AUTOMATED LINGUISTIC ANNOTATION OF LATIN

Margherita Fantoli
margherita.fantoli@kuleuven.be
KU Leuven

Encode Workshop, Bologna, 25.01.2023

STRUCTURE OF THE PRESENTATION

What, How and Why do we annotate?

Enhanced linguistic information

Annotation schemes

Manual vs automatic annotation

Possible applications

Predicting lemmas, tags and dependencies

The field of NLP

Rule-based approaches, Machine Learning and language model for linguistic annotation

Hands-on

Does it work?

A BIT OF HISTORY

- The work of Father Busa (1913-2011) is considered one of the “origins” of the Digital Humanities field
- Father Busa wanted to build a lemmatized concordance of the texts of Thomas Aquinas
- Quickly, his plans became more ambitious

A BIT OF HISTORY

- The work of Father Busa (1913-2011) is considered one of the “origins” of the Digital Humanities field
- Father Busa wanted to build a lemmatized concordance of the texts of Thomas Aquinas
- Quickly, his plans became more ambitious

Then, however—not yet ten years ago—the men involved in automation began to make the boss lean out from his cabin in the tower of electronics and ask philologists and grammarians, who were busy in the fields selecting the choicest flowers, questions such as these:

Please, how many verbs are there in Russian that are active and transitive, and how many that are active and intransitive? How many are there in English?

Which words or linguistic situations are found within a radius of n-words, only when and always when “faccia” means “face”, and which others only and always when “faccia” is a form of the verb “fare” (“to do/make”)?

Again: Please, would you arrange all the words in the dictionary according to the various morphological and grammatical categories? (Busa 1962, 105)

ENHANCING LINGUISTIC INFORMATION

Arma virumque cano, Troiae qui primus ab oris

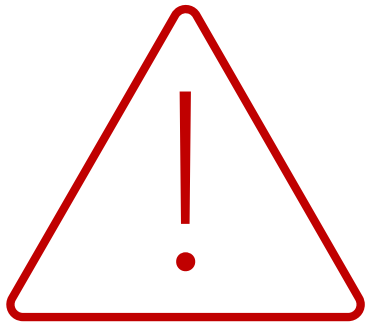
ARMA VIR QVE CANO TROIA QVI PRIMVS AB ORA

NOUN NOUN CONJ VERB PROPN PRONREL ADJ PREP NOUN

PLUR SING X SING SING SING SING X PLUR

TOKENIZATION

Converting a sequence of characters (i.e. your text) into a sequence of lexical tokens (with a meaning)



NOT STRAIGHTFORWARD!

TOKENIZATION

Tantaene animis caelestibus irae?

How many tokens?

TOKENIZATION

Tantaene animis caelestibus irae?

How many tokens?

technique		
white spaces	Tantaene | animis | caelestibus | irae?	4
removing punctuation	Tantaene | animis | caelestibus | irae	4
Separating punctuation	Tantaene | animis | caelestibus | irae | ?	5
Splitting clitics	Tantae | ne | animis | caelestibus | irae | ?	6

LEMMATISATION

Tantaene animis caelestibus irae?

Vocabulary entry

token	Lemma
tantae	tantus
ne	ne
animis	animus
caelestibus	caelestis
irae	ira

LEMMATISATION

CAELESTIS: how many lemmas?

Lemma	PoS
caelestis	proper noun
caelestis	common noun
caelestis celestis coelestis	adjective

https://lila-erc.eu/#page-top

LEMMATISATION

CAELESTIS: how many lemmas?

cælestis,[9] *e (cælum),* **¶ 1** du ciel, céleste : CIC. *Rep. 6, 17 ; Nat. 2, 120* **¶ 2** d'origine céleste, qui se rapporte aux dieux d'en haut : CIC. *Har. 20* ‖ [fig.] divin, excellent, merveilleux : CIC. *Phil. 5, 28 ;* QUINT. *10, 2, 18* ‖ **cælestis,** *is,* subst. m., ordin[t] au plur., habitant du ciel, dieu : CIC. *Off. 3, 25 ;* au fém., déesse : TERT. *Apol. 24* ‖ **cælestĭa,** *ium,* n. pl., choses célestes : CIC. *CM 77.* ➵ *-tior* SEN. *Ep. 66, 11 ; -tissimus* VELL. *2, 66, 3* ‖ abl. *-te* au lieu de *-ti* OV. *M. 15, 743 ;* gén. pl. *-tum* au lieu de *-tium* VARRO *L. 6, 53 ;* LUCR. *6, 1272.*

Gaffiot via https://outils.biblissima.fr/fr/collatinus-web/

LEMMATISATION

CAELESTIS: how many lemmas?

caelestis (**coel-**), **e** (*gen. sing.* CAELESTAE, Inscr. Neapol. 2602; *abl. sing.* regularly, caelesti: caeleste, Ov. H. 16, 277; id. M. 15, 743; cf.: bimestris, cognominis, perennis, patruelis, etc.; *gen. plur.* caelestum, but caelestium, Enn. Epigr. v. 9 Vahl.; Att. ap. Cic. N. D. 3, 26, 68, or id. Trag. Rel. v. 209 Rib.; Varr. L. L. 6, § 53 Müll.; Lucr. 6, 1274; Cat. 64, 191; 64, 205; Verg. A. 7, 432; Ov. M. 1, 150), adj. [caelum], *pertaining to heaven* or *to the heavens*, *found in heaven*, *coming from heaven*, etc., *heavenly*, *celestial* (class. and very freq.): ignis fulminis, Lucr. 2, 384; cf.: turbine correptus et igni, id. 6, 395: flammae, id. 5, 1093: urbes igne caelesti flagrasse, Tac. H. 5, 7: arcus, *the rainbow*, Plin. 11, 14, 14, § 37; Suet. Aug. 95: nubes, Ov. A. A. 2, 237: aqua, *rain*, Hor. C. 3, 10, 20; cf. aquae, id. Ep. 2, 1, 135; Liv. 4, 30, 7; Col. 3, 12, 2; 7, 4, 8; Plin. 17, 2, 2, § 14; Dig. 39, 3, 1: imbres, Col. 3, 13, 7: templa, Lucr. 5, 1203; 6, 388; 6, 671: solum, Ov. M. 1, 73: plagae, id. ib. 12, 40 al.: astra, id. ib. 15, 846: aërii mellis dona, Verg. G. 4, 1: prodigia, Liv. 1, 34, 9; cf. minae, Tac. H. 1, 18: caelestia auguria vocant cum fulminat aut tonat, Paul. ex Fest. p. 64, 8 Müll.: fragor, Quint. 12, 10, 4: orbes, quorum unus est caelestis, Cic. Rep 6, 17, 17.—*Subst.*: **caelestĭa**, **ĭum**, n., *the heavenly bodies*: cogitantes supera atque caelestia, haec nostra, ut exigua et minima, contemnimus, Cic. Ac. 2, 41, 127; Tac. H. 5, 4; id. A. 4, 58.—

II. Meton.

A. *Divine;* and *subst.*, *the deity* (most freq. like caeles in plur.), *the gods.*

1. Adj., numen, Cat. 66, 7; Tib. 3, 4, 53; Ov. M. 1, 367: animi, Verg. A. 1, 11: aula, Ov F 1, 139: irae. Liv. 2, 36, 6: ira, Sen. Herc. Oet. 441: origo, Verg. A. 6, 730: ortus, Quint. 3, 7, 5: stirps, Ov. M. 1, 760; cf. species, id. ib. 15, 743: nectar, id. ib. 4, 252; cf. pabula, id. ib. 4, 217: sapientia, Hor. Ep 1, 3, 27: auxilium, *of the gods*, Ov. M. 15, 630: dona, id. ib. 13, 289 al.: cognitio caelestium et mortalium, Quint. 1, 10, 5; cf. id. 10, 1, 86.— * *Comp neutr.*: nihil est caelesti caelestius, Sen. Ep. 66, 11—

2. *Subst.*: **caelestis**, **is**, m., *a deity*: quicumque dedit formam caelestis avarae, Tib 2, 4, 35.—Mostly plur., *the gods*: divos et eos qui caelestes semper habiti colunto, Cic. Leg. 2, 8, 19: caelestum templa, Lucr. 6, 1273: in concilio caelestium, Cic. Off. 3, 5, 25; so id. Phil. 4, 4, 10; Liv. 1, 16, 7; 9, 1, 3; Tac. G. 9; id. H. 4, 84; Cat. 64, 191; 64, 205; 68, 76; Tib. 1, 9, 5; Verg. A. 1, 387; 7, 432; Ov. M. 1, 150; 4, 594; 6, 72, 6, 171.—

3. Caelestis, **is**, f., *a female divinity in Carthage*, Tert. Apol. 24, Capitol. Pert. 4, 2; Macrin. 3, 1; Treb. Pol. Trig. Tyr. 29, 1.—

4. caelestĭa, **ĭum**, n., *heavenly objects*, *divine things*: haec caelelestia semper spectato, illa humana con-t emnito, Cic. Rep. 6, 19, 20: sapientem non modo cognitione caelestium vel mortalium putant instruendum, Quint. 1, 10, 5; Tac. H. 5, 5.—

B. As in most languages, an epithet of any thing splendid or excellent, *celestial*, *divine*, *god-like*, *magnificent*, *preeminent*, etc. (so most freq. since the Aug. per., esp. as a complimentary term applied to eminent persons and their qualities; in Cic. only once): caelestes divinaeque legiones, Cic. Phil. 5, 11, 28: quem prope caelestem fecerint, Liv. 6, 17, 5: ingenium, Ov. A. A. 1, 185: mens, id. F. 1, 534: in dicendo vir (sc. Cicero), Quint. 10, 2, 18; cf.: caelestissimum os (Ciceronis), Vell. 2, 66, 3: ju dicia, Quint. 4 prooem. § 4 Spald.: praecepta, Vell. 2, 94, 2: anima, id. 2, 123: animus, id. 2, 60, 2: caelestissimorum ejus operum, id. 2, 104, 3: quos Elea domum reducit Palma caelestes, *glorified*, *like the gods*, Hor C. 4, 2, 18.—*Adv.* not in use.

L&S via https://outils.biblissima.fr/fr/collatinus-web/

PART-OF-SPEECH (POS) TAGGING

Tantaene animis caelestibus irae?

token	PoS
tantae	ADJ? DET?
ne	PART? ADV?
animis	NOUN
caelestibus	ADJ? PROPN? NOUN?
irae	NOUN

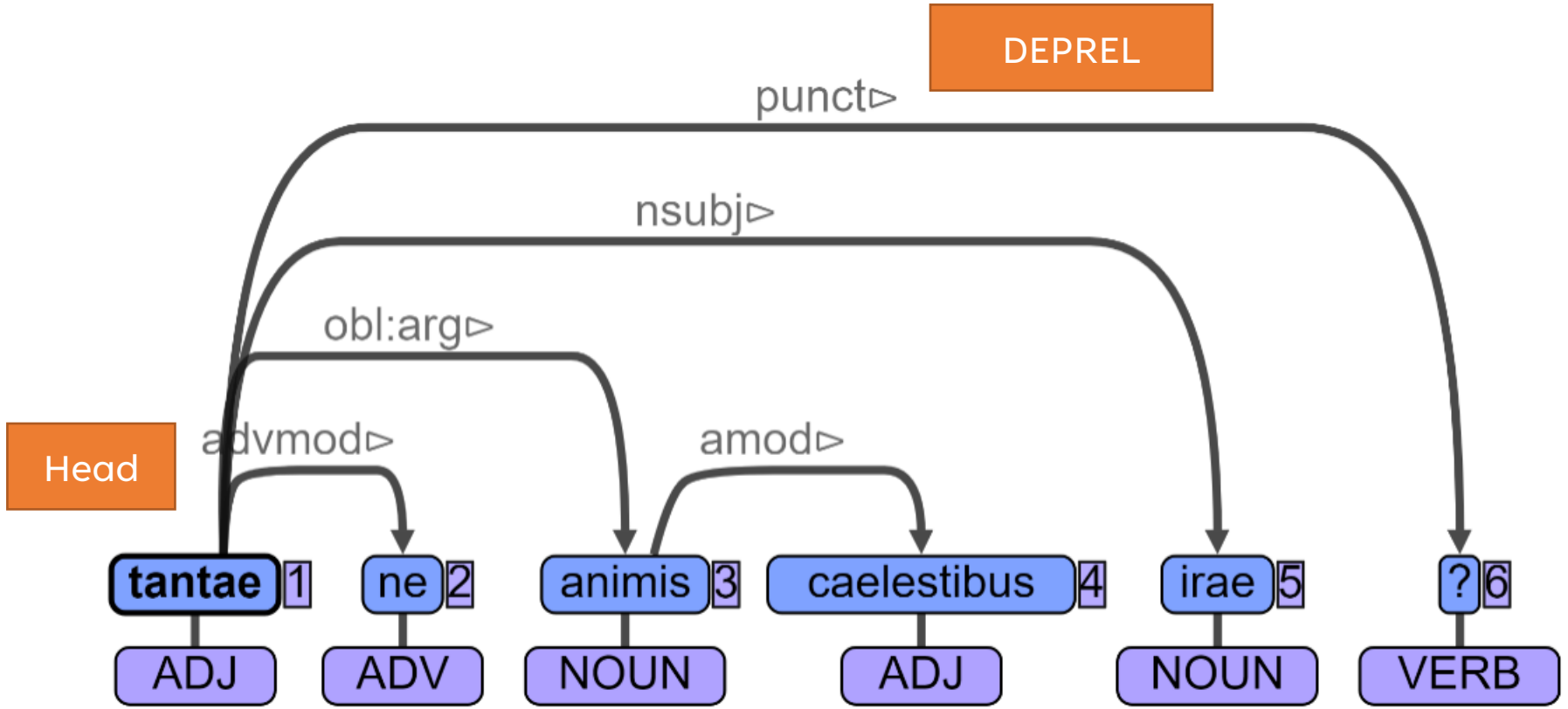
MORPHOLOGICAL FEATURES

Tantaene animis caelestibus irae?

token	PoS
tantae	Class1|nom|plur|fem
ne	#
animis	decl2|dat|plur|masc
caelestibus	Class2|dat|plur
irae	decl1|nom|sing|fem

SYNTACTIC TREE

Tantaene animis caelestibus irae?



ANNOTATION STYLES

- Different projects follow different conventions for linguistic annotation

- Despite many common points, they differ on some crucial choices, such as:
  - How to annotate “particles” (*enim*, *autem*, *uero* etc.)?
  - How to treat words used in a different function, e.g. nominalized adjectives etc.?
  - How to deal with spelling variation and irregular declensions?

- For classical studies we find, for instance:
  - Different dictionaries as lemmatization reference (Forcellini, L&S etc.)
  - Different tagsets for PoS and morphology
  - Different parsing styles (constituent vs dependency)

ANNOTATION STYLES FOR CLASSICAL TEXTS

- Perseus (morphology + syntax)
  - Prague Dependency Treebank + Pinkster Latin Grammar
  - https://static.perseus.tufts.edu/docs/guidelines.pdf
  - H. Pinkster, *The Oxford Latin Syntax*. Vol I and II. Oxford: Oxford University Press. 2015-2021

- Universal Dependencies:
  - Universal PoS tagset: https://universaldependencies.org/u/pos/index.html
  - https://universaldependencies.org/u/feat/index.html (features based on existing treebanks)
  - Dependency relations (https://universaldependencies.org/u/dep/)

- LASLA Annotation tagset (see shared documents): traditional grammatical categories:
  - Currently only limited documentation online
  - Morphology + some syntactic features (limited)

MANUALLY ANNOTATED CORPORA

- LASLA corpus:
  - 2.5 million tokens
  - Classical literary texts
  - Lemmatization & morphology

- PROIEL (Pragmatic Resources of Old Indo-European Languages)
  - 179 000 Latin tokens
  - Caesar, Cicero, Jerome's New Testament..
  - Initially using Perseus formalism, now available as UD

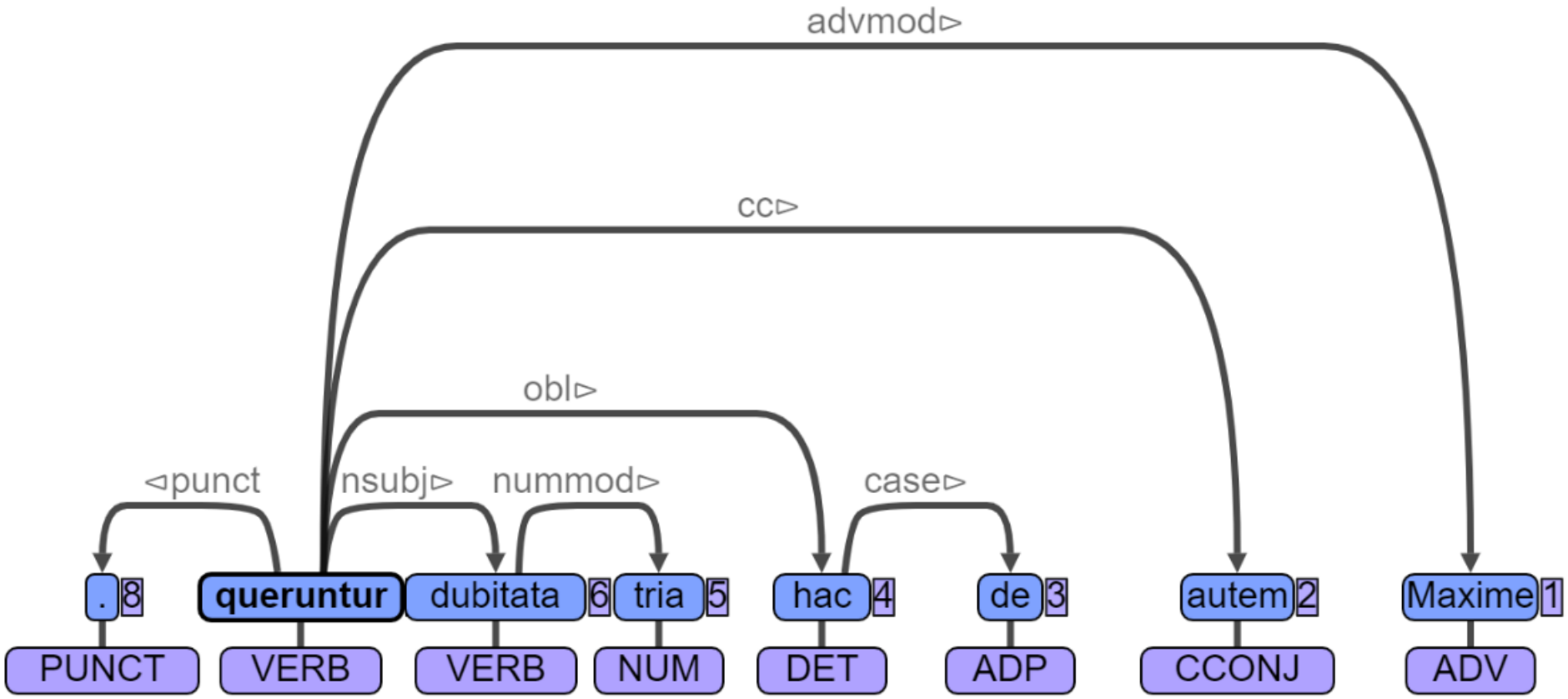
- Latin Dependency Treebank (LDT, Perseus)
  - 53000 Latin tokens
  - Classical authors
  - Initially using Perseus formalism, now available as UD

- Index Thomisticus Treebank (ITTB):
  - 350 000 000 nodes (17 000 sentences)
  - Works of Thomas Aquinas (esp. *Summa Contra Gentiles*)
  - Initially using Perseus formalism, now available as UD
- LLDT (Late Latin Charter Treebank)
  - 521 Early Medieval Latin charters written in Tuscany, 8th-9th centuries
  - 242411 tokens
  - Initially using Perseus formalism, now available as UD
- UDante:
  - Latin works by Dante (*De vulgari eloquentia, Monarchia, Letters, Questio de aqua et terra, Eclogues*)
  - Ca 55000 tokens
  - UD formalism

EXAMPLE OF UD FILE (CONLL-U)

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
1	Maxime	magis	ADV		Degree=Abs	7	advmod
2	autem	autem	CCONJ		_	7	cc
3	de	de	ADP		AdpType=Prep	4	case
4	hac	hic	DET		Case=Abl|Gender=Fem|InflClass=LatPron|Number=Sing|PronType=Dem	7	obl
5	tria	tres	NUM		Case=Nom|Gender=Neut|InflClass=IndEurI|Number=Plur|NumForm=Word|NumType=Card	6	nummod
6	dubitata	dubito	VERB		Case=Nom|Degree=Pos|Gender=Neut|InflClass=LatA|InflClass[noun]=IndEurO|Number=Plur|Tense=Past|VerbForm=Part|Voice=Pass	7	nsubj
7	queruntur	quaero	VERB		InflClass=LatX|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Pass	0	root
8	.	.	PUNCT			7	punct

EXAMPLE OF UD TREE



UNIVERSAL POS TAGSET

https://universaldependencies.org/u/pos/index.html

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

UD DEPRELS

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

WHY DO WE ANNOTATE?

- (Latin) Linguistic studies
  - Diachronic synchronic study of linguistic features
  - Comparison of Latin linguistic features with other languages
- Literary studies:
  - characterization of the writing of writers
  - identification of trends in literary corpora
- Educational purposes:
  - Training (university and high school) students
  - More engaging approach to linguistic analysis

EXISTING CHALLENGES

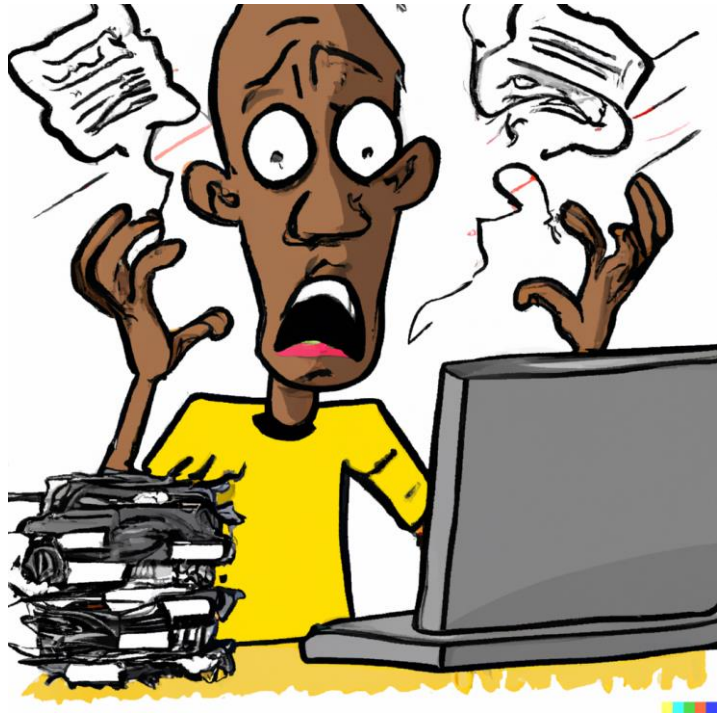
- Multiple projects -> different choices (cf. LiLa project)

- Very restricted canon of texts
  - Classical literary authors
  - “Big” medieval literary/philosophical authors
  - One treebank with non-literary texts

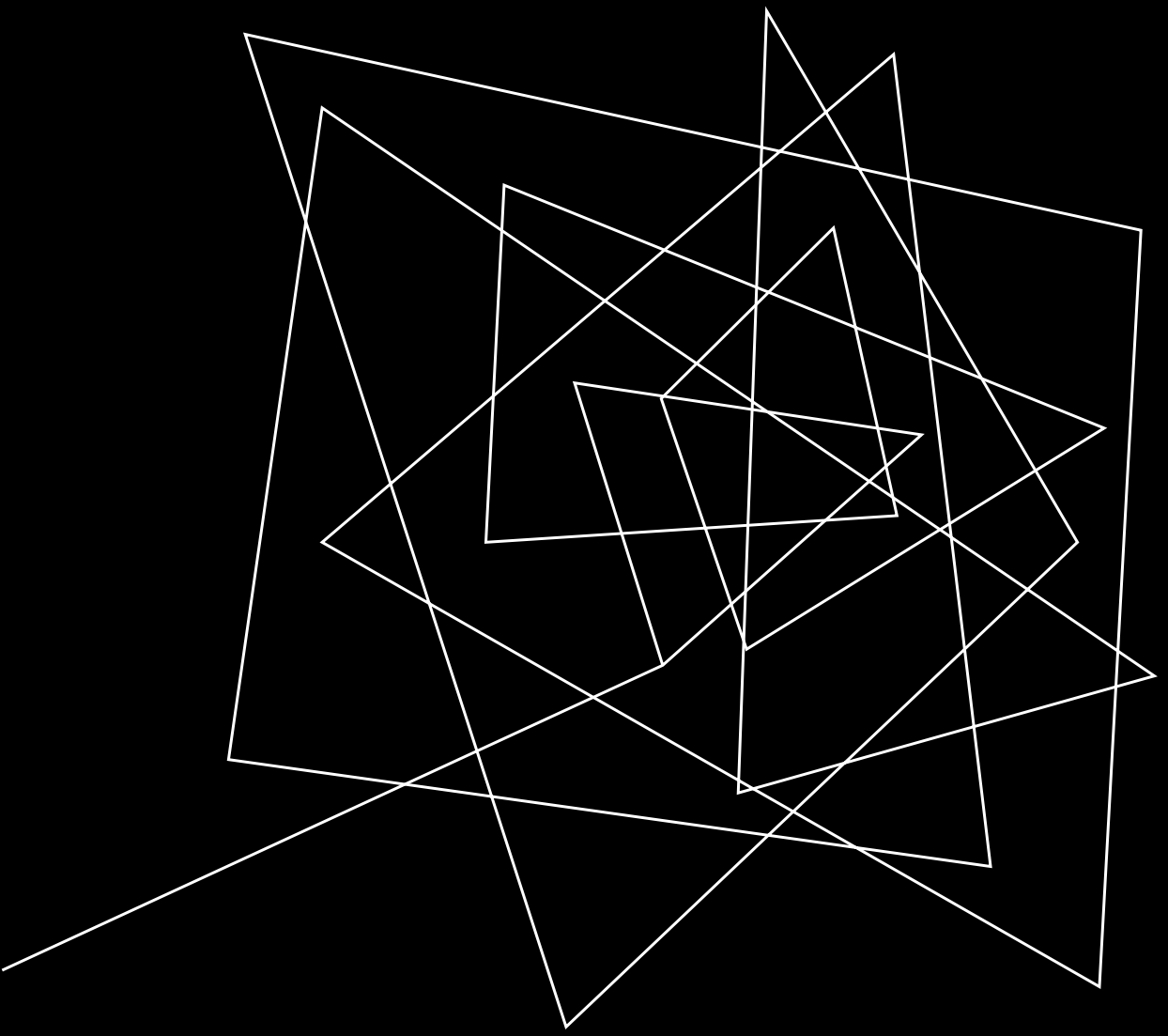
- Limited availability of tools for exploring the annotated texts

EXISTING CHALLENGES

MANUAL ANNOTATION IS PAINFUL



Generated with Dall-E 2



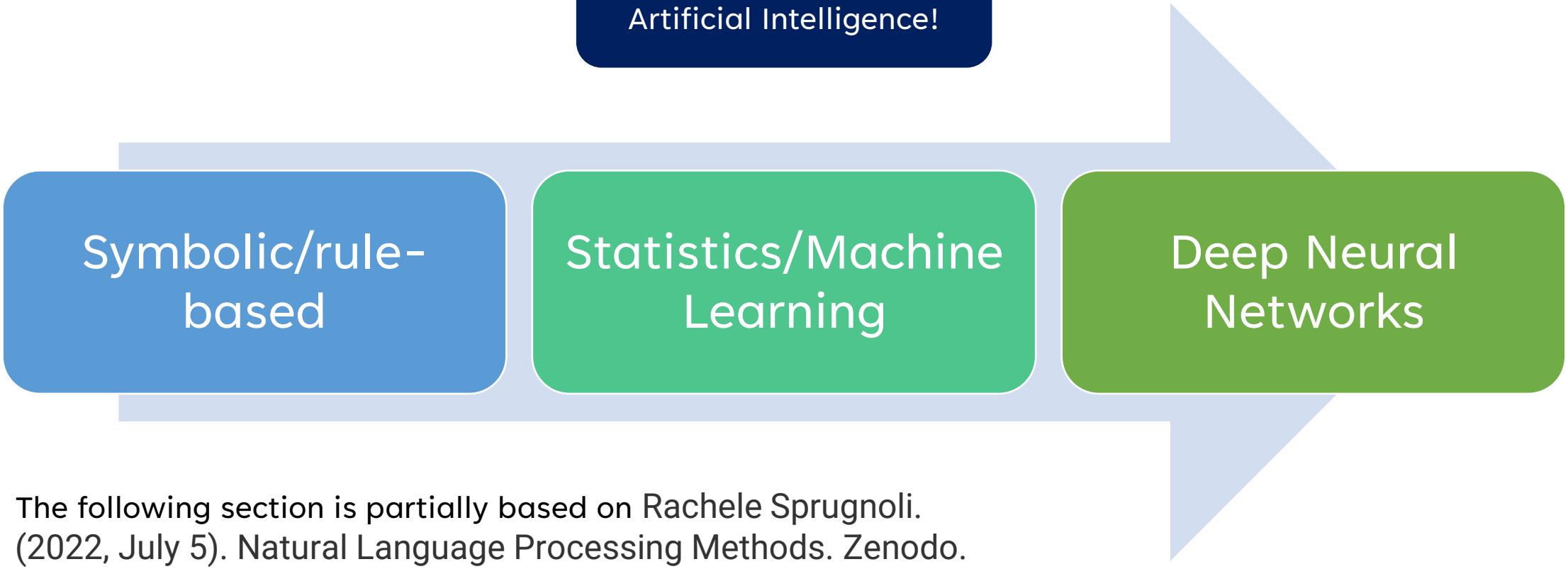
PREDICTING LEMMAS, TAGS AND DEPENDENCIES

AUTOMATING ANNOTATION

- Scholar have been trying to automate the process as much as possible
- In the last years, incredible acceleration in terms of tools availability
- However, major challenges remain that hinder the dissemination of automatically annotated texts within the community: we will discuss this later on!

NATURAL LANGUAGE PROCESSING

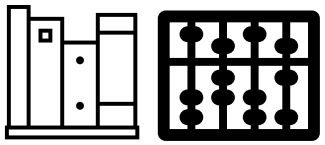
Technically, it is all Artificial Intelligence!



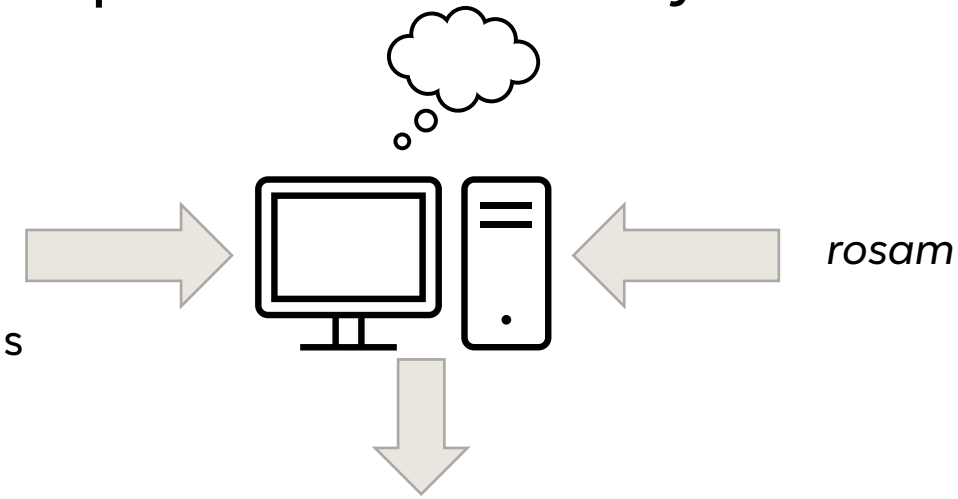
The following section is partially based on Rachele Sprugnoli. (2022, July 5). Natural Language Processing Methods. Zenodo. https://doi.org/10.5281/zenodo.6798390

SYMBOLIC/RULE-BASED

- Human experts design a set of rules that the machine applies to the data, and produces the analysis



- List of Latin words
- List of Latin word endings
- Inflectional rules
- Word-splitting rules



rosam

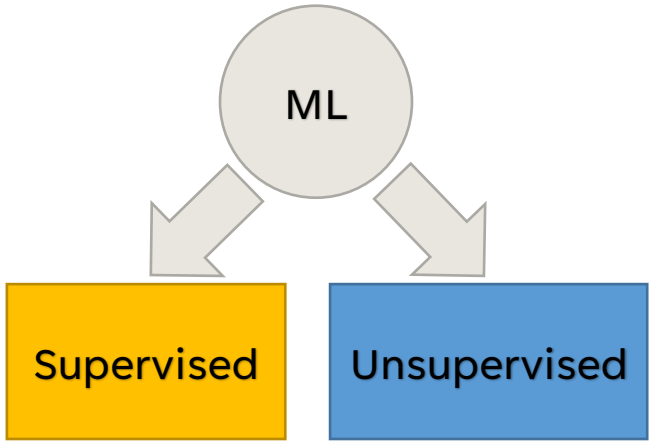
LEMMA: *rosa*
Morphology: First declension, accusative, singular, feminine

Examples for Latin: LEMLAT, Collatinus.

What might be the issues?

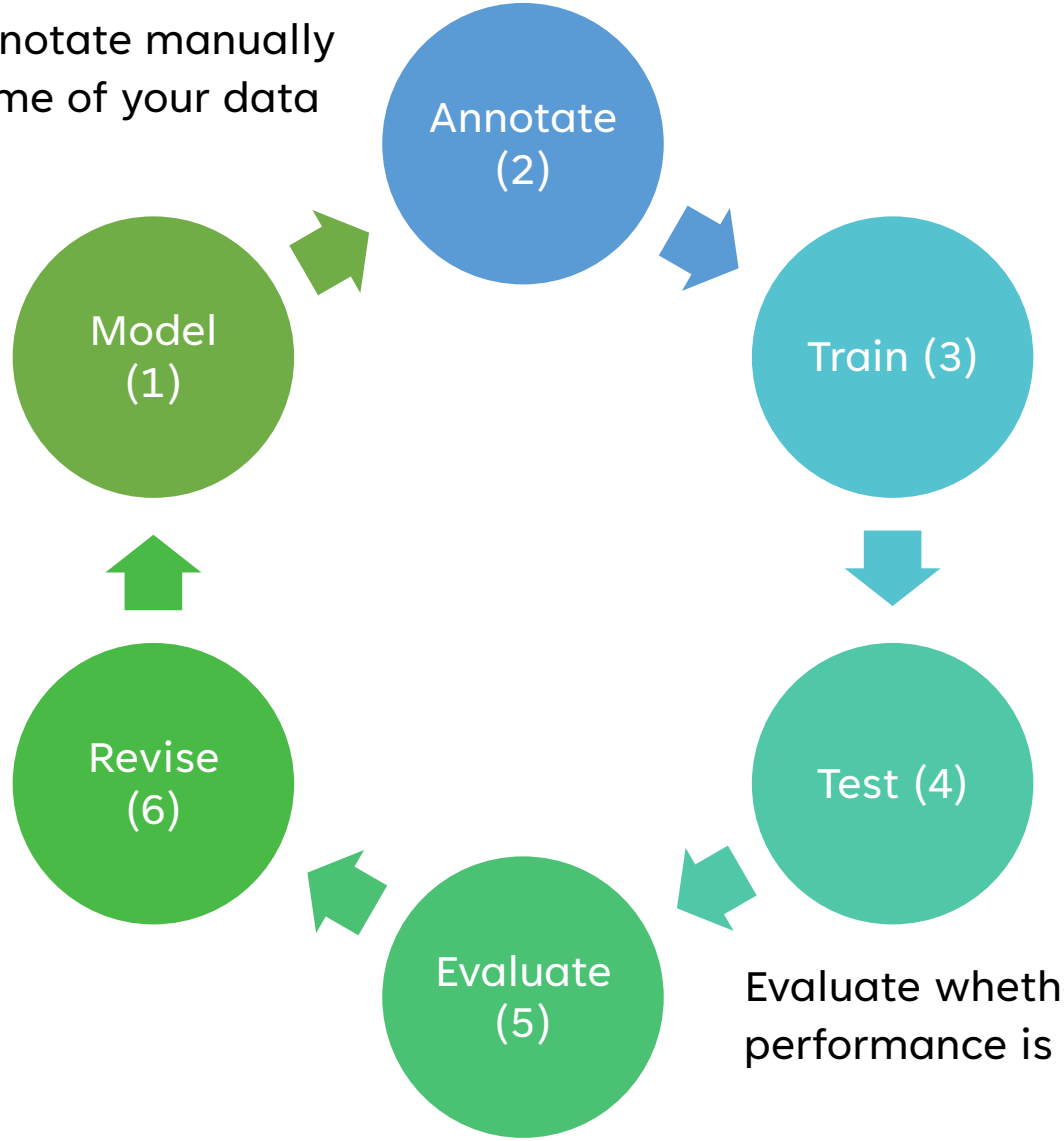
STATISTICAL/MACHINE LEARNING (ML)

- The machine tries to identify patterns/statistical rules in the data
- Large variety of techniques, applying different statistical models and algorithms
- One very important distinction
  - Supervised learning
  - Unsupervised learning



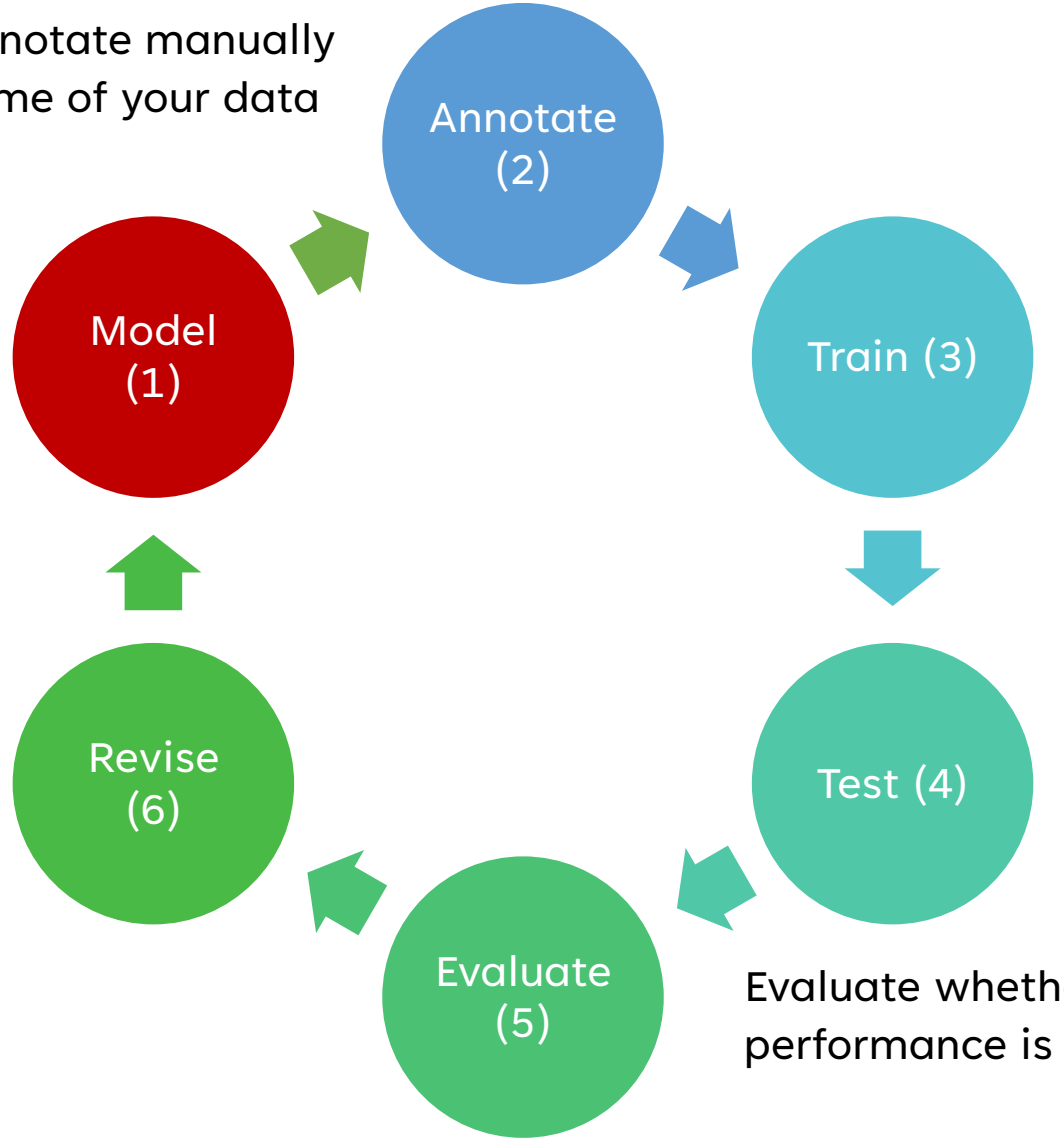
Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

SUPERVISED LEARNING



Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

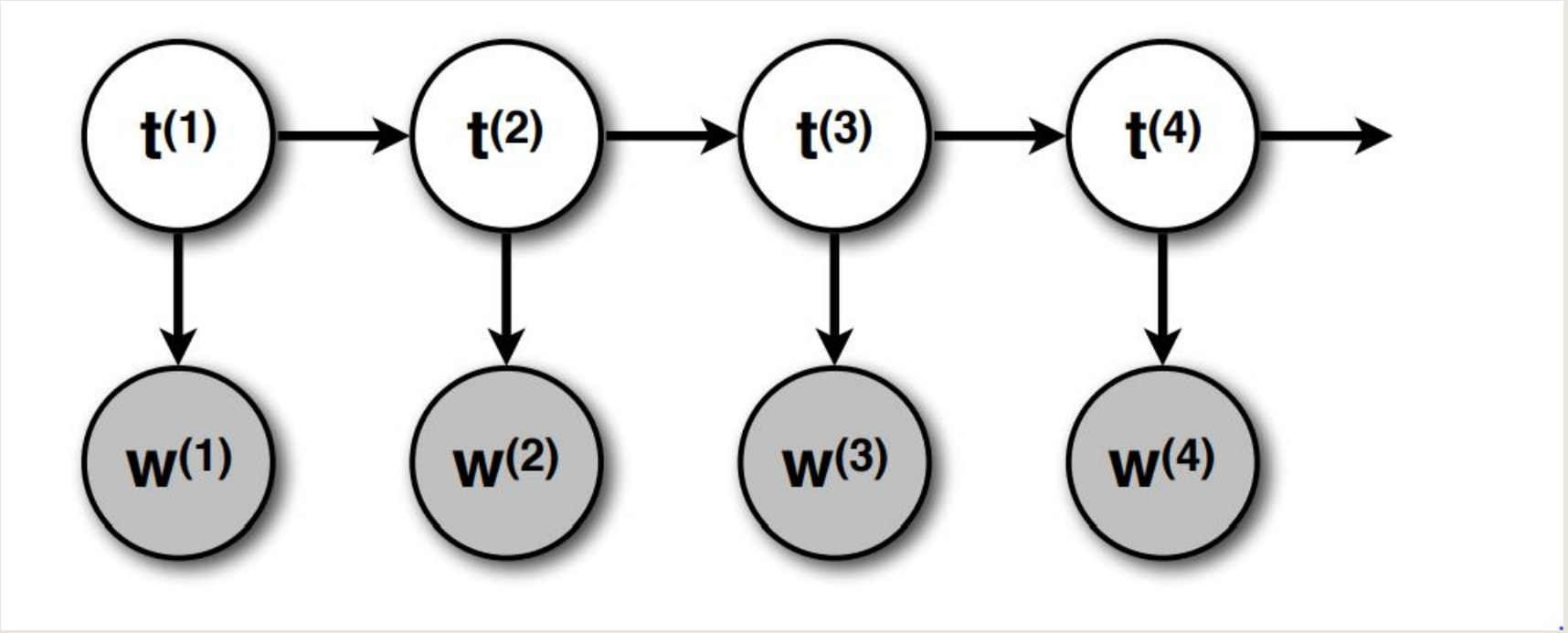
SUPERVISED LEARNING



AN EXAMPLE: SEQUENCE LABELLING MET HMM

- A text can be seen as a sequence of “**states**”, which are in part **observable** (the words that we concretely see) and in part **unobservable** (their linguistic properties, for instance PoS)
- Hidden Markov Models (HMM) estimate, given a sequence of hidden labels, what is the likelihood that those specific words appear
- For instance the sequence ADJ – NOUN – VERB is more likely to generate “paruam pecuniam dat” than “quousque tandem Catilina”
- To decide what is the right sequence of PoS we look for the one that maximizes the likelihood for the words we have

GRAPHICAL REPRESENTATION OF HMM

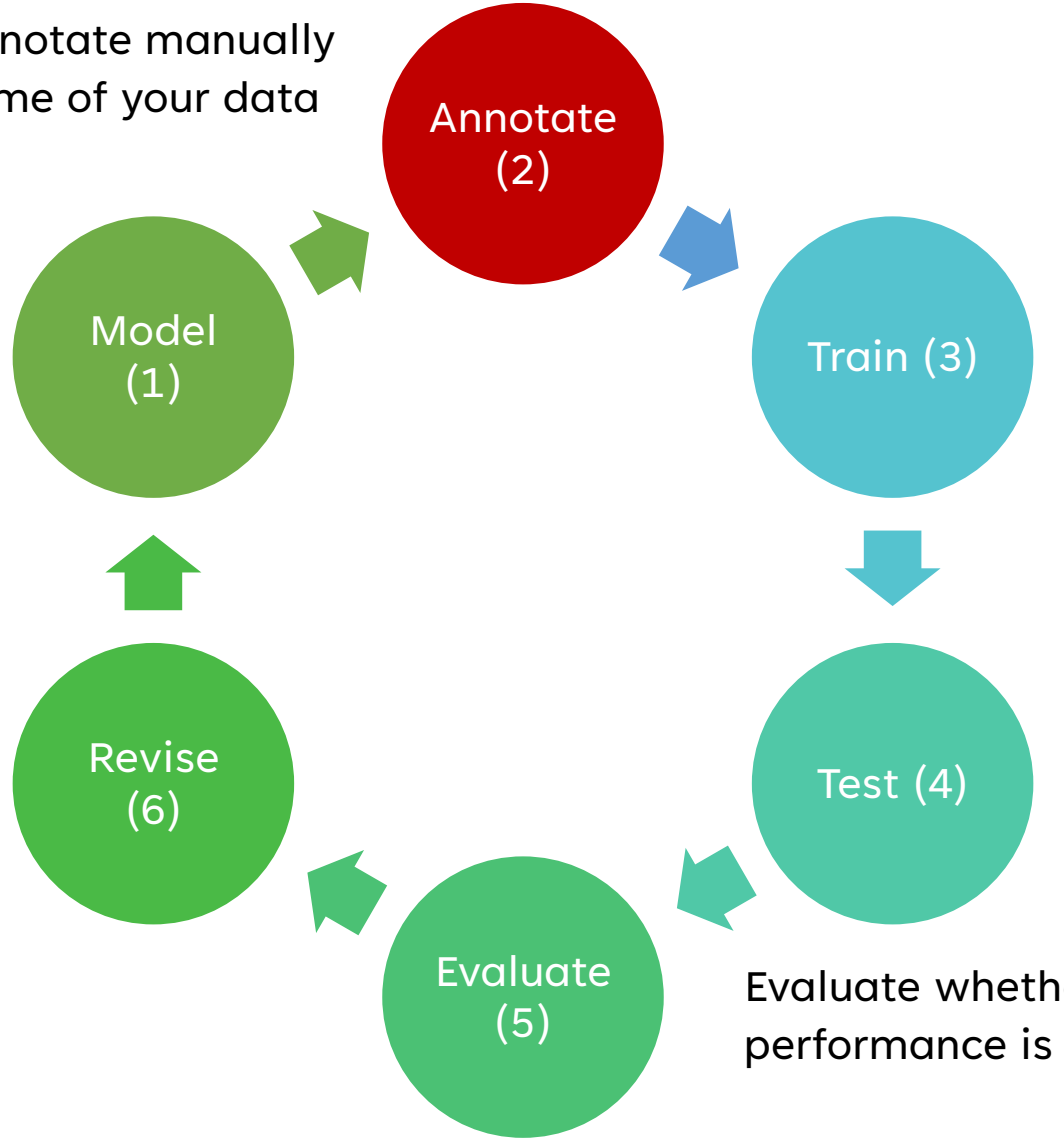


https://courses.engr.illinois.edu/cs447/fa2018/Slides/Lecture07.pdf

GENERATIVE MODEL
(DISCRIMINATIVE MODELS SUCH AS CONDITIONAL RANDOM FIELDS CURRENTLY MORE USED:
arrows from bottom to top!)

Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

SUPERVISED LEARNING

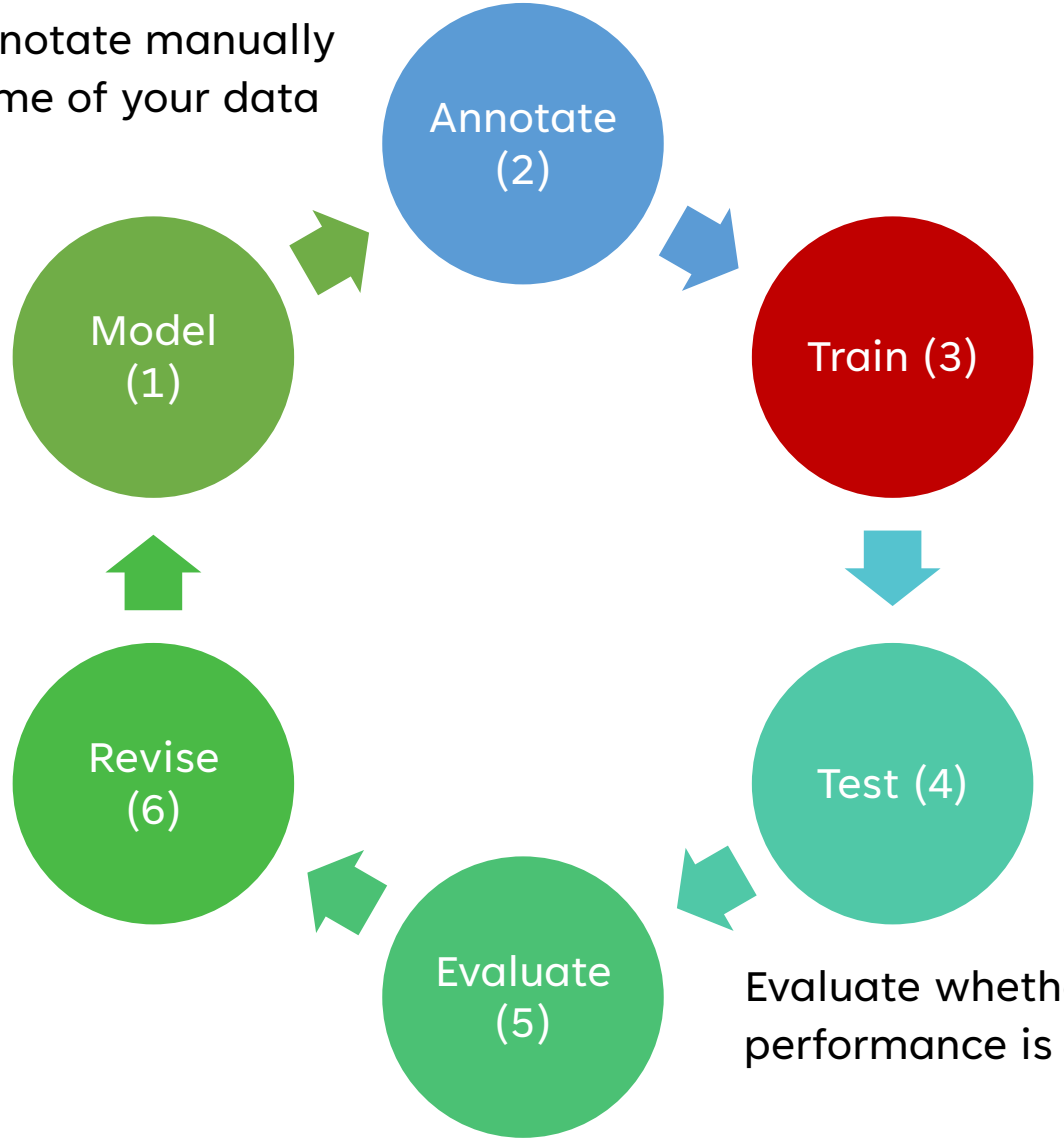


ANNOTATED DATA

- Pick a formalism (for instance UD), annotation guidelines and representative sample of data to annotate
- E.g., if you are interested in late-antique Christian text, pick one text of the corpus you want to study and annotate the PoS of every word
- Manually annotated data are called GOLD data: apart from human errors, they represent the optimal standard
- Ideally, annotated (gold) data are already available: for instance, the annotated corpora listed before
- However, inconsistencies in the annotated data are very dangerous -> harmonization

Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

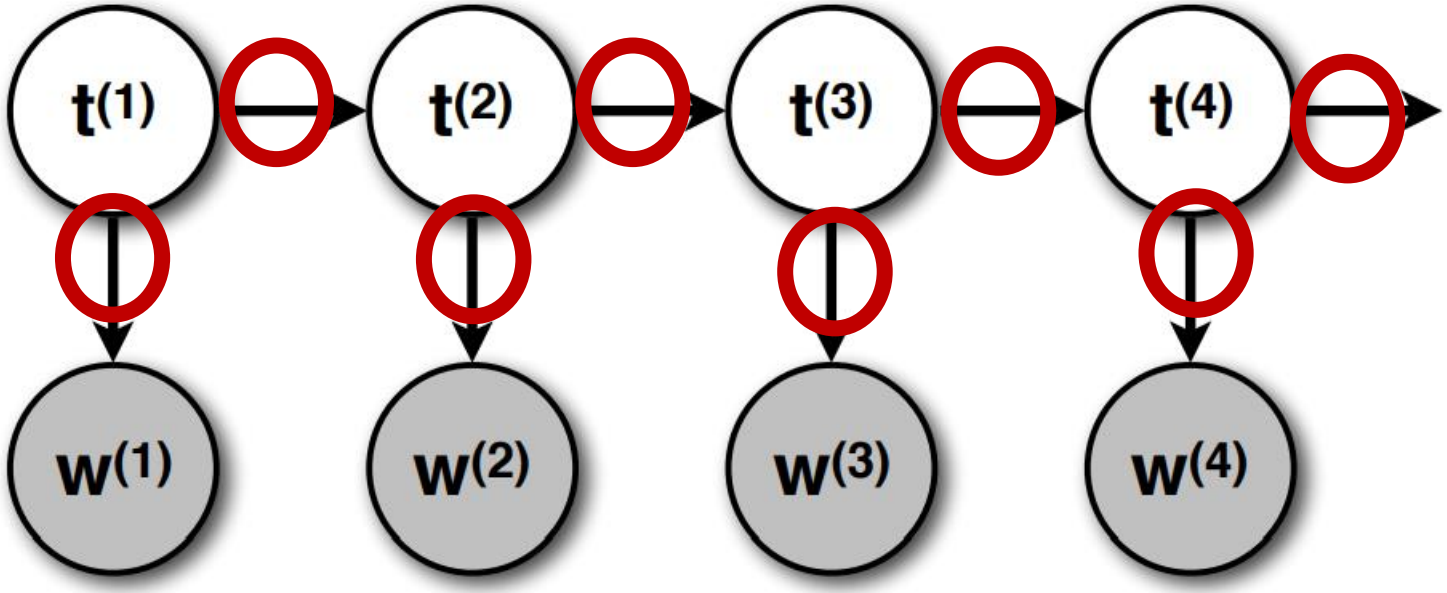
SUPERVISED LEARNING



TRAIN THE MODEL

Computationally heavy

Crucial step

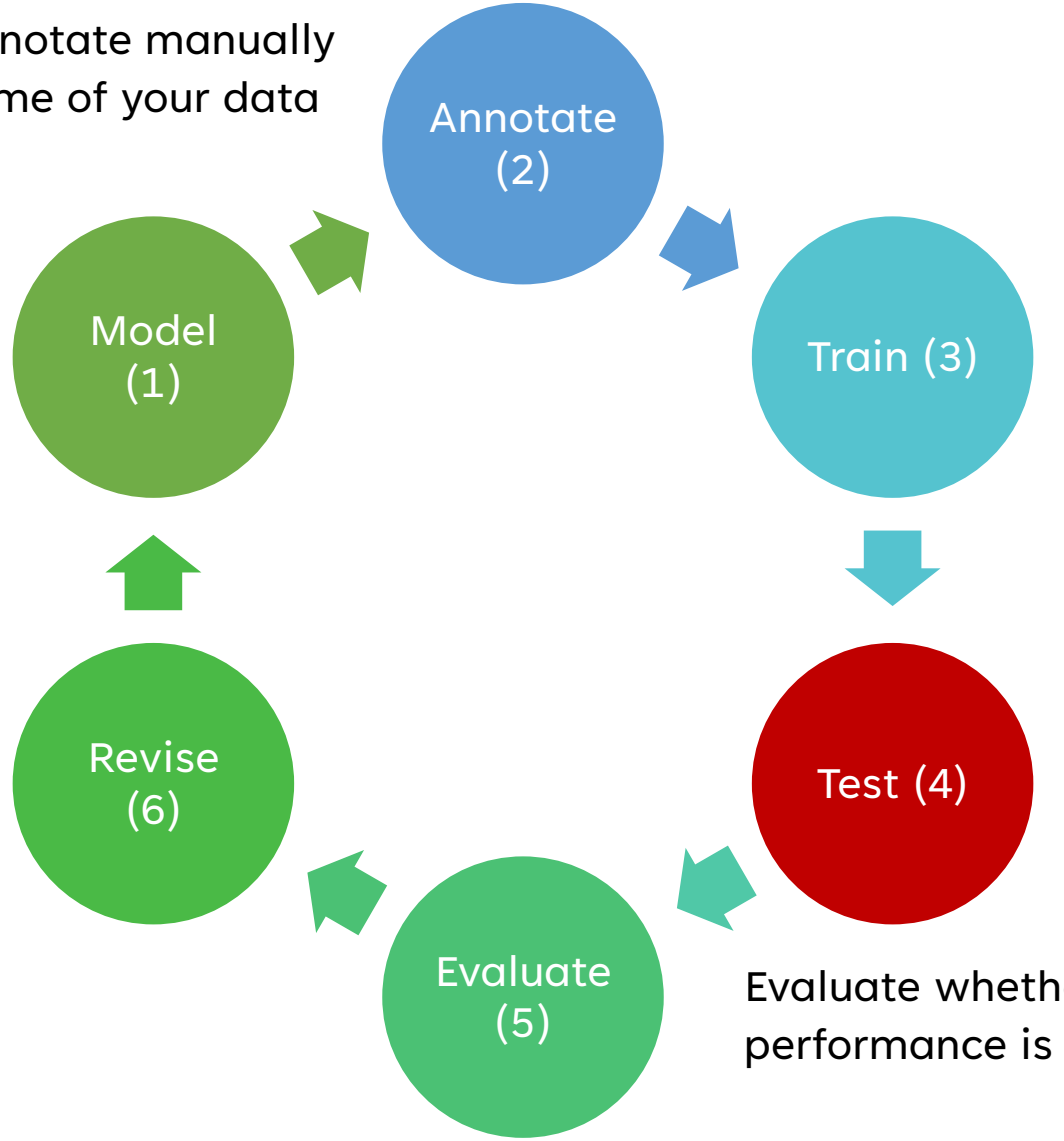


Often “finetuning” and not training from scratch

By looking at the data, where each word is associated to its PoS, the program learns the probabilities that link the PoS and the words

Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

SUPERVISED LEARNING

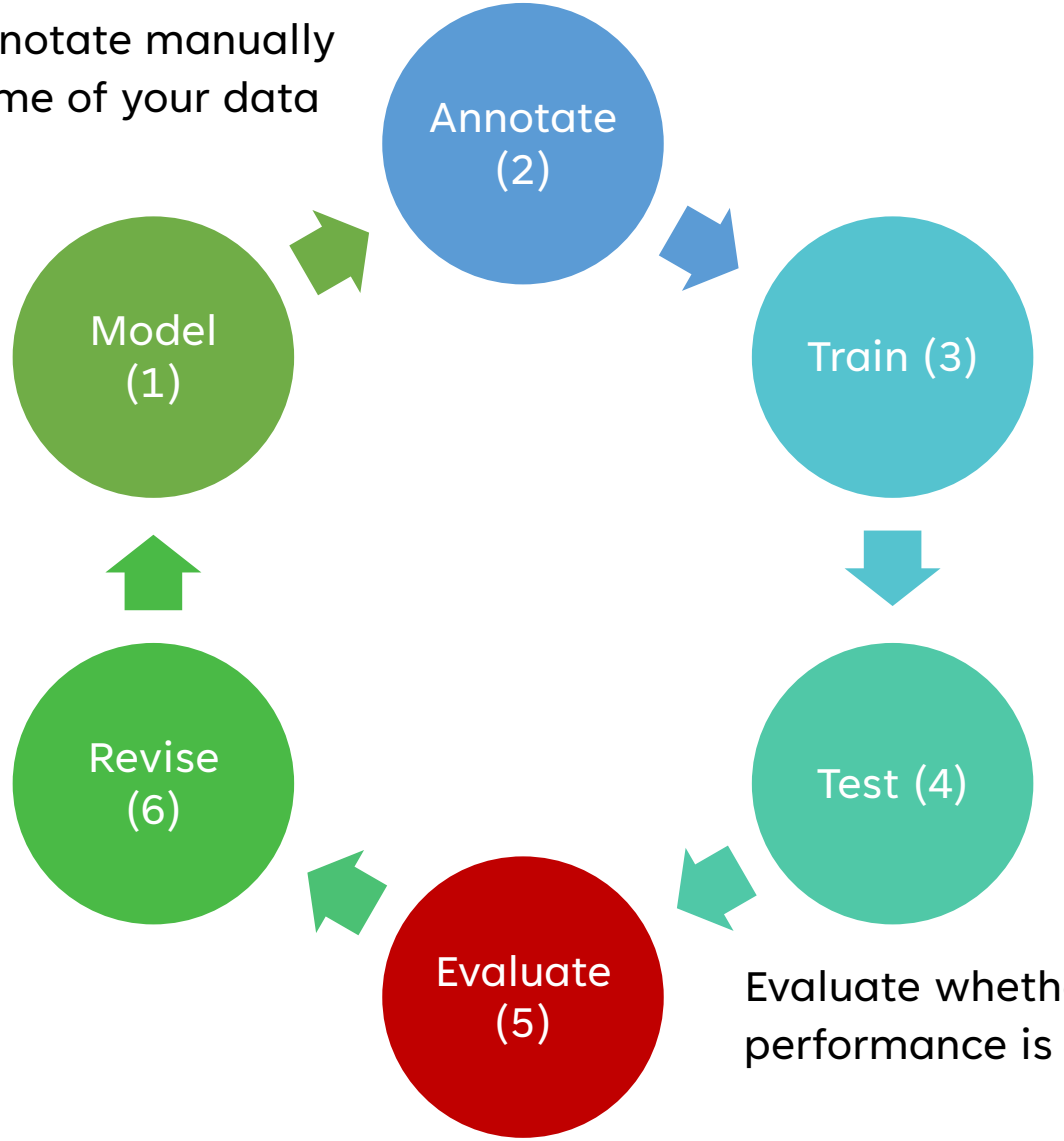


TEST DATA

- Use the trained model to annotate new data
- They need to be **unseen** data: not used during training!!
- In this way, you assess whether the model is able to **generalize** beyond the data it was trained on
- When you are annotating it is thus good practice to split the “gold” data in: train data and test data, and keep one NOT ANNOTATED version of the test data (test set)
- This is needed to compare the annotation PREDICTED by your model and the ACTUAL annotation
- Ex: GitHub - UniversalDependencies/UD_Latin-UDante

Pustejovsky and Stubbs (2012) “Natural Language Annotation for Machine Learning”. O'Reilly Media.

SUPERVISED LEARNING



HOW DO YOU EVALUATE A MODEL?

- So, does my model predict well?
- Standardized ways to evaluate the performance of the model
- Critical to provide it so that users can pick the best model
- Based on the comparison between the predicted values on the test set and the manually annotated values for the same test

Word	Predicted PoS	Gold PoS
tantae	ADJ	DET
ne	ADV	CONJ
animis	NOUN	NOUN

CONFUSION MATRIX

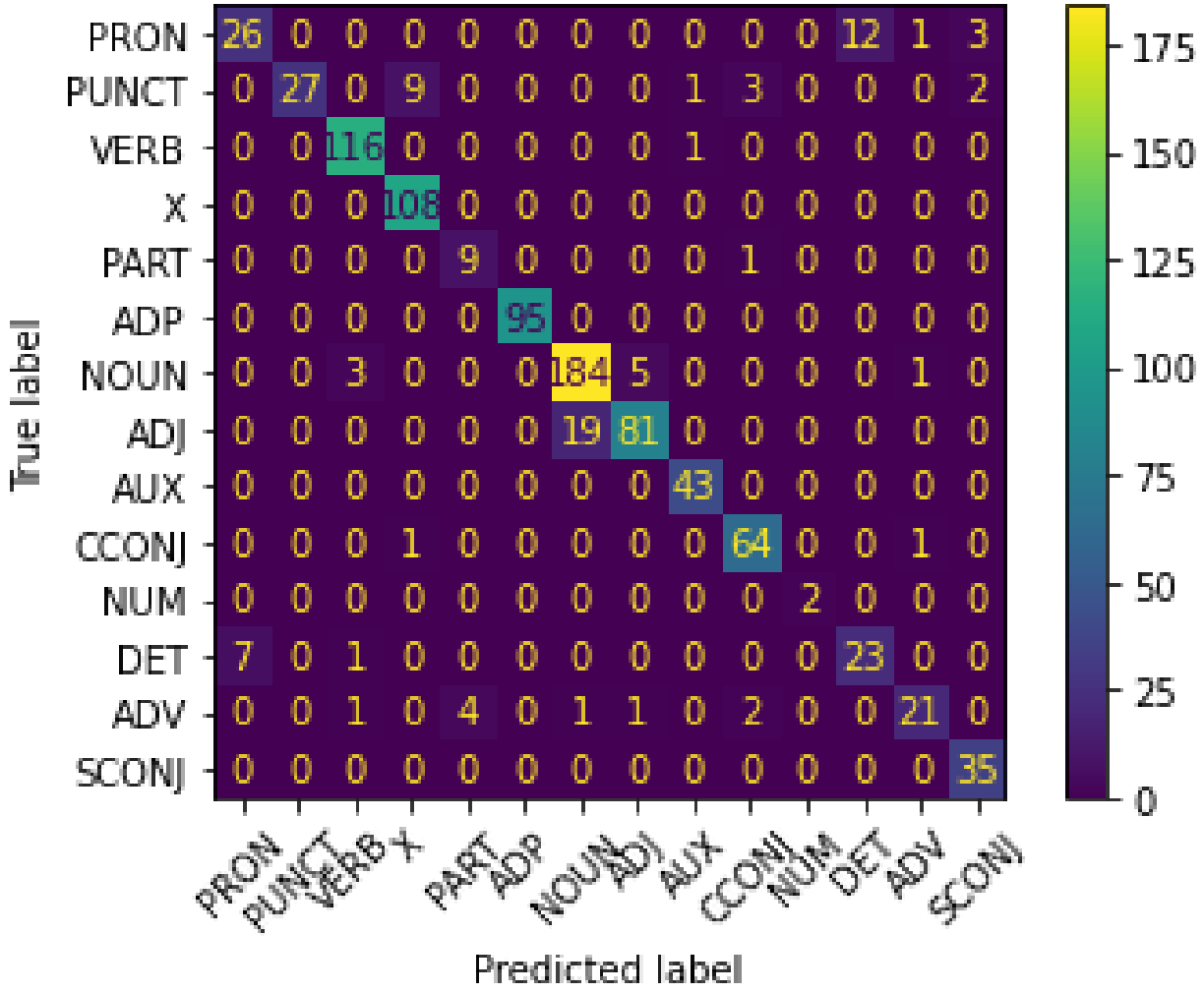
- For each PoS you count how the predicted value distributed
- Example
  - you annotated 100 NOUNS
  - For these words the model predicts 80 times NOUN, 10 ADJ, 5 DET and 5 VERB
  - The confusion matrix captures this information
- A way to visualize immediately the most problematic cases

predicted

gold

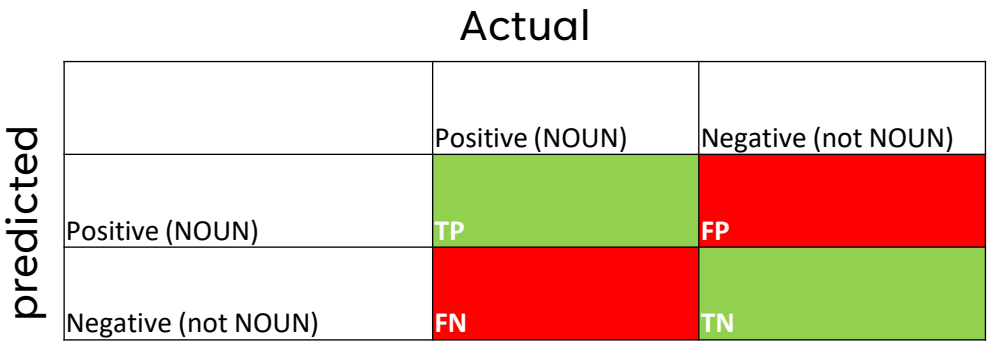
	NOUN	ADJ	PRON	DET	ADV	VERB	ADP
NOUN	80%	10%		5%		5%	
ADJ							
PRON							
DET							
ADV							
VERB							
ADP							

CONFUSION MATRIX



PER CLASS METRICS

- For each class (here, PoS) you can assess the quality of the prediction by looking at the true/false positive/negative
- E. g. if we want to evaluate the prediction of NOUNs we consider:
  - **TRUE POSITIVES (TP):** how many NOUNS were predicted as NOUNS
  - **FALSE POSITIVES (FP):** how many times NOUN was predicted when the word was not a NOUN
  - **TRUE NEGATIVES (TN):** the cases in which non NOUNS were correctly predicted as non NOUNS
  - **FALSE NEGATIVES (FN):** the cases in which NOUNs were labelled as something else



PRECISION, RECALL, F1

- **PRECISION**: ratio between the number of correct prediction and the number of total predictions, i.e. how many times NOUN is predicted correctly over the total number of times in which it is predicted

- TP/ TP+FP

- Example : 1/1+1 = 1/2

Word	Predicted PoS	Gold PoS
tantae	ADJ	DET
ne	ADV	CONJ
animis	NOUN	NOUN
caelestibus	NOUN	ADJ
irae	ADJ	NOUN
?	PUNCT	PUNCT

PRECISION, RECALL, F1

- **RECALL**: ratio between the number of correct predictions and the effective number of nouns in the text, i.e. how many times NOUN is predicted correctly over the total number of nouns in the text

- TP/ TP+FN

- Example : 1/1+1 = 1/2

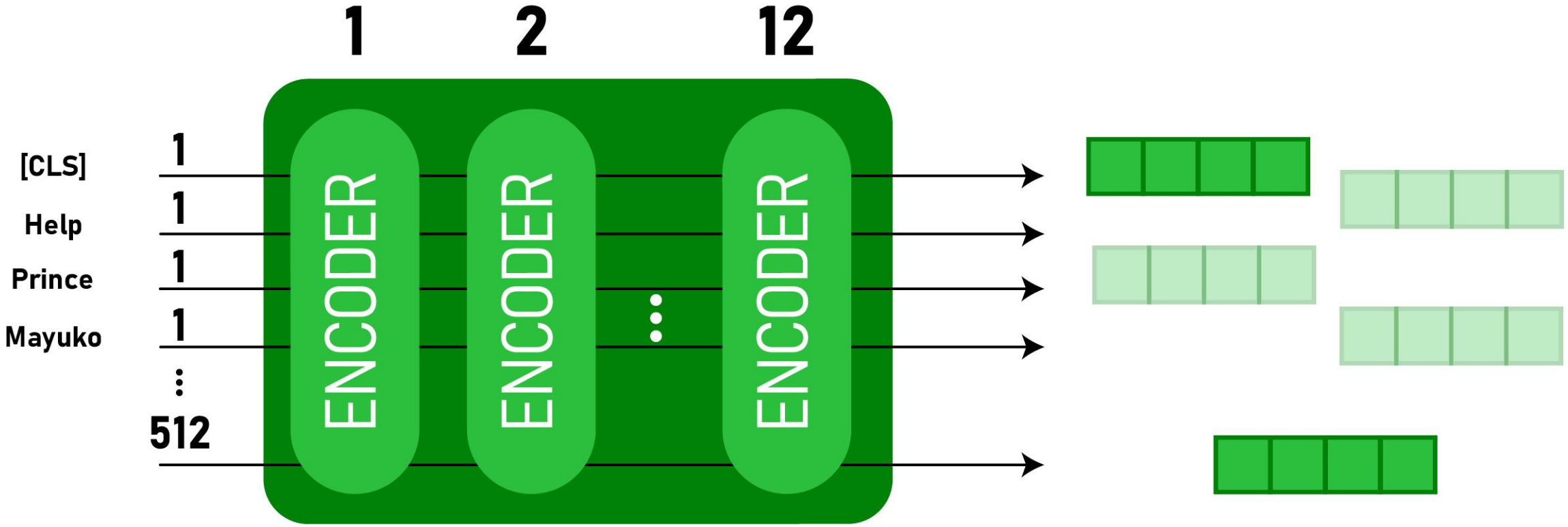
Word	Predicted PoS	Gold PoS
tantae	ADJ	DET
ne	ADV	CONJ
animis	NOUN	NOUN
caelestibus	NOUN	ADJ
irae	ADJ	NOUN
?	PUNCT	PUNCT

PRECISION, RECALL, F1

- **F1**: combination of precision and recall: harmonic mean between precision and recall
- 2* Precision*recall (precision+recall)
- Example : (2*1/2*1/2) / (1/2 + 1/2) = 1/2

LANGUAGE MODELS & DEEP LEARNING

- HMM (Hidden Markov Models) are one technique for processing sequential data, such as language
- However, much more complex approaches have been proposed, namely architectures that are able to better capture long-distance relations in the sequence and bidirectional relations
- **Transformers** were the last step of this evolution
  - Transformers read all the input text at once and assign to each token a vector (embedding)
  - This vector is built based on the contextual information of the token
  - Attention-mechanisms are used to capture the most relevant elements of the context
  - Essentially, tokens found in similar contexts are assigned similar vectors
  - In this way, it is possible to capture the semantics of terms
- Vectors can then be used for classifications task, i.e. linguistic processing



[Explanation of BERT Model - NLP - GeeksforGeeks]

LET'S TEST SOME EXISTING TOOLS

Pyrrha (psl.eu): requires login, but allows to tokenize, lemmatize & morphologically annotate a text. In addition, it also provides an interface for the postcorrection of the output (recommended!)

UDPipe (cuni.cz): UDPipe models are trained to tokenize, lemmatize, tag and syntactically parse various languages. The output are valid conllu files. Every model is trained on a different available treebank. It is easy to upload a text, select the required tasks and visualize the results in different formats. It is a pipeline that can also be trained on your specific data.

By using UDPipe, upload one of the texts (the latin version) found in the dropbox and generate the annotation. Try to test different models. Are there differences? How is the quality of the annotation?