A first attempt to use the Parsed Linguistic Atlas of Early Middle English as an atlas:
Middle English V2

Rob Truswell
rob.truswell@ed.ac.uk

DAAD workshop: Mapping Language Variation and Change
Cambridge, 18/3/19
In brief

- We took part of a Linguistic Atlas of Early Middle English (Laing 2013)...
- ... and made it into a parsed corpus (the Parsed Linguistic Atlas of Early Middle English, Truswell et al. 2019).
- Our primary goal in doing this was diachronic-syntactic, in the Penn tradition of parsed historical corpora.
- Have we obliterated the dialectological virtues of LAEME?
- Or have we extended them by allowing easier investigation of syntactic variation?
Roadmap

1. Introduction to PLAEME.
2. Replication of classic ME dialect syntax results from Kroch & Taylor (1997).
The Penn–Helsinki Parsed Corpus of Middle English, 2nd edition (Kroch & Taylor 2000) is now the industry-standard resource for Middle English syntactic research.

- > 1m words, spanning 1150–1500.
- Annotated with POS and constituency information.
- Allows retrieval of large amounts of high-quality data in minutes.
Sample query

```
node: CP-REL*|CP-CAR*
print_indices: [1]

query: (CP-REL*|CP-CAR* iDomsMod W* WD)
```
and with that digs a hole, which hole hee bids him make for his graue,

(AARMIN-E2-P1,25.343)

*/

21 CP-CAR: 21 CP-CAR, 23 WD

*/

(0 (1 IP-MAT (2 CONJ and)
 (4 NP-SBJ *con*)
 (6 PP (7 P with)
   (9 NP (10 D that)))
 (12 VBP digs)
 (14 NP-OBJ (15 D a) (17 N hole))
 (19 , ,)
 (21 CP-CAR (22 WPN-1 (23 WD which) (25 N hole))
 (27 C 0)
 (29 IP-SUB (30 NP-OBJ *T*-1)
 (32 NP-SBJ (33 PRO hee))
 (35 VBP bids)
 (37 IP-.Inf (38 NP-SBJ (39 PRO him))
 (41 VB make)
 (43 PP (44 P for)
 (46 NP (47 PRO$ his) (49 N

graue))))))

(51 . ,))

(53 ID AARMIN-E2-P1,25.343))
Sample query: counts
## The data gap: PPCME2, 1150–1350

<table>
<thead>
<tr>
<th>Filename</th>
<th>Title</th>
<th>Date</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>cmkentho</td>
<td>Kentish Homilies</td>
<td>c12a2–b1</td>
<td>4048</td>
</tr>
<tr>
<td>cmpeterb</td>
<td>Peterborough Chronicle</td>
<td>c.1131, c.1154</td>
<td>6757</td>
</tr>
<tr>
<td>cmlambx1</td>
<td>Lambeth Homilies</td>
<td>c12b2</td>
<td>20752</td>
</tr>
<tr>
<td>cmtrinit</td>
<td>Trinity Homilies</td>
<td>c12b2</td>
<td>41844</td>
</tr>
<tr>
<td>cmorm</td>
<td>Ormulum</td>
<td>c12b2</td>
<td>50579</td>
</tr>
<tr>
<td>cmlamb1</td>
<td>Lambeth Homilies</td>
<td>c12b2</td>
<td>6459</td>
</tr>
<tr>
<td>cmvices1</td>
<td>Vices and Virtues</td>
<td>c13a1</td>
<td>27677</td>
</tr>
<tr>
<td>cmsawles</td>
<td>Sawles Warde</td>
<td>c13a2</td>
<td>4111</td>
</tr>
<tr>
<td>cmhali</td>
<td>Hali Meiðhad</td>
<td>c13a2</td>
<td>8495</td>
</tr>
<tr>
<td>cmkathe</td>
<td>St. Katherine</td>
<td>c13a2</td>
<td>8699</td>
</tr>
<tr>
<td>cmjulia</td>
<td>St. Juliana</td>
<td>c13a2</td>
<td>6810</td>
</tr>
<tr>
<td>cmmarga</td>
<td>St. Margaret</td>
<td>c13a2</td>
<td>8069</td>
</tr>
<tr>
<td>cmancriw</td>
<td>Ancrene Riwle</td>
<td>c13a2</td>
<td>63790</td>
</tr>
<tr>
<td>cmkentse</td>
<td>Kentish Sermons</td>
<td>c13b2?</td>
<td>3534</td>
</tr>
<tr>
<td>cmayenbi</td>
<td>Ayenbite of Inwyty</td>
<td>1340</td>
<td>45944</td>
</tr>
<tr>
<td>cmearlps</td>
<td>Earliest Prose Psalter</td>
<td>c.1350</td>
<td>44521</td>
</tr>
</tbody>
</table>
The data gap: PPCME2, 1150–1350
LAEME complements PPCME2

- LAEME:
  - covers 1150–1325;
  - includes a much broader range of texts:
    - Verse/prose;
    - Fragmentary/whole;
    - Long/short;
    - Multiple versions of same text.
Sample of 68 texts (172,624 words): Single version of all texts meeting the following:

1. Manuscript is from 1250–1325;
2. No parsed version currently exists;
3. > 100 words.

Where multiple versions of a text meet these criteria:

1. Aim to balance across dialect areas;
2. All else being equal, take the longest version.

Small amount of text (8 files, all short) unlocalized, mainly excluded today.
LAEME has an incredibly detailed (in principle infinite!) tagset, including information about grammatical function, some nonlocal dependencies, and some meaning distinctions, as well as part of speech.

This formed the basis of preliminary labelled bracketing.

LAEME ‘lexels’ are also provided for all content words (and can be automatically added for all function words), so PLAEME could be lemmatized for free, eliminating challenges relating to orthographic variation.

LAEME marks rhymes, and these annotations are transferred to PLAEME, potentially useful for investigating effect of verse and metre on word order.
Automatic bracketing: Examples

\textit{be iuele man} ‘the evil man’ \hspace{1cm} \textit{ner be se} ‘near the sea’

\begin{align*}
\$/TN\_yE \hspace{1cm} & \$/near/pr\_NER \\
$\text{evil/aj\_IUELE} \hspace{1cm} & $\text{T<pr_yE} \\
$\text{man/n\_MAN} \hspace{1cm} & $\text{sea/n<pr\_SE} \\
\text{(NP-SBJ (D +te-the))} \hspace{1cm} & \text{(PP (P ner-near))} \\
\text{(ADJ iuele-evil)} \hspace{1cm} & \text{(NP (D +te-the))} \\
\text{(N man-man))} \hspace{1cm} & \text{(N se-sea))}
\end{align*}
Automatic bracketing: Examples

"dat ghe ne migte hĩ bringen on ‘what she might not prove against him’

(CP-REL (WNP 0)
  (C +dat-that)
  (IP-SUB (NP-OB1 *T*)
    (NP-SBJ (PRO ghe-she))
    (NEG ne-ne)
    (MD migte-may)
    (NP (PRO hiM-him))
    (VB bring+en-bring)
    (PP (P-RH on-on))
    (NP *ICH*)))))
Manual correction

All structures corrected, indices added, etc., using Annotald (Beck et al. 2011).

\[
\text{(CP-FRL (WNP-1 0))}
\text{(C +dat-that)}
\text{(IP-SUB (NP-OB1 *T*-1))}
\text{(NP-SBJ (PRO ghe-she))}
\text{(NEG ne-ne)}
\text{(MD migte-may)}
\text{(NP-2 (PRO hiM-him))}
\text{(VB bring+en-bring)}
\text{(PP (P-RH on-on)}
\text{(PP (P-RH on-on)}
\text{(NP *ICH*-2))}
\]
PLAEME largely fills the gap in PPCME2
The extra data is helpful
Time course of Jespersen’s Cycle in English
The extra data is helpful

Time course of recipient–theme ditransitives with *to*

Corpus
- PLAEME+PCMEP
- PPCHE

Tokens
- 10
- 25
The extra data is helpful

Emergence of argument-gap *wh*-relatives
But still

- LAEME is more than a corpus: it’s an atlas.
- It has been used for dialect syntax work, e.g. on the expression of negation (Laing 2002, Walkden & Morrison 2017).
- Some of our construction decisions (esp. no parallel texts) may militate against using PLAEME in the same way.
- Today: an exploration of PLAEME as a syntactic atlas.
Kroch & Taylor (1997)

- We will attempt to replicate Kroch & Taylor’s (1997) results about dialect contact in Middle English V2.
- This is an ideal case study for several reasons:
  1. The analysis revolves around dialectal differences.
  2. The analysis is irreducibly phrase-structural: parsed corpus comes into its own.
  3. The phenomenon involves fine differences in word order: useful testing ground for investigating the effect of verse.
  4. One of the dialects is sparsely represented in PPCME2: lots of inferences are drawn from a single text.
Kroch & Taylor summary

- Southern Early ME was an IP-V2 language.
- Subject pronouns are clitics: target IP/CP border.
  - Some embedded V2;
  - Matrix V3 orders with subject pronouns but not with full NP subjects.
- Northern Early ME was a CP-V2 language, though subject pronouns are still clitics.
  - No embedded V2?
  - No differentiation between pronouns and full NPs w.r.t. placement of subjects.
### Kroch & Taylor’s numbers

<table>
<thead>
<tr>
<th>Preposed</th>
<th>Southern NP % inv.</th>
<th>Southern Pronoun % inv.</th>
<th>Ayenbite NP % inv.</th>
<th>Ayenbite Pronoun % inv.</th>
<th>cmbenrul NP % inv.</th>
<th>cmbenrul Pronoun % inv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP compl.</td>
<td>93</td>
<td>5</td>
<td>82</td>
<td>8</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>PP compl.</td>
<td>75</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Adj. compl.</td>
<td>95</td>
<td>33</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>67</td>
</tr>
<tr>
<td><em>pa/then</em></td>
<td>95</td>
<td>72</td>
<td>25</td>
<td>58</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td><em>now</em></td>
<td>92</td>
<td>27</td>
<td>100</td>
<td>50</td>
<td>NA</td>
<td>100</td>
</tr>
<tr>
<td>PP adj.</td>
<td>75</td>
<td>2</td>
<td>36</td>
<td>3</td>
<td>89</td>
<td>91</td>
</tr>
<tr>
<td>Other adv.</td>
<td>57</td>
<td>1</td>
<td>56</td>
<td>10</td>
<td>96</td>
<td>91</td>
</tr>
</tbody>
</table>
Sanity check I

Replicate Kroch & Taylor’s counts

- Worthwhile because volume of PPCME2 data has roughly tripled for the southern dialects.
- Also checks that my queries do roughly what theirs do.
- Collapsed PP complement/adjunct because not coded in PPCME2.
- Results pretty much hold up.

<table>
<thead>
<tr>
<th>Preposed</th>
<th>Southern (&lt; 1250)</th>
<th>Ayenbite (1340)</th>
<th>cmbenrul (a1425)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NP % inv.</td>
<td>Pronoun % inv.</td>
<td>NP % inv.</td>
</tr>
<tr>
<td>NP compl.</td>
<td>87</td>
<td>13</td>
<td>80</td>
</tr>
<tr>
<td>PP</td>
<td>76</td>
<td>15</td>
<td>78</td>
</tr>
<tr>
<td>Adj. compl.</td>
<td>94</td>
<td>27</td>
<td>100</td>
</tr>
<tr>
<td><em>pa/then</em></td>
<td>93</td>
<td>80</td>
<td>42</td>
</tr>
<tr>
<td><em>now</em></td>
<td>75</td>
<td>17</td>
<td>75</td>
</tr>
<tr>
<td>Other adv.</td>
<td>63</td>
<td>11</td>
<td>65</td>
</tr>
</tbody>
</table>
The ‘southern’ pattern

(1) Efter þe þridde fiue þe schule seggen [. . .] Kirieleyson [etc.]
    after the third five you shall say Kyrie eleison
    ‘After the third five, you shall say Kyrie eleison, etc.’
    (cmancriw-1-m1,l.60.193)

(2) cheos þenne of þeos twa for þoðer þu most leten.
    choose then of those two for the other thou must let
    ‘Choose, then, between those two, because you must leave the other.’
    (cmancriw-1-m1,II.81.978–9)

(3) Nu þu hauest iseid tus.
    now thou hast said thus
    ‘Now you have said thus.’
    (cmhali-m1,147.276)
The ‘northern’ pattern

(4) Lauerd, of me haue I noht, bot þu sende it me.
    lord of me have I naught but thou send it me
    ‘Lord I have nothing of myself unless you send it to me.’
    (cmbenrul-m3,3.60)

(5) Mi scole wil i stablis to godis seruise.
    My school will I establish to God’s service
    ‘I will establish my school to serve God.’
    (cmbenrul-m3,4.84)

(6) now wil I blinne to speke of þaim, for it ne helpis noht
    now will I cease to speak of them for it NEG helps not
    ‘Now I will stop speaking of them, because it doesn’t help.’
    (cmbenrul-m3,5.118)
Sanity check II

Effect of verse

- I used the Parsed Corpus of Middle English Poetry (Zimmermann 2015) to get a sample with more verse (also *Ormulum* from PPCME2).
- Rephrasing the question: Is verse vs. prose a significant predictor of inversion?
- Several choices about model structure, mixed effects vs. classical logistic regression, etc. Under pretty much every choice, verse vs. prose isn’t significant

Formula:

```r
ifelse(Inv == "Inv", 1, 0) ~ ClauseType + SbjType + Year + Genre +
(1 | File)
```

Fixed effects:

| Estimate   | Std. Error | df   | t value | Pr(>|t|) |
|------------|------------|------|---------|----------|
| (Intercept)| 1.130e+00  | 2.844e-01 | 8.240e+01 | 3.974    | 0.000151 *** |
| ClauseTypeSub| -8.538e-02| 1.124e-02 | 2.363e+04 | -7.595   | 3.2e-14 ***   |
| SbjTypePronoun| -3.079e-01| 5.706e-03 | 2.365e+04 | -53.960  | < 2e-16 ***   |
| Year       | -3.987e-04| 2.065e-04 | 8.250e+01 | -1.931   | 0.056933 .    |
| GenreVerse | -3.294e-02| 4.772e-02 | 8.556e+01 | -0.690   | 0.491904      |
Sanity check II

Effect of verse

- This does not mean that verse has no effect on word order.
- It means that we can’t see a systematic effect on word order (within the rest of our framework of assumptions).
- Interpreting any one example requires analytical sensitivity to such factors.
- But the verse nature of most PLAEME texts shouldn’t be construed as a barrier to drawing quantitative inferences about dialectal variation in V2.
Sanity check III
Can we detect nonsyntactic dialectal differences in PLAEME?

- Good representation of several broad dialect areas, though geographical coverage inevitably patchy.
- Yorkshire texts all relatively late in period, but still significantly earlier than first prose texts from the north in PPCME2.
Sanity check III

They vs. hi
Major differences in functional vocabulary are well-represented in PLAEME.

(Though not every northern text robustly shows all ‘northern’ features).

This should increase our hopes that regional differences w.r.t. V2 will be interpretable.
On to V2

No general diachronic pattern across PLAEME

- *Cursor Mundi* and *Northern Homilies* (in red) are outliers.
- Large, late, northern texts, with distinctive syntax.
- To be continued.
On to V2

Distribution of inversion in matrix clauses with one (or more) of the preposed elements identified by Kroch & Taylor.

V2 concentrated in north.

(All such statements supported by series of mixed-effects models, details skipped).
But V2 with full NPs is everywhere

- 69% of matrix clauses with preposed elements have inversion.
- No geographical pattern (no significant predictors at all).
- English c.1300 is still largely a V2 language.
The distinctive pronoun pattern

Regional differences driven by inversion around pronouns.

And inversion around pronouns in matrix clauses is a significantly northern thing.

So Kroch & Taylor’s main conclusion is supported by the PLAEME data.
But there’s more
Embedded V2 with pronouns

▶ Kroch & Taylor use pronoun subjects to diagnose CP-V2 vs. IP-V2.
▶ But inversion around pronominal subjects is also well attested in embedded clauses in some northern texts.
▶ Not originally taken to be a CP-V2 property (though we have a lot more projections to play with now).
edincXt is different again

- Three text languages, two texts (*Cursor Mundi* in two hands, *Northern Homilies*), in one manuscript (edicmat/bt/ct).
- Main vs. subordinate clause is *not* a significant predictor of inversion.
- Can’t tell in *Rule of St. Benet* because virtually no relevant contexts in subordinate clauses (only 7 vs. 343 matrix; compare edincXt 123 embedded, 1458 matrix).

Formula: `ifelse(Inv == "Inv", 1, 0) ~ SbjType + ClauseType + (1 | Filename)`

Fixed effects:

|                     | Estimate | Std. Error | df     | t value | Pr(>|t|) |
|---------------------|----------|------------|--------|---------|----------|
| (Intercept)         | 0.69664  | 0.08123    | 2.25662| 8.576   | 0.00912  **|
| SbjTypePronoun      | -0.21237 | 0.02635    | 1576.11410 | -8.058 | 1.51e-15 ***|
| ClauseTypeSub       | -0.05887 | 0.04457    | 1576.21652 | -1.321 | 0.18675   |
(7) For soruīg al dūb war þai
for sorrowing all dumb were they

Swaþat a word miht þai noht sai
so.that a word might they not say

Na stand apō þair fete
nor stand upon their feet

(6) Na stand apō þair fete
nor stand upon their feet

(8) For he suar bi þe kīg of heuī
for he swore by the king of heaven

Þat harald slahtir suld he heuī
that Harold’s slaughter should he avenge

(6) Þat harald slahtir suld he heuī
that Harold’s slaughter should he avenge


What’s going on?

▶ In Kroch & Taylor’s terms, the natural analysis would be that the edincmXt texts show IP-V2 with nonclitic subjects.

▶ They give two arguments why this couldn’t be the correct analysis of the *Rule of St. Benet*. At least one also holds for edincmXt.

1. Sensitivity of inversion to preposed element.
   ▶ Southern EME inverts a lot more with preposed NP/AP/*then* than with preposed PP/AdvP/*now*.
   ▶ *Benet’s rule* inverts with preposed anything, regardless of whether the subject is a pronoun.
   ▶ edincmXt is different again: near-categorical inversion after *then* and *now*, variable otherwise.

2. Stylistic fronting with pronominal subjects.
   ▶ Stylistic fronting requires empty [Spec,IP].
   ▶ Occurs with apparently in situ pronominal subjects, including in *Rule of St. Benet*.
   ▶ Makes sense if those subjects have left [Spec,IP] by cliticization.
   ▶ No shortage of stylistic fronting with pronominal subjects in edincmXt.
edincmXt stylistic fronting examples

(9) Astank It cald es of sain Ion
A.stank it called is of Saint John

(10) Bot þar es nan þat gernis mar
but there is none that yearns more

Pan þai ĵ (er)uis worþi war
than they in service worthy were

(11) Bot als þaime vp help wit hād
but as they.me up helped with hand

Vnbū was ik of bote
unbound was I of mercy
Conclusions
From most to least general

- PLAEME is useable as a syntactic atlas as well as a diachronic corpus.
- Verse texts are useable for investigation of word order change.
- Kroch & Taylor’s (1997) claims about dialectal variation in ME V2 largely survive testing against new data.
- But the syntax of one northern text (Rule of St. Benet) doesn’t match that of another set of texts (edincmXt).
- And we don’t understand the syntax of the latter perfectly.
Next steps

- Use PLAEME to investigate any of the many changes c.1300 where inflection arguably plays a role.
  - Ditransitives
  - Relatives
  - ...
- Expand PLAEME, but in which direction?
  - Back in time?
  - Parallel versions?
- Parallel corpora/atlases?
  - LAOS?
Construction of PLAEME was funded by British Academy/Leverhulme small research grant SG150315. Thanks to them, to my collaborators Rhona Alcorn and Joel Wallenberg, RA Jim Donaldson, and to our indefatigable sources of advice and encouragement, Meg Laing, Beatrice Santorini, Susan Pintzuk, and Aaron Ecay.
References


