

Automatic discovery of Latin syntactic changes

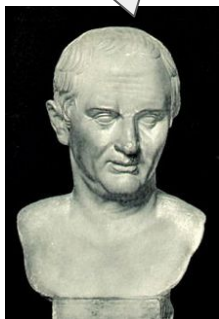
Micha Elsner and Emily Lane



THE OHIO STATE UNIVERSITY

Running a variationist study

This construction sounds odd...



Intuitions about a variant

Let's see who uses it!



Gather and analyze data

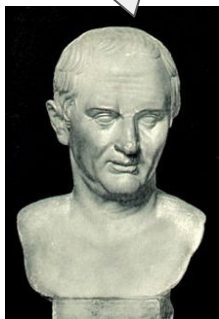
Where and when?



Social and historical conclusions

Initial question relies on human intuition

This construction sounds odd...



**Intuitions about
a variant**



**Gather and
analyze data**



**Social and historical
conclusions**

Intuitions can be tricky...

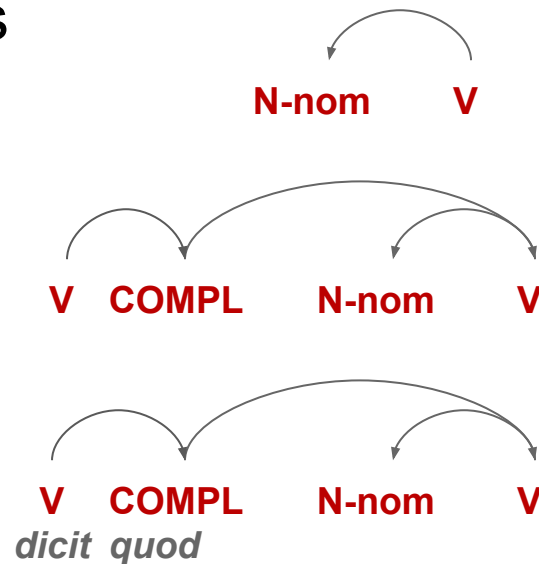
- Recently emerging variant
- Dead language or dialect
- Gradient effect

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

- Exists for **lexical** variation (e.g. Eisenstein 2014)
- What about syntax?

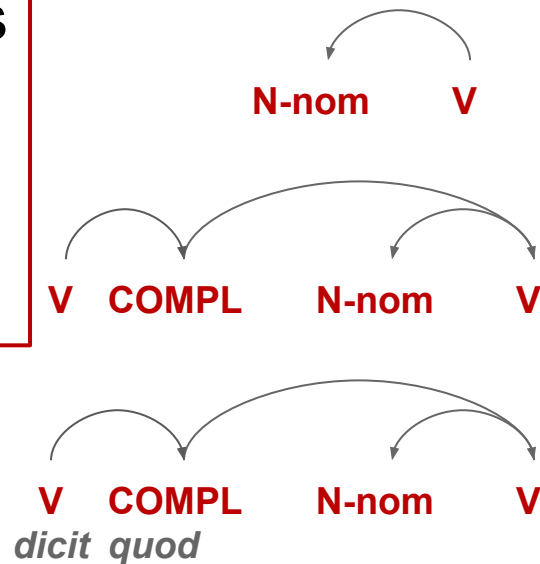
Syntax is hard, because:

- 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?



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?

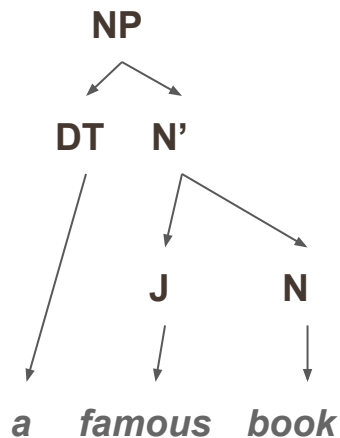
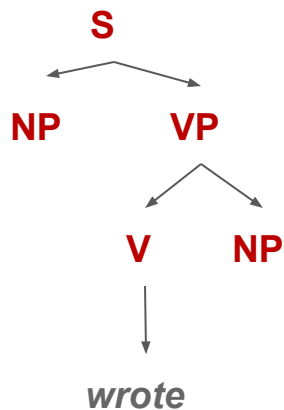


Representing syntax: tree fragments

Grammar formalism generalizes context-free grammar (see Cohn et al. 2009)

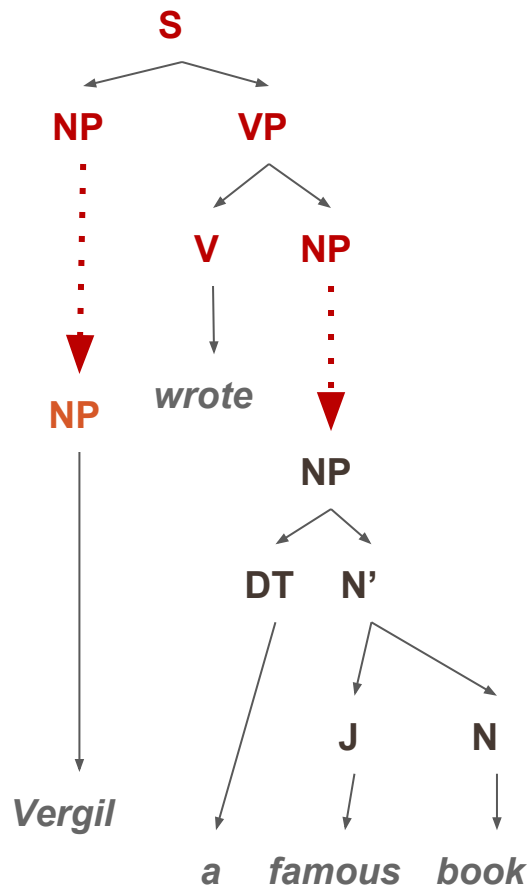
Used in native language identification

(Swanson and Charniak 2012 and subseq., Wong and Dras 2011)

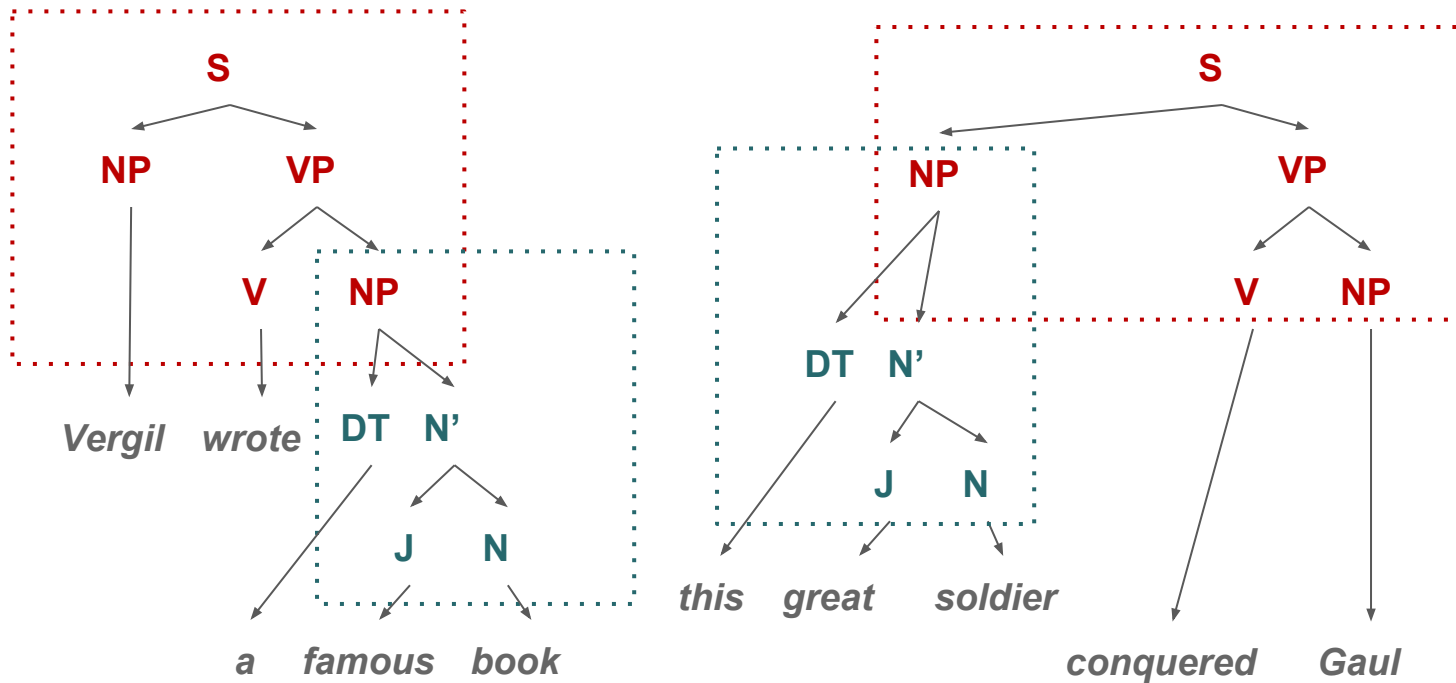


But which TSG fragments?

- 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



Lexicalization: What is “grammar”?

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

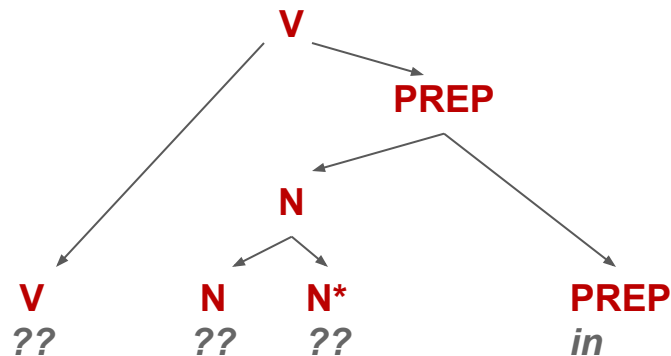
- Caesar’s grammar: (NP $\rightarrow$ *Gallia*)
- Aquinas’ grammar: (Adj $\rightarrow$ *Christiana*)

These effects aren’t historical language change



How can we separate
cultural difference from
linguistic difference?



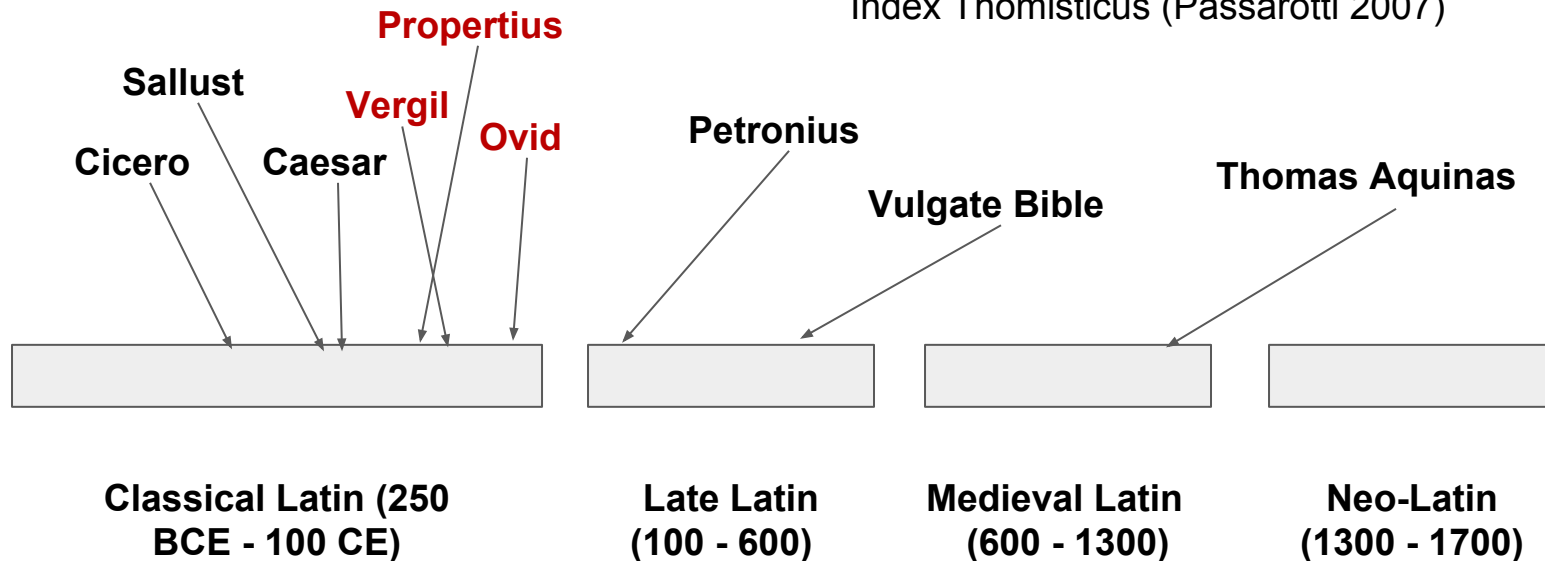


How to detect change (following Swanson and Charniak 2014)

- 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* $\times$ *fragment* co-occurrence matrix for socio-historical patterns
 - Use χ -squared statistic to rank

Why Latin? Parsed corpus available across time

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



dates following Lind 1941

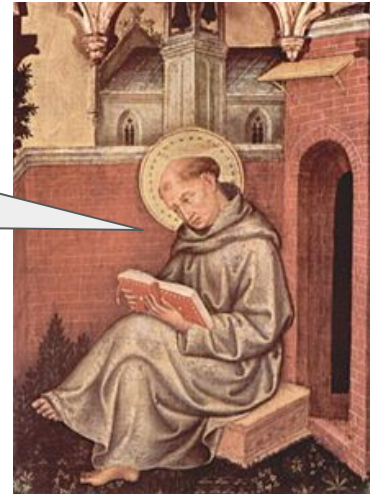
Canonical authors validate the methodology

- 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

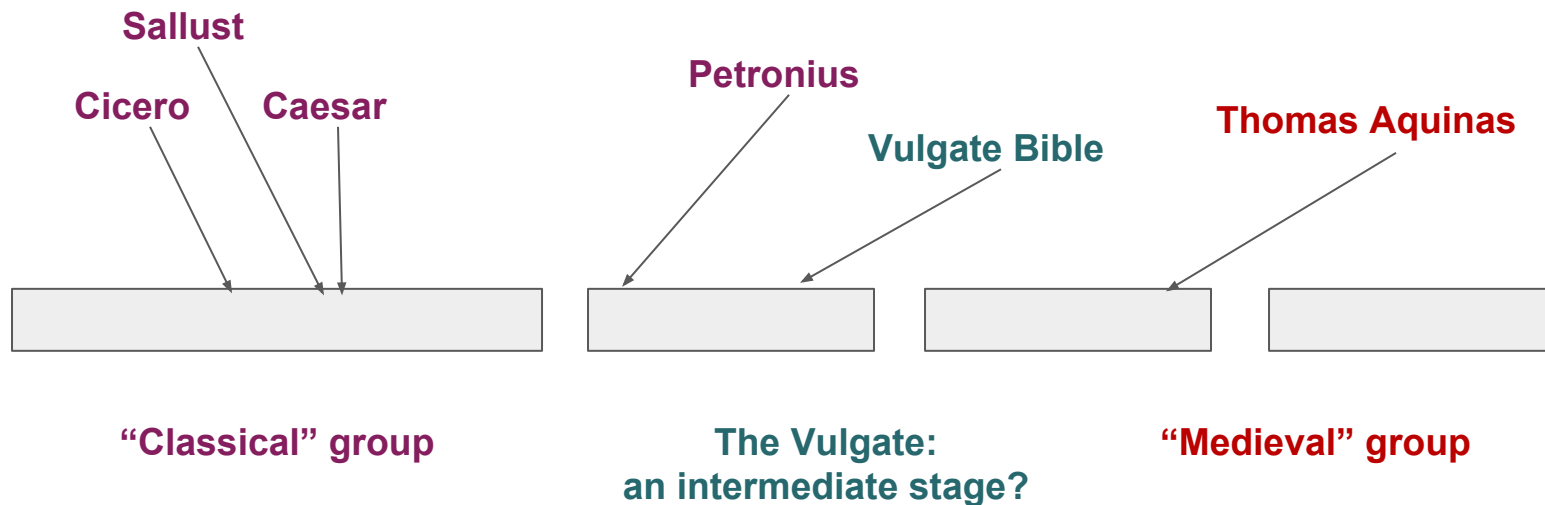
Löfstedt 1959 ch. 3

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

A full-scale computational method would be useful!

Case study: Classical vs. Medieval prose

Also looked at prose vs. poetry



Can we tell them apart?

Yes!

- Selected rules with χ -squared $p < .00001$ ($n=357$)
- Testing 2414 unseen sentences
(442 classical, 1972 Thomas)
- Can correctly mark:
 - 341 classical sentences (77%)
 - 1931 Thomas sentences (98%)

Latin complement clauses: a well-known change

Cicero:

e.g. Sidwell 1990 p368

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:

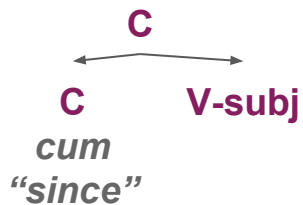
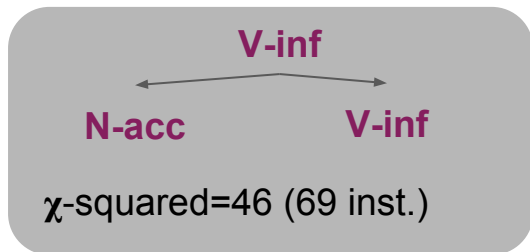
dicitur quod sapientia infinitus thesaurus est

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

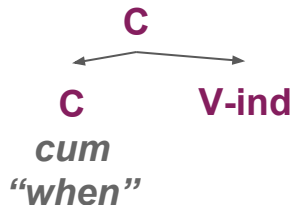
“It is said that wisdom is an infinite treasury”

Our system: complementizers

Classical authors

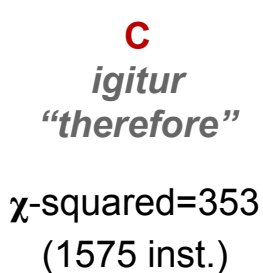


χ -squared=299 (68 inst.)

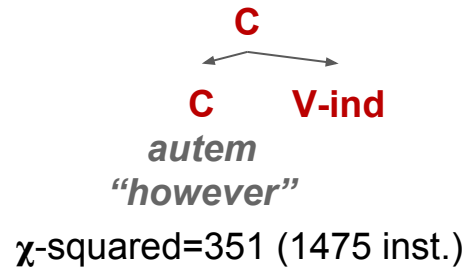


χ -squared=102 (24 inst.)

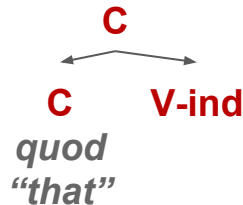
Thomas Aquinas



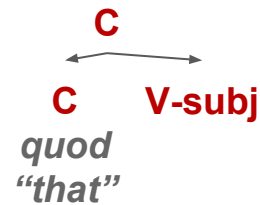
χ -squared=353
(1575 inst.)



χ -squared=351 (1475 inst.)



χ -squared=161 (990 inst.)



χ -squared=150 (738 inst.)

Why are the rules so small?

TSG has trouble with adjuncts:

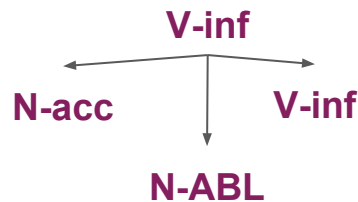
dico te [priore nocte] venisse
say-1 you-ACC [previous night]-ABL come-INF
“I say that you came on the previous night”

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

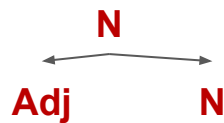
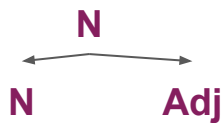
Rule for classical subclause
after *dico* “say”



Rule with added temporal
modifier



Distinguishing feature: adjective placement



	Classical	Thomas
Nom	53% (69 : 60)	26% (45 : 129)
Gen	56% (53 : 42)	25% (29 : 89)
Dat	65% (24 : 13)	8% (2 : 34)
Acc	57% (138 : 104)	34% (42 : 82)
Abl	36% (87 : 158)	36% (35 : 62)

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

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?

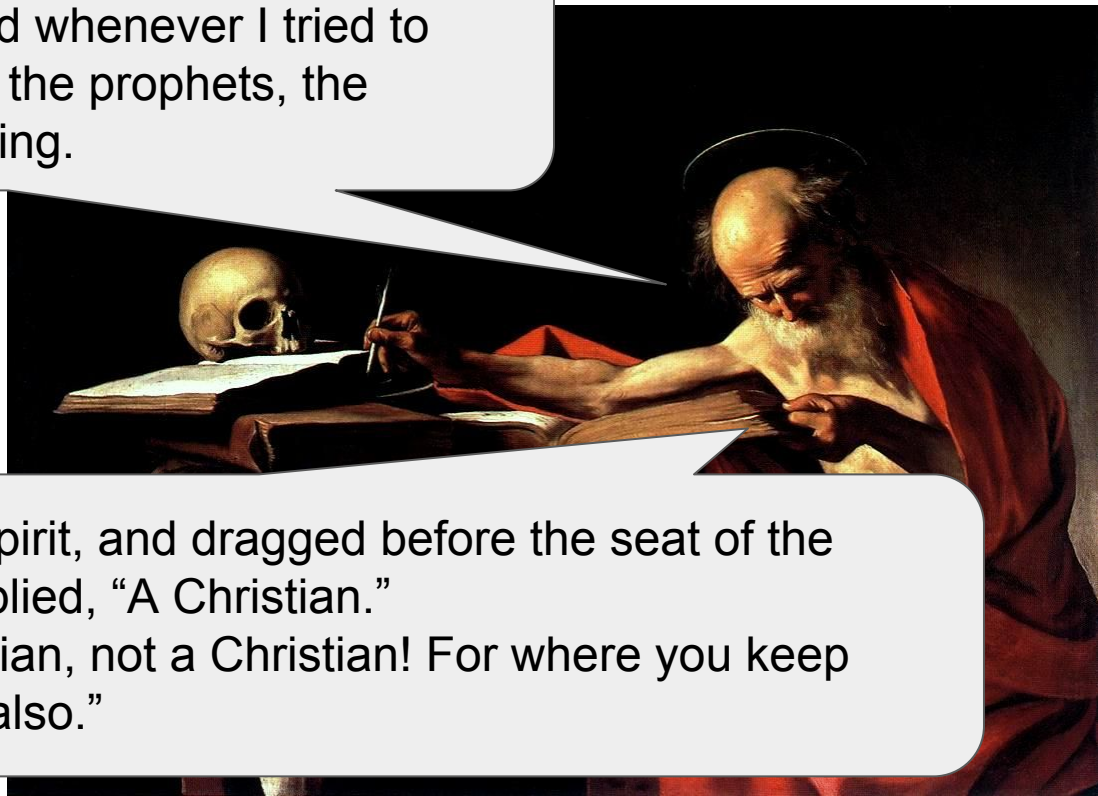
What about the Vulgate?

- 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 is your heart also."

How classical is the Vulgate?

According to the classifier

- 258 more **classical**
- 147 more **Thomist**

Actually, you're close to 60% Ciceronian!



Miserere mei, Domine!

Which features make the difference?

More **classical**

- **Post-nominal adj. (abl)**
- Indicative verbs
- **Postnominal adj. (acc)**
- Preposition *super* “on”
- Misc. complementizers
- **Conjunction *que* “and”**
- **Complementizer *cum* “when/since”**

More **Thomistic**

- Pronouns (gen.)
- Adverbials
- Preposition *in* “in”
- Clause-initial *et* “and”
- Pronouns (nom)
- **Postnominal adj. in PP**
- Conjunction *sicut* “just as”

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
 - **Some proposals:** McGillivray 2014, Passarotti et al 2010, et al.

Can't handle semantics

Changes to tense system undetectable as structural rules:

- Imperfect for perfect
- Perfect for pluperfect
- Pluperfect for perfect (*sed ego dixeram* : “but I said”)

Sidwell 1995

Detecting these requires the *sense* as well as the form

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!

Gratias agimus:

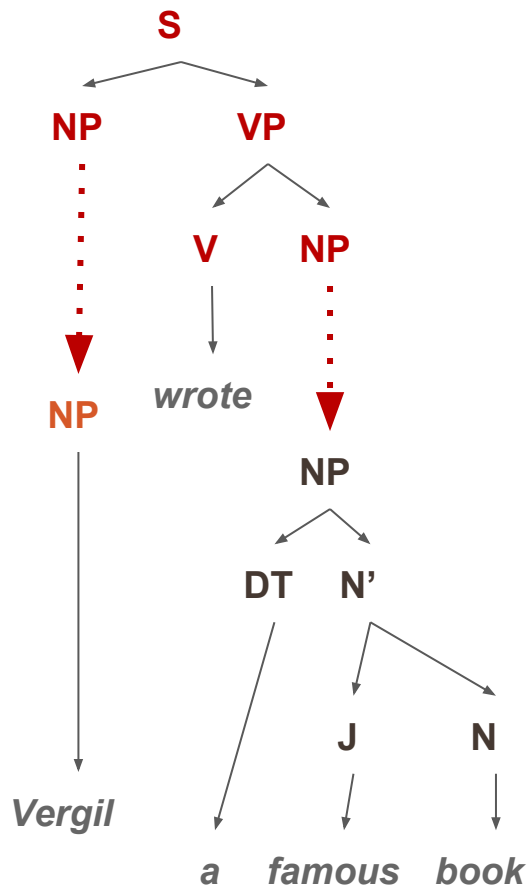
Marco Passarotti, Ben Swanson,
Brian Joseph, Alex Erdmann, Julia
Papke and our reviewers.

Questions?



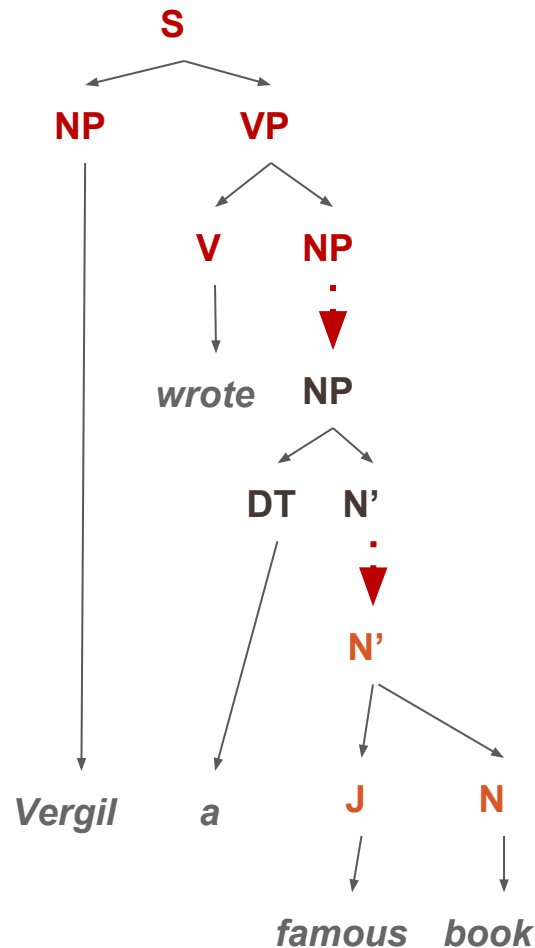
Tree substitution rules

- 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



χ -squared ranking

- Depends on both frequency and predictive power

Rule 1:
Frequent and predictive
(complementizer *autem*)
 χ -squared = 246

	classics	Thomas
has rule	11	1035
no rule	1539	5867

Rule 2:
Rare and predictive
(locative noun)
 χ -squared = 151

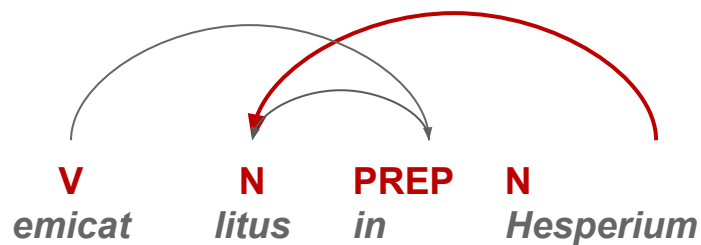
	classics	Thomas
has rule	35	0
no rule	1515	6902

Rule 3:
**Frequent, not
predictive**
(infinitive verb)
 χ -squared = 67

	classics	Thomas
has rule	1176	4488
no rule	1550	2414

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”

