

Part-of-Speech:

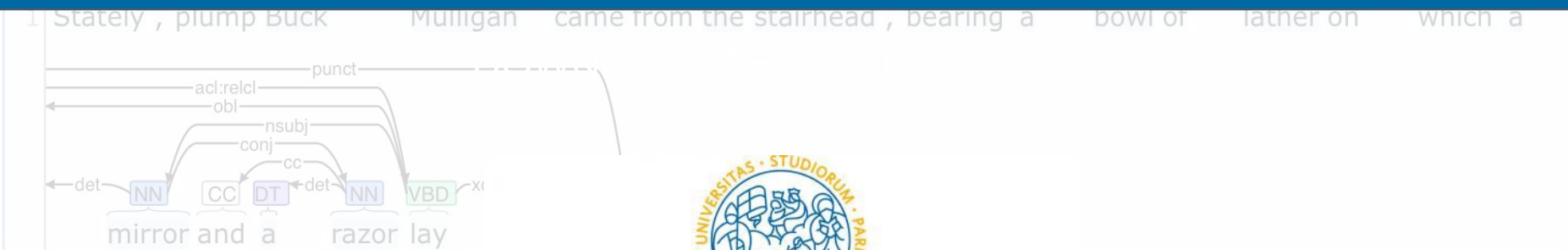


Named Entity Recognition:

Natural Language Processing Methods

Rachele Sprugnoli

rachele.sprugnoli@unipr.it



UNIVERSITÀ DI PARMA



SCHEDULE

PART 1: THEORY

- What is NLP?
- How NLP and DH interact with each other?
- Why NLP is challenging?
- 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?
- How to georeference automatically extracted place names?

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

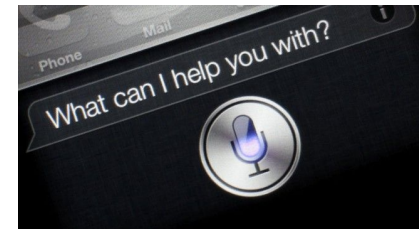
*Bender, Emily M. 2016. "Linguistic Typology in Natural Language Processing".
Linguistic Typology 20(3), 645-660.*

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 IN THE HUMANITIES

- 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 artworks, 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:



Thomas Müller  @esmuellert_ · May 28

Congratulations to @realmadrid with my companions @David_Alaba,
@ToniKroos and @MrAncelotti 🙌🏆 #ucl #RealMadrid 🏆 #RMALIV
#UCLfinal #esmuellert

- 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

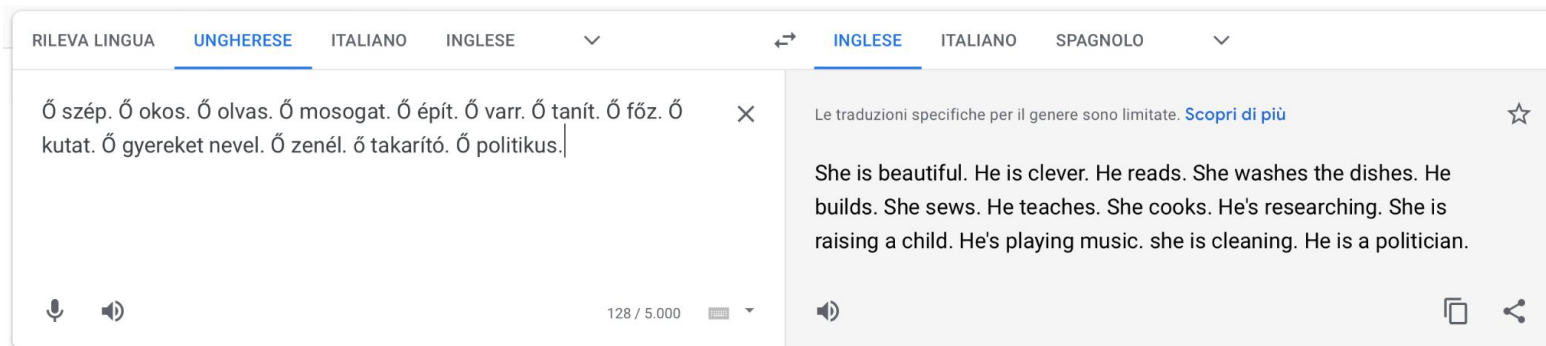
“She has a face like a Botticelli Madonna!”

“He looks like a Picasso painting!”



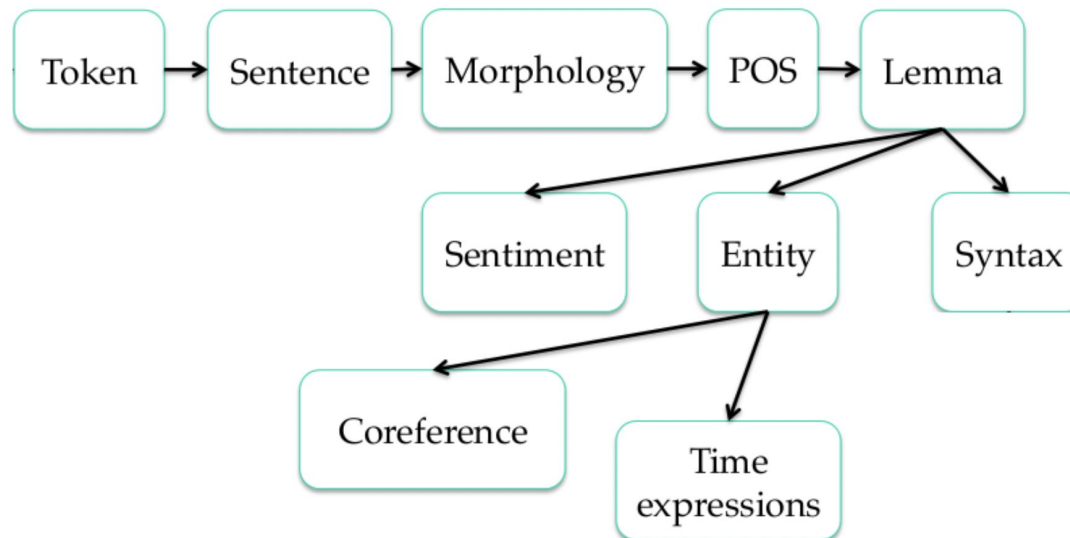
WHY IS NLP A CHALLENGE?

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

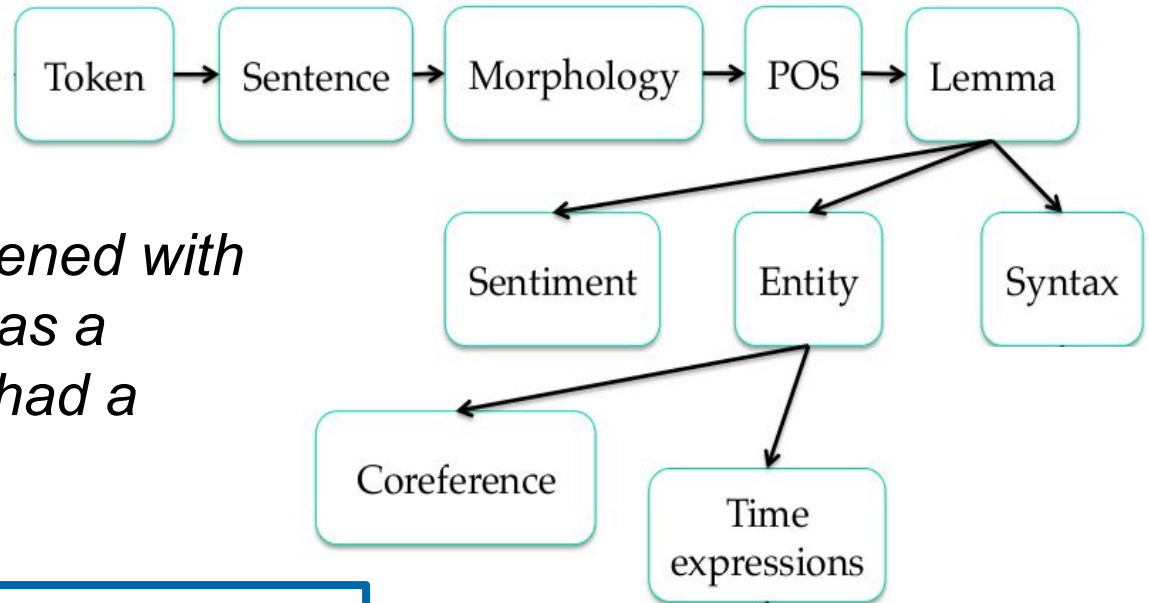


HOW TO PROCESS LANGUAGE

- 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 of a hypothetical pipeline:



Images in the following slides are taken from the CoreNLP demo: <https://corenlp.run>

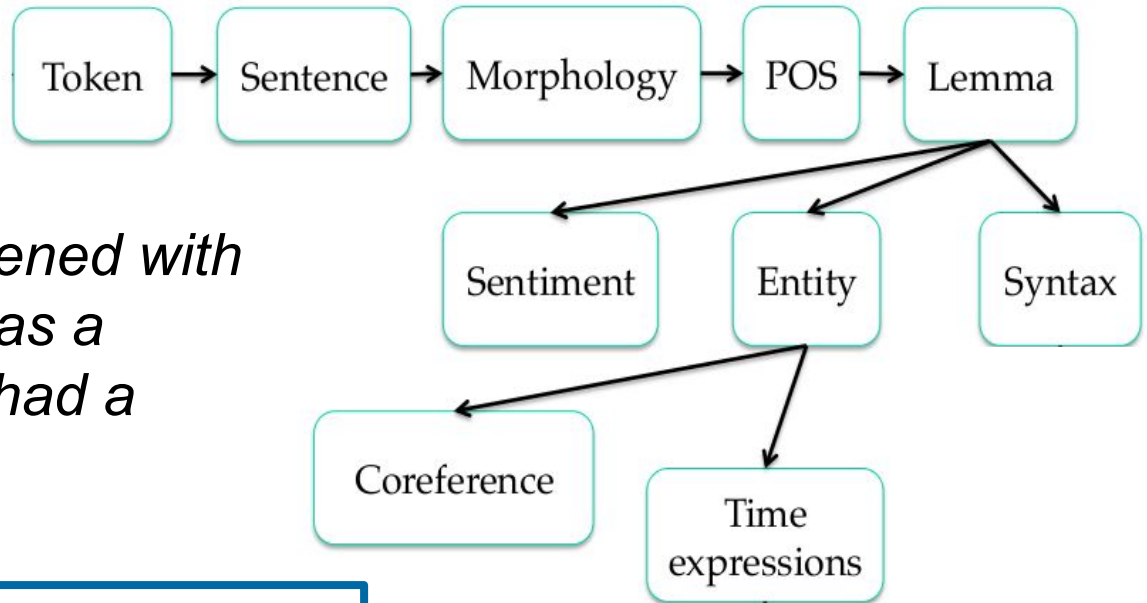


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

	WRB	PRP	VBP	WP	VBD	IN	JJ	NNP	NN	,	PRP	VBD	DT	NN	.
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.

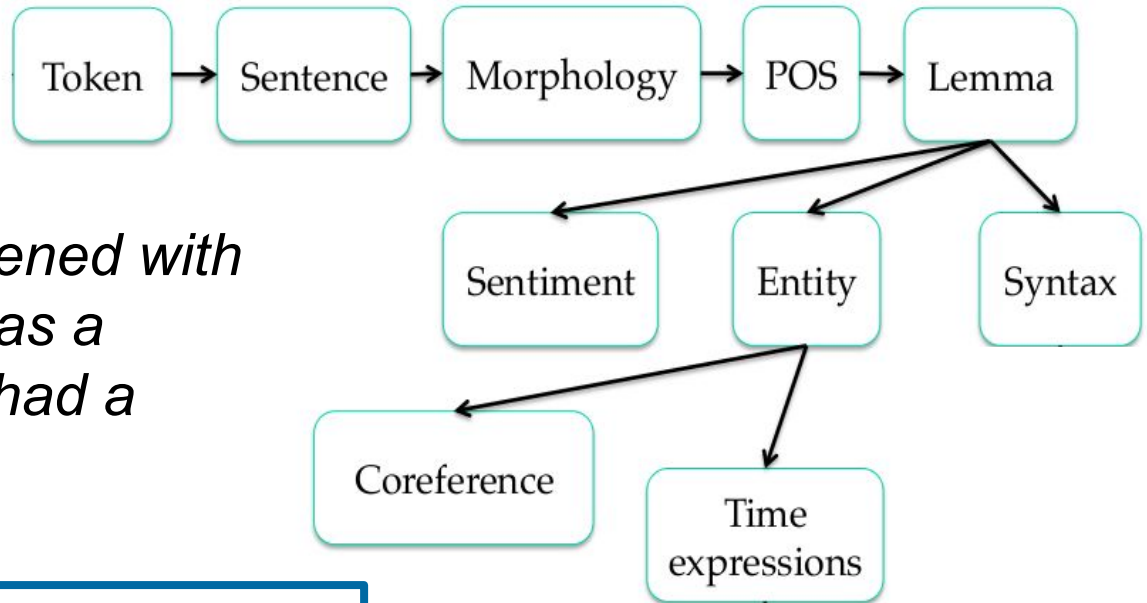
Trump, 2016-08-05

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?



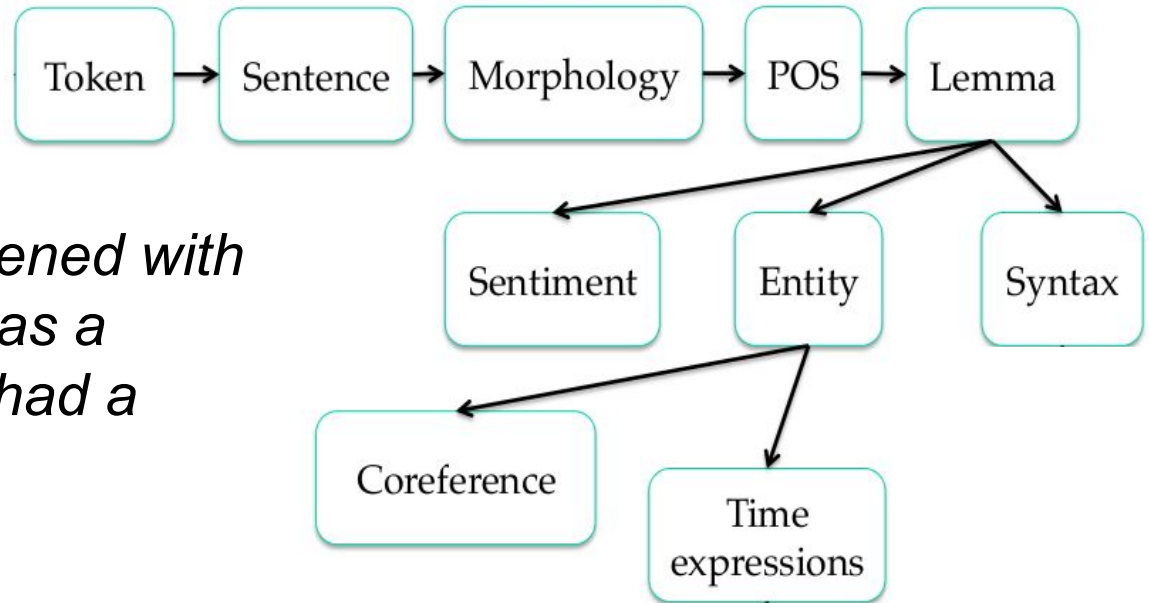
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

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

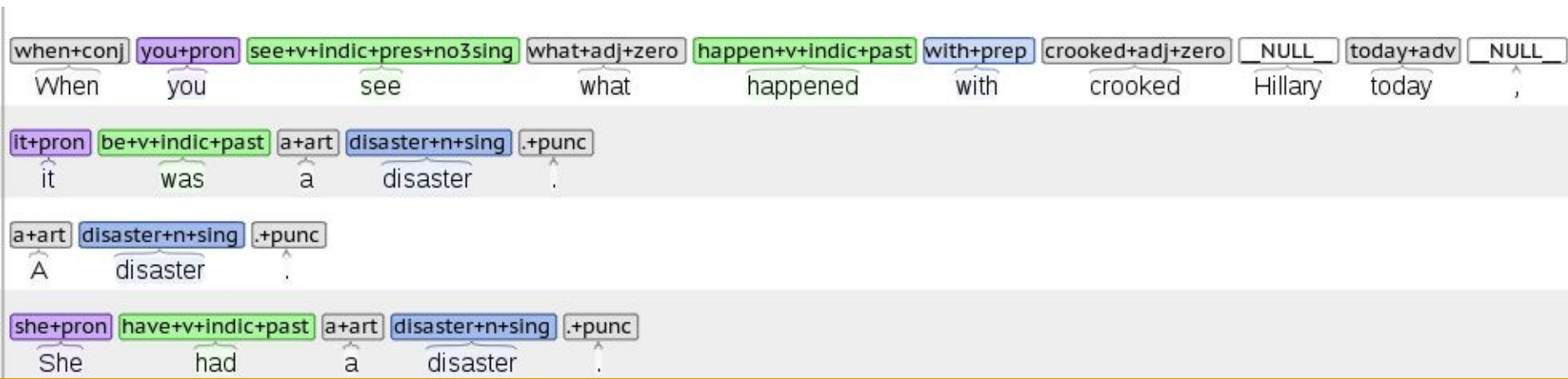
	WRB	PRP	VBP	WP	VBD	IN	JJ	NNP	NN	,	PRP	VBD	DT	NN	.
1	When	you	see	what	happened	with	crooked	Hillary	today	,	it	was	a	disaster	.
2	A	disaster	.												
3	She	had	a	disaster	.										

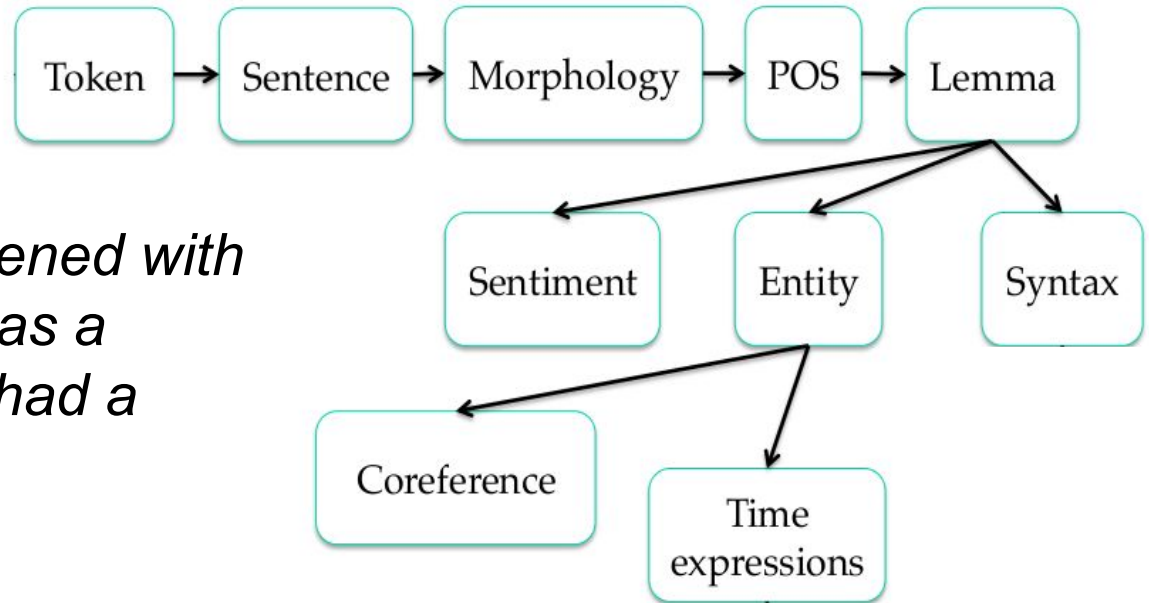


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

MORPHOLOGY





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

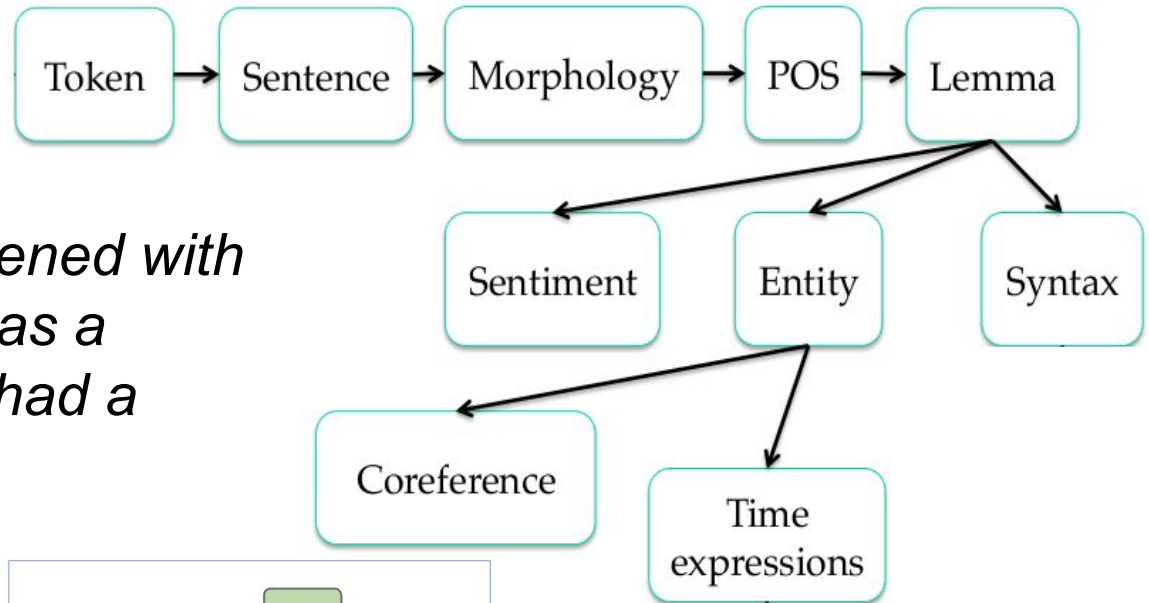
Trump, 2016-08-05

LEMMA

1 when you see what happen with crooked Hillary today , it be a disaster .

2 a disaster .

3 she have a disaster .

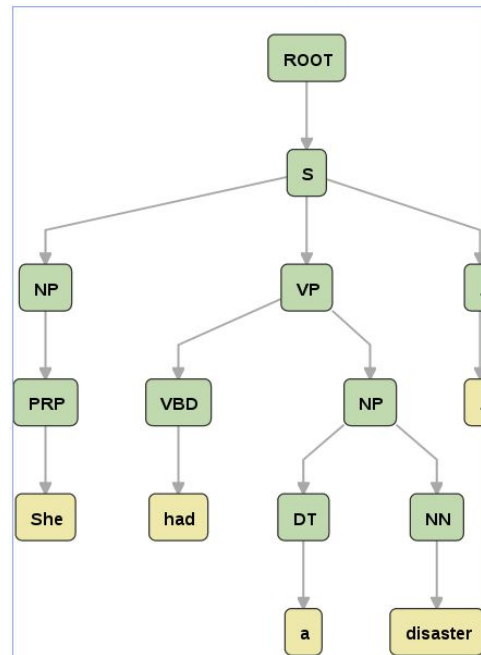


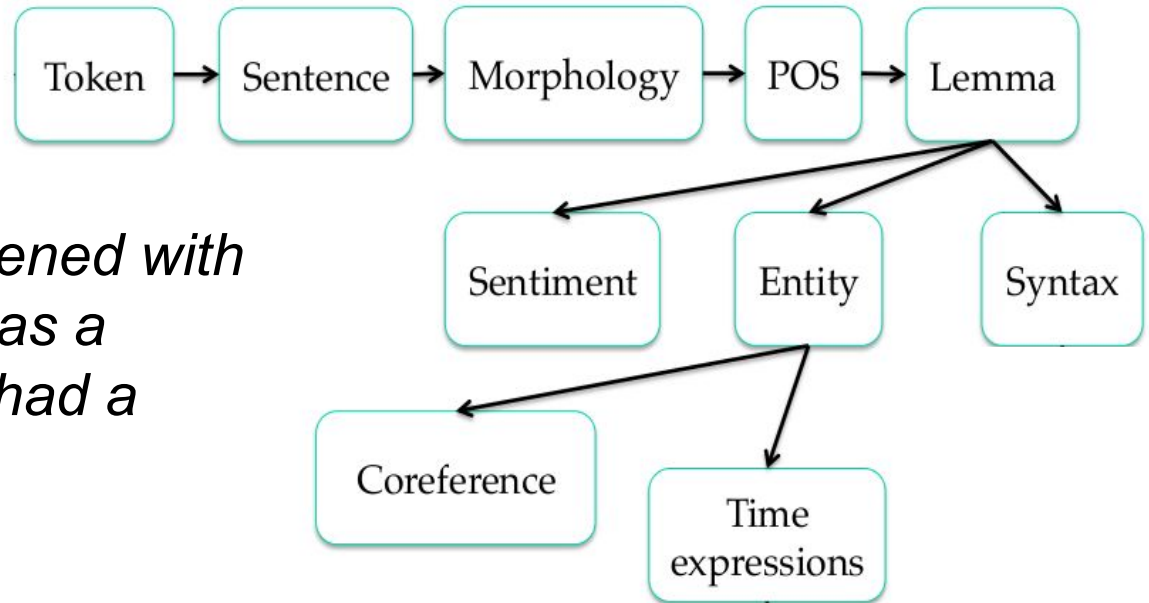
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

SYNTAX / PARSING

- CONSTITUENCY PARSING



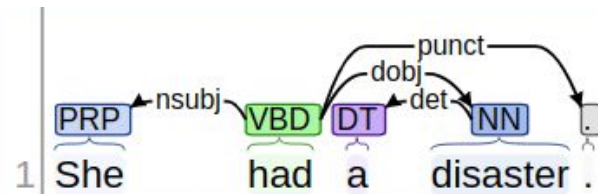


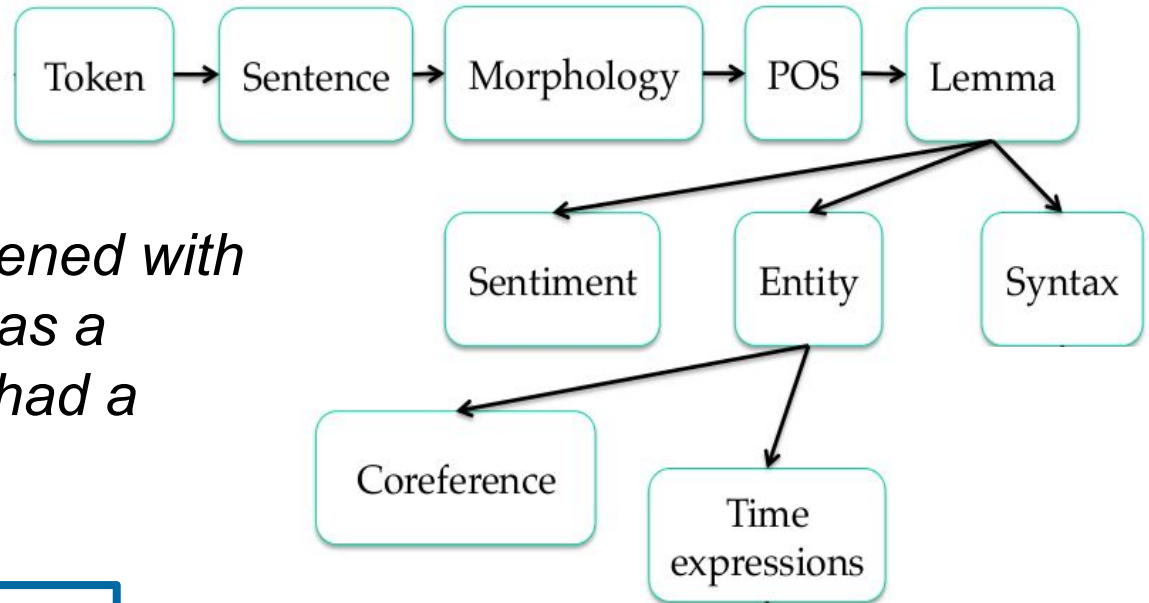
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

SYNTAX / PARSING

- DEPENDENCY PARSING



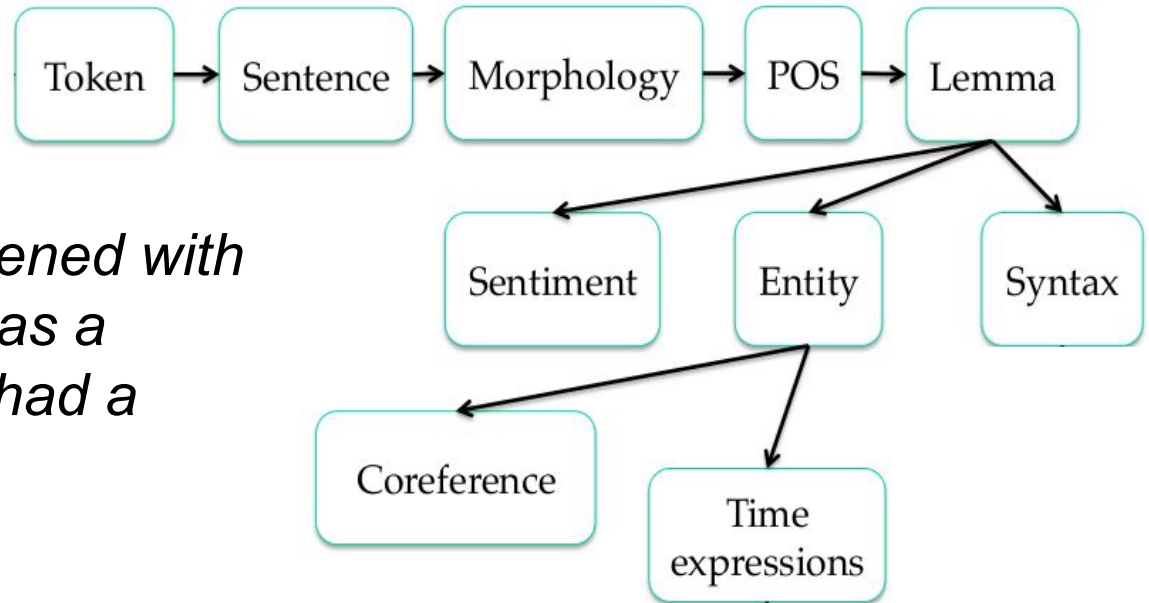


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

NAMED ENTITY RECOGNITION

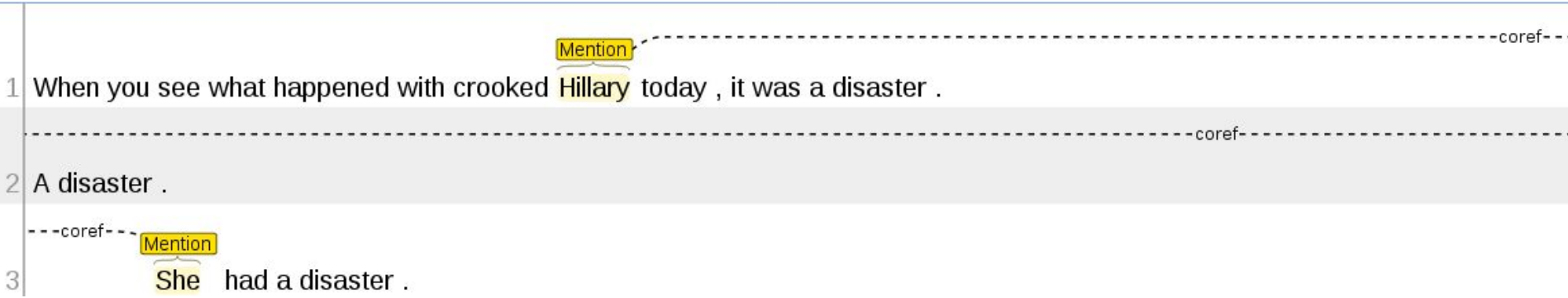
- 1 When you see what happened with crooked PER Hillary today , it was a disaster .
- 2 A disaster .
- 3 She had a disaster .

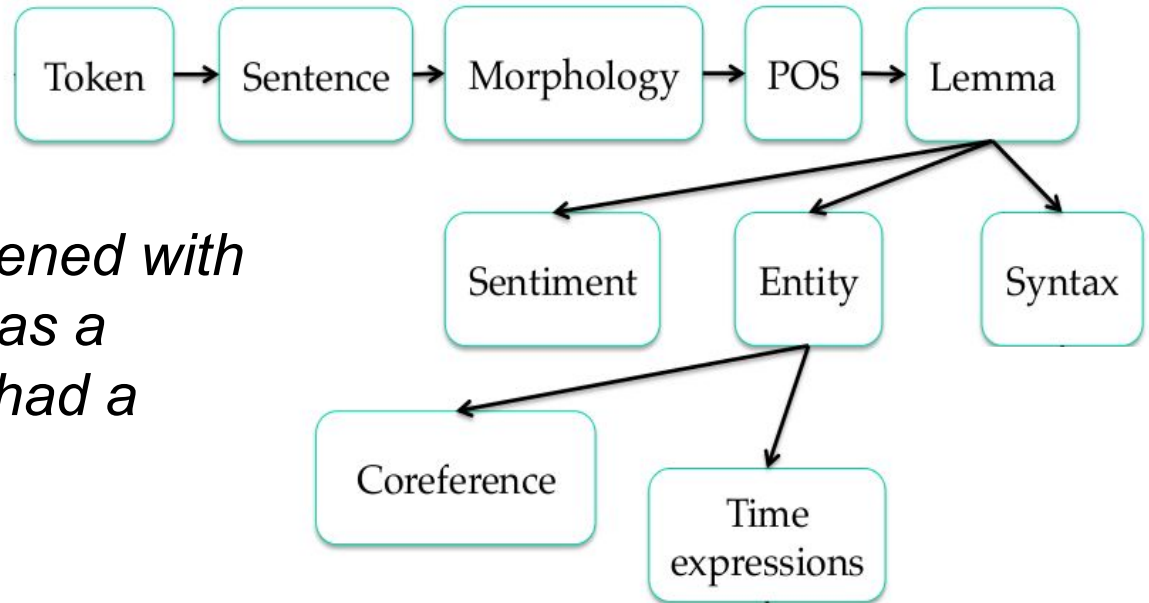


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

COREFERENCE



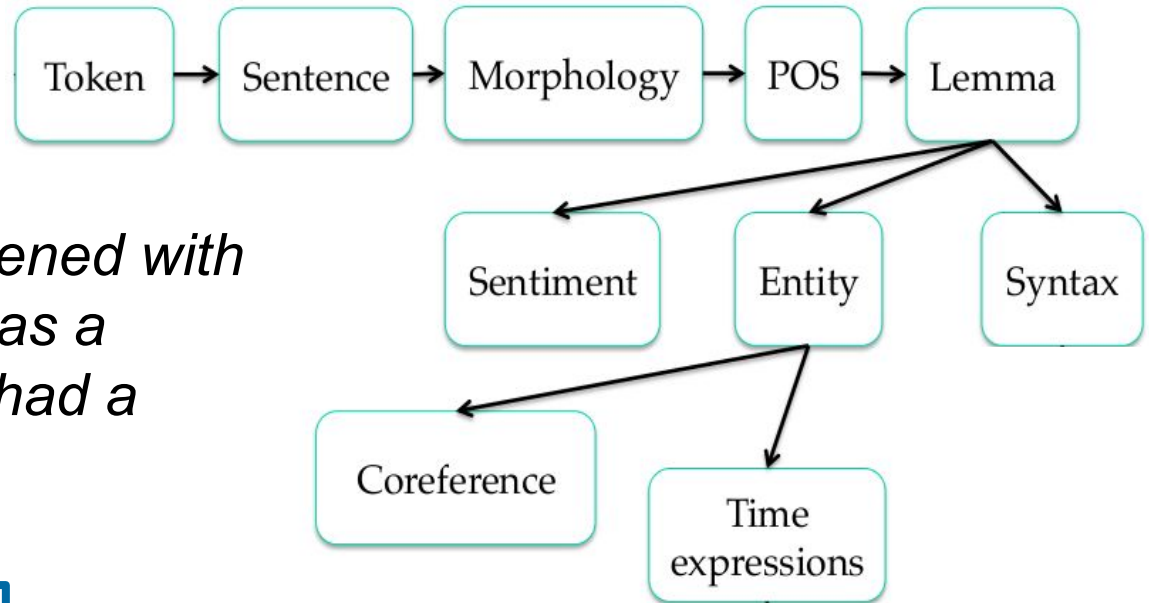


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

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 .	



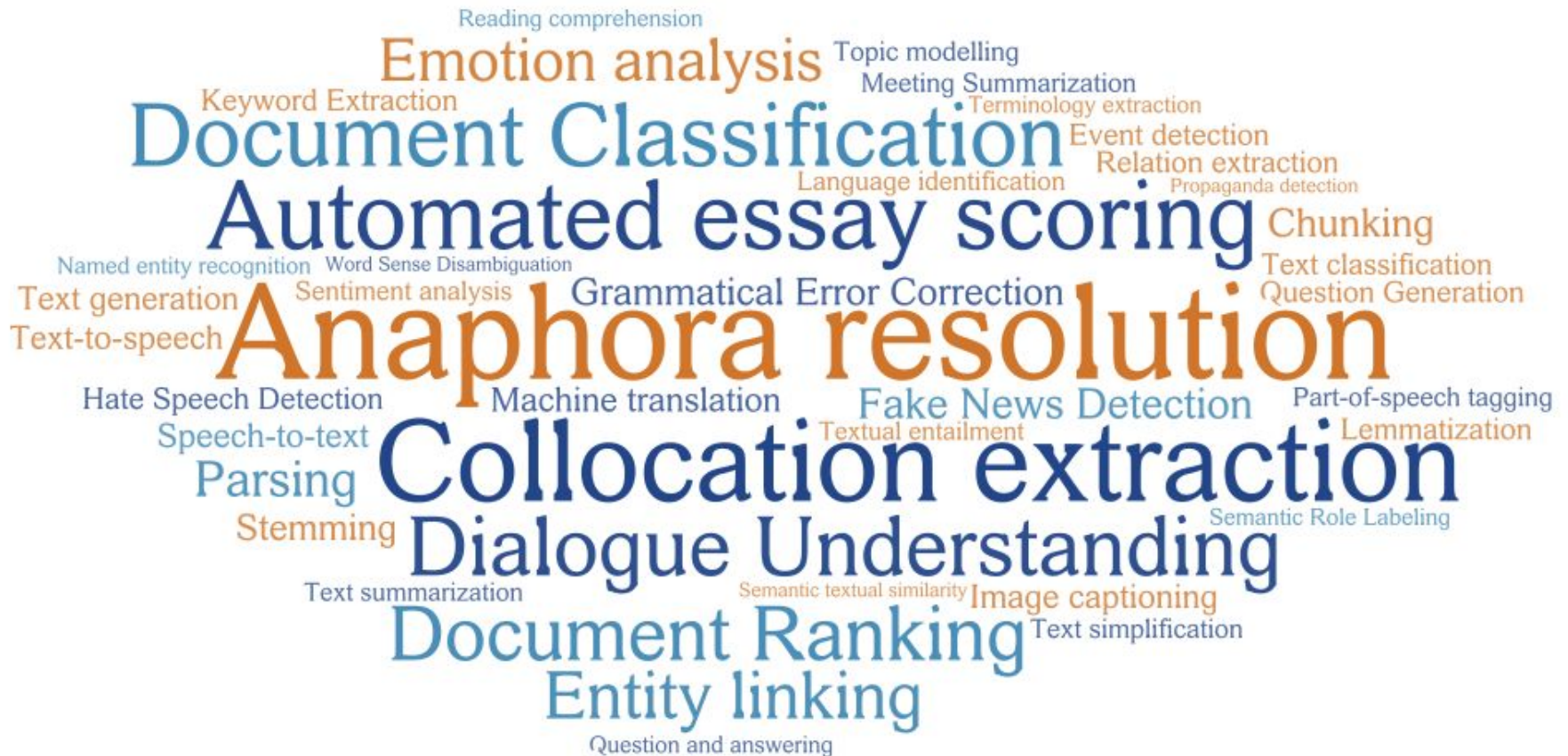
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

SENTIMENT

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

SO MANY TASKS...



HOW TO DEVELOP A MODULE

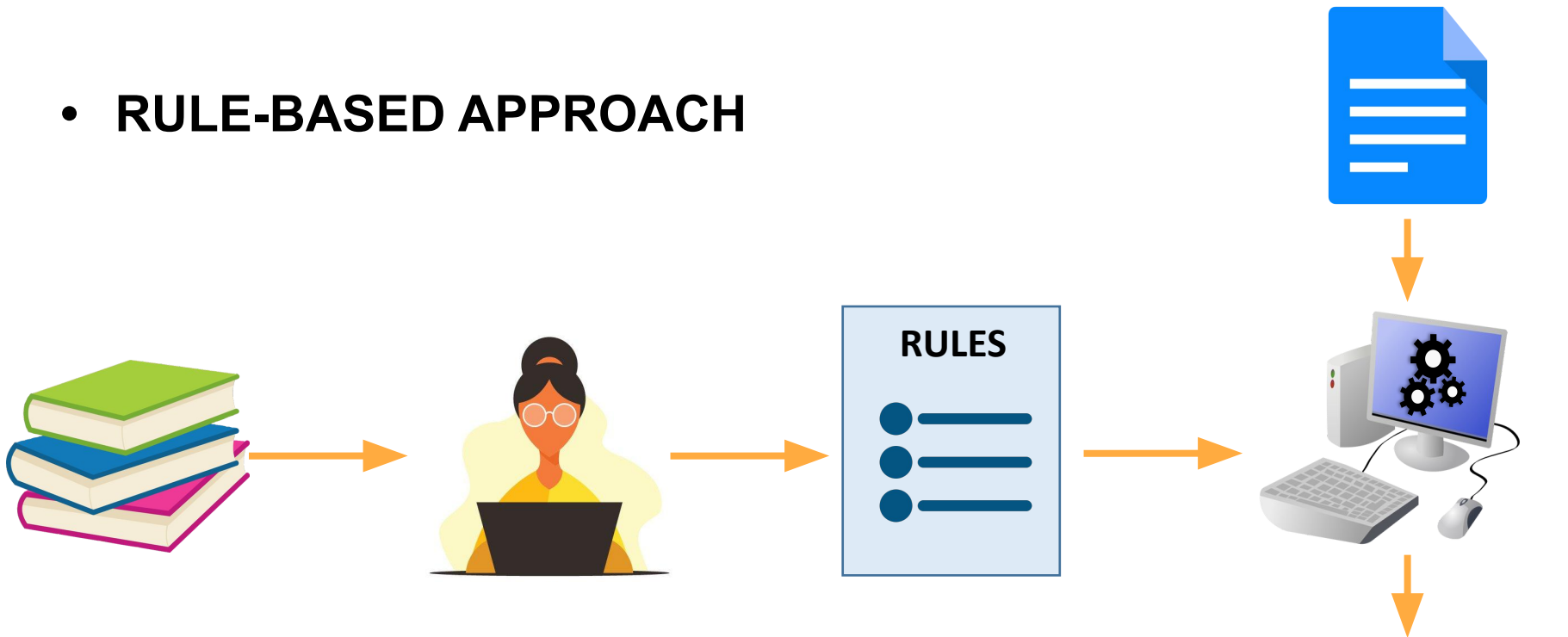
- **LOOKUP LIST APPROACH**

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

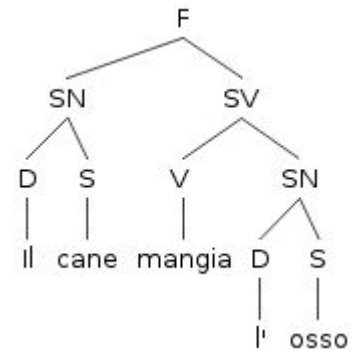
CURRENCIES	CITIES
Euro, euros dollar, dollars, pound, \$, €...	http://download.geonames.org/export/dump/

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
NOUN		ADJ
VERB	DET	ADV
« <i>pity</i> »	<i>the</i>	<i>dead</i> »

2) application of rules to remove ambiguous labels

- «choose NOUN if preceded by DET»

		NOUN
NOUN		ADJ
VERB	DET	ADV
« <i>pity</i> »	<i>the</i>	<i>dead</i> »

HOW TO DEVELOP A MODULE

- **RULE-BASED APPROACH**

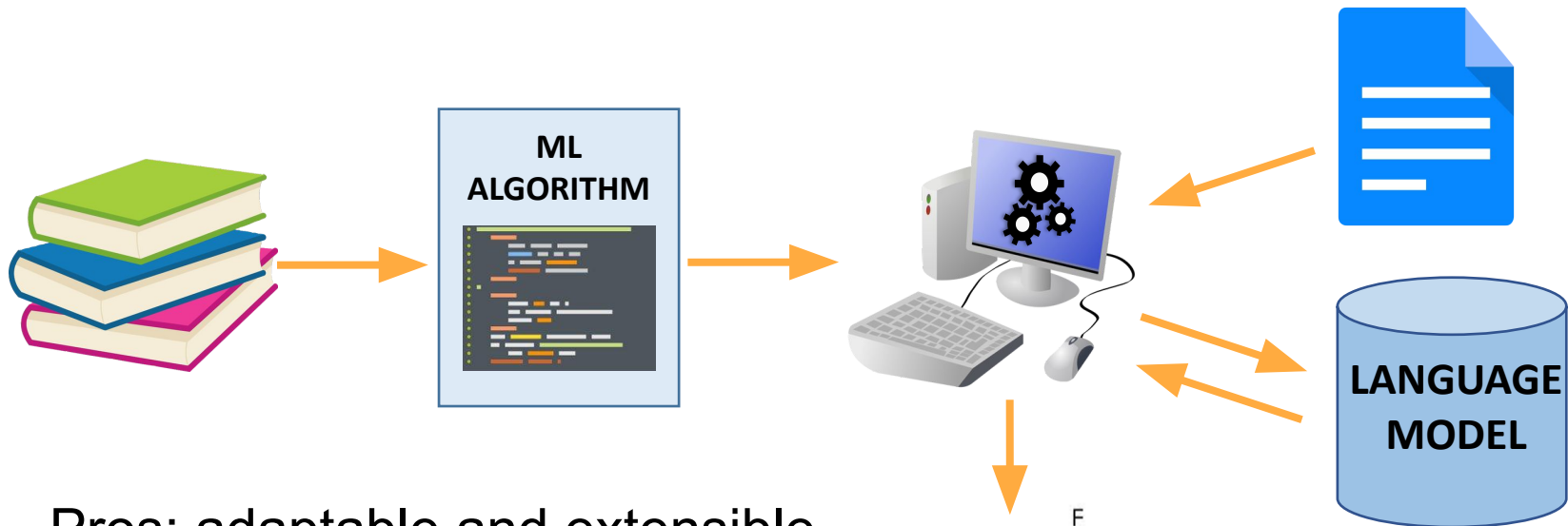
Named Entity Recognition without Gazetteers,

Mikheev et al. 1999

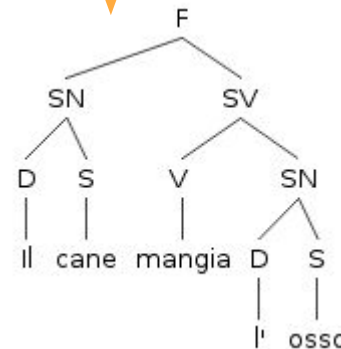
Context Rule	Assign	Example
Xxxx+ is? a? JJ* PROF	PERS	Yuri Gromov, a former director
Xxxx+ is? a? JJ* REL	PERS	John White is beloved brother
Xxxx+ himself	PERS	White himself
Xxxx+, DD+, shares in Xxxx+	PERS	White, 33, shares in Trinity Motors
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
Xxxx+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

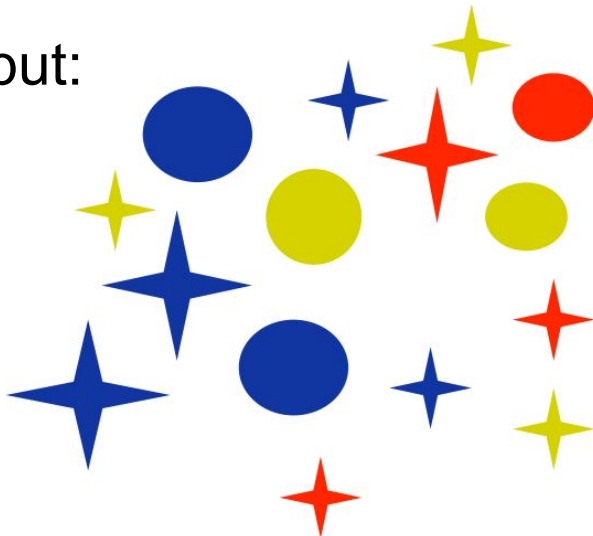
1. **UNSUPERVISED**: they do not need an annotated corpus for training the model
2. **SUPERVISED**: they use an annotated corpus for training the model
3. **SEMI-SUPERVISED**: they combine information from both annotated and non-annotated data for training the model

HOW TO DEVELOP A MODULE

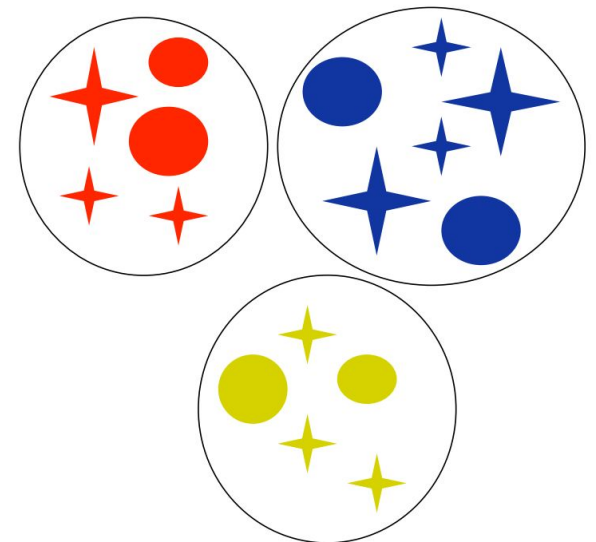
- **MACHINE LEARNING APPROACH**
- **UNSUPERVISED ALGORITHM, example**

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

Input:



Color-based clustering:

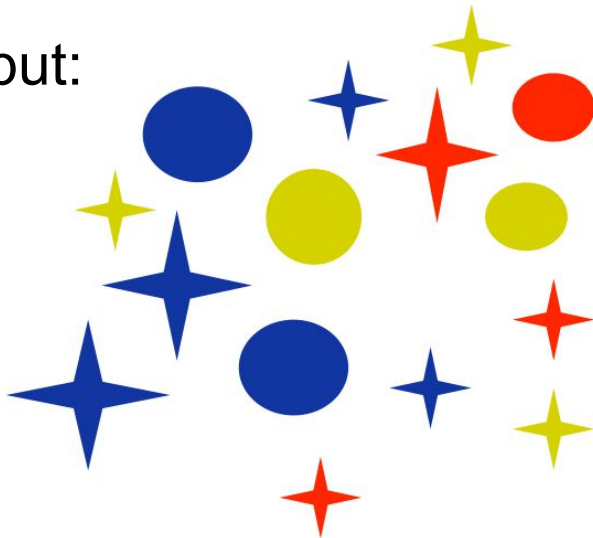


HOW TO DEVELOP A MODULE

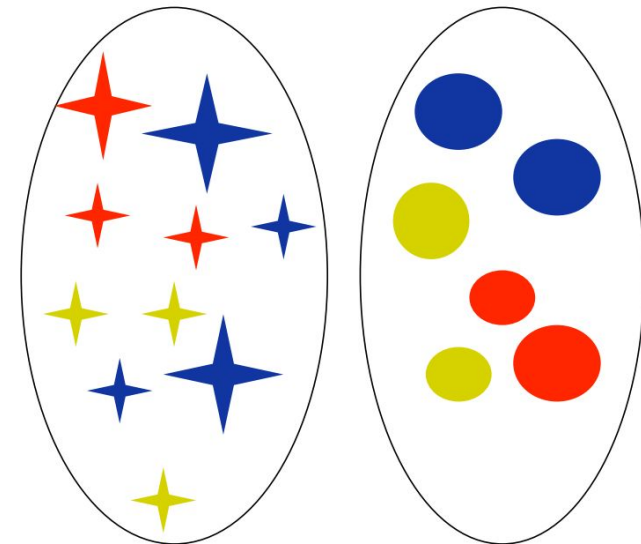
- **MACHINE LEARNING APPROACH**
- **UNSUPERVISED ALGORITHM, example**

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

Input:



Form-based clustering:

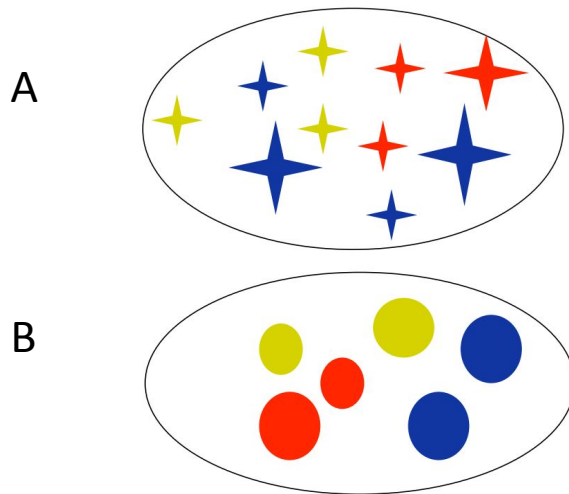


HOW TO DEVELOP A MODULE

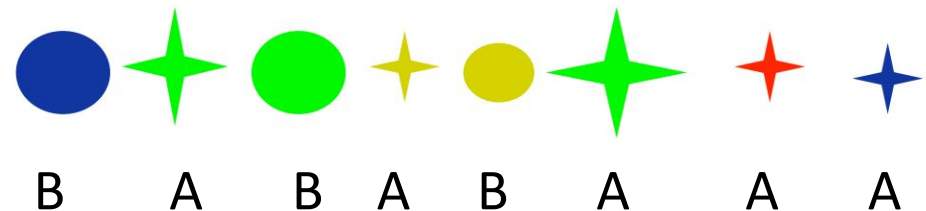
- MACHINE LEARNING APPROACH
- SUPERVISED ALGORITHM, example

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

Input (training):

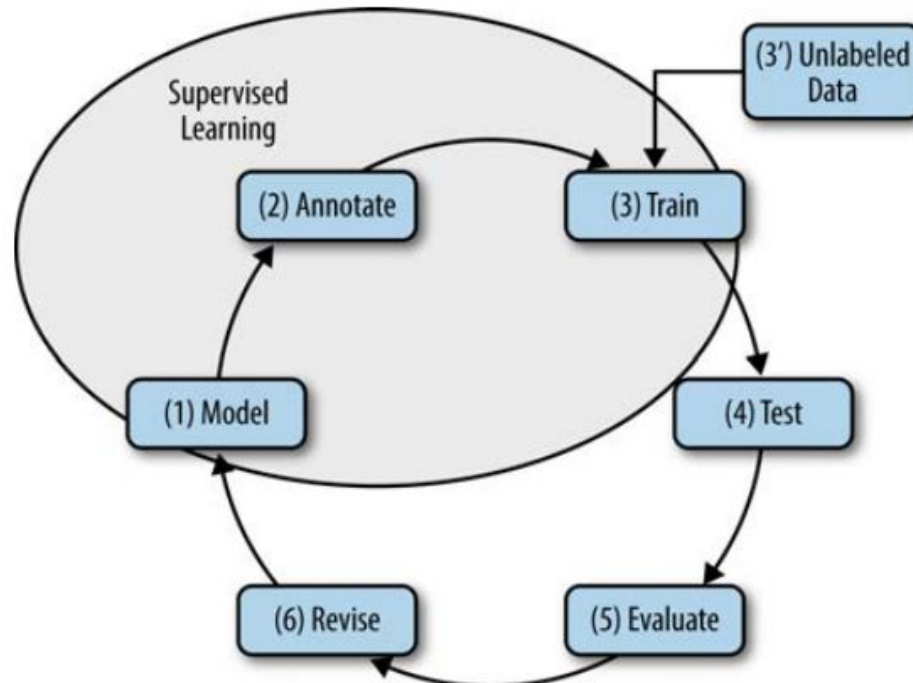


Classification of unseen data (test):



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 following 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 model, annotation, algorithm

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,** <https://universaldependencies.org>**): 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
<u>ADJ</u>	<u>ADP</u>	<u>PUNCT</u>
<u>ADV</u>	<u>AUX</u>	<u>SYM</u>
<u>INTJ</u>	<u>CCONJ</u>	<u>X</u>
<u>NOUN</u>	<u>DET</u>	
<u>PROPN</u>	<u>NUM</u>	
<u>VERB</u>	<u>PART</u>	
	<u>PRON</u>	
	<u>SCONJ</u>	

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

ANNOTATION: EXAMPLE

- UNIVERSAL DEPENDENCIES (UD): Features

Lexical features	Inflectional features	
	<i>Nominal*</i>	<i>Verbal*</i>
<u>FronType</u>	<u>Gender</u>	<u>VerbForm</u>
<u>NumType</u>	<u>Animacy</u>	<u>Mood</u>
<u>Poss</u>	<u>NounClass</u>	<u>Tense</u>
<u>Reflex</u>	<u>Number</u>	<u>Aspect</u>
<u>Foreign</u>	<u>Case</u>	<u>Voice</u>
<u>Abbr</u>	<u>Definite</u>	<u>Evident</u>
<u>Type</u>	<u>Degree</u>	<u>Polarity</u>
		<u>Person</u>
		<u>Polite</u>
		<u>Clusivity</u>

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

ANNOTATION: EXAMPLE

- UNIVERSAL DEPENDENCIES (UD): Syntactic Relations

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

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

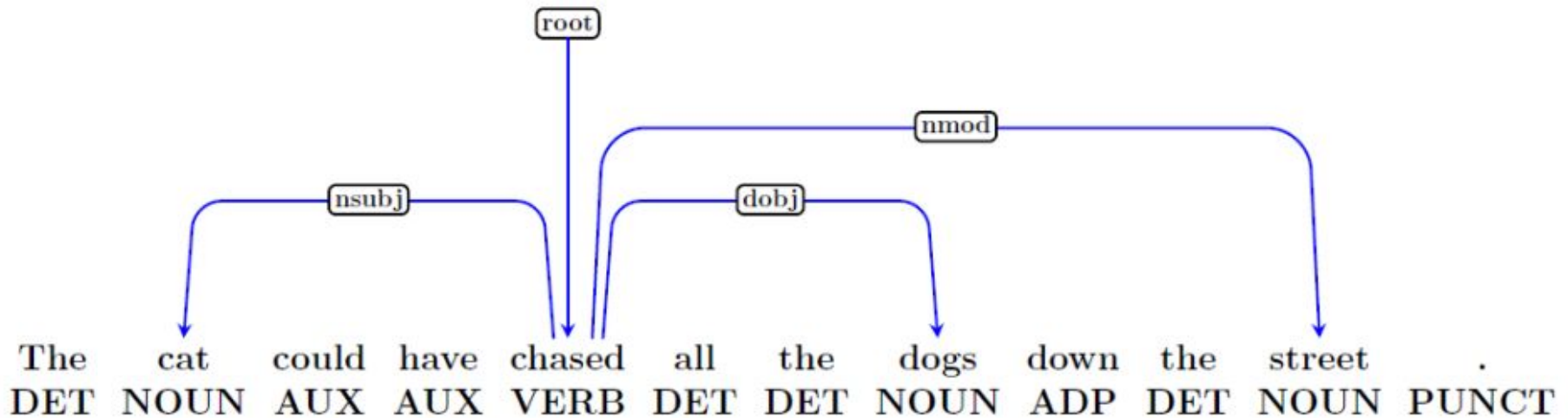
ANNOTATION: EXAMPLE

- **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

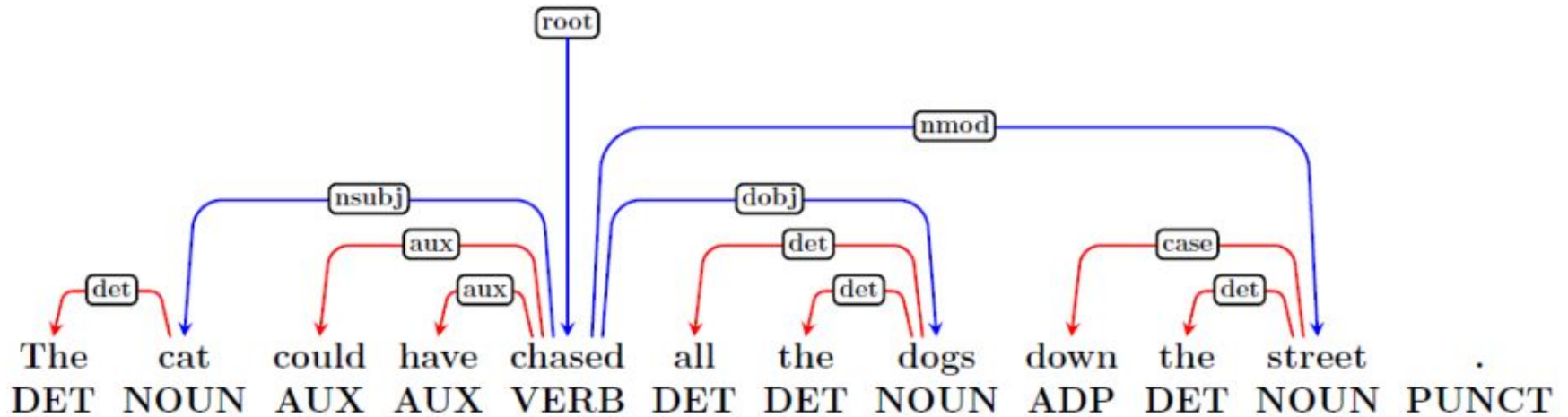
- UNIVERSAL DEPENDENCIES (UD): annotation



1. Content words are linked with dependencies relations

ANNOTATION: EXAMPLE

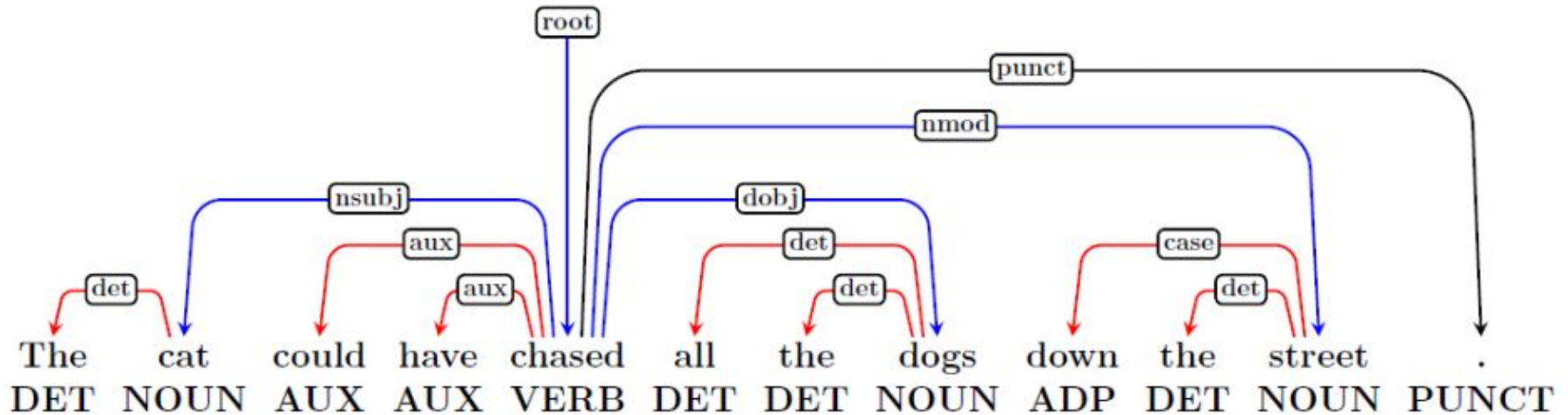
- UNIVERSAL DEPENDENCIES (UD): annotation



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

ANNOTATION: EXAMPLE

- UNIVERSAL DEPENDENCIES (UD): annotation

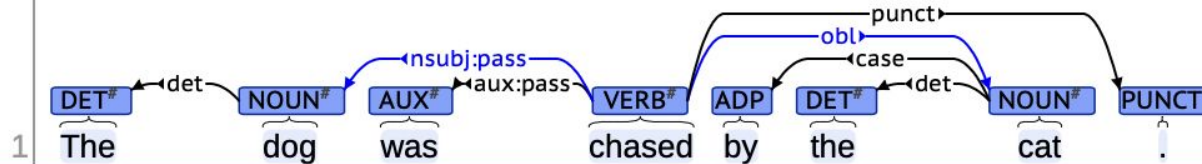


1. Content words are linked with dependencies relations
2. Function words depend on the content word they modify
3. Punctuation is attached to content words and can never has dependents

ANNOTATION: EXAMPLE

- UNIVERSAL DEPENDENCIES (UD): annotation

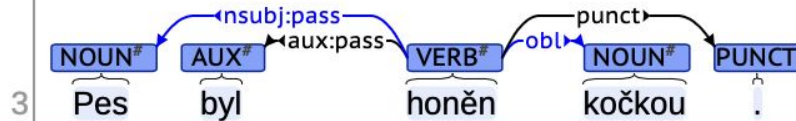
English



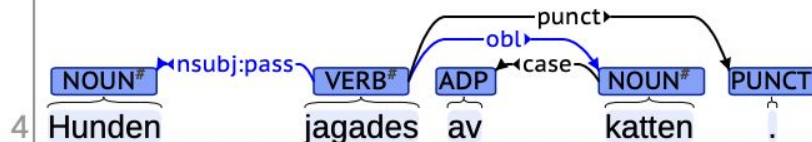
Bulgarian



Czech



Swedish



ANNOTATION: FORMATS

- CoNLL-U

```
# sent_id = DVE-124
# text = Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improprium quendam cecinisse recolimus Enter l' ora del vesper cio
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 improprium improprium 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 recole     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    vesper    X        zi       Foreign=Yes                                  18 flat:foreign --
23 cio       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   ochiover  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

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

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">.</tc:token>
</tc:tokens>
```

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>
```

POS TAGS

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 OF CAMONICA.  
</tc:text>
```

```
<tc:tokens xmlns="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>

Annotation

1 ὅτι Ἀρχέστρατος ὁ Συρακούσιος ἢ Γελῶος ἐν τῇ ὥσ' Χρυσίππος ἐπιγράφει Γαστρονομίᾳ, ὥς δὲ Λυγκεύς καὶ Καλλιμαχος ἠδυσπαθεῖα, ὥς δὲ Κλέαρχος Δειπνολογία, ὥς δ' ἄλλοι Ὀψοποία ἐπιπικόν δὲ τὸ ποίημα, οὐ ἡ ἀρχή· Ἑλλάς

2 ἱστορίας ἐπίδειγμα ποιούμενος Ἑλλάδι πάση — φησί·
3 πρὸς δὲ μὴ πάντας δειπνεῖν ἀβρόδοται τραπέζῃ.
4 ἔστωσαν δ' ἢ τρεῖς ἢ τέσσαρες οἱ ξυνάπαντες ἢ τῶν πέντε γε μὴ πλείους·
5 ἤδη γὰρ ἂν εἴη μισθοφόρων ἀρπαξιβίων σκηνὴ στρατιωτῶν.

BRAT, <https://brat.nlplab.org>

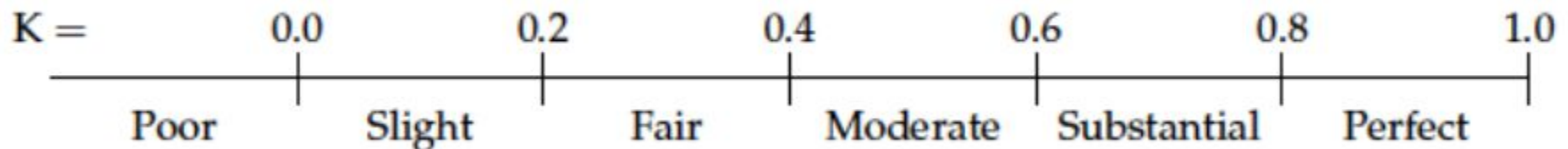
1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about \$100 million for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

Entities: Organization, Org, Person, GPE, Money

Relations: Giver, Recipient, Beneficiary, Money, Family, Origin

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**

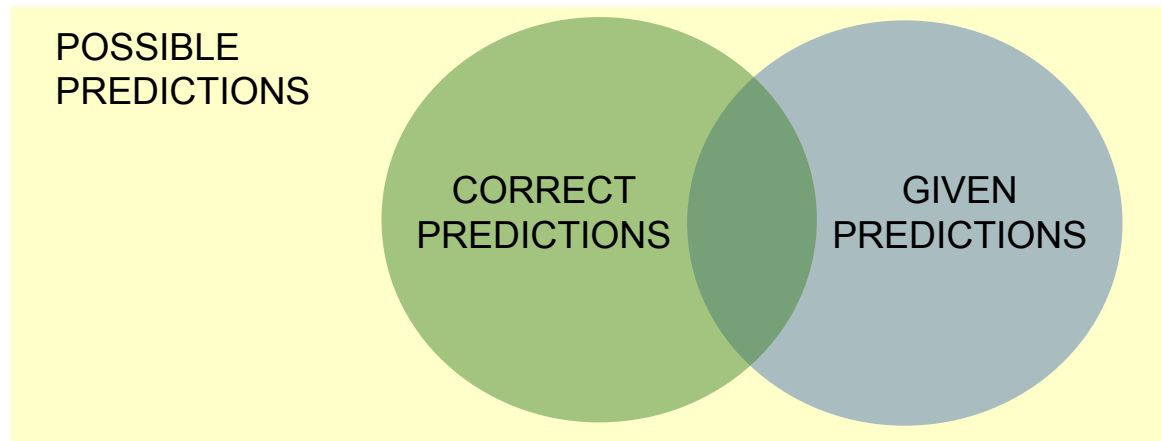
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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

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




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

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

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



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

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
 - $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$
- Alternative metric: parameterized average, which allows to choose to give more importance to P or R: when beta = 1 we speak of F-1

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

$\beta = 1$: P and R they have the same weight

$\beta > 1$: R is more important

$\beta < 1$: P is more important

$\beta = 0$: 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 text 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 - <https://weblicht.sfs.uni-tuebingen.de/>*
 - 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

UDPipe

- *UDPipe is a **trainable** pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files - <https://lindat.mff.cuni.cz/services/udpipe/>*
 - trainable
 - 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 (CLI) or a web based interface: CLI is the only option if you want to train a new model

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. - <http://lindat.mff.cuni.cz/services/nametag/>*
 - 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

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 - <http://nlp.d4science.org/hub/>*
 - 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

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*

The screenshot displays the 'Switchboard' interface. At the top, the title 'Switchboard' is followed by the text: 'Switchboard helps you find tools that can process your data. The data will be shared with the tools via public links. For more details, see the [FAQ](#).' Below this text are two buttons: 'Upload files or text' (highlighted in blue) and 'Tool inventory'. Underneath these buttons is the section 'Add your data', which contains three tabs: 'Upload File', 'Submit URL', and 'Submit Text'. Below the tabs is a large dashed rectangular box containing the text 'Drop files here, or click to select file'. Two blue arrows originate from the text in the first list item: one points to the 'Upload files or text' button, and the other points to the dashed file selection area.

USING SWITCHBOARD

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

Resources

[Pisa_Italian_Days_and_Ways.txt](#) 6.78 KiB

Show content

Mediatype

text/plain

Language

English

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

▼ Constituency Parsing



> WebLicht Const Parsing EN

Open

Requires authentication

▼ Dependency Parsing



> UDPipe

Open



> WebLicht Dep Parsing EN

Open

Requires authentication

USING SWITCHBOARD

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

The screenshot displays the UDPipe web interface. At the top, under 'Model:', there are two radio buttons: 'UD 2.6 (description)' (selected) and 'EvaLatin20 (description)'. Below this is a language dropdown menu showing 'english-ewt-ud-2.6-200830'. Under 'Actions:', there are two checked checkboxes: 'Tag and Lemmatize' and 'Parse'. A grey bar labeled 'Advanced Options' is partially visible. Below the settings are two input options: 'Input Text' (selected) and 'Input File'. The 'Input Text' area contains a large text box with the following text: 'Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear and eminently sane, but you seem to have strangely perverted my meaning; then you revert to an earlier letter from La Cava, and are pleased to imagine that we are taking risks all the time and leading a reckless life generally. I shall really hesitate to tell you again of any of our adventures such as that drive home from Pæstum, which I merely related as an amusing incident. There is no danger of brigands in these days and we did not "need a protector," especially as kind Providence looked after us. That drunken driver would not have surrendered his reins to you or to any one except the padrone; and then "all's well that ends well," and we returned from our excursion with nothing worse than a grievance. I was so vexed with you for two whole days that I wrote you not one line from Siena or Pisa. Now your indiscretion is partially atoned for by a letter which has just reached me here, and I am trying to forgive you and "be friends again," as we used to say when we were children. But the charms of Siena are already so eclipsed by those of Florence that it is quite impossible for me to give you an atmospheric description of its streets and churches, above all of the shining cathedral, rich from dome to pavement with colored marbles, frescoes, and mosaics. This may be no loss to you, who are doubtless well tired of my Italian rhapsodies; but your respite is only temporary, as I quite missed writing you that letter. I wanted to tell you that the campanile at Pisa

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

A Output Text Show Table Show Trees

Save Output File

```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipe_model = english-ewt-ud-2.6-200830
# udpipe_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# 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
```


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

The screenshot shows the UDPipe Switchboard interface with the following options and callouts:

- 1** Model: UD 2.6 (description) EvaLatin20 (description)
- 1**
- 2** Actions: Tag and Lemmatize Parse
- 3** Advanced Options: UDPipe version: UDPipe 2 UDPipe 1
- 4** Input [?]: Tokenize plain text CoNLL-U Horizontal Vertical
- 5** Tokenizer [?]: Normalize spaces Presegmented input Save token ranges

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

The screenshot shows the UDPipe interface. At the top, there are three buttons: 'Output Text', 'Show Table', and 'Show Trees'. Below these is a green bar with a 'Save Tree as SVG' button, which is highlighted by a blue arrow. Below the bar is a pagination control with buttons for 'Previous', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '...', and 'Next'. The main content area displays the sentence: "Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact ?". Below the sentence is a parse tree diagram. The root node is '<root>', which branches into 'realize' (root VERB) and 'pact ?' (ccomp NOUN, punct PUNCT). 'realize' branches into 'Do' (aux AUX) and 'you' (nsubj PRON). 'pact ?' branches into 'that' (mark SCONJ), 'letter' (nsubj NOUN), 'was' (cop AUX), 'not' (advmod PART), 'quite' (advmod ADV), 'within' (case ADP), 'the' (det DET), and 'pact ?' (ccomp NOUN, punct PUNCT). 'letter' branches into 'your' (nmod:poss PRON), 'answer' (nmod NOUN), 'March' (nmod PROPN), and 'Rome' (nmod PROPN). 'answer' branches into 'in' (case ADP), 'mine' (nmod PRON), 'of' (case ADP), and '18th' (nummod NOUN). 'mine' branches into 'to' (case ADP). '18th' branches into 'from' (case ADP).

USING SWITCHBOARD

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

Model: NameTag 2 (description) NameTag 1 (description)

Input: Plain text Vertical CoNLL-U

Output: XML (original text with annotations) Vertical (retrieved named entities only) CoNLL (?) CoNLL-U+NE (?)

Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear and eminently sane, but you seem to have strangely perverted my meaning; then you revert to an earlier letter from La Cava, and are pleased to imagine that we are taking risks all the time and leading a reckless life generally. I shall really hesitate to tell you again of any of our adventures such as that drive home from Pæstum, which I merely related as an amusing incident. There is no danger of brigands in these days and we did not "need a protector," especially as kind Providence looked after us. That drunken driver would not have surrendered his reins to you or to any one except the padrone; and then "all's well that ends well," and we returned from our excursion with nothing worse than a grievance. I was so vexed with you for two whole days that I wrote you not one line from Siena or Pisa. Now your indiscretion is partially atoned for by a letter which has just reached me here, and I am trying to forgive you and "be friends again," as we used to say when we were children. But the charms of Siena are already so eclipsed by those of Florence that it is quite impossible for me to give you an atmospheric description of its streets and churches, above all of the shining cathedral, rich from dome to pavement with colored marbles, frescoes, and mosaics. This may be no loss to you, who are doubtless well tired of my Italian rhapsodies; but your respite is only temporary, as I quite missed writing you that letter. I wanted to tell you that the campanile at Pisa leans quite as much as the little Parian

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

Output: XML (original text with annotations) Vertical (retrieved named entities only) CoNLL (?) CoNLL-U+NE (?)

The screenshot shows the NameTag tool interface. At the top, the 'Output' format is set to 'Vertical (retrieved named entities only)'. Below this, there are two tabs: 'A Raw Output' and 'Highlighted Output'. The 'Highlighted Output' tab is active, showing a table of named entities. A green bar with a download icon and the text 'Save Output File' is visible above the table. The table has three columns: 'Token Range', 'Entity Type', and 'Entity Text'. The data rows are: 15 LOC Rome, 30 LOC Siena, and 67,68 PER La Cava.

Token Range	Entity Type	Entity Text
15	LOC	Rome
30	LOC	Siena
67,68	PER	La Cava

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

- **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

TRY IT YOURSELF!

From UDPipe to NameTag



From NameTag to UDPipe



USING SWITCHBOARD

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

Input and Chain Selection

Title [Plain Text]
Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear ...

SFS: To TCF Converter
Language: English
Document Type: TCF
TCF Version: 5
Text

SFS: Stanford Tokenizer
Sentences
Tokens

SFS: Jitar POS Tagger
Part of Speech: Penn Treebank Te

SFS: TurboParser
Parsing (Dep): No Empty Tokens
Parsing (Dep): With Multi Govs
Parsing (Dep): pennnb

Run Tools Clear Results Download chain

Available Annotations for:
English Plain Text

- Pos Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities

- Some chains are ready to be used, just select another task 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)

The screenshot displays the TüNDRA interface for processing the sentence: "Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact?".

Available Annotations for: English Plain Text

- Pos Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities

Visualization

The visualization shows the sentence with dependency arcs and labels:

- mine (NN) → of (IN) → March (NNP) → 18th (JJ) (NMOD)
- mine (NN) → from (IN) → Rome (NNP) (PMOD)
- was (VBD) → not (RB) → quite (RB) → within (IN) → the (DT) (VMOD, AMOD)

Table view

Input and Chain Selection

Buttons: Run Tools, Clear Results, Download chain

Title [Plain Text]	SFS: To TCF Converter	SFS: Stanford Tokenizer	SFS: Jitar POS Tagger	SFS: TurboParser
Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear ...	Language: English Document Type: TCF TCF Version: 5 Text	Sentences Tokens	Part of Speech: Penn Treebank Te	Parsing (Dep): No Empty Tokens Parsing (Dep): With Multi Govs Parsing (Dep): pennnb

Two blue arrows point from the text in the first bullet point to the 'Download chain' button and the download icons in the 'SFS: Jitar POS Tagger' and 'SFS: TurboParser' columns.

USING SWITCHBOARD

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



The screenshot shows the WebLicht interface. At the top, there are three tabs: 'Main Page', 'Chain 1 ✕', and '+ New Chain'. A blue arrow points from the text 'New Chains' in the list above to the '+ New Chain' tab. Below the tabs, on the left, is a sidebar titled 'Available Annotations for: English Plain Text' with radio buttons for 'Pos Tags/Lemmas', 'Morphology', 'Constituent Parses', 'Dependency Parses' (which is selected), and 'Named Entities'. The main area is titled 'Query' and contains a text input field with the placeholder text 'Enter either a TIGERSearch query, or simply a word in quotation marks.' and a search icon. Below the input field are two buttons: 'History' and 'Query Language Help'. At the bottom of the main area, there is a navigation bar for the query results, showing 'Sentence' followed by navigation buttons '<<', '<', '1', '>', and '>>', and 'out of 41'. Below this, the query result is displayed: 'Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact ?'.

USING SWITCHBOARD

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

Input Selection

Enter your text here.

Choose a sample input:

Preview of Sample Input

OR

OR

Upload a file:

Pisa_Italian_Days_and_Ways.txt

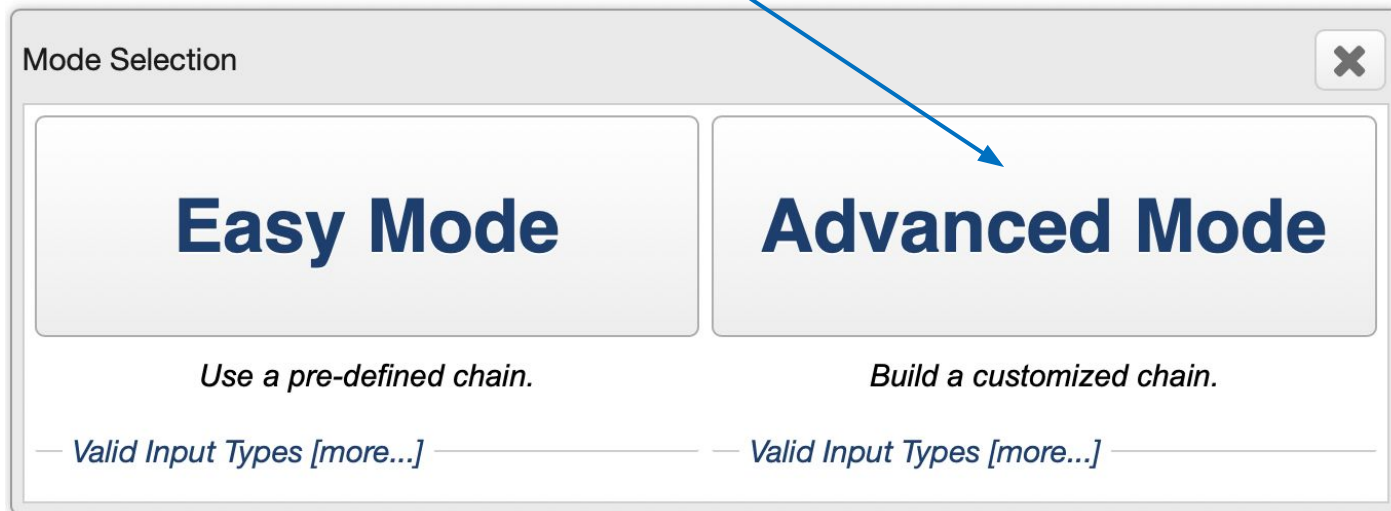
— Valid Input Types [more...] —

Document Type: Plain Text

Language: English

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



Next Choices (Double-click on an icon to add it to the chain)


SFS: Morphology analysis service Morphology: nupos 	SFS: OpenNLP Named Entity Named Entities: OpenNLP 	SFS: Illinois Named Entity Named Entities: CoNLL-2002 	CLAR: TextCorpus2Lexicon Language: English Document Type: Lexicon Format TCF Version: 0.4 entries.type: types 	SFS: POS Tagger - OpenNLP Part of Speech: Penn Treebank Te 	SFS: Charniak Parser +POS Part of Speech: Penn Treebank Te Parsing: Penn Treebank Tagset 	IMS: Constituent Parser Parsing: Penn Treebank Tagset 	SFS: Berkeley Parser - Berkeley Part of Speech: Penn Treebank Te Parsing: Penn Treebank Tagset 
IMS: TreeTagger Part of Speech: Penn Treebank Te Lemmas	SFS: Jitar POS Tagger Part of Speech: Penn Treebank Te	SFS: Constituent Parser - Part of Speech: Penn Treebank Te Parsing: Penn Treebank Tagset	SFS: TCF Text cleanup Language: English Document Type: TCF TCF Version: 5 Text	***	***		

Input and Chain Selection

Title [Plain Text]
Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear ...

SFS: To TCF Converter
TCF Version: 5
Language: English
Document Type: TCF
Text
 

SFS: Stanford Tokenizer
Sentences
Tokens
 



Run Tools Clear Results Download chain

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

The screenshot displays the 'Input and Chain Selection' interface. On the left, there is an input field with the text: "Do you realize that your letter in answer to mine of March 18th from Rome was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear ...". Below the input field are navigation icons: a blue double-headed arrow, a blue download icon, a blue home icon, a blue information icon, and a red 'X' icon.

The main area contains a horizontal chain of tool modules, each with its own set of navigation icons at the bottom. From left to right, the modules are:

- SfS: To TCF Converter**: TCF Version (5), Language: English, Document Type: TCF Text.
- SfS: Stanford Tokenizer**: Sentences, Tokens.
- SfS: Jitar POS Tagger**: Part of Speech: Penn Treebank Te
- SfS: MorphAdorner Lemmatizer**: Lemmas.
- SfS: Illinois Named Entity**: Named Entities: CoNLL-2002.
- SfS: Geolocation**: Geo - Capitals: Name, Geo - Continents: Name, Geo - Coordinates: Decimal Degre, Geo - Countries: 2-Letter Country.
- SfS: Lemma Frequencies**: out (tsv).

At the top right of the interface are three buttons: "Run Tools", "Clear Results", and "Download chain". Two blue arrows originate from the text in the list above: one points from "Clear Results" to the "Clear Results" button, and the other points from "click on the red 'x'" to the red 'X' icon in the bottom right corner of the "SfS: Lemma Frequencies" module.

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

The screenshot displays the NLPHub interface with several key components:

- Language selection:** A dropdown menu currently set to "English".
- Input text:** A section with instructions: "Drag a .TXT file on the Upload box, or select a file from your PC, or paste a text." It includes an "UPLOAD" button, an upload icon, and a "CANCEL" button. A message indicates a file named "Pisa_Italian_Days_and_Ways.txt" has been uploaded.
- Text area:** A large text input field with the placeholder "paste your text here" and a character limit of "(max 4000 characters)".
- Annotations:** A grid of checkboxes for various entity types, all of which are checked (indicated by green checkmarks):
 - Keyword, Date, Location, Organization
 - Person, Money, Percentage, Time
 - Sentence, Token, Event, Taxon
 - Asfa
- ANALYSE:** A green button at the bottom center to execute the analysis.

Blue arrows from the text above point to the language selection dropdown, the input text area, the annotations grid, and the ANALYSE button.

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

NER

You can download the overall result as a JSON file [here](#) Location occurs 19 times.

Do you realize that your letter in answer to mine of March 18th from **Rome** was not quite within the pact? I found it awaiting me at Siena, with a number of others. I thought my explanation quite clear and eminently sane, but you seem to have strangely perverted my meaning; then you revert to an earlier letter from **La Cava**, and are pleased to imagine that we are taking risks all the time and leading a reckless life generally. I shall really hesitate to tell you again of any of our adventures such as that drive home from **Pæstum**, which I merely related as an amusing incident. There is no danger of brigands in these days and we did not "need a protector," especially as kind **Providence** looked after us. That drunken driver would not have surrendered his reins to you or to any one except the padrone; and then "all's well that ends well," and we returned from our excursion with nothing worse than a grievance. I was so vexed with you for two whole days that I wrote you not one line from Siena or Pisa. Now your indiscretion is partially atoned for by a letter which has just reached me here, and I am trying to forgive you and "be friends again," as we used to say when we were children. But the charms of Siena are already so eclipsed by those of **Florence** that it is quite impossible for me to give you an atmospheric description of its streets and churches, above all of the shining cathedral, rich from dome to pavement with colored marbles, frescoes, and mosaics. This may be no loss to you, who are doubtless well tired of my Italian rhapsodies; but your respite is only temporary, as I quite missed writing you that letter. I wanted to tell you that the campanile at Pisa leans quite as much as the little Parian model on your desk, and about the famous Campo Santo with its interesting paintings, and many other things. The habit of relieving my mind of the burden of surplus impressions, or of what I might call my "oversoul," has become second nature. Do you remember, Allan, the man in **Frank Stockton's** story who, on his return from abroad, found his friends and acquaintances so much interested in their own affairs that he

Algorithms

Annotations

- Keyword
- Date
- Location
- Organization
- Person
- Money
- Percentage
- Time
- Sentence
- Token
- Event
- Taxon
- Asfa

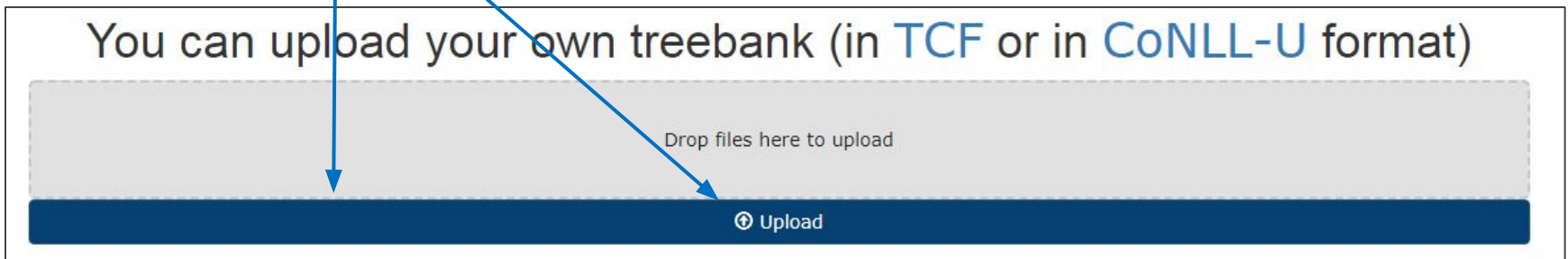
USING TüNDRA

- **TüNDRA** can be used also outside WebLicht to query treebanks:
 - CON: the interface is not updated with the last version of Universal Dependencies treebanks
 - Alternatives are PML Tree Query (<http://lindat.mff.cuni.cz/services/pmltq/>) and Grew-match (<http://universal.grew.fr/>): please note that each interface has its own query language
 - PRO: you can upload treebanks saved in CoNLL-U files

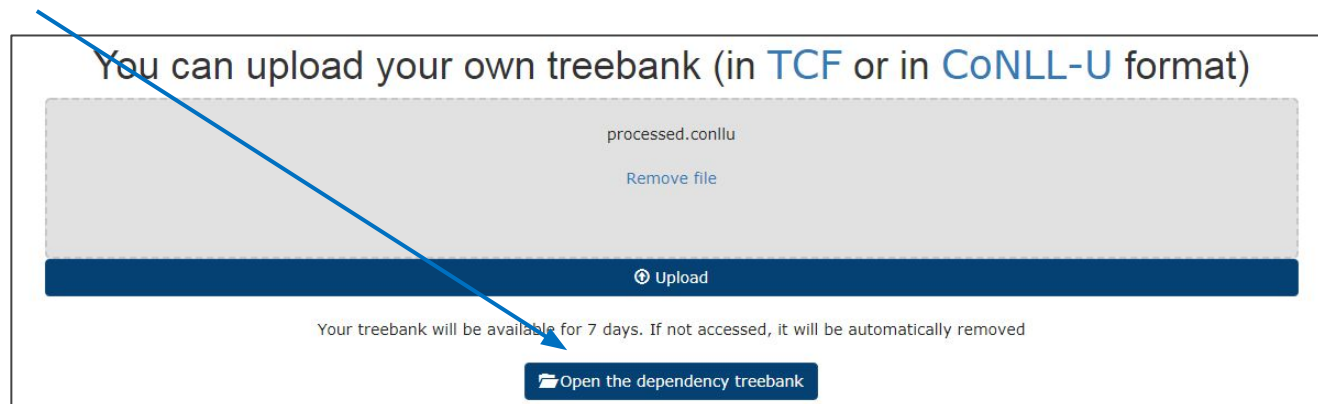
USING TüNDRA

- Upload your treebank: use the CoNLL-U file produced with UDPipe

- 1) Drop a file or click in the grey area to upload a treebank and then click on “Upload”

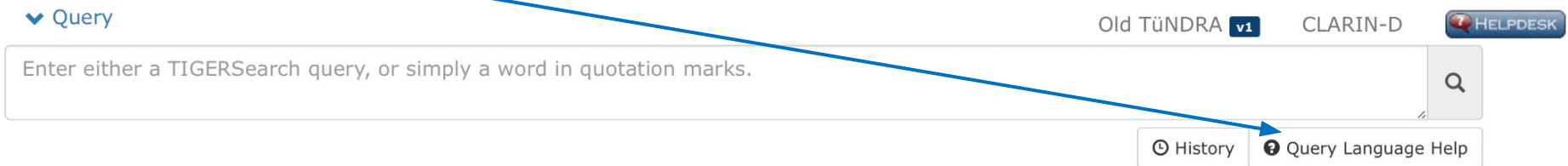


- 2) Once the upload is finished, click on “Open the dependency treebank”



USING TüNDRA

3) Use the text area for querying the treebank: click on “Query Language Help” for a detailed tutorial

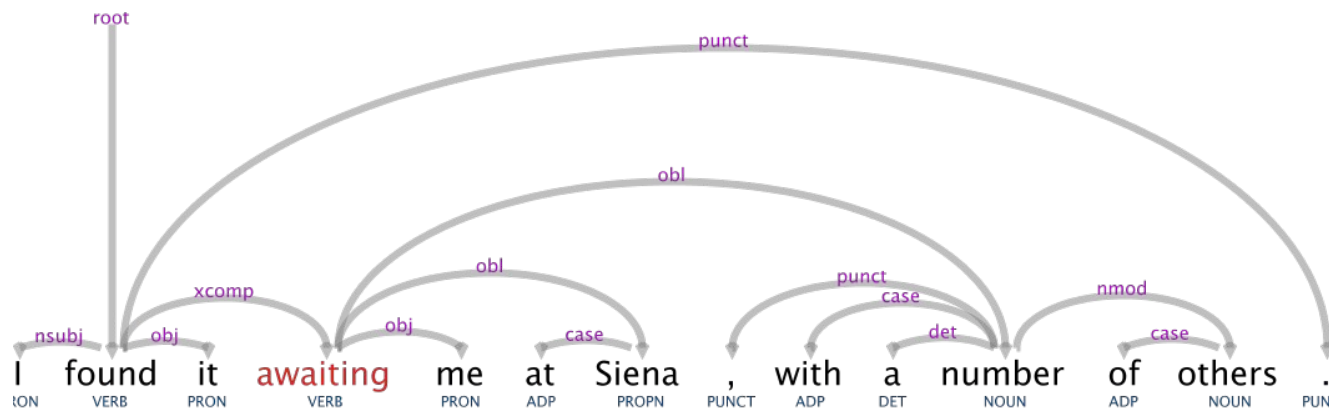
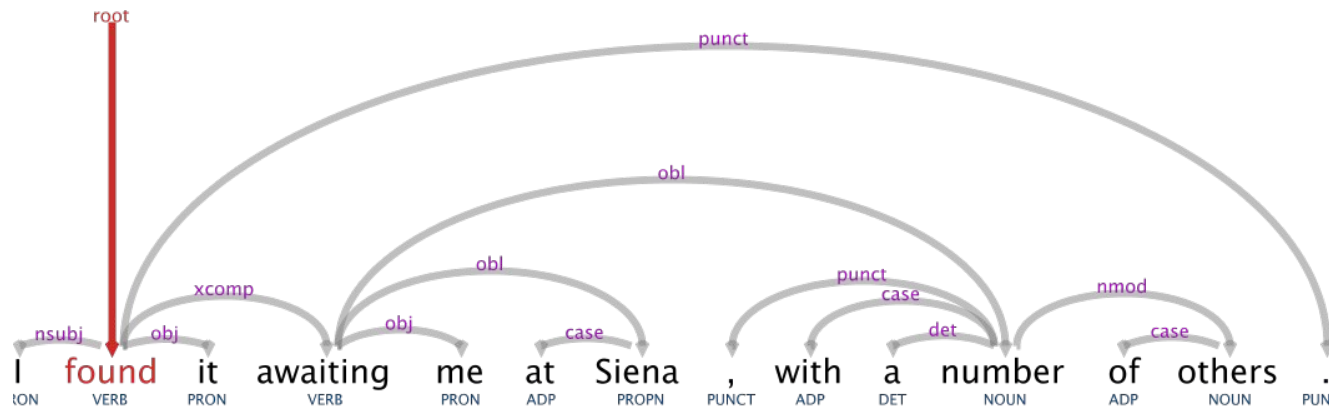


Examples:

- [pos="VERB"]: retrieves all verbs
- #1:[pos="VERB"] & #1:[morph tense="Past"]: retrieves all past tense verbs
- [lemma="say"]: retrieves the occurrences of the lemma “say”
- [token=/[Nn]ow/]: retrieves all tokens that are “Now” or “now”
- [pos="NOUN"] >amod [pos="ADJ"]: retrieves adjectival modifier relations between a noun and an adjective

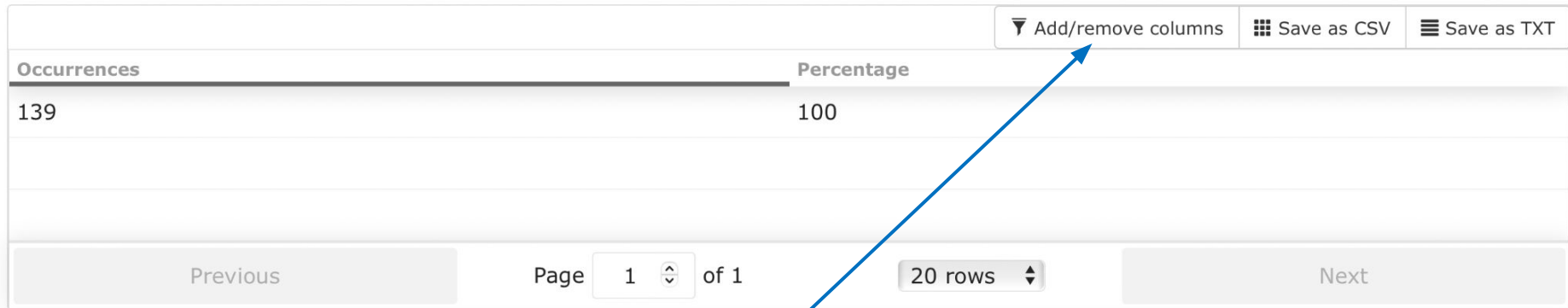
USING TüNDRA

4) Explore the graphical visualization for each sentence: the result of the query is highlighted in red. For example, for the query [pos="VERB"], each verb of each sentence is highlighted in red



USING TüNDRA

4) The section “Statistics” appears only after running a query. The following image shows that the uploaded treebank has 139 tokens annotated as VERB



Occurrences	Percentage
139	100

Previous Page 1 of 1 20 rows Next

Add/remove columns Save as CSV Save as TXT

5) Click on “Add/remove columns” to refine the available statistics

USING TüNDRA

6) Select the information you want to add to the table (e.g. lemma) then “Apply” and “OK” and check the new statistics

Columns can be sorted clicking on the header

Add/Remove Columns ×

Variable _query:

- edge
- pos
- lemma
- text
- token
- morphverbform
- morphmood
- morphtense
- morphnumber
- morphperson

Apply

Ok

Close

_query: lemma	Occurrences	Percentage
tell	5	3.597
say	4	2.878
reach	3	2.158
look	3	2.158

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. Extract 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: set tab as delimiter
- Filter the columns so to have only LOC

FROM NLP TO VISUALIZATION

3. Use DARIAH-DE Geo-Browser

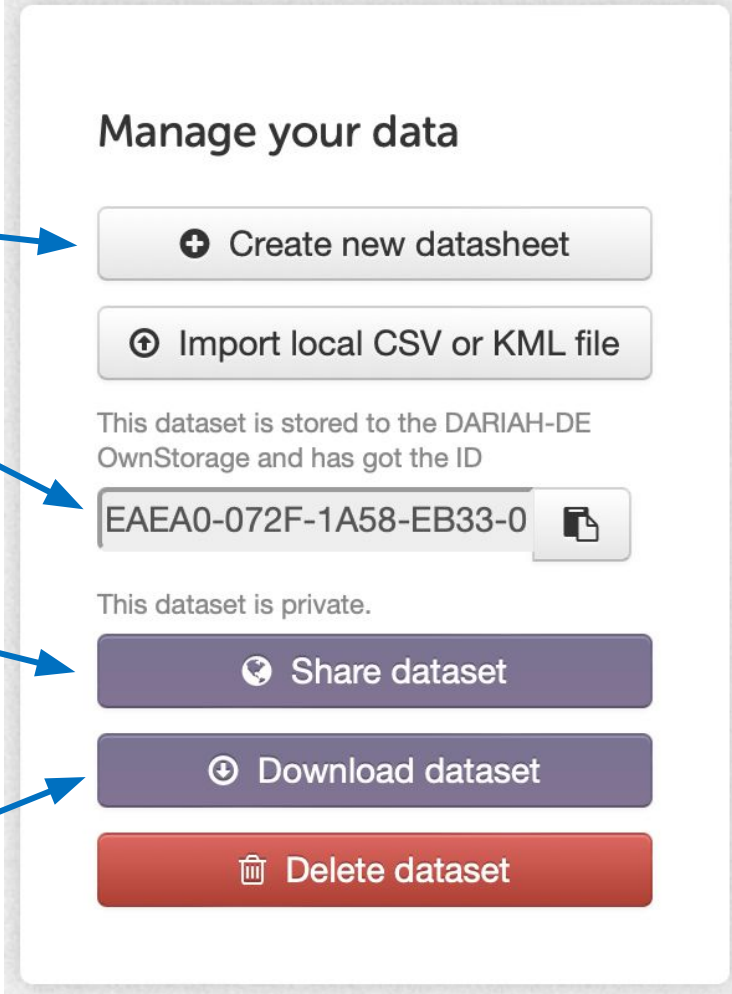
2 free services offered by the German branch of DARIAH (Digital Research Infrastructure for the Arts and Humanities):

- 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. Geolocate 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



The screenshot displays the 'Manage your data' section of the DARIAH-DE Datasheet Editor. It features a vertical list of actions: 'Create new datasheet' (light grey button with a plus icon), 'Import local CSV or KML file' (light grey button with a plus icon), a text box containing the dataset ID 'EAEA0-072F-1A58-EB33-0' with a copy icon, 'Share dataset' (dark purple button with a globe icon), 'Download dataset' (dark purple button with a plus icon), and 'Delete dataset' (red button with a trash icon). The text 'This dataset is stored to the DARIAH-DE OwnStorage and has got the ID' is positioned above the ID field, and 'This dataset is private.' is positioned above the 'Share dataset' button.

FROM NLP TO VISUALIZATION

4. Geolocate LOC using DARIAH-DE Datasheet Editor
- the spreadsheet can be filled in directly on the browser
 - the only required column is Address, it cannot be empty: we paste there 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. Add geocoordinates

- 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

Add geocoordinates

Geocoordinates for places from the **Address** field can be added with the Getty Thesaurus Service.

 Geolocation completion

Place selection

Geolocation can be based either on the *Getty Thesaurus of Geographic Names (TGN)*, *OpenStreetMap (OSM)*, or *GeoNames*. Choose which geolocator to use:

TGN


TGN search can be based on column *Address* or on column *GettyID* in your datasheet. Choose which one to use:

Address



The coordinates for the first place found are automatically added to the sheet if no existing entries are available. If there are multiple results, you may have to adjust the places manually!

- ✎ sets all places in the sheet with the highlighted address to the coordinates of the chosen place. This will overwrite existing entries!
- 👁 shows the chosen place in the *Map selection* and allows you to correct and refine coordinates or set unrecognized places.

Pæstum

No results, please try map selection.  


Campo Santo

Campo Santo: Salta | Argentina | South A  

Pisa

Pisa: Tuscany | Italy | Europe | World  

Siena

Siena: Tuscany | Italy | Europe | World  

Map selection

GettyID / Longitude / Latitude

GettyID Longitude Latitude
Enter exact place name as in "Address" field... 

Note: 'Set' will overwrite coordinates for all table rows matching this address unless you only select specific rows or cells first.



Map search

Search the Getty Thesaurus of Geographical Names, OpenStreetMap, or GeoNames to fill your map selection:

Enter your TGN/OSM/GeoNames search term here...

 TGN

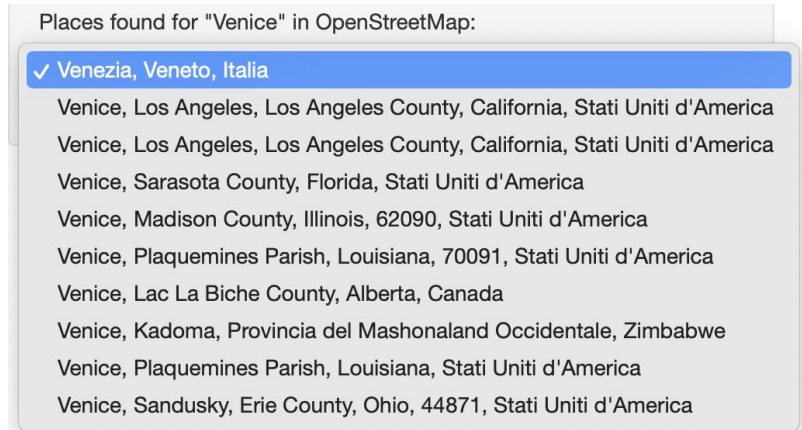
 OSM

 GeoNames

FROM NLP TO VISUALIZATION

5. Correct geolocation

- if the automatic geolocation is NOT correct, choose another option from the drop-down menu in the "Place selection" section
- 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. Correct 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

Map selection

GettyID / Longitude / Latitude

GettyID	12.3345898	45.4371908
Venice		

*Note: 'Set' will overwrite coordinates for **all** table rows matching this address unless you only select specific rows or cells first.*

Open with Geo-Browser

Your dataset will be validated and sent to the **Geo-Browser**.





Thank you!

Email: rachele.sprugnoli@unipr.it

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)