Comparing the Evolution of V2 in English and French

Anthony Kroch and Beatrice Santorini
February 2012

www.ling.upenn.edu/~kroch/handouts/thurs-thoughts.pdf

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

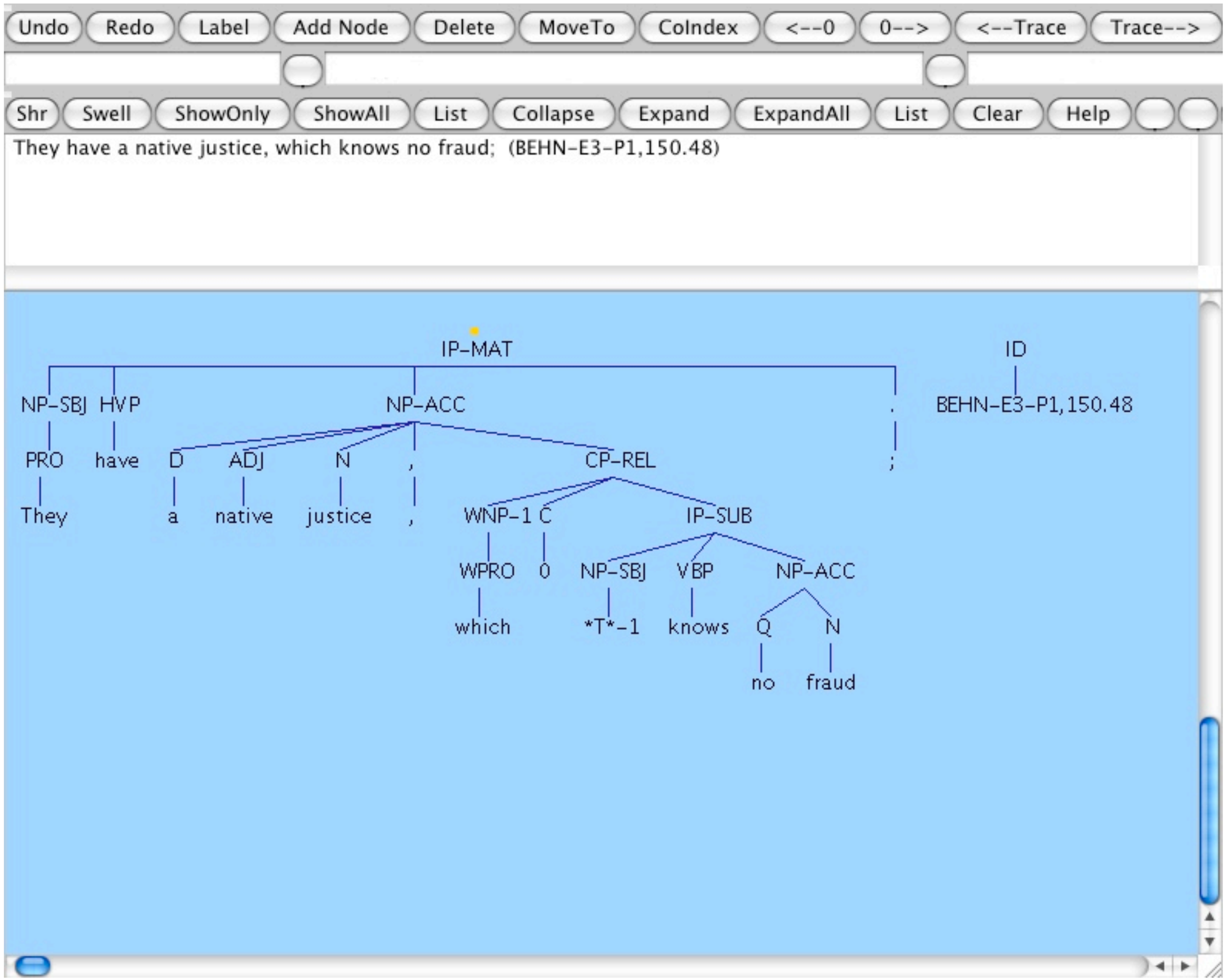
  grammatical function of phrases/clauses

- 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-ACC (D a)
                 (ADJ native)
                 (N justice)
                 (, ,)
                 (CP-REL (WNP-1 (WPRO which))
                         (C 0)
                         (IP-SUB  (NP-SBJ *T*-1)
                                  (VBP knows)
                                  (NP-ACC (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
NP-SBJ
HVP
NP-ACC
.
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
ID
BEHN-E3-P1,150.48

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.

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.

  1.0 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

Other languages

- 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/)

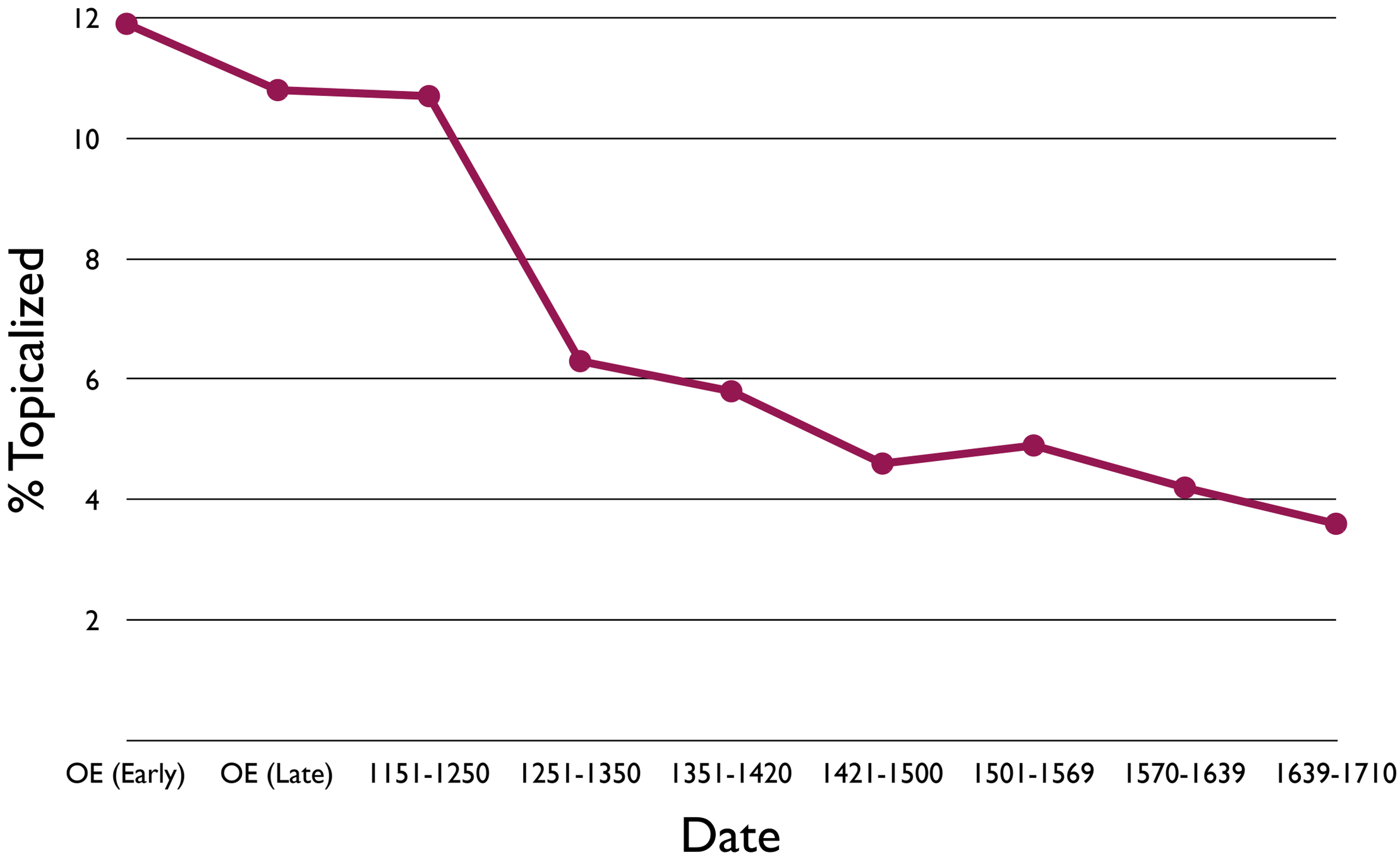
  ≈1 million words

- 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, .5 million parsed to date

The loss of verb-second word order and the decline of topicalization in English

Decline of direct object topicalization in English

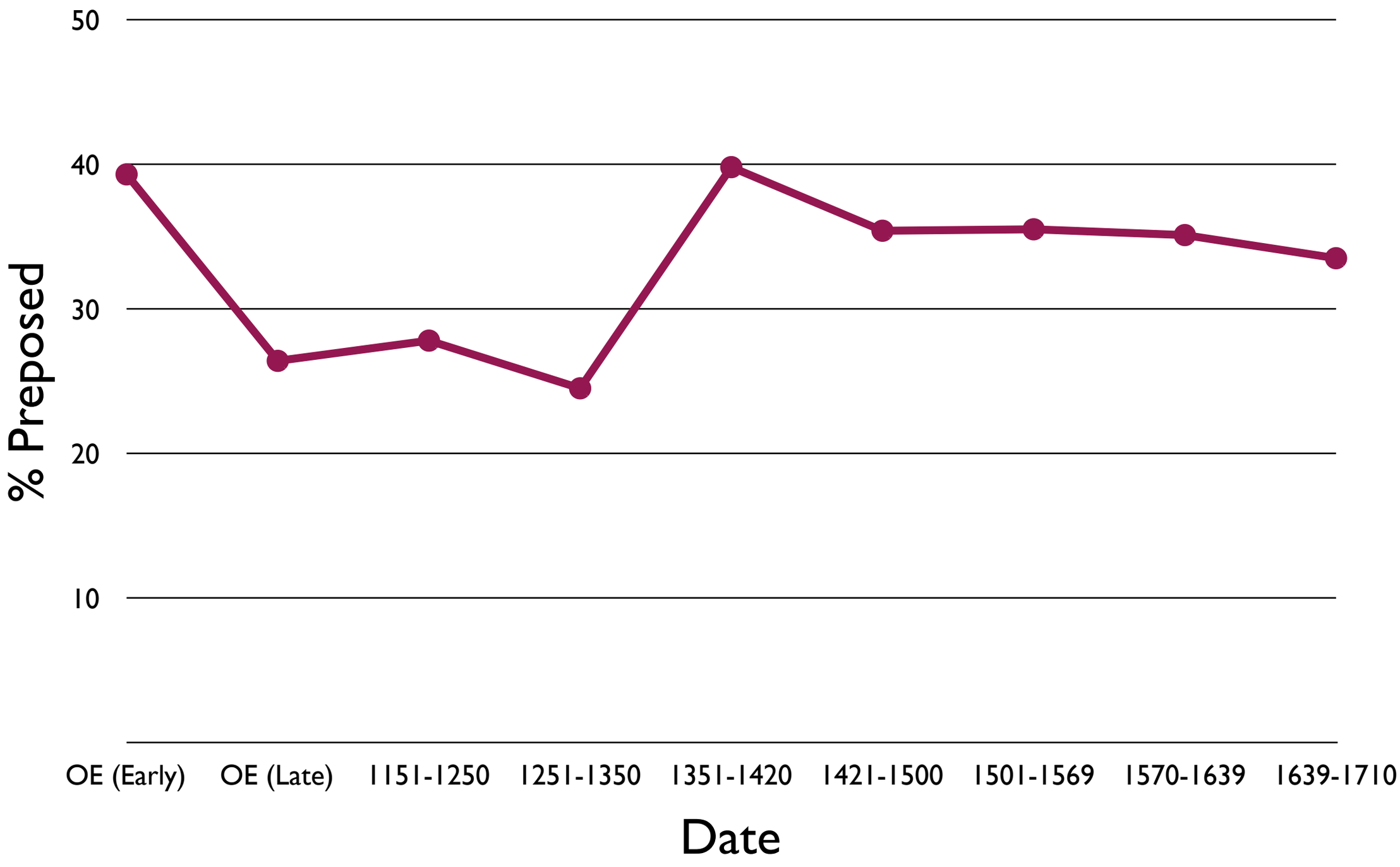


Frequency of direct object topicalization in modern spoken Dutch (Bouma 2008)

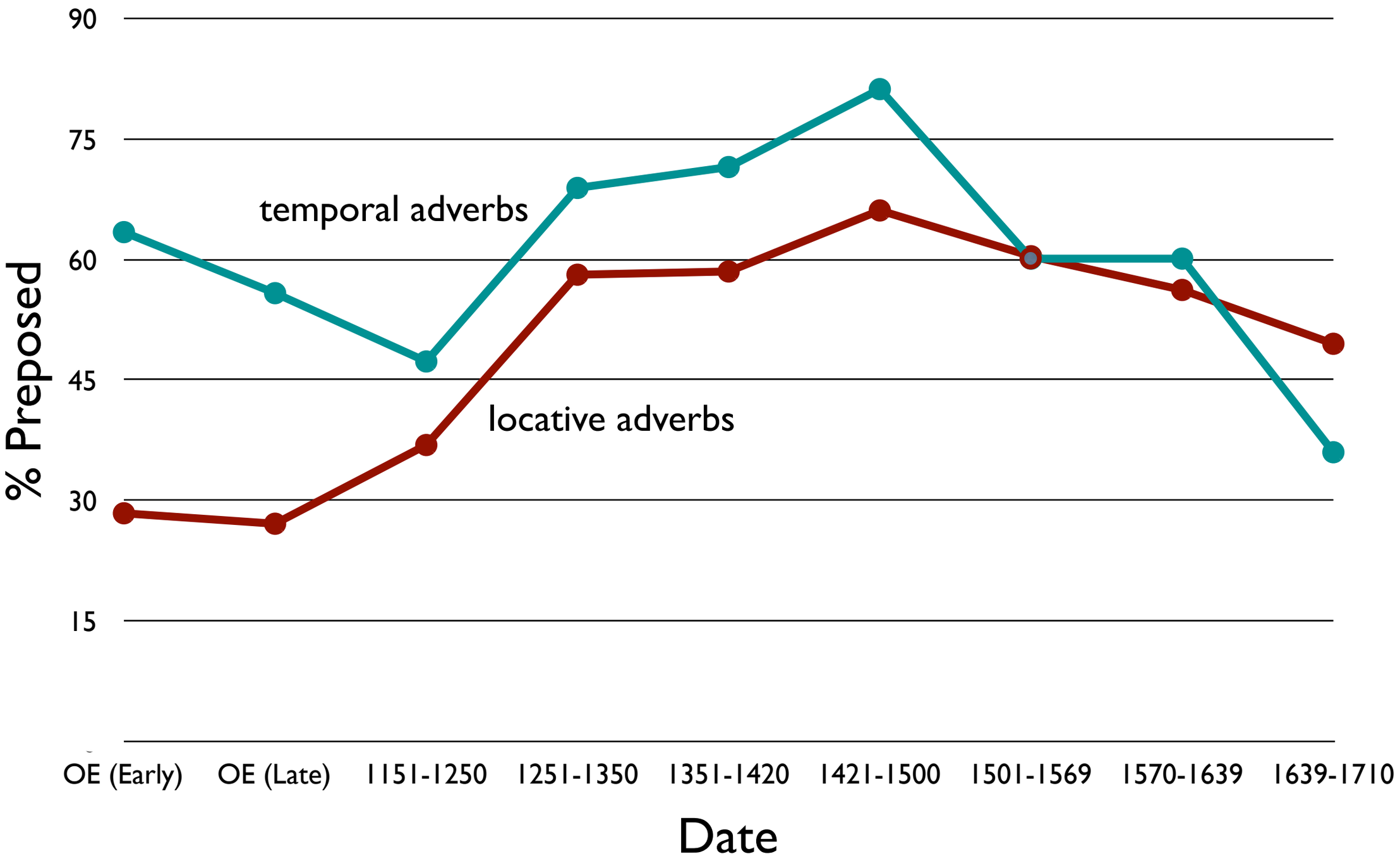
Table 4.2: Summary of Vorfeld occupation of arguments.

	Vorfeld		Prop est (%)
Argument	yes	no	pt
subject	43 523	18 597	**70.1**
direct object	3 418	20 432	**14.3**
indirect object	38	815	**4.5**

Evolution of PP preposing in English



Evolution of adverb fronting in English



The history of topicalization in English (Speyer 2008)

- Why does topicalization decline in Middle English but not disappear? If the change is parametric, it should go to completion. Otherwise, topicalization, a clear case of stylistic variation, might be expected to be stable in frequency over time.

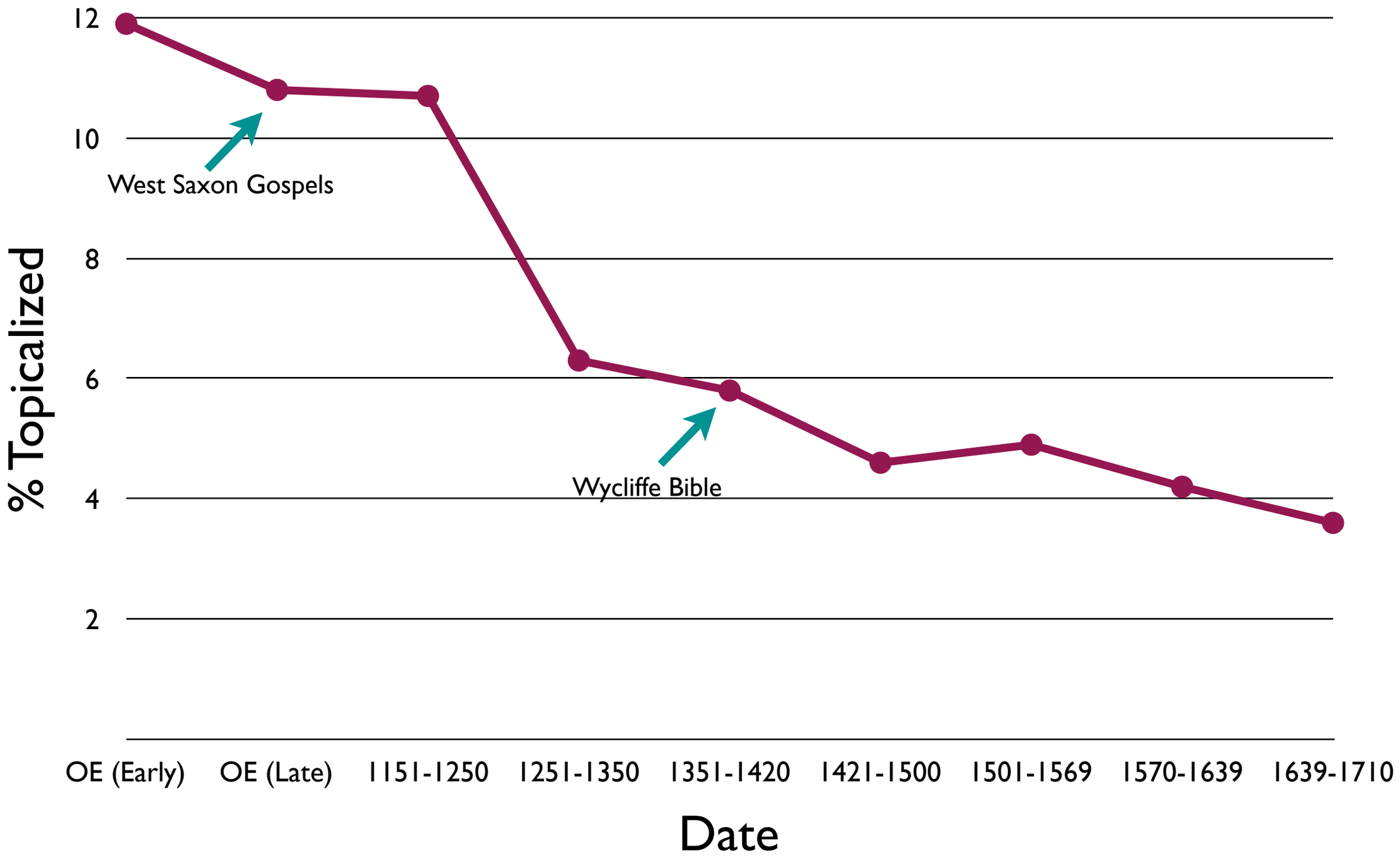
- This question finds an answer in the specific interaction between parametric settings and stylistic variation in the history of English.

An illustrative case in the New Testament
Matthew 13.28

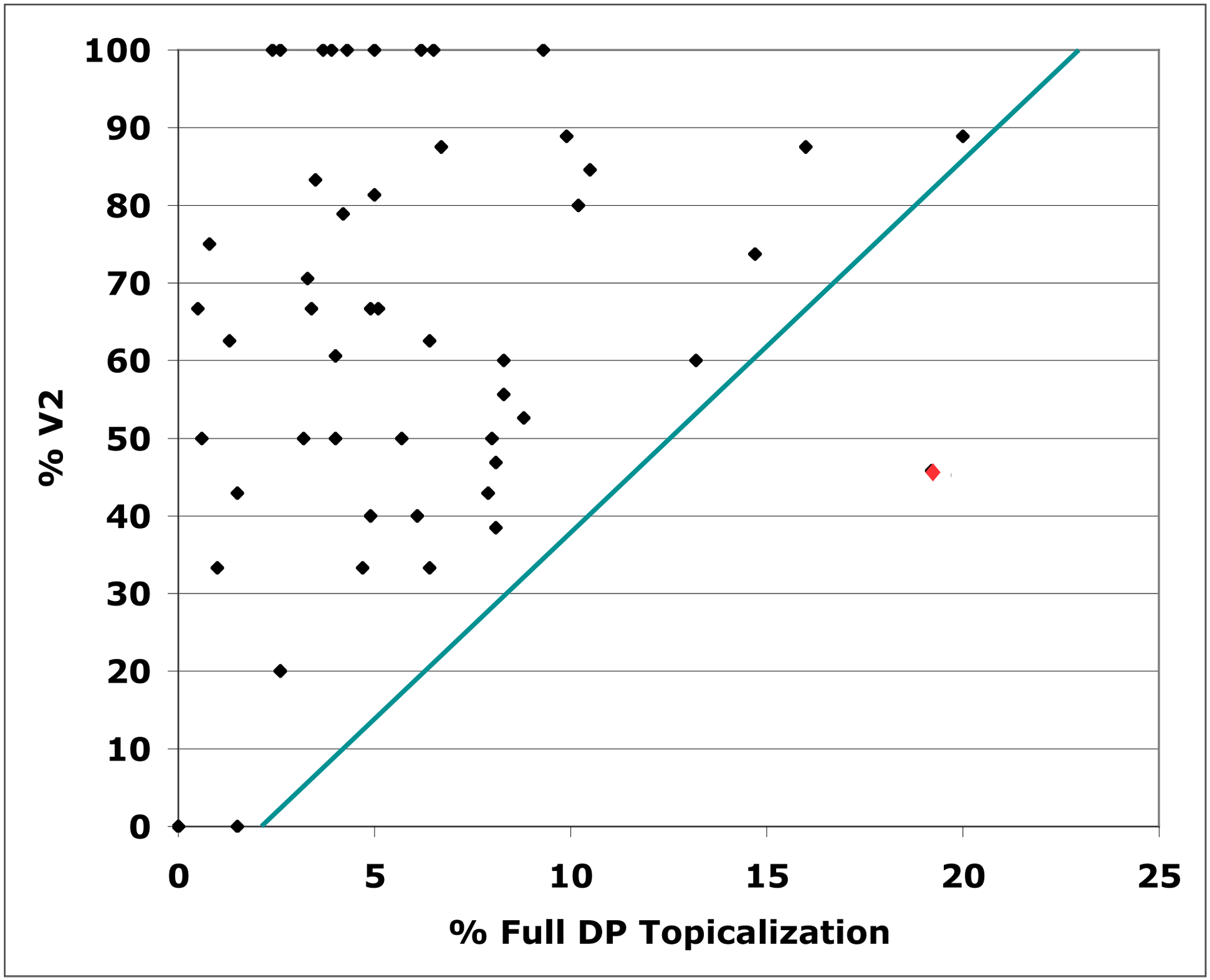
(1) Þā cwæþ hē, Þæt dyde unhold mann.
West Saxon Gospels ca. 1000

(2) And he seide to hem, An enemy hath do˜ this thing.
Wycliffe Bible ca. 1380

Decline of direct object topicalization in English



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



Distribution of subject types in a corpus of topicalized and non-topicalized sentences in natural speech

personal pronoun	demonstrative pronoun	full noun phrase
140	20	142
46.4	**6.6**	**47.0**

Subject type in sentences with *in situ* objects

personal pronoun	demonstrative pronoun	full noun phrase
181	2	17
90.5%	**1%**	**8.5 %**

Subject type in sentences with topicalized objects

Clash avoidance

- The type of topicalization that declines:

(1) The nèwspaper Jóhn read; the nòvel Máry did.

(Compare: The nèwspaper read Jóhn.)

- The type of topicalization that doesn't:

(2) The nèwspaper I réad; the nòvel I dídn't.

Translating German topicalized arguments into English in three modern German novels [by Böll, Dürrenmatt and Grass]

Topicalized to topicalized:

G: Mahlkes Haupt bedeckte dieser Hut besonders peinlich.

E: On Mahlke's head this hat made a particularly painful impression.

Topicalized to non-topicalized:

G: Zu den sechs kamen noch drei weitere.

E: Three others joined these six in the afternoon.

Accent placement and topicalization frequencies in translating German topicalized arguments into English

	focus accent on the German subject	accent elsewhere
topicalization in the English translation	0	31
no topicalization in the English	25	100

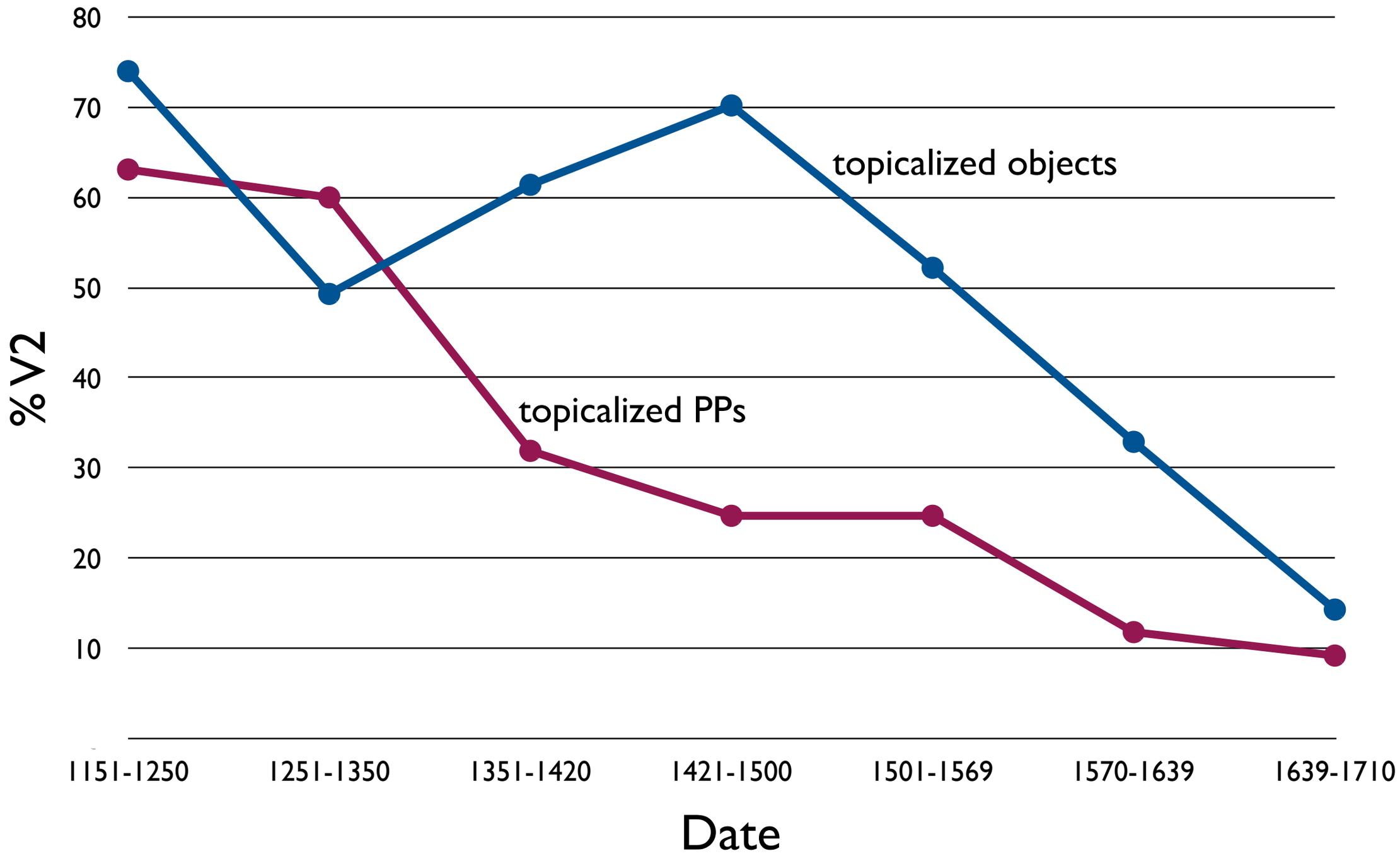
Distribution of contrastive topicalization by focus (second accent) placement in Middle English

focus / position / distribution of cases	focus on subject	focus on tensed verb	focus elsewhere
N (total= 207)	113	29	65
% inversion	89	14	71
% of cases	55	14	31

Frequency of matrix clause topicalization and V2 in Middle and Early Modern English

time / period / sentence type	me1	me2	me3	me4	eme1	eme2	eme3
# sent. with DO	2855	1300	4615	2271	3229	3584	2544
# topicalized	219	69	145	66	67	82	28
% topicalized	7.7	5.3	3.1	2.9	2.1	2.3	1.1
#V2	162	34	89	46	35	27	4
%V2	74	49.3	61.4	70.2	52.2	32.9	14.3

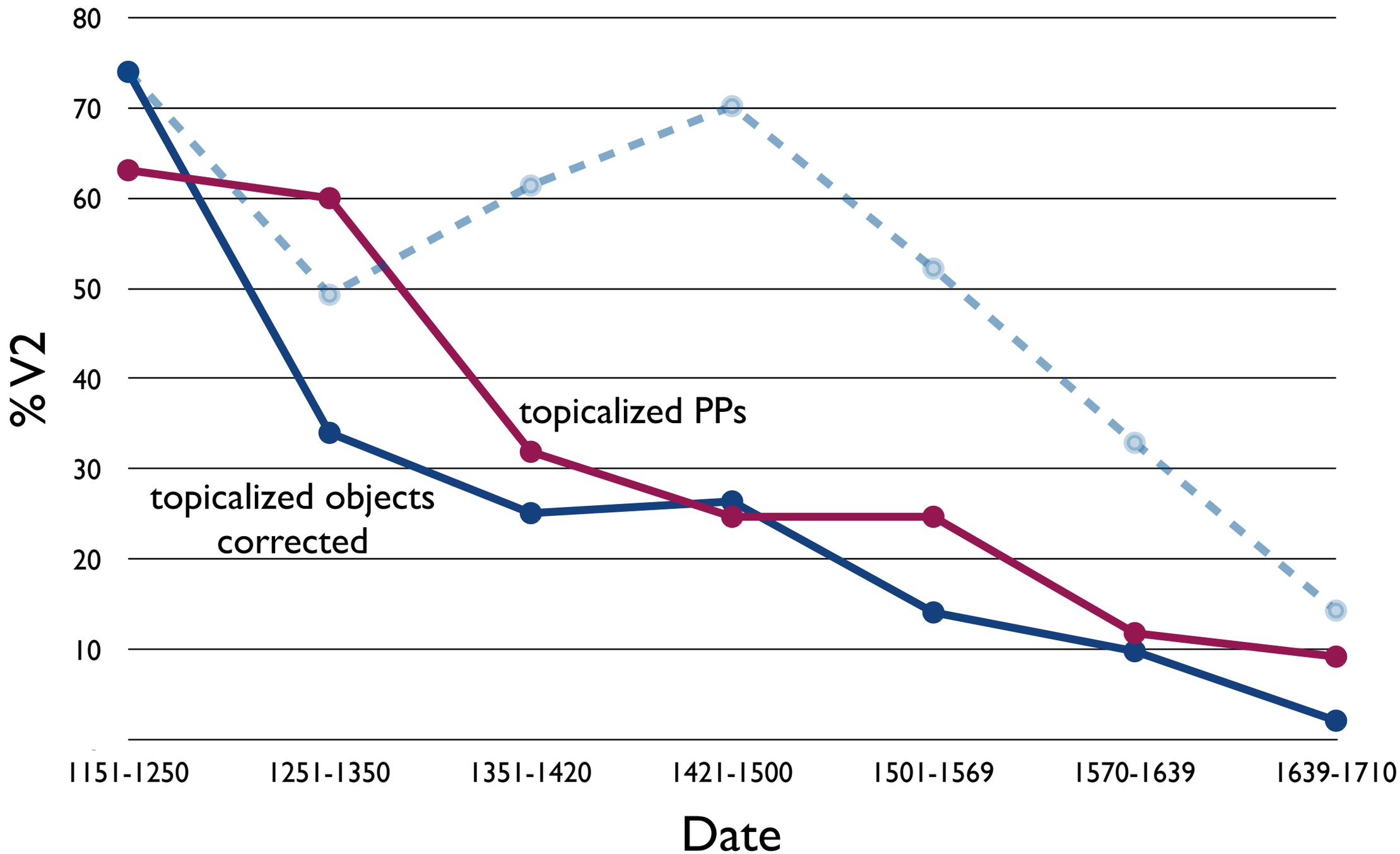
Rate of V2 loss in English with topicalized objects and PPs



Corrected frequency of matrix clause topicalization and V2 in Middle and Early Modern English

time period / sentence type	me1	me2	me3	me4	eme1	eme2	eme3
# sent. with DO	2855	1300	4615	2271	3229	3584	2544
# topicalized	219	69	145	66	67	82	28
would have been topicalized	219	100	354	174	248	275	195
actual rate of V2	74	49.3	61.4	70.2	52.2	32.9	14.3
corrected rate V2	74.0	34.0	25.1	26.4	14.1	9.8	2.1

Rate of V2 loss in English corrected for clash avoidance



Was Old English a V2 language?

The V2 constraint in Old English: the pronoun exception

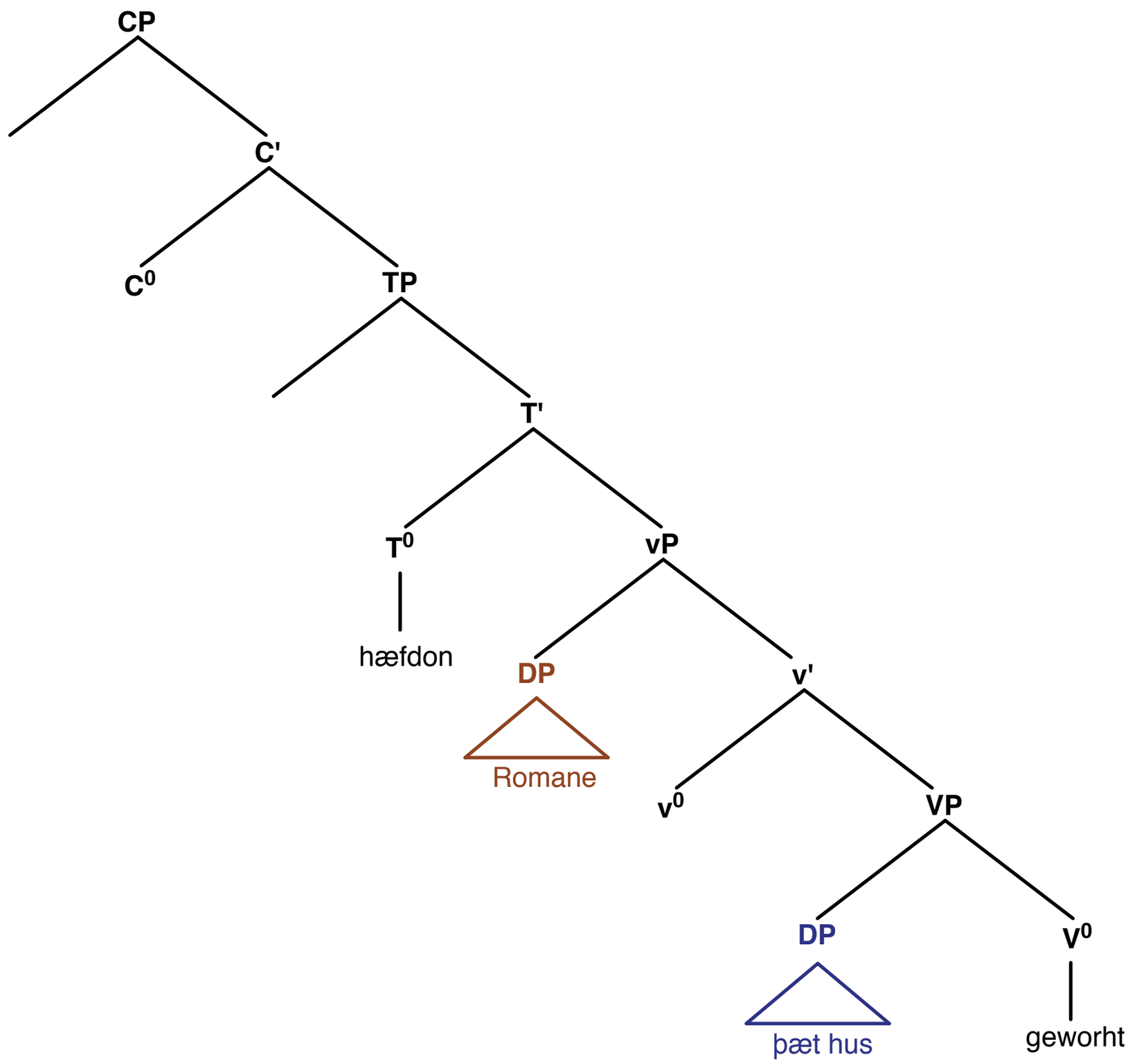
(1) Þæt hus hæfdon Romane to ðæm anum tacne geworht.

(2) Ælc yfel he mæg don.

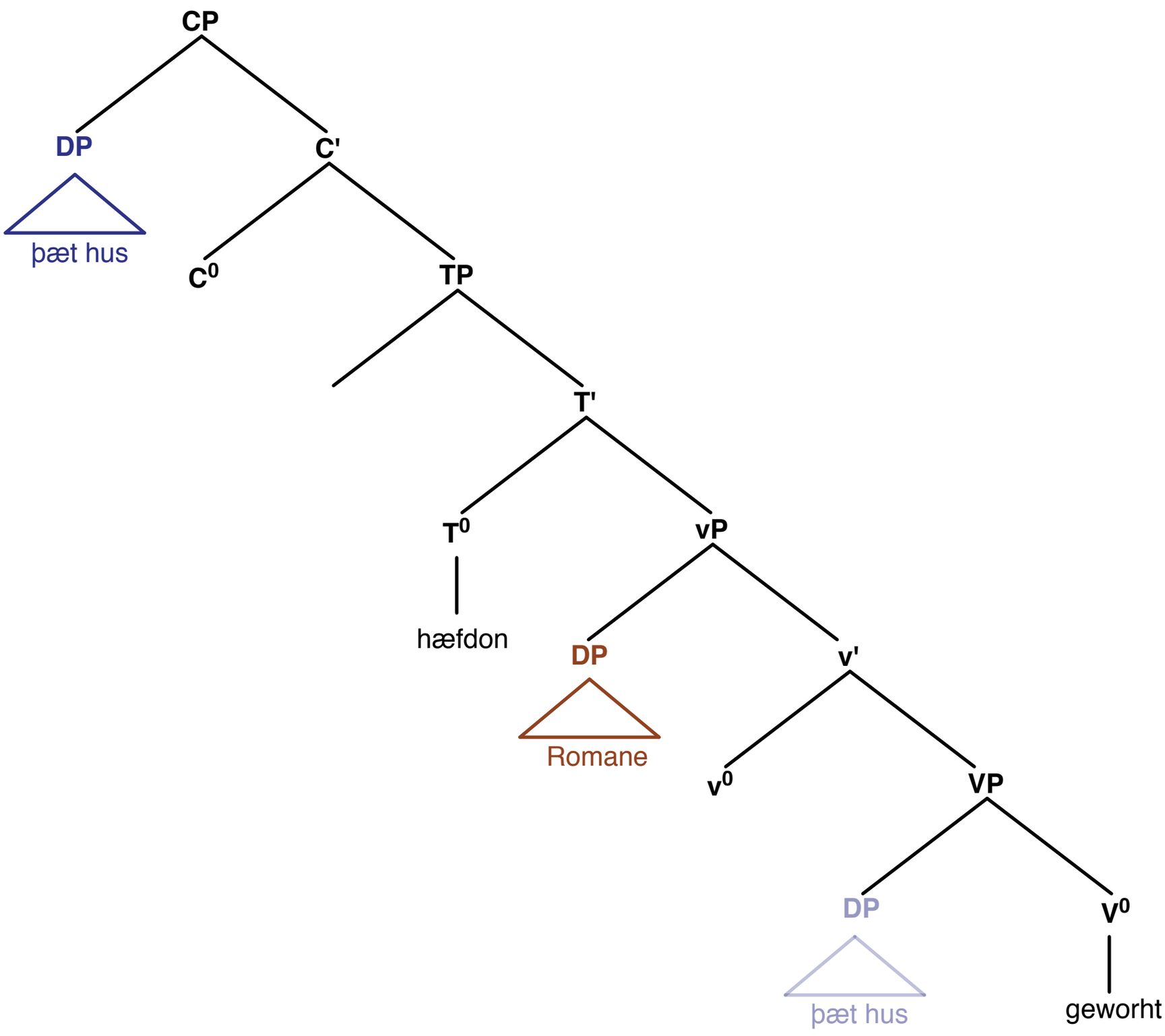
(3) Þin agen geleafa þe hæfþ gehæledne.

(4) & seofon ærendracan he him hæfde to asend.

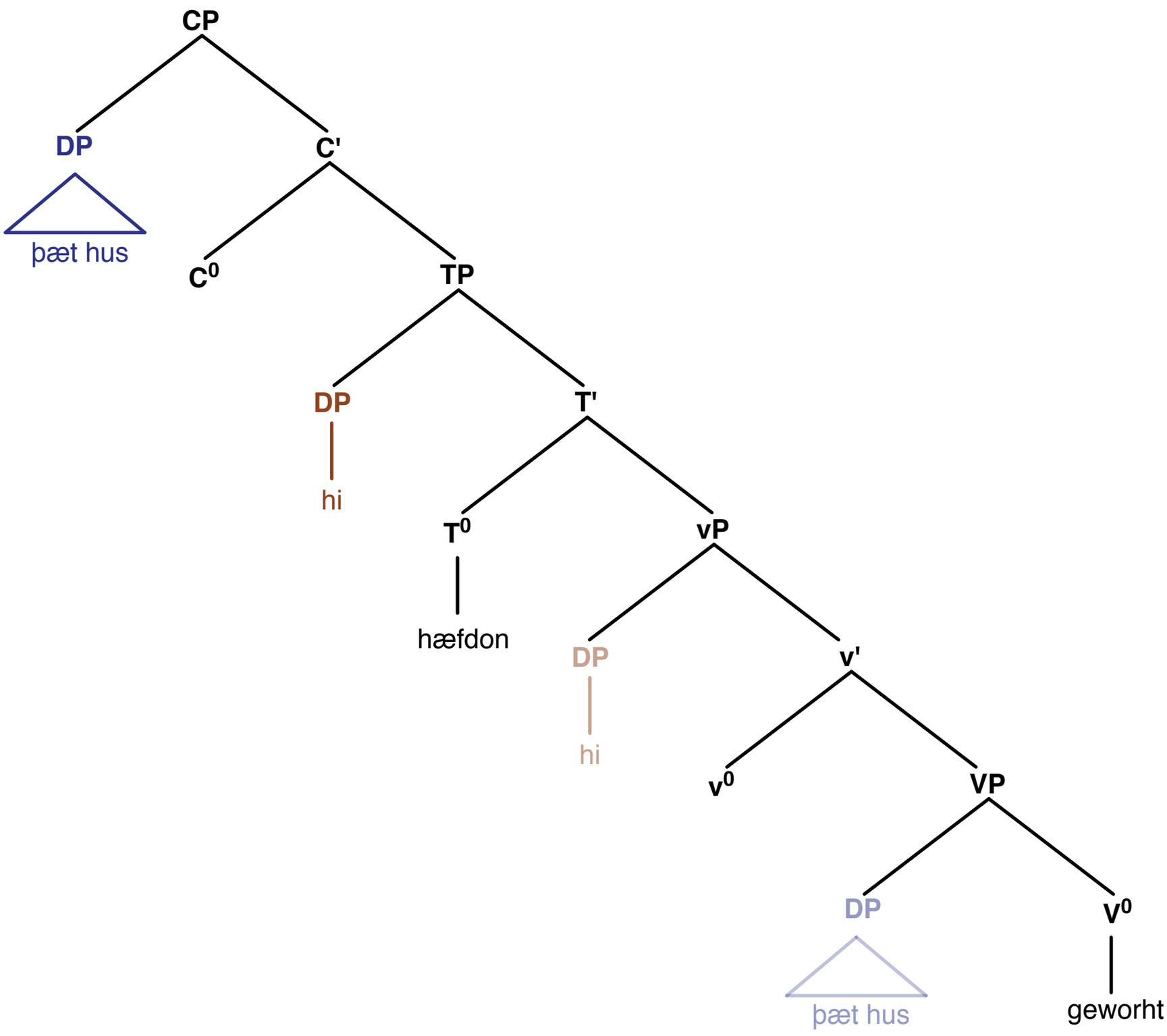
Phrase structure of an English V2 clause



Phrase structure of an English V2 clause



Phrase structure of an English V2 clause



Unambiguous V3 clauses with topicalized objects

(1) þæne se geatweard let in
that-one the doorkeeper let in
(cowsgosp, Jn_[WSCp]:10.3.6596)

(2) and him se innoþ eac geopenode ongean
and him the heart also opened again
(coælive, +ALS_[Vincent]:170.7907)

Frequency of unambiguous V3 clauses against all particle verb cases

	full DP subjects	pronoun subjects
V2 cases	74	6
V3 cases	20	45
frequency V3	0.21	0.88

V3 clauses with topicalized objects ambiguous due to West Germanic verb raising

(3) ac þone yfelan fæstrædan willan folneah nan wind ne mæg
but the evil constant will almost no storm not may
awecggean
awaken

(cocuraC,CP_[Cotton]:33.224.4.85f.)

(4) ac folneah nan wind þone yfelan fæstrædan willan awecggean ne mæg

(5) ac folneah nan wind þone yfelan fæstrædan willan ne mæg awecggean

(6) ac þone yfelan fæstrædan willan folneah nan wind ne mæg awecggean

Expected versus observed number of V3 clauses with topicalized objects given verb raising

number SOVI main clauses with full noun phrase subjects	20
ratio of SOVI to SOIV in unambiguous verb-raising environments	0.7
rate of object topicalization in verb-final clauses	0.2
predicted number of OSIV cases due to verb-raising with topicalization	2.8
actual number of OSIV cases	22

The evolution of word order in French

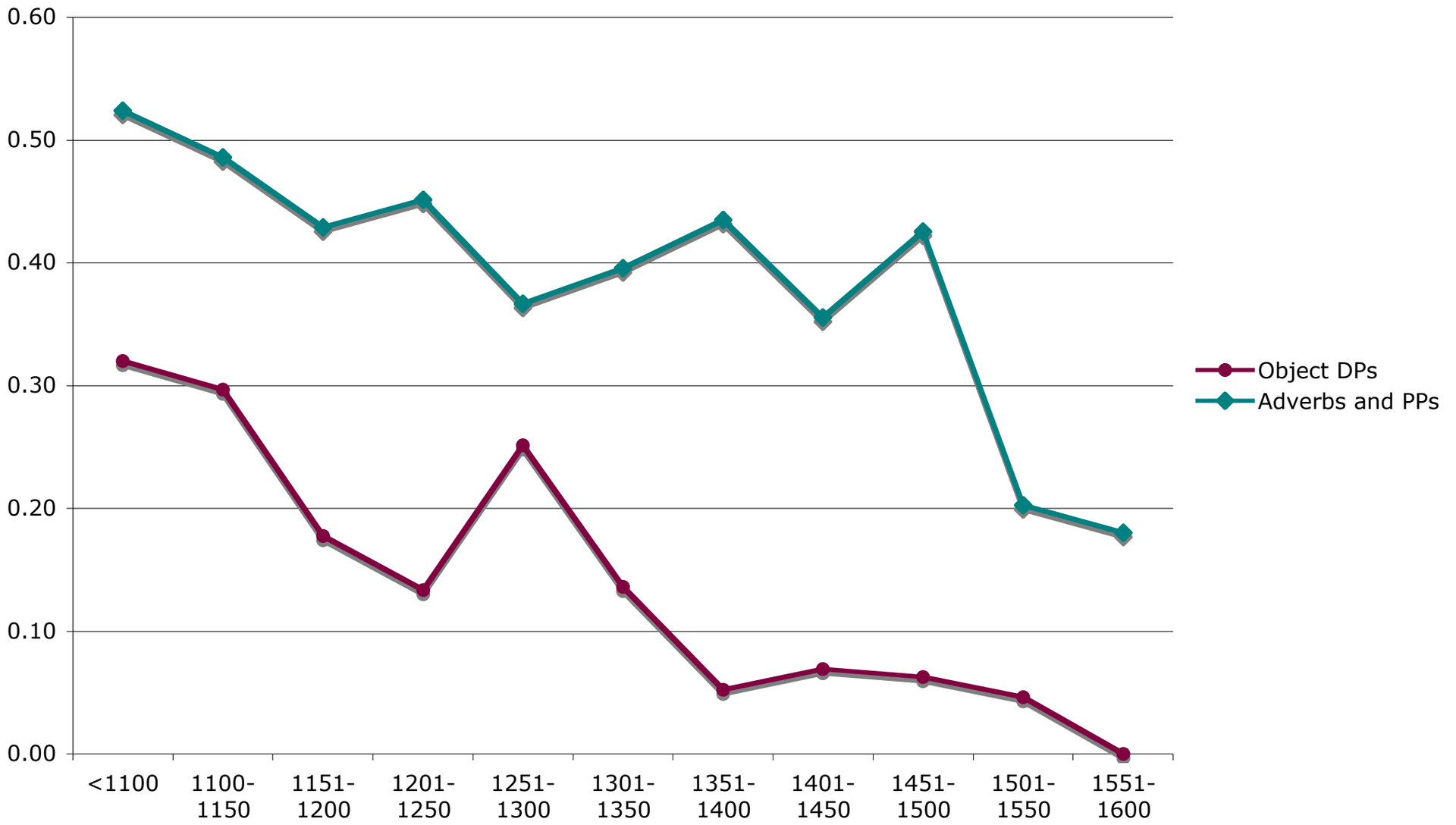
V2 in Old and Middle French

(1) l'estreu li tint sun uncle Guinemer
the stirrup him held his uncle Guinemer
Roland 27.329

(2) Espaigne vus durat il en fiet
Spain you will-give he in fief
Roland, 36.446

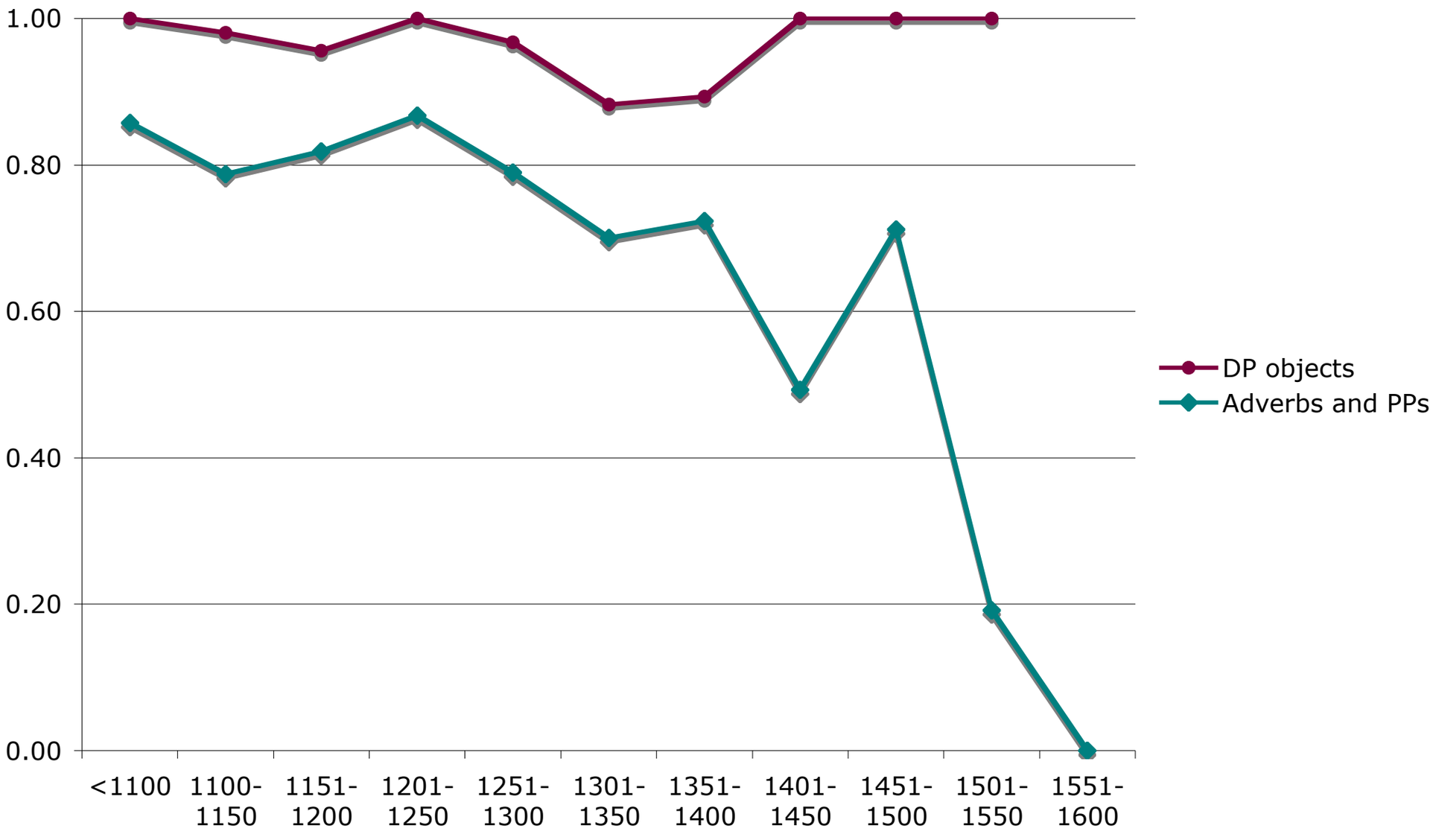
(3) or est ele bien venue
now is she welcome
Yvain 43.1440

Decline of XP fronting in French in sentences with overt subjects



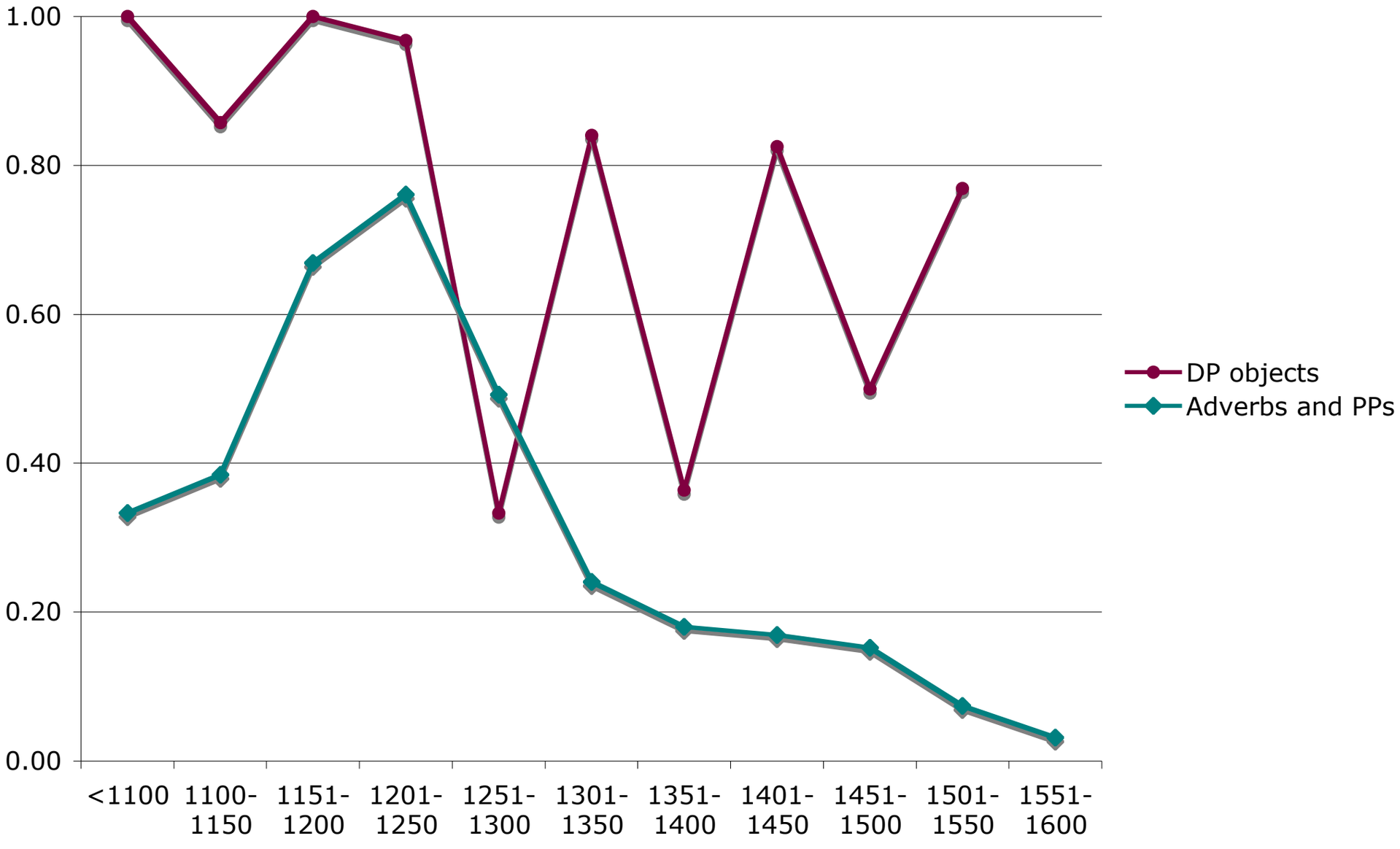
Old French — Middle French — Modern French

V2 frequency: sentences with full NP subjects



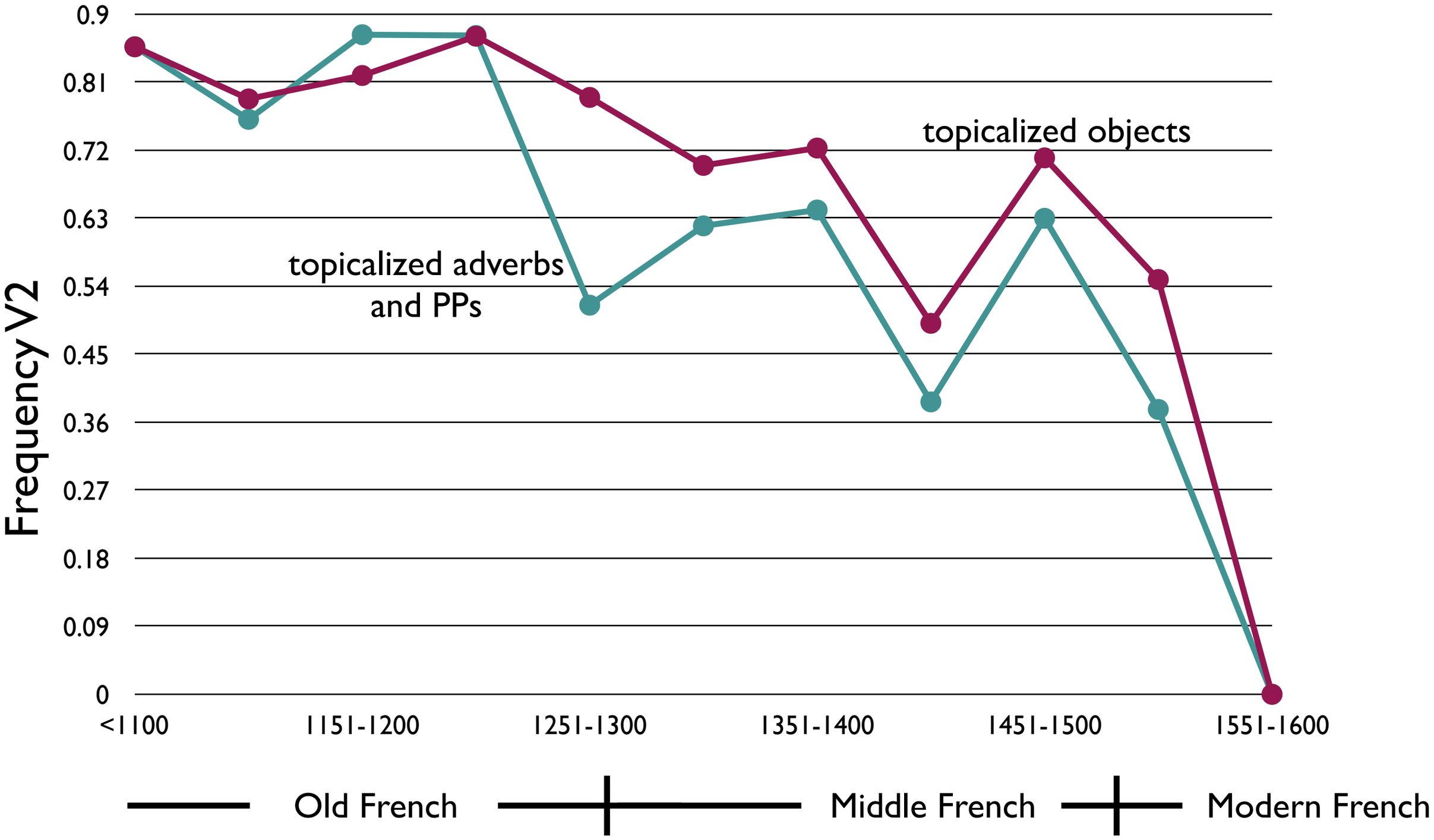
Old French — Middle French — Modern French

V2 frequency: sentences with overt pronoun subjects



Old French — Middle French — Modern French

Corrected evolution of V2 word order in French in sentences with full DP subjects



Germanic and Romance inversion in Old and Middle French

“Germanic” inversion in Old and Middle French

(1) messe e matines ad li reis escultet
mass and matins has the king heard
Roland 11.139

(2) chars avoient ils assés
meat had they enough
Froissart, 135.569

(3) une chose ont-ilz asez honneste
one thing have-they enough honest
Commynes, 120.1634

“Romance” inversion in Old French

(1) ... puis si chevalchet od sa grant ost li ber
then so rides with his great army the baron
Roland, 179.2438

(2) ... ço ad tut fait Rollant
that has all done Roland
Roland, 24.301

(3) ceste parole ot escoutee li seneschax
this speech has heard the seneschal
Yvain 134.4663

Ambiguous cases

(1) Après parlat ses filz envers Marsilies
then spoke his son to Marsilies
Roland 37.466

(2) Bien fiert nostre guarent
well fights our guardian
Roland 124.1665

(3) Mult fierement chevalchet li emperere
very proudly rides the emperor
Roland 23.3296

Temporal evolution of V2 with full DP subjects for all types of preposed XP

	sentences with an auxiliary verb	sentences with a single verb
Old French	0.86 [218]	0.83 [2163]
Middle French	0.69 [402]	0.70 [3633]
Modern French	0.27 [33]	0.22 [160]

Temporal evolution of Germanic and Romance inversion in V2 sentences with auxiliary verbs (sentences with topicalized XPs and full DP subjects)

	frequency of Germanic inversion	frequency of Romance inversion	Romance + Germanic inversion
Old French	0.50	0.36	0.86
Middle French	0.32	0.37	0.69
Modern	0.03	0.24	0.27

An independence result

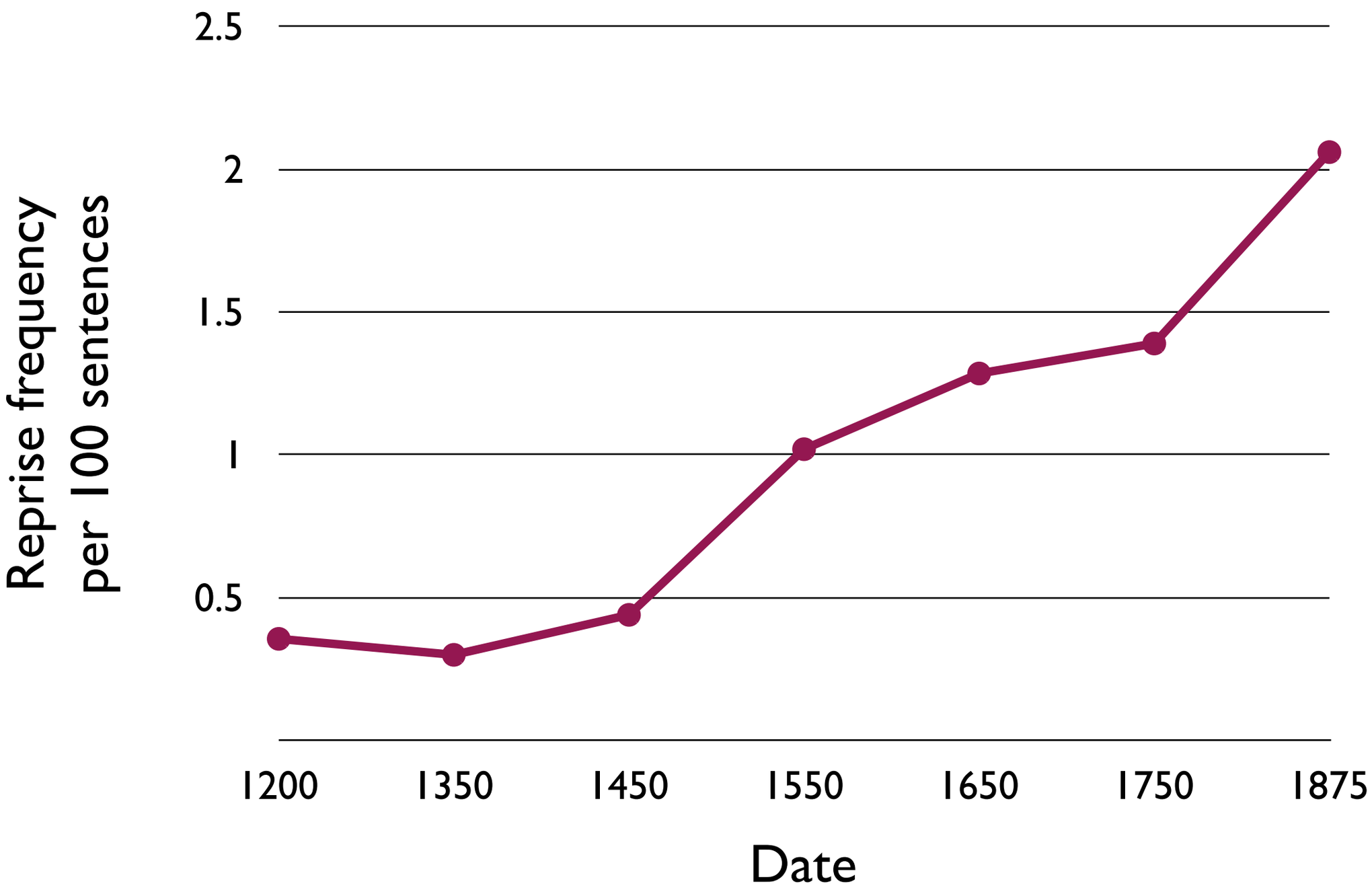
	Romance + Germanic inversion	sentences with a single verb
Old French	0.86	0.83
Middle French	0.69	0.70
Modern French	0.27	0.22

Frequency of *pro* subjects in sentences with fronted XPs

Old French — Middle French — Modern French

Why does French completely lose object topicalization?

Rise of clitic left-dislocation and loss of topicalization (Priestley 1955)



Modern French clitic left dislocation

(1) *Le Figaro*$_i$, Jean *(le$_i$) lit tous les jours.
The Figaro John it reads every day

(2) Ma femme$_i$, elle$_i$ travaille à la Bibliothèque Nationale.
My wife she works at the library national

Temporal evolution of subject and object left dislocation frequencies per thousand sentences

	frequency of subject left dislocation	frequency of object left dislocation	number of matrix clauses
Old French	2.6	2.2	12022
Middle French	3.8	1.8	24634
Early Modern	28	4.3	3514

Cleft sentences in Modern French

(1) C'est *Le Figaro*$_i$ que Jean lit t_i tous les jours.
It's *The Figaro* that John reads every day

(2) C'est ma femme$_i$ qui t_i travaille à la BN.
It's my wife that works at the BN

(3) Il y a un an$_i$ qu'elle travaille à la BN t_i.
It's one year that-she works at the BN

Temporal evolution of cleft sentence frequencies per thousand sentences

	frequency of temporal clefts	frequency of subject and object clefts	number of matrix clauses
Old French	1.2	0.25	12022
Middle French	0.41	0.61	24634
Early Modern	0.56	5.4	3514

Finis