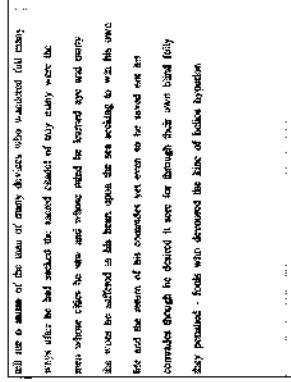
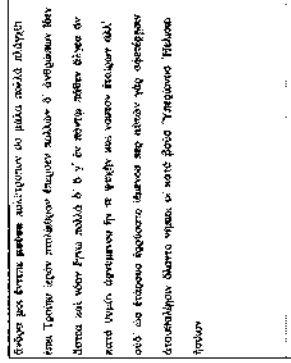


Interlinear Translations

" Interlinear *Odyssey* (Alpheios Project) using Perseus alignment data.



Automatically Building Bilingual Dictionaries for Greek and Latin

9

Automatically Building Bilingual Dictionaries for Greek and Latin

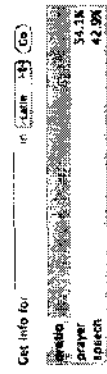
10

Alignment Postprocessing

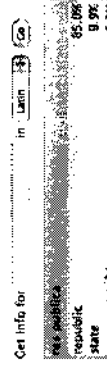
- " Harmonize the two alignments (Latin/Greek-English, English-Greek/Latin) by taking their intersection and adding additional neighbors (Och and Ney 2003).
- " Re-project the original word tokens onto the stemmed alignments and lemmatize ambiguous forms using evidence from unambiguous tokens:
 - e.g., *est* can be derived from two lemmas: *sum* and *edo*.
 - edint* is an inflection of only *edo*.
- " if we have an alignment that *edint* = eat then we have evidence that the lemma *edo* = eat
- " if *est* is aligned to eat then we have evidence that *est* = *edo*.



Sense Induction



This sense inventory has been automatically created using parallel texts. The probability associated with each word in the inventory is the probability that the sense word is associated with the word. This word can be derived from the lemma 55323.



This sense inventory has been automatically created using parallel texts. The probability associated with each word in the inventory is the probability that the sense word is associated with the word. This word can be derived from the lemma 55323.

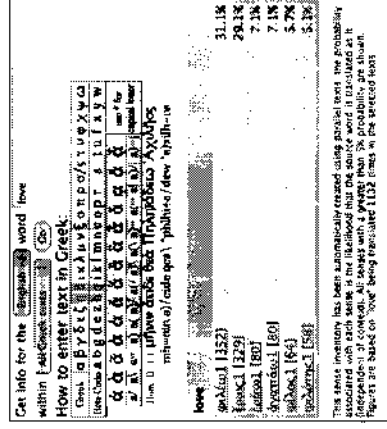


Automatically Building Bilingual Dictionaries for Greek and Latin

11

Lexicon in a Digital Library

- " Explore the different Greek concepts that correspond to the English word love



Lemma	Probability
φιλοτιμία (55323)	51.1%
φιλοτιμία (55323)	29.1%
φιλοτιμία (55323)	7.1%
φιλοτιμία (55323)	7.1%
φιλοτιμία (55323)	5.2%



Automatically Building Bilingual Dictionaries for Greek and Latin

12

Lexicon in a Digital Library

" Find passages in a digital library where love **ἀγαπάω**

<p>Matthew 5: εἰν γὰρ ἀγαπήσατε τοὺς ἀγαπῶντας ὑμᾶς, τίνα μισθὸν ἔχετε; (for if you love those who love you, what reward do you have?)</p>
<p>Matthew 24: καὶ διὰ τὸ πληθυνθῆναι τὴν ἀνομίαν ποιεῖται ἡ ἀγάπη τῶν πολλῶν, (because iniquity will be multiplied, the love of many will grow cold.)</p>
<p>Mark 10: ὁ δὲ Ἰησοῦς ἀμβλέψας αὐτῷ ἠγάπησεν αὐτὸν καὶ εἶπεν αὐτῷ "὘ν σε ὑπερέπε- (Jesus looking at him loved him, and said to him, one thing you lack.)</p>



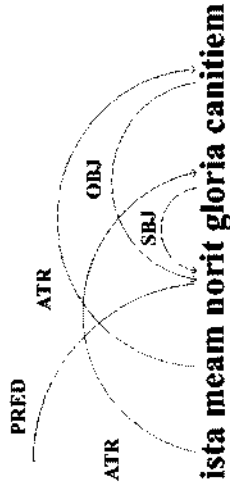
II. Syntax

<p>Βίβλος: ἐν, πρῶμ, ἰ Latin texts</p> <p>I. Translation equivalents</p> <ul style="list-style-type: none"> ► accusative (49,376) (1320) ► dative (17,581) (438) ► Genitive (3) ► Substantive (9) ► Adjective (59) ► Vulgata (12) ► Cypriota (3) ► Gen. A. B. ► Gen. U. L. L. ► Gen. M. L. I. ► Encheiridion (12) <p>► accusative (8,766) (143)</p> <p>II. Suborganization</p> <ul style="list-style-type: none"> ► SSB OBL (4,681) (143) ► SSB OBL (9,084) (564) <p>III. Schoolroom preferences</p> <ul style="list-style-type: none"> ► SB ► OBL ► All auxiliary ► Caesar ► Caesari ► periphrasis (3,061) (143) ► metre (14,110) ► metre (14,110) ► metre (4,871) (143) ► metre (4,871) (143) <p>► Jaeger</p> <ul style="list-style-type: none"> ► name (4,161) (39) ► name (4,161) (39) ► name (4,161) (39) ► name (4,161) (39)
--



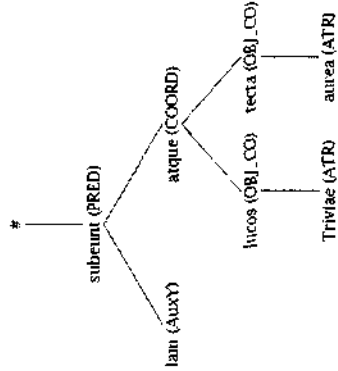
Latin and Ancient Greek Dependency Treebanks

" Large collections of syntactically parsed sentences in which the syntactic relation of every word is explicitly annotated. Given the flexible word order in both Latin and Greek, we have based our annotation style on the dependency grammar of the Prague Dependency Treebank (Hajic 1998) for Czech.

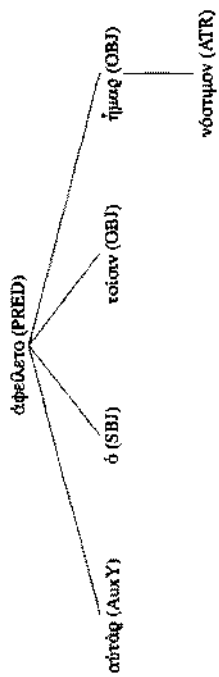


LDT 1.5

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, <i>Odyssey</i>	48,123
Homer, <i>Iliad</i>	3,956
Total	52,079

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: 82% (Nivre and McDonald 2008)
- Accuracy (fied to treebank size: larger is better)
- " English: Penn treebank (+1 million words)
 - " Czech: Prague Dependency 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.

	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

" 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 Greek and Latin treebanks
- " Parsed all the texts in our 3.5M word Latin corpus, 4.9M word Greek corpus
- " To find selectional preferences from this noisy data, we 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 $p = \frac{c_{12}}{N}$, $p_1 = \frac{c_1}{N - c_1}$, $p_2 = \frac{c_2 - c_{12}}{N - c_1}$, $N = \text{corpus count}$
 and $L(a, b, c) = c^a(1 - c)^{b-a}$.

Dynamic Lexicon: do (to give)

Latin	English	OLD def.	Log score
Opera	Service (= take pains)	22c	254.2
Obses	Hostage	11a	21.8
Signum	Sign	-	12.6
Velum	Sail (= set sail)	18f	7.9
Pecunia	Money (= pay)	6a	7.3
Negotium	Business	-	6.2
Poenā	Penalty	7b	5.6
Possessio	Possession	1c	4.8
Littera	Letter (for delivery)	10a	4.3

Dynamic Lexicon: do (to give)

Latin	English	Log score
Obses	Hostage	18.4
Opera	Service	11.9
Suspicio	Suspicion	2.2
Facultas	Faculty	1.8
Signum	Sign	1.5

Strongest OBL of do in Caesar

Latin	English	Log score
Osculum	Kiss	8.5
Velum	Sail	5.9
Munus	Gift	3.5
Signum	Sign	2.6

Strongest OBL of do in Oros

Dynamic Lexicon

δύναμις

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

Attributes:

- ναυτική ("naval force": 15.01/31, Polybios)
- στρατός ("land army": 12.45/12, Polybios)
- κράτος ("great power": 4.32/15, Diochris)
- δύναμις ("power": 4.49/15, Diochris)
- στρατός ("this power": 8.24/102)

Object of:

- ἔχει ("having as much power": 8.93/239, Plinio)
- ἀφίξει ("to army": 2.40/16, Polybios)
- ἀδελφεὸς ("gather all together army": 2.32/15)
- ἐπὶ ("pointing": 2.16/25, Epictetus, Plinio)

Example sentences:

- ἡ δύναμις ἡ ἀναστὰς ("the reasoning faculty": Epict, 1.1)
- ἀφίξει ὁ ἐπὶ τοῦ ἀσπίδος, καὶ ἐπιβήσεται, Ἐρσίου Νόμου, ἡμετέρας καὶ ἀσπεράτης, (Cicero, Aristot. Met. 3.104.5b)
- ἀπὸς ἀδελφῶν περὶ τὴν πόλιν, ἡ δύναμις ἡ ἀσπίδος, (Cicero, Aristot. Met. 3.104.5b)