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

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

```plaintext
((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))
```

Wednesday March 3 2010
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

  1.3 million words

  1.8 million words

  1.0 million words
EnglishParsedCorporaII

  1.5 million words

  2.2 million words

Otherlanguages

  \(\approx\) 2 million words

  \(\approx\) 1 million words

TotalCurrentlyAvailableParsedHistoricalText

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>7.8 million</td>
</tr>
<tr>
<td>Portuguese</td>
<td>(\approx) 2 million</td>
</tr>
<tr>
<td>French</td>
<td>(\approx) 1 million</td>
</tr>
</tbody>
</table>
The loss of verb-second word order and the decline of topicalization in English

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

Table 4.2: Summary of Vorfeld occupation of arguments.

<table>
<thead>
<tr>
<th>Argument</th>
<th>Vorfeld</th>
<th>Prop est (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>subject</td>
<td>43 523</td>
<td>18 597</td>
</tr>
<tr>
<td>direct object</td>
<td>3 418</td>
<td>20 432</td>
</tr>
<tr>
<td>indirect object</td>
<td>38</td>
<td>815</td>
</tr>
</tbody>
</table>

Decline of direct object topicalization in English

Evolution of PP preposing 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.
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

<table>
<thead>
<tr>
<th></th>
<th>focus accent on the German subject</th>
<th>accent elsewhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>topicalization in the English translation</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>no topicalization in the English</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

Distribution of contrastive topicalization by focus accent placement in Middle English

<table>
<thead>
<tr>
<th>focus position of cases</th>
<th>focus on subject</th>
<th>focus on tensed verb</th>
<th>focus elsewhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (total= 207)</td>
<td>113</td>
<td>29</td>
<td>65</td>
</tr>
<tr>
<td>% inversion</td>
<td>89</td>
<td>14</td>
<td>71</td>
</tr>
<tr>
<td>% of cases</td>
<td>55</td>
<td>14</td>
<td>31</td>
</tr>
</tbody>
</table>
**V2 loss in English sentences with topicalized objects and PPs**

- % V2
- Date
- 1151-1250, 1251-1350, 1351-1420, 1421-1500, 1501-1569, 1570-1639, 1639-1710

**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 fiel
   - 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**

- Frequency of DP topicalization
- Old French, Middle French, Modern French

- Date
- 1101-1150, 1151-1200, 1201-1250, 1251-1300, 1301-1350, 1351-1400, 1401-1450, 1451-1500, 1501-1550, 1551-1600
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*, Jean *(le)* lit tous les jours.
    *The Figaro* John it reads every day

(2) Ma femme, elle 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

<table>
<thead>
<tr>
<th></th>
<th>frequency of subject left dislocation</th>
<th>frequency of object left dislocation</th>
<th>number of matrix clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old French</td>
<td>2.6</td>
<td>2.2</td>
<td>12022</td>
</tr>
<tr>
<td>Middle French</td>
<td>3.8</td>
<td>1.8</td>
<td>24634</td>
</tr>
<tr>
<td>Early Modern</td>
<td>28</td>
<td>4.3</td>
<td>3514</td>
</tr>
</tbody>
</table>

Cleft sentences in Modern French

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

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

(3) Il y a un an qu’elle travaille à la BN.
    It’s one year that-she works at the BN
Temporal evolution of cleft sentence frequencies per thousand sentences

<table>
<thead>
<tr>
<th></th>
<th>frequency of temporal clefts</th>
<th>frequency of subject and object clefts</th>
<th>number of matrix clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old French</td>
<td>1.2</td>
<td>0.25</td>
<td>12022</td>
</tr>
<tr>
<td>Middle French</td>
<td>0.41</td>
<td>0.61</td>
<td>24634</td>
</tr>
<tr>
<td>Early Modern</td>
<td>0.56</td>
<td>5.4</td>
<td>3514</td>
</tr>
</tbody>
</table>

“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

<table>
<thead>
<tr>
<th></th>
<th>sentences with an auxiliary verb</th>
<th>sentences with a single verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old French</td>
<td>0.86 [218]</td>
<td>0.83 [2163]</td>
</tr>
<tr>
<td>Middle French</td>
<td>0.69 [402]</td>
<td>0.70 [3633]</td>
</tr>
<tr>
<td>Modern French</td>
<td>0.27 [33]</td>
<td>0.22 [160]</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>frequency of Germanic inversion</th>
<th>frequency of Romance inversion</th>
<th>Romance + Germanic inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old French</td>
<td>0.50</td>
<td>0.36</td>
<td>0.86</td>
</tr>
<tr>
<td>Middle French</td>
<td>0.32</td>
<td>0.37</td>
<td>0.69</td>
</tr>
<tr>
<td>Modern</td>
<td>0.03</td>
<td>0.24</td>
<td>0.27</td>
</tr>
</tbody>
</table>
An independence result

<table>
<thead>
<tr>
<th></th>
<th>Romance + Germanic inversion</th>
<th>sentences with a single verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old French</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Middle French</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>Modern French</td>
<td>0.27</td>
<td>0.22</td>
</tr>
</tbody>
</table>