TREEBANKS II: QUERYING TREEBANKS

by Francesco Mambrini
Università cattolica del Sacro Cuore,
Milan (Italy)

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

OUTLINE OF THE SESSION

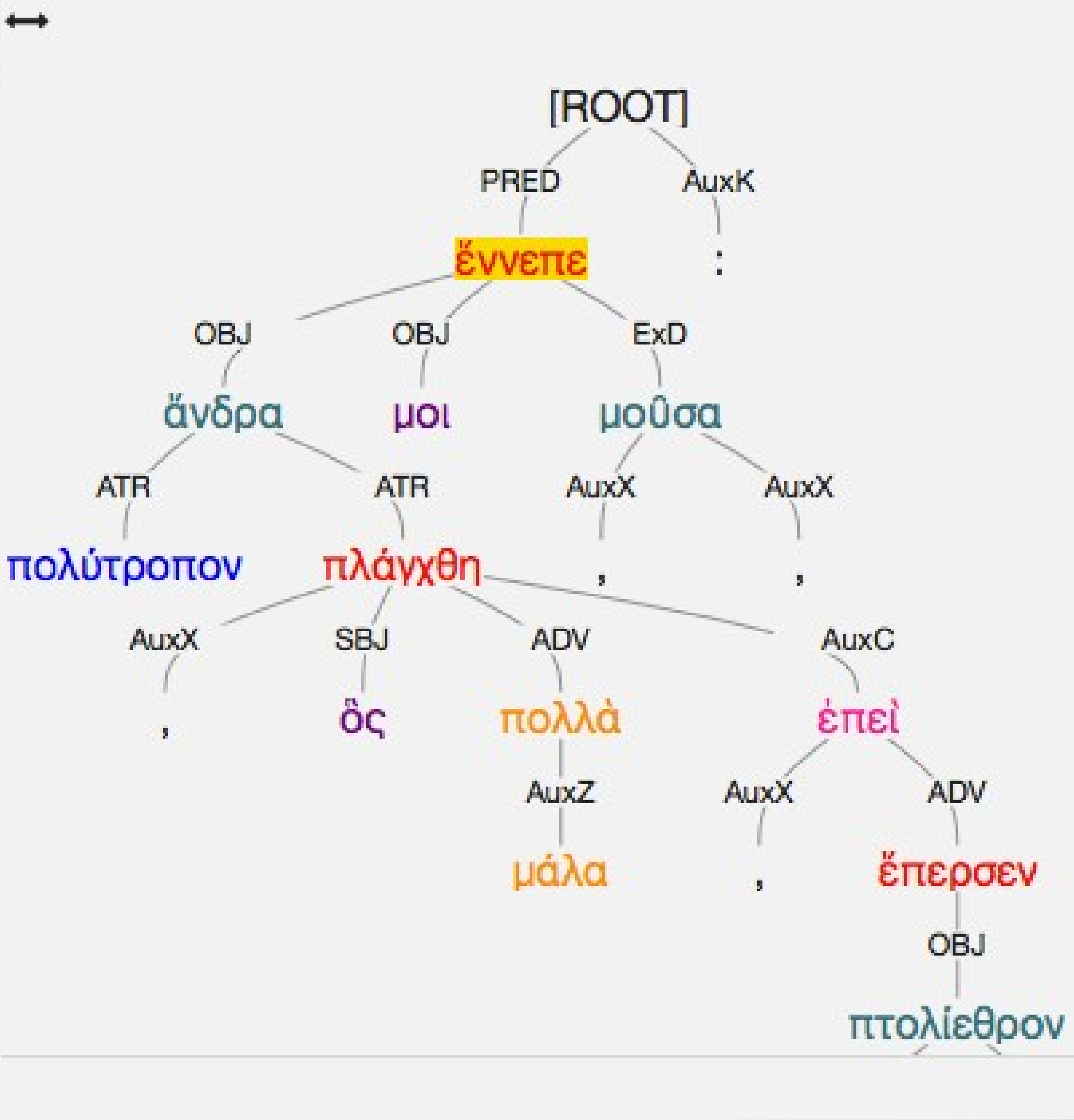
1. General introduction:
   - Applications of treebanks (overview)
   - Linguistic/textual research: querying

2. Natural Language Processing applications (Toon Van Hal)

3. Example of queries

TREEBANK

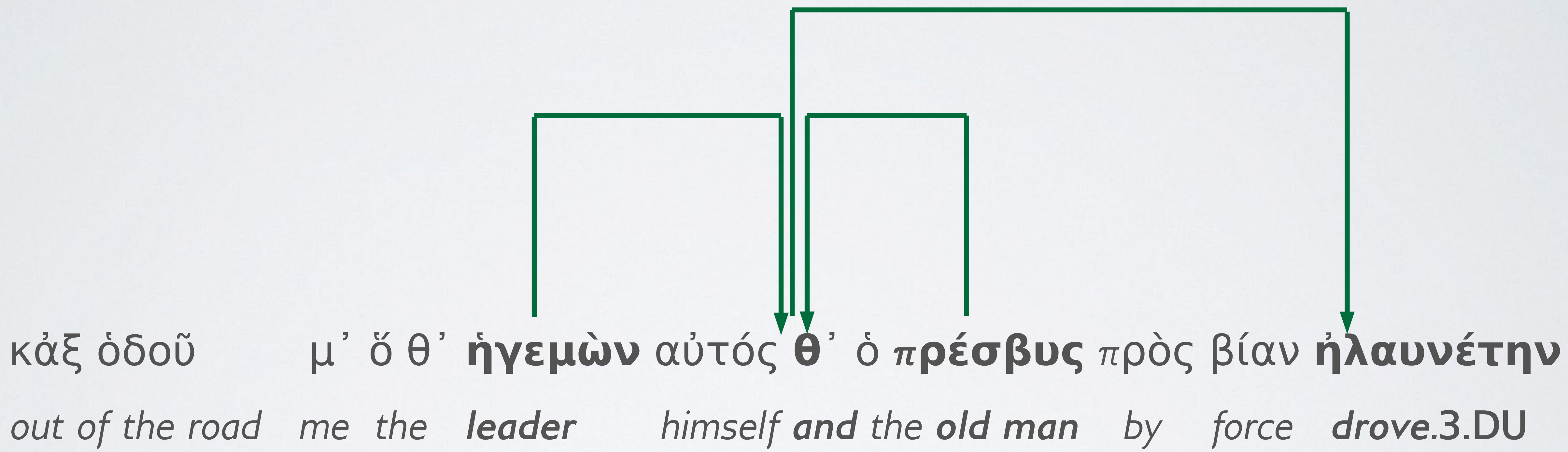
- (a type of) a linguistically **annotated corpus** (in digital format)
- **Morphology** (PoS and feats)
- A representation of the syntactic structure of the sentences
- Available in:
  - **several standards** and annotation styles
  - for **several languages** (and language strata/modalities)



WHAT I DID WITH THEM

- Non-projectivity (discontinuous phrases) in the AGDT
- Agreement pattern with coordinated subjects
- Nominal VS Copular clauses in Hdt., Thuc., and Polybius
- (in progress) The syntax of the Sophoclean characters

Soph. OT 805-6



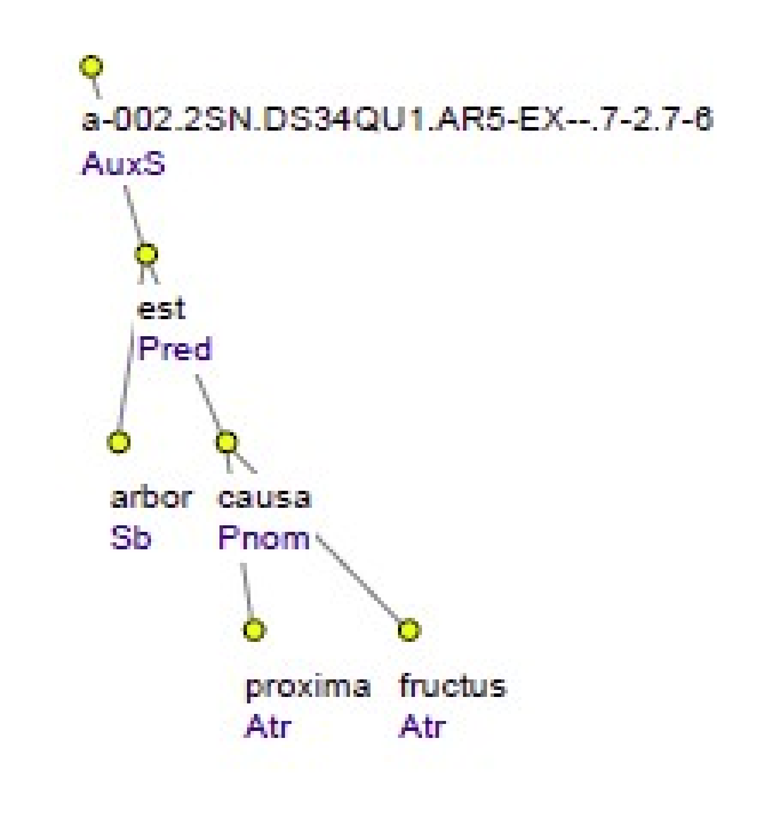
κἀξ ὁδοῦ μ᾽ ὅ θ᾽ **ἡγεμὼν** αὐτός **θ**᾽ ὁ π**ρέσβυς** πρὸς βίαν **ἠλαυνέτην**

out of the road *me* *the* ***leader*** *himself* ***and*** *the* ***old man*** *by* *force* ***drove***.3.DU

1. CHOOSE YOUR DATASET

- Starts as **a parallel TB** of IE languages
- For each: translation of the New Testament + some prose texts for comparison
- Guidelines are similar to that of Perseus’ Treebanks, but not quite identical!
- Greek: NT, Herodotus, Sphrantzes (15th CE)
- Latin: Vulgate, Caesar, Cicero, *Peregrinatio Aetheriae*, Palladius
- http://clarino.uib.no/iness/treebanks

Index Thomisticus Treebank



https://itreebank.marginalia.it/

- Latin works of Thomas Aquinas (13th CE)
- Currently about 350k tokens
- (almost) same tagset and guidelines as Perseus
- Can be queried online via PMLTQ

Universal Dependencies

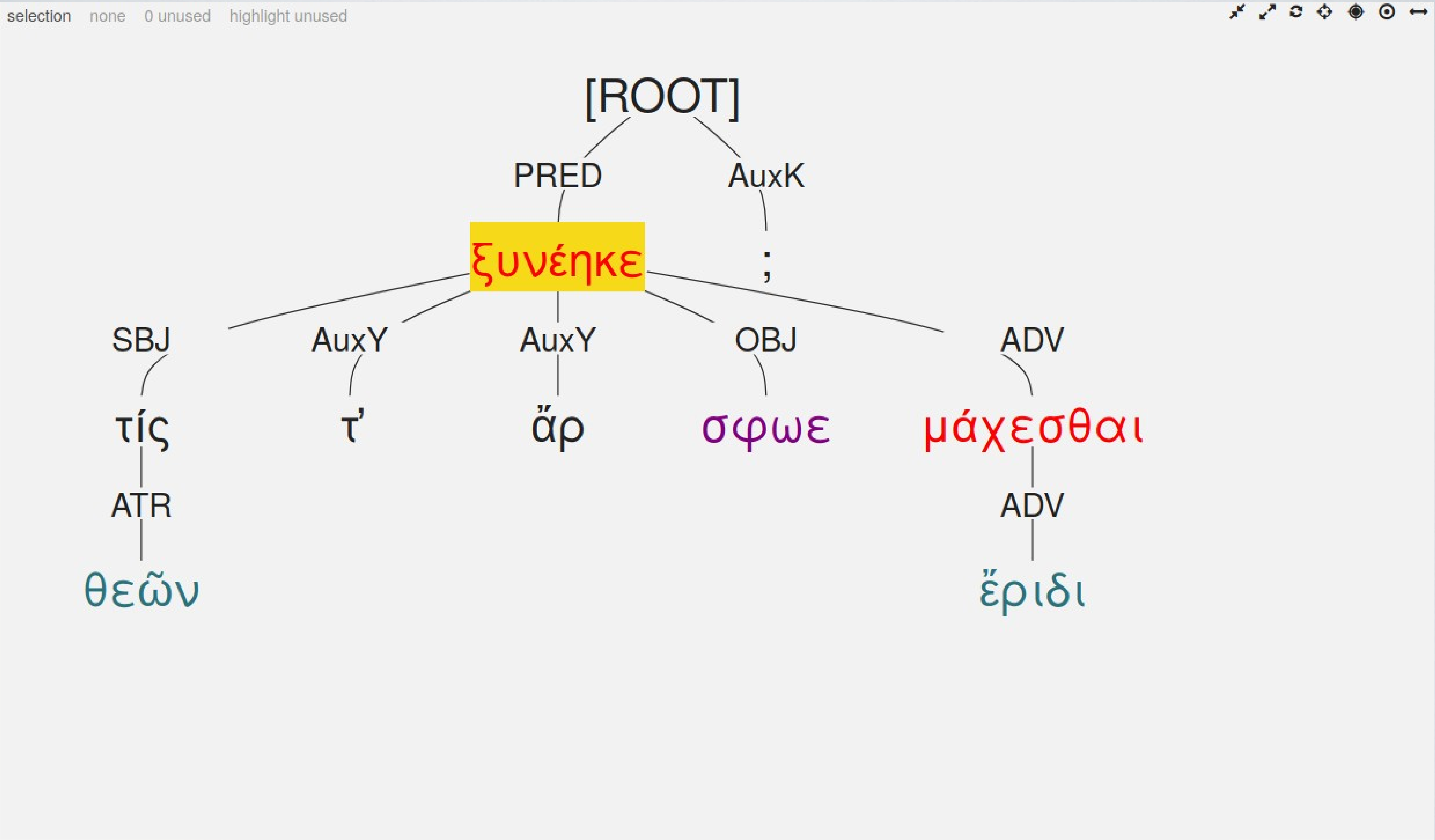
https://universaldependencies.org/

- Unified guidelines
- More than 70 languages
- Growing community
- Efforts to go beyond dependency syntax
- Lots of tools and software available (also for querying)
- Major TBs for Latin and Greek have a UD version (sort of…)

2. FORMULATE YOUR RESEARCH QUESTION

I’d like to know…

What are the most frequent subjects of verbs composed with the pre-verb “*syn*” *(“xyn”)* in Homer?


selection
none
0 unused
highlight unused
[ROOT]
PRED
AuxK
ξυνέηκε
;
SBJ
AuxY
AuxY
OBJ
ADV
τίς
τ’
ἄρ
σφωε
μάχεσθαι
ATR
ADV
θεῶν
ἔριδι

Formalize your query

- A verb (morphology)
- That must start with the letters συν
- And has at least one subject
- We want to extract its subject(s) and count them

Exercise

We know that in Ancient Greek, neuter plural subjects trigger either plural or singular agreement with the verb. This is supposed to be a relic of an old Indo-European collective number. How frequently does this happen in Homer? And in Aeschylus? Which agreement pattern is more frequent?

Try to use “Iliados” to answer these questions

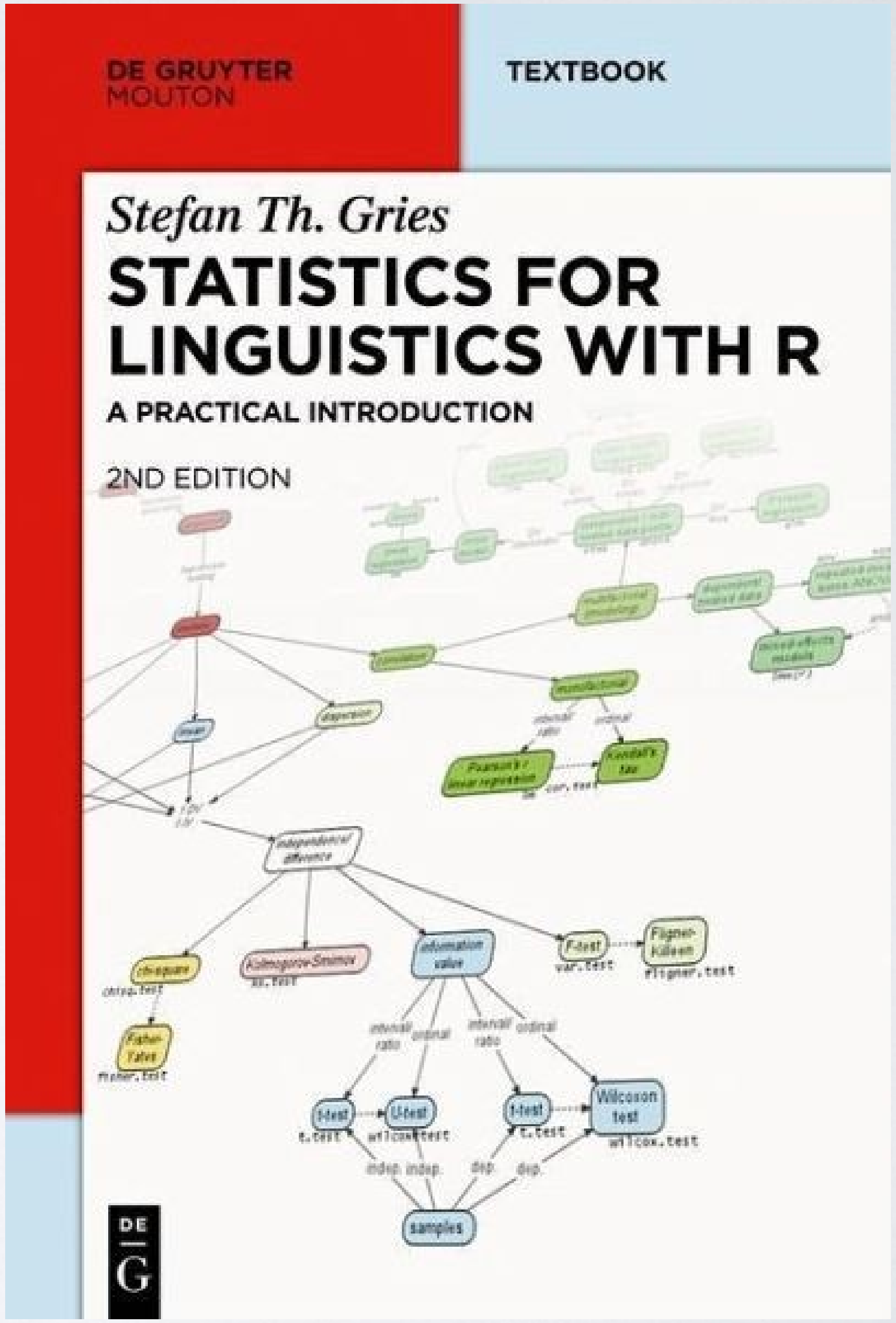
3. FAMILIARIZE WITH THE QUERY LANGUAGE/SOFTWARE

Check if your tool support...

- Unlimited **nesting** of trees (e.g. NOUN > ADJ > NOUN > ADJ...)
- **Word-order-based** queries (NOUN > ADJ where ADJ precedes/follows the NOUN)
- **Negative** constraints (all NOUN that govern no ADJ)
- **GUI** to build the query graphically
- **Boolean** operators (AND, OR)
- Some **math** operations (count, mean...)
- Also, check out what format you can output your results to (txt, csv, json, html...)

Wait, there is more!

- Corpus linguistics
- Methodology of quantitative research
- Statistics…



AUTOMATING THE LINGUISTIC ANALYSIS OF ANCIENT GREEK

Alek Keersmaekers KU Leuven & FWO

Toon Van Hal (presenter) KU Leuven

SunoikisisDC 2020 session 8

Using Treebanks

CONTENTS

Three questions:

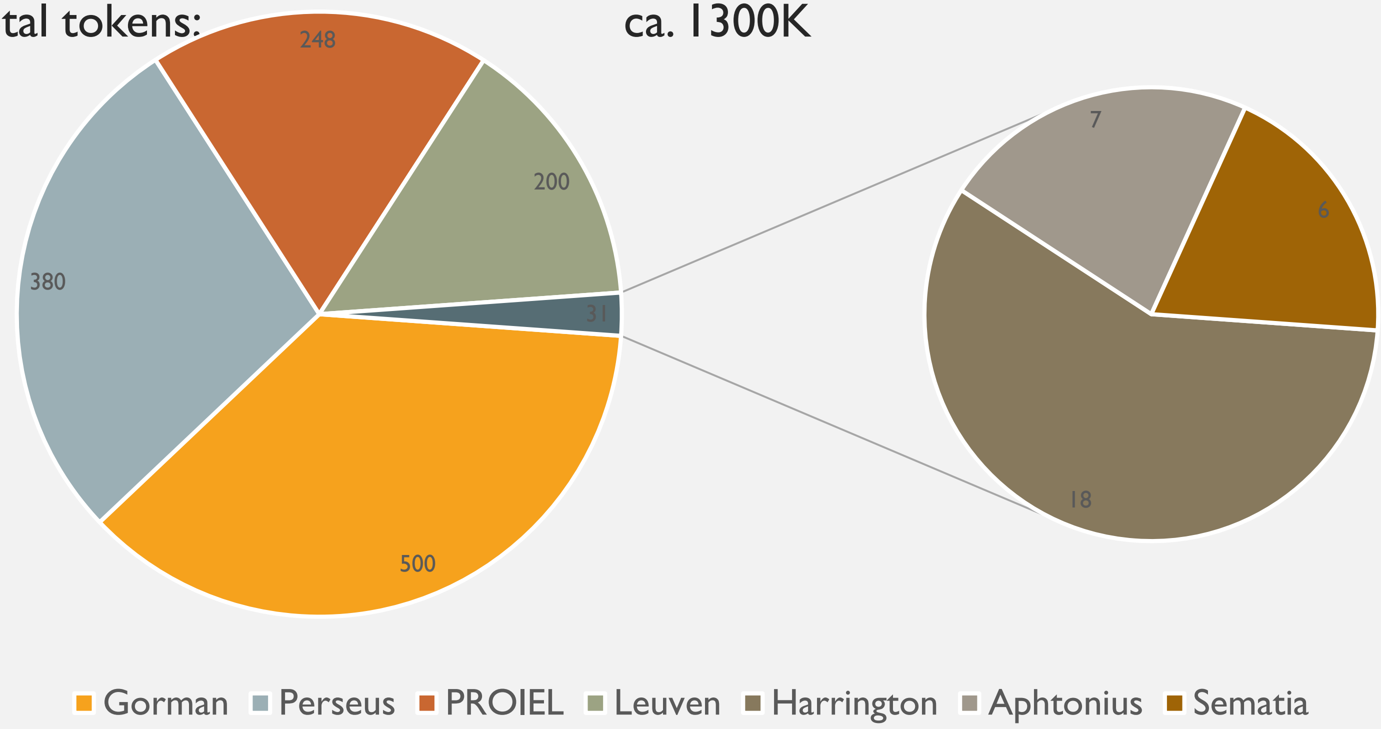
- What are the starting points of automated analysis and how does automated analysis interlock with this course?
- What is the way of proceeding in generating automated treebanks?
- What can we do with such automated (and hence: imperfect) treebanks?

I. STARTING POINTS

- This is a course about "how and why using treebanks"
- Manually treebanks can be used to
  - come to a better understanding of Ancient Greek (cf. Francesco Mambrini's presentation)
  - create … more treebanks
- The data of these treebanks are used as example data ("training data") from which machines can learn.
- We are making use of present-day technology
  - Machine learning, Artificial Intelligence, Neural networks, Natural Language Processing
- Our focus is not on the development of software, but on using existent software and on working with the Ancient-Greek data

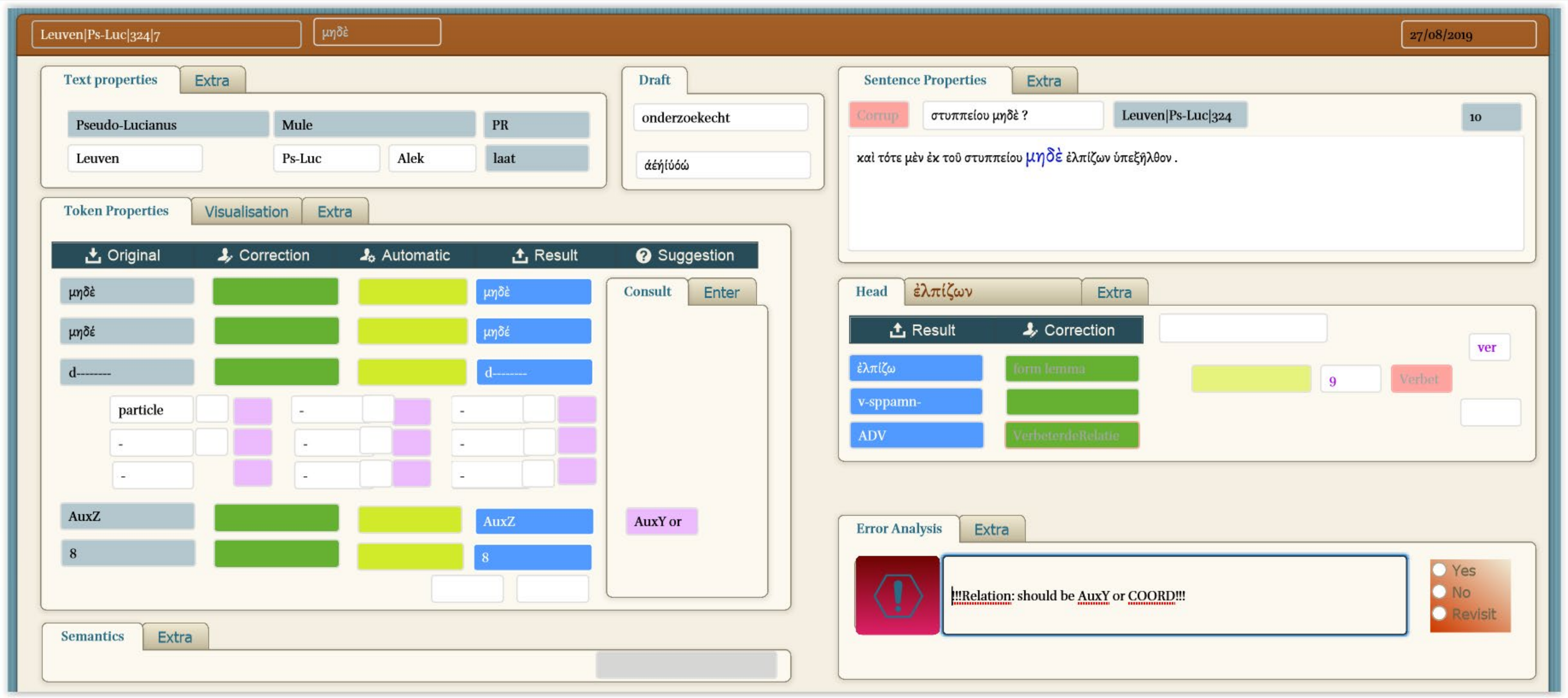
1. STARTING POINTS

- The training data must be **extensive**: integration of several existing treebanks
- Total tokens: ca. 1300K



I. STARTING POINTS

- The training data must be **reliable** and **homogenous**
- This project makes use of several existing corpora of Ancient Greek, each with their own differences in the annotation of specific Greek constructions
- As a result, there are a lot of inconsistencies (even sometimes in the same text from the same annotator!)
- Consistency important for NLP tasks as well as corpus linguistic research
- Therefore we integrated all these treebanks into a database (FileMaker) and are systematizing the data as much as possible



Leuven|Ps-Luc|324|7 μηδὲ 27/08/2019

Text properties Extra

Pseudo-Lucianus	Mule	PR
Leuven	Ps-Luc Alek	laat

Draft

onderzoekecht

άέήίὑόώ

Sentence Properties Extra

Corrup στυππείου μηδὲ ? Leuven|Ps-Luc|324 10

καὶ τότε μὲν ἐκ τοῦ στυππείου μηδὲ ἐλπίζων ὑπεξῆλθον .

Token Properties Visualisation Extra

Original	Correction	Automatic	Result	Suggestion
μηδὲ			μηδὲ	
μηδέ			μηδέ	
d--------			d--------	
particle	-	-		
-	-	-		
-	-	-		
AuxZ			AuxZ	AuxY or
8			8	

Consult Enter

Head ἐλπίζων Extra

Result	Correction
ἐλπίζω	form lemma
v-sppamn-	
ADV	VerbeterdeRelatie

ver 9 Verbet

Error Analysis Extra

!!!Relation: should be AuxY or COORD!!!

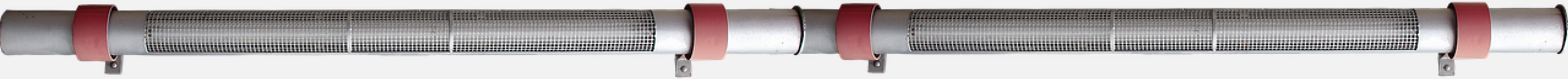
○ Yes
○ No
○ Revisit

Semantics Extra

I. STARTING POINTS

- The training data must be **<u>easy to process</u>** for a computer
- Thorny issues are, for example,
  - ellipsis
    - words or constituents are missing in the sentence, even though they are implied
  - coordination structures:
    - e.g. "He ate big burgers and sandwiches."
    - 'horizontal elements', difficult to represent in a ('vertical') tree structure
- Finding workarounds in the back-office environment: manipulating the data in such a way that they become more 'digestible' for computers

2. CREATING A PIPELINE



- What is a Natural Language Processing Pipeline?
  - The design of a process where the output of module A feeds to the input of module B, whose output feeds to the input of module C, etc.

2. CREATING A PIPELINE

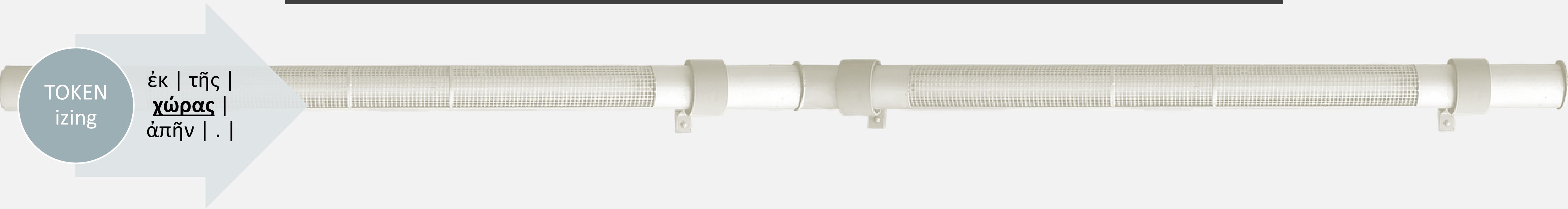
2. CREATING A PIPELINE

καὶ τοῦτ’ ἐποίουν ἕως ἐκ τῆς **χώρας** ἀπῆν.

(sentence in Goodwin’s syntax after Xenophon)

And this they continued to do until he had quitted their borders

2. CREATING A PIPELINE

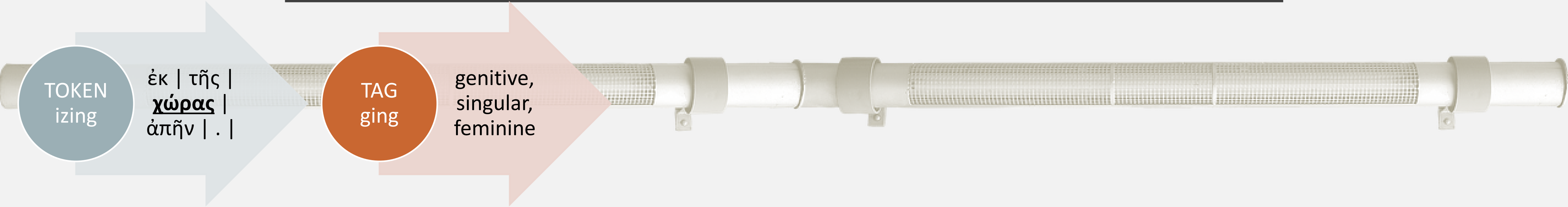


TOKEN izing

ἐκ | τῆς | **<u>χώρας</u>** | ἀπῆν | . |

- Tokenization is the process of converting a string of written language into a sequence of tokens (‘words’, interpunction)
- The process is rule-based (based on spaces, interpunction, *krasis*)

2. CREATING A PIPELINE



TOKEN izing

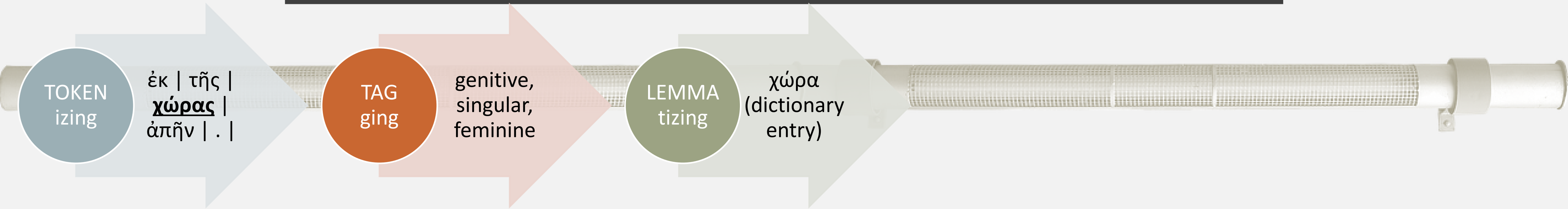
ἐκ | τῆς | **<u>χώρας</u>** | ἀπῆν | . |

TAG ging

genitive, singular, feminine

- Part-of-speech (POS) tagging is the process of assigning each word in a text a specific POS-tag (and specific attributes)
- Technology used: RFTagger
- Accuracy: about 90% (at worst) to 96% (at best)

2. CREATING A PIPELINE



TOKEN izing

ἐκ | τῆς | **χώρας** | ἀπῆν | . |

TAG ging

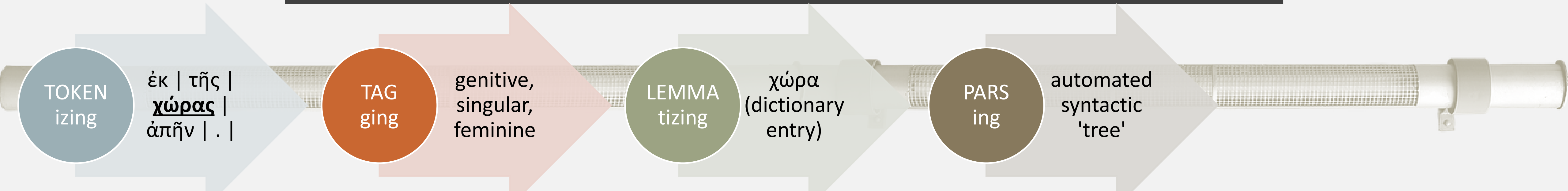
genitive, singular, feminine

LEMMA tizing

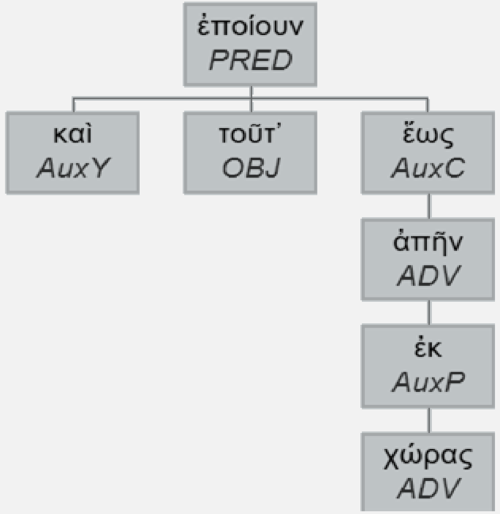
χώρα (dictionary entry)

- Lemmatizing is the process of assigning each token one specific lemma in a dictionary
- Technology used: MarMoT
- Possibilities of integrating existing dictionaries
- Accuracy: 96% (at worst) to 99.5% (at best)

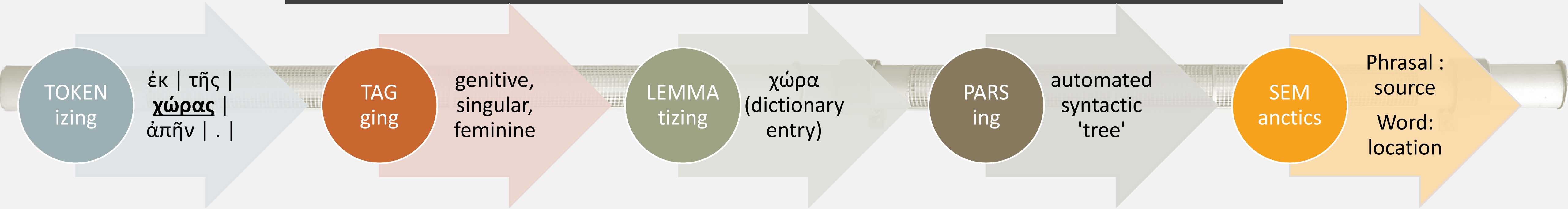
2. CREATING A PIPELINE


TOKEN
izing
ἐκ | τῆς | **χώρας** | ἀπῆν | . |
TAG
ging
genitive, singular, feminine
LEMMA
tizing
χώρα (dictionary entry)
PARS
ing
automated syntactic 'tree'

- Parsing is the process of structurally representing sentences
  - Relations: ADV, OBJ, …
  - Heads
- Technology used: MaltParser in first tests, and recently Turku Neural Parser
- Accuracy: somewhere between 80-90%. Difficult to assess and to a large dependent on authors (Aristotle is e.g. notoriously difficult)



2. CREATING A PIPELINE



TOKEN izing

ἐκ | τῆς | **<u>χώρας</u>** | ἀπῆν | . |

TAG ging

genitive, singular, feminine

LEMMA tizing

χώρα (dictionary entry)

PARS ing

automated syntactic 'tree'

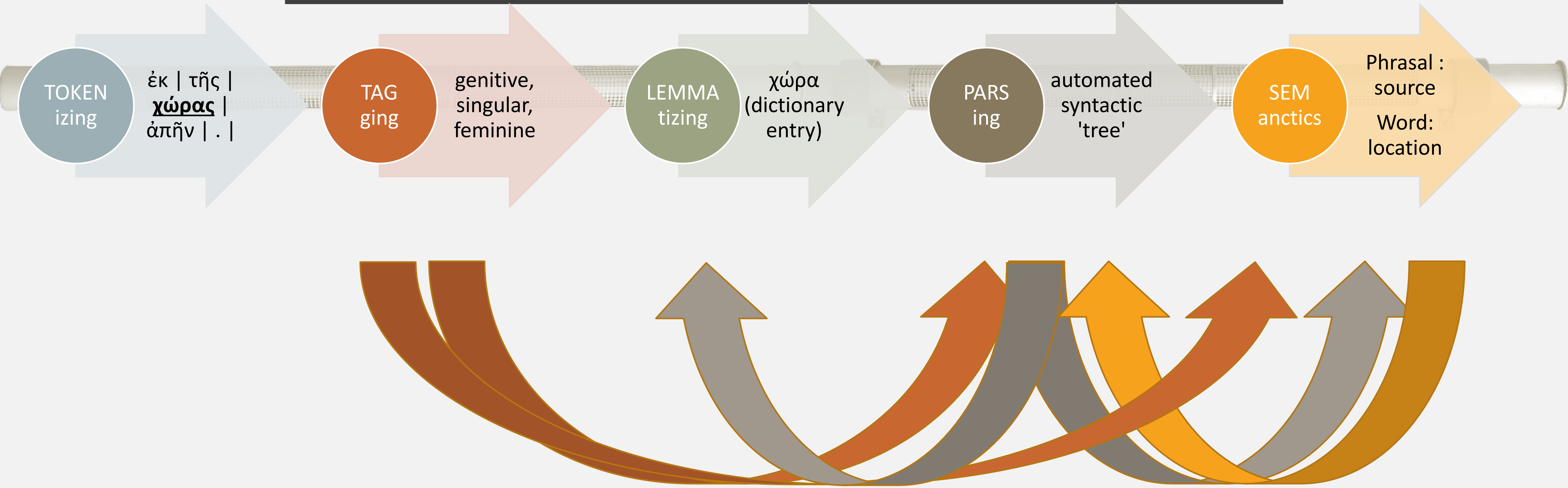
SEM anctics

Phrasal : source

Word: location

- Semantic annotation on various levels (ongoing work by Alek Keersmaekers)
- (Ia) On the <u>word</u> level: word vectors, using a large corpus (37 million tokens) as input material. This allows us to find synonyms: e.g. ἤμερας ~ ἔτη, ἐνιαυτούς etc.
- (Ib) Annotation of noun categories (e.g. animal, person, non-concrete etc.), verb categories (e.g. emotion, cognition, motion etc.), adjective categories (e.g. quantifier/qualifier), also using word vectors as input
- (II) On the <u>phrasal</u> level: semantic roles

2. CREATING A PIPELINE



TOKEN izing

ἐκ | τῆς | **χώρας** | ἀπῆν | . |

TAG ging

genitive, singular, feminine

LEMMA tizing

χώρα (dictionary entry)

PARS ing

automated syntactic 'tree'

SEM anctics

Phrasal : source

Word: location

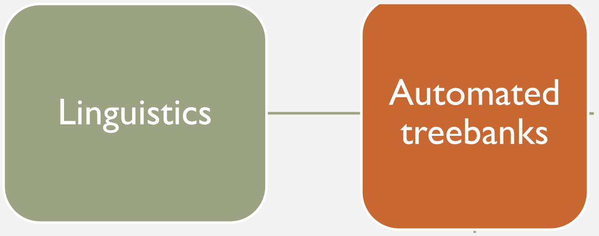
3. WHAT CAN WE DO WITH IT?

- With this pipeline, we were able to automatically analyze
  - The Greek literary corpus (about 32 million tokens)
  - The papyrus corpus (about 4.5 million tokens)
- We can speed up manual annotation by correcting preprocessed data
  - See exercise 2
  - The Leuven treebanks (200K tokens) are all (except for one) first automatically analyzed and then manually corrected
  - They are annotated with Arethusa by ourselves, job students and thesis students

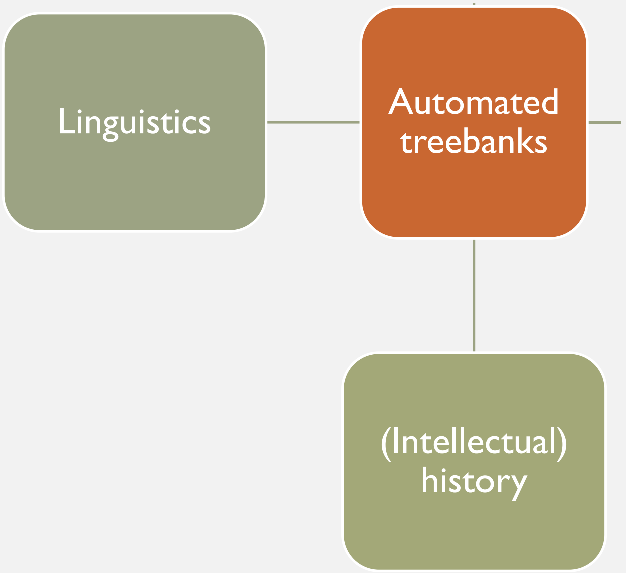
3. WHAT CAN WE DO WITH IT?

Automated treebanks

3. WHAT CAN WE DO WITH IT?



3. WHAT CAN WE DO WITH IT?

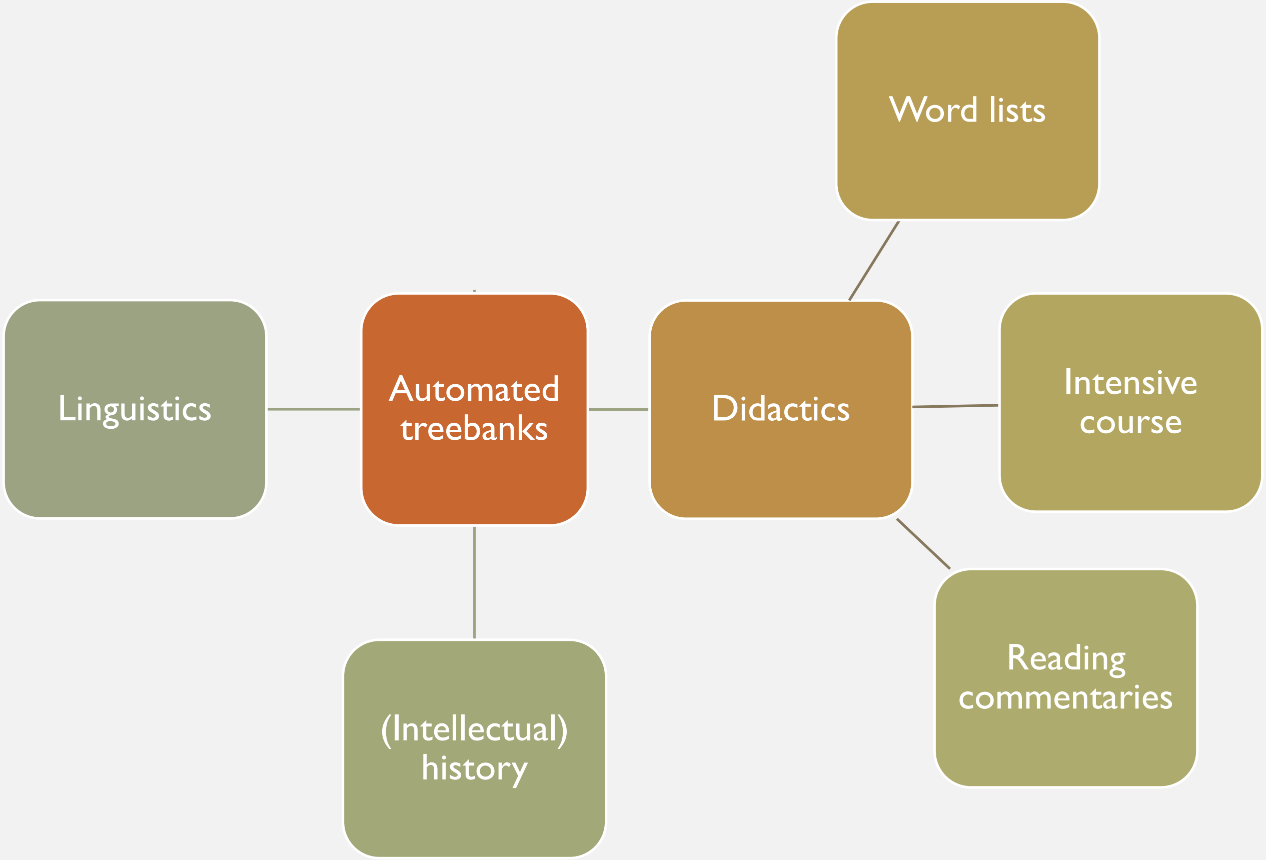


Linguistics

Automated treebanks

(Intellectual) history

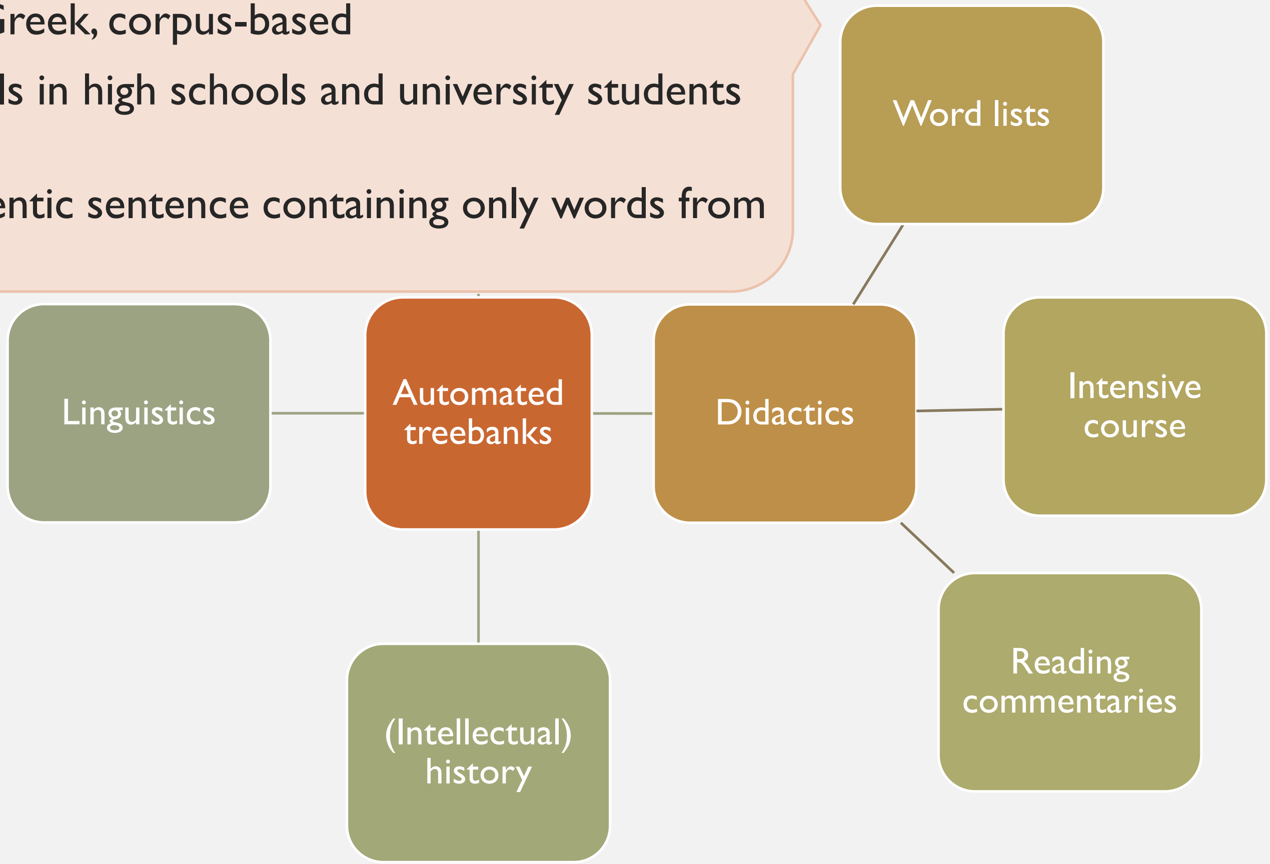
3. WHAT CAN WE DO WITH IT?



3. WHAT CAN WE DO WITH IT?

Chilia

- A list of 1000 ‘key words’ of Ancient Greek, corpus-based
- Context: pedagogical material for pupils in high schools and university students without previous knowledge of Greek
- Every word is illustrated with an authentic sentence containing only words from this Chilia list.



Word lists

Linguistics

Automated treebanks

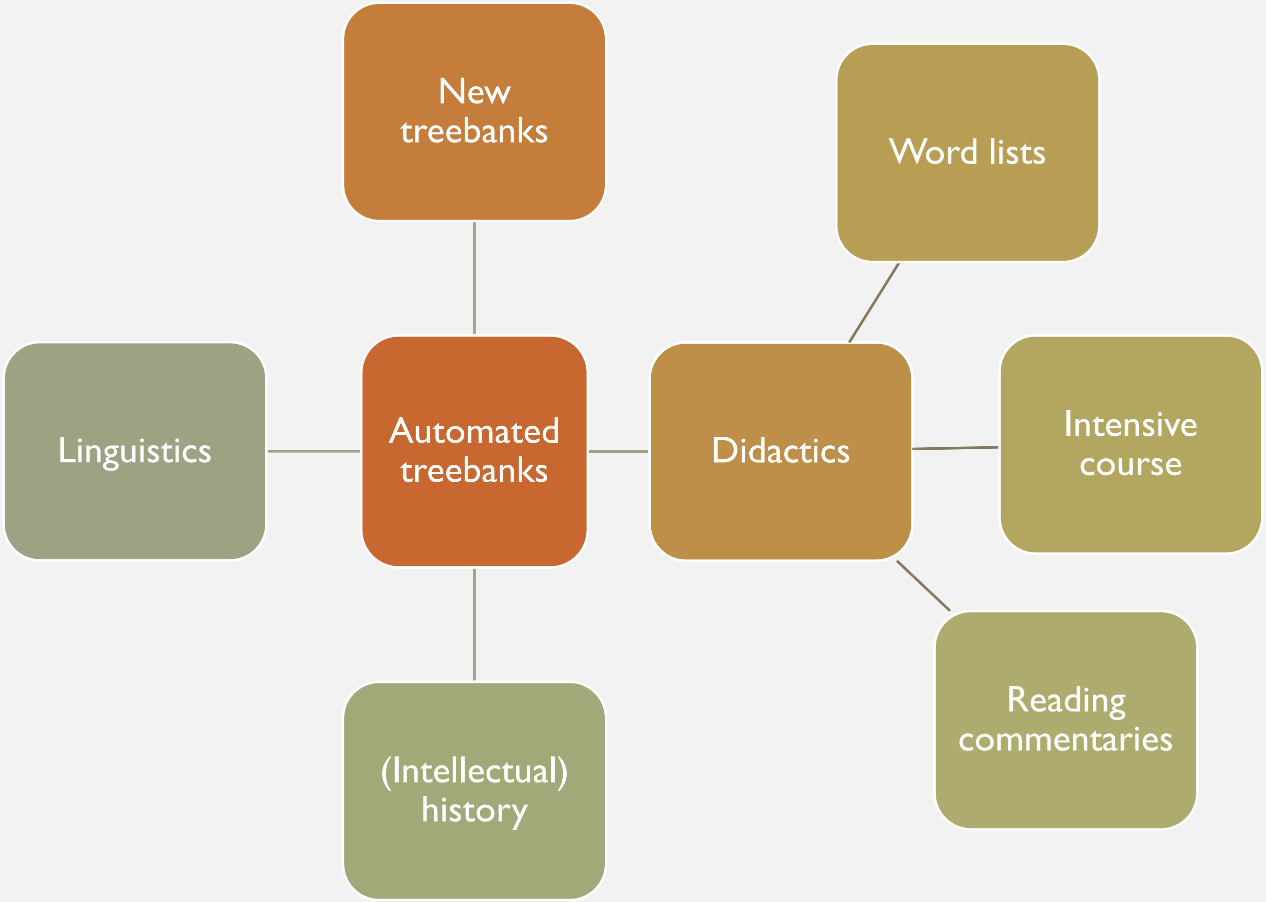
Didactics

Intensive course

Reading commentaries

(Intellectual) history

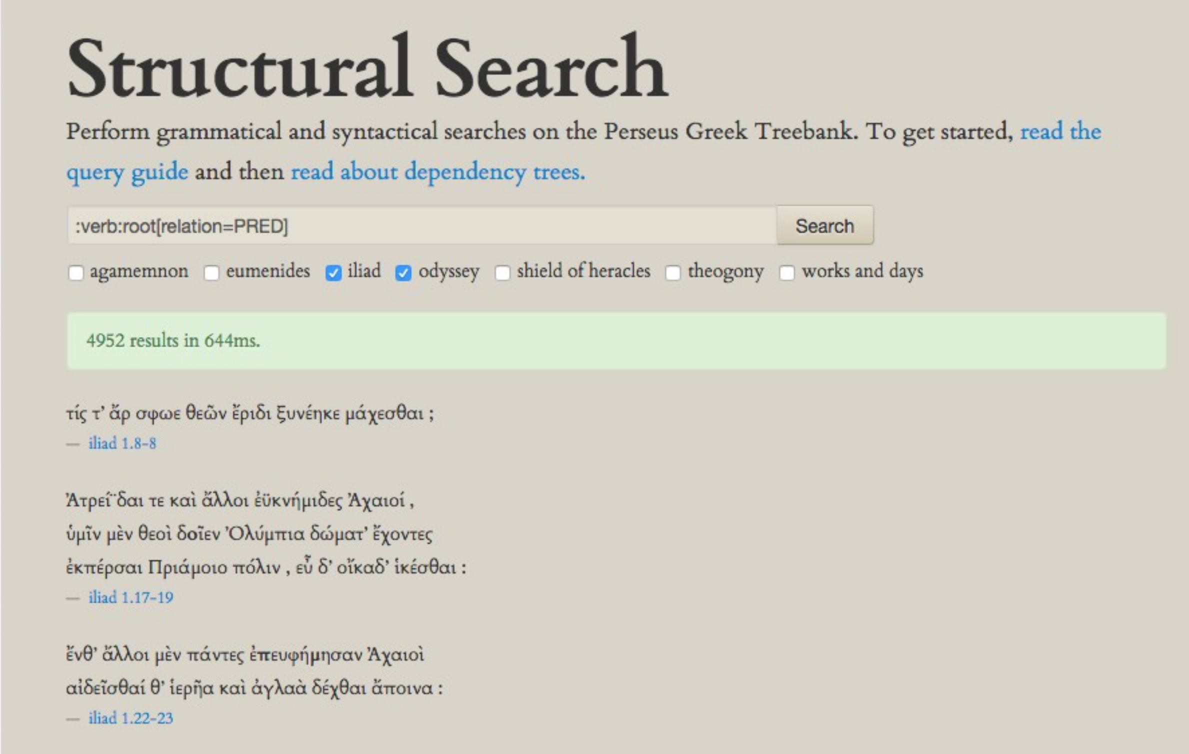
3. WHAT CAN WE DO WITH IT?



PART II
PRACTICAL EXAMPLES

STRUCTURAL SEARCH

A simple but very powerful solution to query (some of) the AGDT, based on CSS3 selector syntax



Structural Search

Perform grammatical and syntactical searches on the Perseus Greek Treebank. To get started, read the query guide and then read about dependency trees.

:verb:root[relation=PRED]

☐ agamemnon ☐ eumenides ☑ iliad ☑ odyssey ☐ shield of heracles ☐ theogony ☐ works and days

4952 results in 644ms.

τίς τ’ ἄρ σφωε θεῶν ἔριδι ξυνέηκε μάχεσθαι ;
— iliad 1.8-8

Ἀτρεί¨δαι τε καὶ ἄλλοι ἐϋκνήμιδες Ἀχαιοί ,
ὑμῖν μὲν θεοὶ δοῖεν Ὀλύμπια δώματ’ ἔχοντες
ἐκπέρσαι Πριάμοιο πόλιν , εὖ δ’ οἴκαδ’ ἱκέσθαι :
— iliad 1.17-19

ἔνθ’ ἄλλοι μὲν πάντες ἐπευφήμησαν Ἀχαιοὶ
αἰδεῖσθαί θ’ ἱερῆα καὶ ἀγλαὰ δέχθαι ἄποινα :
— iliad 1.22-23

http://www.iliados.com/