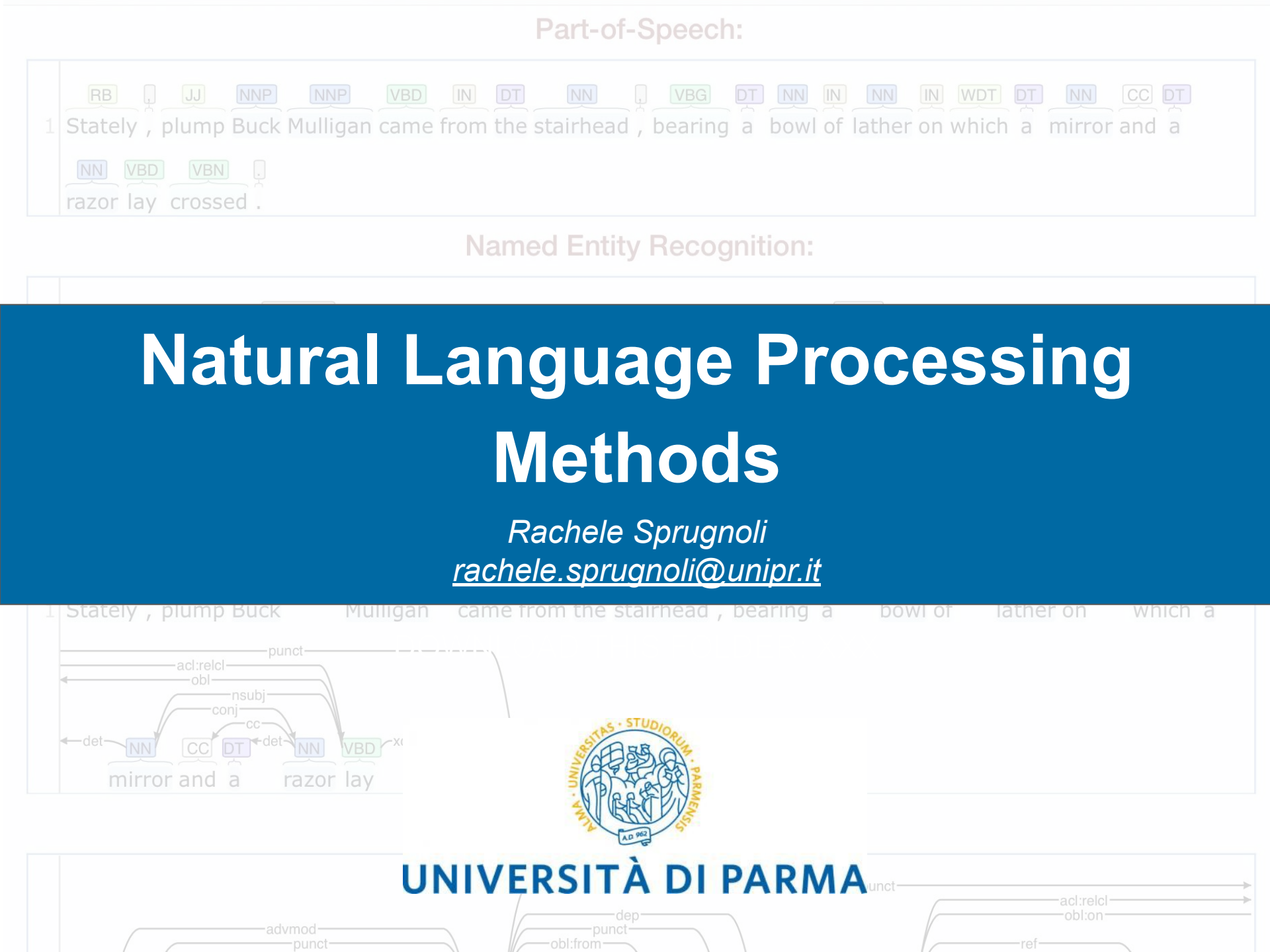
Natural Language Processing Methods

Rachele Sprugnoli
rachele.sprugnoli@unipr.it



SCHEDULE

PART 1: THEORY

- What is NLP?
- How NLP and DH interact with each other?
- Why NLP is a challenge?
- What is a NLP pipeline?
- How to develop an NLP module?
- What is manual text annotation and why is it important?
- How are NLP systems evaluated?

PART 2: HANDS-ON SESSION

- How to use CLARIN-ERIC tools for text processing?

PART 1
A LITTLE BIT OF THEORY…

COMPUTATIONAL LINGUISTICS

versus

NATURAL LANGUAGE PROCESSING

Computational linguistics and natural language processing […] are sometimes used interchangeably to describe the field concerned with the processing of human language by computers

- **Computational Linguistics** is used to describe research interested in answering linguistic questions using computational methodology
- **Natural Language Processing** describes research on automatic processing of human language for practical applications

Bender, Emily M. 2016. "Linguistic Typology in Natural Language Processing". Linguistic Typology 20(3), 645-660.

BUT…

Computers do NOT know natural language!

Natural Language Processing (NLP) aims to equip the computer with linguistic knowledge, to create machines that understand (and even reproduce) natural language, to develop programs that assist human beings in linguistic tasks, such as:

- automatic speech recognition
- speech synthesis
- machine translation
- sentiment analysis

APPLICATIONS

- LIBRARIES and PUBLISHING: recognize authors / bibliographic references, identify relevant articles, suggest reading paths, monitor the opinion of readers
- HISTORY: extracting events from sources, identifying sources on similar topics, improving the quality of OCR for the digitization of sources
- LITERATURE: identify linguistic and stylistic characteristics
- MUSEUMS: generate (semi-) automatically the descriptions of artworks, enrich the descriptions, identify similar works, create personalized museum visits

NLP & DIGITAL HUMANITIES

- DH is the field in which Humanities and NLP can interact and support each other

- 2 directions of research:
  - Humanities for NLP
  - NLP for Humanities

- Roots of this interaction: Father Roberto Busa pioneering work "*Index Thomisticus*"

NLP & DIGITAL HUMANITIES

LET'S READ BUSA

"L'Analisi Linguistica nell'Evoluzione Mondiale dei Mezzi d'Informazione" 1962

- The advent of automation: a monster for humanism

"[...] *At this point a nightmare intervened, technology triumphant with its latest creation: automation.*

People shuddered, considering it a crude, hard bulldozer that goes roaring ahead, crushing and shredding flowers, amongst which, a delicate and gentle victim, is humanism."

LET'S READ BUSA

"L'Analisi Linguistica nell'Evoluzione Mondiale dei Mezzi d'Informazione" 1962

- New questions for humanists

"[...] *the men involved in automation began to […] 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?*
- *Please, would you arrange all the words in the dictionary according to the various morphological and grammatical categories?*
- *Would you please tell me which words may be omitted, and when, so as to shorten a text without any detriment to its meaning?"*

LET'S READ BUSA

"L'Analisi Linguistica nell'Evoluzione Mondiale dei Mezzi d'Informazione" 1962

- New questions for humanists

"[...] *the men involved in automation began to […] 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?* ⇒ **PARSING**
- *Please, would you arrange all the words in the dictionary according to the various morphological and grammatical categories?* ⇒ **PoS TAGGING**
- *Would you please tell me which words may be omitted, and when, so as to shorten a text without any detriment to its meaning?* ⇒ **TEXT SUMMARIZATION**

LET'S READ BUSA

"L'Analisi Linguistica nell'Evoluzione Mondiale dei Mezzi d'Informazione" 1962

- Too little humanism!

"[...] *a machine made us realize that no humanist has such command of his own language as to be able to answer such questions. A machine […] has revealed that there is still too little humanism of the serious and systematic type."*

"Not only do computers invite us to wider, deeper, and more systematic research, they also make it possible."

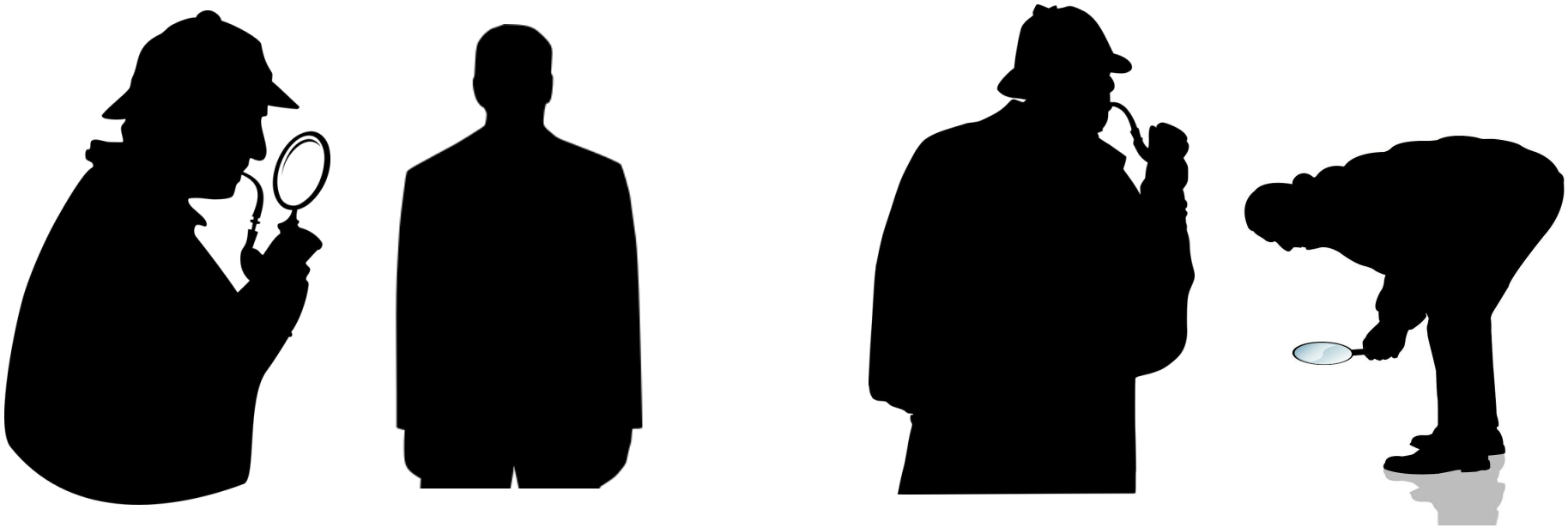
WHY IS NLP A CHALLENGE?

1. Grammatical ambiguity

PAROLA	CATEGORIA GRAMMATICALE
Do	VERB/NOUN
not	ADVERB/NOUN
pity	NOUN/VERB
the	ARTICLE
dead	ADJECTIVE/NOUN/ADVERB
,	PUNCTUATION
Harry	PROPER NAME
.	PUNCTUATION

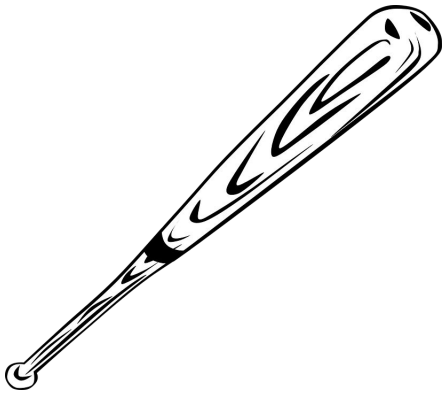
WHY IS NLP A CHALLENGE?

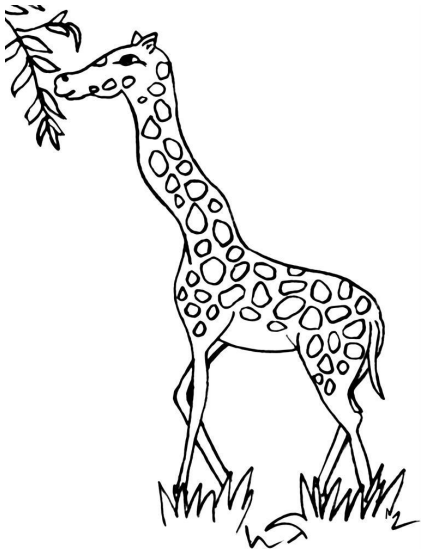
2. Syntactic ambiguity: «*Sherlock saw a man with a magnifying glass*»

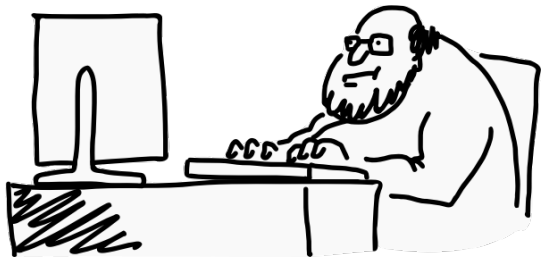


WHY IS NLP A CHALLENGE?

3. Semantic ambiguity: «*bat*» / «*browsing*»







WHY IS NLP A CHALLENGE?

4. The language changes

   - Classical / historical languages:

Ahi quanto a dir qual era è cosa dura
esta selva selvaggia e aspra e forte
che nel pensier rinova la paura!

   - Non-standard languages:

   - Neologisms: *Brexit*

WHY IS NLP A CHALLENGE?

5. Multi-word expressions, or "2 + 2 is not always 4"

Their meaning does not correspond to the lexical combination of the words that compose them, examples:

- metaphorical expressions: “we sailed the seven seas”
- light-verb constructions: “to take a shower”
- phrasal verbs: “to give up”
- idioms: “it’s raining cats and dogs”

WHY IS NLP A CHALLENGE?

6. We need contextual information and world knowledge

«Elsa and Anna are sisters»

WHY IS NLP A CHALLENGE?

7. We need to understand irony

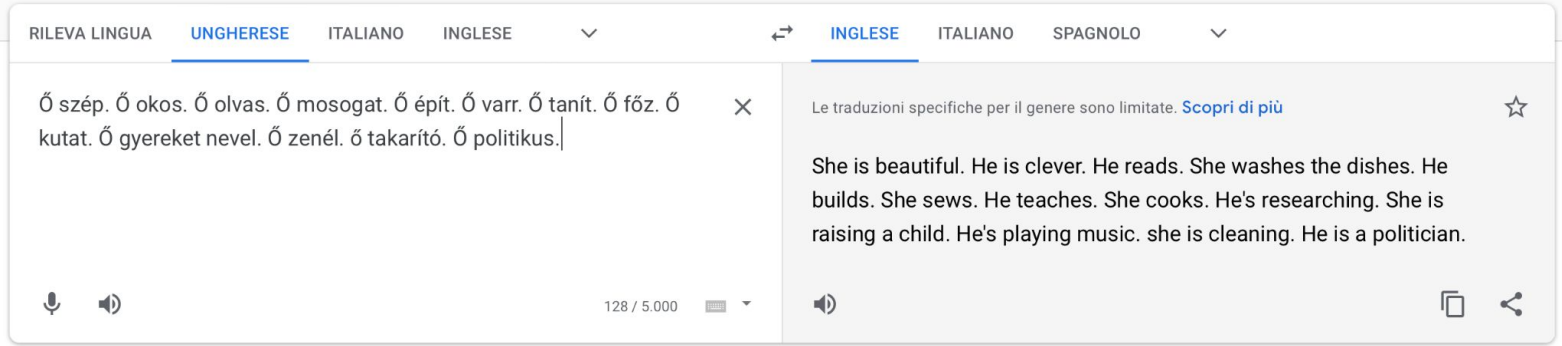
“She has a face like a Botticelli Madonna!”
“He looks like a Picasso painting!”

WHY IS NLP A CHALLENGE?

8. Our texts are full of bias: examples, Microsoft's Tay chatbot (2016) and Google Translate (from Hungarian to English)

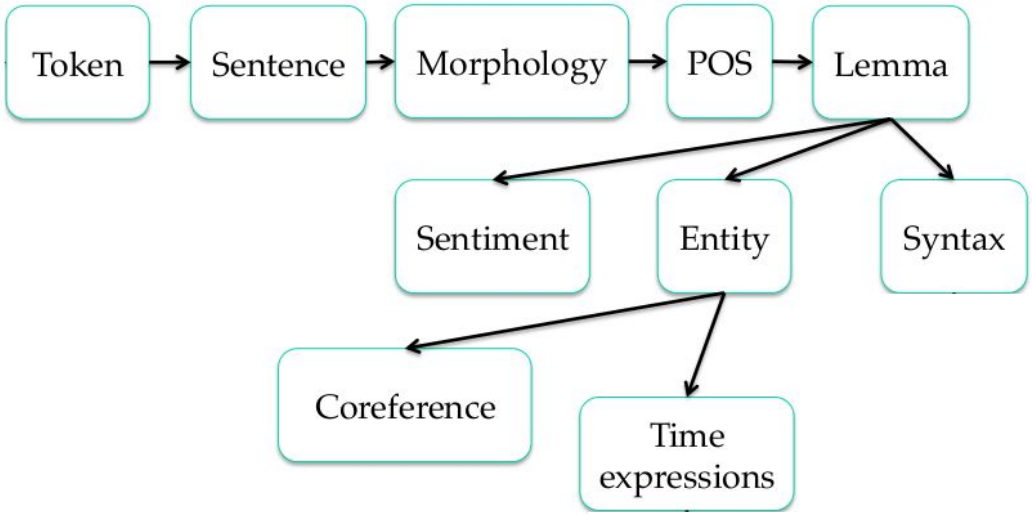
The Telegraph

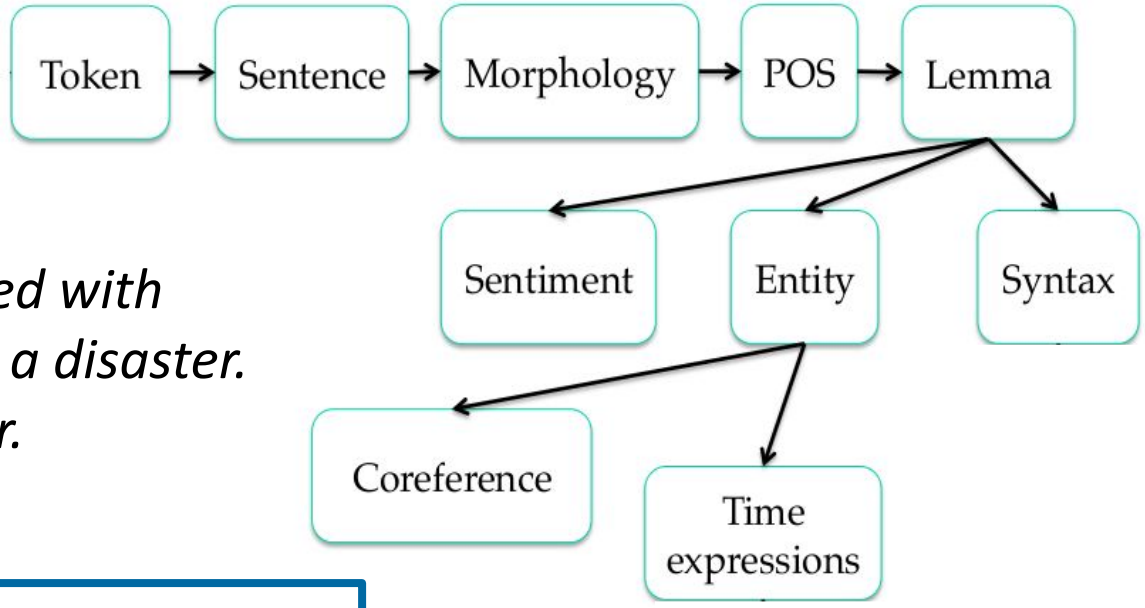
Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours



HOW TO PROCESS LANGUAGE

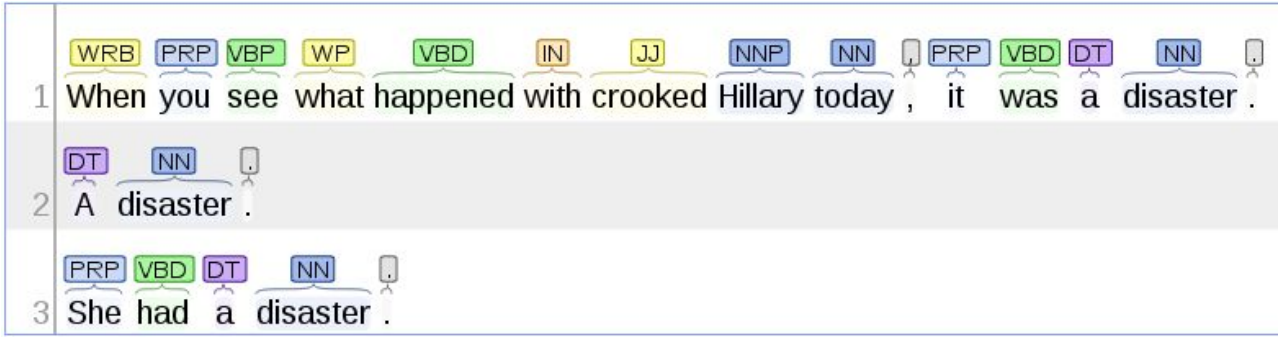
- PIPELINE structure: chain whose modules each describe a different level of linguistic analysis and where the output of one module becomes the input for the next module. Example:

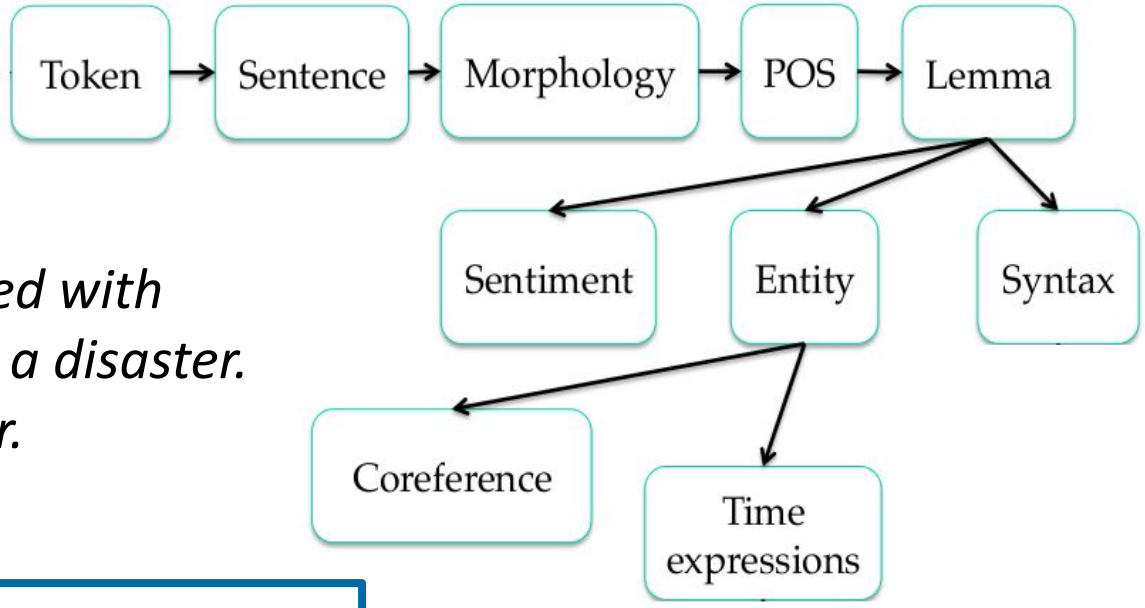




When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH





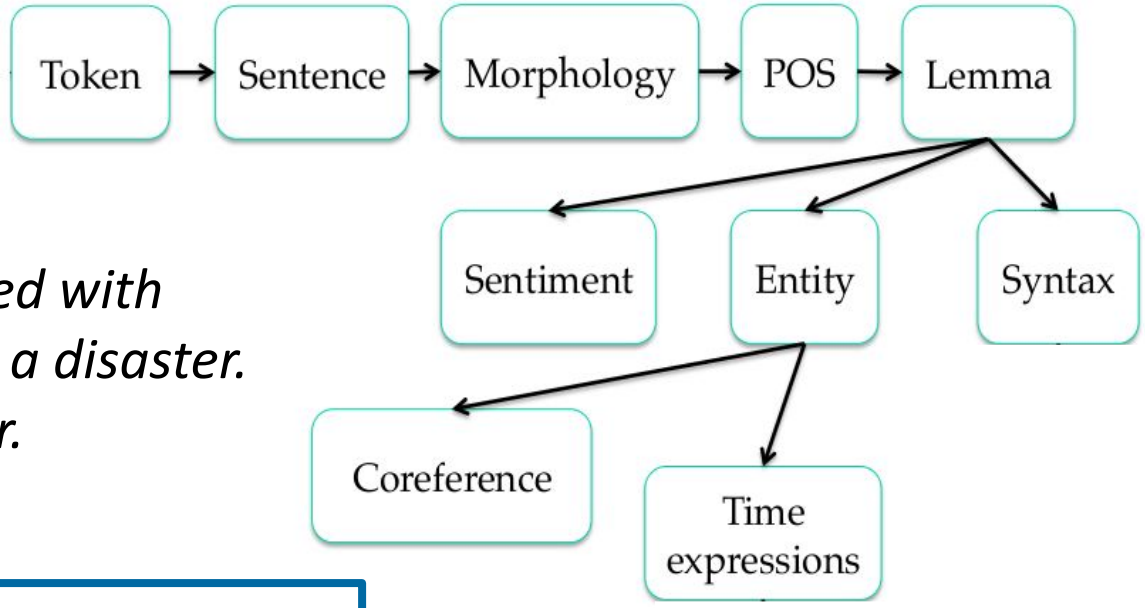
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

Do not pity the dead, Harry.

- HOW MANY TOKENS?

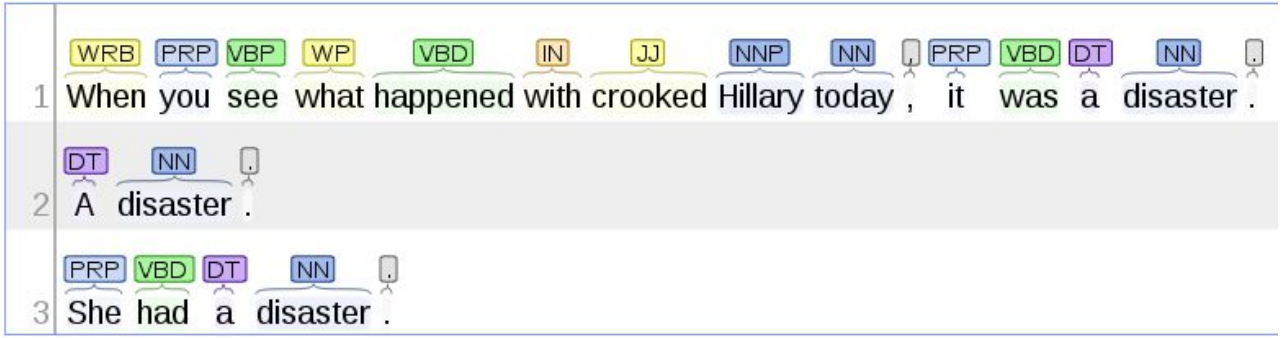
Do | not | pity | the | dead, | Harry. → 6?
Do | not | pity | the | dead | , | Harry | . → 8?

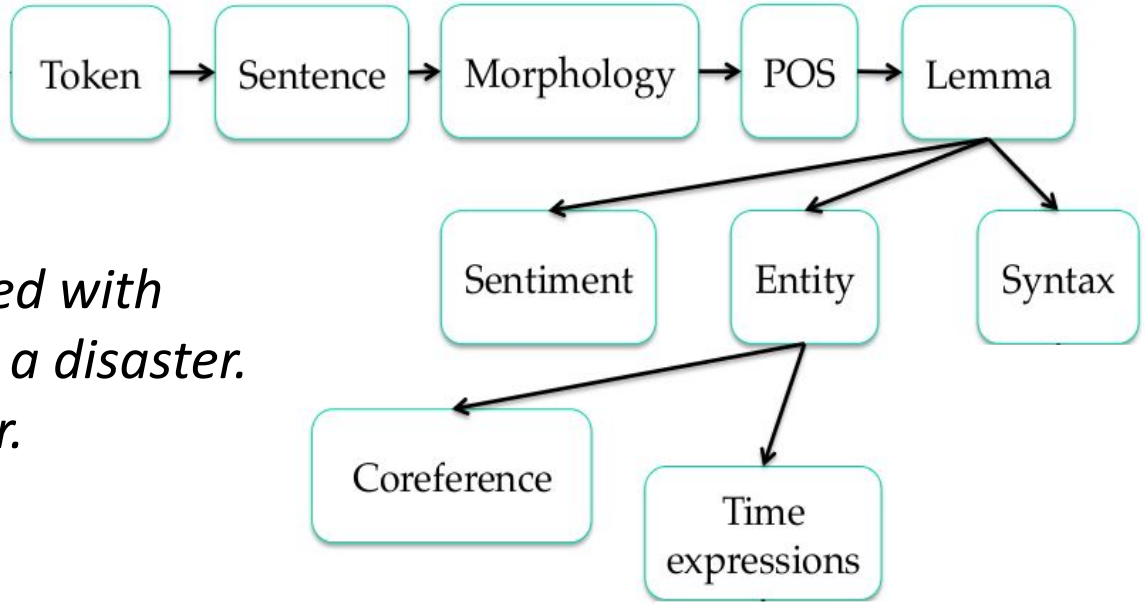


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

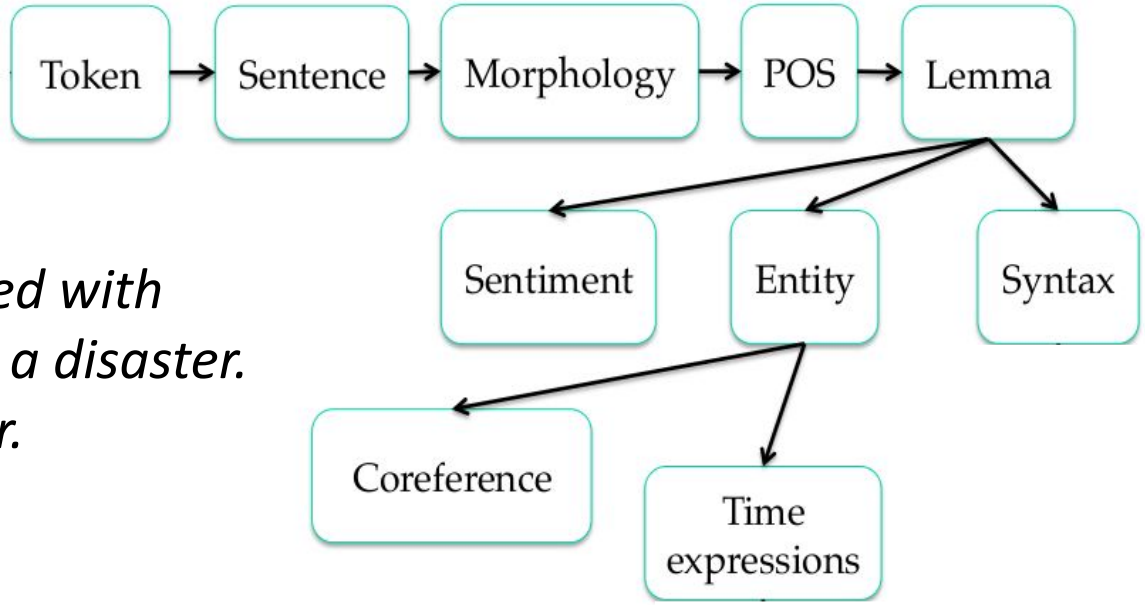
PoS Tags: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html





When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

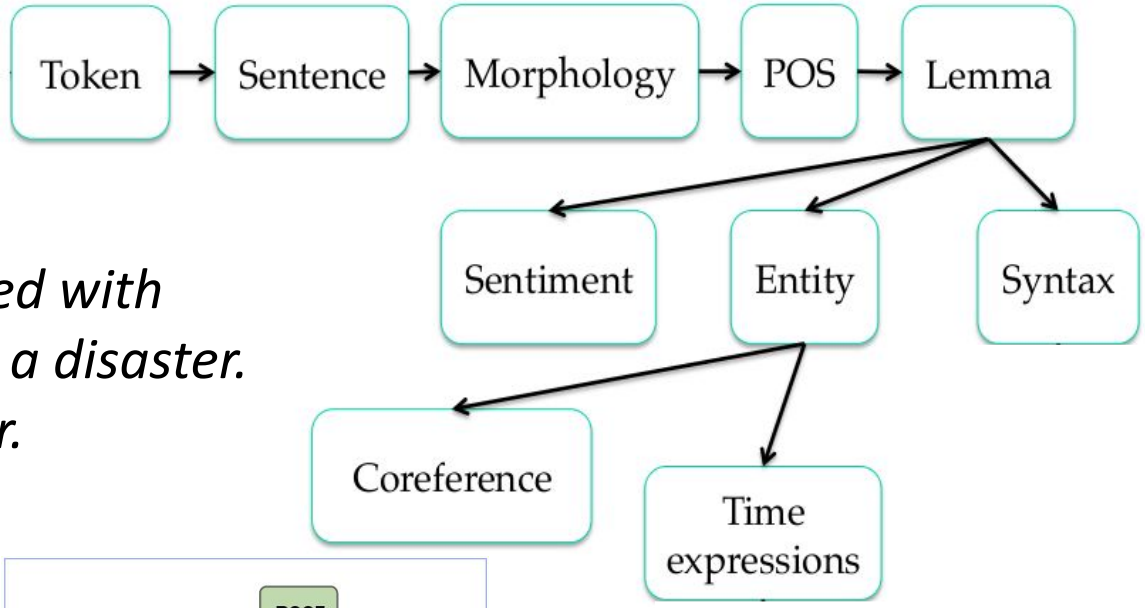
MORPHOLOGY



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

LEMMA

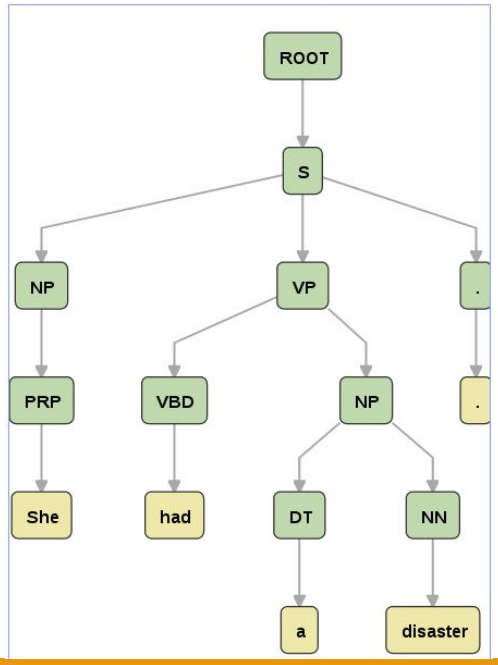
1	when you see what happen with crooked Hillary today , it be a disaster .	When you see what happened with crooked Hillary today , it was a disaster .
2	a disaster .	A disaster .
3	she have a disaster .	She had a disaster .

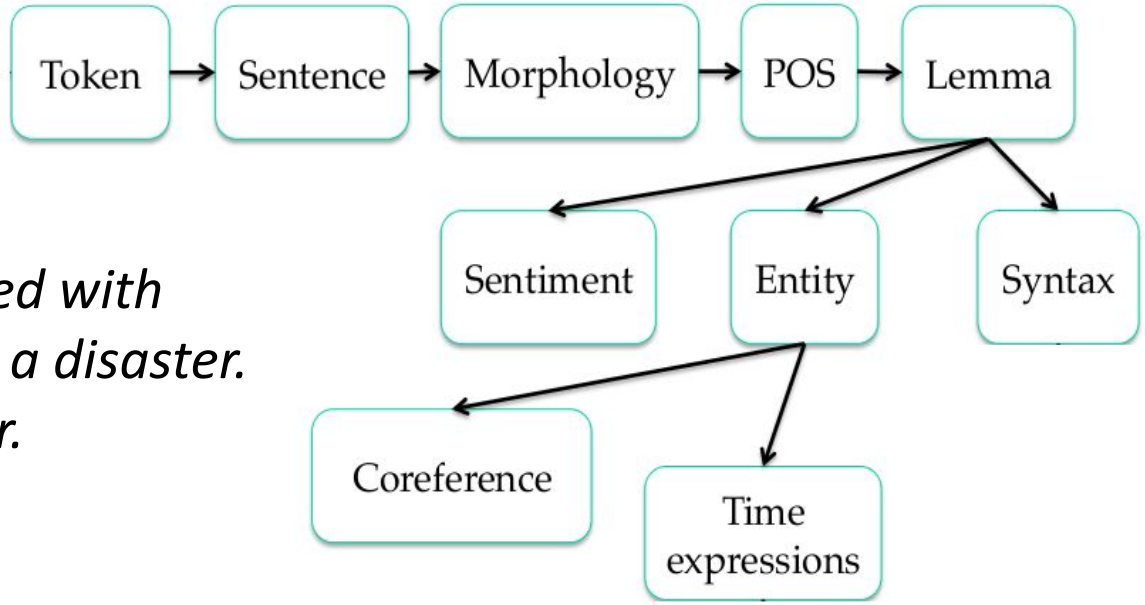


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX / PARSING

- CONSTITUENCY PARSING

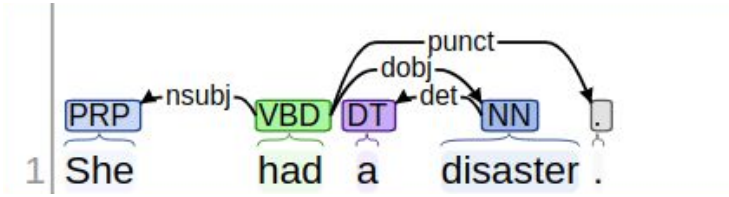


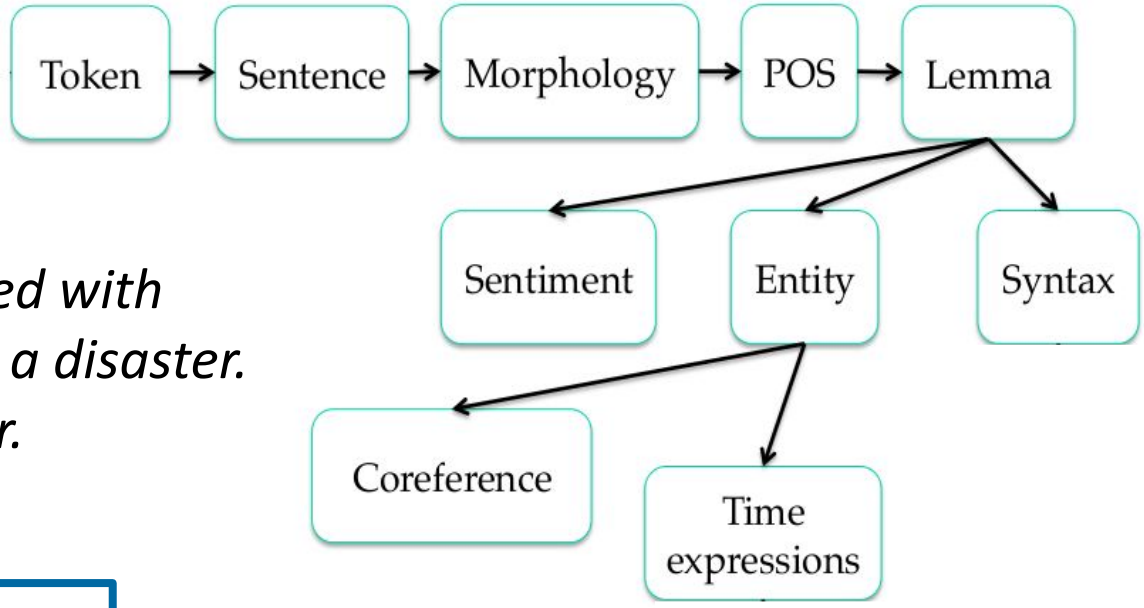


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX / PARSING

- DEPENDENCY PARSING



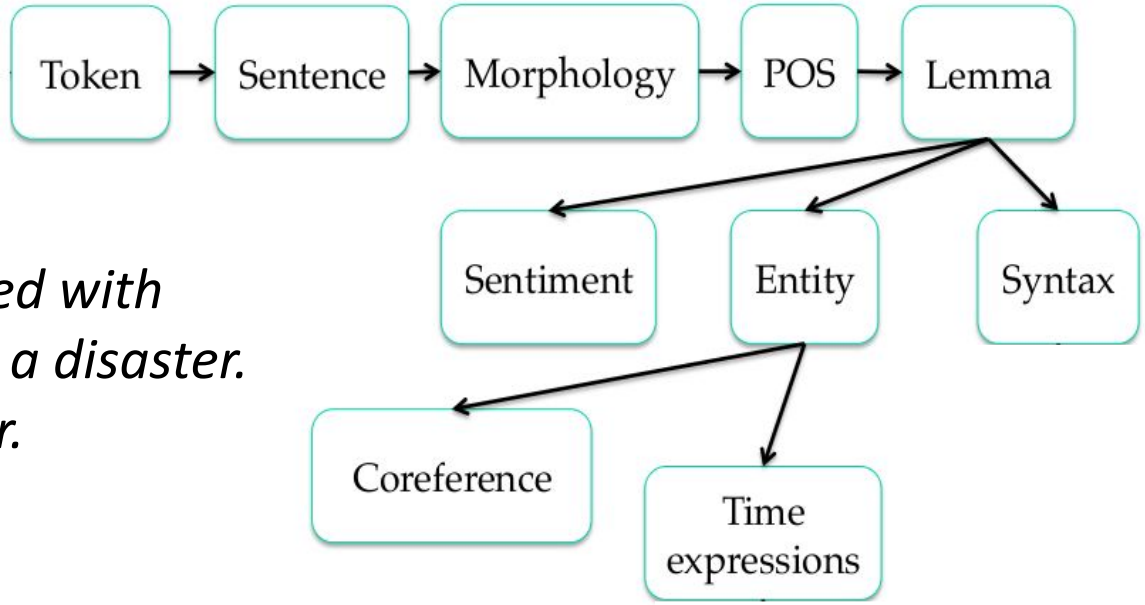


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

NAMED ENTITY RECOGNITION

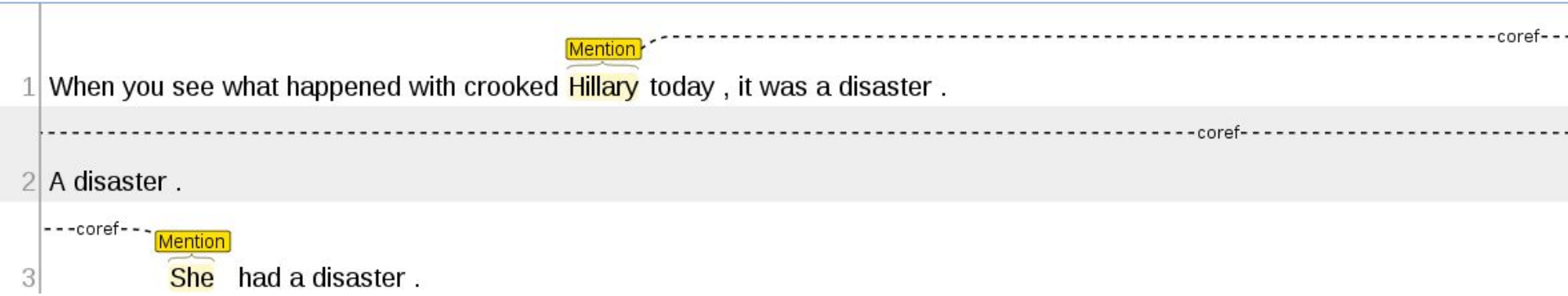
PER

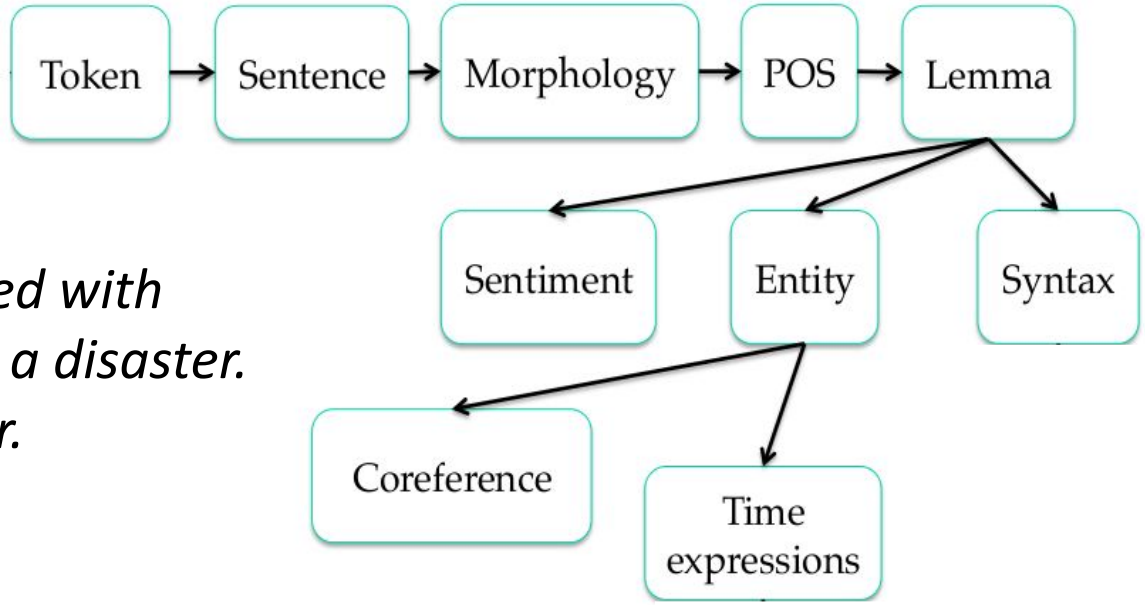
1 When you see what happened with crooked Hillary today , it was a disaster .
2 A disaster .
3 She had a disaster .



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

COREFERENCE



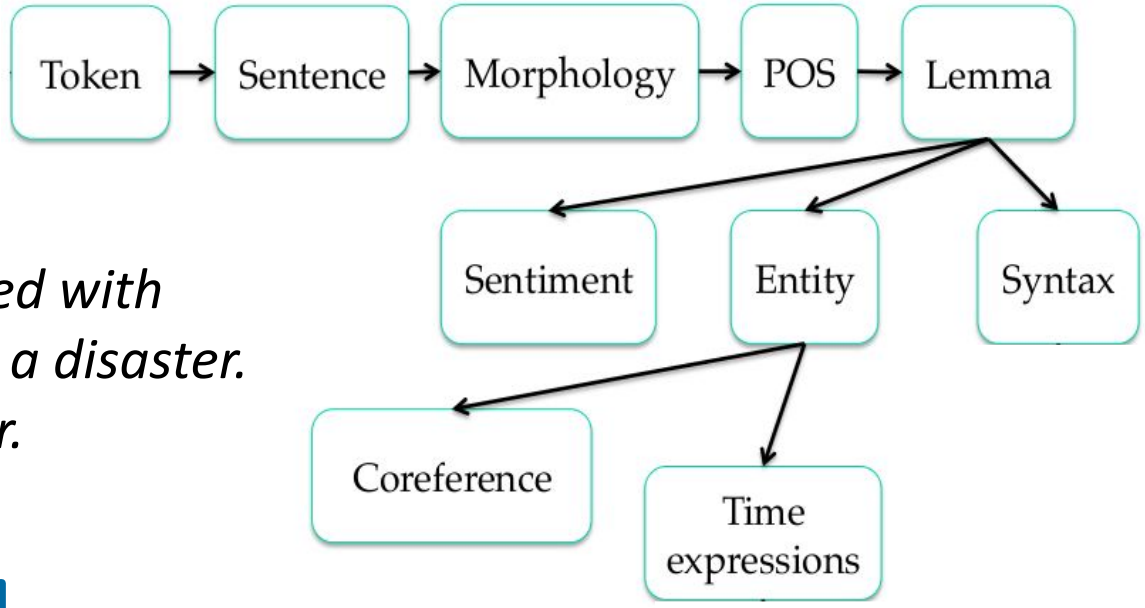


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TIME EXPRESSIONS

2016-08-05

1 When you see what happened with crooked Hillary today , it was a disaster .
2 A disaster .
3 She had a disaster .



When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SENTIMENT

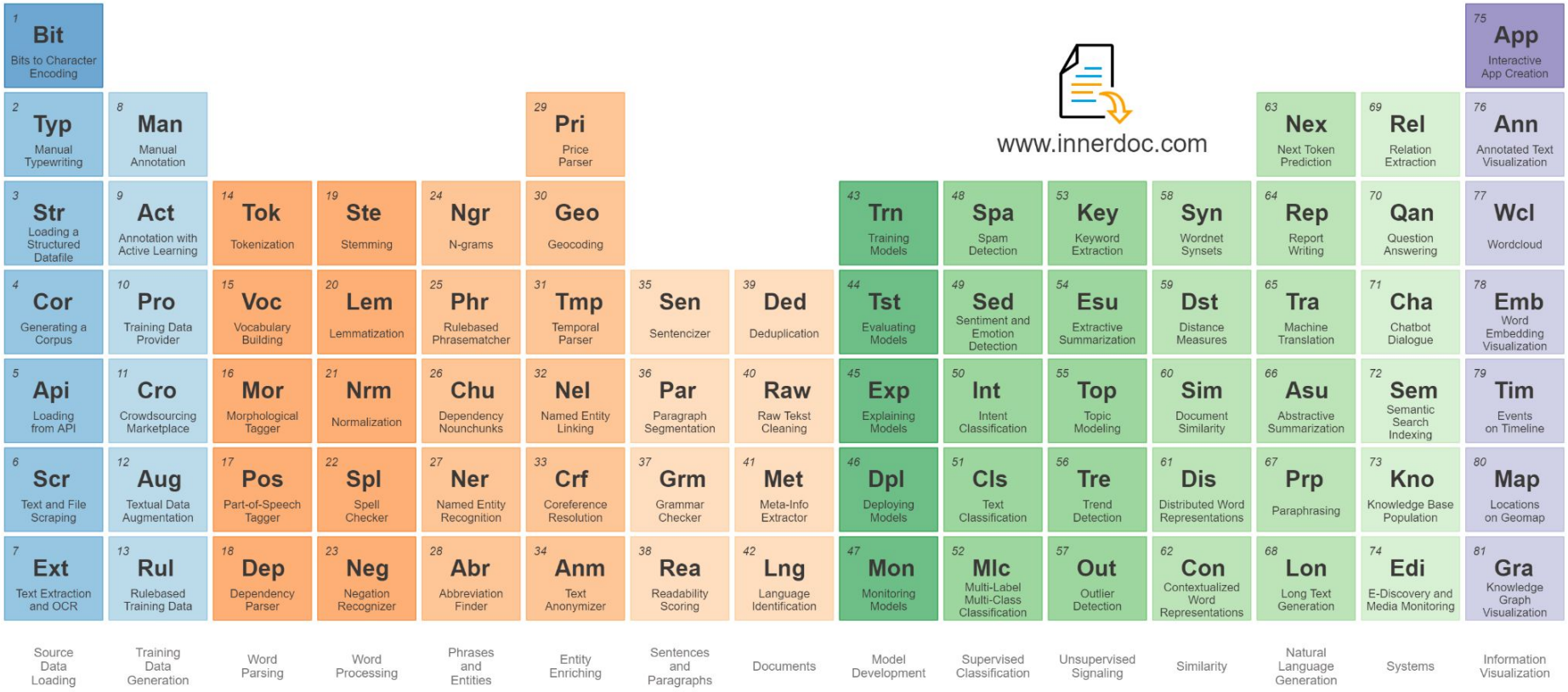
NEGATIVE
1 When you see what happened with crooked Hillary today , it was a disaster .

VERY NEGATIVE
2 A disaster .

NEGATIVE
3 She had a disaster .

SO MANY TASKS…

Periodic Table of Natural Language Processing Tasks



HOW TO DEVELOP A MODULE

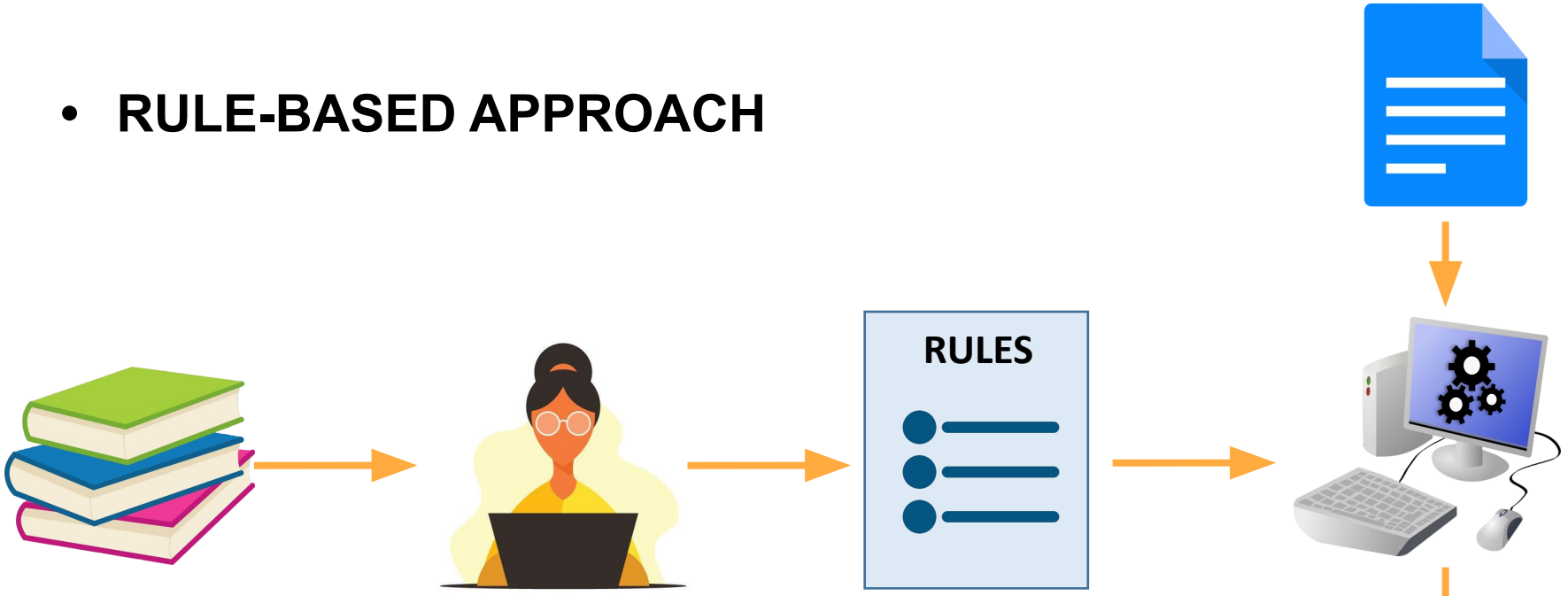
- **LOOKUP LIST APPROACH**

- Systems that recognize only the words stored in lists called "gazetteers"
- Pros: simple, fast, easy to use
- Cons: collecting and maintaining lists takes time, lists do not handle all possible variations of words and cannot resolve ambiguity or make any kind of inference

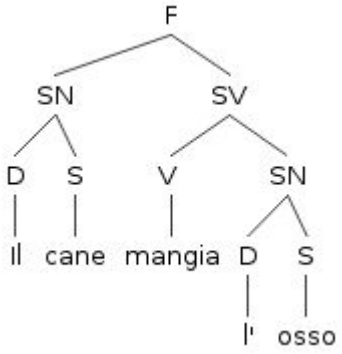
CURRENCIES	CITIES
Euro, euros dollar, dollars, pound, $, €...	http://download.geonames.org/export/dump/

HOW TO DEVELOP A MODULE

- **RULE-BASED APPROACH**



- Pros: based on linguistic evidence, accurate
- Cons: difficult to extend or adapt to new domains, slow development



HOW TO DEVELOP A MODULE

- **RULE-BASED APPROACH**
  - Example: Part-of-Speech tagging

1) assignment to each word of all possible PoS using a dictionary

NOUN VERB *«pity*	DET *the*	NOUN ADJ ADV *dead»*

2) application of rules to remove ambiguous labels

- «choose NOUN if preceded by DET»

NOUN VERB *«pity*	DET *the*	NOUN ~~ADJ~~ ~~ADV~~ *dead»*

HOW TO DEVELOP A MODULE

- **RULE-BASED APPROACH**

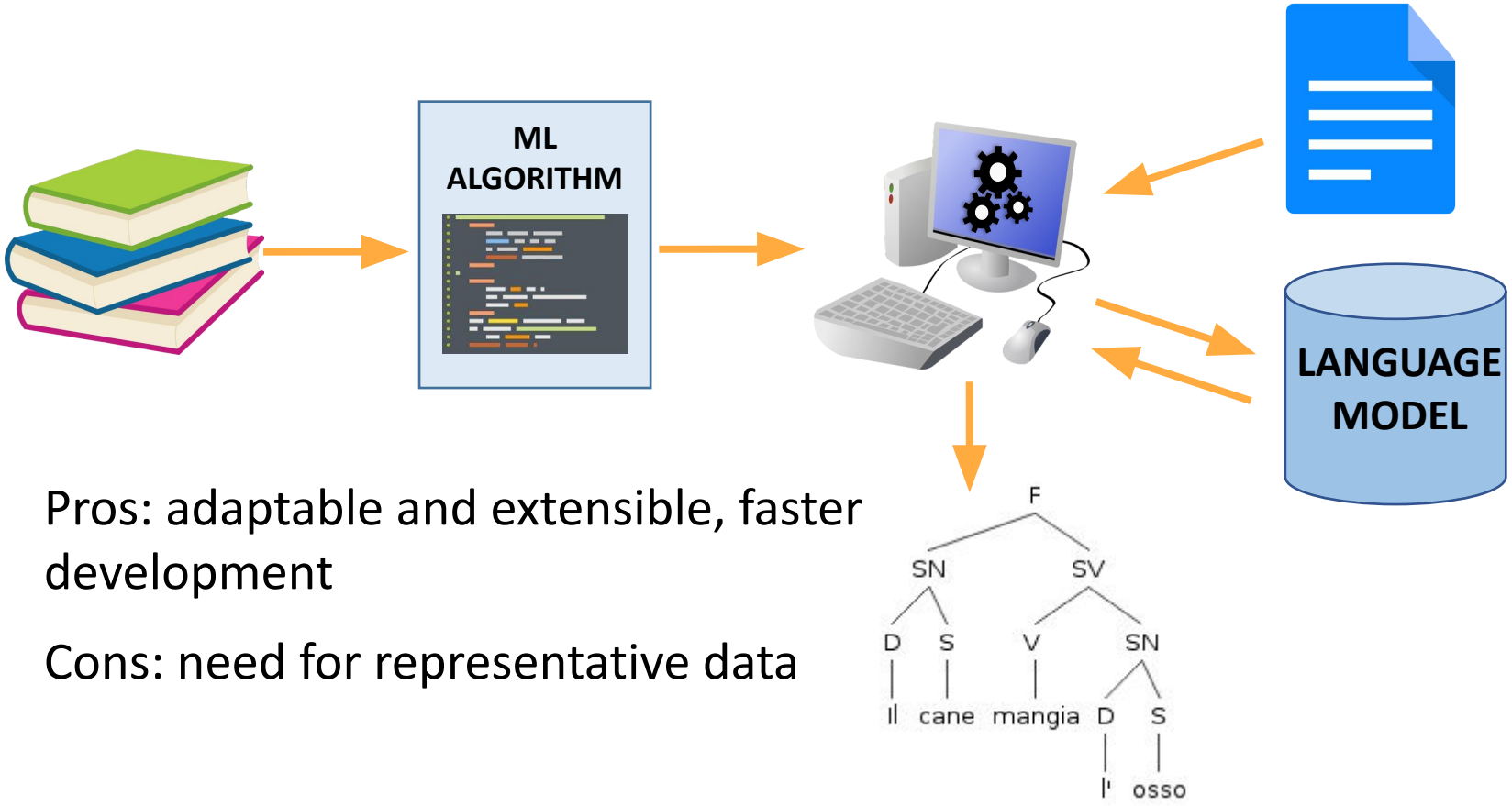
Named Entity Recognition without Gazetteers,

Mikheev et al. 1999

Context Rule	Assign	Example
Xxxx+ is? a? JJ* PROF	PERS	Yuri Gromov, a former director
Xxxx+ is? a? JJ* REL	PERS	John White is beloved brother
Xxxx+ himself	PERS	White himself
Xxxx+, DD+,	PERS	White, 33,
shares in Xxxx+	ORG	shares in Trinity Motors
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
Xxxx+ area	LOC	Beribidjan area

HOW TO DEVELOP A MODULE

- **MACHINE LEARNING (ML) APPROACH**



- Pros: adaptable and extensible, faster development
- Cons: need for representative data

HOW TO DEVELOP A MODULE

- **MACHINE LEARNING APPROACH**

- **3 main types of ML algorithms**

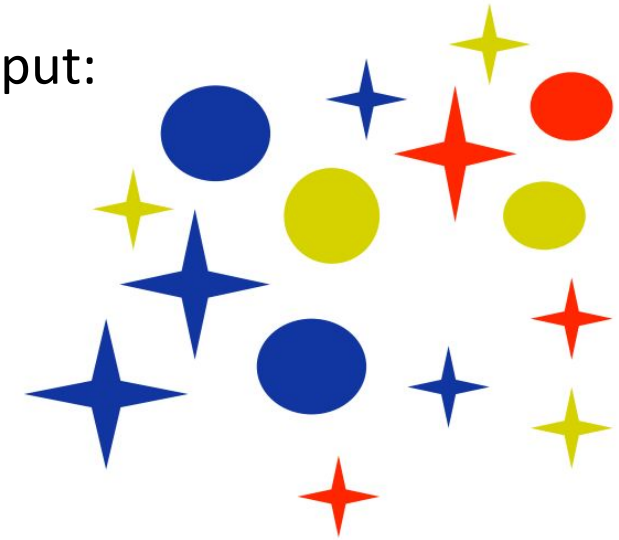
1. UNSUPERVISED: they do not need a hand-annotated corpus for training the model
2. SUPERVISED: they use a hand-annotated corpus for training the model
3. SEMI-SUPERVISED: they combine information from both annotated and non-annotated data for training the model

HOW TO DEVELOP A MODULE

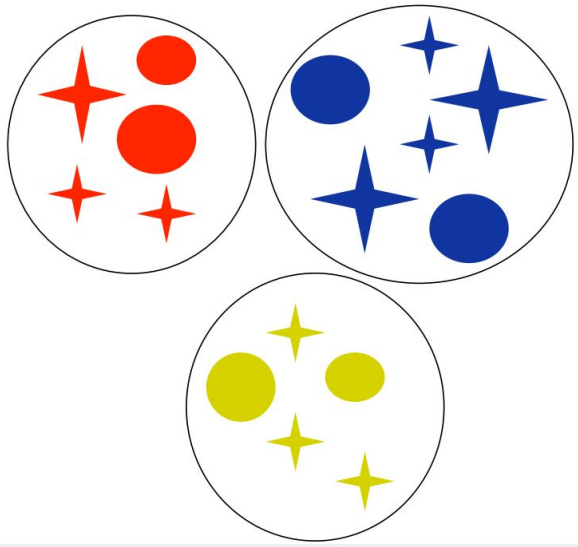
- **MACHINE LEARNING APPROACH**
- **UNSUPERVISED ALGORITHM, example**

- CLUSTERING: grouping of the input based on some relationship of similarity between the data

Input:



Color-based output:

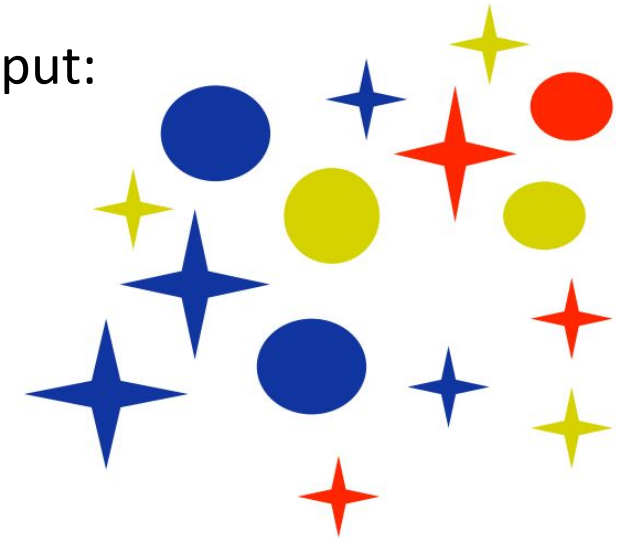


HOW TO DEVELOP A MODULE

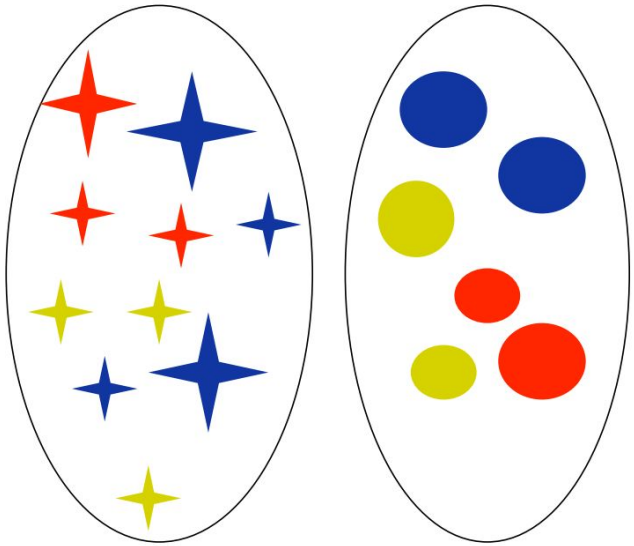
- **MACHINE LEARNING APPROACH**
- **UNSUPERVISED ALGORITHM, example**

- CLUSTERING: grouping of the input based on some relationship of similarity between the data

Input:



Form-base output:



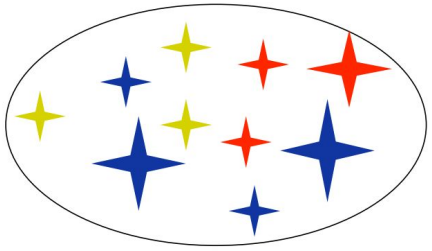
HOW TO DEVELOP A MODULE

- **MACHINE LEARNING APPROACH**
- **SUPERVISED ALGORITHM, example**

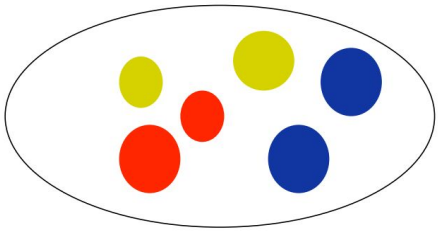
- CLASSIFICATION: given a set of predefined classes, determine which class a certain linguistic element belongs to

Input (training):

A



B

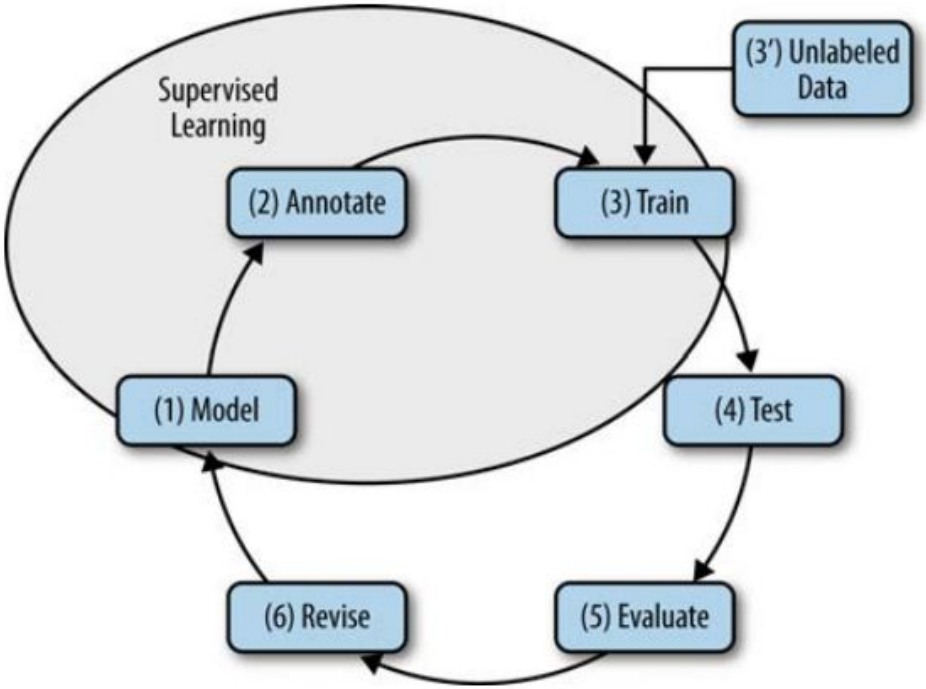


Classification of unseen data (test):

B A B A B A A A

HOW TO DEVELOP A MODULE

- **MACHINE LEARNING APPROACH**
- **SUPERVISED ALGORITHM**



MATTER CYCLE

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

HOW TO DEVELOP A MODULE

- **MACHINE LEARNING APPROACH**
- **SUPERVISED ALGORITHM**

The MATTER cycle:

1. **Model**: theoretical description of a linguistic phenomenon
2. **Annotate**: data annotation with a model-based annotation scheme
3. **Train**: training of an ML algorithm on the annotated corpus
4. **Test**: test the trained system on a new sample of data
5. **Evaluate**: system performance evaluation
6. **Revise**: revision of the annotation model and scheme

ANNOTATION

- adding (linguistic) information to text via labels (tags)
- it covers every aspect of linguistic analysis
- it makes explicit the linguistic structure implicit in the text

- **ANNOTATION SCHEME**

- repertoire of categories for annotation: list of tags and attributes

- **ANNOTATION GUIDELINES**

- document explaining the way in which the annotation is projected on the text

ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): principles**

1) Dependency Parsing
   - available in many treebanks and many languages
2) Lexicalism
   - the fundamental units of the annotation are the syntactic words: split off clitics, undo contractions
   - syntactic words have morphological properties and enter into syntactic relationships
3) Recoverability
   - transparent mapping between input text and segmentation into syntactic words
4) Universality
   - universal inventory of categories and guidelines

ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): UPOS tags**

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

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

ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): Features**

Lexical features	Inflectional features	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

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

ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): Syntactic Relations**

	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

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

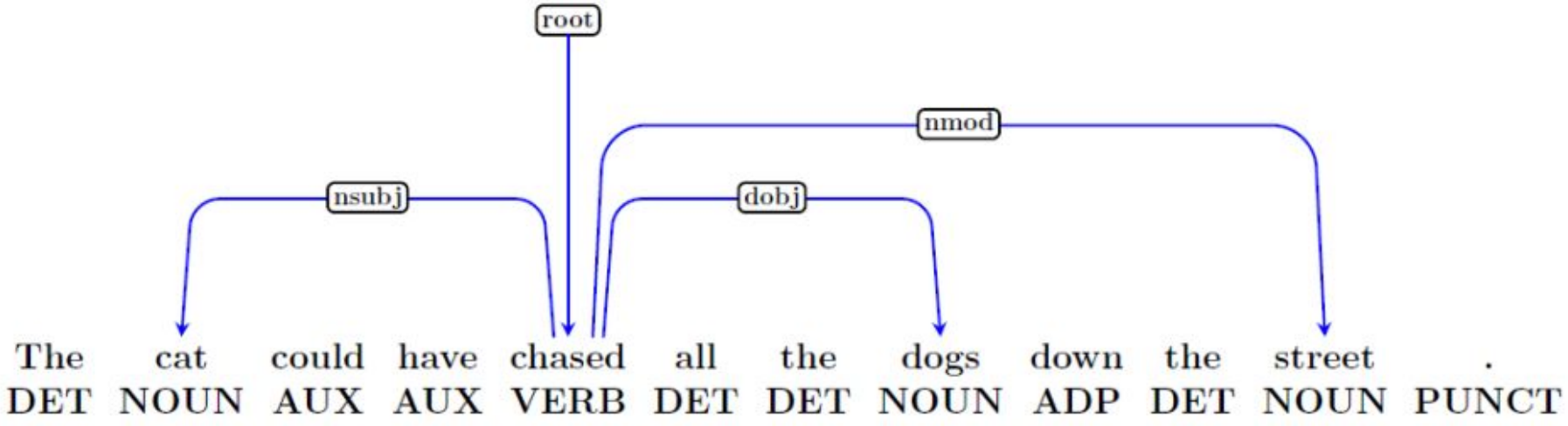
ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): annotation**

The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

ANNOTATION: EXAMPLE

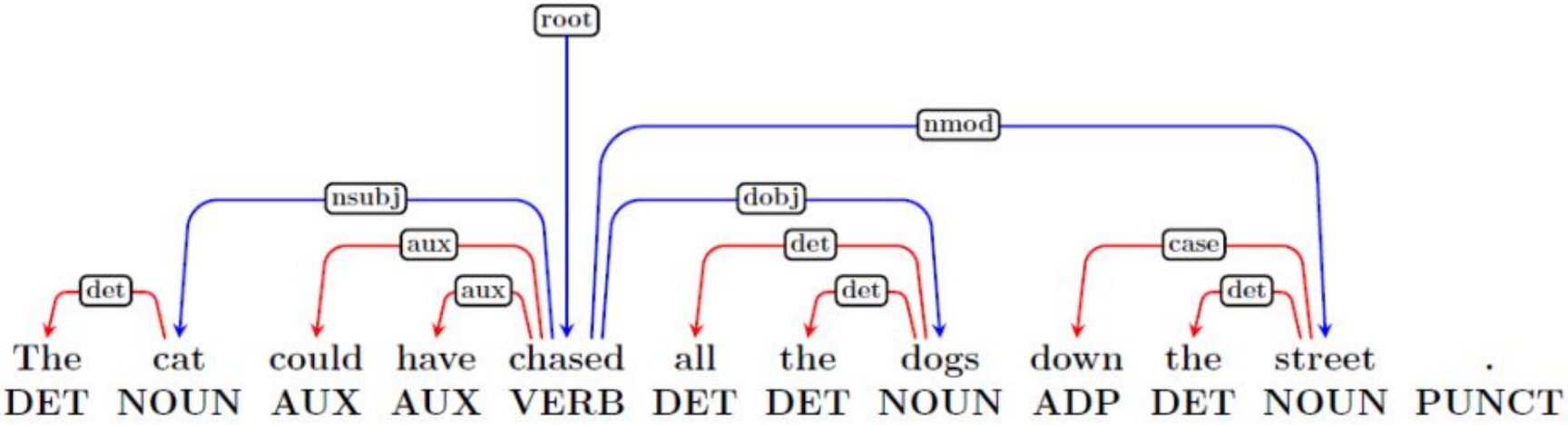
- **UNIVERSAL DEPENDENCIES (UD): annotation**



1. Content words are linked with dependencies relations

ANNOTATION: EXAMPLE

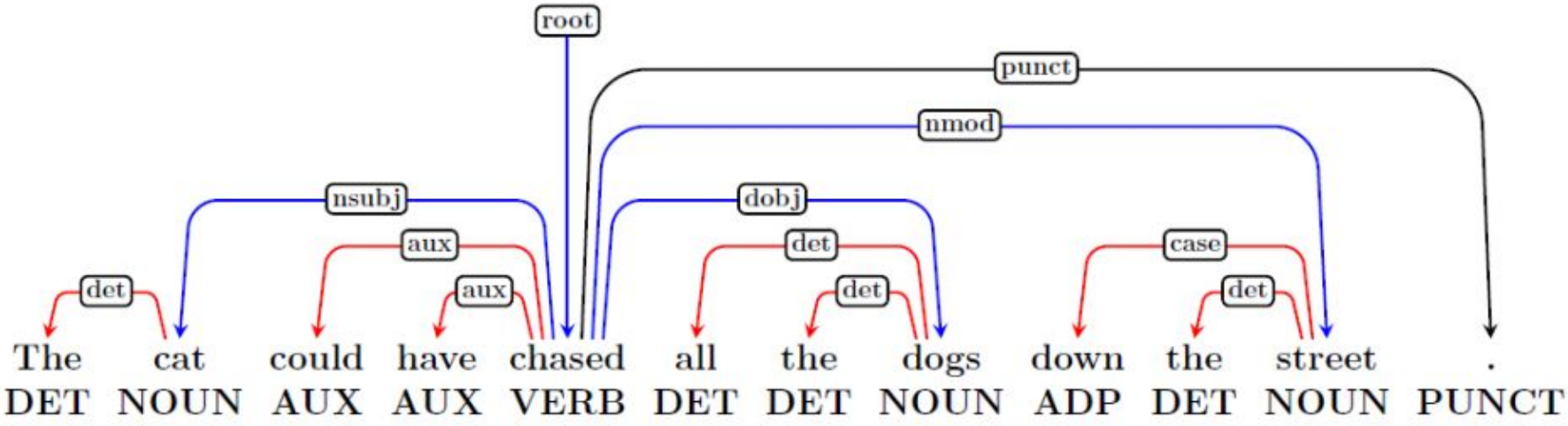
- **UNIVERSAL DEPENDENCIES (UD): annotation**



1. Content words are linked with dependencies relations
2. Function words depend on the content word they modify

ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): annotation**

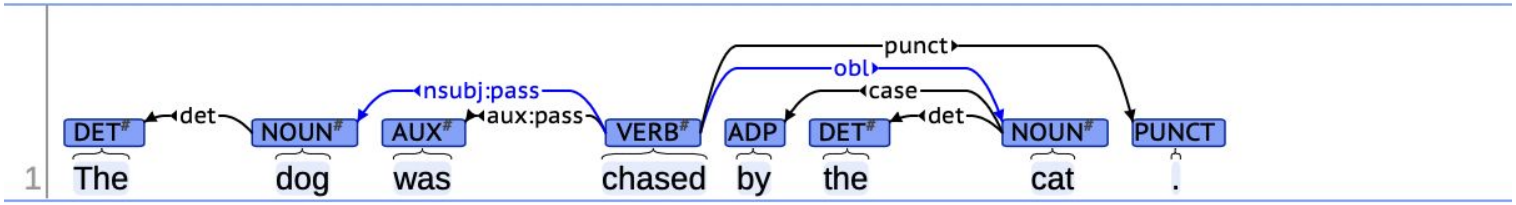


1. Content words are linked with dependencies relations
2. Function words depend on the content word they modify
3. Punctuation is attached to content words and can never has dependents

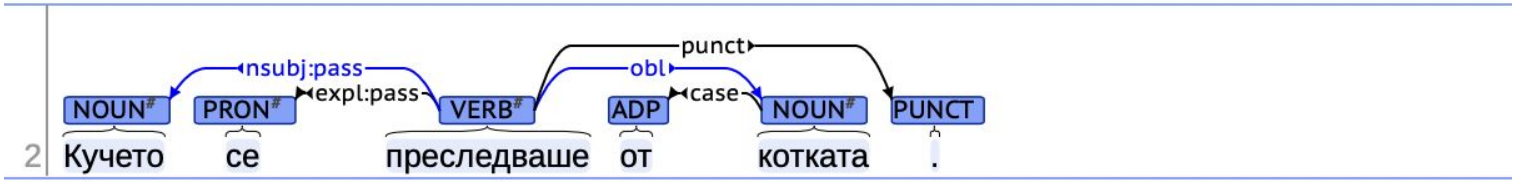
ANNOTATION: EXAMPLE

- **UNIVERSAL DEPENDENCIES (UD): annotation**

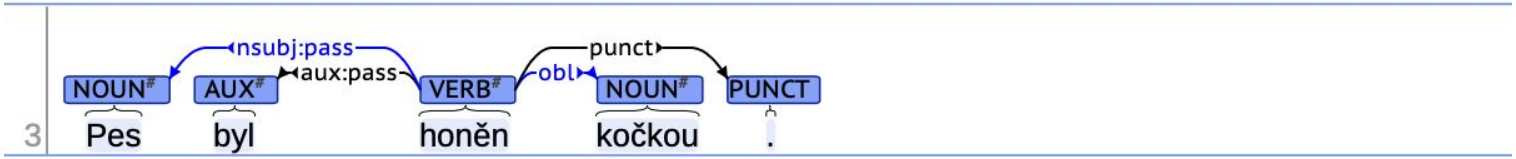
English



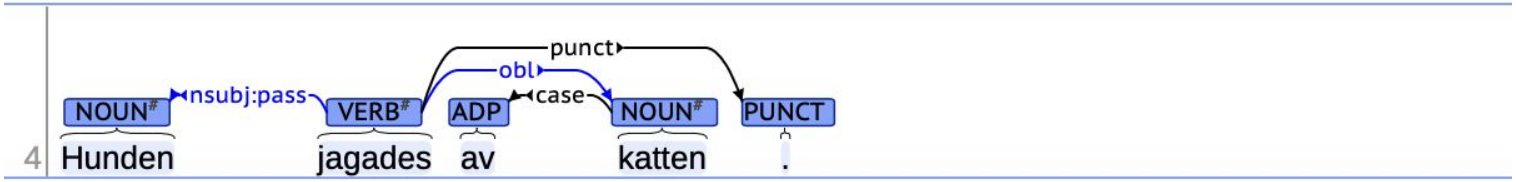
Bulgarian



Czech



Swedish



ANNOTATION: FORMATS

- CoNLL-U

```
# sent_id = DVE-124
# text = Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l' ora del vesper ciò
fu del mes d' ochiover.
# citation_hierarchy = Liber_Primus,xi,Paragraphus_5
1 Post          post          ADP   e       AdpType=Prep                                                                                  2  case         _ _
2 quos          qui           PRON  prepma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                               9  obl          _ _
3 Mediolanenses mediolanensis ADJ   Smp3a   Case=Acc|Gender=Masc|InflClass=IndEurI|NameType=Nat|Number=Plur                              9  obj           _ _
4 atque         atque         CCONJ co      Emphatic=Yes                                                                                  5  cc            _ _
5 Pergameos     pergameus     ADJ   Smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|NameType=Nat|Number=Plur                              3  conj         _ _
6-7 eorumque    _             _     _       _                                                                                             _  _            _ _
6 eorum         is            PRON  ddepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|Person=3|PronType=Prs                     8  nmod         _ _
7 que           que           CCONJ co9     Clitic=Yes                                                                                    6  cc           _ _
8 finitimos     finitimus     ADJ   smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|Number=Plur                                          3  conj         _ _
9 eruncemus     erunco        VERB  va1cpp1 Aspect=Imp|InflClass=LatA|Mood=Sub|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 0  root         _ SpaceAfter=No
10 ,            ,             PUNCT Pu      _                                                                                             17 punct        _ _
11 in           in            ADP   e       AdpType=Prep                                                                                  14 case         _ _
12 quorum       qui           PRON  prepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                               14 nmod         _ _
13 etiam        etiam         ADV   co      Compound=Yes                                                                                  12 advmod:emph  _ _
14 improperium  improperium   NOUN  sns2a   Case=Acc|Gender=Neut|InflClass=IndEurO|Number=Sing                                          17 obl          _ _
15 quendam      quidam        DET   dinsma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Sing|PronType=Ind                             16 nsubj        _ _
16 cecinisse    cano          VERB  va3fr   Aspect=Perf|InflClass=LatX|InflClass[noun]=Ind|Tense=Past|VerbForm=Inf|Voice=Act          17 ccomp        _ _
17 recolimus    recolo        VERB  va3ipp1 Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 3  acl:relcl    _ _
18 Enter        enter         X     zi      Foreign=Yes                                                                                   16 obj          _ _
19 l'           l             X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
20 ora          ora           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
21 del          del           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
22 vesper       uesper        X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
23 ciò          cio           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
24 fu           fu            X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
25 del          del           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
26 mes          mes           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
27 d'           d             X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
28 ochiover     ochiouer      X     zi      Foreign=Yes                                                                                   18 flat:foreign _ SpaceAfter=No
29 .            .             PUNCT Pu      _                                                                                             9  punct        _ _
```

ANNOTATION: FORMATS

- IOB

Words	IOB Label
American	B-ORG
Airlines	I-ORG
,	O
a	O
unit	O
of	O
AMR	B-ORG
Corp.	I-ORG
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B-PER
Wagner	I-PER
said	O
.	O

Tags	Description
B-PER	The beginning of a Person's name
I-PER	Part of a person's name
B-LOC	The beginning of a Location name
I-LOC	Part of a Location name
B-ORG	The beginning of a Organization name
I-ORG	Part of a Organization name
O	Not named-entity

ANNOTATION: FORMATS

- XML stand-off

TEXT

```
<?xml version='1.0' encoding='UTF-8'?>
<TextCorpus xmlns="http://www.dspin.de/data/textcorpus" lang="en">
IN THE VAL CAMONICA.
</tc:text>
```

TOKENS

```
<tc:tokens xmlns:tc="http://www.dspin.de/data/textcorpus">
      <tc:token ID="t_0">IN</tc:token>
      <tc:token ID="t_1">THE</tc:token>
      <tc:token ID="t_2">VAL</tc:token>
      <tc:token ID="t_3">CAMONICA</tc:token>
      <tc:token ID="t_4">.</tc:token>
</tc:tokens>
```

POS TAGS

```
<tc:POStags xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="penntb">
      <tc:tag tokenIDs="t_0">IN</tc:tag>
      <tc:tag tokenIDs="t_1">DT</tc:tag>
      <tc:tag tokenIDs="t_2">NNP</tc:tag>
      <tc:tag tokenIDs="t_3">NNP</tc:tag>
      <tc:tag tokenIDs="t_4">.</tc:tag>
</tc:POStags>
```

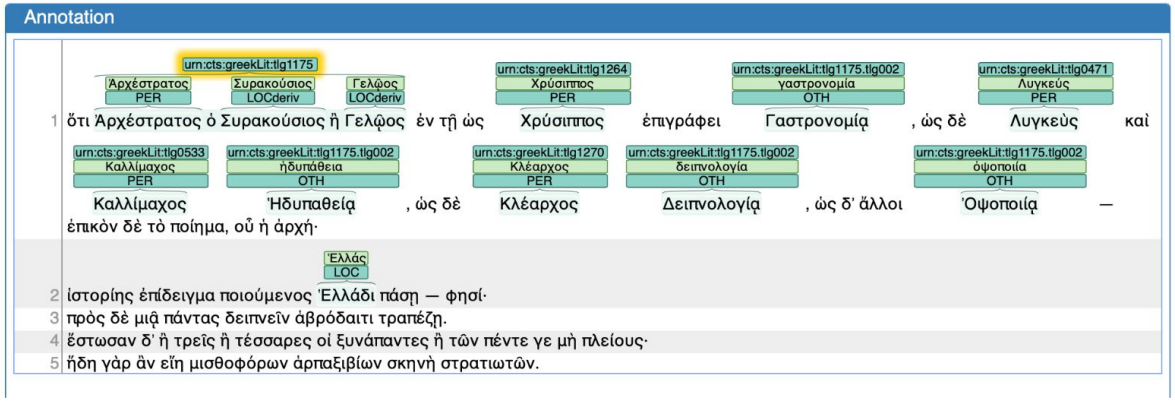
ANNOTATION: FORMATS

- XML stand-off

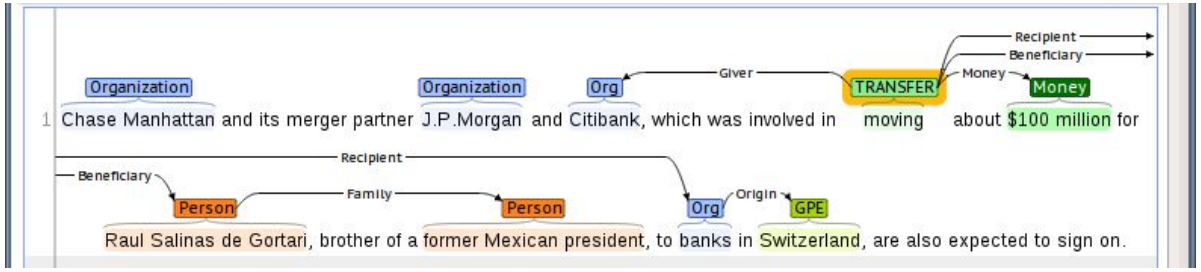
```
<?xml version='1.0' encoding='UTF-8'?>
<TextCorpus xmlns="http://www.dspin.de/data/textcorpus" lang="en">
IN THE VAL CAMONICA.
</tc:text>


<tc:tokens xmlns:tc="http://www.dspin.de/data/textcorpus">
      <tc:token ID="t_0">IN</tc:token>
      <tc:token ID="t_1">THE</tc:token>
      <tc:token ID="t_2">VAL</tc:token>
      <tc:token ID="t_3">CAMONICA</tc:token>
      <tc:token ID="t_4">CAMONICA</tc:token>
</tc:tokens>


<tc:POStags xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="penntb">
      <tc:tag tokenIDs="t_0">IN</tc:tag>
      <tc:tag tokenIDs="t_1">DT</tc:tag>
      <tc:tag tokenIDs="t_2">NNP</tc:tag>
      <tc:tag tokenIDs="t_3">NNP</tc:tag>
      <tc:tag tokenIDs="t_4">.</tc:tag>
</tc:POStags>
```

ANNOTATION: TOOLS

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



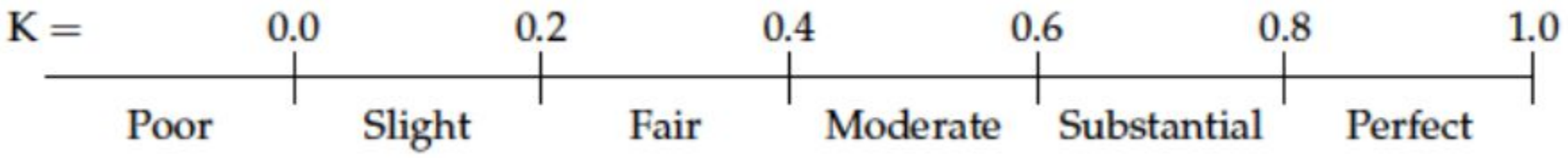
BRAT, https://brat.nlplab.org



ANNOTATION: AGREEMENT

- **Inter-Annotator Agreement (IAA) =** agreement between at least 2 annotators on the same text

  - consistency of the annotation
  - cognitive plausibility of the model
  - a broad agreement between the annotators is considered to guarantee the validity of the scheme and the high quality of annotated data
  - Cohen's Kappa (annotators = 2) o Fleiss's Kappa (annotators > 2)



ANNOTATION & MACHINE LEARNING

- **WHAT WE NEED:**
  - *training set*: annotated data for training the model (for supervised algorithms)
  - *test set*: NOT annotated data, other than training data, on which to apply the trained model
  - *gold standard*: annotated test data on which to evaluate the performance of the trained model

EVALUATION

- The evaluation of the prediction of the classifier (output) is based on manually annotated data: *gold standard*
- The simplest metric: **ACCURACY**

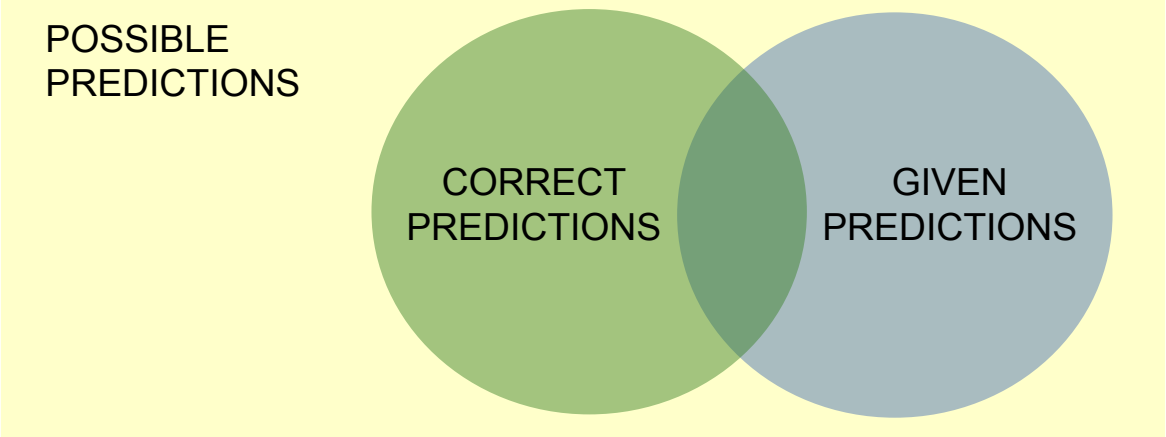
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Example:

- 150 NEs annotated in the gold standard
- 120 NEs correctly predicted
- accuracy = 120/150 = 0,8 (80%)
- it can be calculated on a general level or by class/tag

EVALUATION

- Confusion matrix

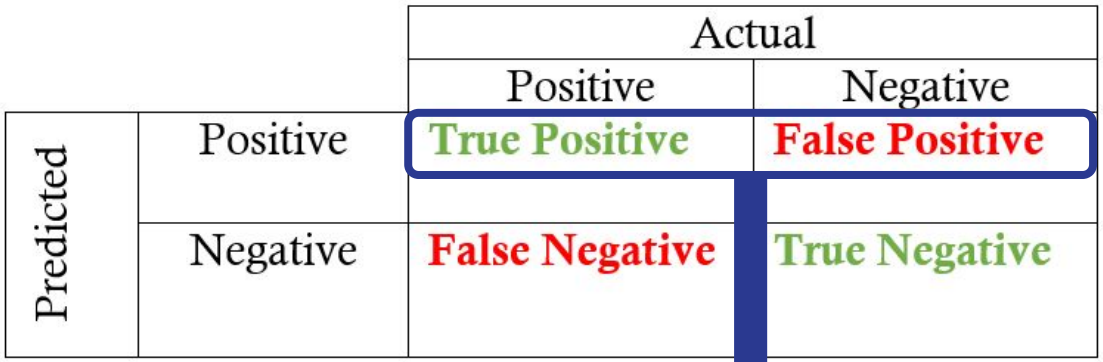


		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

EVALUATION

- **PRECISION (P)**: it measures the ratio between the elements correctly predicted by the system and the total of predicted elements

  - # correct predictions / # predictions given

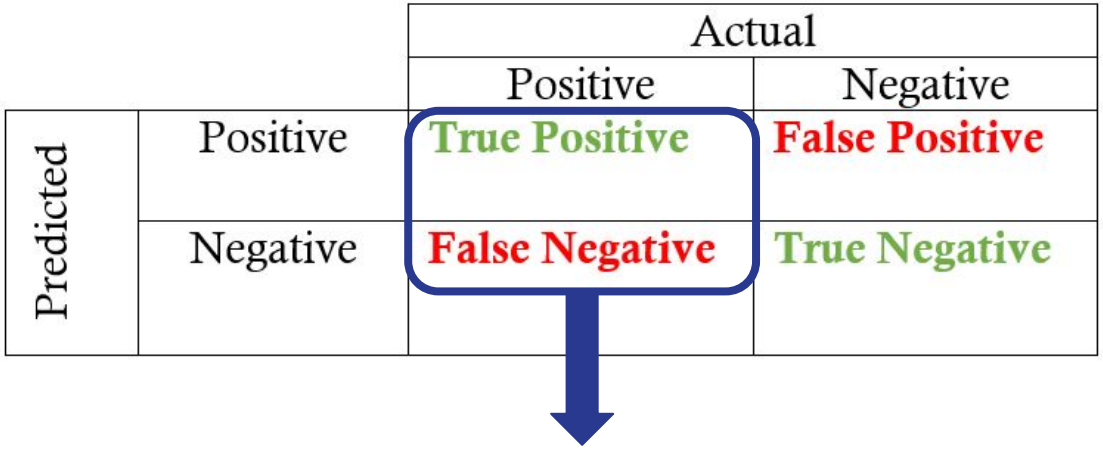


$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

EVALUATION

- **RECALL (R)**: it measures the ratio between the elements correctly predicted by the system and the total of the correct elements

- # correct predictions / # possible correct elements



$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

EVALUATION

- Sometimes there is a gap between precision and recall. As precision increases, recall often drops (and vice versa)
- **F-MEASURE**: harmonic mean between precision and recall
  - 2*precision*recall / (precision + recall)
- Alternative metric: parameterized average, which allows to choose to give more importance to P or R: when beta = 1 we speak of F-1

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P) + R}$$

β = 1: P and R they have the same weight
β > 1: R is more important
β < 1: P is more important
β = 0: Only P is taken into consideration

EVALUATION

- Example:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

EVALUATION

- Example:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

- Precision: 70 / (70+15) = 70 / 85 = 0.82

EVALUATION

- Example:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

- Precision: 70 / (70+15) = 70 / 85 = 0.82
- Recall: 70 / (70+30) = 70 / 100 = 0.70

EVALUATION

- Example:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

- Precision: 70 / (70+15) = 70 / 85 = 0.82
- Recall: 70 / (70+30) = 70 / 100 = 0.70
- F-measure: 2*0.82*0.7 / (0.82+0.70) = 0.75

PART 2
A LITTLE BIT OF PRACTICE…

WHAT WE ARE GOING TO USE

- **Data** from the ILC4CLARIN repository: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-976
    - txt_V1.zip: unzip the folder
    - create a file with a chapter of a book of your choice or use *Pisa_Italian_Days_and_Ways.txt* (https://www.gutenberg.org/files/44418/44418-h/44418-h.htm)
- **Tools** from the Language Resource Switchboard (LRS): https://switchboard.clarin.eu
    - upload a file (only a single resource can be processed)
    - check the full list of available tools

WebLicht

- *WebLicht consists of a collection of **web-based linguistic annotation tools**, distributed repositories for storing and retrieving information about the tools, and this web application, which allows you to easily create and execute **tool chains** without downloading or installing any software on your local computer*
  - pre-built chains (easy mode) or make-your-own chains (advanced mode)
  - sentence splitting, tokenisation, PoS tagging, lemmatisation, morphological analysis, parsing, NER, geolocation
  - possibility to: download the chain, the output of each module, the final output or to check the output in an interface
  - https://weblicht.sfs.uni-tuebingen.de/

UDPipe

- *UDPipe is a **trainable** pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files*
  - based on the Universal Dependencies framework
  - UDPipe v1 (C++) or UDPipe v2 (Python)
  - last models: v2.6, 91 different languages - modern (English, Russian ...), ancient (Latin, ancient Greek, Gothic ...), very widespread (Chinese, Spanish ...), not very widespread (Wolof, Uyghur), of different genres (spoken, social media ...)
  - command line interface or a web based interface: CLI is the only option if you want to train a new model
  - https://lindat.mff.cuni.cz/services/udpipe/

NameTag

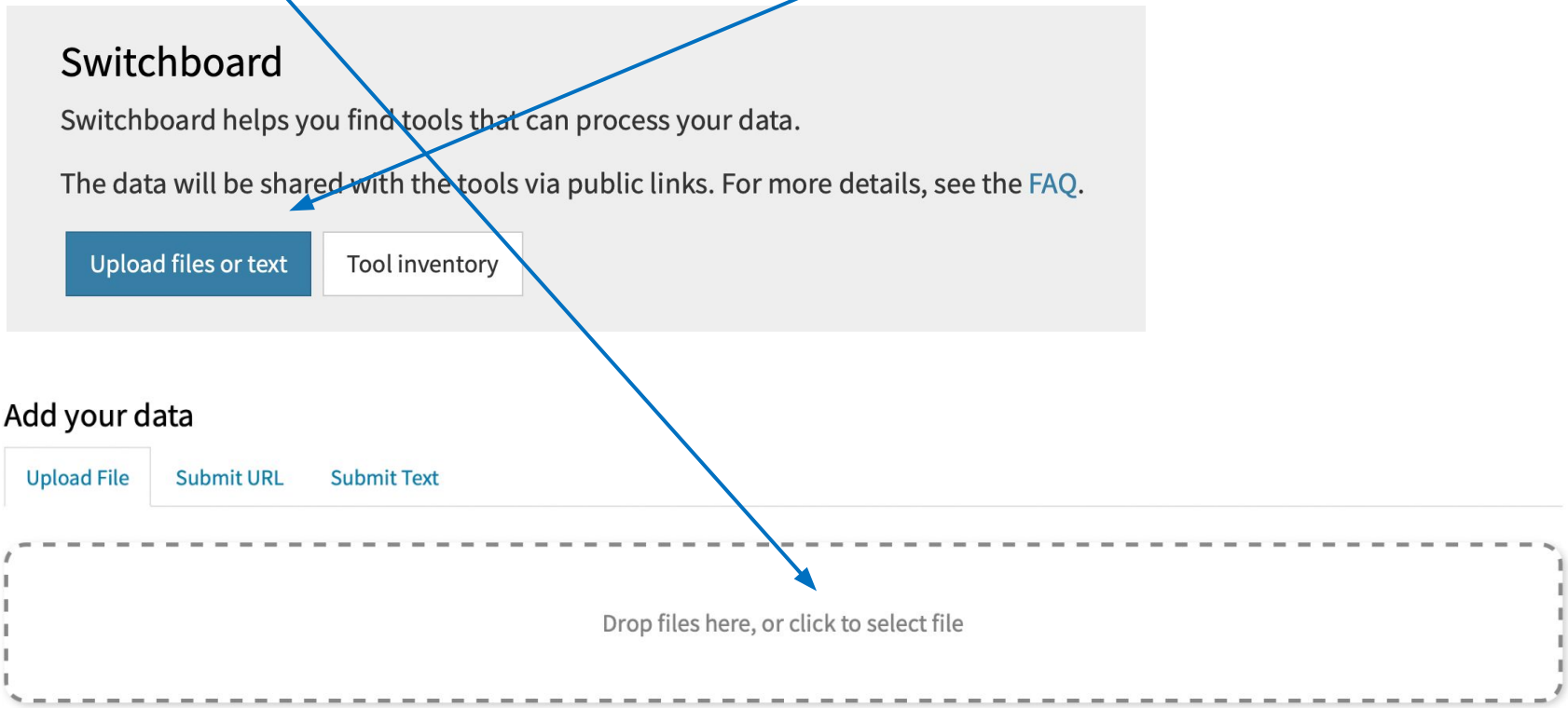
- *NameTag is an open-source tool for **named entity recognition** (NER). NameTag identifies proper names in text and classifies them into predefined categories, such as names of persons, locations, organizations, etc.*
  - trainable
  - NameTag v1 (Czech and English) and NameTag v2 (Czech, English, Spanish, German, Dutch)
  - PER, ORG, LOC and MISC
  - command line interface or a web based interface: CLI is the only option if you want to train a new model
  - http://lindat.mff.cuni.cz/services/nametag/

NLPHub

- *A distributed system that orchestrates and combines several state-of-the-art text mining services that recognize spatiotemporal events, keywords, and a large set of named entities.*
  - it merges the results of different NER tools run in parallel
  - names of persons, locations, organizations, money amounts, time and date expressions, but also keywords and events
  - English, French, Italian, Spanish and German
  - http://nlp.d4science.org/hub/

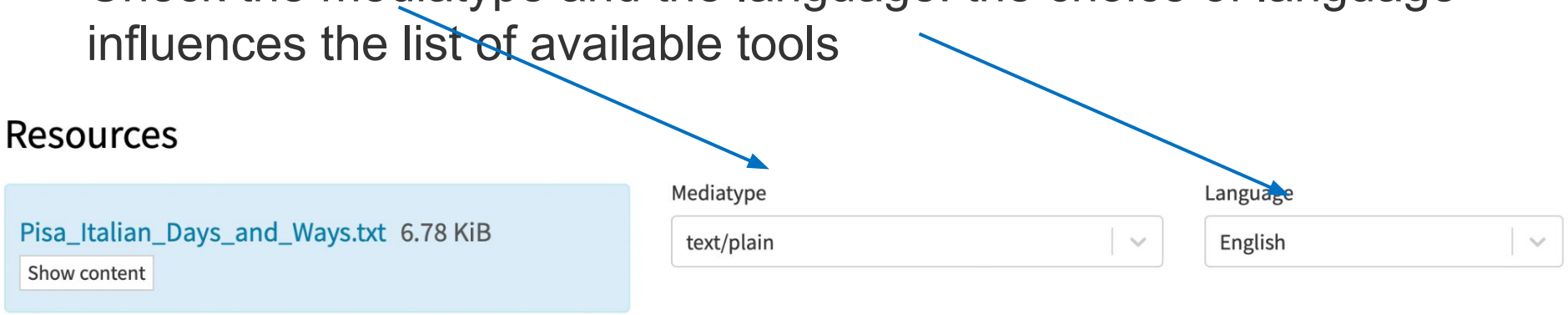
USING SWITCHBOARD

- Go to https://switchboard.clarin.eu, click on “Upload files or text”, click to select the file, select *Pisa_Italian_Days_and_Ways.txt*

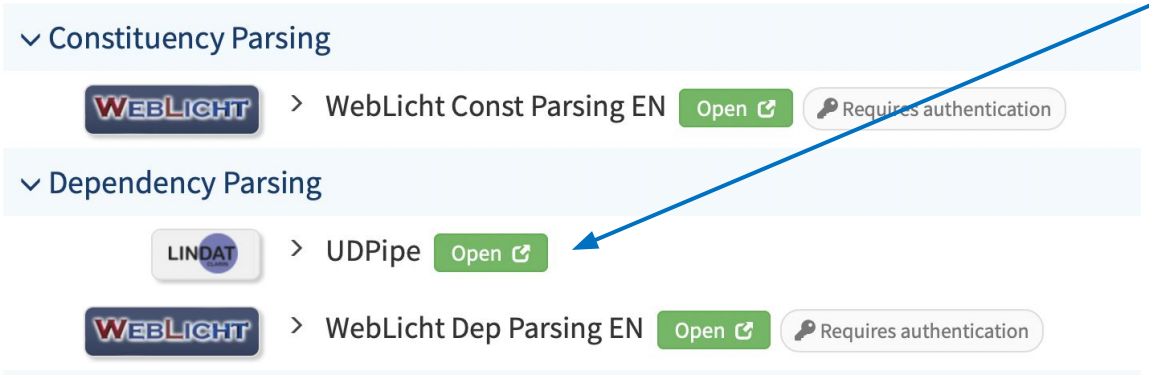


USING SWITCHBOARD

- Check the mediatype and the language: the choice of language influences the list of available tools

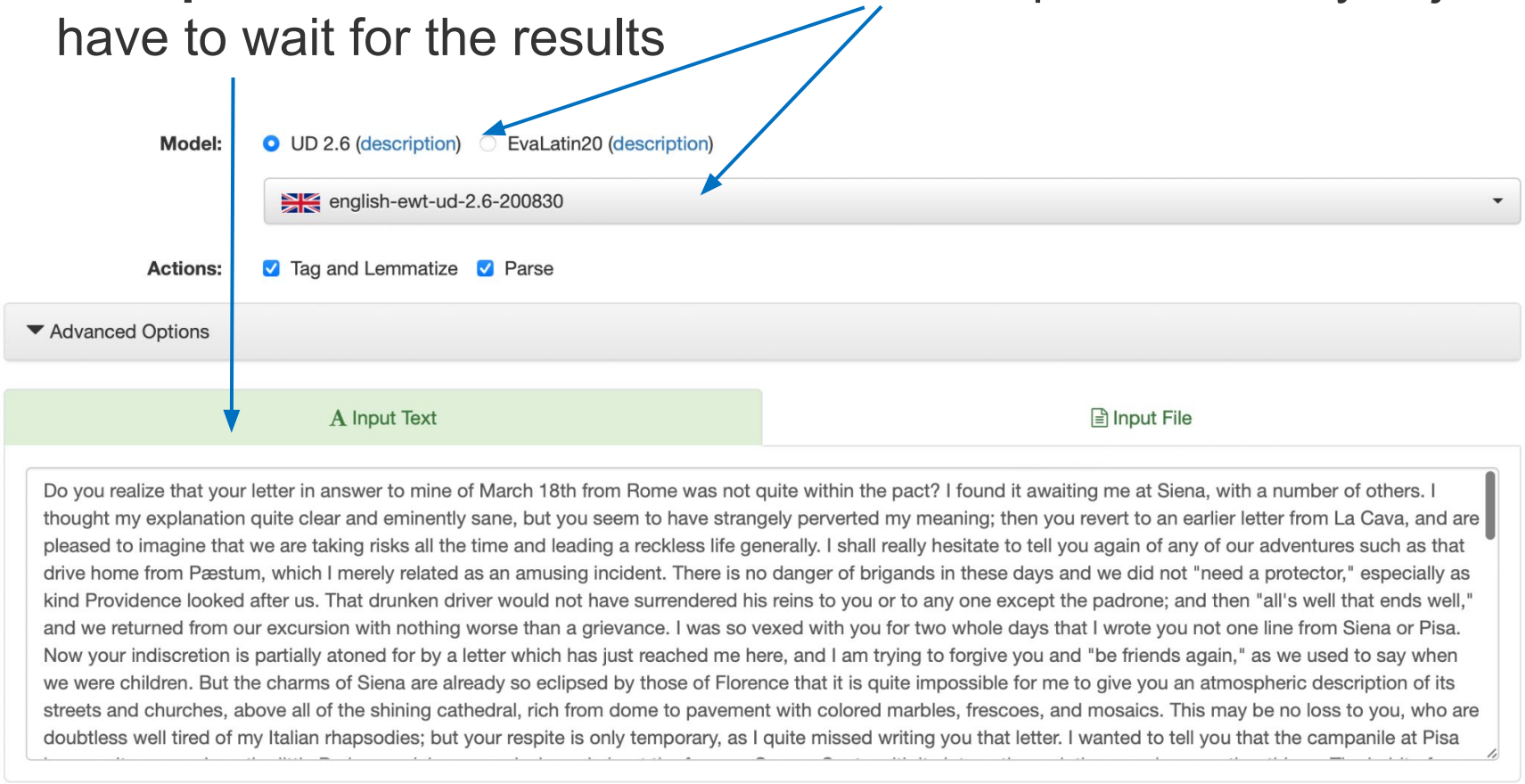


- Choose the task of interest and click on the green button “Open”



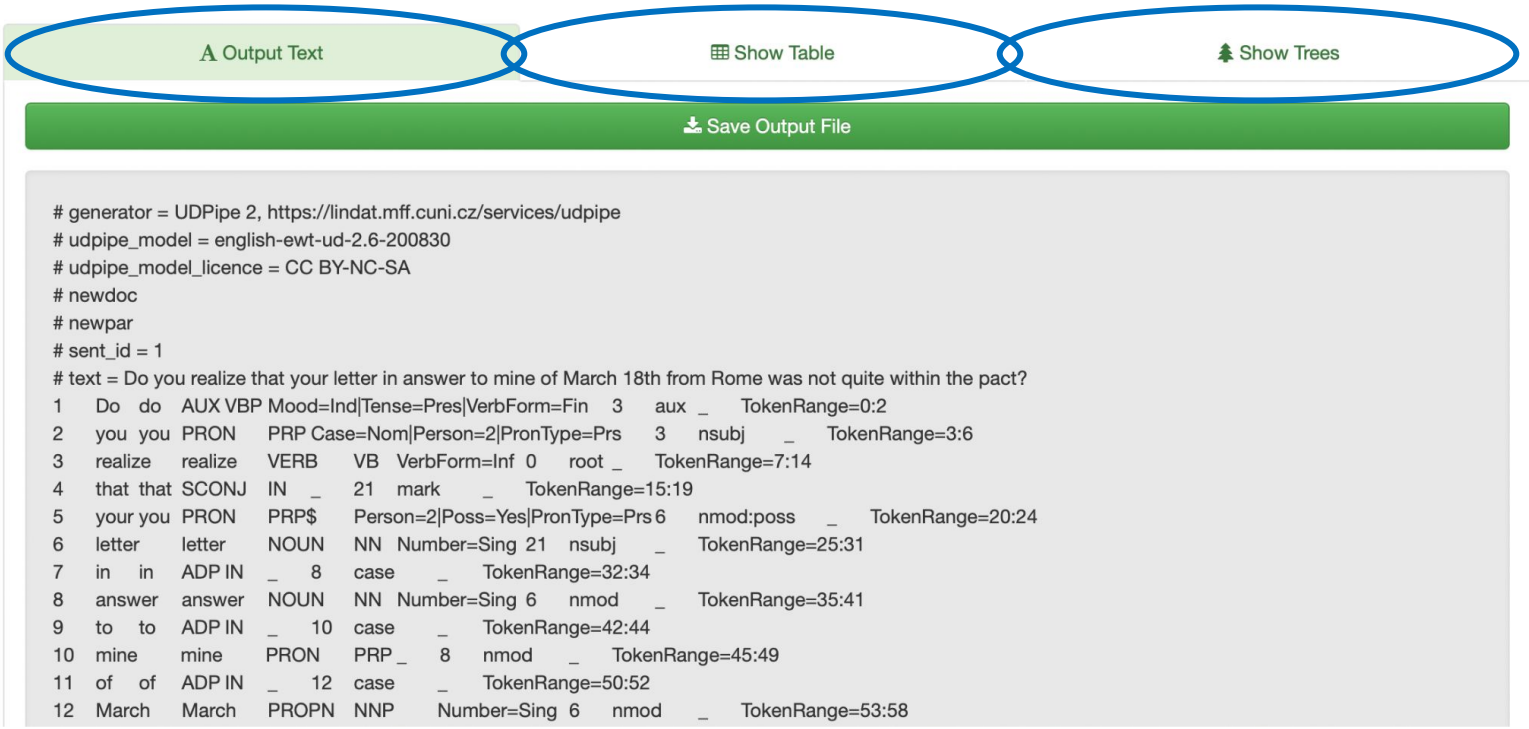
USING SWITCHBOARD

- **UDPipe**: the model is chosen, the text is uploaded and you just have to wait for the results



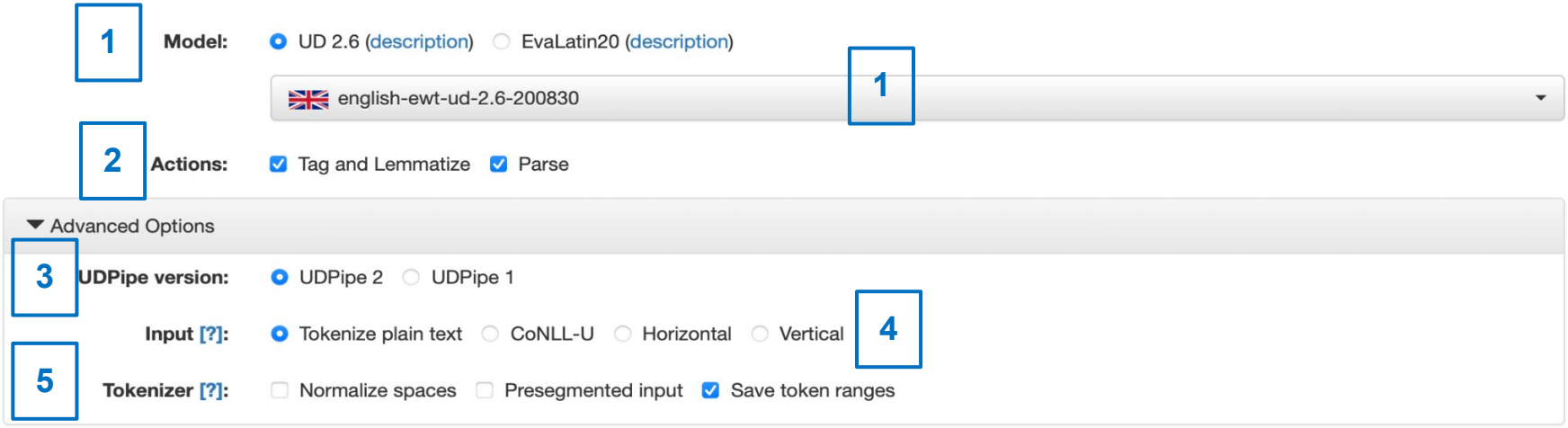
USING SWITCHBOARD

- **UDPipe**: results are displayed as a text, a table or an image (tree)
  - in the first two cases, you can download the CoNLL-U file otherwise you can save the svg file of the image.



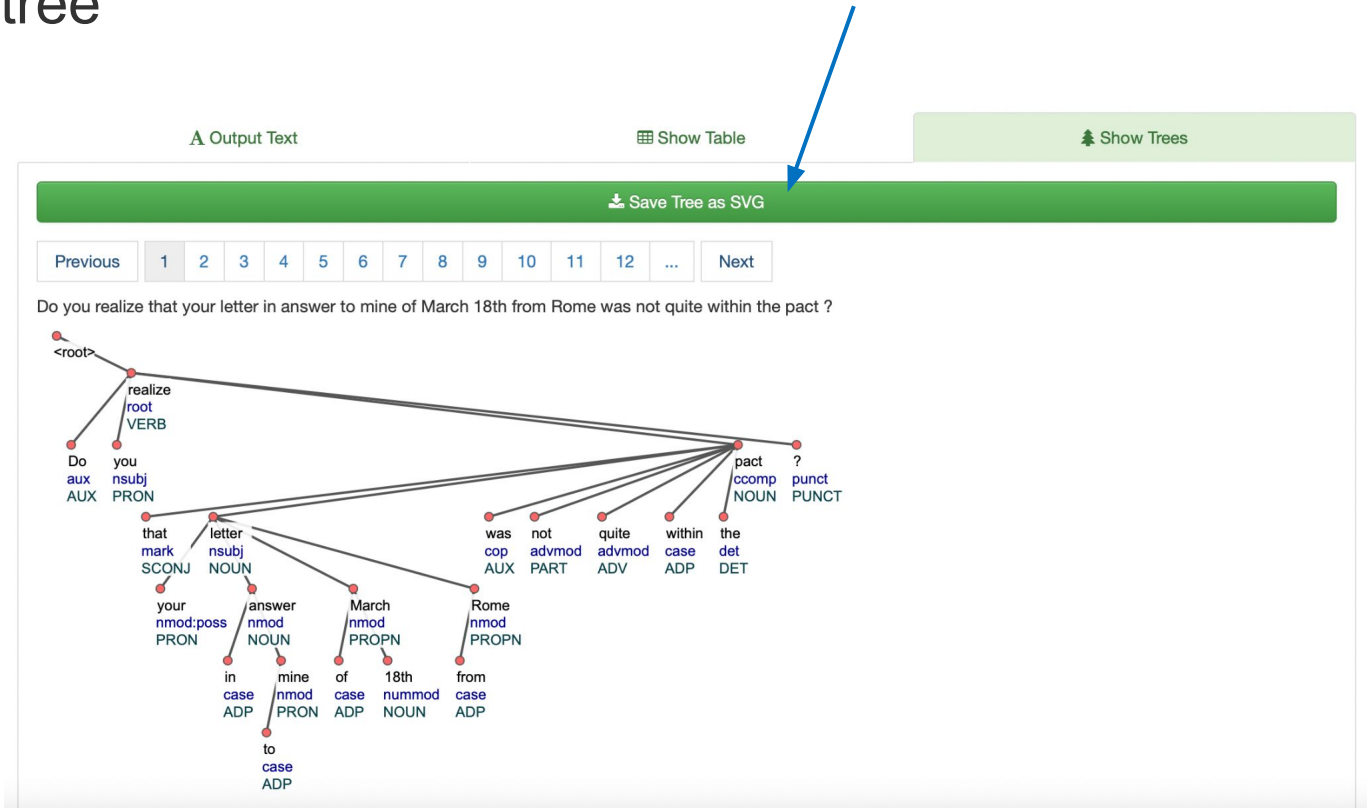
USING SWITCHBOARD

- **UDPipe**: several options can be chosen
  1. the model
  2. the tasks (with or without parsing)
  3. the tool version
  4. the format of text input: use “Horizontal” if you have text already splitted by sentences, use “Vertical” if you have text already tokenized
  5. the type of tokenization



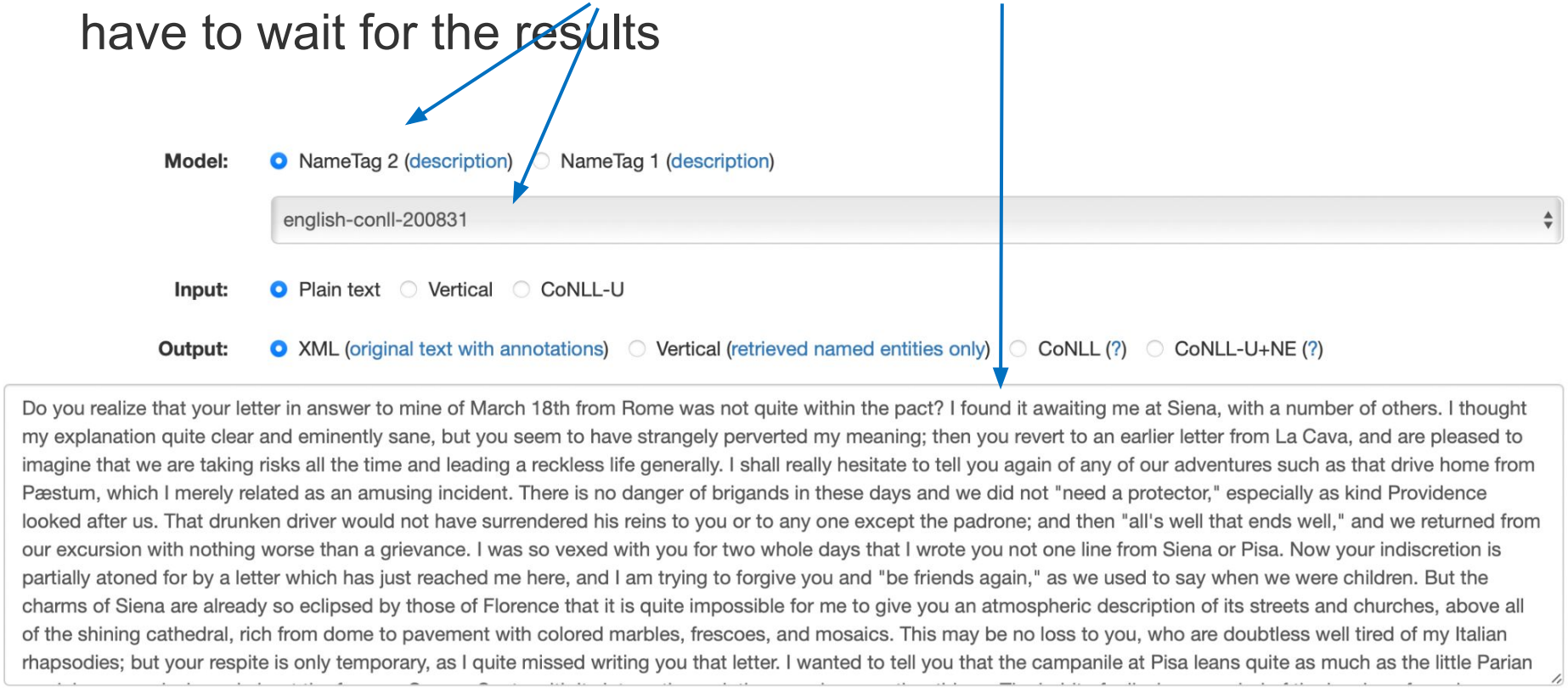
USING SWITCHBOARD

- **UDPipe**: results are displayed as a text, a table or an image (tree)
  - in the first two cases, you can download the CoNLL-U file otherwise you can save the vector image (SVG format) of the tree



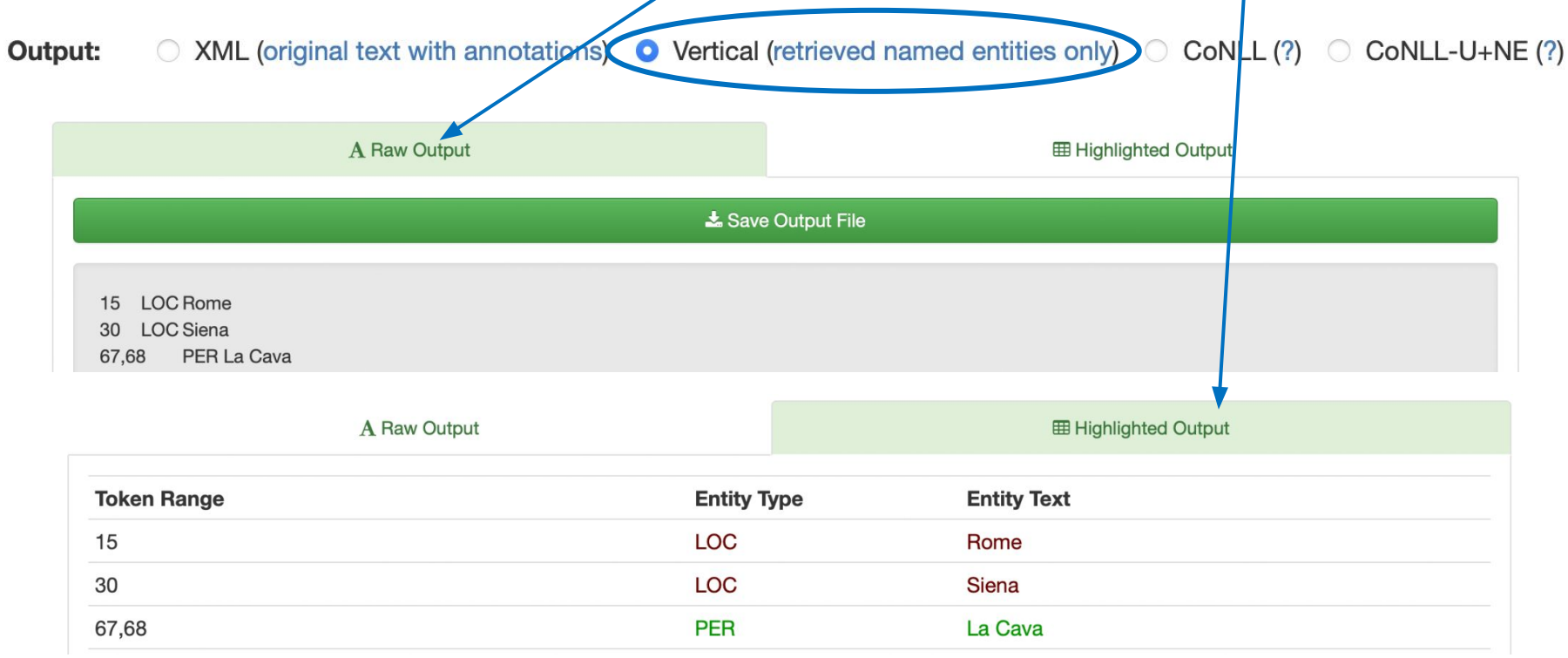
USING SWITCHBOARD

- **NameTag**: the model is chosen, the text is uploaded and you just have to wait for the results



USING SWITCHBOARD

- **NameTag**: results are displayed using a raw or highlighted output (easier to read)
  - the format of the results changes depending on the type of output selected before running the tool



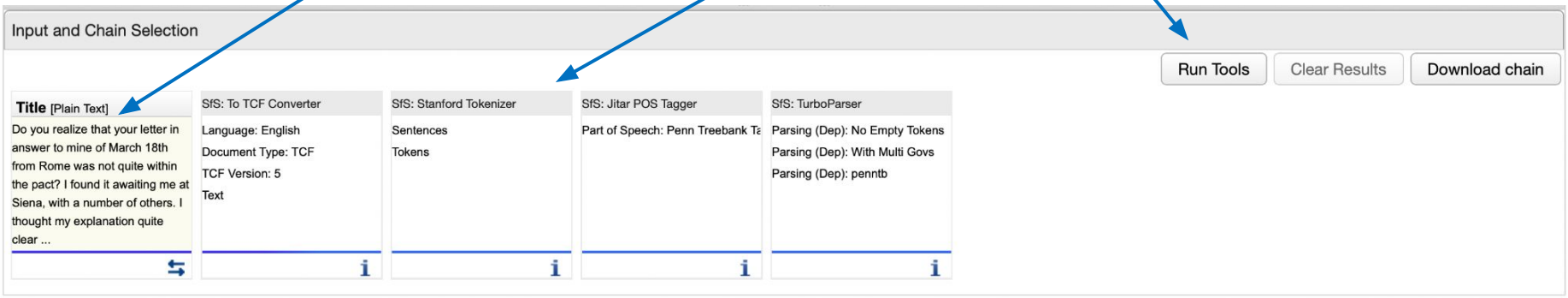
USING SWITCHBOARD

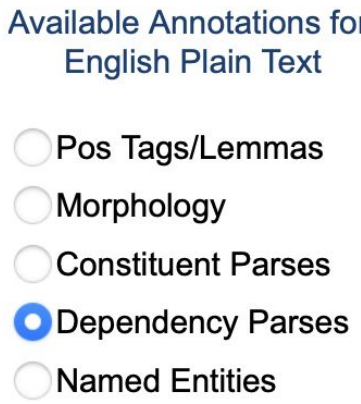
- **NameTag**: you can upload a CoNLL-U file pre-processed with UDPipe so to add the NE annotation
  - or vice versa, you can process a file with NameTag, save the output in CoNLL-U format and upload it in UDPipe to add lemmas, UPoS tags and syntactic annotation

```
# text = Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact?
1   Do      do      AUX     VBP     Mood=Ind|Tense=Pres|VerbForm=Fin    3   aux     _   TokenRange=0:2
2   you     you     PRON    PRP     Case=Nom|Person=2|PronType=Prs      3   nsubj   _   TokenRange=3:6
3   realize realize VERB    VB      VerbForm=Inf    0   root    _   TokenRange=7:14
4   that    that    SCONJ   IN      _   21  mark    _   TokenRange=15:19
5   your    you     PRON    PRP$    Person=2|Poss=Yes|PronType=Prs  6   nmod:poss   _   TokenRange=20:24
6   letter  letter  NOUN    NN      Number=Sing 21  nsubj   _   TokenRange=25:31
7   in      in      ADP     IN      _   8   case    _   TokenRange=32:34
8   answer  answer  NOUN    NN      Number=Sing 6   nmod    _   TokenRange=35:41
9   to      to      ADP     IN      _   10  case    _   TokenRange=42:44
10  mine    mine    PRON    PRP     _   8   nmod    _   TokenRange=45:49
11  of      of      ADP     IN      _   12  case    _   TokenRange=50:52
12  March   March   PROPN   NNP     Number=Sing 6   nmod    _   TokenRange=53:58
13  18th    18th    NOUN    NN      Number=Sing 12  nummod  _   TokenRange=59:63
14  from    from    ADP     IN      _   15  case    _   TokenRange=64:68
15  Rome    Rome    PROPN   NNP     Number=Sing 6   nmod    _   TokenRange=69:73|NE=LOC_1
16  was     be      AUX     VBD     Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   21  cop     _   TokenRange=74:77
17  not     not     PART    RB      _   21  advmod  _   TokenRange=78:81
18  quite   quite   ADV     RB      _   21  advmod  _   TokenRange=82:87
19  within  within  ADP     IN      _   21  case    _   TokenRange=88:94
20  the     the     DET     DT      Definite=Def|PronType=Art   21  det     _   TokenRange=95:98
21  pact    pact    NOUN    NN      Number=Sing 3   ccomp   _   SpaceAfter=No|TokenRange=99:103
22  ?       ?       PUNCT   .       _   3   punct   _   TokenRange=103:104
```

USING SWITCHBOARD

- **WebLicht**: the text is uploaded, the chain is chosen (easy mode), and you just have click on the “Run Tools” button and wait

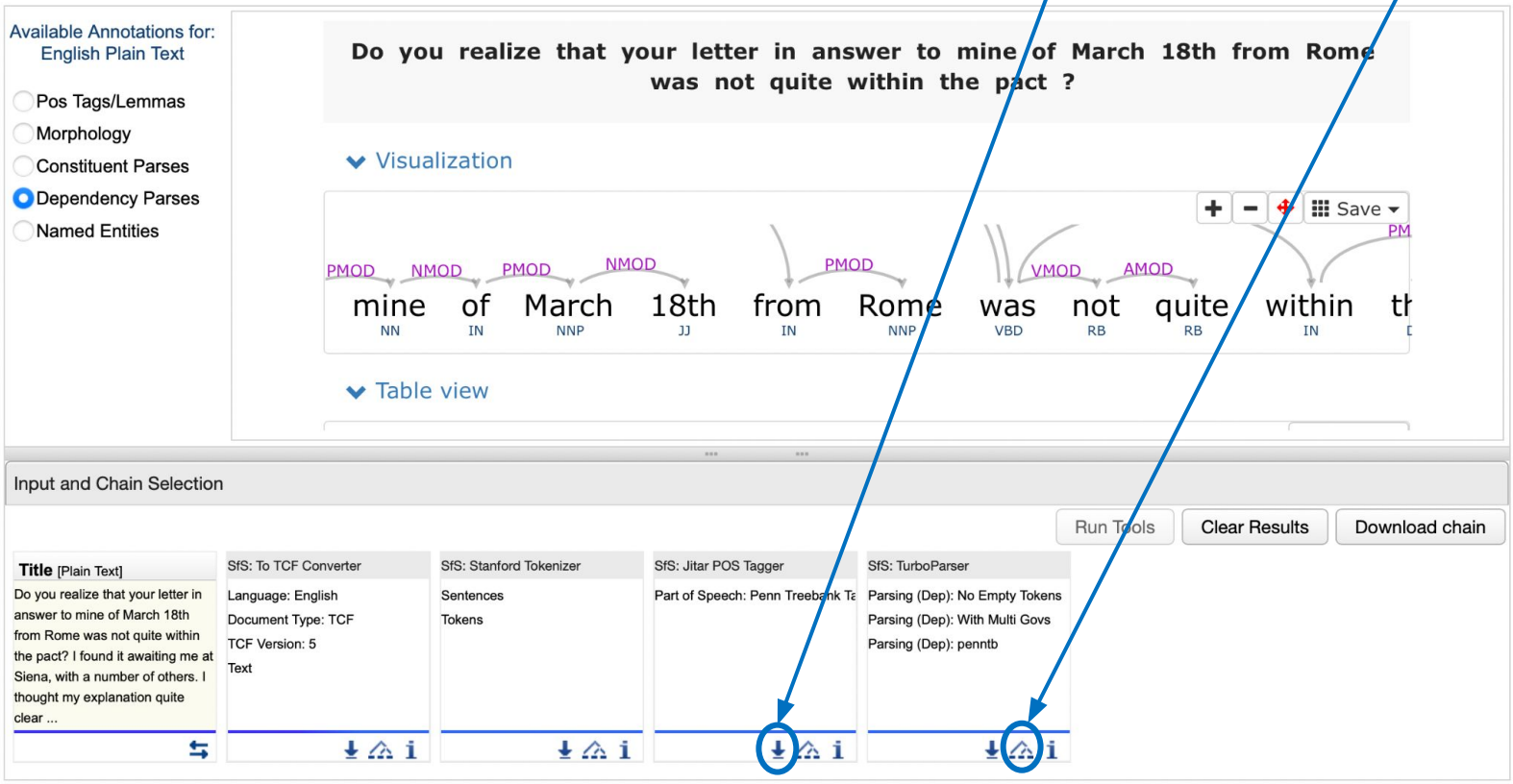




- Some chains are ready to be used, just select another task and to change the chain

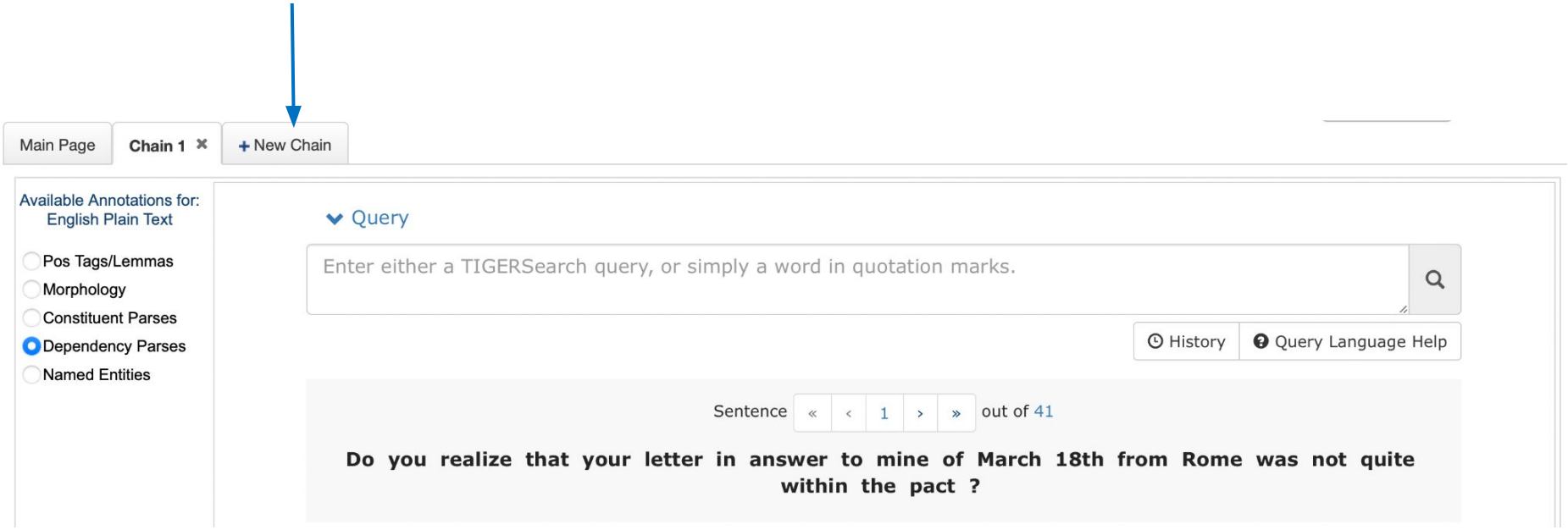
USING SWITCHBOARD

- **WebLicht**: after running the tools it is possible to save (XML stand-off or CoNLL-U format depending on the module) or visualize the results in an integrated interface (TüNDRA)



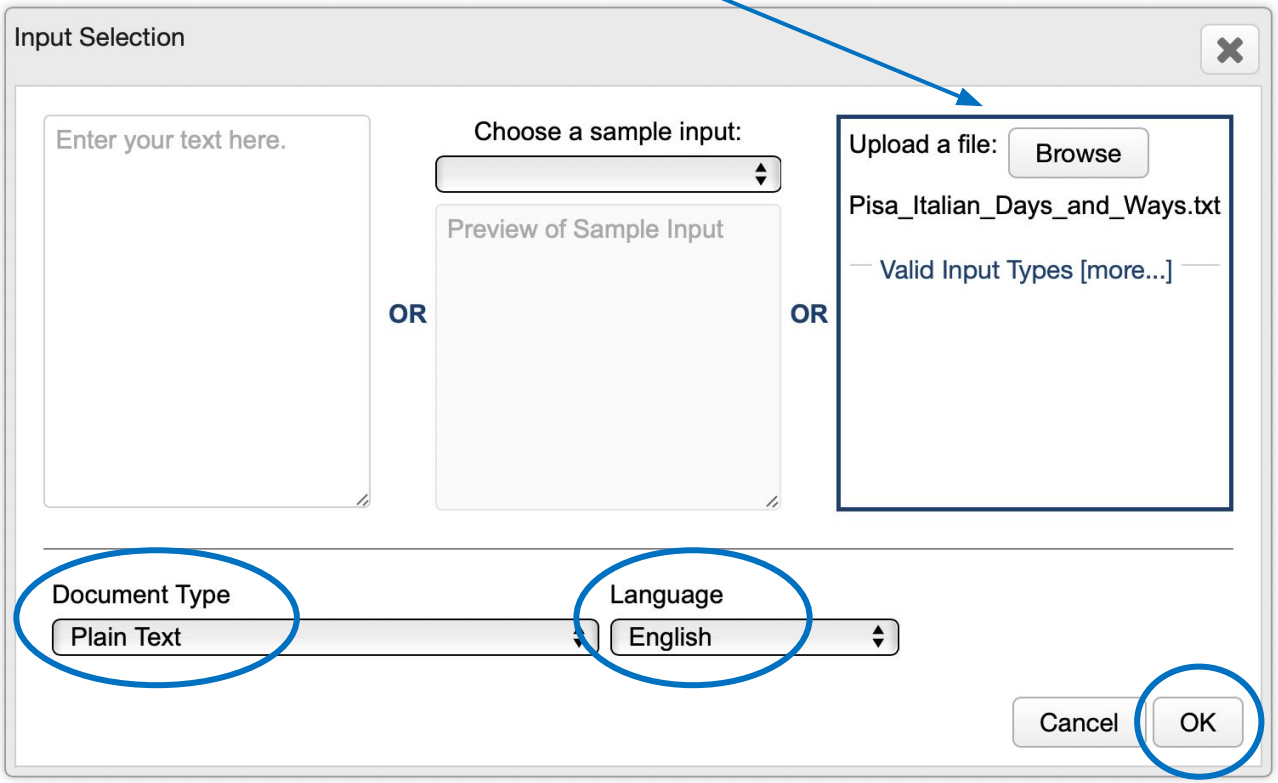
USING SWITCHBOARD

- **WebLicht**: it is possible to create new chains by clicking on the tab "New Chains"



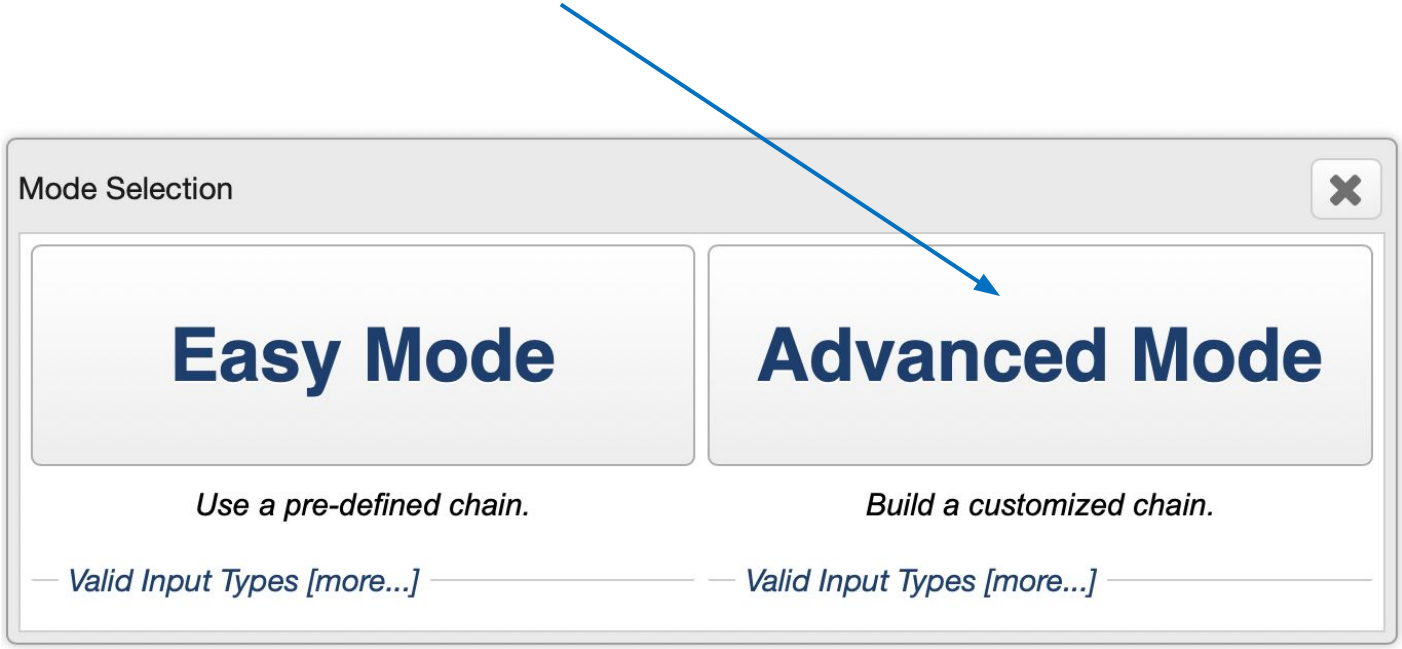
USING SWITCHBOARD

- **WebLicht**: upload the file, choose the document type (Plain Text) and the language (English), click on “OK”



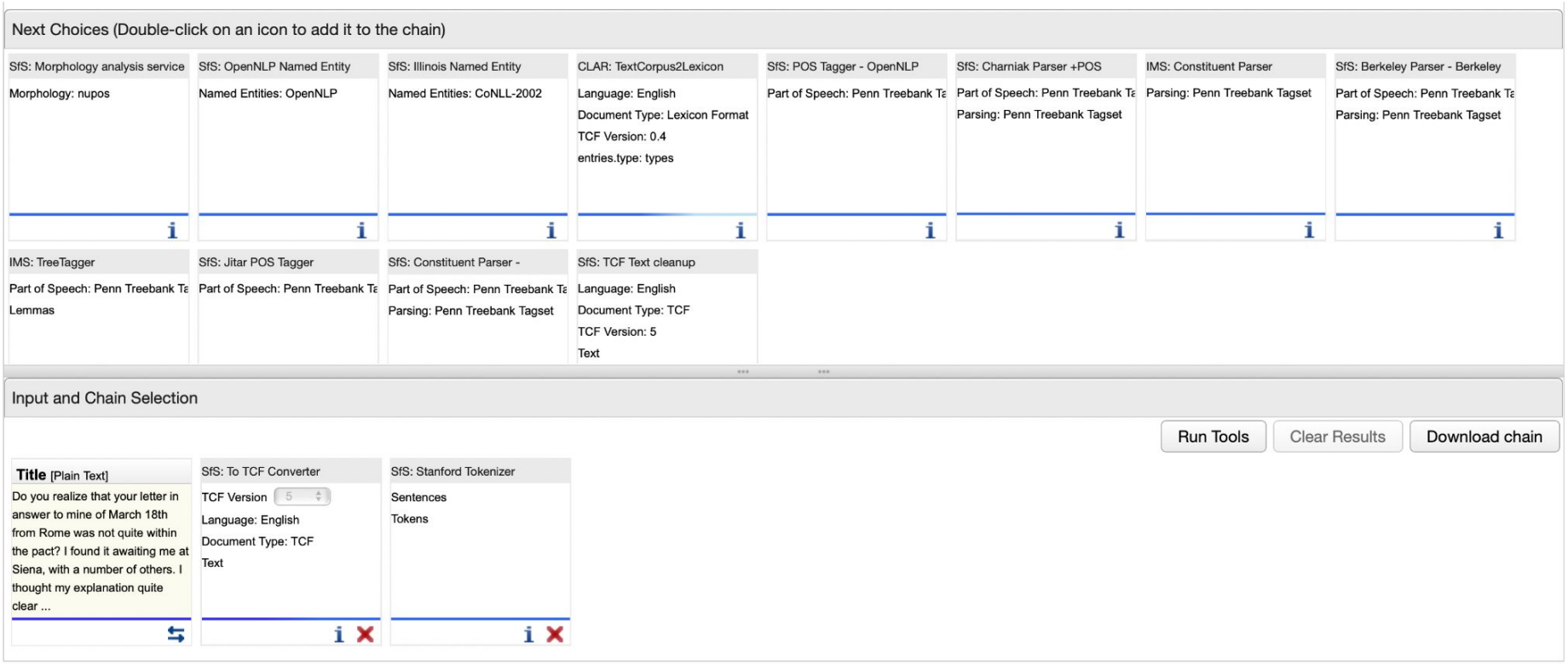
USING SWITCHBOARD

- **WebLicht**: click on “Advanced Mode”



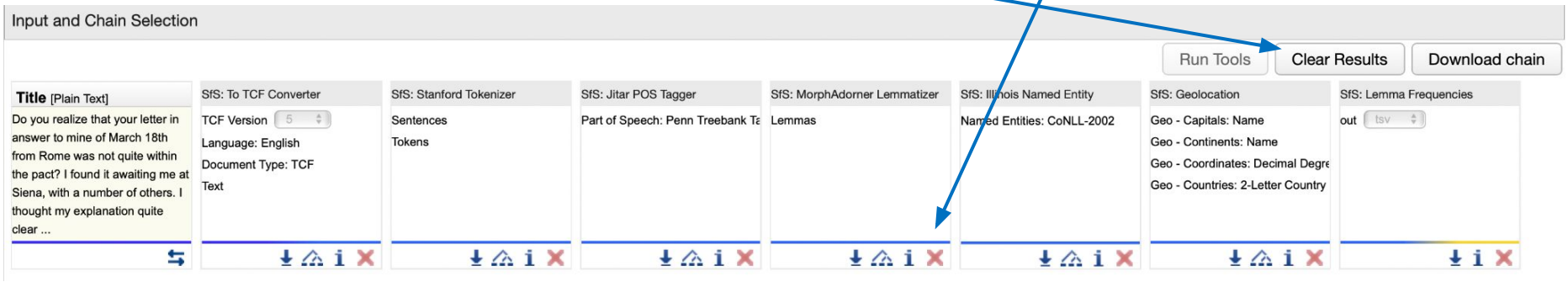
USING SWITCHBOARD

- **WebLicht**: double-click on the module, each choice has an impact on the modules available in the next step



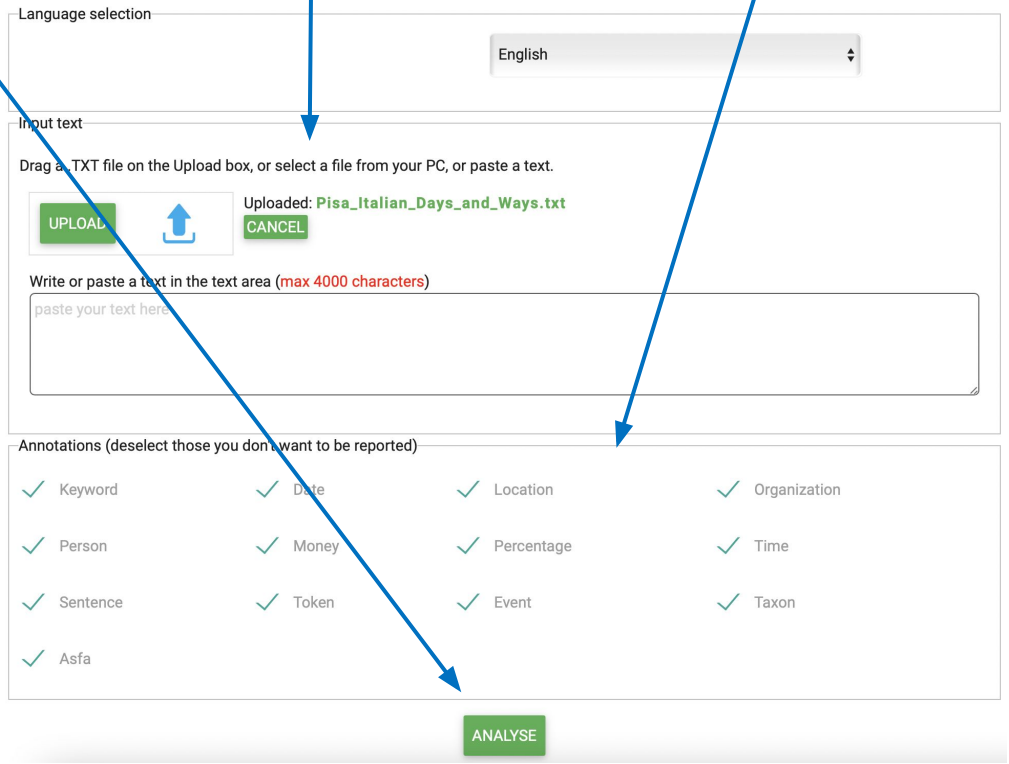
USING SWITCHBOARD

- **WebLicht**: example of chain below
  - if lemmatization is performed, the module “Lemma Frequencies” is available
  - if NER is performed, the module “Geolocation” is available: it only does the georeferencing of capitals and continents
  - if you want to change the chain after running the tools, click on “Clear Results” and then click on the red “x” for the module you want to remove



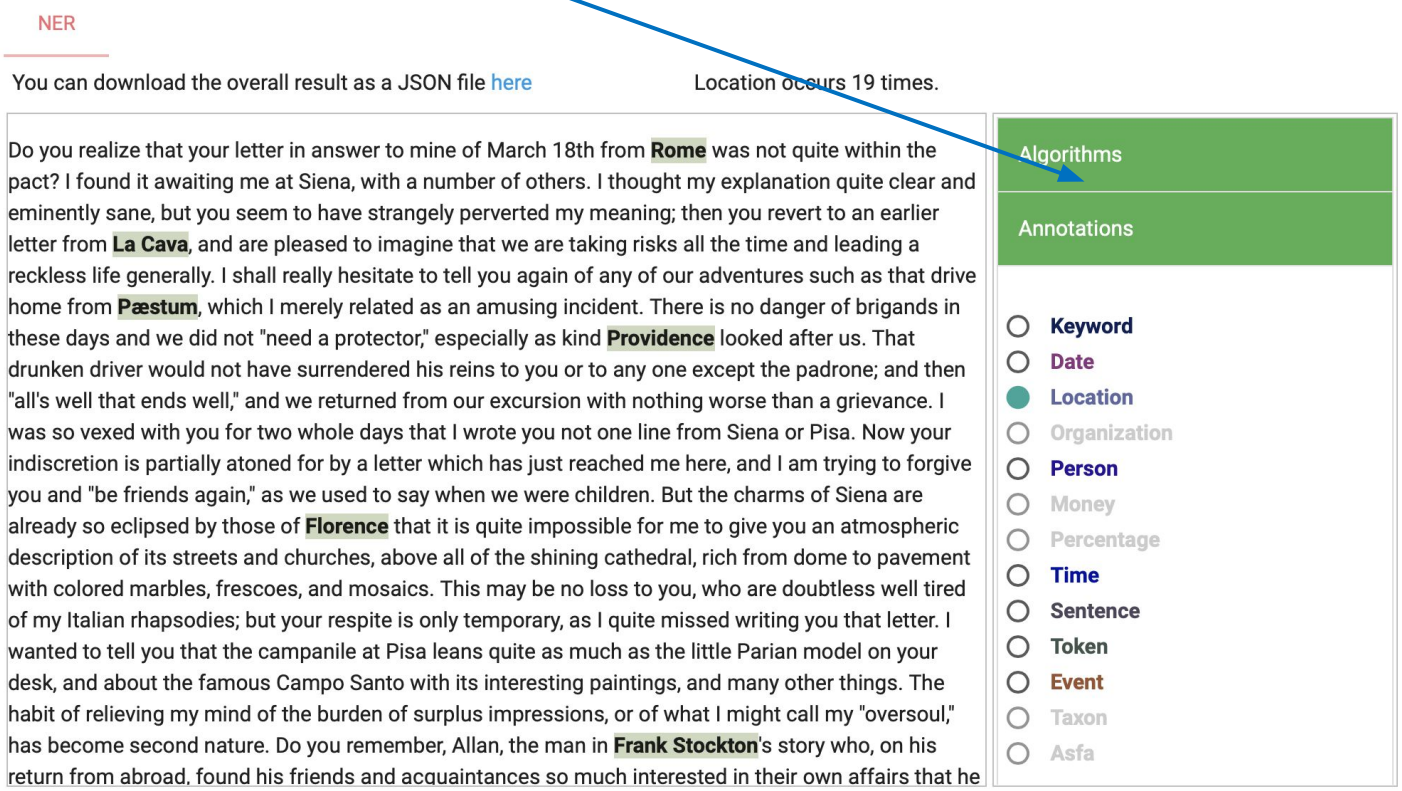
USING SWITCHBOARD

- **NLPHub**: the text is uploaded, some annotations are already selected but the selection can be changed -- sometimes an error occurs and the text should be uploaded or pasted in the text area -- click “Analyse” to run the tool



USING SWITCHBOARD

- **NLPHub**: the results are highlighted in a web interface, you can select one type of annotation at a time or download the output in JSON format



FROM NLP TO VISUALIZATION

1. Process *Pisa_Italian_Days_and_Ways.txt* with UDPipe v2 and then with NameTag 2
   - Open the file with a text editor
   - Copy and paste the text in the text area of UDPipe (https://lindat.mff.cuni.cz/services/udpipe/), choose a model and click on “Process Input”
   - Click on “Save Output File” to download the CoNLL-U file
   - Go to NameTag (http://lindat.mff.cuni.cz/services/nametag/), choose the model (english-conll-200831), select CoNLL-U as input format and vertical as output format, paste the output of UDPipe in the text area and click on “Process Input”

FROM NLP TO VISUALIZATION

2. Extraction of LOC from the output

- Download the results of NameTag by clicking on “Save Output File” (“Raw Output” should be selected): the txt file has three columns separated with tabs (you can check the structure of the file by opening it with a text editor)
- Open the output file with a spreadsheet editor (Excel, Numbers, Google sheets, but my favorite choice is LibreOffice Calc!): set tabs as delimiters
- Filter the columns so to have only LOC

FROM NLP TO VISUALIZATION

3. Use of DARIAH-DE Geo-Browser

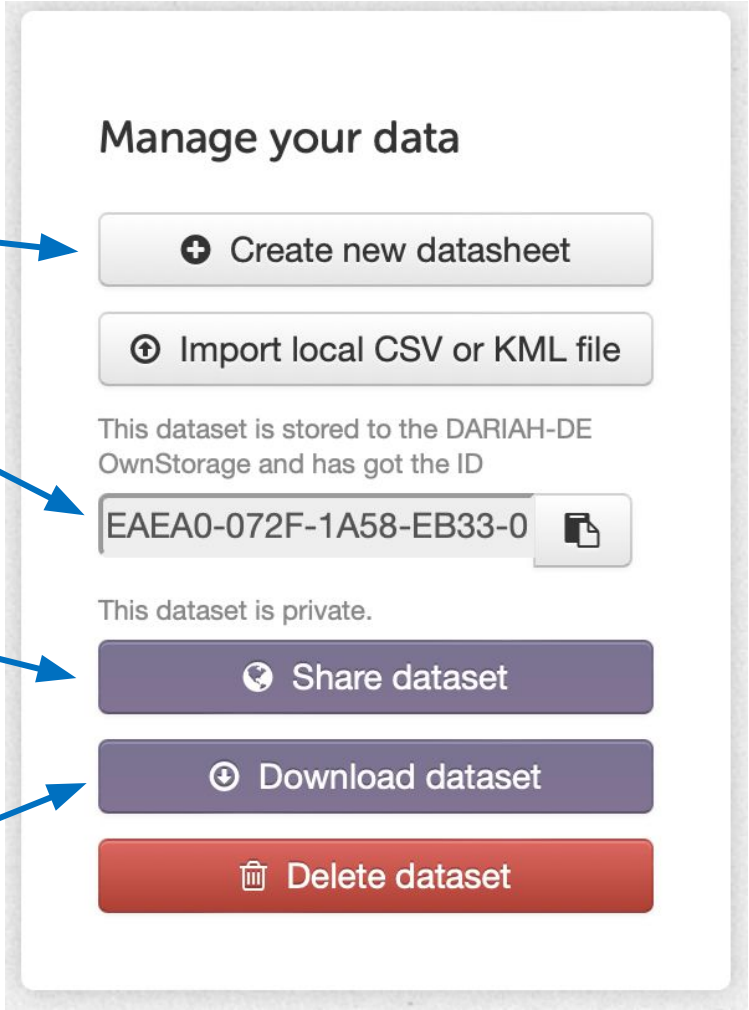
2 free services offered by the German branch of DARIAH (Digital Research Infrastructure for the Arts and Humanities):

- Browser: https://geobrowser.de.dariah.eu/index.html
- Editor: https://geobrowser.de.dariah.eu/edit/index.html → login with your institutional account or with a CLARIN-ERIC account
- Documentation: https://geobrowser.de.dariah.eu/doc/index.html

FROM NLP TO VISUALIZATION

4. Geolocation of LOC using DARIAH-DE Datasheet Editor

- Click on "Create a new datasheet" on the right of the interface
- Each new dataset is associated to an ID to be copied to return to later
- The dataset is private but can be made publicly visible: modification and cancellation remain the right of whoever created the dataset
- The dataset can be downloaded in csv format



FROM NLP TO VISUALIZATION

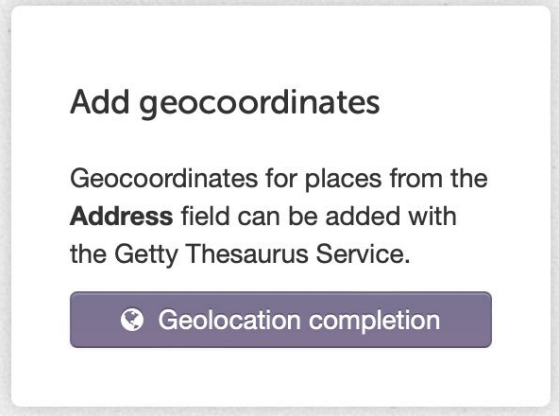
4. Geolocation of LOC using DARIAH-DE Datasheet Editor
   - the spreadsheet can be filled in directly on the browser
   - the fundamental column is Address, it cannot be empty: we paste the list of places extracted from the output of NameTag
   - alternatively, it is possible to load a csv but it must have the same required columns

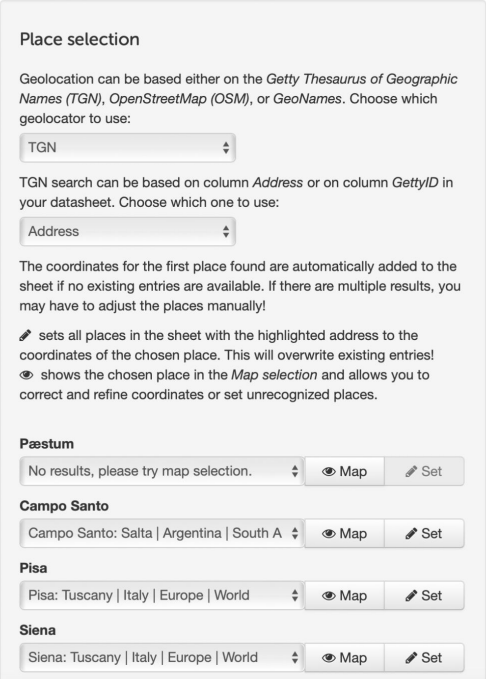
	A	B	C	D	E	F	G	H	I	J
1	**Name**	**Address**	**Description**	**Longitude**	**Latitude**	**TimeStamp**	**TimeSpan:begin**	**TimeSpan:end**	**GettyID**	
2		Rome								
3		Siena								
4		Pæstum								
5		Siena								
6		Pisa								

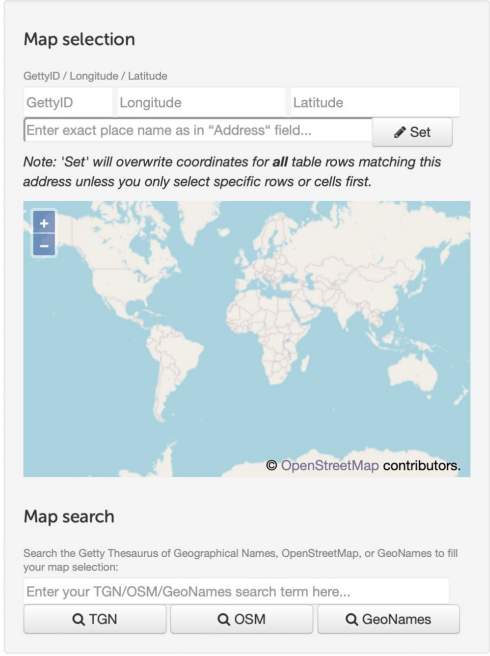
FROM NLP TO VISUALIZATION

5. Correction of geolocation

- if latitude and longitude are missing, they can be added automatically by clicking on "Geolocation completion"
- each entry can be checked in the "Place selection" section



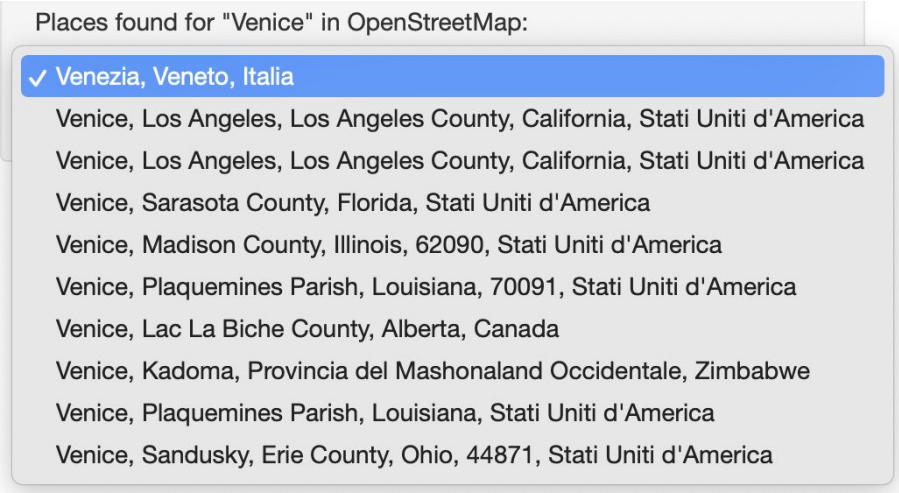




FROM NLP TO VISUALIZATION

5. Correction of geolocation

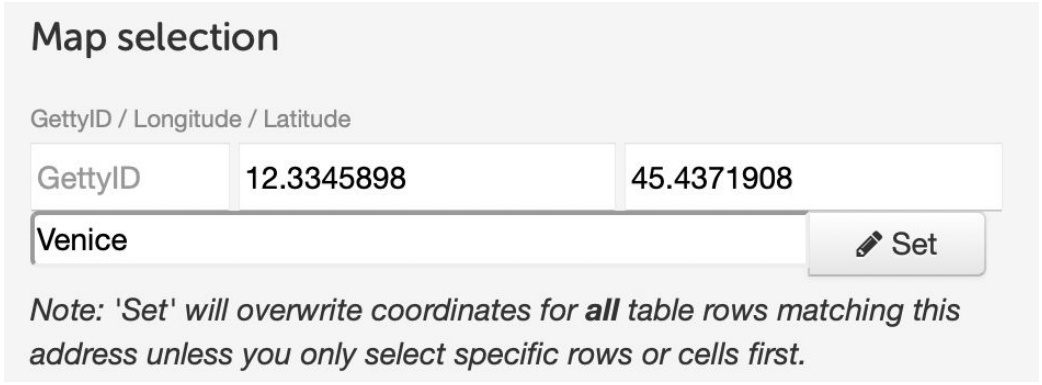
- if the automatic geolocation is NOT correct, choose another option from the drop-down menu
- if there is no correct place in the drop-down menu or no results appear: click on "Map", go to "Map selection", click on the name of another georeferencing system (OSM or GeoNames), choose the right option from the drop-down menu

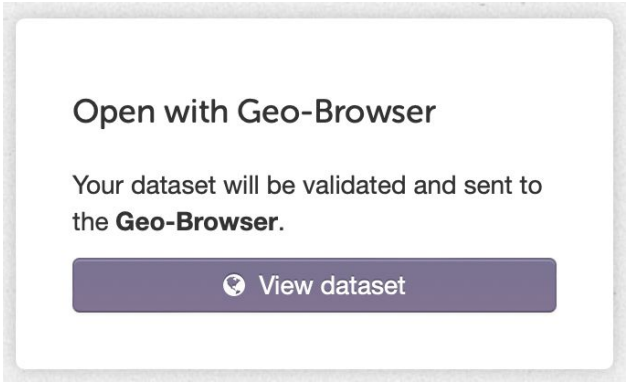


FROM NLP TO VISUALIZATION

5. Correction of geolocation

- once you have chosen the right location, click on "Set" in the "Map selection" section
- latitude and longitude are added to the spreadsheet
- to see the mapped locations, click on "Open with Geo-Browser" button on the right of the interface





Thank you!

Email: rachele.sprugnoli@unipr.it
Twitter: @RSprugnoli