A Parsed Linguistic Atlas of Early Middle English

Rob Truswell, Rhona Alcorn, and James Donaldson
University of Edinburgh
rob.truswell@ed.ac.uk, r.alcorn@ed.ac.uk,
james.donaldson@ed.c.uk

AMC symposium, 10/6/16
An initial report on a new, BA/Leverhulme-funded, corpus project.

Goal: use data from the Linguistic Atlas of Early Middle English to plug a gap in the PennParsed Corpora of Historical English.

Agenda:
1. Beginner’s guide to PPCHE
2. Beginner’s guide to LAEME
3. Plans for the project
4. Beyond this project
Section 1

PPCHE
The Penn historical corpus format

- Family of parsed historical corpora covering English (7.7m words, –1914) and several other Germanic and Romance languages.
- Annotated with detailed labelled brackets.
  - Explicit annotation of grammatical function.
  - Explicit representation of many cases of “movement”, extraposition, etc.
- The PPCHE philosophy:
  - no claim to being an accurate theory of syntax.
  - aim to include useful information about constituency in a way that can be consistently implemented.
    - So (almost) no VP nodes, default high attachment, etc.
- Major virtue: very easy to query information about phrasal syntax.
- Very easy to make (some kind of) crosslinguistic comparison.
Sample query

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

query: (CP-REL*|CP-CAR* iDomsMod W* WD)
```

and with that digs a hole, which hole hee bids him make for his graue,
(ARMIN-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-OB1 (15 D a) (17 N hole))
(19 , ,)
(21 CP-CAR (22 WNP-1 (23 WD which) (25 N hole))
 (27 C 0)
(29 IP-SUB (30 NP-OB1 *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 ARMIN-E2-P1,25.343))
Sample query: counts

```
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0/0/996</td>
<td>0/0/341</td>
<td>0/0/104</td>
<td>0/0/37</td>
<td>0/0/30</td>
<td>0/0/12</td>
<td>0/0/21</td>
<td>1/1/629</td>
<td>0/0/12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>whole search, hits/tokens/total</td>
<td>1834/1737/257436</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

PPCME2 weaknesses

- Built from published editions, not manuscripts.
- Limited metadata.
- Not lemmatized.
- Significant data gap c.1250–1340.
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>cmmancriw</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 Inwyt</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
The data gap in action

Ecay & Tamminga (2016): Expression of negation

Bacovcin (2016): Double object word order

Gisborne & Truswell (2016): Argument relatives
Help is at hand

- Luckily, the Linguistic Atlas of Early Middle English (Laing 2013) complements PPCME2 remarkably well.
  - Built from manuscripts.
  - Extensive metadata.
  - Lemmatized.
  - Significant amounts of material from 1250–1325.

- A major motivation for constructing PLAEME is that LAEME complements PPCME2 so well.

- By transferring features from LAEME, we also have the opportunity to expand on the PPCHE format.
Section 2

LAEME
Compilation began early '90s; published online 2008

Specimens of written English 1150 – 1325:

- Documents, e.g. Chertsey Cartulary
- Prose, e.g. *Vices and Virtues*
- Poetry, e.g. *Poema Morale*
- Lyrics, e.g. *Love Song of Our Lady*

- c.650,000 words (shortest 18; longest 30,237)
Transcription

- Upper case for 'plain text' ms letters
- * for capitals
- Lower case for special characters:
  - $y = \beta$, $d = \delta$, $w = \rho$, $ae = \ae$, $g = \varsigma$, $z = \tilde{z}$

MS: <For it is caldore ner $\beta e$ se> (The Infancy of Christ)

LAEME: *FOR IT IS CALDORE NER $\gamma E$ SE
Tagging

- Most morphemes (incl. inflections & affixes) are tagged
- $\text{lexel/grammel\_form}$, e.g. $\text{for/cj\_*FOR}$
  - lexel: semantic identifier
  - grammel: grammatical identifier
- All tags can be translated into lemmas
- Word-level tags arranged vertically in tagged text files

$\text{for/cj\_*FOR}$
$\text{/P13NI\_IT}$
$\text{be/vps13\_IS}$
$\text{cold/aj-cpv\_CALD+ORE}$
$\text{-er/xs-cpv-aj\_+ORE}$
$\text{near/pr\_NER}$
$\text{/T<pr\_yE}$
$\text{sea/n<pr\_SE}$
Syntactic clues: NPs

$near/pr\_NER$

$/T<pr\_yE$

$sea/n<pr\_SE$

- $/T(...) = \text{Definite determiner}$
- $xxx/n(...) = \text{Noun}$
- $/T(...) + xxx/n(...) = \text{NP}$
Syntactic clues: NP function

$near/pr\_NER
$/T<pr\_yE
$sea/n<pr\_SE

- <pr = ‘in the scope of a preceding preposition’
- $/T<pr + $xxx/n<pr = NP object of P
Syntactic clues: NP function

<To teche þe volk þe riȝte lawe>
‘to teach the folk the right law’

$to/im+C_*TO$
$teach/vi-m_TECH+E $/vi-m+E$
$/TOi_yE$
$folk/nOi_VOLK$
$/TOd_yE$
$right/ajOd_RIzTE$
$law/nOd_LAWE$

- $/T(...) + $folk/n(...)$ = NP
- Oi in grammel indicates indirect object function
Syntactic clues: NP function

- $to/im+C_*TO$
- $teach/vi-m_TECH+E$ $/vi-m+_E$
- $/TOi_yE$
- $folk/nOi_VOLK$
- $/T0d_yE$
- $right/aj0d_RIzTE$
- $law/n0d_LAWE$

- $/T(...)$ + $right/aj(...)$ + $law/n(...)$ = NP
- Od in grammel indicates direct object function
Non-local dependency: discontinuous XPs

<... ḍat ghe ne migte **him** bringen **on**>
'... which she may not prove **against** him'

- >pr = ‘object of a following preposition’
- pr< = ‘preposition with a preposed object’

$/RTIOd\_dAT$
$/P13NF\_GHE$
$/neg-v\_NE$
$/may/vpt13\_MIGTE$
$/P13>prM\_HIm$
$/bring/vi\_BRING+EN$ $/vi\+_EN$
$/on\{p\}/pr<\{rh\}\_ON$
Non-local dependency: 'braced' negation

- >= points forward to a non-contiguous coordinating item
- <= points back to a non-contiguous coordinating item

<br>

<thou ne sselt naȝt consenti ...>
'thou shall not consent ...'

$/P12N_yOU$
$/neg-v>=_NE$
$shall/vps12_SSELF$
$/not/neg-v<=_NAzT$
$consent/viFir_conSENT+I $/viFir_+I$
LAEME: Text selection

Prioritised texts:

1. 1250 - 1325
2. > 100 words
3. No pre-existing parsed version

Multiple versions of same text:

4. Choose on basis of date, length and county
Our top 8

<table>
<thead>
<tr>
<th>Text</th>
<th>#</th>
<th>Date</th>
<th>Words</th>
<th>County</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. English Legendary</em></td>
<td>286</td>
<td>ca.1310-20</td>
<td>30,237</td>
<td>Berks</td>
</tr>
<tr>
<td>Nthn Homily Coll’n</td>
<td>298</td>
<td>C14a</td>
<td>22,164</td>
<td>Yorks NR</td>
</tr>
<tr>
<td><em>Havelok</em></td>
<td>285</td>
<td>C14a1</td>
<td>17,089</td>
<td>Norfolk</td>
</tr>
<tr>
<td><em>Cursor Mundi, Hand A</em></td>
<td>297</td>
<td>C14a</td>
<td>15,107</td>
<td>Yorks ER</td>
</tr>
<tr>
<td><em>Cursor Mundi, Hand C</em></td>
<td>296</td>
<td>C14a</td>
<td>14,087</td>
<td>York</td>
</tr>
<tr>
<td><em>Infancy of Christ</em></td>
<td>283</td>
<td>ca.1300</td>
<td>12,489</td>
<td>Oxon</td>
</tr>
<tr>
<td><em>Genesis and Exodus</em></td>
<td>155</td>
<td>C14a1</td>
<td>12,467</td>
<td>Norfolk</td>
</tr>
<tr>
<td><em>Life of Christ</em></td>
<td>281</td>
<td>ca.1300</td>
<td>10,547</td>
<td>Oxon</td>
</tr>
</tbody>
</table>

These texts alone supplement PPCHE by 134,187 words
Section 3

The parsing process
Distinctive properties

- Unusual parsing approach:
  - LAEME grammels stand in a many–many relationship with PPCHE POS tags, so will not be directly visible in the output.
  - Nevertheless, LAEME grammels contain a large amount of syntactically relevant information:
    - Category
    - Grammatical function
    - Some meaning relations
    - Some nonlocal dependencies
      - `$/neg-v([neither]<av>norC>nor>norC)_NE`
        (laud108bt.tag, line 1161)
  - Plan in essence: project the LAEME grammels into labelled brackets, while also replacing them with PPCHE tags.
  - Allows for fairly accurate and very detailed automatic structure generation.
Workflow: Format conversion

1. Take LAEME text
2. Store initial metadata as separate file.
3. For lines with textual material (initial character $, ’, ;):
   3.1 Find full word form, most informative grammel, etc.
   3.2 Segment that material into $lexel/grammel_form$
   3.3 Convert LAEME orthography into PPCHE orthography.
   3.4 Add lemmata for function words (no lexels in LAEME).
   3.5 Reformat as (POS@grammel@ form-lexel), where POS is a best-guess PPCHE equivalent of the LAEME grammel, with some lexel-by-lexel correction for e.g. different treatment of relativizers, subordinating conjunctions, etc.).
   3.6 Split compound forms (ifthat, nolde, shalþu, etc.), retag, etc.
4. For other nodes (detailing deletion, insertion, commentary, etc., also punctuation), figure out an appropriate tag (e.g. COMMENT, INS), reformat slightly, and pass to output file.
5. Use punctuation as (very) rough guide to sentence boundaries, insert IP-MAT brackets accordingly.
So far
Workflow: Parsing

1. Use corpus revision queries to incrementally add/adjust bracketing, relabel, etc.
2. Transfer rhyme information to word-level tags (not yet implemented).
3. Strip out LAEME grammels.
4. Manually correct automatically generated parse.
Sample corpus revision query

node: $METAROOT

query: (\{1\}NP-SBJ iPrecedes \{2\}IP-NoSbj) AND (NP-SBJ HasSister IP-NoSbj)

extend_span\{2,1\}:
truncat\{label\{1\}: -}
Hand-correction: Annotald
Section 4

What next?
Prospects

- Initial grant covers $\frac{1}{3}$ of LAEME.
- Longer-term goals: the rest of LAEME, and LAOS.
- Practical issues:
  - Up-to-speed annotators can correct 4-500 words/hour, so working through the rest of LAEME and LAOS involves c.1500 hours of someone’s time (Jim’s?).
  - Any way to “reuse parses” across parallel texts? Rough calculation: c.390k words of LAEME are versions of texts already parsed in PPCME2/PCMEP or slated for inclusion in our first sample ($\approx 87\%$ of the remainder of LAEME).
- Potential for new research uses of parsed historical corpora?
  - Traditionally, PPCHE people have had limited interest in dialectal variation.
  - Our initial choice of texts for parsing was motivated by helping PPCHE people to find fuller answers to PPCHE-style questions.
  - Any hope for using parallel texts in a fully parsed version of LAEME to investigate dialectal variation in syntactic structure?

