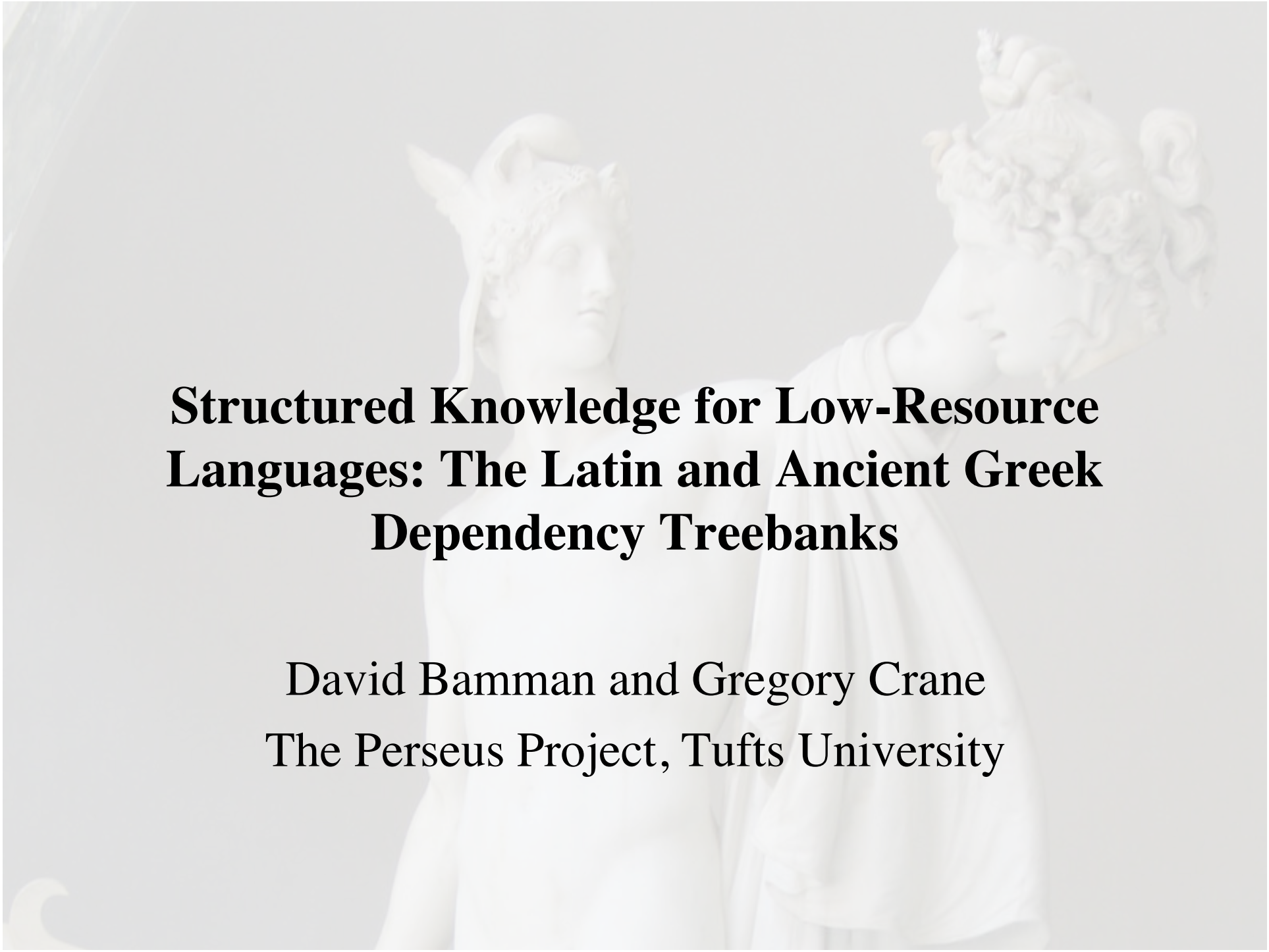


Structured Knowledge for Low-Resource Languages: The Latin and Ancient Greek Dependency Treebanks

David Bamman and Gregory Crane

The Perseus Project, Tufts University

The Problem: Classical Philology

"We should not fail to hear the almost benevolent nuances which for a Greek noble, for example, lie in all the words with which he set himself above the lower people—how a constant type of pity, consideration, and forbearance is mixed in there, sweetening the words, to the point where **almost all words which refer to the common man** finally remain as expressions for "unhappy," "worthy of pity" (compare *deilos [cowardly]*, *deilaios [lowly, mean]*, *ponêros [oppressed by toil, wretched]*, *mochthêros [suffering, wretched]*—the last two basically designating the common man as a slave worker and beast of burden)."

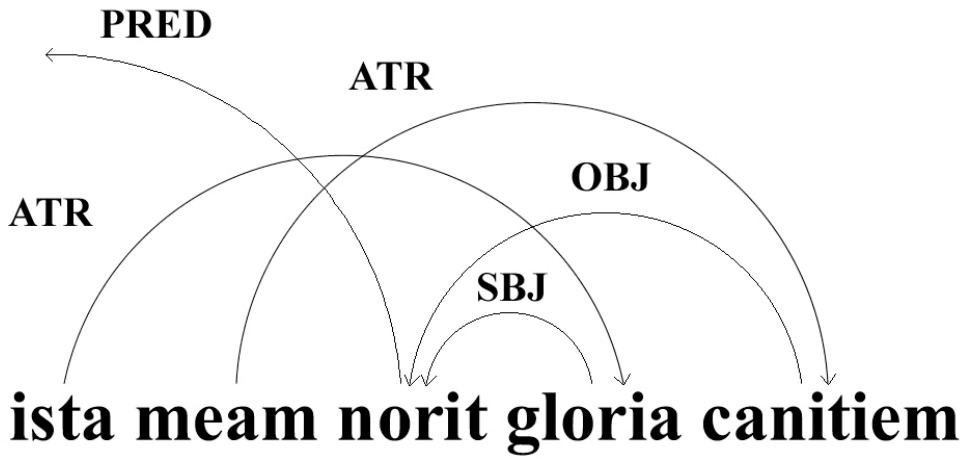
F. Nietzsche, *Genealogy of Morals* 1.10

Background

- Most recent research and labor in treebanks has focused on modern languages, but recent scholarship has seen the rise of treebanks for historical languages as well:
  - Middle English (Kroch and Taylor 2000)
  - Early Modern English (Kroch et al. 2004)
  - Old English (Taylor et al. 2003)
  - Early New High German (Demske et al. 2004)
  - Medieval Portuguese (Rocio et al. 2000)
  - Latin/Thomas Aquinas (Passarotti 2008ff.)
  - Latin/Greek/Gothic/Slavonic (PROIEL 2008ff.)

Design

- Latin and Ancient Greek are heavily inflected languages with a high degree of variability in word order: constituents of sentences are often broken up with elements of other constituents, as in *ista meam norit gloria canitiem* ("that glory will know my old age").



Design

- This high level of non-projectivity has encouraged us to base our annotation style on that used by the Prague Dependency Treebank (PDT) for Czech (another non-projective language), while tailoring it for Latin via the grammar of Pinkster (1990). In contrast to the phrase-structure style annotation of other treebanks (e.g., the Penn Treebank), the PDT annotation is based on the dependency grammar of Mel'cuk (1988), which links words to their immediate heads without any intervening non-terminal phrasal categories.

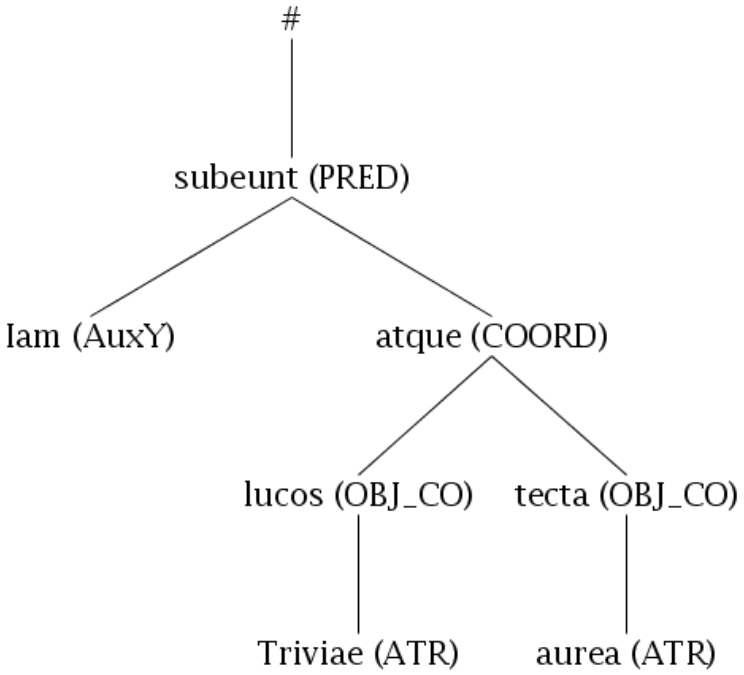
Tagset

PRED	predicate
SBJ	subject
OBJ	object
ATR	attribute
ADV	adverbial
ATV/AtvV	complement
PNOM	predicate nominal
OCOMP	object complement
COORD	coordinator
APOS	apposing element

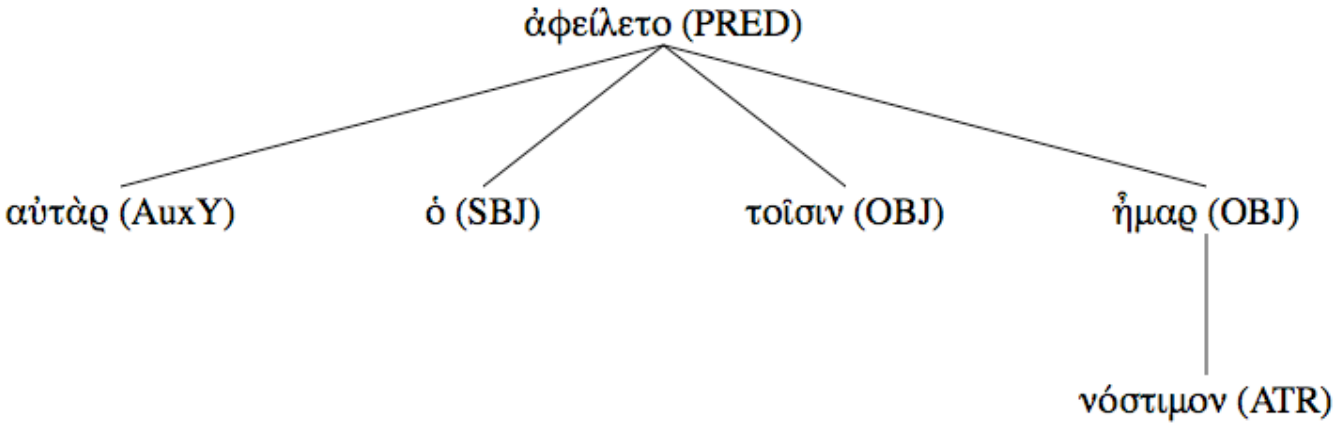
AuxP	preposition
AuxC	conjunction
AuxR	reflexive passive
AuxV	auxiliary verb
AuxX	commas
AuxG	bracketing punctuation
AuxK	terminal punctuation
AuxY	sentence adverbials
AuxZ	emphasizing particles
ExD	ellipsis

LDT 1.5 Composition

Author	Words
Caesar	1,488
Cicero	6,229
Sallust	12,311
Vergil	2,613
Jerome	8,382
Ovid	4,789
Petronius	12,474
Propertius	4,857
Total	**53,143**



AGDT 1.1 Composition



Work	**Words**
Homer, *Odyssey*	18,790
Homer, *Iliad*	3,945
Total	22,735

Annotation Process

- Annotation process
  - All texts are annotated two independent annotators and reconciled by a third.

- Distributed online annotation through the Perseus Digital Library

Perseus Digital Library

Homer, *Odyssey*

Agamemnon Search

("Agamemnon", "Hom. Od. 9.1", "denarius")

[advanced search] [view abbreviations]

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position. Hide browse bar

book:

This text is part of:

Greek and Roman Materials
Greek Hexameter
Greek Poetry
Greek Texts
Homer
Homer, Odyssey

View text chunked by:

book : line

Table of Contents:

book 1
lines 1-43
lines 44-79
lines 80-124
lines 125-177

1.1

Click on a word to bring up parses, dictionary entries, and frequency statistics

ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ
πλάγχθη, ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσεν:
πολλῶν δ᾽ ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω,
πολλὰ δ᾽ ὅ γ᾽ ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν,
ἀρνύμενος ἥν τε ψυχὴν καὶ νόστον ἑταίρων.
ἀλλ᾽ οὐδ᾽ ὣς ἑτάρους ἐρρύσατο, ἱέμενός περ:
αὐτῶν γὰρ σφετέρῃσιν ἀτασθαλίῃσιν ὄλοντο,
νήπιοι, οἳ κατὰ βοῦς Ὑπερίονος Ἠελίοιο
ἤσθιον: αὐτὰρ ὁ τοῖσιν ἀφείλετο νόστιμον ἦμαρ.
τῶν ἁμόθεν γε, θεά, θύγατερ Διός, εἰπὲ καὶ ἡμῖν.

ἔνθ᾽ ἄλλοι μὲν πάντες, ὅσοι φύγον αἰπὺν ὄλεθρον,
οἴκοι ἔσαν, πόλεμόν τε πεφευγότες ἠδὲ θάλασσαν:
τὸν δ᾽ οἶον νόστου κεχρημένον ἠδὲ γυναικὸς
νύμφη πότνι᾽ ἔρυκε Καλυψὼ δῖα θεάων
ἐν σπέσσι γλαφυροῖσι, λιλαιομένη πόσιν εἶναι.
ἀλλ᾽ ὅτε δὴ ἔτος ἦλθε περιπλομένων ἐνιαυτῶν,
τῷ οἱ ἐπεκλώσαντο θεοὶ οἰκόνδε νέεσθαι

English (1919) focus load

English (Samuel Butler) focus load

Notes (W. Walter Merry, James Riddell, D. B. Monro, 1886) focus show

References (23 total) show

Vocabulary Tool load

Syntax hide

See a syntactic parse of this sentence:

- ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ πλάγχθη, ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσεν
- πολλῶν δ᾽ ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω, πολλὰ δ᾽ ὅ γ᾽ ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν, ἀρνύμενος ἥν τε ψυχὴν καὶ νόστον ἑταίρων
- ἀλλ᾽ οὐδ᾽ ὣς ἑτάρους ἐρρύσατο, ἱέμενός περ
- αὐτῶν γὰρ σφετέρῃσιν ἀτασθαλίῃσιν ὄλοντο, νήπιοι, οἳ κατὰ βοῦς Ὑπερίονος Ἠελίοιο ἤσθιον
- αὐτὰρ ὁ τοῖσιν ἀφείλετο νόστιμον ἦμαρ
- τῶν ἁμόθεν γε, θεά, θύγατερ Διός, εἰπὲ καὶ ἡμῖν
- ἔνθ᾽ ἄλλοι μὲν πάντες, ὅσοι φύγον αἰπὺν ὄλεθρον, οἴκοι ἔσαν, πόλεμόν τε πεφευγότες ἠδὲ θάλασσαν
- τὸν δ᾽ οἶον νόστου κεχρημένον ἠδὲ γυναικὸς νύμφη πότνι᾽ ἔρυκε Καλυψὼ δῖα θεάων ἐν σπέσσι γλαφυροῖσι, λιλαιομένη πόσιν εἶναι

Treebank Annotation

Latin and Greek Dependency Treebanks: david

- **document_id**: Perseus:text:1999.02.0066
- **subdoc**: book=1;poem=8B [context]
- **span**: ista0:canitiem0
- Image: [SVG] [PNG]
- Return to list
- Logoff

ista meam norit gloria canitiem

index	word	head	relation	lemma + morph	add new lemma	add new morph	notes
0	ista	3	ATR	pron sg fem nom			
1	meam	4	ATR	adj sg fem acc			
2	norit	-1	PRED	verb 3rd sg perf subj act			
3	gloria	2	SBJ	noun sg fem nom			
4	canitiem	2	OBJ	noun sg fem acc			

Save Re-Parse

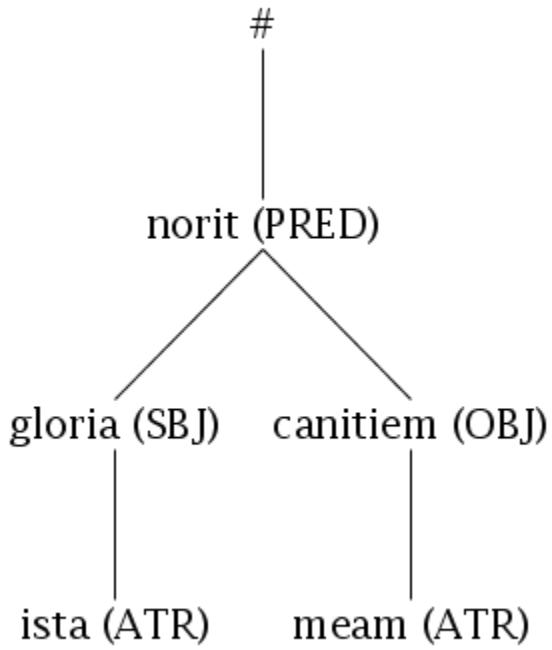
```
ista meam norit gloria canitiem
  +-----ATR>-------+
        +-------ATR>--------+
              +----<OBJ------+
              +<SBJ--+
ista meam norit gloria canitiem
```

Treebank Annotation

Try it yourself!

http://nlp.perseus.tufts.edu/hopper

Serialization: SVG/PNG



Serialization: XML

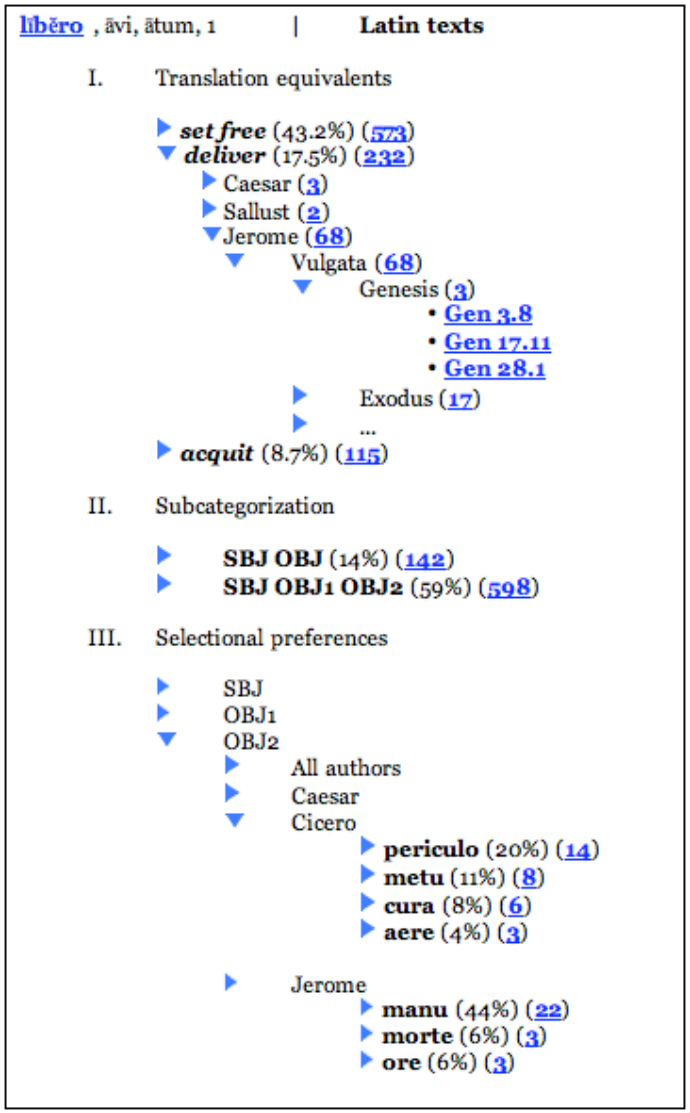
```
– <treebank version="1.4" xsi:schemaLocation="http://nlp.perseus.tufts.edu/syntax/treebank/1.4 treebank-1.4.xsd">
  – <sentence id="1" document_id="Perseus:text:1999.02.0066" subdoc="book=1:poem=8B" span="ista0:canitiem0">
      <word id="1" form="ista" lemma="iste1" postag="p-s---fn-" head="4" relation="ATR"/>
      <word id="2" form="meam" lemma="meus1" postag="a-s---fa-" head="5" relation="ATR"/>
      <word id="3" form="norit" lemma="nosco1" postag="v3srsa---" head="0" relation="PRED"/>
      <word id="4" form="gloria" lemma="gloria1" postag="n-s---fn-" head="3" relation="SBJ"/>
      <word id="5" form="canitiem" lemma="canities1" postag="n-s---fa-" head="3" relation="OBJ"/>
    </sentence>
  </treebank>
```

Applications

- NLP tasks pertaining directly to syntax
  - grammar induction
  - automatic parsing

- Downstream applications for which syntax is one feature among many
  - machine translation (Charniak et al. 2003; SSST workshops)
  - measuring text reuse/similarity (Bamman and Crane 2008b)
  - **lexicography**

Background: *Dynamic Lexicon*



Automatic parsing

- Supervised learning: train a parser on human annotated data and use that trained model to parse unannotated data

  - Labeled dependency parsing accuracy (with gold tags)
    - English: 86% (Nivre et al. 2007)
    - Czech: 80% (Collins et al. 1999)
  - Accuracy tied to treebank size: larger is better
    - English: Penn treebank (+1 million words)
    - Czech: Prague Depenency Treebank (1.5 million words)

Morphological tagging

- Tested with TreeTagger (Schmid 1994) analyzer - performed in a 10-fold test with an accuracy of **83%** in disambiguating the full morphological analysis (Bamman and Crane 2008a).

	Accuracy
Case	90.10%
Gender	92.90%
Mood	98.68%
Number	95.15%
POS	95.11%
Person	99.56%
Tense	98.62%
Voice	98.89%
All	**83.10%**

Automatic parsing

- LDT 1.4 contains 30,537 words. Parsing evaluation with MSTParser (McDonald et al. 2005): with gold morphological tags, **54.34%;** with automatically assigned tags, **50.00%.** By author:

	Gold	Automatic
Jerome	61.44%	58.15%
Sallust	53.04%	46.99%
Caesar	51.34%	46.24%
Cicero	49.97%	44.41%
Vergil	48.99%	40.60%

Automatic parsing

- By tag (with automatically assigned tags):

	Precision	Recall
ATR	63.09%	62.41%
AuxP	63.66%	66.81%
SBJ	50.93%	51.10%
OBJ	50.90%	55.12%
ADV	49.24%	55.31%

	Precision	Recall
AuxC	34.80%	36.04%
SBJ_CO	26.58%	29.04%
OBJ_CO	31.84%	30.85%
ATR_CO	30.35%	25.17%
ADV_CO	30.29%	22.22%

Automatic parsing: summary

- Parsing accuracy on a 30K-word training set isn't that great
  - 54.34% with gold morphological tags
  - 50.00% with automatically assigned tags (83% accurate tagging)
- Better performance on prose than poetry
- With automatic morphological tagging, better precision/recall (~60%) on ATR, AuxP, SBJ, OBJ, ADV than on long-distance relationships (AuxC etc.)

- Automatic parsing isn't really viable as an end in itself (for pedagogy etc.), but it can be offset by a large enough volume of unstructured data for other tasks (like automatically building dictionaries).

Inducing selectional preferences

- Trained a parser on our 30K word Latin treebank
- Parsed all the texts in our 3.5 million word Latin corpus
- To find selectional preferences from this noisy data, we've used the same hypothesis tests (log likelihood etc.) used to find *syntactic* collocations in completely unstructured texts.

$$\log \lambda = \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)$$

where $p = \frac{c_2}{N}$, $p_1 = \frac{c_{12}}{c_1}$, $p_2 = \frac{c_2 - c_{12}}{N - c_1}$, $N = \textit{corpus count}$

and $L(a, b, c) = c^a (1 - c)^{b-a}$.

Greek Treebank = Greek Lexicon

δύναμις

(noun): power, force, army (Flavius Josephus)

Attributes:

- ναυτικός ("naval force"): 15.01/31. (Polybius)
- πεζικός ("land army"): 12.45/12. (Polybius)
- μέγας ("great power"): 4.52/115. (Isocrates)
- τηλικοῦτος ("so great power"): 4.49/25. (Isocrates)
- ἑαυτοῦ ("his power"): 3.24/102.

Object of:

- ἔχω ("having as much power"): 8.93/239. (Plato)
- ἐξάγω ("to army"): 2.40/16. (Polybius)
- ἀθροίζω ("gather all together army"): 2.32/15.
- ἕχις ("potency"): 2.16/25. (Epictetus, Plato)

URLs

- Treebank data:
  http://nlp.perseus.tufts.edu/syntax/treebank/

- Treebank annotation environment
  http://nlp.perseus.tufts.edu/hopper/