How treebanks are used in the WoPoss project to automatically annotate Latin (and Greek) texts

Francesca Dell’Oro (PI, presenter)

Helena Bermúdez Sabel

Paola Marongiu

FNSNF

Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

1. The WoPoss project and automatic annotation
2. Stanza: an NLP package to automatically annotate texts
3. Some examples
4. Conclusion: advantages and disadvantages

WoPoss. A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language

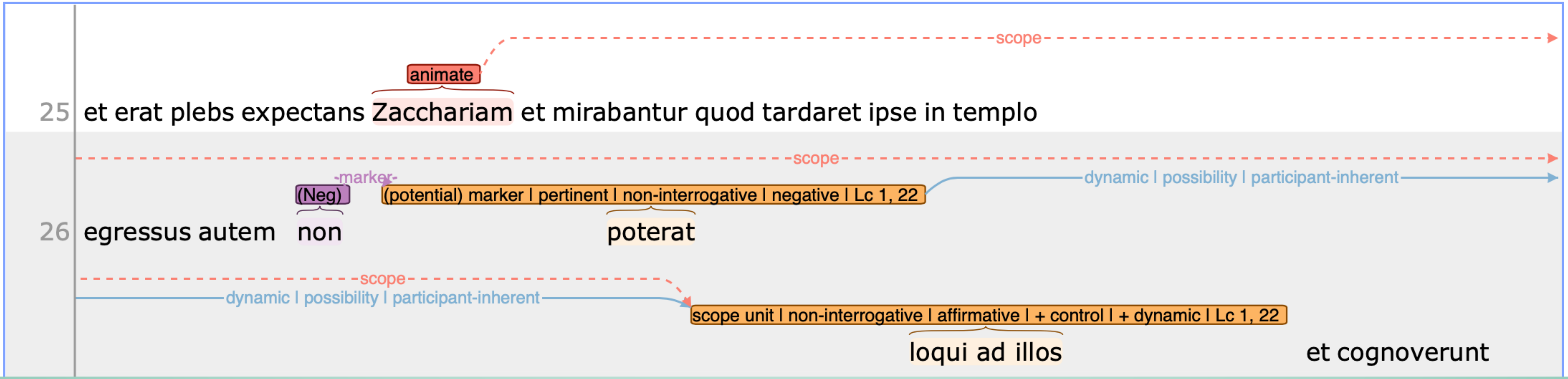
What are we interested in? In analysing modal passages

Goal: to find and analyse modal paths in the diachrony of the Latin language

How? By analysing modal passages, i.e. (to put it in a simple way) passages containing words that indicate possibility, necessity or probability: *possum*, *debeo*, *risibilis*, *faciendus*, *forsitan* etc.

Manual annotation of a modal passage

- *Non poterat loqui ad illos…* ‘He couldn’t speak to them’ (Luc 1, 22)
- What someone (something) can do: ‘to speak to them’
- What type of entity can do (or be) something: animate (human being)
- Negation: *non*



25 et erat plebs expectans Zacchariam et mirabantur quod tardaret ipse in templo

26 egressus autem non poterat loqui ad illos et cognoverunt

scope
animate
scope
-marker-
(Neg)
(potential) marker I pertinent I non-interrogative I negative I Lc 1, 22
dynamic I possibility I participant-inherent
scope
dynamic I possibility I participant-inherent
scope unit I non-interrogative I affirmative I + control I + dynamic I Lc 1, 22

In order to analyse modality…

- We need more linguistic information!

- *potest loqui* ‘someone can speak’ (= someone has now the possibility to speak OR someone has now the ability/capacity to speak)
- *potest locutus esse* ‘he may have spoken’ (= there is now the possibility that he spoke in the past)

- Cf. *poterat loqui* ‘he could speak’ (= he had in the past the ability/capacity to speak)

What we need…

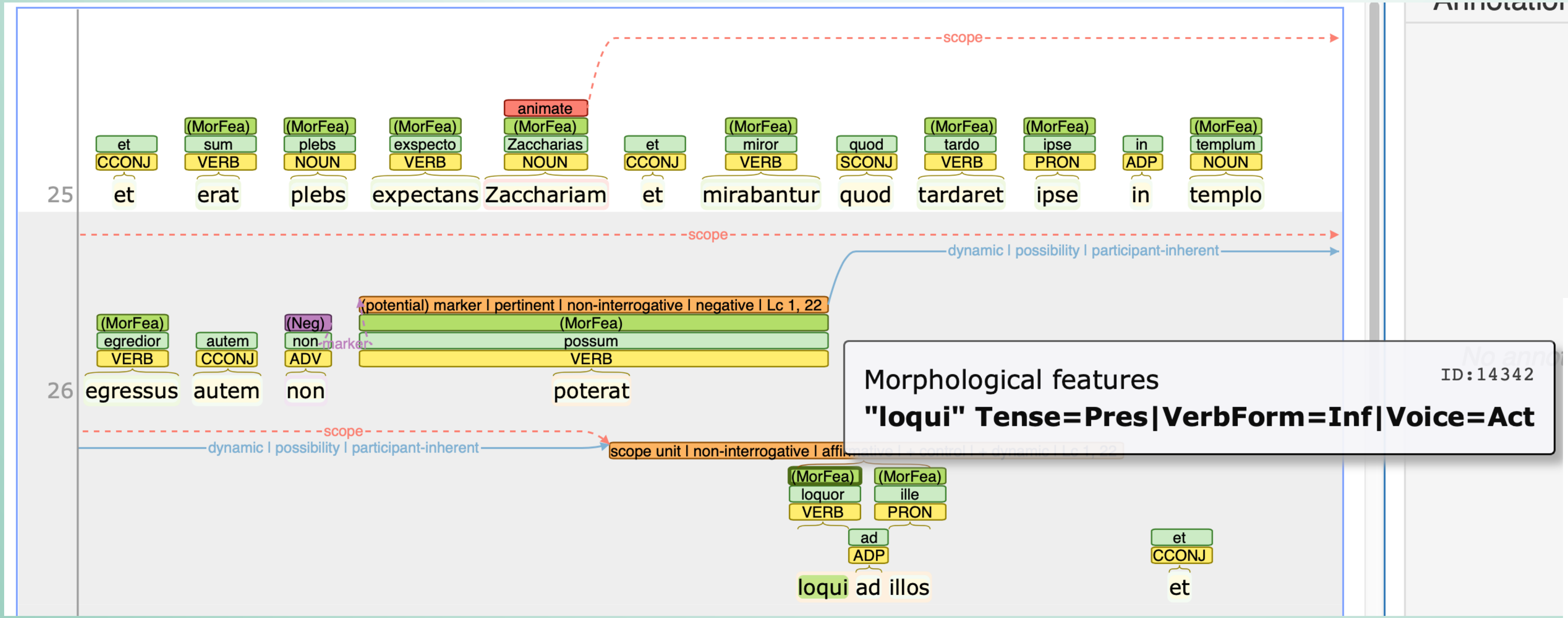
1) Lemmatisation: *loqui* –> *loquor* (light green box)

2) PoS tagging: *loquor* is verb (yellow box)

3) Morphological features: *loqui* –> present tense, infinitive form, active voice (dark green box)

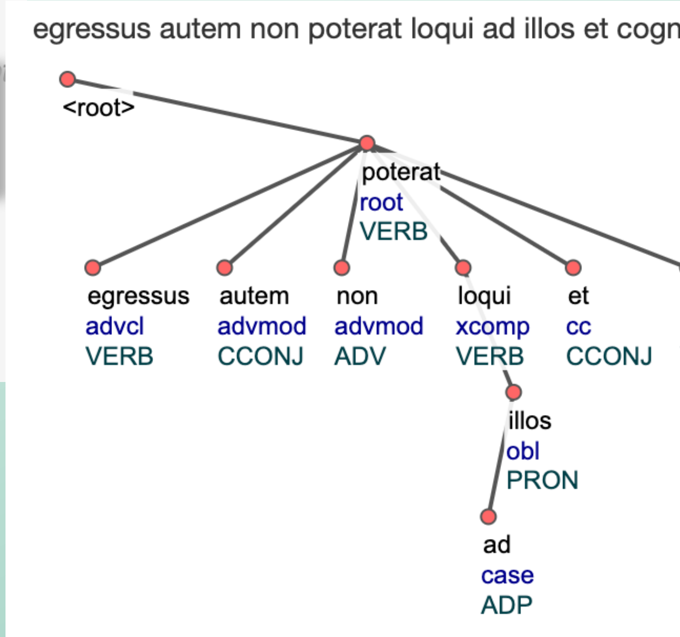
4) Dependencies

–> add the semantic analysis of modality


Annotation
25
et CCONJ et
(MorFea) sum VERB erat
(MorFea) plebs NOUN plebs
(MorFea) exspecto VERB expectans
animate (MorFea) Zaccharias NOUN Zacchariam
et CCONJ et
(MorFea) miror VERB mirabantur
quod SCONJ quod
(MorFea) tardo VERB tardaret
(MorFea) ipse PRON ipse
in ADP in
(MorFea) templum NOUN templo
scope
26
(MorFea) egredior VERB egressus
autem CCONJ autem
(Neg) non ADV non
marker
(potential) marker I pertinent I non-interrogative I negative I Lc 1, 22
(MorFea) possum VERB poterat
dynamic I possibility I participant-inherent
scope
dynamic I possibility I participant-inherent
scope unit I non-interrogative I affirmative I + control I + dynamic I Lc 1, 22
(MorFea) loquor VERB loqui
(MorFea) ille PRON illos
ad ADP ad
et CCONJ et

Morphological features ID:14342

"loqui" Tense=Pres|VerbForm=Inf|Voice=Act


egressus autem non poterat loqui ad illos et cogn

NO TIME TO CARRY OUT THE ANNOTATION FOR ALL LEVELS OF ANALYSIS

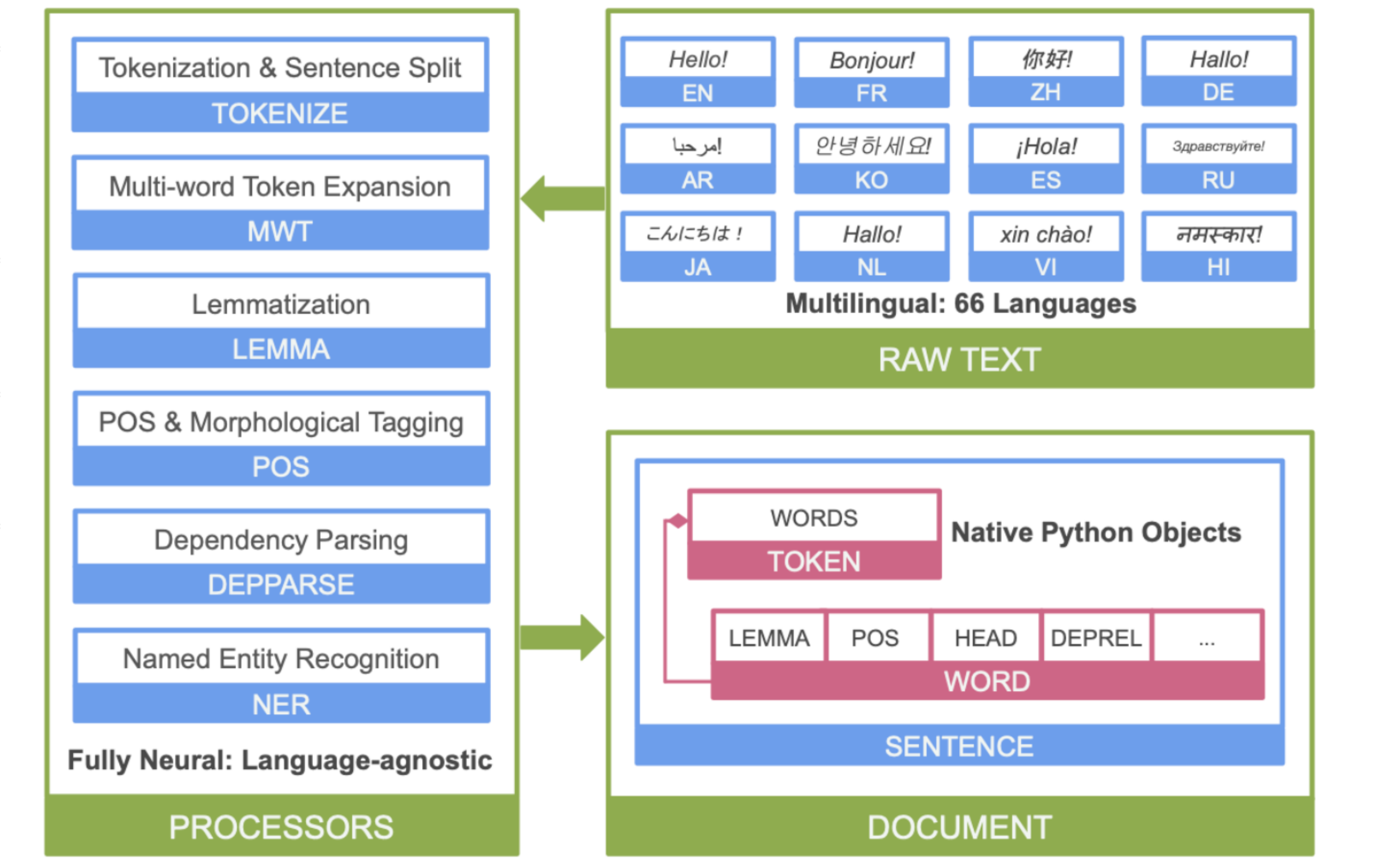
The solution: use models trained on the treebank data in order to carry out an automatic annotation of PoS, lemmas, morphological features and dependencies

Stanza

Stanza – A Python NLP Package for Many Human Languages

“Stanza is a Python natural language analysis package. It contains tools, which can be used in a pipeline, **to convert a string containing human language text into lists of sentences and words, to generate base forms of those words, their parts of speech and morphological features, to give a syntactic structure dependency parse**, and to recognize named entities. The toolkit is designed to be parallel among more than 60 languages, using the Universal Dependencies formalism.”

How Stanza works…



Tokenization & Sentence Split

TOKENIZE

Multi-word Token Expansion

MWT

Lemmatization

LEMMA

POS & Morphological Tagging

POS

Dependency Parsing

DEPPARSE

Named Entity Recognition

NER

Fully Neural: Language-agnostic

PROCESSORS

Hello! EN	*Bonjour!* FR	*你好!* ZH	*Hallo!* DE
مرحبا! AR	*안녕하세요!* KO	*¡Hola!* ES	*Здравствуйте!* RU
こんにちは！ JA	*Hallo!* NL	*xin chào!* VI	*नमस्कार!* HI

Multilingual: 66 Languages

RAW TEXT

WORDS

TOKEN

Native Python Objects

LEMMA	POS	HEAD	DEPREL	...

WORD

SENTENCE

DOCUMENT

Processors work based on pretrained models

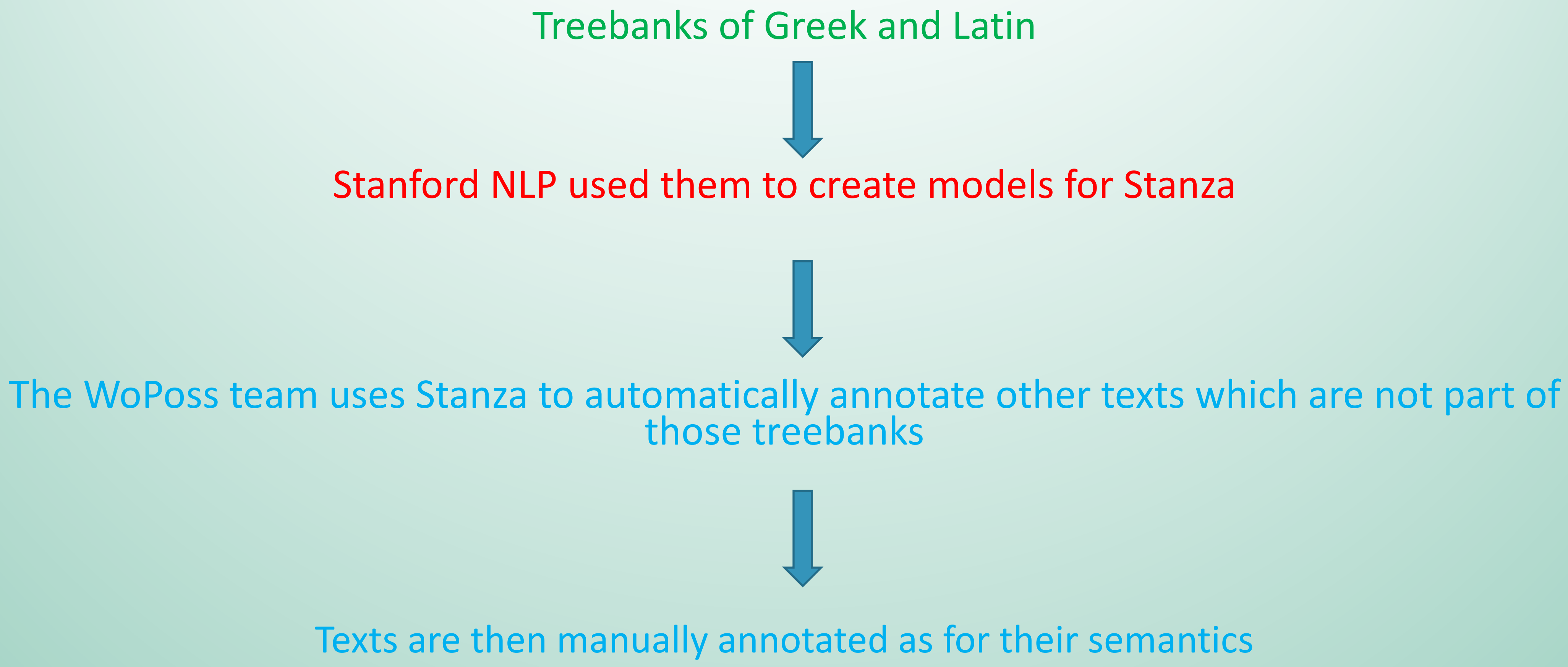
Models were trained on the Universal Dependencies treebanks

Latin: ITTB, PROIEL and Perseus

Ancient Greek: PROIEL and Perseus

For all models:
https://stanfordnlp.github.io/stanza/available_models.html

An indirect way to use treebanks: from treebanks to WoPoss



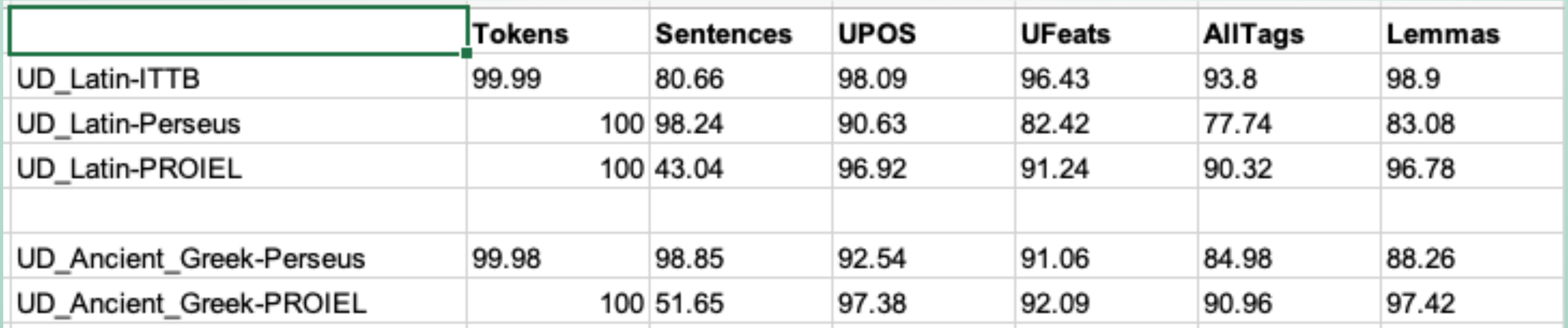
Treebanks of Greek and Latin

Stanford NLP used them to create models for Stanza

The WoPoss team uses Stanza to automatically annotate other texts which are not part of those treebanks

Texts are then manually annotated as for their semantics

Performance of the models



	Tokens	Sentences	UPOS	UFeats	AllTags	Lemmas
UD_Latin-ITTB	99.99	80.66	98.09	96.43	93.8	98.9
UD_Latin-Perseus	100	98.24	90.63	82.42	77.74	83.08
UD_Latin-PROIEL	100	43.04	96.92	91.24	90.32	96.78
UD_Ancient_Greek-Perseus	99.98	98.85	92.54	91.06	84.98	88.26
UD_Ancient_Greek-PROIEL	100	51.65	97.38	92.09	90.96	97.42

Some examples, mainly from our corpus of texts, but not necessarily about modality

You can try it yourself:
https://github.com/WoPoss/automatic_annotation

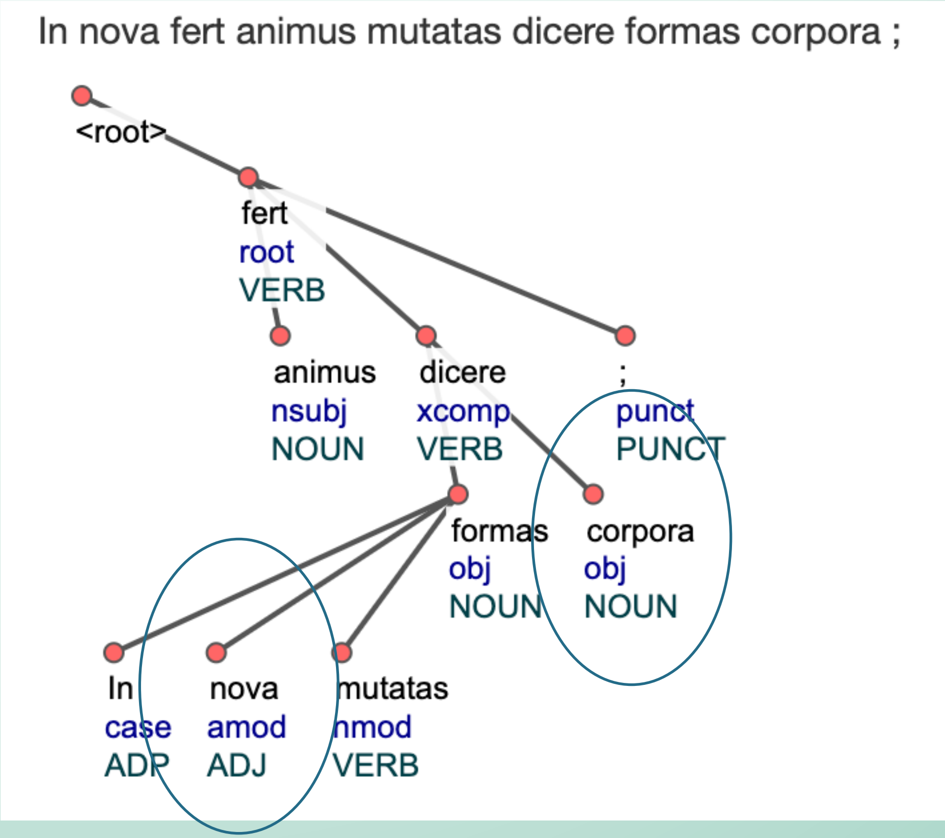
(exercise to compare the models for Ancient Greek and Latin)

Examples

Types of texts	**Model**
• Latin poetic work	With Perseus
• Latin inscription	
• Ancient Greek prose	With PROIEL
• Ancient Greek inscription	

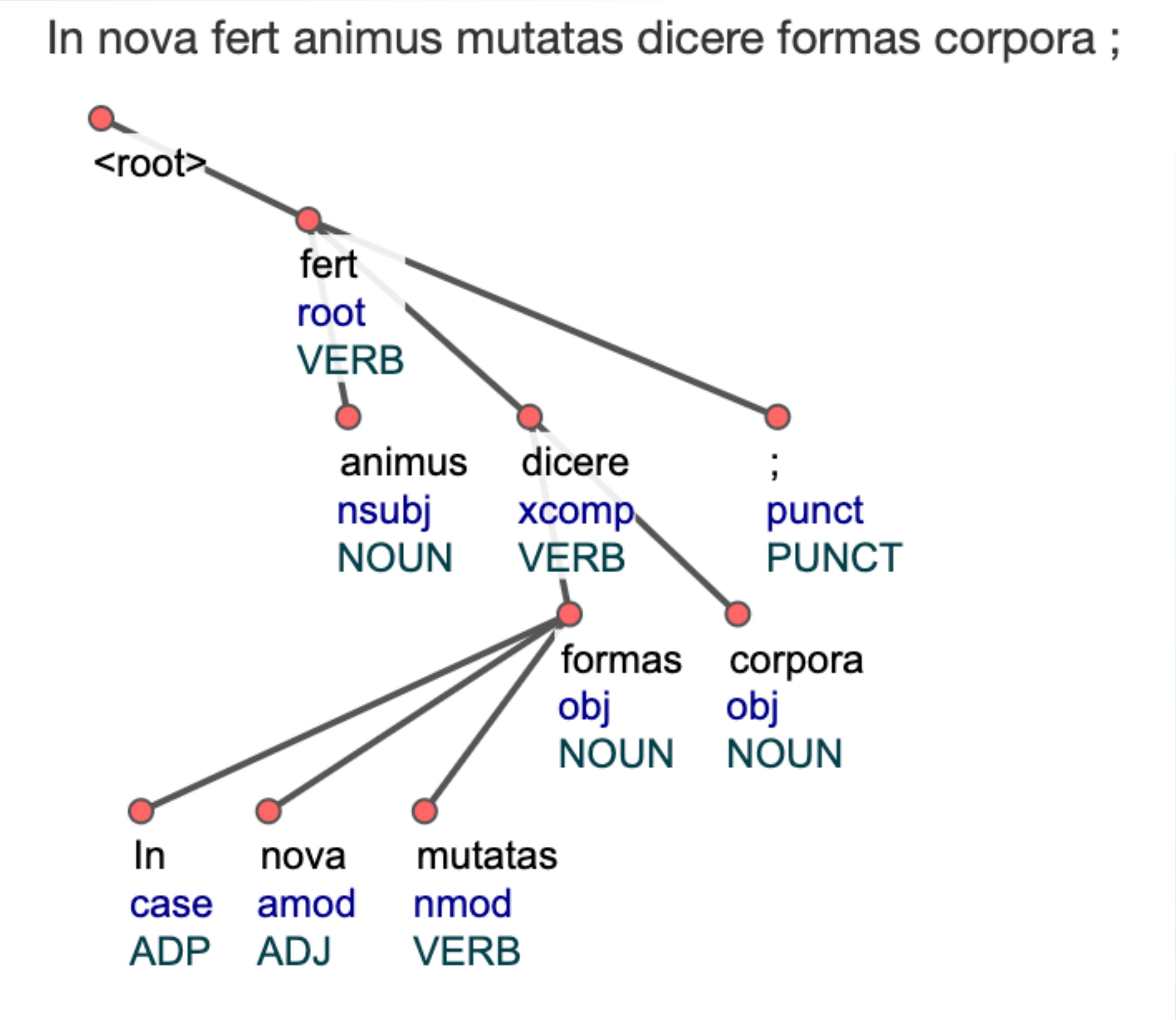
Ovid, Metamorphoses, I, 1-2: *in nova fert animus mutatas dicere formas / corpora* ‘My mind leads me to speak of forms changed into new bodies’

- Results are not at all bad, but they are not perfect: here we need to correct e.g. *nova* which is amod (adjectival modifier) of *corpora* (see also next slide)

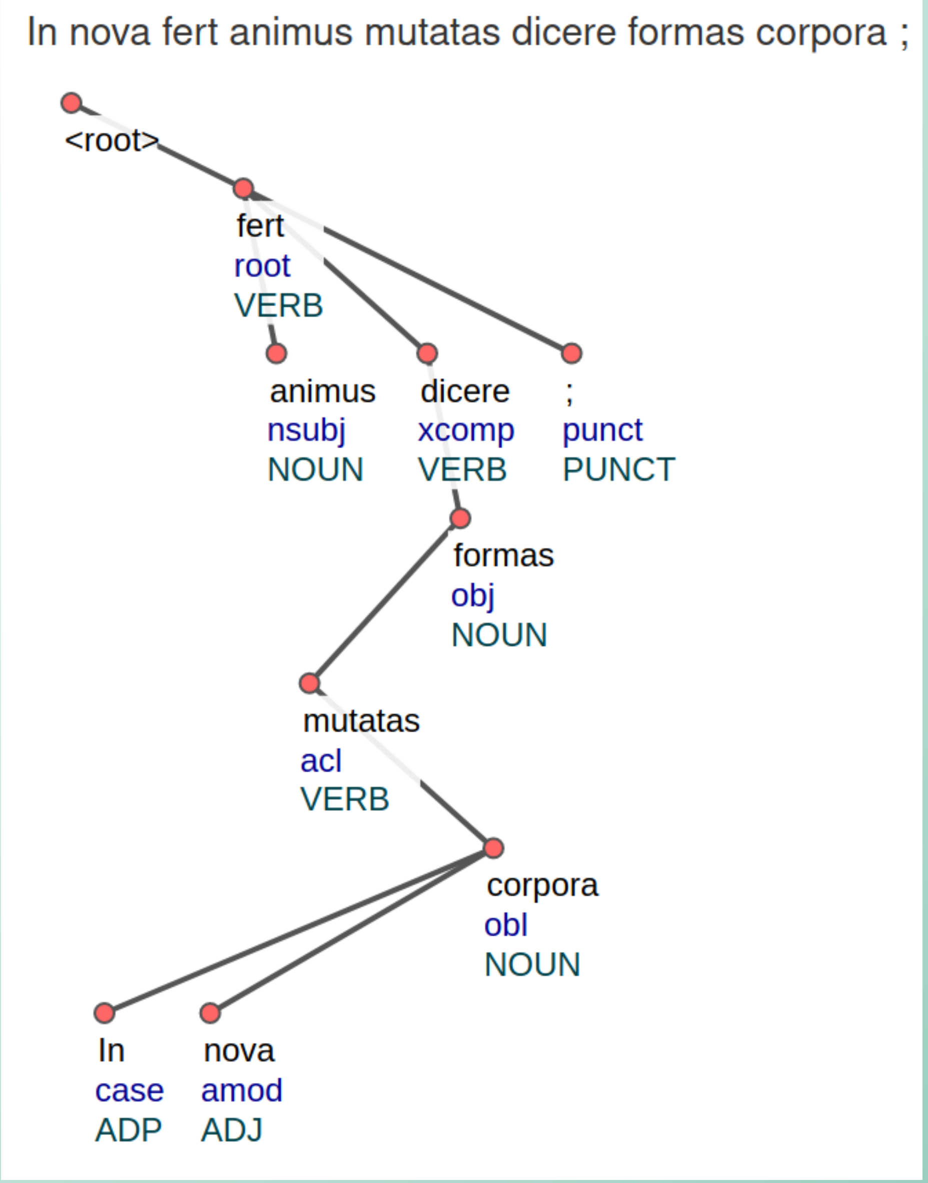


In nova fert animus mutatas dicere formas corpora ;

RESULT (automatic annotation)

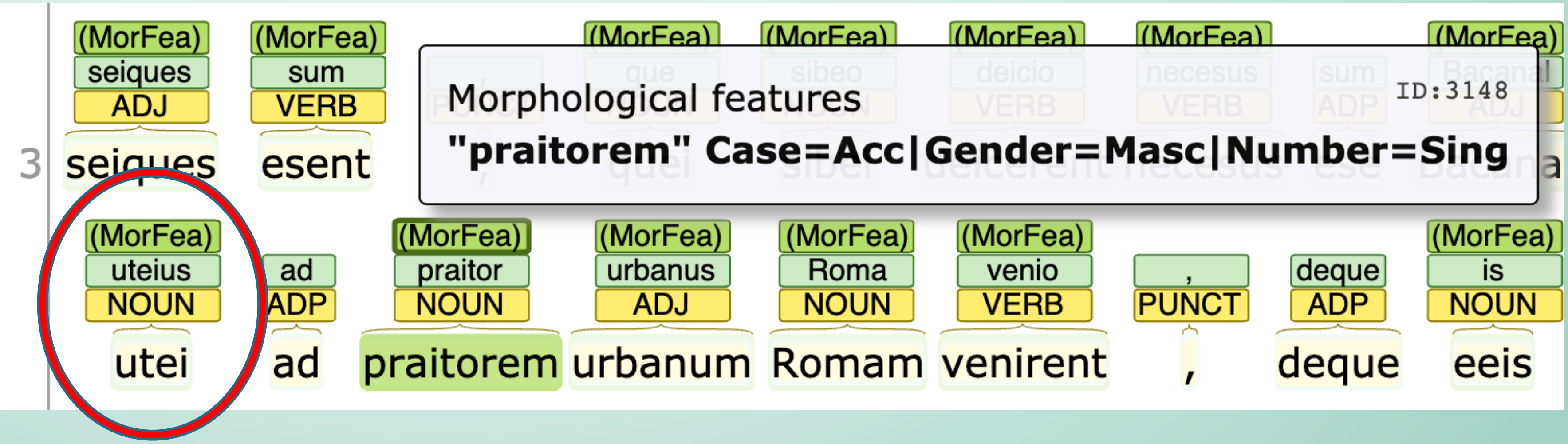

In nova fert animus mutatas dicere formas corpora ;

(manually) CORRECTED


In nova fert animus mutatas dicere formas corpora ;

Senatus consultum de Bacchanalibus, 4-5: …*utei ad praitorem urbanum Romam venirent* ‘they are to come to Rome to the praetor urbanus’

- many archaic forms
  - *utei*
    - PoS SCONJ
    - Lemma *ut*
    - No MorFea
  - *praitor*
    - Lemma *praetor*
    - But PoS and MorFea are correct



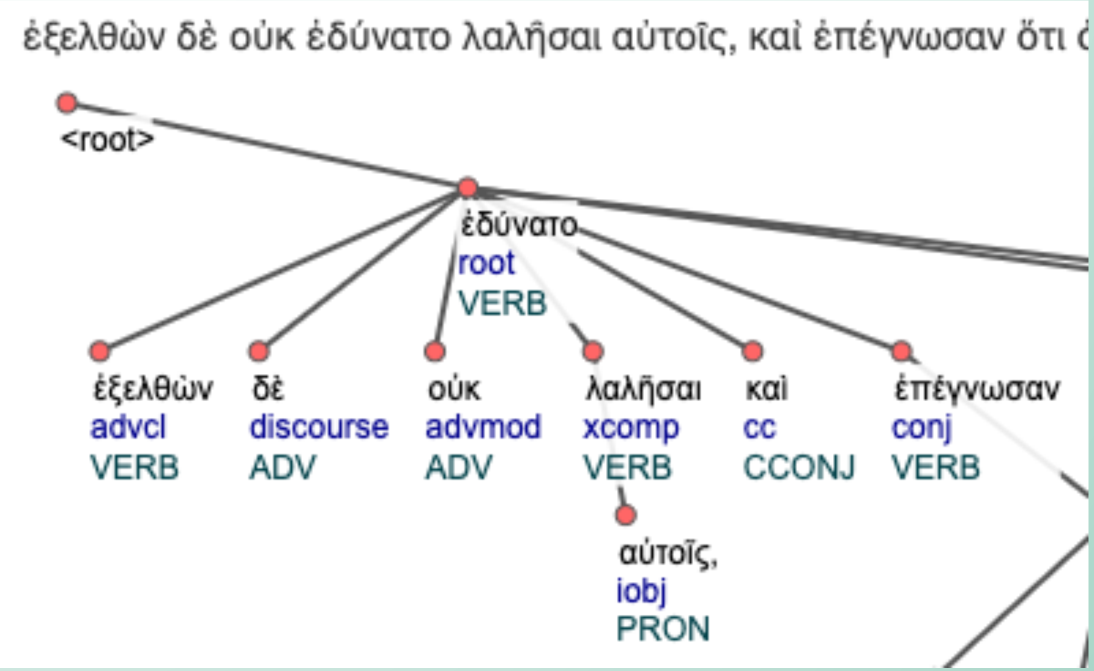
ID:3148

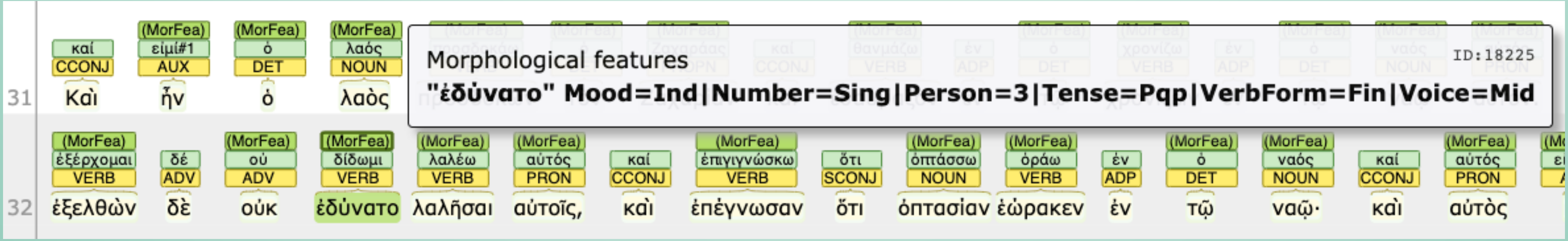
Morphological features

"praitorem" Case=Acc|Gender=Masc|Number=Sing

Luc I, 22:
ἐξελθὼν δὲ οὐκ ἐδύνατο λαλῆσαι αὐτοῖς…
‘when he came out, he couldn’t speak to them’

- All seems fine, but ἐδύνατο has δύναμαι as lemma, the tense is wrong…

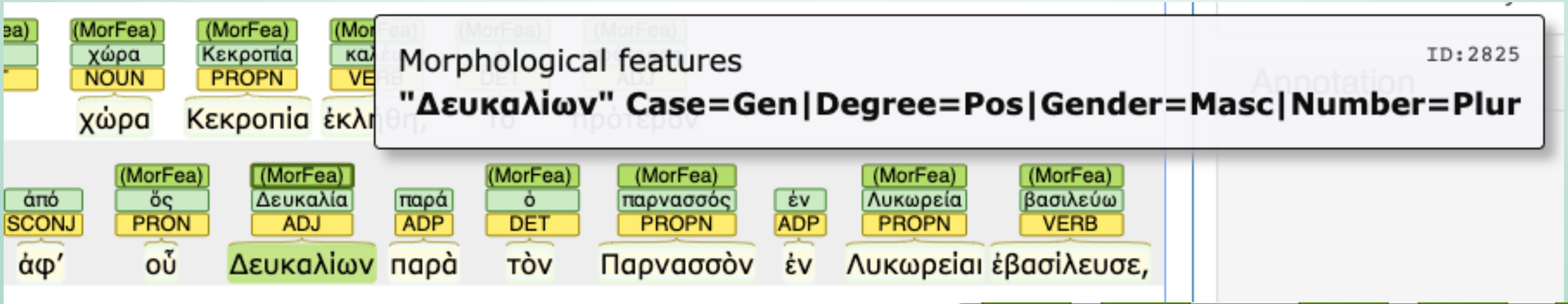



31
(MorFea) καί CCONJ Καὶ
(MorFea) εἰμί#1 AUX ἦν
(MorFea) ὁ DET ὁ
(MorFea) λαός NOUN λαὸς
32
(MorFea) ἐξέρχομαι VERB ἐξελθὼν
δέ ADV δὲ
(MorFea) οὐ ADV οὐκ
(MorFea) δίδωμι VERB ἐδύνατο
(MorFea) λαλέω VERB λαλῆσαι
(MorFea) αὐτός PRON αὐτοῖς,
καί CCONJ καὶ
(MorFea) ἐπιγιγνώσκω VERB ἐπέγνωσαν
ὅτι SCONJ ὅτι
(MorFea) ὀπτάσσω NOUN ὀπτασίαν
(MorFea) ὁράω VERB ἑώρακεν
ἐν ADP ἐν
(MorFea) ὁ DET τῷ
(MorFea) ναός NOUN ναῷ·
καί CCONJ καὶ
(MorFea) αὐτός PRON αὐτὸς

Morphological features ID:18225

"ἐδύνατο" Mood=Ind|Number=Sing|Person=3|Tense=Pqp|VerbForm=Fin|Voice=Mid

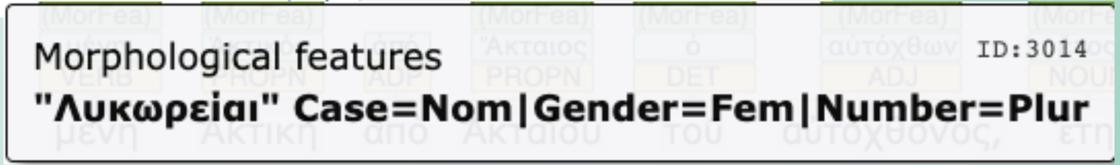
Marmor Parium 2.4b: ἀφ’ οὗ Δευκαλίων παρὰ τὸν Παρνασσὸν ἐν Λυκωρείαι ἐβασίλευσε ‘From the time Deucalion became king on Mount Parnassus in Lycorea’



(MorFea)	(MorFea)	(Mor
χώρα	Κεκροπία	καὶ
NOUN	PROPN	VE
χώρα	Κεκροπια	ἐκλ

Morphological features ID:2825
"Δευκαλίων" Case=Gen|Degree=Pos|Gender=Masc|Number=Plur

	(MorFea)	(MorFea)		(MorFea)	(MorFea)		(MorFea)	(MorFea)
ἀπό	ὅς	Δευκαλία	παρά	ὁ	παρνασσός	ἐν	Λυκωρεία	βασιλεύω
SCONJ	PRON	ADJ	ADP	DET	PROPN	ADP	PROPN	VERB
ἀφ’	οὗ	Δευκαλίων	παρὰ	τὸν	Παρνασσὸν	ἐν	Λυκωρειαι	ἐβασίλευσε,



Morphological features ID:3014
"Λυκωρειαι" Case=Nom|Gender=Fem|Number=Plur

In the same sentence, proper nouns can be correctly or incorrectly recognised.

Conclusion

Advantages in using Stanza

- Tokenisation, lemmatisation, PoS tagging, morphological analysis, dependency parsing are already done
- Annotators can focus on the semantic annotation of modal passages

Disadvantages

- The annotation is not perfect and it is not possible to predict the errors

Acknowledgements

- Swiss National Science Foundation: project n° 176778 (http://p3.snf.ch/project-176778)
- EAGLE-IDEA (https://www.eagle-network.eu/category/idea-association/)

The WoPoss annotators

Francesca Dell’Oro

Paola Marongiu

Helena Bermúdez Sabel

Maël Graa

Séverine Reymond

Thomas von Kaenel

Kevin Duc

References

Helena Bermúdez Sabel, Digital tools for semantic annotation: the WoPoss use case. Bulletin de linguistique et des sciences du langage. In press, 30. [Preprint version: https://zenodo.org/record/3572410]

EAGLE: https://www.eagle-network.eu/basic-search/

INCEpTION: https://inception-project.github.io/

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. [pdf][bib]

Marmor Parium: https://www.dh.uni-leipzig.de/wo/dmp/

Universal Dependencies: https://universaldependencies.org/

UD Greek Perseus: https://universaldependencies.org/treebanks/grc_perseus/index.html

UD Greek PROIEL: https://universaldependencies.org/treebanks/grc_proiel/index.html

UD Latin ITTB: https://universaldependencies.org/treebanks/la_ittb/index.html

UD Latin Perseus: https://universaldependencies.org/treebanks/la_perseus/index.html

UD Latin PROIEL: https://universaldependencies.org/treebanks/la_proiel/index.html

Thank you