihten eft of deaþe hem aræreþ,
but our Lord again of death them arises

So he alle men deþ, þonne domes dai cumeþ.
as he all men does when Doomsday comes

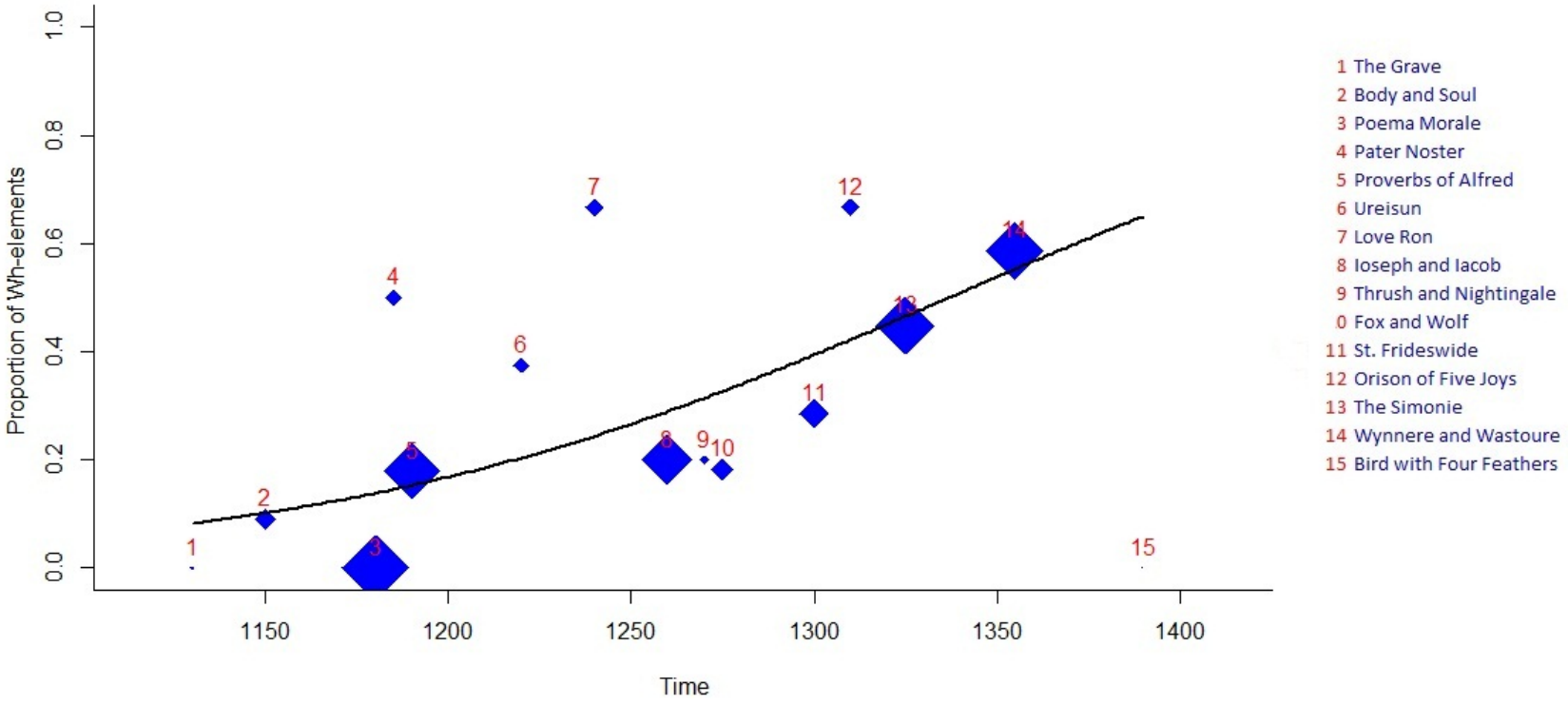
'But our Lord will raise them up from death, as he will all men, when Doomsday comes.'
(BodySoul,185.7.13.FragE)

b. and al bi-fuliþ he his frend
and all befouls he his friend

hwen he him vnfoldiþ.
when he him embraces

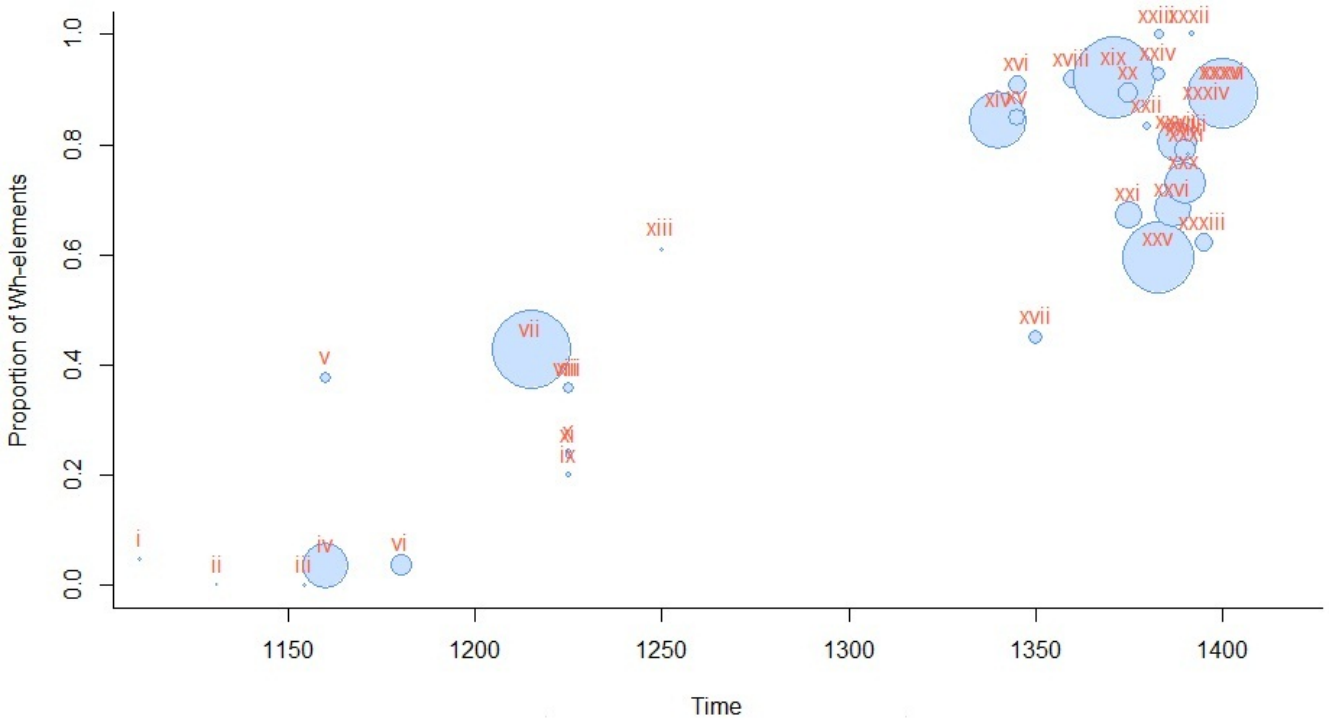
'He wholly befouls his friend when he embraces him.'
(ProvAlf,224.50.659.B32)

Development of *Th* and *Wh* elements



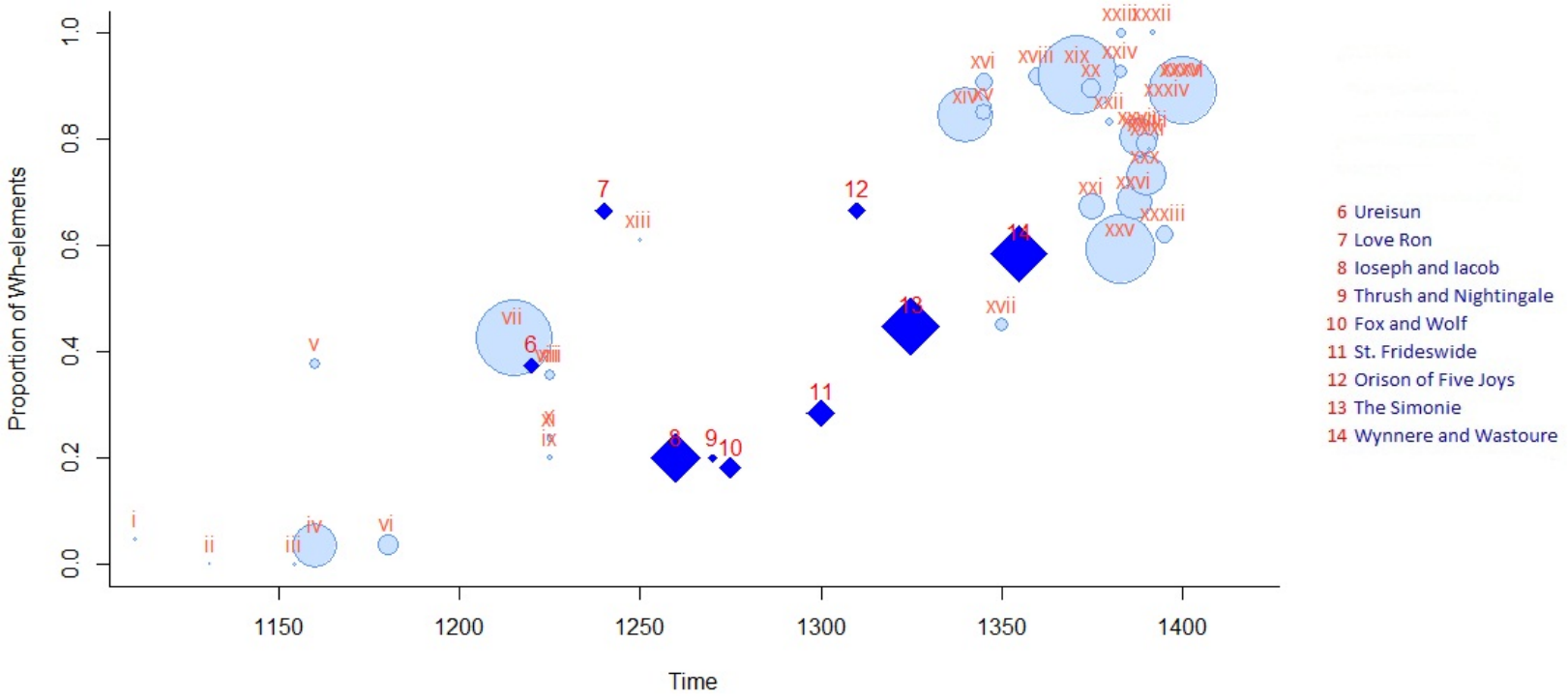
→ Significant increase of *Wh*- relative and adverbial clauses (c. 10% to c. 70%) in Middle English verse 1150-1400

Comparison verse - prose ...



i	In FestisSancti Marie	x	Saint Juliana	xix	Travels of Sir Mandeville	xxviii	Julian of Norwich's Revelations
ii	ASC Continuation1	xi	Saint Margaret	xx	Mirror of Edmund(Vernon)	xxix	Chaucer Tale of Melibee
iii	ASC Continuation 2	xii	Hali Meidhad	xxi	Rievaulx De Institutione	xxx	Chaucer Parson's Tale
iv	Trinity Homilies	xiii	Kentish Sermons	xxii	Chaucer Boethius Translation	xxxi	Chaucer Treatise Astrolabe
v	Lambeth Homilies	xiv	Ayenbite of Inwyt	xxiii	Wycliffe Old Testament	xxxii	Equatorie of the Planets
vi	Vices and Virtues	xv	Richard Rolle Treatises	xxiv	Wycliffe New Testament	xxxiii	Cloud of Unknowing
vii	Ancrene Riwle	xvi	Richard Rolle Epistles	xxv	Wycliffe Sermons	xxxiv	Hilton's Eight Chapters on Perfection
viii	Sawles Warde	xvii	Earliest Prose Psalter	xxvi	John Trevisia's Polychronicon	xxxv	Book of Vices and Virtues
ix	Saint Katherine	xviii	Mirror of Edmund (Thornton)	xxvii	Purvey's Prologue to Wycliffe	xxxvi	The Chronicles of England

... shows that verse can close the gap in prose texts.



i	In FestisSancti Marie	x	Saint Juliana	xix	Travels of Sir Mandeville	xxviii	Julian of Norwich's Revelations
ii	ASC Continuation1	xi	Saint Margaret	xx	Mirror of Edmund(Vernon)	xxix	Chaucer Tale of Melibee
iii	ASC Continuation 2	xii	Hali Meidhad	xxi	Rievaulx De Institutione	xxx	Chaucer Parson's Tale
iv	Trinity Homilies	xiii	Kentish Sermons	xxii	Chaucer Boethius Translation	xxxi	Chaucer Treatise Astrolabe
v	Lambeth Homilies	xiv	Ayenbite of Inwyt	xxiii	Wycliffe Old Testament	xxxii	Equatorie of the Planets
vi	Vices and Virtues	xv	Richard Rolle Treatises	xxiv	Wycliffe New Testament	xxxiii	Cloud of Unknowing
vii	Ancrene Riwle	xvi	Richard Rolle Epistles	xxv	Wycliffe Sermons	xxxiv	Hilton's Eight Chapters on Perfection
viii	Sawles Warde	xvii	Earliest Prose Psalter	xxvi	John Trevisia's Polychronicon	xxxv	Book of Vices and Virtues
ix	Saint Katherine	xviii	Mirror of Edmund (Thornton)	xxvii	Purvey's Prologue to Wycliffe	xxxvi	The Chronicles of England

```
( (IP-MAT-SPE (NP-SBJ *pro*)
              (VBP Thank)
              (NP-OB2 (PRO you))
              (PP (P for)
                  (NP (PRO$ your) (N attention)))
              (. !))
  (ID EndOfTalk))
```

Special thanks to Beatrice Santorini for running her queries on the GeCMEP files to find annotation mistakes. Thanks are also due to Benjamin Börschinger and Paola Merlo for useful advice on chunking.

Abney, S. (1991), Parsing by chunks, *in* R. Berwick and C. Tenny, eds, 'Principle-Based Parsing', Kluwer Academic Publishers, Dordrecht, pp. 257–278.

Kroch, A. and Taylor, A. (2000), *Penn-Helsinki Parsed Corpus of Middle English*, http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3 (Accessed 10 April 2013), 2 edn, Department of Linguistics, University of Pennsylvania.

Randall, B. (2004), *CorpusSearch 2*, http://corpussearch.sourceforge.net/ (Accessed 7 November 2013), Sourceforge.net.

Schmid, H. (1995), 'Improvements in part-of-speech tagging with an application to german', *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland* pp. 47–50.