60% Ciceronianus es: Automatic discovery of Latin syntactic changes

Micha Elsner, Ben Swanson and Emily Lane
Running a variationist study

This construction sounds odd...

Let’s see who uses it!

Where and when?

Intuitions about a variant

Gather and analyze data

Social and historical conclusions
Initial question relies on human intuition

Intuitions about a variant

This construction sounds odd...

Gather and analyze data

Social and historical conclusions
Intuitions can be tricky...

- Recently emerging variant
- Dead language or dialect
- Gradient effect

What we want: **data-driven** method to suggest variants

- Exists for **lexical** variation (e.g. Eisenstein 2014)
- What about syntax?
Syntax is hard, because:

- Parsers unreliable outside training domain (McClosky 2010)
  - *Especially* for variant constructions we care about!
- Have to choose correct unit of analysis
  - Single phrasal rules?
  - Bigger subtrees?
  - Lexicalized subtrees?

```
N-nom  V

V   COMPL   N-nom  V

dicit  quod
```
Focus here on representation

- Parsers unreliable outside training domain (McClosky 2010)
  - *Especially* for variant constructions we care about!
- Have to choose correct unit of analysis
  - Single phrasal rules?
  - Bigger subtrees?
  - Lexicalized subtrees?
Representing syntax: tree fragments

Grammar formalism generalizes context-free grammar (see Cohn et al. 2009)
Used in native language identification

(Swanson and Charniak 2012 and subsq., Wong and Dras 2011)
But which TSG fragments?

- Single phrase structure tree has many TSG derivations
- Can use Bayesian analysis (Cohn et al. 2009)
- "Double-DOP" technique (Sangati and Zuidema 2011)
  - If two trees share a maximal fragment, add it to the grammar
Double-DOP extracts shared subtrees

Vergil wrote a famous book.

Conquered Gaul by a great soldier.
Lexicalization: What is “grammar”? 

Naive TSG learning will pick up topic effects: (cf. Sarawgi et al 2011)

- Caesar’s grammar: (NP → Gallia)
- Aquinas’ grammar: (Adj → Christiana)

These effects aren’t historical language change

How can we separate cultural difference from linguistic difference?
De-lexicalize most of the sentence

A weak point of the approach…
(I have some ideas about how to avoid this in future work)

Retain only:
● Conjunctions (et, vel… )
● Prepositions (in, ad… )
● Complementizers (ut, quia… )
● Some adverbials (non… )
How to detect change  (following Swanson and Charniak 2014)

- Create TSG grammar from corpus
  - Using Bayesian extractor or double-DOP
- Use grammar to parse each sentence
  - Find TSG fragments which occur in any derivation
- Examine text × fragment co-occurrence matrix for socio-historical patterns
  - Use $\chi^2$-squared statistic to rank
Why Latin? Parsed corpus available across time

data from Perseus (Bamman and Crane 2011); Index Thomisticus (Passarotti 2007)

Classical Latin (250 BCE - 100 CE)

Late Latin (100 - 600)

Medieval Latin (600 - 1300)

Neo-Latin (1300 - 1700)

dates following Lind 1941

Cicero

Sallust

Caesar

Propertius

Vergil

Ovid

Petronius

Vulgate Bible

Thomas Aquinas
Canonical authors validate the methodology

- May not tell us much that is really surprising
- But can compare what we find to known answers

My book is the most canonical!

Well, I’ve actually been canonized!
Medieval Latin *does* have mysteries left to solve…

- “Regional” Latins? (Afro-Latin, Germano-Latin)
- Standards of education in Medieval world

Comprehensive picture requires comparison across non-canonical texts (e.g. monastery records)

A full-scale computational method would be useful!

Löfstedt 1959 ch. 3
Case study: Classical vs. Medieval prose

Also looked at prose vs. poetry

“Classical” group:
- Sallust
- Cicero
- Caesar

“Medieval” group:
- Petronius
- Vulgate Bible
- Thomas Aquinas

The Vulgate: an intermediate stage?
Can we tell them apart?

Yes!

- Selected rules with $\chi^2$-squared $p < .00001$ (n=357)
- Testing 2414 unseen sentences
  - (442 classical, 1972 Thomas)
- Can correctly mark:
  - 341 classical sentences (77%)
  - 1972 Thomas sentences (98%)
Latin complement clauses: a well-known change

e.g. Sidwell 1990 p368

Cicero:

**Lepidum te habitare velle dixisti**

Lepidus-ACC you-ACC live-INF want-INF say-2PERF

“You said that you wanted to live with Lepidus”

Thomas:

**dicitur quod sapientia infinitus thesaurus est**

say-3PASSV that wisdom infinite treasury be-3PRES

“It is said that wisdom is an infinite treasury”
Our system: complementizers

Classical authors

V-inf

N-acc

V-inf

χ-squared=46 (69 inst.)

Thomas Aquinas

igitur

“therefore”

χ-squared=353 (1575 inst.)

autem

“however”

χ-squared=351 (1475 inst.)

cum

“since”

χ-squared=299 (68 inst.)

cum

“when”

χ-squared=102 (24 inst.)

quod

“that”

χ-squared=161 (990 inst.)

quod

“that”

χ-squared=150 (738 inst.)
Why are the rules so small?

TSG has trouble with adjuncts:

\[
dico \quad te \quad [priore \ nocte] \quad venisse
\]

say-1 you-ACC [previous night]-ABL come-INF

“I say that you came on the previous night”

- No way of marking optionality
- Worsened by flat structure in dependency trees

Rule for classical subclause after \textit{dico} “say”

\[
V\text{-inf} \quad N\text{-acc} \quad V\text{-inf}
\]

Rule with added temporal modifier

\[
V\text{-inf} \quad N\text{-acc} \quad V\text{-inf} \quad N\text{-ABL}
\]
Distinguishing feature: adjective placement

Classical authors use more post-nominal adjectives
But Thomas prefers prenominals

<table>
<thead>
<tr>
<th>Case</th>
<th>Classical</th>
<th>Thomas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom</td>
<td>52% (101 : 93)</td>
<td>27% (65 : 174)</td>
</tr>
<tr>
<td>Gen</td>
<td>55% (72 : 58)</td>
<td>24% (41 : 131)</td>
</tr>
<tr>
<td>Dat</td>
<td>64% (30 : 17)</td>
<td>8% (3 : 34)</td>
</tr>
<tr>
<td>Acc</td>
<td>54% (187 : 157)</td>
<td>32% (55 : 115)</td>
</tr>
<tr>
<td>Abl</td>
<td>35% (113 : 211)</td>
<td>34% (45 : 86)</td>
</tr>
</tbody>
</table>
Is this change, or something else?

Classical Latin:

● Change in progress from Adj-N to N-Adj (Ledgeway 2012)
● N-Adj claimed to be classical unmarked order

Medieval Latin:

● N-Adj persists into Romance

Why the Adj-N preference in Thomas?
What about the Vulgate?

- Latin bible, compiled in 380s by Saint Jerome
  - New Testament based on existing vernacular versions
- Important forerunner of Medieval Latin:
  - “sanctified… changes in the use of the cases and the subjunctive… It is linguistically a central text.”

Sidwell, 1995
Jerome thought his own Latin was classical...

I would fast, and then read Cicero. After sleepless nights, after tears… I took up Plautus. And whenever I tried to change my wicked ways and read the prophets, the crudity of the language was shocking.

Suddenly I was caught up in the spirit, and dragged before the seat of the Judge. And asked who I was, I replied, “A Christian.” “Liar,” he said, “You are a Ciceronian, not a Christian! For where you keep your treasure, there also is your heart.”
How classical is the Vulgate?

According to the classifier

- 258 more classical
- 147 more Thomist

Actually, you’re close to 60% Ciceronian!
Which features make the difference?

<table>
<thead>
<tr>
<th>More classical</th>
<th>More Thomistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Post-nominal adj. (abl)</td>
<td>● Pronouns (gen.)</td>
</tr>
<tr>
<td>● Indicative verbs</td>
<td>● Adverbials</td>
</tr>
<tr>
<td>● Postnominal adj. (acc)</td>
<td>● Preposition <em>in</em> “in”</td>
</tr>
<tr>
<td>● Preposition <em>super</em> “on”</td>
<td>● Clause-initial <em>et</em> “and”</td>
</tr>
<tr>
<td>● Misc. complementizers</td>
<td>● Pronouns (nom)</td>
</tr>
<tr>
<td>● Conjunction <em>que</em> “and”</td>
<td>● Postnominal adj. in PP</td>
</tr>
<tr>
<td>● Complementizer <em>cum</em> “when/since”</td>
<td>● Conjunction <em>sicut</em> “just as”</td>
</tr>
</tbody>
</table>

Some possible change, some stylistic features
Subclauses in the Vulgate Apocalypse

**Classical subclause:**

*his, qui se dicunt Judæos esse, et non sunt, sed sunt synagoga Satanæ*

“of these, who say they are Jews, and are not, but are the synagogue of Satan”

**Direct quote with quod, parallel tensed subclause:**

*quia dicis quod dives sum... et nescis quia tu es miser*

“because you say this: I am rich, and you do not know that you are poor”

**Tensed subclause:**

*diabolus ad vos habens iram magnam, sciens quod modicum tempus habet*

“the devil has great wrath against you, knowing that he has but a short time”
So, what’s still missing?

- Lexically specific constructions
  - Nearly all Medieval Latin changes are lexico-syntactic
- A way to handle adjuncts
- Good automatic parsing
Can’t handle semantics

Changes to tense system undetectable as structural rules:

- Imperfect for perfect
- Perfect for pluperfect
- Pluperfect for perfect (\textit{sed ego dixeram} : “but I said”)

Detecting these requires the \textit{sense} as well as the form

\textit{Sidwell 1995}
In conclusion

- Tree substitution grammar represents constructions
- Finds several major changes in history of Latin
- The Vulgate retains many classical features
- Good automatic analysis still requires innovation in:
  - Distinguishing topic from grammar
  - Handling adjuncts
  - Cross-domain parsing
Thanks for listening!

Questions?
Tree substitution rules

- Tree fragments represent constructions
- Can vary in size:
  - Single context-free rule...
  - To entire sentence
- A flexible way of capturing syntactic variation
But which TSG fragments?

- Single phrase structure tree has many TSG derivations
- Can use Bayesian analysis (Cohn et al. 2009)
- “Double-DOP” technique (Sangati and Zuidema 2011)
  - If two trees share a maximal fragment, add it to the grammar
\( \chi \)-squared ranking

- Depends on both frequency and predictive power

<table>
<thead>
<tr>
<th>Rule 1: Frequent and predictive (complementizer <em>autem</em>)</th>
<th>Rule 2: Rare and predictive (locative noun)</th>
<th>Rule 3: Frequent, not predictive (infinitive verb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi )-squared = 246</td>
<td>( \chi )-squared = 151</td>
<td>( \chi )-squared = 67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>classics</th>
<th>Thomas</th>
<th></th>
<th>classics</th>
<th>Thomas</th>
<th></th>
<th>classics</th>
<th>Thomas</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>has rule</td>
<td>11</td>
<td>1035</td>
<td>has rule</td>
<td>35</td>
<td>0</td>
<td>has rule</td>
<td>1176</td>
<td>4488</td>
<td></td>
</tr>
<tr>
<td>no rule</td>
<td>1539</td>
<td>5867</td>
<td>no rule</td>
<td>1515</td>
<td>6902</td>
<td>no rule</td>
<td>1550</td>
<td>2414</td>
<td></td>
</tr>
</tbody>
</table>
Some technical issues

- Latin non-projective dependencies converted to phrase structure trees
  - Put a projection over every head
  - Mark and reorder elements with crossing arcs

"leapt out on the Hesperian shore"