

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?

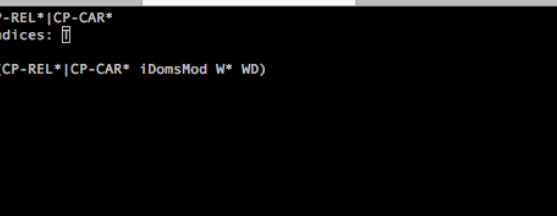
Roadmap

1. Introduction to PLAEME.
2. Replication of classic ME dialect syntax results from Kroch & Taylor (1997).

Background: PPCME2

- ▶ 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

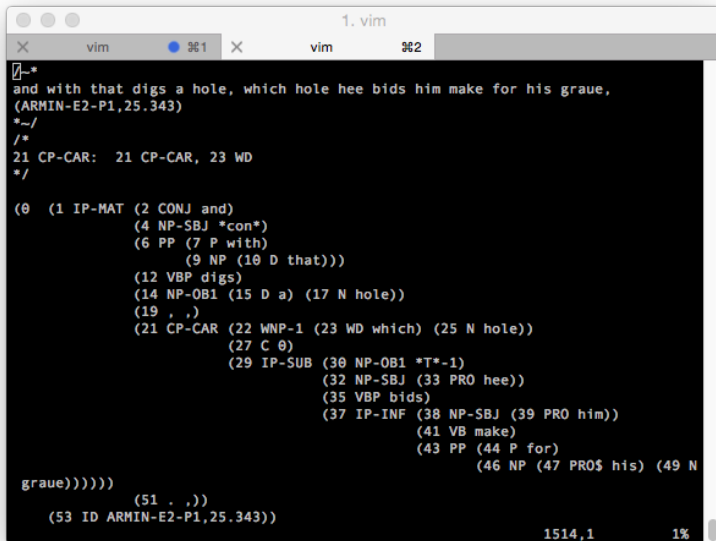


The screenshot shows a vim editor window with the title "1. vim". The editor has two tabs: "vim" (active) and "vim". The active tab shows a SQL query and its execution results. The query is:

```
node: CP-REL*|CP-CAR*
print_indices: [ ]
query: (CP-REL*|CP-CAR* iDomsMod W* WD)
```

The execution results are displayed as a list of 16 lines, each starting with a tilde (~). The bottom status bar shows the file name "WhNRel.q", the line and column number "4L, 80C", and the total number of lines "2,16".

Sample query: sample output

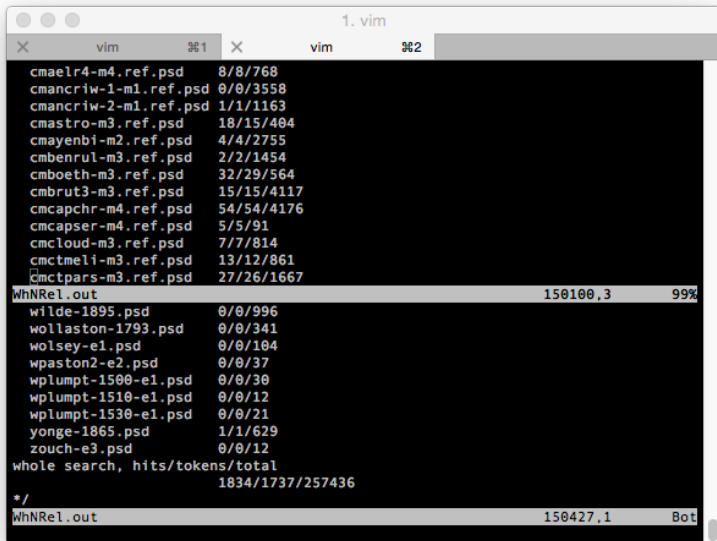


```
vim 1 2
and with that digs a hole, which hole hee bids him make for his graue,
(ARMIN-E2-P1,25.343)
*/
/*
21 CP-CAR: 21 CP-CAR, 23 WD
*/

(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))

1514,1 1%
```

Sample query: counts

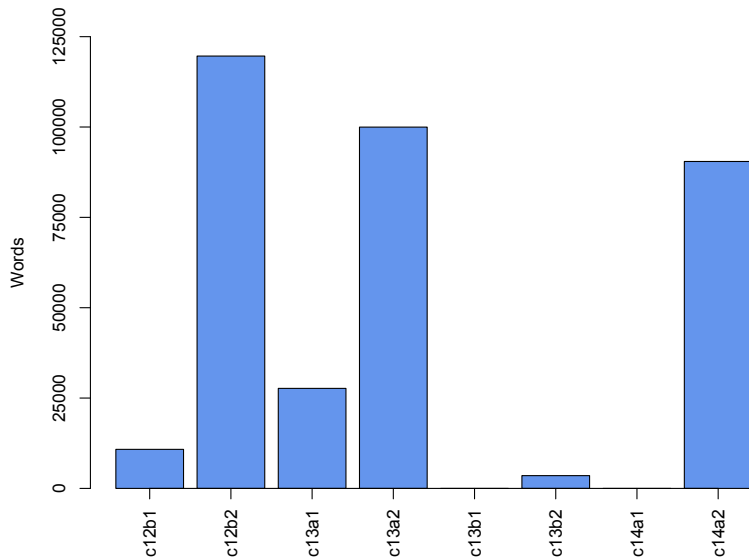


```
1. vim
vim 1 vim 2
cmaelr4-m4.ref.psd 8/8/768
cmancriw-1-m1.ref.psd 0/0/3558
cmancriw-2-m1.ref.psd 1/1/1163
cmastro-m3.ref.psd 18/15/404
cmayenbi-m2.ref.psd 4/4/2755
cmbenrul-m3.ref.psd 2/2/1454
cmboeth-m3.ref.psd 32/29/564
cmbrut3-m3.ref.psd 15/15/4117
cmcapchr-m4.ref.psd 54/54/4176
cmcapser-m4.ref.psd 5/5/91
cmcloud-m3.ref.psd 7/7/814
cmctmeli-m3.ref.psd 13/12/861
cmctpars-m3.ref.psd 27/26/1667
WhNRel.out 150100,3 99%
wilde-1895.psd 0/0/996
wollaston-1793.psd 0/0/341
wolsey-e1.psd 0/0/104
wpaston2-e2.psd 0/0/37
wplumpt-1500-e1.psd 0/0/30
wplumpt-1510-e1.psd 0/0/12
wplumpt-1530-e1.psd 0/0/21
yonge-1865.psd 1/1/629
zouch-e3.psd 0/0/12
whole search, hits/tokens/total
*/ 1834/1737/257436
WhNRel.out 150427,1 Bot
```

The data gap: PPCME2, 1150–1350

Filename	Title	Date	Words
cmkenth	Kentish Homilies	c12a2–b1	4048
cmpeterb	Peterborough Chronicle	c.1131, c.1154	6757
cmlambx1	Lambeth Homilies	c12b2	20752
cmtrinit	Trinity Homilies	c12b2	41844
cmorm	Ormulum	c12b2	50579
cmlamb1	Lambeth Homilies	c12b2	6459
cmvices1	Vices and Virtues	c13a1	27677
cmsawles	Sawles Warde	c13a2	4111
cmhali	Hali Meiðhad	c13a2	8495
cmkathe	St. Katherine	c13a2	8699
cmjulia	St. Juliana	c13a2	6810
cmmarga	St. Margaret	c13a2	8069
cmancriw	Ancrene Riwle	c13a2	63790
cmkentse	Kentish Sermons	c13b2?	3534
cmayenbi	Ayenbite of Inwyte	1340	45944
cmearlps	Earliest Prose Psalter	c.1350	44521

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.

Building PLAEME from LAEME

Text selection

- ▶ 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.

Building PLAEME from LAEME

LAEME annotations

- ▶ 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

be iuele man 'the evil man'

$\$/\text{TN_yE}$

$\$/\text{evil/aj_IUELE}$

$\$/\text{man/n_MAN}$

(NP-SBJ (D +te-the)
 (ADJ iuele-evil)
 (N man-man))

ner be se 'near the sea'

$\$/\text{near/pr_NER}$

$\$/\text{T<pr_yE}$

$\$/\text{sea/n<pr_SE}$

(PP (P ner-near)
 (NP (D +te-the)
 (N se-sea)))

Automatic bracketing: Examples

đat ghe ne migte hĩ bringen on 'what she might not prove against him'

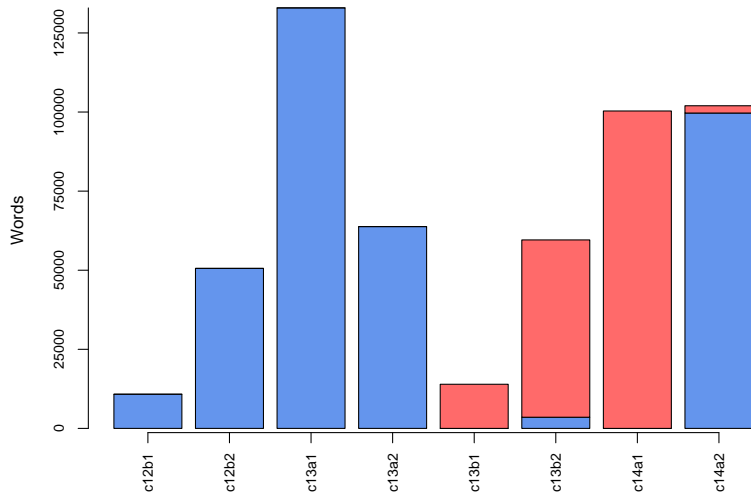
	(CP-REL (WNP 0)
	(C +dat-that)
\$/RTIOd_dAT	(IP-SUB (NP-OB1 *T*)
\$/P13NF_GHE	(NP-SBJ (PRO ghe-she))
\$/neg-v_NE	(NEG ne-ne)
\$may/vpt13_MIGTE	(MD migte-may)
\$/P13>prM_HIm	(NP (PRO hiM-him))
\$bring/vi_BRING+EN \$/vi_+EN	(VB bring+en-bring)
\$on{p}/pr<{rh}_ON	(PP (P-RH on-on)
	(NP *ICH*))))

Manual correction

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

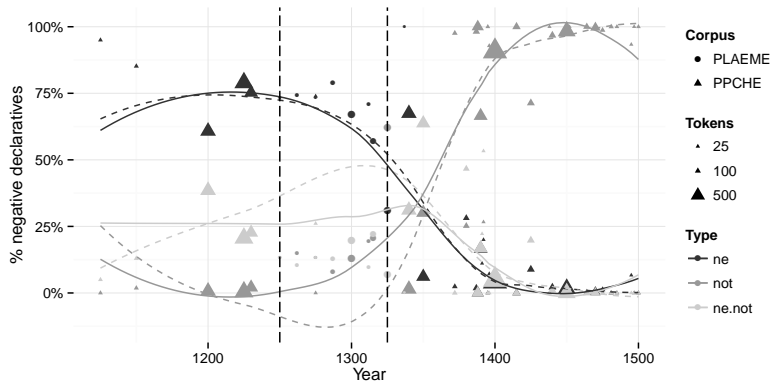
```
(CP-FRL (WNP-1 0)
  (C +dat-that)
  (IP-SUB (NP-OB1 *T*-1)
    (NP-SBJ (PRO ghe-she))
    (NEG ne-ne)
    (MD migte-may)
    (NP-2 (PRO hiM-him))
    (VB bring+en-bring)
    (PP (P-RH on-on)
      (NP *ICH*-2))))
```

PLAEME largely fills the gap in PPCME2



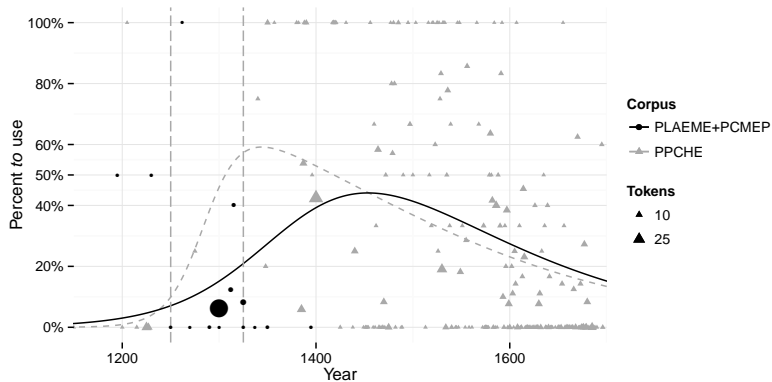
The extra data is helpful

Time course of Jespersen's Cycle in English



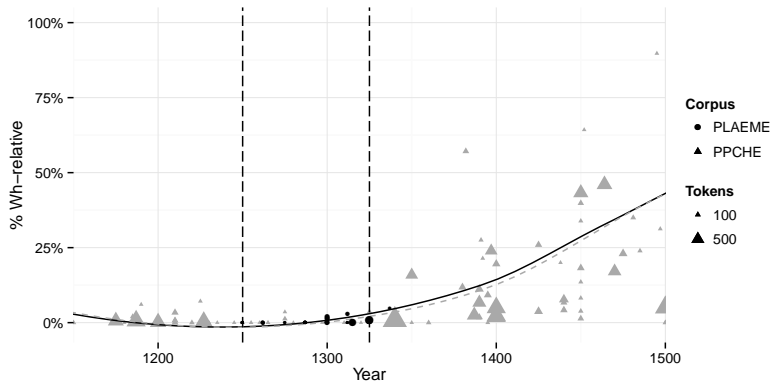
The extra data is helpful

Time course of recipient-theme ditransitives with *to*



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

	Southern (< 1250)		<i>Ayenbite</i> (1340)		cmbenrul (a1425)	
	NP	Pronoun	NP	Pronoun	NP	Pronoun
Preposed	% inv.	% inv.	% inv.	% inv.	% inv.	% inv.
NP compl.	93	5	82	8	100	95
PP compl.	75	0	100	0	100	100
Adj. compl.	95	33	100	0	100	67
<i>ba/then</i>	95	72	25	58	100	97
<i>now</i>	92	27	100	50	NA	100
PP adj.	75	2	36	3	89	91
Other adv.	57	1	56	10	96	91

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.

	Southern (< 1250)		<i>Ayenbite</i> (1340)		<i>cmbenrul</i> (a1425)	
	NP	Pronoun	NP	Pronoun	NP	Pronoun
Preposed	% inv.	% inv.	% inv.	% inv.	% inv.	% inv.
NP compl.	87	13	80	9	88	95
PP	76	15	78	6	96	86
Adj. compl.	94	27	100	0	NA	60
<i>pa/then</i>	93	80	42	57	100	96
<i>now</i>	75	17	75	37	NA	100
Other adv.	63	11	65	14	87	85

The 'southern' pattern

- (1) Efter þe þridde fiue ze 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,ll.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 haue 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:

```
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	

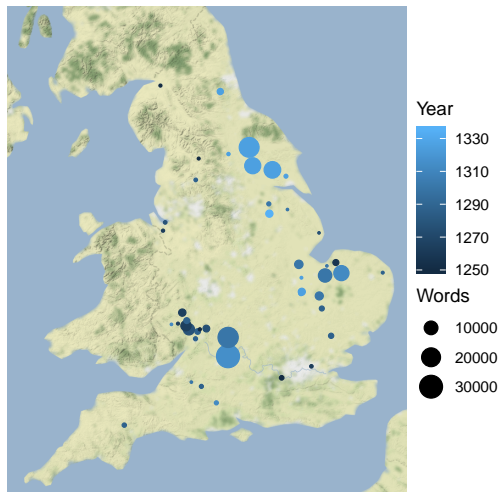
Sanity check II

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

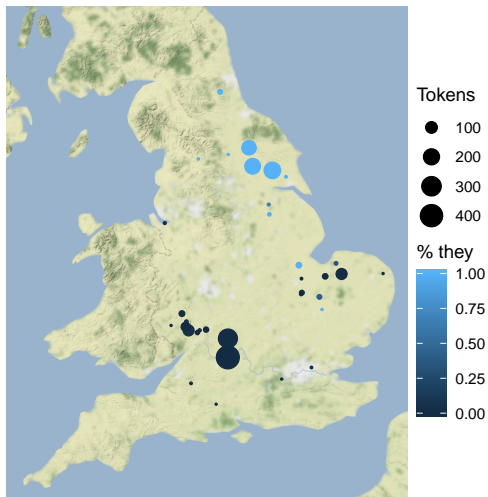
Can we detect nonsyntactic dialectal differences in PLAEME?



- ▶ 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.

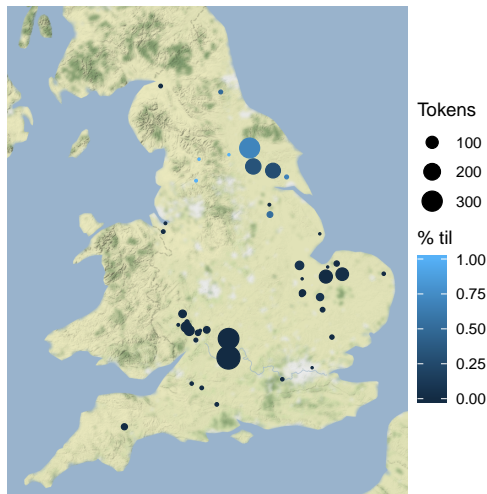
Sanity check III

They vs. *hi*



Sanity check III

To vs. *til*

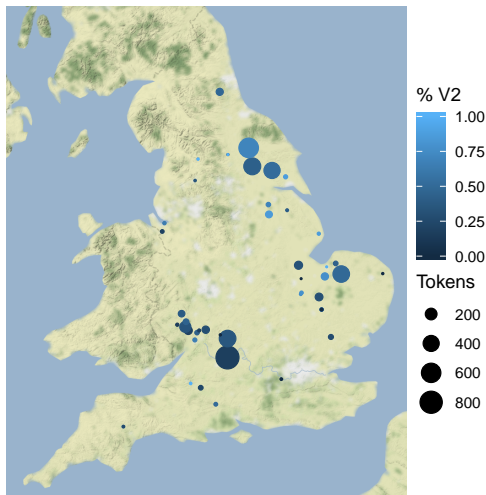


- ▶ Major differences in functional vocabulary are well-represented in PLAEME.
- ▶ (Though not every northern text robustly shows all 'northern' features).
- ▶ This should increase our hopes that regional differences w.r.t. V2 will be interpretable.

No general diachronic pattern across PLAEME

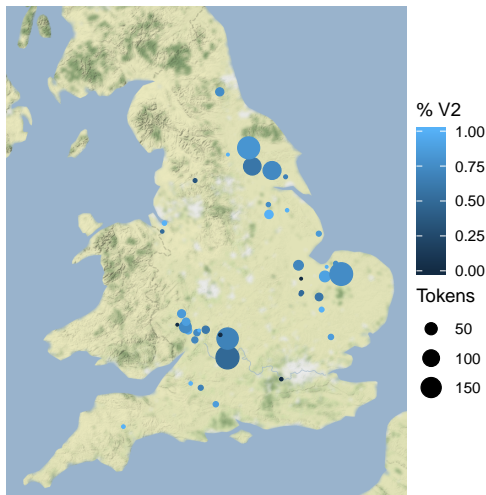
-

On to V2



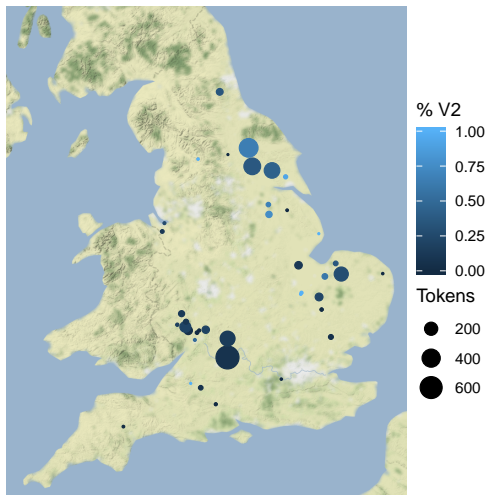
- ▶ Distribution of inversion in matrix clauses with one (or more) of the preposed elements identified by Kroch & Taylor.
- ▶ V2 concentrated in north.
- ▶ (All such statements supported by series of mixed-effects models, details skipped).

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.

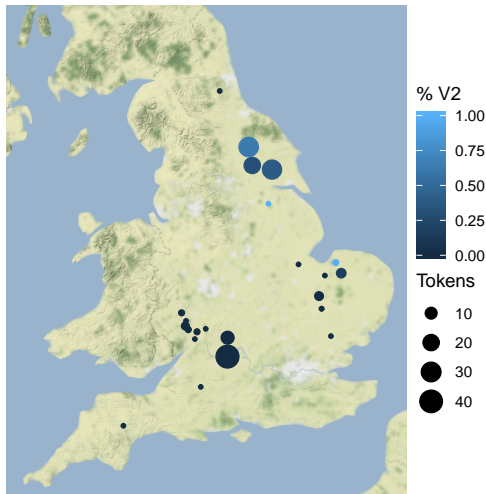
The distinctive pronoun pattern



- ▶ 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).

edincmXt 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 edincmXt 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	

Embedded V2 in edincmXt

- (7) For soruĩg al dũb war þai
for sorrowing all dumb were they

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

Na stand apõ þair fete
nor stand upon their feet

(edincmat.931)

- (8) For he suar bi þe kĩg of heuĩ
for he swore by the king of heaven

Pat harald slahtir suld he heuĩ
that Harold's slaughter should he avenge

(edincmat.1108)

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 lon
 A.stank it called is of Saint John

(edincmat.371)

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

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

(edincmat.509)

- (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

(edincmat.765)



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/atlasses?
 - ▶ LAOS?

Acknowledgements

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

- Beck, J., Ecay, A., & Ingason, A. K. (2011). Annotald, version 1.3.8.
<https://annotald.github.io/>.
- Kroch, A. & Taylor, A. (1997). Verb movement in Old and Middle English: Dialect variation and language contact. In A. van Kemenade & N. Vincent (Eds.), *Parameters of Morphosyntactic Change* (pp. 297–325). Cambridge: Cambridge University Press.
- Kroch, A. & Taylor, A. (2000). Penn-Helsinki parsed corpus of Middle English (2nd edition).
- Laing, M. (2002). Corpus-provoked questions about negation in Early Middle English. *Language Sciences*, 24, 297–321.
- Laing, M. (2013). A Linguistic Atlas of Early Middle English, 1150–1325. Version 3.2, <http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html>.
- Truswell, R., Alcorn, R., Donaldson, J., & Wallenberg, J. (2019). A Parsed Linguistic Atlas of Early Middle English. In R. Alcorn, J. Kopaczyk, B. Los, & B. Molineaux (Eds.), *Historical Dialectology in the Digital Age* (pp. 19–38). Edinburgh: Edinburgh University Press.
- Walkden, G. & Morrison, D. A. (2017). Regional variation in Jespersen's Cycle in Early Middle English. *Studia Anglica Posnaniensia*, 52, 173–201.
- Zimmermann, R. (2015). The Parsed Corpus of Middle English Poetry. University of Geneva.