UNIVERSITÄT MANNHEIM
Philosophische Fakultät

Enriching annotated corpora to identify contact-induced change: new tools and methods

Carola Trips (University of Mannheim)

Corpora and Diachrony: Influential Texts, Text Types, and Genres
Workshop of Athens Digital Glossa Chronos
November 26-29, 2025
Delphi, Greece

DFG Deutsche Forschungsgemeinschaft

FOR 5157: *Structuring the Input in Language Processing, Acquisition, and Change*
https://silpac.uni-mannheim.de/

© C. Trips 2021

Outline

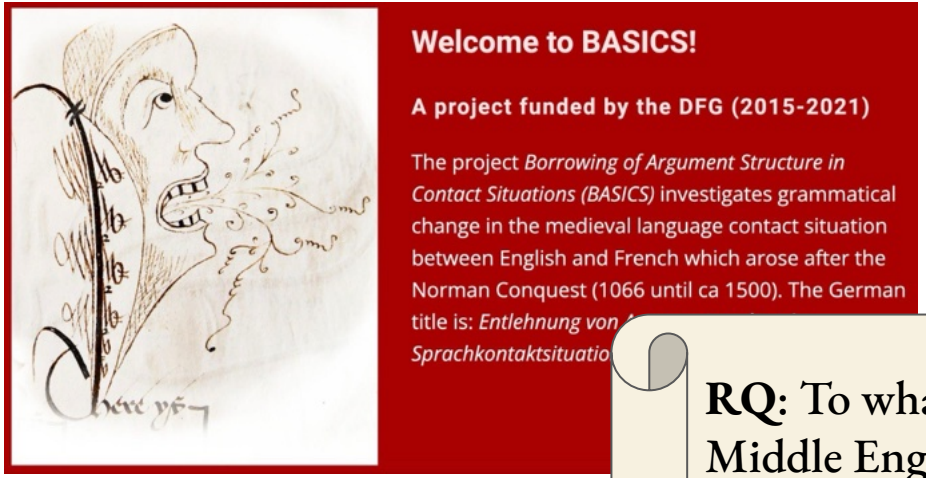
1. Introduction to the BASICS project: contact-induced change in medieval times (Old French, Middle English)
2. The format of the Penn Parsed Corpora
3. Enrichment of Penn annotation
4. Working with the enriched Middle English corpora
5. Using verb lemmatization and etymological information: two case studies
6. Conclusions

1. Contact-induced change in medieval times: Old French influence on Middle English grammar

DFG project
Borrowing of Argument Structure in Contact Situations
(BASICS, 2015-2021)
(Link to GEPRIS)

BASICS Toolkit
combines the software tools and resources created and used in the framework of the project.

1. Contact-induced change in medieval times: Old French influence on Middle English grammar



Welcome to BASICS!

A project funded by the DFG (2015-2021)

The project *Borrowing of Argument Structure in Contact Situations (BASICS)* investigates grammatical change in the medieval language contact situation between English and French which arose after the Norman Conquest (1066 until ca 1500). The German title is: *Entlehnung von*
Sprachkontaktsituatio

RQ: To what extent was the argument structure of Middle English (ME) verbs affected by the borrowing of Old French (OF) verbs, and with that English grammar?

Hypothesis: massive borrowing of OF verbs has at least accelerated grammatical change (if not initiated it).

1. Contact-induced change in medieval times: Old French influence on Middle English grammar

We work(ed) with syntactically annotated and verb lemmatized Penn corpora of Middle English

- the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME2, release 4); 1.2 mil. words, 56 texts
- the *Parsed Linguistic Atlas of Early Middle English* (PLAEME); c. 173,000 words, 68 texts
- the *Parsed Corpus of Middle English Poetry* (PCMEP), 216,649 words, 51 texts

and added additional layers of annotation:

- etymology of all lexical verbs (e= French/non-French origin)

We also annotated the meta data according to origin of text base (en=native, fr=french) that we took from the descriptions of the Penn corpora

1. Contact-induced change in medieval times: Old French influence on Middle English grammar

Title	Period	MsDateSimplified	Base	Title	Period	MsDateSimplified	Base
Aelred of Rievaulx's De Institutione Inclusarum (Vernon ms.)	M23	1400	en	St. Katherine	M1	1225	en
Aelred of Rievaulx's De Institutione Inclusarum (Bodley ms.)	M4	1450	fr	The Book of Margery Kempe	M4	1450	en
Ancrene Riwle	M1	1230	en	Kentish Homilies	M1	1150	en
Treatise on the Astrolabe	M3	1450	en	Kentish Sermons	M2	1275	fr
Ayenbite of Inwyt	M2	1340	fr	Lambeth Homilies	M1	1225	en
The Rule of St. Benet	M3	1425	en	Lambeth Homilies	MX1	1225	en
Boethius	M3	1425	fr	Malory's Morte Darthur	M4	1470	fr
The Brut or The Chronicles of England	M3	1400	fr	Mandeville's Travels	M3	1425	fr
Capgrave's Chronicle	M4	1464	en	St. Margaret	M1	1225	en
Capgrave's Sermon	M4	1452	en	Mirk's Festial	M34	1500	en
The Cloud of Unknowing	M3	1425	en	The New Testament (Wycliffe)	M3	1388	en
Tale of Melibee	M3	1390	fr	The Ormulum	M1	1200	en
The Parson's Tale	M3	1390	fr	The Old Testament (Wycliffe)	M3	1425	en
Earliest Prose Psalter	M2	1350	fr	Peterborough Chronicle	M1	1150	en
Life of St. Edmund	M4	1450	fr	John of Trevisa's Polychronicon	M3	1387	en
Mirror of St. Edmund (Thornton ms.)	M34	1440	en	Purvey's General Prologue to the Bible	M3	1388	en
Mirror of St. Edmund (Vernon ms.)	M3	1390	en	Caxton's History of Reynard the Fox	M4	1481	en
The Equatorie of the Planets	M3	1392	en	The Commonplace Book of Robert Reynes	M4	1470	en
Fitzjames' Sermo die Lune	M4	1495	en	Richard Rolle's Epistles	M24	1450	en
Dan Jon Gaytridge's Sermon	M34	1440	en	Richard Rolle's Prose Treatises	M24	1440	en
Gregory's Chronicle	M4	1475	en	Middle English Sermons, Royal Ms.	M34	1450	en
Hali Meidhad	M1	1225	en	Sawles Warde	M1	1225	en
Hilton's Eight Chapters on Perfection	M34	1450	en	The Siege of Jerusalem	M4	1500	en
A Late Middle English Treatise on Horses	M3	1450	en	The 'Liber de Diversis Medicinis' in Thornton Ms.	MX4	1440	en
In Die Innocencium	M4	1497	en	Trinity Homilies	MX1	1225	fr
St. Juliana	M1	1225	en	Vices and Virtues	M1	1225	en
Julian of Norwich's Revelations of Divine Love	M34	1450	en	The Book of Vices and Virtues	M34	1450	fr
				English Wycliffite Sermons	M3	1400	en

Text base (English, French) in the PPCME2

2. The format of the Penn Parsed corpora

https://penn-historical-corpora.uni-mannheim.de/

2. The format of the Penn Parsed corpora

Parsed corpora of other languages

This list is updated from time to time, but does not aim to be exhaustive. We hope it is useful nonetheless.

- Germanic
  - **Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE)** - Christina Tortora (City University of New York) and collaborators
  - **Corpus of Historical Low German** - Anne Breitbarth (University of Gent) and collaborators
  - **HeliPaD** (Old Saxon Heliand) - George Walkden (University of Konstanz)
  - **Icelandic Parsed Historical Corpus (IcePaHC)** - Eiríkur Rögnvaldsson (University of Iceland) and collaborators
  - **Indiana Parsed Corpus of Historical High German** - Chris Sapp (Indiana University) and collaborators
  - **Penn Parsed Corpus of Historical Yiddish** - Beatrice Santorini (University of Pennsylvania)
  - **The Parsed Corpus of Scottish Correspondence** - Lisa Gotthard (University of Edinburgh)
- Romance
  - **CORDIAL-SIN Corpus**, a syntax-oriented corpus of European Portuguese dialects - Ana Maria Martins (Centro de Linguística da Universidade de Lisboa) and collaborators
  - **Modéliser le changement: les voies du français (Modelling change: the paths of French)**, a parsed corpus of historical French - France Martineau (University of Ottawa) and collaborators
  - **Penn-BFM Parsed Corpus of Historical French** - Tony Kroch (University of Pennsylvania) and collaborators
  - **P.S. Post Scriptum - A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain)** - Rita Marquilhas (Centro de Linguística da Universidade de Lisboa) and collaborators
  - **Tycho Brahe Corpus**, a parsed corpus of historical European Portuguese - Charlotte Galves (University of Campinas, Brazil) and collaborators
  - **Word order and word order change in Western European languages (WOChWEL) Corpus**, a growing parsed corpus of Old Portuguese - Ana Maria Martins and Sandra Pereira (Centro de Linguística da Universidade de Lisboa) and collaborators
- Japanese
  - **NINJAL Parsed Corpus of Modern Japanese (NPCMJ)** - Prashant Pardeshi (National Institute of Japanese Language and Linguistics) and collaborators
  - **Oxford-NINJAL Corpus of Old Japanese (ONCOJ)** - Bjarke Frellesvig (Oxford University) and an **international committee**

2. The format of the Penn Parsed corpora

```
( (LATIN (FW Millesimo) (FW cxx=o=ii=o=) (. .))
  (ID CMPETERB,41.2))

( (IP-MAT (PP (P On)
          (NP (D +tis) (N geare)))
      (BED w+as)
      (NP-SBJ (D se)
          (N king)
          (NP-PRN (NPR Heanri)))
      (PP (P on)
          (NP (NPR$+NPR Cristesm+assan)))
      (PP (P on)
          (NP (NPR Norhtwic)))
      (. ,))
  (ID CMPETERB,41.3))
```


← linguistic annotation: POS tagging, syntactic tagging, syntactic parses (flat structure)

← tree structure (svg)

(illustration from Toolbox Anglistik https://anglistik-toolbox.uni-mannheim.de/)

2. The format of the Penn Parsed corpora

Three formats per text: txt (1), pos (2), psd (3)

1

```
On +tis geare w+as se king Heanri on Cristesm+assan on Norhtwic ,
CMPETERB,41.3
```

2

```
On/P
+tis/D
geare/N
w+as/BED
se/D
king/N
Heanri/NPR
on/P
Cristesm+assan/NPR$+NPR
on/P
Norhtwic/NPR
,/.
CMPETERB,41.3/ID
```

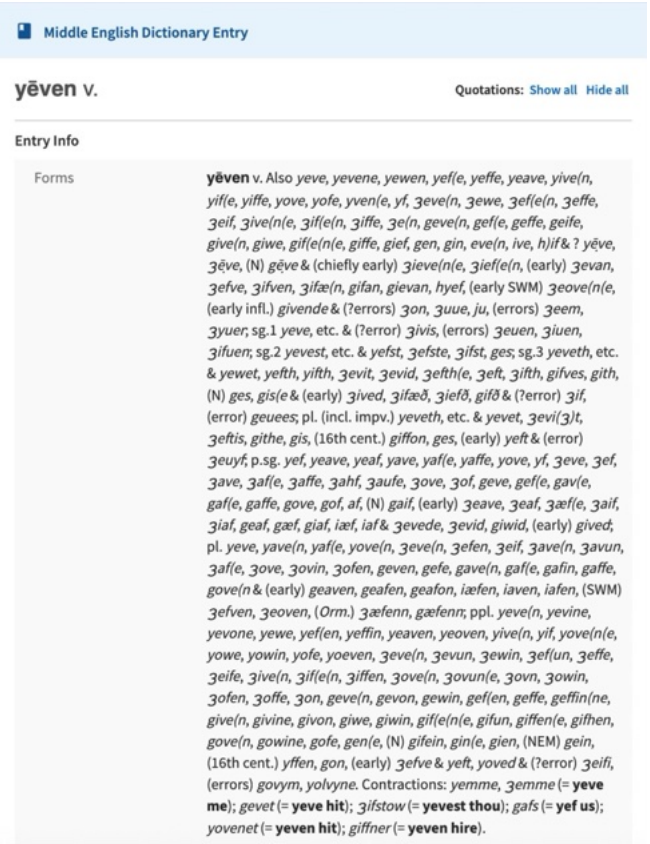
3

```
( (IP-MAT (PP (P On)
              (NP (D +tis) (N geare)))
          (BED w+as)
          (NP-SBJ (D se)
              (N king)
              (NP-PRN (NPR Heanri)))
          (PP (P on)
              (NP (NPR$+NPR Cristesm+assan)))
          (PP (P on)
              (NP (NPR Norhtwic)))
          (. ,))
  (ID CMPETERB,41.3))
```

3. Enriching the Penn Parsed corpora of Middle English with verb lemmatization and etymological information

Spelling variation of *yeven* according to the Middle English Dictionary (MED, Schaffner et al., 2018)

Why we need verb lemmatization



Middle English Dictionary Entry

yēven v.

Quotations: Show all Hide all

Entry Info

Forms

yēven v. Also *yeve, yevene, yewen, yef(e, yeffe, yeave, yive(n, yif(e, yiffe, yove, yofe, yven(e, yf, ȝeve(n, ȝewe, ȝef(e(n, ȝeffe, ȝeif, ȝive(n(e, ȝif(e(n, ȝiffe, ȝe(n, geve(n, gef(e, geffe, geife, give(n, giwe, gif(e(n(e, giffe, gief, gen, gin, eve(n, ive, h)if* & *? yēve, ȝēve,* (N) *gēve* & (chiefly early) *ȝieve(n(e, ȝief(e(n,* (early) *ȝevan, ȝefve, ȝifven, ȝifæ(n, gifan, gievan, hyef,* (early SWM) *ȝeove(n(e,* (early infl.) *givende* & (?errors) *ȝon, ȝuue, ju,* (errors) *ȝeem, ȝyuer;* sg.1 *yeve,* etc. & (?error) *ȝivis,* (errors) *ȝeuen, ȝiuen, ȝifuen;* sg.2 *yevest,* etc. & *yefst, ȝefste, ȝifst, ges;* sg.3 *yeveth,* etc. & *yewet, yefth, yifth, ȝevit, ȝevid, ȝefth(e, ȝeft, ȝifth, gifves, gith,* (N) *ges, gis(e* & (early) *ȝived, ȝifæð, ȝiefð, gifð* & (?error) *ȝif,* (error) *geuees;* pl. (incl. impv.) *yeveth,* etc. & *yevet, ȝevi(ȝ)t, ȝeftis, githe, gis,* (16th cent.) *giffon, ges,* (early) *yeft* & (error) *ȝeuyf;* p.sg. *yef, yeave, yeaf, yave, yaf(e, yaffe, yove, yf, ȝeve, ȝef, ȝave, ȝaf(e, ȝaffe, ȝahf, ȝaufe, ȝove, ȝof, geve, gef(e, gav(e, gaf(e, gaffe, gove, gof, af,* (N) *gaif,* (early) *ȝeave, ȝeaf, ȝæf(e, ȝaif, ȝiaf, geaf, gæf, giaf, iæf, iaf* & *ȝevede, ȝevid, giwid,* (early) *gived;* pl. *yeve, yave(n, yaf(e, yove(n, ȝeve(n, ȝefen, ȝeif, ȝave(n, ȝavun, ȝaf(e, ȝove, ȝovin, ȝofen, geven, gefe, gave(n, gaf(e, gafin, gaffe, gove(n* & (early) *geaven, geafen, geafon, iæfen, iaven, iafen,* (SWM) *ȝefven, ȝeoven,* (*Orm.*) *ȝæfenn, gæfenn;* ppl. *yeve(n, yevine, yevone, yewe, yef(en, yeffin, yeaven, yeoven, yive(n, yif, yove(n(e, yowe, yowin, yofe, yoeven, ȝeve(n, ȝevun, ȝewin, ȝef(un, ȝeffe, ȝeife, ȝive(n, ȝif(e(n, ȝiffen, ȝove(n, ȝovun(e, ȝovn, ȝowin, ȝofen, ȝoffe, ȝon, geve(n, gevon, gewin, gef(en, geffe, geffin(ne, give(n, givine, givon, giwe, giwin, gif(e(n(e, gifun, giffen(e, gifhen, gove(n, gowine, gofe, gen(e,* (N) *gifein, gin(e, gien,* (NEM) *gein,* (16th cent.) *yffen, gon,* (early) *ȝefve* & *yeft, yoved* & (?error) *ȝeifi,* (errors) *govym, yolvyne.* Contractions: *yemme, ȝemme* (= **yeve me**); *gevet* (= **yeve hit**); *ȝifstow* (= **yevest thou**); *gafs* (= **yef us**); *yovenet* (= **yeven hit**); *giffner* (= **yeven hire**).

3. Enriching the Penn Parsed corpora of Middle English with verb lemmatization and etymological information

Why we need verb lemmatization

```
65648  ( (IP-MAT-SPE (CONJ for)
65649             (NP-SBJ (PRO He))
65650             (HVP hath)
65651             (VBN yevyn)
65652             (NP-OB2 (PRO you))
65653             (NP-OB1 (NP (N beaute`))
65654                 (, ,)
65655                 (CONJP (NP (N bownte`)))
65656                 (, ,)
65657                 (CONJP (NP (N semelynes)))
65658                 (, ,)
65659                 (CONJP (CONJ and)
65660                    (NP (ADJ grete)
65661                    (N strengthe)
65662                    (PP (P over)
65663                       (NP (Q all) (OTHER other) (NS knyghtes))))))
65664             (. .))
65665    (ID CMMALORY,655.4476))
```

```
2010   ( (IP-MAT (ADVP-TMP (ADV Sone))
                 (VBD come)
                 (NP-SBJ (NPR Merlyn))
                 (PP (P unto)
                     (NP (D the) (N kyng))))
       (ID CMMALORY,5.135))
```

only POS information
"hidden" valency,
argument structure

3. Enriching the Penn Parsed corpora of Middle English with verb lemmatization and etymological information

Lemmatization and addition of etymological information

- **Verb lemmatisation** is based on a list of form-lemma correspondences
  - extraction of all verb forms of PPCME2 (and then the PLAEME and PCMEP) and manual assignment to lemmas from the MED that possess an individual ID.
- **Etymological information** was gained from an advanced search in the OED (extraction of all verbs between 1066 to 1500 with French as immediate etymon and possibly other etymons).
  (for further info see Percillier & Trips 2020, Percillier et al 2024)

3. Enriching the Penn Parsed corpora of Middle English with verb lemmatization and etymological information

```
/~*
Sone come@l=comen@m=8522@e=nonfrench@ Merlyn unto the
kyng@l=king_N|king@a=animate@
(CMMALORY,5.135)
*~/
/*
1 IP-MAT:  1 IP-MAT, 2 ADVP-TMP, 5 VBD, 7 NP-SBJ
*/

(0  (1 IP-MAT (2 ADVP-TMP (3 ADV Sone))
              (5 VBD come@l=comen@m=8522@e=nonfrench@)
              (7 NP-SBJ (8 NPR Merlyn))
              (10 PP (11 P unto)
                     (13 NP (14 D the) (16 N kyng@l=king_N|king@a=animate@))))
    (18 ID CMMALORY,5.135))

/~*
```

lemmatized verb

Lemmatisation of verb *comen*: @l=, MED-ID @m=, @e=nonfrench

3. Enriching the Penn Parsed corpora of Middle English with verb lemmatization and etymological information

```
/~*
So departed@l=departen@m=11123@e=french@ Merlyon,
(CMMALORY,34.1096)
*~/
/*
1 IP-MAT:  1 IP-MAT, 2 ADVP, 5 VBD, 7 NP-SBJ
*/

(0  (1 IP-MAT (2 ADVP (3 ADV So))
              (5 VBD departed@l=departen@m=11123@e=french@)
              (7 NP-SBJ (8 NPR Merlyon))
              (10 . ,))
    (12 ID CMMALORY,34.1096))

/~*
```

lemmatized verb

Lemmatisation of verb *departen*: @l=, MED-ID @m=, @e=french

4. Working with the enriched Middle English corpora

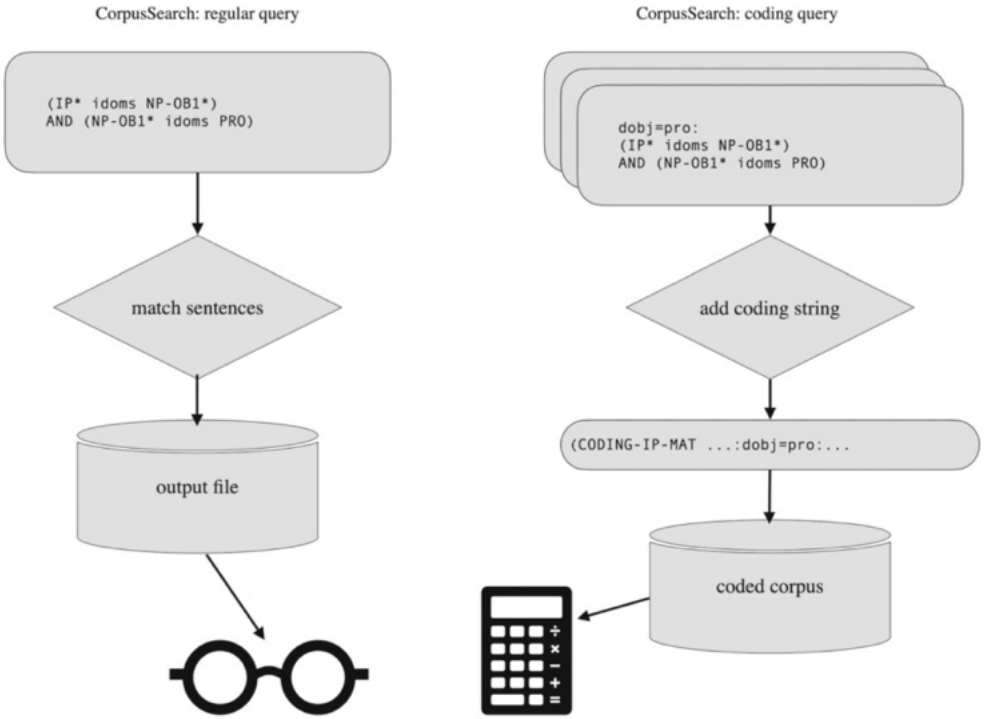


Fig. 2.1 CorpusSearch workflows: regular queries and coding queries

(Percillier et al 2024: 27)

4. Working with the enriched Middle English corpora

How to query the Penn corpora

- **Corpus Search: java programme**
- **runs in a terminal window**

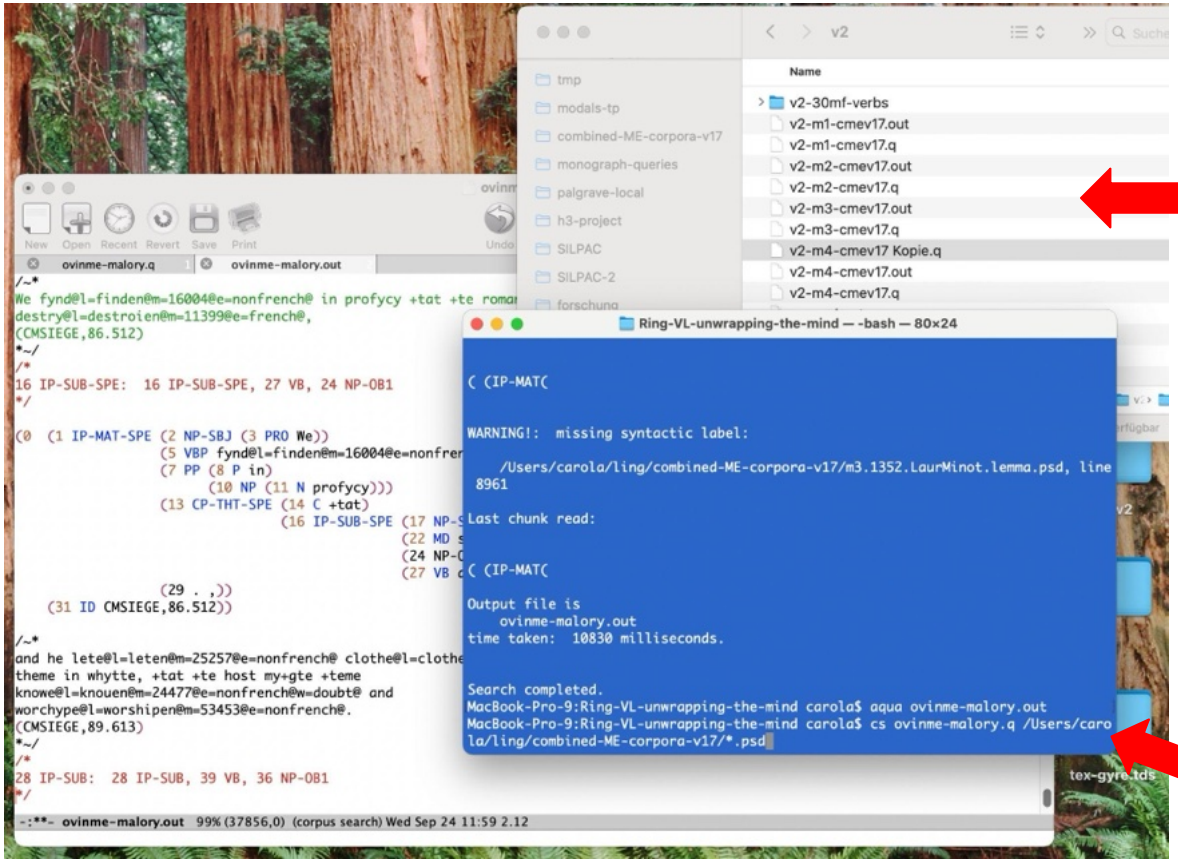
```
(V* iDominates .*)
AND (V* hasSister PP)
AND (PP iDominates !*ADV*)
AND (PP iDominates P)
AND (P iDominates {as|for|into|to})
```

(Percillier et al 2024: 29)

4. Working with the enriched Middle English corpora

Regular query

outfile of search



query files and output files

query command in shell

4. Working with the enriched Middle English corpora

Coding query

```
3: {
/* ------------------------------------------------
   subject
   lex includes tag ME (tag for 'men' used as impersonal subject)
   ------------------------------------------------ */
subj=pro: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms PRO)
subj=trace-exp: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms \*exp\*)
subj=trace: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms \**)
subj=lex: (IP* idoms NP-SBJ*)
subj=0: ELSE
}

4: {
/* ------------------------------------------------
   subject animacy
   - pronouns: only it is inanim, other pronouns are anim (TODO CHECK)
   - lexical: anim, inanim, panim
   ------------------------------------------------ */
subj-anim=inanim: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms PRO) AND (PRO idoms hit|it)
subj-anim=anim: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms PRO)
subj-anim=anim: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms .*) AND (.* idoms *@a=animate*)
subj-anim=inanim: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms .*) AND (.* idoms *@a=inanimate*)
subj-anim=panim: (IP* idoms NP-SBJ*) AND (NP-SBJ* idoms .*) AND (.* idoms *@a=panimate*)
subj-anim=0: ELSE
}
```

4. Working with the enriched Middle English corpora

Coding query

```
6: {
/* ------------------------------------------------
   direct object animacy
   - pronouns: only it is inanim, other pronouns are anim (TODO CHECK)
   - lexical: anim, inanim, panim
   ------------------------------------------------ */
dobj-anim=inanim: (IP* idoms NP-OB1*) AND (NP-OB1* idoms PRO) AND (PRO idoms hit|it)
dobj-anim=anim: (IP* idoms NP-OB1*) AND (NP-OB1* idoms PRO)
dobj-anim=anim: (IP* idoms NP-OB1*) AND (NP-OB1* idoms .*) AND (.* idoms *@a=animate*)
dobj-anim=inanim: (IP* idoms NP-OB1*) AND (NP-OB1* idoms .*) AND (.* idoms *@a=inanimate*)
dobj-anim=panim: (IP* idoms NP-OB1*) AND (NP-OB1* idoms .*) AND (.* idoms *@a=panimate*)
dobj-anim=0: ELSE
}

7: {
/* ------------------------------------------------
   indirect object
   ------------------------------------------------ */
iobj=pro: (IP* idoms NP-OB2*) AND (NP-OB2* idoms PRO)
iobj=lex: (IP* idoms NP-OB2*)
iobj=0: ELSE
}
```

4. Working with the enriched Middle English corpora

Coded corpora

transitive structure

```
26218  /~*
26219  and ye shall dey@l=dien@m=11564@e=nonfrench@ a worshipfull dethe. '
26220  (CMMALORY,35.1118)
26221  *~/
26222
26223  ( (IP-MAT-SPE (CODING-IP-MAT-SPE
       ipHead=verb:etym=e:subj=pro:subj-anim=anim:dobj=lex:dobj-anim=0:iobj=0:iobj-a
       nim=0:pobj=0:2pobj=0:pobj-anim=0:cobj=v-bare:spc=0:pass=0:reflinher=0:reflsel
       f=0:Neg=0:NegPosPre=0:NegPosPost=0:NegModOrder=0:iobjOrder=0:pobjOrder=0:preV
       erbPron=0)
26224              (CONJ and)
26225              (NP-SBJ (PRO ye))
26226              (MD shall)
26227              (VB dey@l=dien@m=11564@e=nonfrench@)
26228              (NP-OB1 (D a) (ADJ worshipfull) (N dethe))
26229              (. .)
26230              (' '))
26231    (ID CMMALORY,35.1118))
```

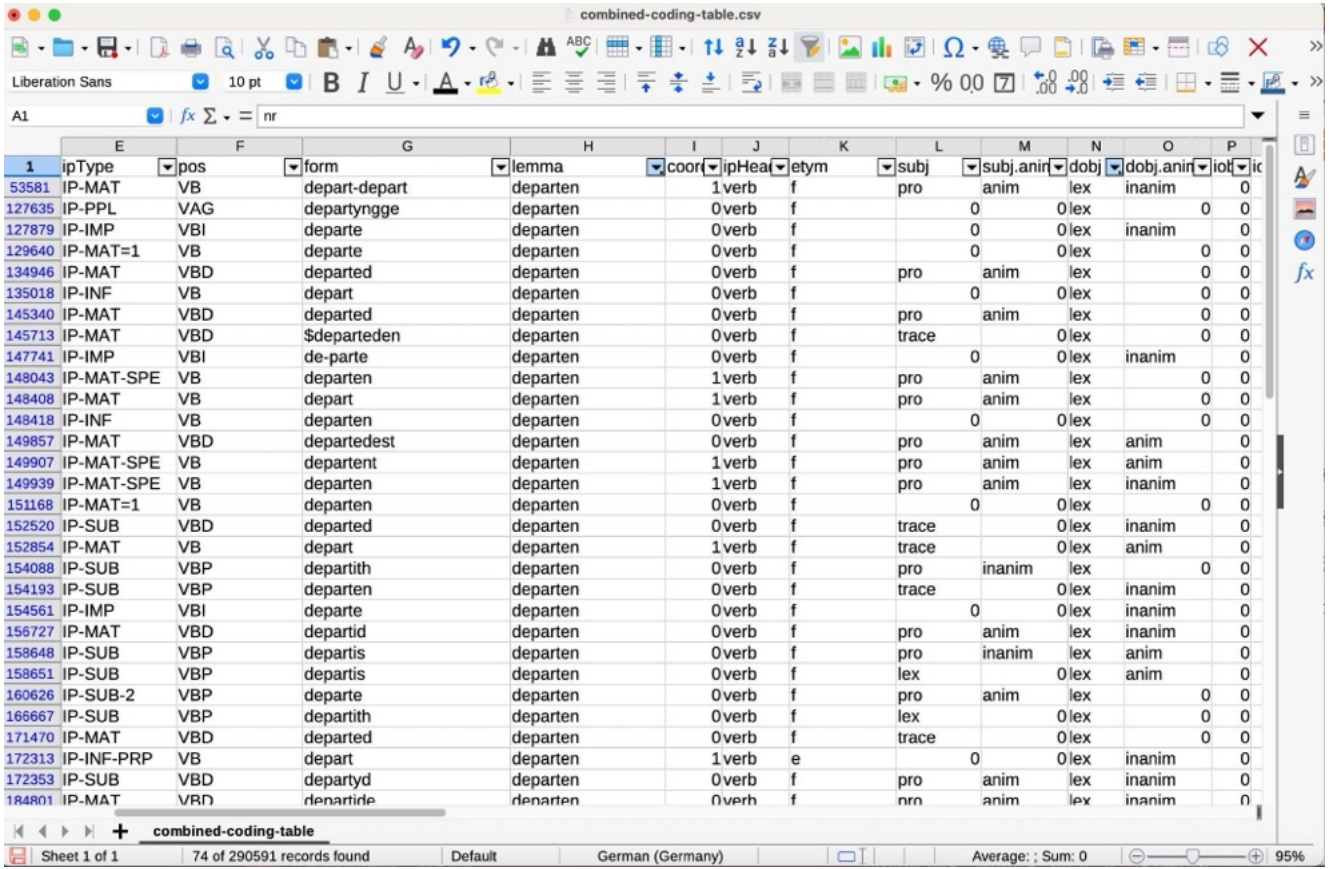
4. Working with the enriched Middle English corpora

Coded corpora intransitive structure

```
25721   So departed@l=departen@m=11123@e=french@ Merlyon,
25722   (CMMALORY,34.1096)
25723   *~/
25724
25725   ( (IP-MAT (CODING-IP-MAT
        ipHead=verb:etym=f:subj=lex:subj-anim=0:dobj=0:dobj-anim=0:iobj=0:iobj-anim=0
        :pobj=0:2pobj=0:pobj-anim=0:cobj=0:spc=0:pass=0:reflinher=0:reflself=0:Neg=0:
        NegPosPre=0:NegPosPost=0:NegModOrder=0:iobjOrder=0:pobjOrder=0:preVerbPron=0)
25726         (ADVP (ADV So))
25727         (VBD departed@l=departen@m=11123@e=french@)
25728         (NP-SBJ (NPR Merlyon))
25729         (. ,))
25730     (ID CMMALORY,34.1096))
```

4. Working with the enriched Middle English corpora

Coding table (csv) input to R etc.



	E	F	G	H	I	J	K	L	M	N	O	P
1	ipType	pos	form	lemma	coor	ipHea	etym	subj	subj.anin	dobj	dobj.anin	iob
53581	IP-MAT	VB	depart-depart	departen	1	verb	f	pro	anim	lex	inanim	0
127635	IP-PPL	VAG	departyngge	departen	0	verb	f		0	0	lex	0
127879	IP-IMP	VBI	departe	departen	0	verb	f		0	0	lex	inanim
129640	IP-MAT=1	VB	departe	departen	0	verb	f		0	0	lex	0
134946	IP-MAT	VBD	departed	departen	0	verb	f	pro	anim	lex	0	0
135018	IP-INF	VB	depart	departen	0	verb	f		0	0	lex	0
145340	IP-MAT	VBD	departed	departen	0	verb	f	pro	anim	lex	0	0
145713	IP-MAT	VBD	$departeden	departen	0	verb	f	trace		0	lex	0
147741	IP-IMP	VBI	de-parte	departen	0	verb	f		0	0	lex	inanim
148043	IP-MAT-SPE	VB	departen	departen	1	verb	f	pro	anim	lex	0	0
148408	IP-MAT	VB	depart	departen	1	verb	f	pro	anim	lex	0	0
148418	IP-INF	VB	departen	departen	0	verb	f		0	0	lex	0
149857	IP-MAT	VBD	departedest	departen	0	verb	f	pro	anim	lex	anim	0
149907	IP-MAT-SPE	VB	departent	departen	1	verb	f	pro	anim	lex	anim	0
149939	IP-MAT-SPE	VB	departen	departen	1	verb	f	pro	anim	lex	inanim	0
151168	IP-MAT=1	VB	departen	departen	0	verb	f		0	0	lex	0
152520	IP-SUB	VBD	departed	departen	0	verb	f	trace		0	lex	inanim
152854	IP-MAT	VB	depart	departen	1	verb	f	trace		0	lex	anim
154088	IP-SUB	VBP	departith	departen	0	verb	f	pro	inanim	lex	0	0
154193	IP-SUB	VBP	departen	departen	0	verb	f	trace		0	lex	inanim
154561	IP-IMP	VBI	departe	departen	0	verb	f		0	0	lex	inanim
156727	IP-MAT	VBD	departid	departen	0	verb	f	pro	anim	lex	inanim	0
158648	IP-SUB	VBP	departis	departen	0	verb	f	pro	inanim	lex	anim	0
158651	IP-SUB	VBP	departis	departen	0	verb	f	lex		0	lex	anim
160626	IP-SUB-2	VBP	departe	departen	0	verb	f	pro	anim	lex	0	0
166667	IP-SUB	VBP	departith	departen	0	verb	f	lex		0	lex	0
171470	IP-MAT	VBD	departed	departen	0	verb	f	trace		0	lex	0
172313	IP-INF-PRP	VB	depart	departen	1	verb	e		0	0	lex	inanim
172353	IP-SUB	VBD	departyd	departen	0	verb	f	pro	anim	lex	inanim	0
184801	IP-MAT	VBD	departide	departen	0	verb	f	pro	anim	lex	inanim	0

5. Using verb lemmatization and etyomological information: some case studies

Table 2.10 Lemmas and tokens, by origin and by argument structure

Valency	Lemmas e	Lemmas f	Lemma ratio f	Tokens e	Tokens f	Token ratio f
INTR	1,528	569	0.271	40,115	5,058	0.112
DO	1,719	756	0.305	43,497	8,583	0.165
DO.CL	664	362	0.353	6,489	1,712	0.209
DO.IO	334	84	0.201	4,771	414	0.080
DO.PP	330	159	0.325	3,045	508	0.143
IO	297	51	0.147	2,218	361	0.140
IO.CL	170	53	0.238	2,557	681	0.210
IO.PP	29	5	0.147	58	12	0.171
CL	696	285	0.291	17,078	1,784	0.095
PP	362	123	0.254	5,411	807	0.103
PP.CL	123	51	0.293	846	156	0.156
Total	6,252	2,498		126,085	20,076	

DO = direct object, IO = indirect object, PP = to-PP, CL = clause

(Percillier et al (2024): 37)

5. Using verb lemmatization and etyomological information: some case studies

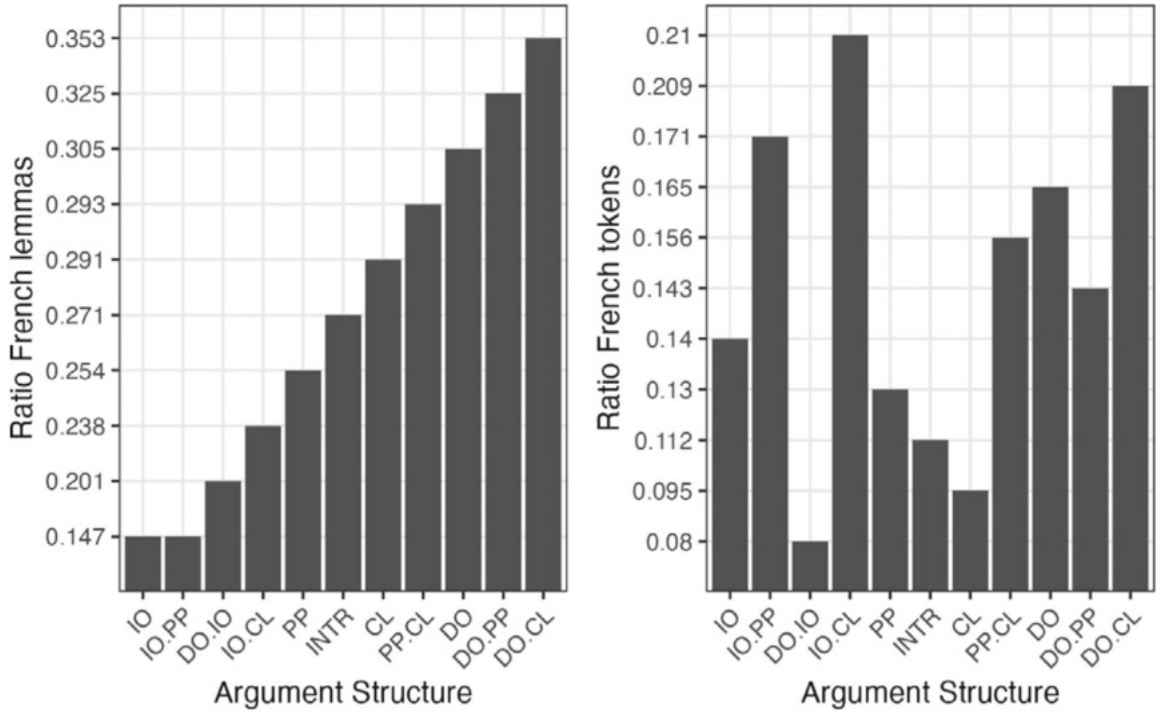


Fig. 2.5 Proportion of French in active verbs by argument structure

(Percillier et al (2024): 38)

5. Using verb lemmatization and etyomological information: two case studies

Study 1: Replication of Haeberli's (2018) study of statistical translation effects (OF influence on ME)

5. Using verb lemmatization and etyomological information: two case studies

Haeberli's (2018) study of statistical translation effects (OF influence on ME) without enriched annotation

Table 2. *The distribution of object pronouns and finite main verbs in Old and Middle English – outliers separated*

Periods	SOproV	SVOpro	Total
Old English	7,979 (81.8%)	1,774 (18.2%)	9,753
m1 1150–1250	467 (43.2%)	615 (56.8%)	1,082
m2 (1250–)1350	5 (1.6%)	304 (98.4%)	309
m2 Ayenbite, Kent. Sermons	*283 (88.2%)*	*38 (11.8%)*	*321*
m3 1350–1420	24 (1.9%)	1,242 (98.1%)	1,266
m3 Brut	*87 (32.8%)*	*178 (67.2%)*	*265*
m4 1420–1500	13 (1.0%)	1,336 (99.0%)	1,349
m4 Siege, Reynes	*28 (36.4%)*	*49 (63.6%)*	*77*

(1) (a) and he *him* **sæde** þas word: … (coaelhom,+AHom_2:276.387)
and he him said these words: …
'and he said these words to him: …'

(b) ær he *hit* **geleornige**. (cowulf,WHom_8c:144.662)
before he it learn
'before he learns it'

(from Haeberli 2018:304)

5. Using verb lemmatization and etyomological information: two case studies

Study 1: Replication of Haeberli's (2018) study of statistical translation effects (OF influence on ME)
based on

- more data (combined ME corpora)
- annotation of verb etymology, text base

Question: Is verb origin or text base the better predictor for disproportional increase of preverbal object pronouns (i.e. a statistical effect)?

5. Using verb lemmatization and etyomological information: two case studies

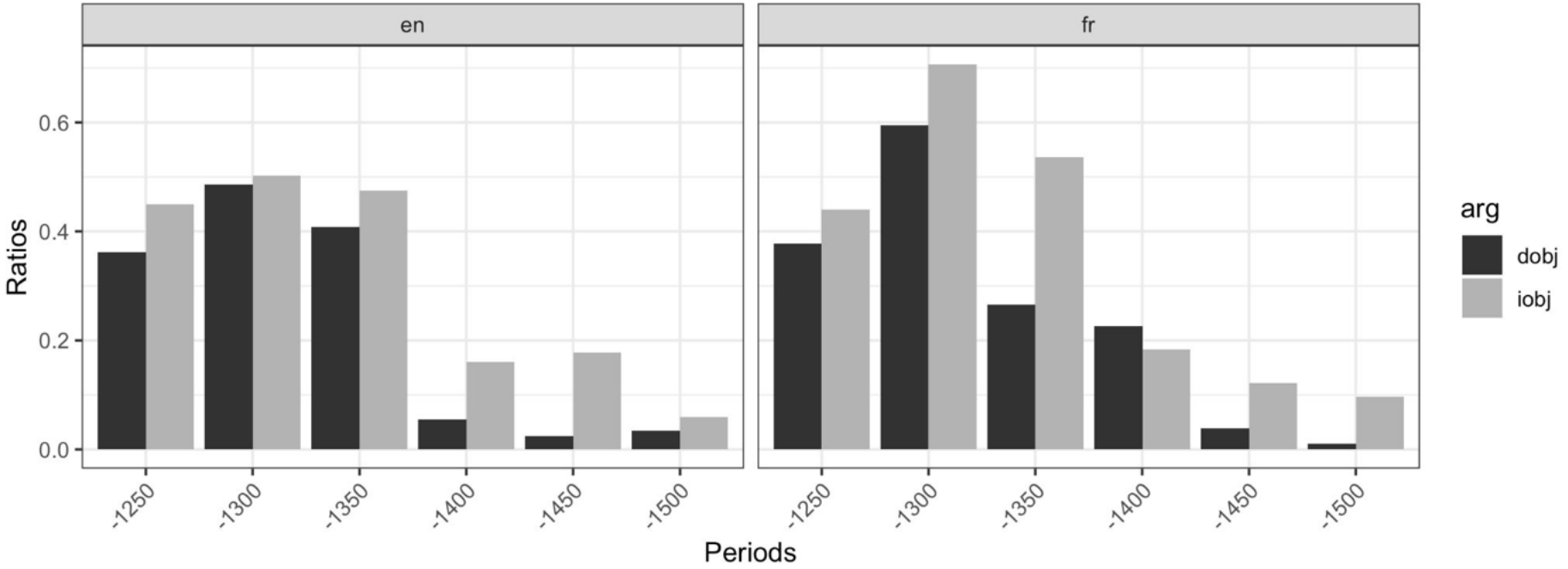


Fig. 9: Ratio of preverbal object pronouns in non-French based texts and French based texts in the combined ME corpora (Percillier et al 2024: 50)

5. Using verb lemmatization and etyomological information: two case studies

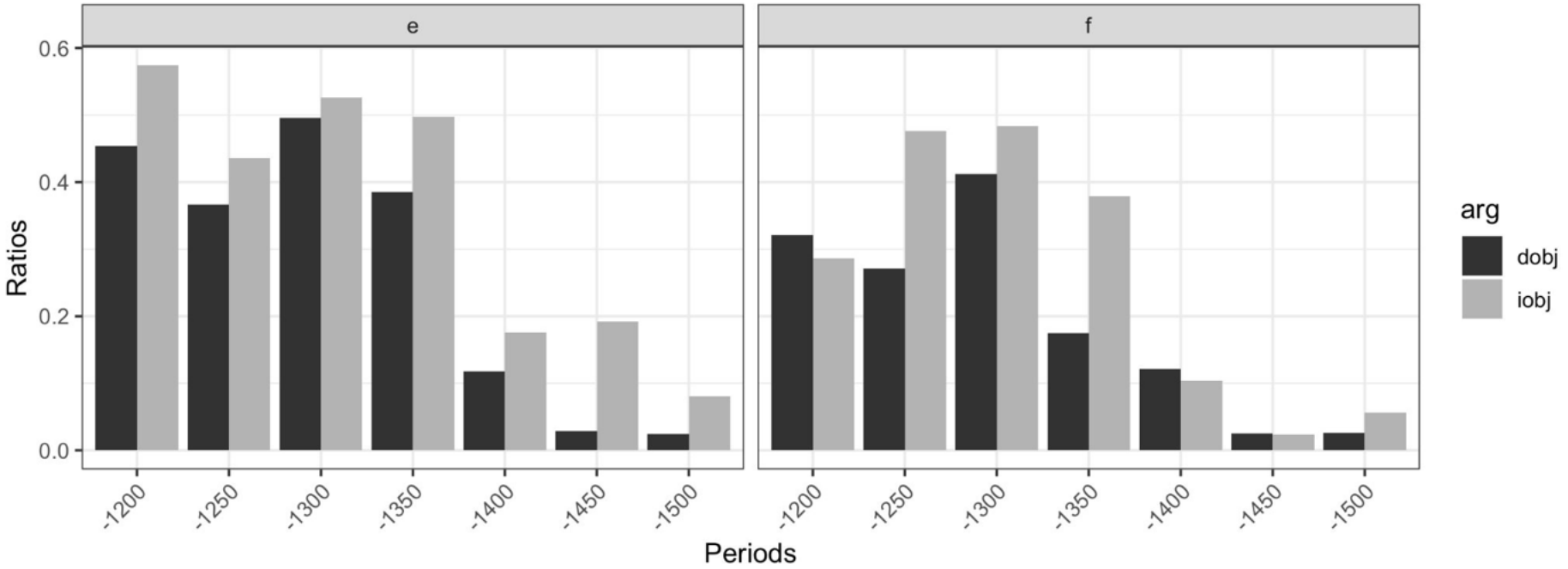


Fig.4: Ratio of preverbal object pronouns with non-French origin verbs and French origin verbs in the combined ME corpora (Percillier et al 2024: 51)

5. Using verb lemmatization and etyomological information: two case studies

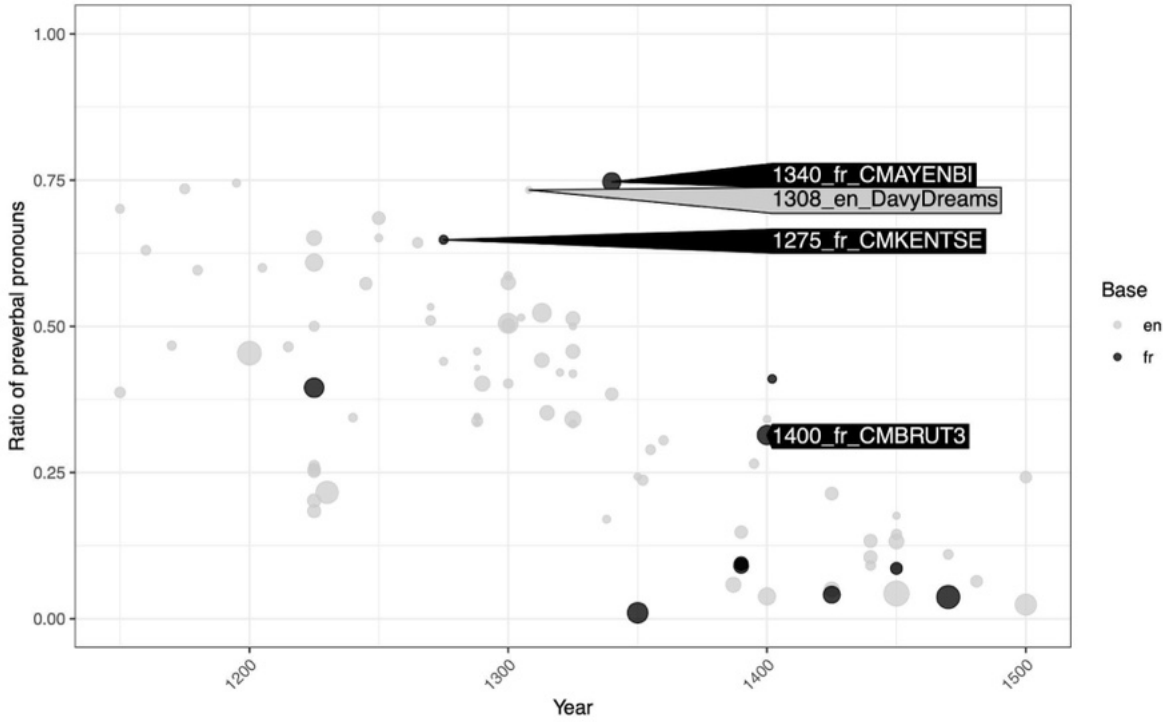


Fig.5: Text with high frequencies of preverbal objects ("outliers") in combined ME corpora

5. Using verb lemmatization and etyomological information: two case studies

Findings:
(i) Higher increase of preverbal object pronouns (esp. IOs) in French-based texts in between 1250 and 1375 (Fig.3)
(ii) IOs are overall more frequent than DOs (Fig.3, Fig.4)
(iii) Higher increase of preverbal object pronouns (esp. IOs) with non-French origin verbs than with French-origin verbs (Fig.4)

Explanation:
(i) and (ii) direct effect: OF preverbal object pronouns including OF dative clitics (in high position) are copied to preverbal position in French-based texts;
(iii) indirect effect: object pronouns of non-French origin verbs more often occur in preverbal position, in both translated and non-translated texts

=> Haeberli's findings are confirmed (Haeberli: full text analysis, BASICS: based on annotation)

5. Using verb lemmatization and etyomological information: two case studies

Study 2: The increase of labile verbs in Middle English under the contact hypothesis

5. Using verb lemmatization and etyomological information: two case studies

What is lability/a labile verb?

A verb may describe the same situation in both causal sense, with a causal agent expressed by the subject and the patient/experiencer expressed by the object of the verb, and inchoative sense, where the verb excludes a causal agent and describes the situation as occurring spontaneously. (Haspelmath 1993, 90-91)

(1) a. Tom broke the window.
b. The window broke.

=> non-directed/labile: same form of verb is used for both argument structures
=> change of state verbs, the new state (the broken window) is a result

(collaborative work with Achim Stein, Uni Stuttgart)

5. Using verb lemmatization and etyomological information: two case studies

Haspelmath (1993): English is quite unique in showing a strong preference for lability (also compared to the other Germanic languages)

(3) a. She **sank** the boat.
b. The boat **sank**.
c. Sie **versenkte** das Boot.
d. Das Boot **(ver)sank**.

(4) a. She **opened** the door.
b. The door **opened**.
c. Sie **öffnete** die Tür.
d. Die Tür **öffnete sich**.

Table 3. Expression types by language

	total	A	C	E	L	S	A/C	% non-dir.
Russian	31	23	0	5	0	3	46.00	26
German	31	14.5	0	4	11.5	1	29.00	53
Greek	31	13.5	0	0	16.5	1	27.00	56
Rumanian	30	24	1	0	3	2	24.00	17
French	31	20.50	2	0	7.5	1	10.25	27
Lithuanian	31	17.5	6	6	0.5	1	2.92	24
Hebrew	31	20.5	7.5	2	1	0	2.73	10
Arabic	31	17	8.5	3	1	1.5	2.00	18
Georgian	31	9	4.5	15.5	0	2	2.00	56
Armenian	31	16	8.5	5.5	0	1	1.88	21
Swahili	31	11	11	8	0	1	1.00	29
Finnish	28	12	13.5	0.5	0.5	1.5	0.88	9
Udmurt	31	10.5	12.5	4.5	2.5	1	0.84	26
Hungarian	31	7	9	12	0	3	0.78	48
Lezgian	31	8	12	6	5	0	0.66	35
Hindi-Urdu	31	7.5	14	7.5	2	0	0.54	31
Turkish	30	9	17.5	2.5	0	1	0.51	12
Mongolian	31	6	22	2	0	1	0.27	10
Indonesian	31	0	14	17	0	0	0.04	55
English	31	2	0	1	25	3		94
Japanese	31	3.5	5.5	20.5	0.5	1		71
total	636	243	164.5	128.5	69	310		

Abbreviations:
A = anticausative alternation
C = causative alternation
E = equipollent alternation
L = labile alternation
S = suppletive alternation
A/C = ratio of anticausative to causative pairs
% non-dir. = percentage of non-directed pairs

5. Using verb lemmatization and etyomological information: two case studies

Old English (OE) showed few labile verbs and used morphology and stem form alternation (as other Germanic languages do) (cf. García García 2020).

stem vowel alternation/morphologically derived

(5) a. *sencan* "sink" (transitive)
b. *sincan* "sink" (intransitive)

(6) a. *drygan* "dry" (transitive)
b. *adrygan/adruwian* "dry" (intransitive)

labile change-of-state verbs (no change of verb form)

(7) a. *bærnan* "burn" (transitive/intransitive)
b. *bigan* "bend" (transitive/intransitive)
c. *godian* "improve" (transitive/intransitive)

5. Using verb lemmatization and etyomological information: two case studies

In Middle English stem vowel alternations and morphologically derived patterns are lost leading to labile verb forms (8 a. and b.), some OE verbs remained labile (9 a.), some became labile (9 b.) and some are copies from OF (c.) (cf. Ingham 2020).

(8) a. *sink* "sink"
b. *drien* "dry"

(9) a. *breken* "break"
b. *cleven* "split"
c. *avauncen* "advance"

5. Using verb lemmatization and etyomological information: two case studies

How does language contact come in here?

- The contact situation between Old French (OF) and Middle English (ME) in medieval times (ca 1150-1500) led to copying of verbs and as a result changes in the grammar of English (cf. Ingham 2012, Percillier et al. 2024)
- Ingham (2020) on lability:
  - Lability was the replication of argument structure properties of French verbs; when OF verbs were copied to ME **their argument structure was copied along (full set)**.
  - **OF/AN exhibited many labile verbs** (change of state/location, cf. Heidinger 2010)
  - between the end of the 12th up to the 14th century individual **bilingualism** existed which favoured grammatical replication

5. Using verb lemmatization and etyomological information: two case studies

Patient/Theme of copied change of state/location verb was given the same alternating argument structural patterns as those of OF verbs:

OF :	{ NP	_____	NP	}
rostir 'roast'	Agent		Patient/Theme	
rouler 'roll'	{ NP	_____	}	
etc.	Patient/Theme			

ME :	{ NP	_____	NP	}
rosten	Agent		Patient/Theme	
rollen	{ NP	_____	}	
etc.	Patient/Theme			

(Ingham 2020: 461)

5. Using verb lemmatization and etyomological information: two case studies

ME rosten 'roast' copied from OF rostir 'roast'

OF: trans: [$_{NP}$ il] **rosterent** [$_{NP}$ phase] (=a sacrificial lamb) sur le feu (*Bible Royal* 300vb)
AGENT PATIENT

intrans: mettez [$_{NP}$ le gars] sur un espé [...] e lessez **roster** molt bien ... (*Five Med MSS* 117.E1)
PATIENT

ME: trans: By a mykel fir [$_{NP}$ he] sat, **Rostyng** [$_{NP}$ a swyn gret & fat].
AGENT PATIENT
(a1450(a1338) Mannyng Chron.Pt.1 (Lamb 131)12342)

intrans: Turbot..when [$_{NP}$ it] **rostis**, springle on salt.
PATIENT
(?a1475 Noble Bk.Cook.(Hlk 674)97)

5. Using verb lemmatization and etyomological information: two case studies

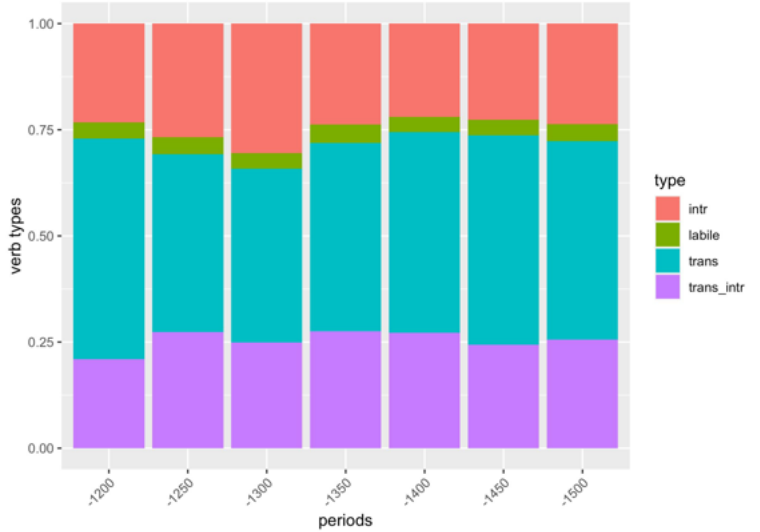
Strong version of the contact hypothesis: French labile verbs "drive the expansion of lability" in English. They are always copied with both structures.

Alternative hypothesis: Lability, and syntactic alternations in general, are rule-governed, i.e. licensed and constrained by the event structure ([[x CAUSE [y BECOME STATE]]; the structure is already there).

Based on our data we find that:

- French lability was not a sufficiently frequent cue for native speakers/learners → labile verbs don't "drive" change
- The English event structure was stable over time. Although French verbs formed a large proportion of the vocabulary, compared to English verbs they were not more liable to alternate.

5. Using verb lemmatization and etyomological information: two case studies



period	trans	intr	trans/intr	labile	labile_ratio
-1200	387	173	156	29	3,89 %
-1250	785	502	513	74	3,95 %
-1300	530	395	323	48	3,70 %
-1350	654	351	406	64	4,34 %
-1400	662	308	382	51	3,64 %
-1450	634	291	313	47	3,66 %
-1500	532	270	292	46	4,04 %

1. Automatic extraction of verbs occurring transitive or intransitive
2. min. 5 occurrences
3. labile = manual verification: causative-anticausative alternation (restricted to CoS verbs) attested in the combined ME corpora (PPCME2, PCMEP, PLAEME) or Middle English Dictionary (MED).
4. intr = always intransitive
5. trans = always transitive
6. trans_intr = both, but not labile

- number of labile verbs is stable
- productive rule of lability already existed in OE
- OF contributed to the increase but did not initiate the change

5. Using verb lemmatization and etyomological information: two case studies

Labile verbs copied from OF often **only show the causative transitive structure** in contrast to the native verb which already used to be labile in OE (event-structure governed lability)

- ME *curen* ('to cure' < OF *curer*)
  - causative 'to cure, restore to health': *He **cured Naaman**, þe prince of Surre', fro seknesse of lepre.* (PPCME2, CMCAPCHR,34.69)
  - anticausative not found in corpus
- ME *helen* ('to cure, heal' < OE *hælan*, labile) => compare German *heilen*
  - causative 'to cure, heal': *he tol=de is fader fore Hov a Naddre him hurte sore And hov Iesus **hel=de him*** 'he told his father how an adder severly hurt him and how Jesus healed him ...' (PLAEME, LAUD108AINFANCY.1154)
  - anticausative: *Þe seste is þe wunde þt eauer wurseð on hond & strengere is to **healen*** 'the sixth is the wound that ever worsens in the hand and is stronger to heal (PPCME2, CMANCRIW-1,II.242.3517)

5. Using verb lemmatization and etyomological information: two case studies

The alternative hypothesis is supported by language acquisition research (e.g. Scott & Fisher 2009).

- Child learners **do not receive sufficient evidence about verb lability** (alternations found in newspaper text are absent in child-directed speech).
- Despite this lack of evidence, children successfully acquire different verb alternations because:
  - they are sensitive to the semantics of verbs and argument roles (e.g. Causer, Patient)
  - they can detect if roles of subject and object overlap or not
  - they use these as **distributional features**

Scott & Fisher (2009):
"Can young children use **distributional features** to assign meanings to novel verbs that occur in the causal and unspecified-object alternations?"

causal alternation
(= causative-anticausative
= P-lability):
x moves y → *y moves*

unspecified-object alternation:
x pushes y → *x pushes*

5. Using verb lemmatization and etyomological information: two case studies

Figure 2. Training and test phases for the novel verb (Experiment 2). Test events: Contact-activity event (left) and caused-motion event (right).

(Scott & Fisher 2009: 790)

The results "add to a growing body of evidence that young children represent verbs somewhat independently of the sentential context in which they occur" (Scott & Fischer 2009:795).

→ Acquisition of lability does not require sentence-structure cues (i.e., the presence of both structures in the input data). It is not acquired "in one package".

5. Using verb lemmatization and etyomological information: two case studies

Summary of case study

- Labile verbs have always been part of the English language, i.e. the event structure for these change of state verbs (causative-anticausative, resultative) has been part of the grammar of English before contact with OF
- Labile verbs copied from OF increased the number to some degree but did not drive the change. (answer to question 3)
- Labile verbs increased because phonological/morphological marking was lost
- **Cognitive aspect**: resultativity is acquired by children when they learn this type of verb, it becomes part of their event structure ([[x CAUSE [y BECOME STATE]])
- children use distributional features (e.g. subject animacy) to learn the alternation => the link between cognitive capability and language acquisition is constant across time (answer to question 1)

5. Conlusions

- (Verb) lemmatization is tedious and time consuming but very helpful especially for languages that do not have a written standard
- We gain systematic insights into verb argument structure via verb lemmas and coding information of arguments
- Advantage of this approach: new texts can easily be added by recoding corpora


THE END OF
THE BEGINNING

References

García García, Luisa. 2020. The basic valency orientation of Old English and the causative *ja-* formation: A synchronic and diachronic approach. *English Language and Linguistics*, *24*(1), 153–177.

Haeberli, E. (2018). Syntactic effects of contact in translations: Evidence from object pronoun placement in Middle English. *English Language and Linguistics, 22(2)*, 301–321.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Comrie, Bernard & Polinsky, Maria (eds.), *Causatives and transitivity*, 87–120. Amsterdam: J. Benjamins.

Ingham, Richard. 2020. How Contact with French Drove Patient-Lability in English. Transactions of the Philological Society 118(3). 447–467.

Percillier, Michael & Schauwecker, Yela & Stein, Achim & Trips, Carola. 2024. *Carrying Verbs Across the Channel - Modelling Change in Bilingual Medieval England*. Palgrave Macmillan.

Percillier, M. & C. Trips. 2020. Lemmatising Verbs in Middle English Corpora: The Benefit of Enriching the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and a Parsed Linguistic Atlas of Early Middle English (PLAEME). *In Proceedings of the 12th Language Resources and Evaluation Conference*, 7172– 7180. Marseille, France: European Language Resources Association. https://www.aclweb.org/ anthology/2020.lrec- 1.886.

References

Scott, Rose M. & Fisher, Cynthia. 2009. Two-year-olds use distributional cues to interpret transitivity-alternating verbs. Language and Cognitive Processes 24(6). 777–803.

Taylor, A. 2008. Contact Effects of Translation: Distinguishing Two Kinds of Influence in Old English. *Language Variation and Change* 20. 341–365.

Trips, C. & A. Stein. Contact-induced changes in the argument structure of Middle English verbs on the model of Old French. In E. Grossman, I. Serzant, and A. Witzlack-Makarevich, editors, *Journal of Language Contact. Special Issue on Valency and Transitivity in Contact*, pages 232–267. Brill, Leiden, 2019.

Yang, Charles. 2016. *The price of linguistic productivity*. Boston, Mass: MIT Press.

Yang, Charles & Trips, Carola. 2025. Manuscript. Distributional Learning and Grammaticalization: Modals in the History of English.

Corpora

Truswell, R., Alcorn, R., Donaldson, J., and Wallenberg, J. (2018). A Parsed Linguistic Atlas of Early Middle English (PLAEME). University of Edinburgh.

Zimmermann, R. (2018). The Parsed Corpus of Middle English Poetry (PCMEP).

Kroch, A. and Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2), release 3. University of Pennsylvania, Philadelphia.