

# Using parsed corpora to compare the evolution of word order in English and French

Anthony Kroch  
University of Pennsylvania  
March 2010

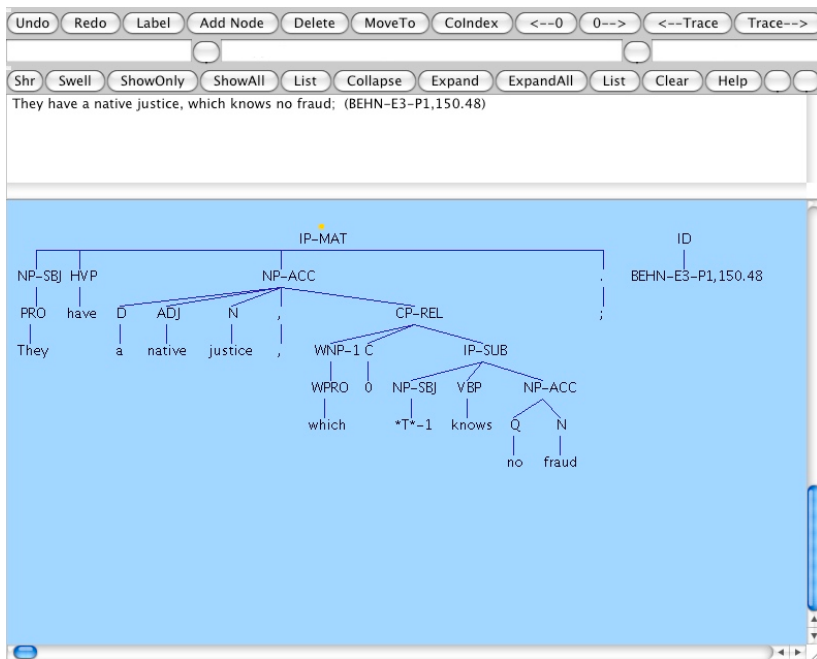
[www.ling.upenn.edu/~kroch/handouts/gcoe.pdf](http://www.ling.upenn.edu/~kroch/handouts/gcoe.pdf)

## 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

## 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-PI,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 historical corpus resources for European languages

### English Parsed Corpora I

- Anthony Kroch and Ann Taylor. *Penn-Helsinki Parsed Corpus of Middle English, second edition*. University of Pennsylvania, 2000.  
1.3 million words
- Anthony Kroch, Beatrice Santorini, and Ariel Diertani. *Penn-Helsinki Parsed Corpus of Early Modern English, first edition*. University of Pennsylvania, 2004.  
1.8 million words
- Anthony Kroch, Beatrice Santorini, and Ariel Diertani. *Penn Parsed Corpus of Modern British English, first edition*. 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.

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

- Charlotte Galves et al. *Tycho Brahe Corpus of Historical Portuguese*, first edition. University of Campinas, São Paulo, Brazil, 2003.

≈2 million words

- France Martineau et al. *MCVF Corpus of Historical French*, first edition. University of Ottawa, 2006.

≈1 million words

Wednesday, March 3, 2010

Wednesday, March 3, 2010

## Total Currently Available Parsed Historical Text

English	7.8 million
Portuguese	≈2 million
French	≈1 million

Wednesday, March 3, 2010

Wednesday, March 3, 2010

### Penn Parsed Corpora of Historical English



The Penn Historical Corpora, including the Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME), and the Penn Parsed Corpus of Modern British English (PPCMBE), are syntactically annotated corpora of prose text samples of English from the indicated time periods. Their syntactic annotation (parsing) permits searching not only for words and word sequences, but also for syntactic structure. The corpora are designed for the use of students and scholars of the history of English, especially the historical syntax of the language, and they are publicly available.

The Penn Parsed Corpora of Historical English are available on CD-ROM at a charge of \$450 for a single-user license for all three corpora. Site licenses for departments/research groups and libraries are also available. See the corpus order form for charges. Corpus license fees go toward improving the corpora and increasing them in size. Upgrades, when completed, are available to corpus license holders at a modest cost.

If you currently hold a license for the two older corpora, the PPCME2 and the PPCEME, and wish to obtain the PPCMBE, please contact akrochATgmailDOTcom for upgrade information.

The Penn Historical Corpora are distributed with a search program CorpusSearch2, written by Beth Randall and released as open source software downloadable from its Sourceforge [project web site](#).

- The PPCME2 was created with the support of the National Science Foundation (Grants BNS89-19701 and SBR95-11368), with supplementary support from the University of Pennsylvania Research Foundation.
- The PPCEME was created with the support of the National Endowment for the Humanities (Grant PA23382-99) and the National Science Foundation (Grant BCS99-05488).
- The PPCMBE was created with the support of the National Science Foundation (Grant BCS05-08731).

<http://www.ling.upenn.edu/hist-corpora/>

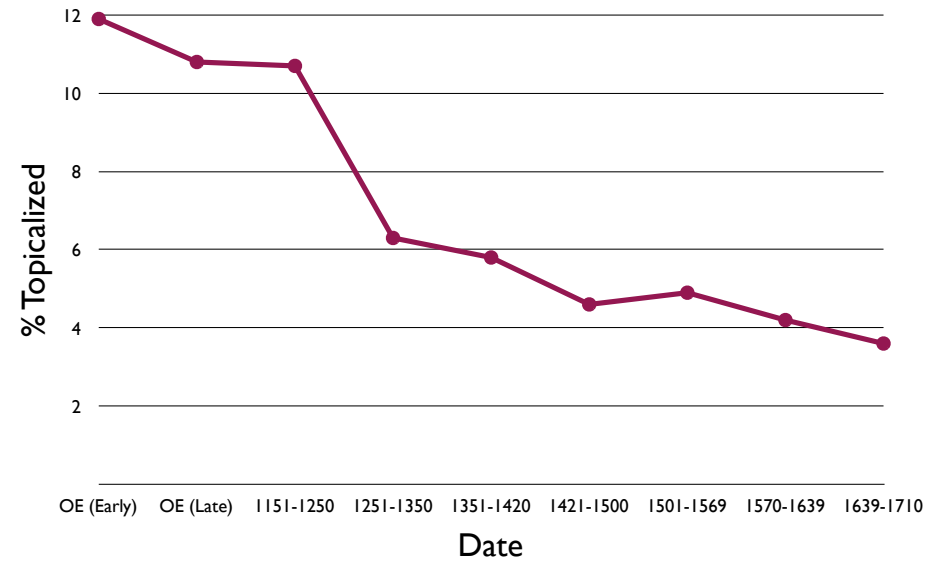
- PPCME2
- PPCEME
- PPCMBE
- Corpus annotation
- CorpusSearch
- Order corpora
- Other corpora
- Contacts / Updates

Wednesday, March 3, 2010

Wednesday, March 3, 2010

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

## Decline of direct object topicalization in English



Wednesday, March 3, 2010

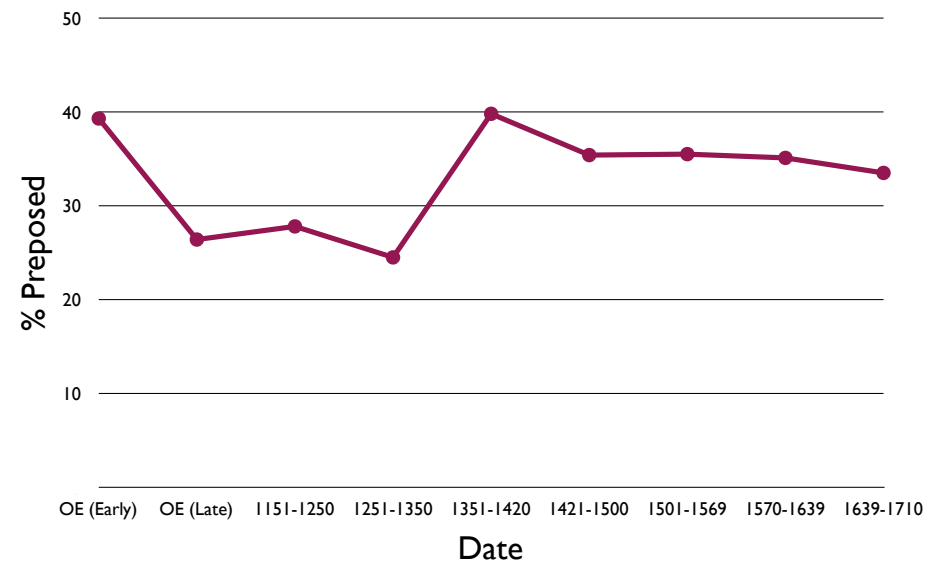
Wednesday, March 3, 2010

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

Table 4.2: Summary of Vorfeld occupation of arguments.

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

## Evolution of PP preposing in English



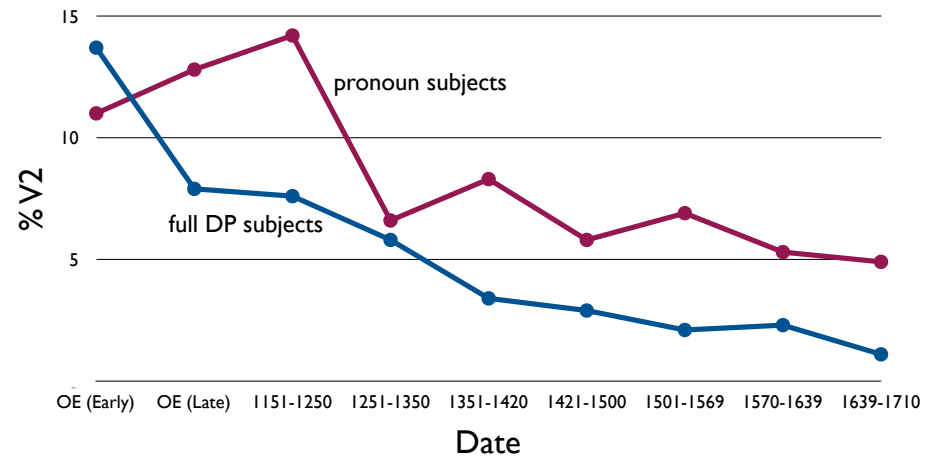
Wednesday, March 3, 2010

Wednesday, March 3, 2010

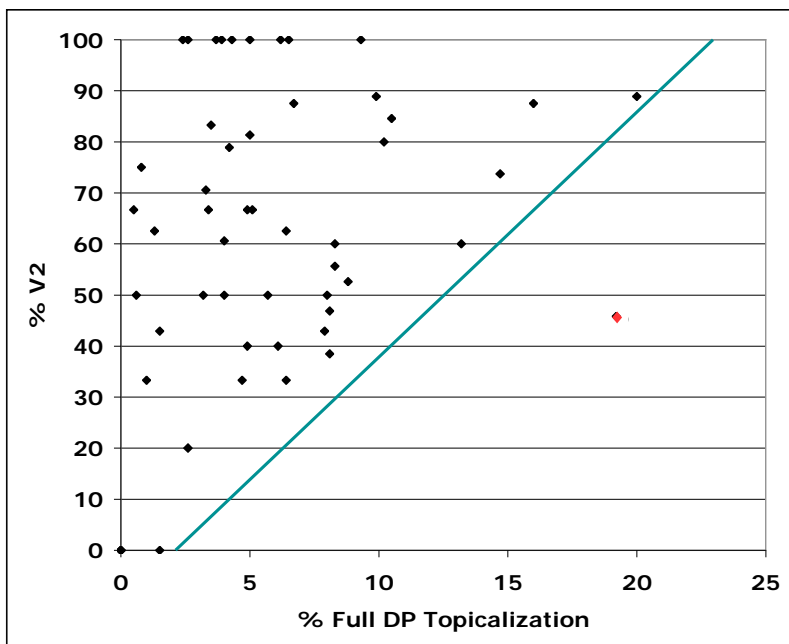
# The history of topicalization in English (Speyer 2008)

- 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.

## Decline of direct object topicalization by subject type



## 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
<b>46.4</b>	<b>6.6</b>	<b>47.0</b>

### Subject type in sentences with *in situ* objects

personal pronoun	demonstrative pronoun	full noun phrase
181	2	17
<b>90.5%</b>	<b>1%</b>	<b>8.5%</b>

### 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.

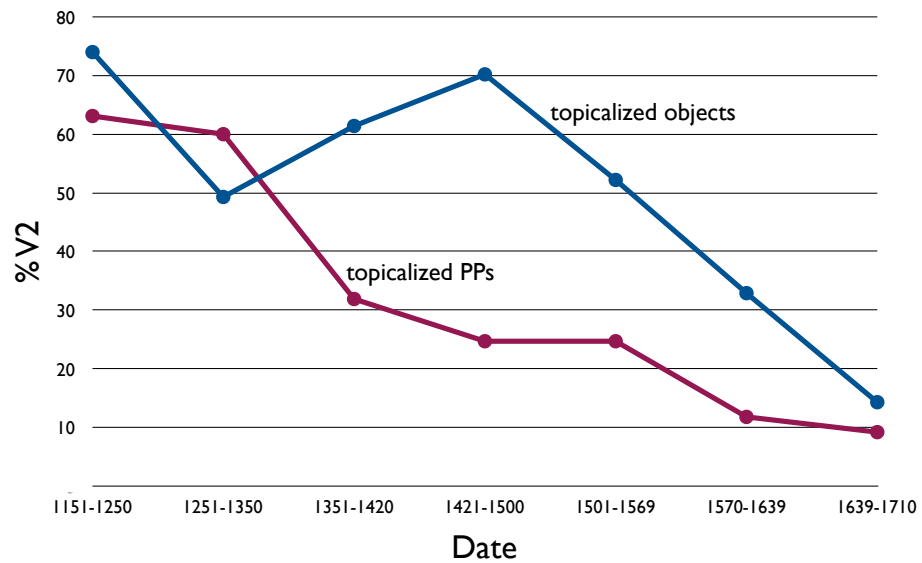
# Accent placement and topicalization frequencies in translating German topicalized arguments into English

	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 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

## V2 loss in English sentences with topicalized objects and PPs

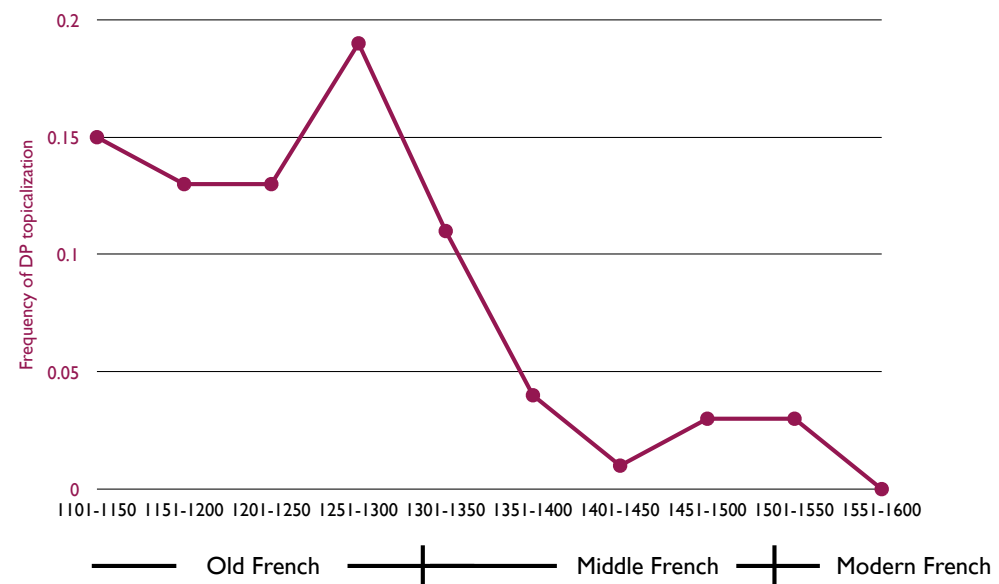


## The loss of verb-second word order in French

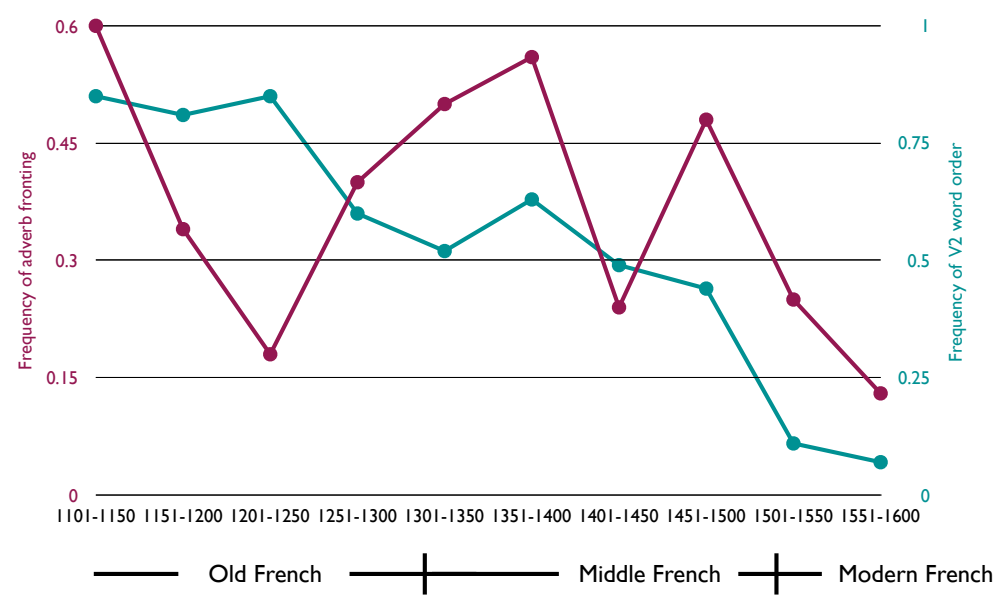
### V2 in Old and Middle French

- (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

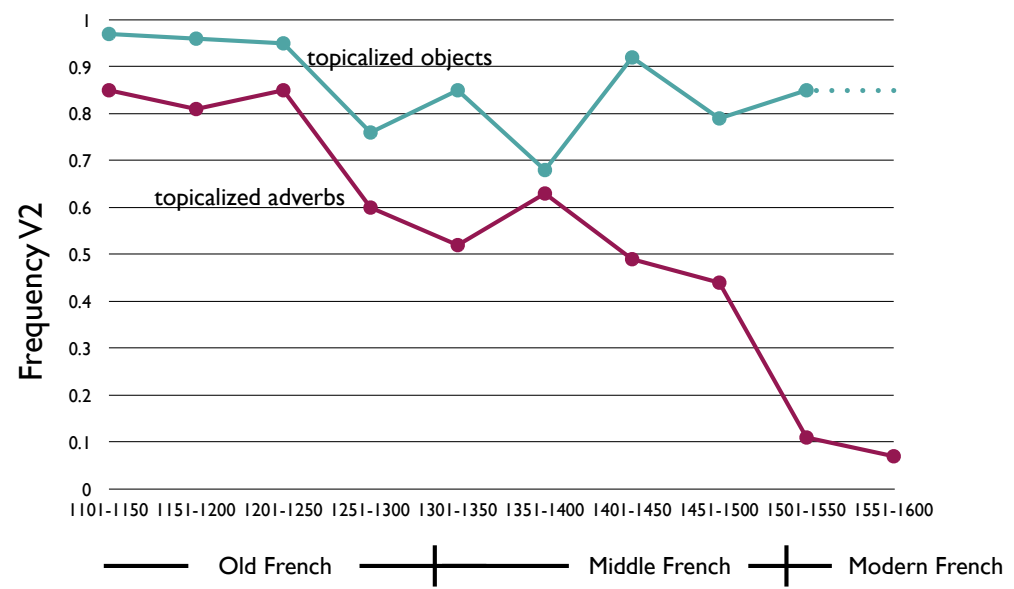
### Decline of direct object topicalization in French



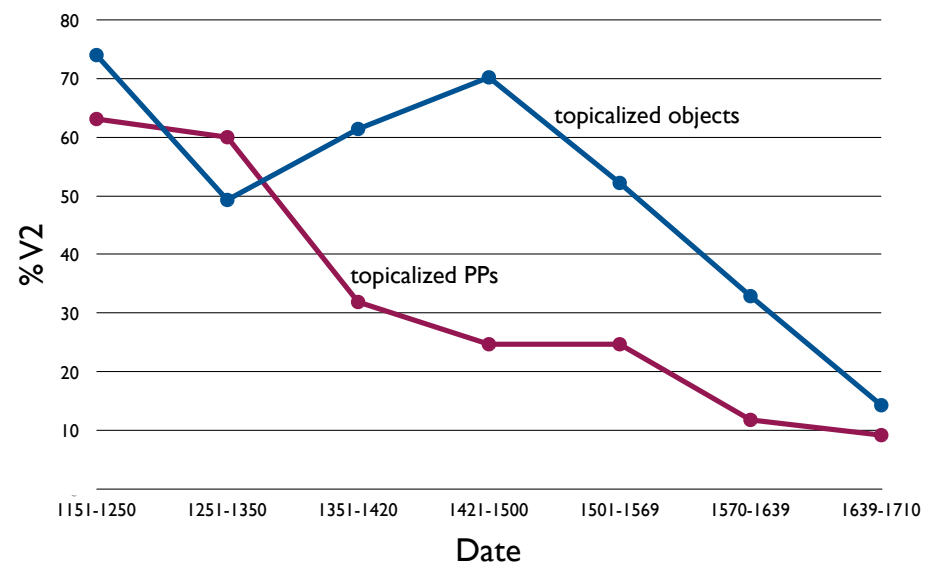
# Evolution of adverb fronting and V2 word order in French



# Evolution of V2 word order in French

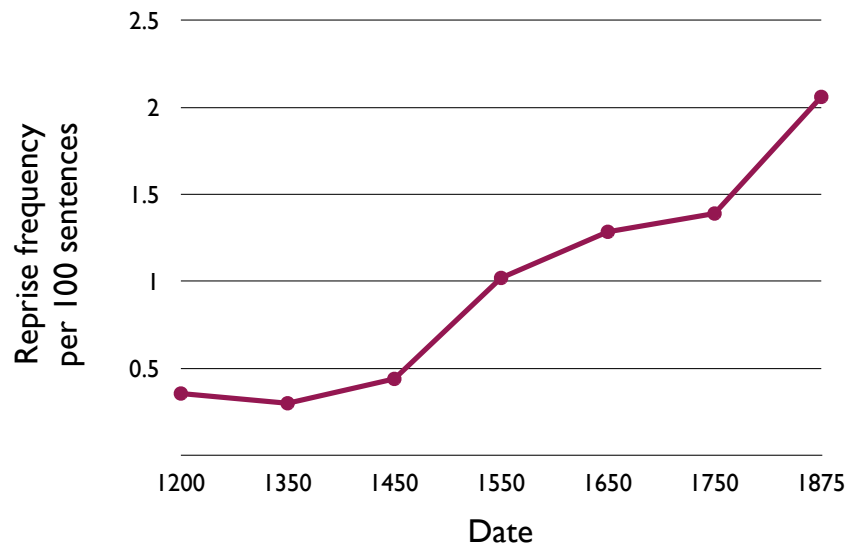


# V2 loss in English sentences with topicalized objects and PPs



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*<sub>i</sub>, Jean \*(le)<sub>i</sub> lit tous les jours.  
The Figaro John it reads every day
- (2) *Ma femme*<sub>i</sub>, elle<sub>i</sub> 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*<sub>i</sub> que Jean lit <sub>t<sub>i</sub></sub> tous les jours.  
It's The Figaro that John reads every day
- (2) C'est *ma femme*<sub>i</sub> qui <sub>t<sub>i</sub></sub> travaille à la BN.  
It's my wife that works at the BN
- (3) Il y a *un an*<sub>i</sub> qu'elle travaille à la BN <sub>t<sub>i</sub></sub>.  
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

Coda

## “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 assez 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	<b>0.86</b> [218]	<b>0.83</b> [2163]
Middle French	<b>0.69</b> [402]	<b>0.70</b> [3633]
Modern French	<b>0.27</b> [33]	<b>0.22</b> [160]

## Temporal evolution of Germanic and Romance inversion in V2 sentences with topicalized XPs and full DP subjects

	frequency of Germanic inversion	frequency of Romance inversion	Romance + Germanic inversion
Old French	<b>0.50</b>	<b>0.36</b>	0.86
Middle French	<b>0.32</b>	<b>0.37</b>	0.69
Modern	<b>0.03</b>	<b>0.24</b>	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