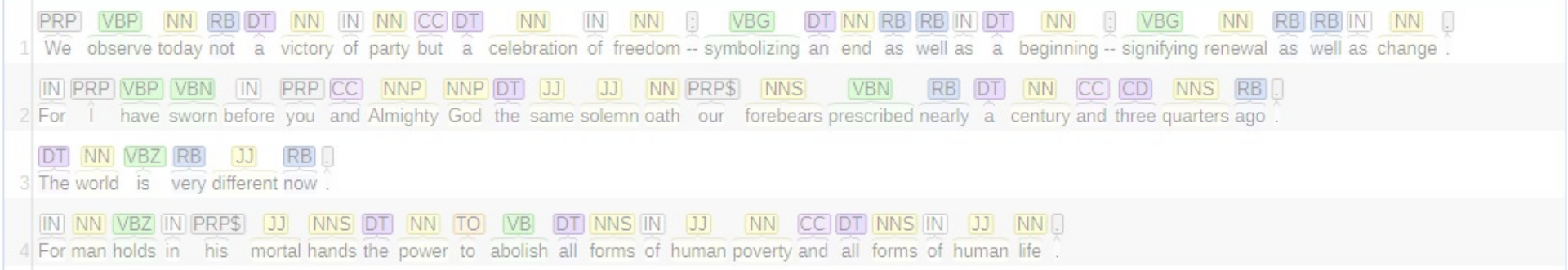


Natural Language Processing

Rachele Sprugnoli
sprugnoli@fbk.eu

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

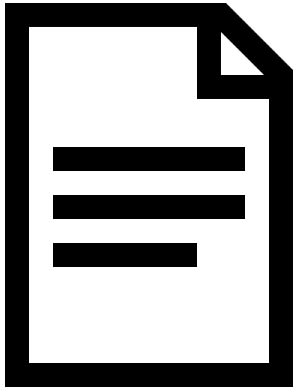
Natural Language Processing (NLP)

- **Goal**: create machines that understand natural languages

- Encompasses many disciplines:
  - linguistics
  - computer science
  - artificial intelligence
  - statistics
  - machine learning

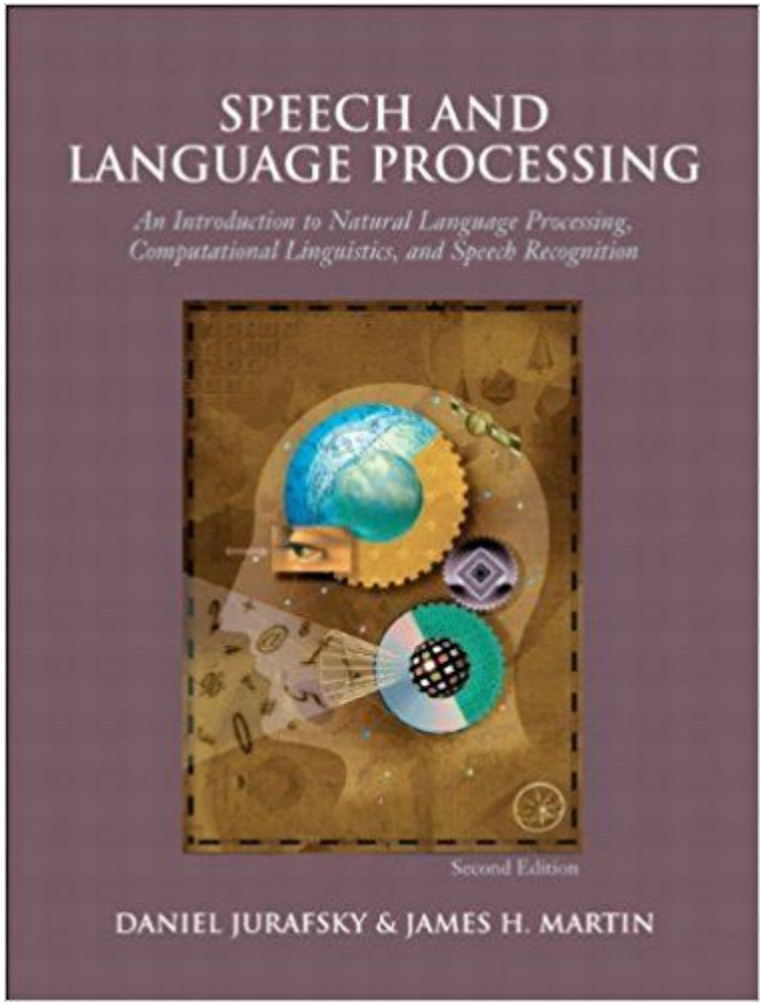
Natural Language Processing

- Oral History and Technology: http://oralhistory.eu/



- “Natural Language Processing for Historical Texts” by Michael Piotrowski

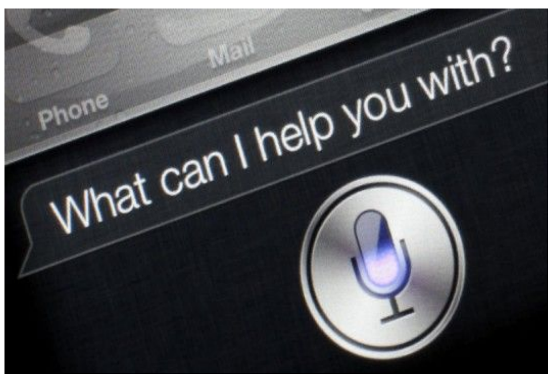
What to read



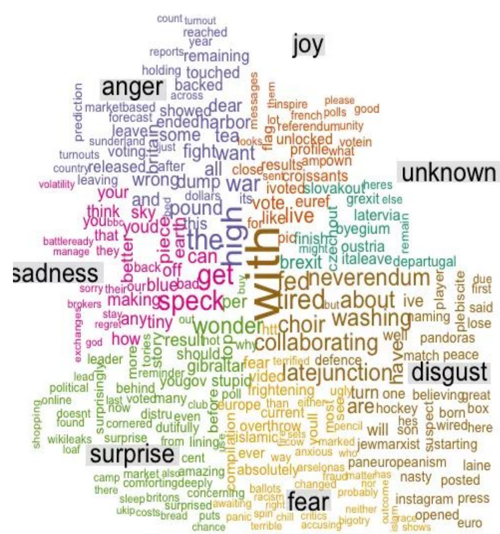
https://web.stanford.edu/~jurafsky/slp3/

Why NLP is interesting?

- Languages involve many human activities
- NLP is used to acquire insights from massive amount of textual/speech data
- NLP has many applications



Google Translate



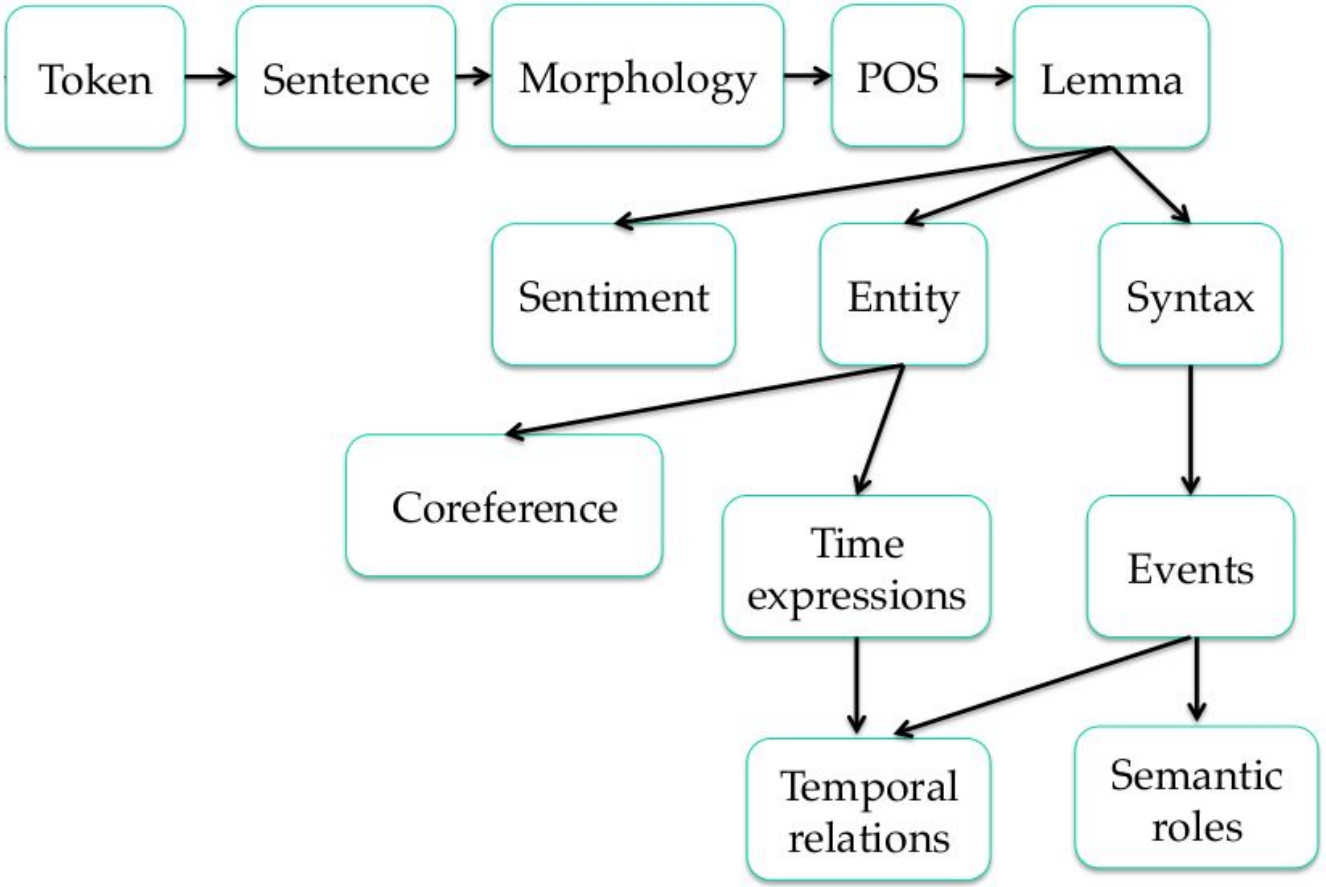
Why NLP is hard but also fun?

- Ambiguity: *Flying planes can be dangerous*
- Non-standard language: *#GreatJob @justinbieber!*
- Old language: *I have thee not, and yet I see thee still*
- Neologisms: *chilax*
- Idioms: *lose face*
- World knowledge: *Mary and Sue are sisters*
- Tricky named entities: *Let It Be was released in 1970*

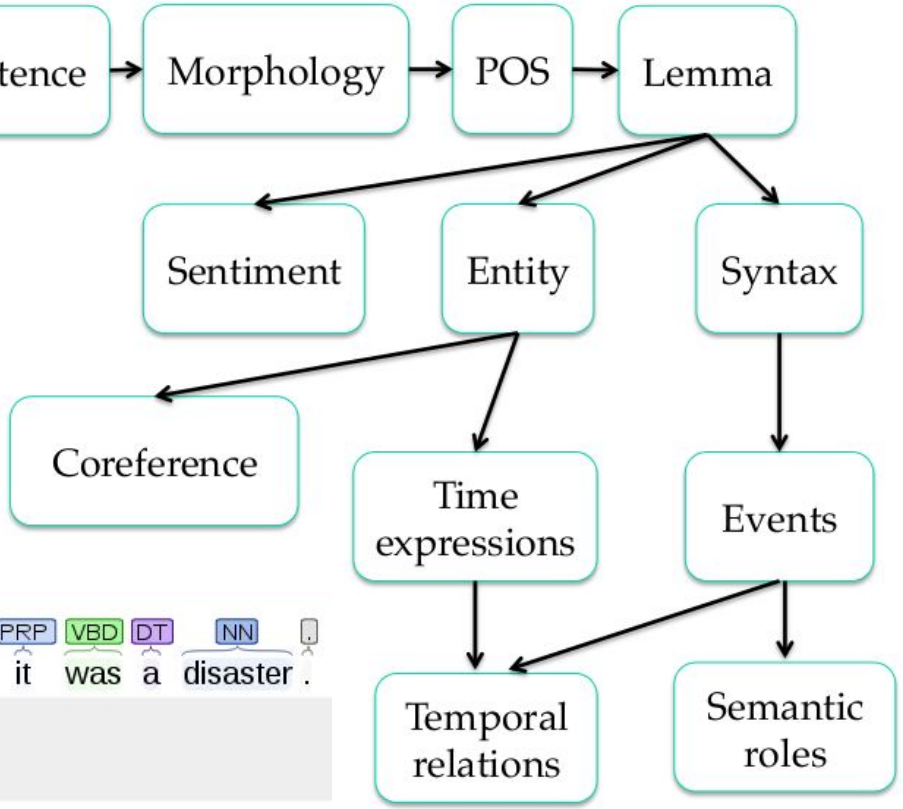
Where do people do research in NLP?

How to process the language?

- Example of a PIPELINE

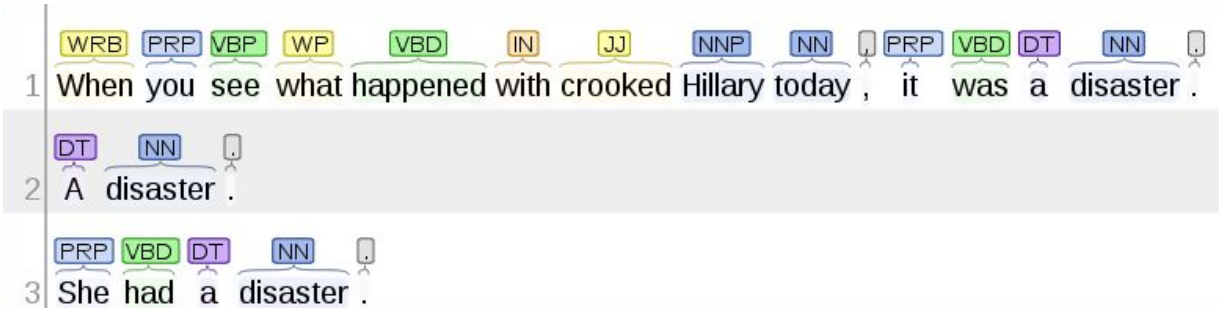


How to process the language?

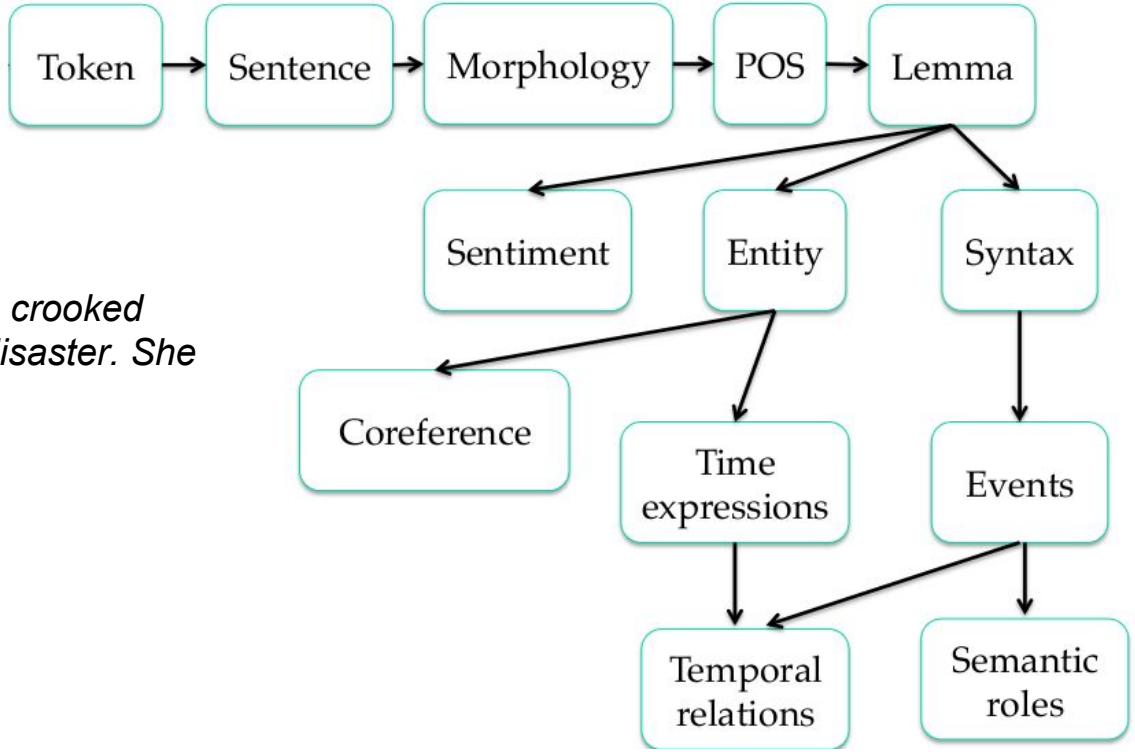


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TOKEN - SENTENCE - PART OF SPEECH

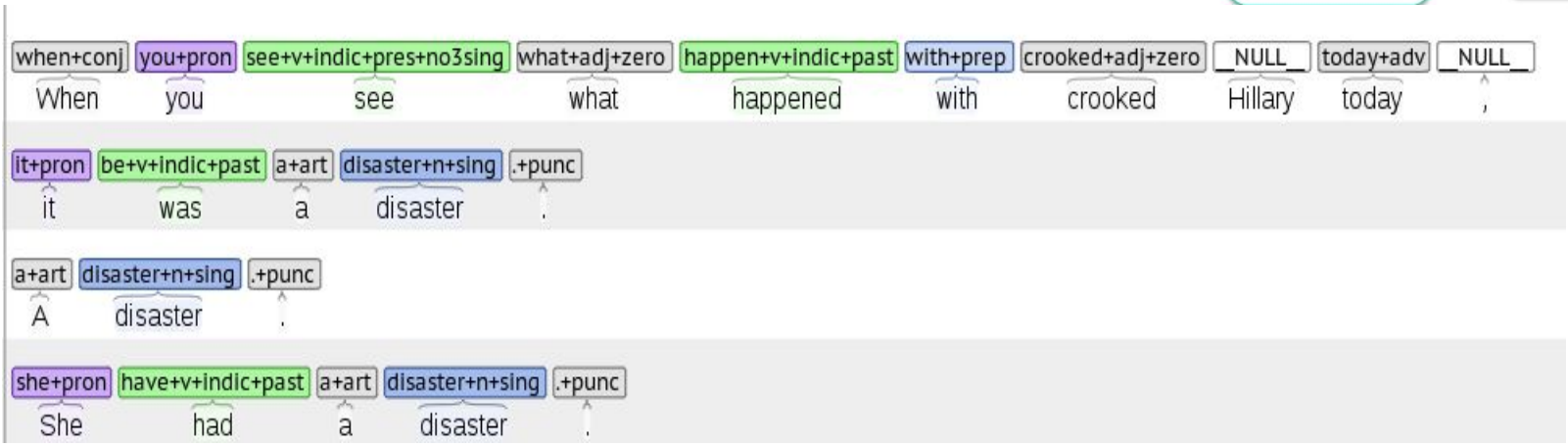


How to process the language?

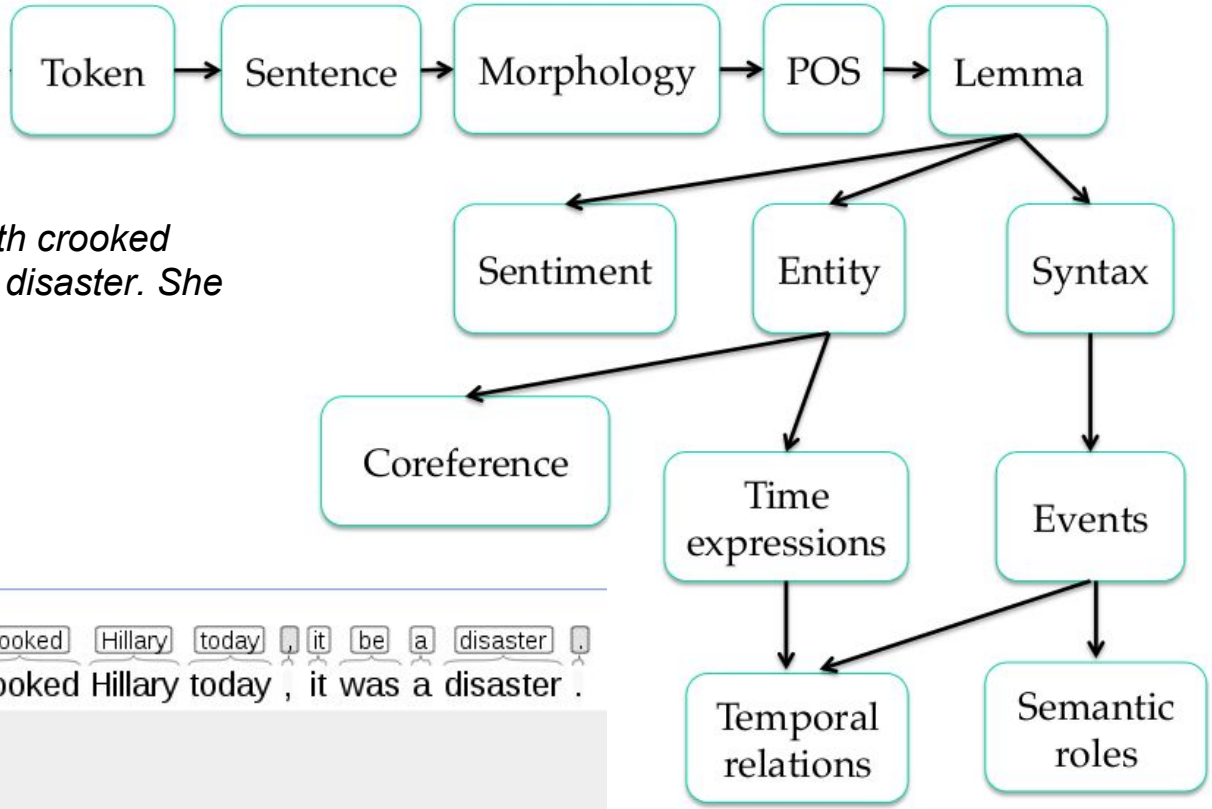


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

MORPHOLOGY



How to process the language?



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

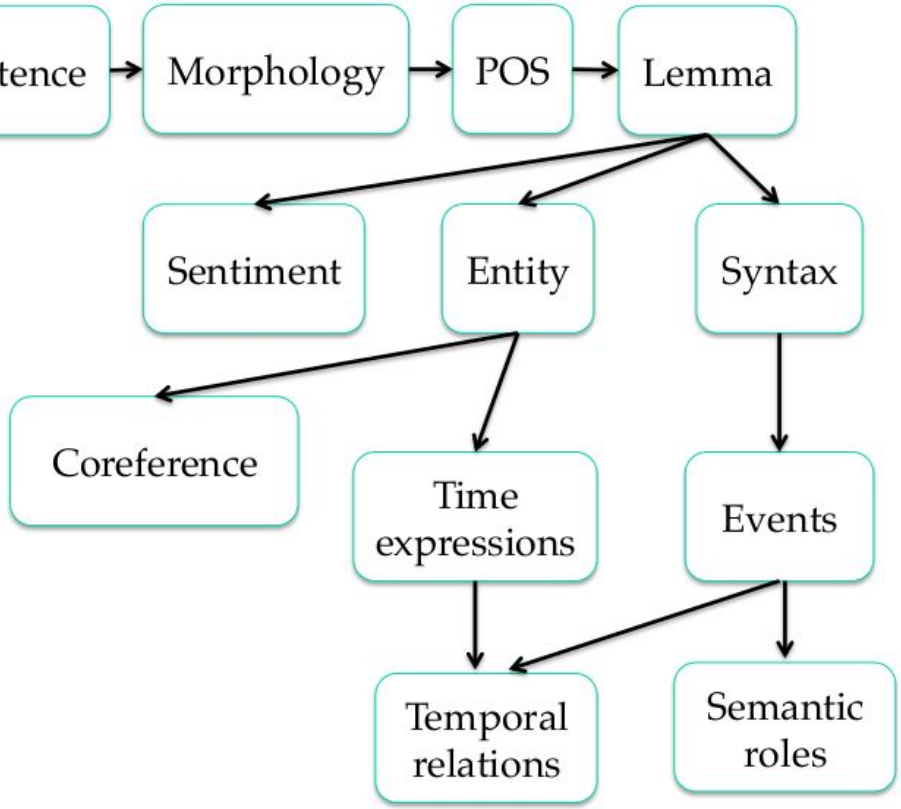
LEMMA

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 .

a disaster .
A disaster .

she have a disaster .
She had a disaster .

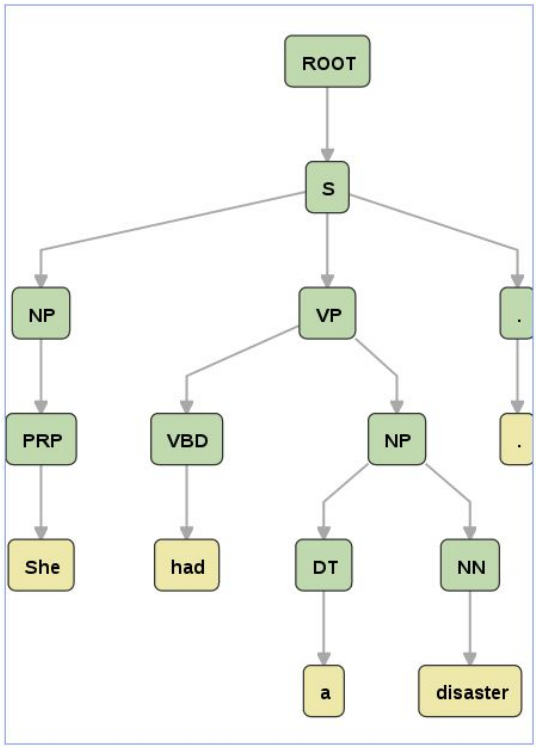
How to process the language?



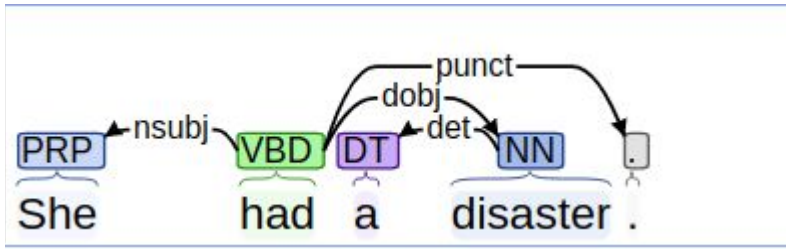
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX
-
PARSING

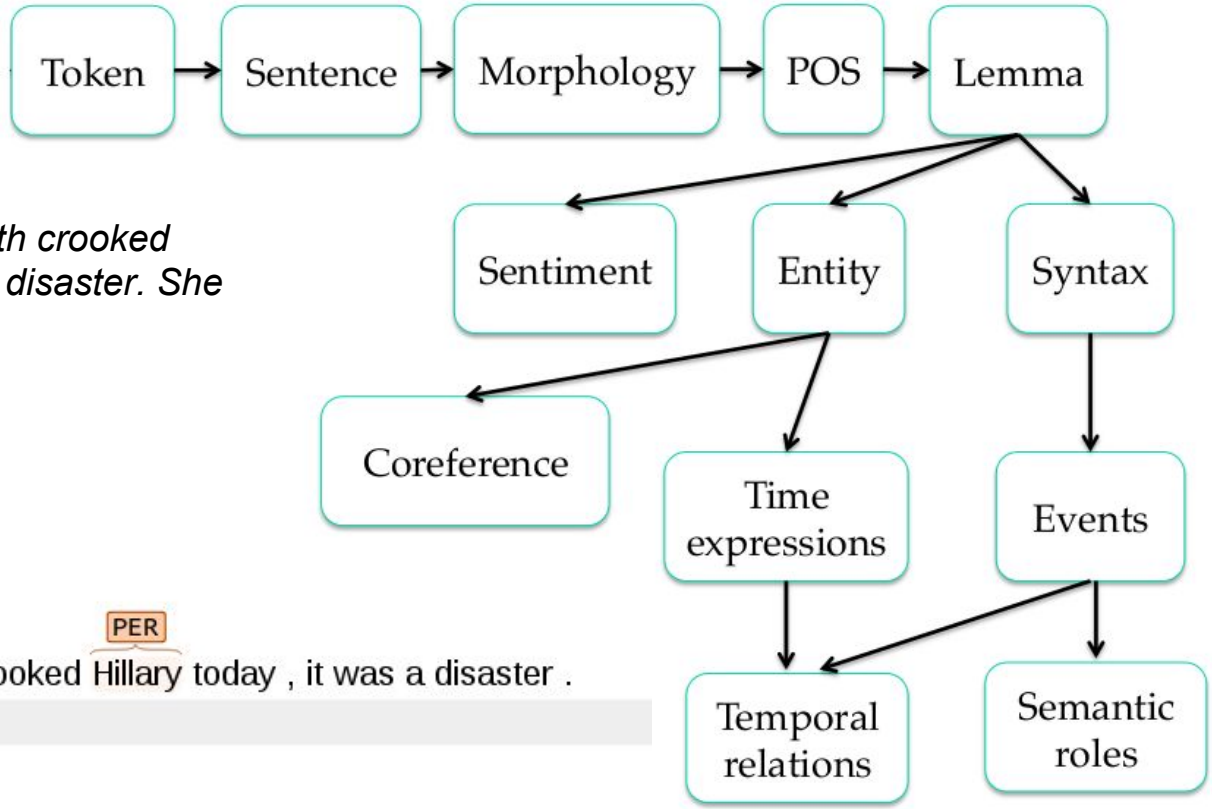
CONSTITUENCY PARSING



DEPENDENCY PARSING



How to process the language?

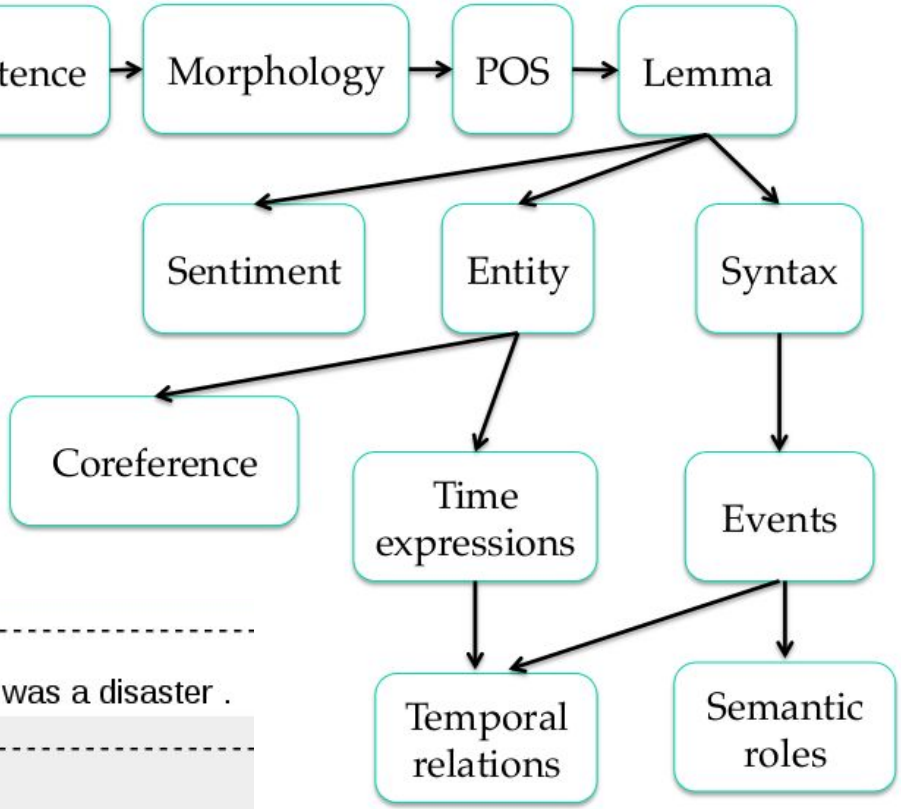


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

ENTITY

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

How to process the language?



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

COREFERENCE

Mention

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

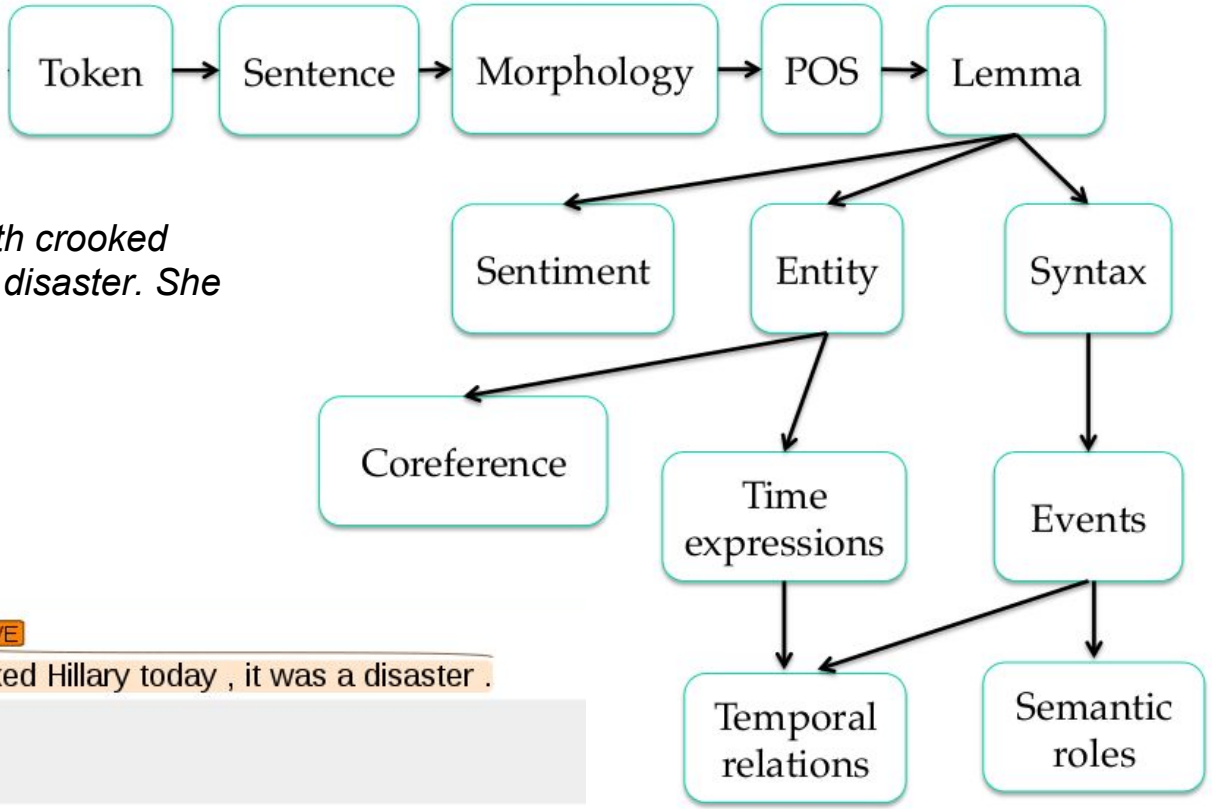
A disaster .

---coref---

Mention

She had a disaster .

How to process the language?



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

SENTIMENT

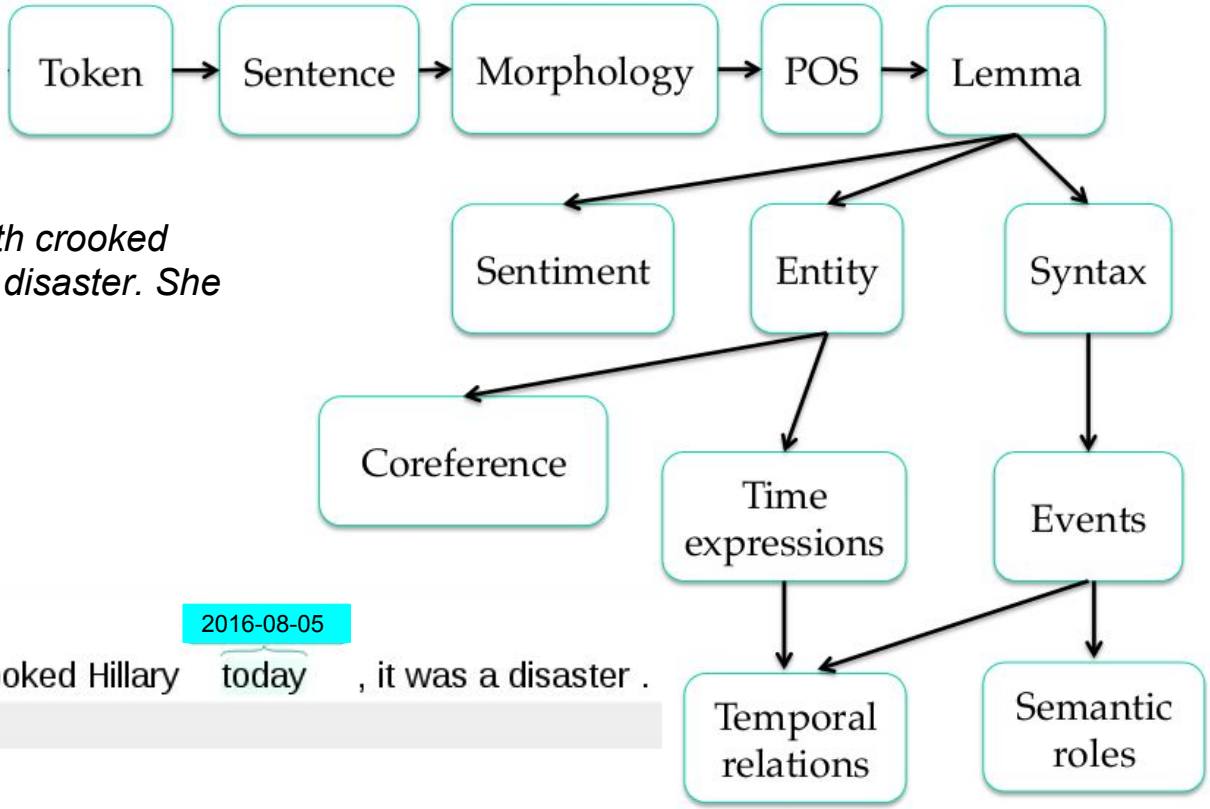
NEGATIVE
When you see what happened with crooked Hillary today , it was a disaster .

VERY NEGATIVE
A disaster .

NEGATIVE
She had a disaster .

Another example: http://www.depechemood.eu/

How to process the language?



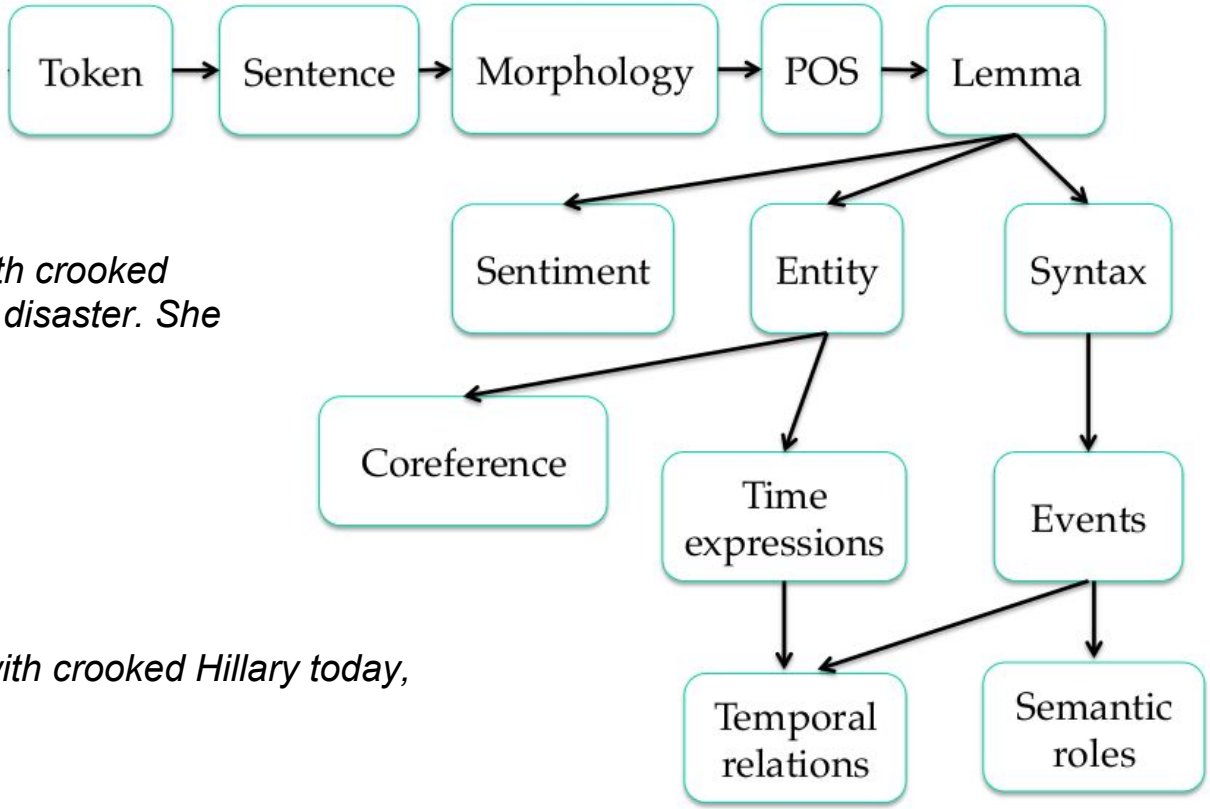
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TIME EXPRESSIONS

2016-08-05

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

How to process the language?

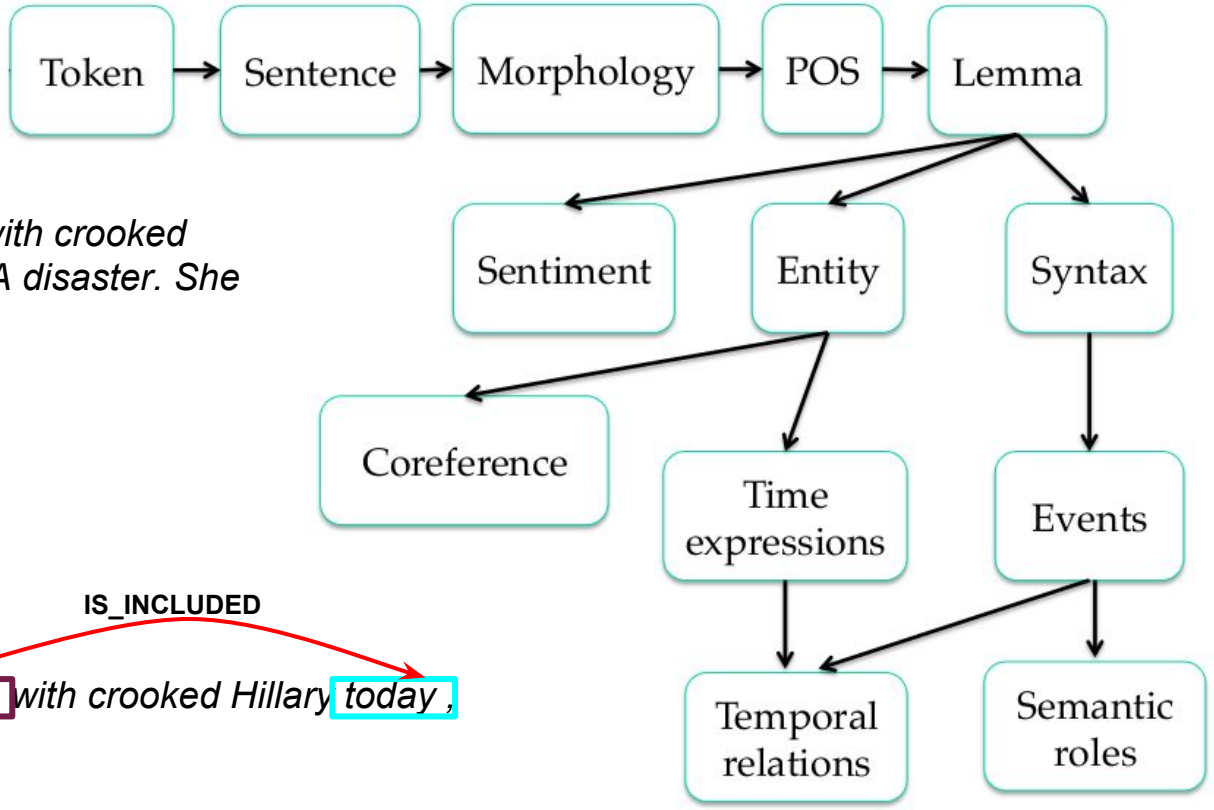


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

EVENTS

PERCEPTION OCCURRENCE
When you see what happened with crooked Hillary today,
STATE
it was a disaster.

How to process the language?



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

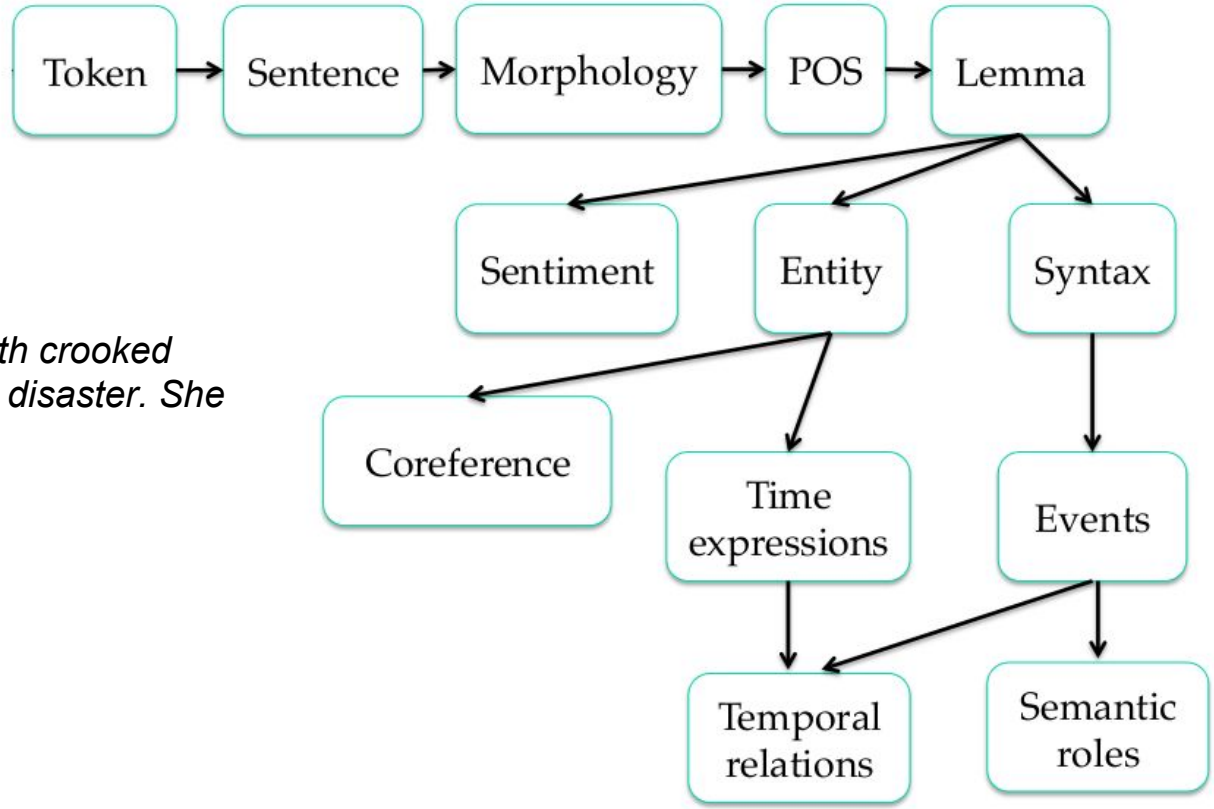
TEMPORAL RELATIONS

IS_INCLUDED

When you see what happened with crooked Hillary today ,

it was a disaster.

How to process the language?



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

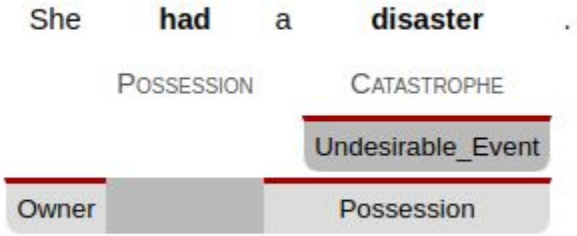
SEMANTIC ROLES

- FrameNet approach

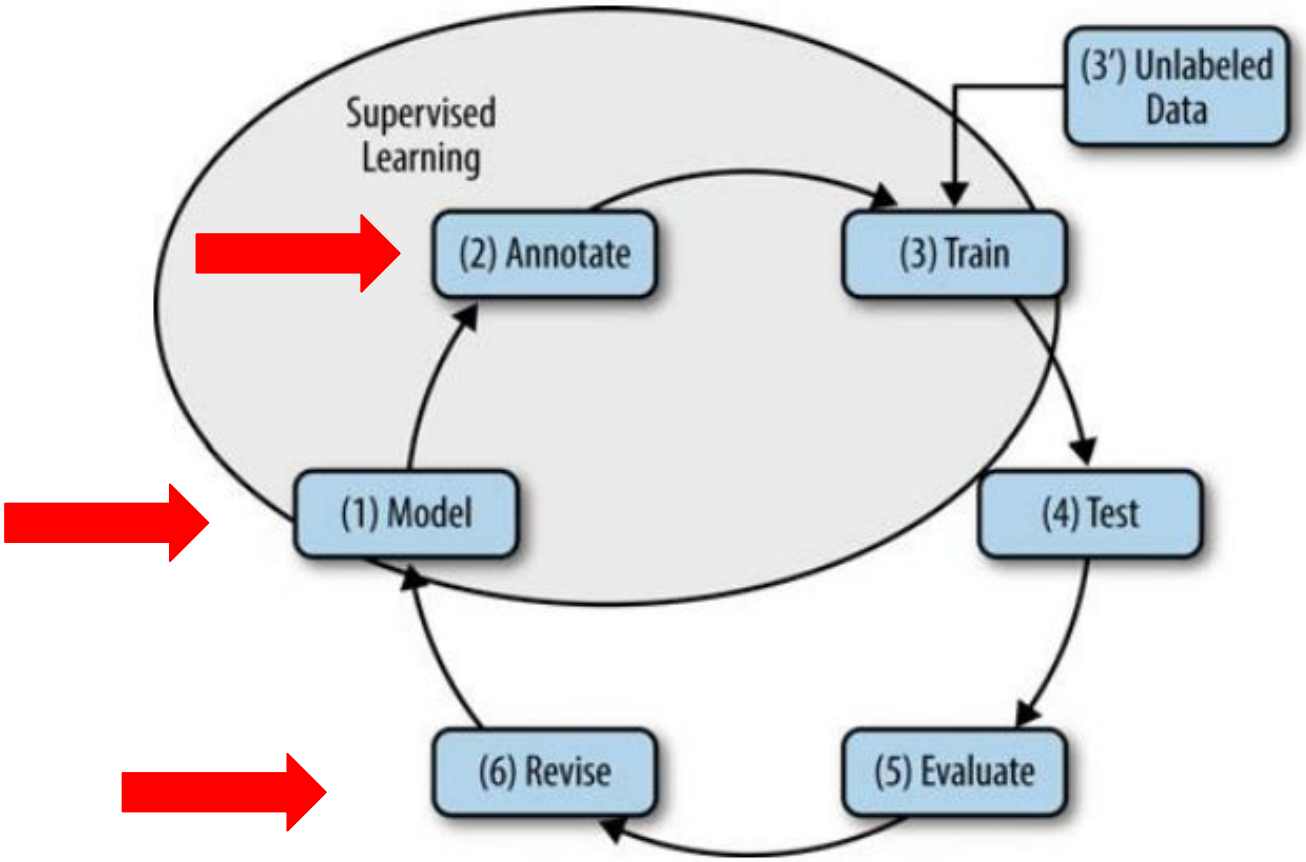
Possession

Definition:

An Owner has (or lacks) a Possession.



How to create an NLP module?



From model to annotated data to automatic systems

(Pustejovsky and Stubbs, 2012)

Applications

- NLP + Humanities:

1. ALCIDE
2. RAMBLE ON
3. Historical Travel Writings

1) The ALCIDE Platform

ALCIDE: *Analysis of Language and Content In a Digital Environment*

ISIG
ISTITUTO STORICO ITALO-GERMANICO
ITALIENISCH-DEUTSCHES HISTORISCHES INSTITUT

- **General goal**: give means to investigate who, where, when, what and how in large Humanities corpora

http://celct.fbk.eu:8080/Alcide_Demo/

- **NLP modules**: tokenization, sentence splitting, PoS tagging, lemmatization, NER

2) RAMBLE ON

- Tracing Movements of Popular Historical Figures

“Perhaps people will soon be persuaded that there is no patriotic art and no patriotic science. Both belong, like everything good, to the whole world and can be promoted only through general, ***free interaction among all who live at the same time****.”*

Goethe, 1826

- **NLP modules:** tokenization, sentence splitting, PoS tagging, lemmatization, NER, temporal expressions detection, semantic role labelling

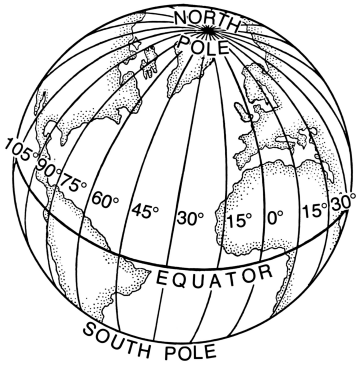
2) RAMBLE ON

- Our approach: automatic extraction of motion trajectories from Wikipedia biographies

In 1914, the couple separated; **Einstein** moved to Berlin

He settled in the U.S., becoming an American citizen in 1940

In 1946 **Einstein** visited Lincoln University



http://dh.fbk.eu/technologies/rambleon

3) Historical Travel Writings

- **NLP modules**: NER (only locations)

https://sites.google.com/view/travelwritingsonitaly/

Conclusions of the First Part

- In “traditional” NLP research, final users are not in the loop. In Digital Humanities, final users (humanities scholars) play a central role

- Friendly suggestions:
  - Be curious
  - Stay updated
  - Look for interdisciplinary collaborations

Stay hungry, stay foolish.
-Steve Jobs

Hands-on Session

- TODAY
  - Topic modeling
  - Key-concept extraction
  - CoreNLP

DOWNLOAD THE SHARED FOLDER:

https://drive.google.com/drive/folders/1hxHpmFKhhhthmBYecN7aRp6-EPaDSd6N?usp=sharing

Topic Modeling

- What topics a corpus of documents contain?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

Topic Modeling

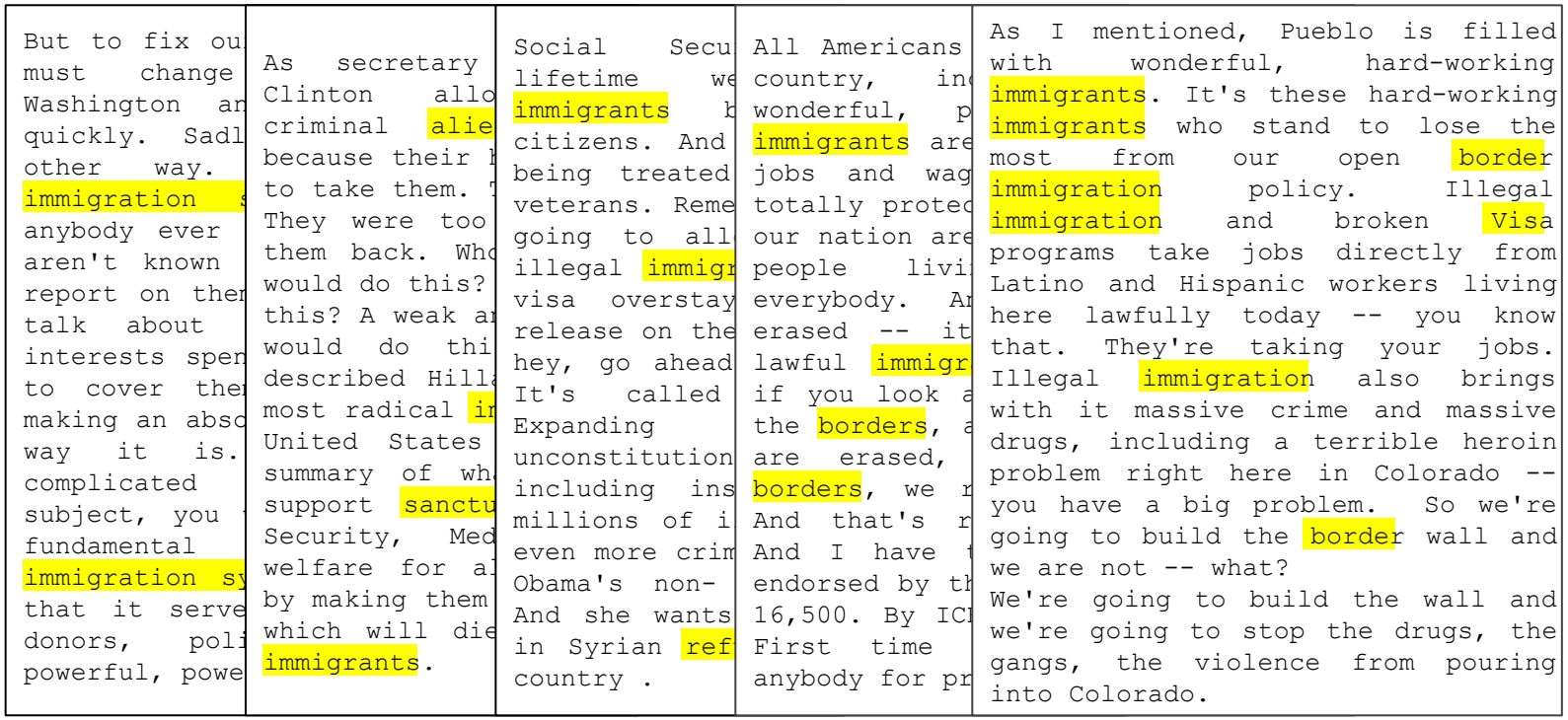
- What topics a corpus of documents contain?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

- IMMIGRATION
- POLITICS
- ECONOMY

Topic Modeling



"That's how topic modeling works in practice. You assign words to topics randomly and then just keep improving the model, to make your guess more internally consistent, until the model reaches an equilibrium that is as consistent as the collection allows."

Ted Underwood, 2012

Topic Modeling : Pros and Cons

“Essentially, all models are wrong, but some are useful.”

George Box, 1987

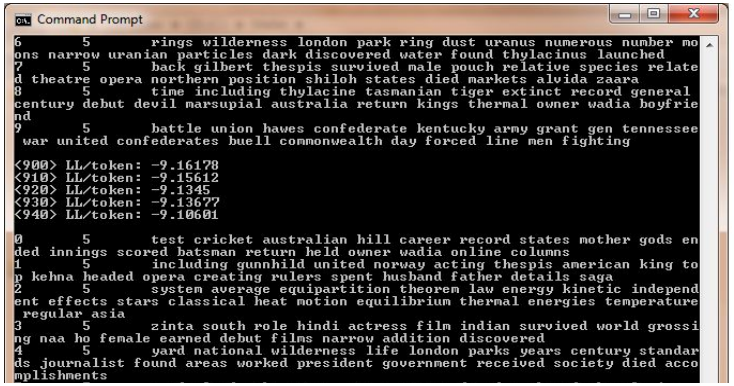
- No easy method to evaluate the output
- No way to automatically determine the best number of topics for a corpus
- Too ambiguous and configurable

- Good starting point to explore data
- Generates new way of looking to big amount of texts

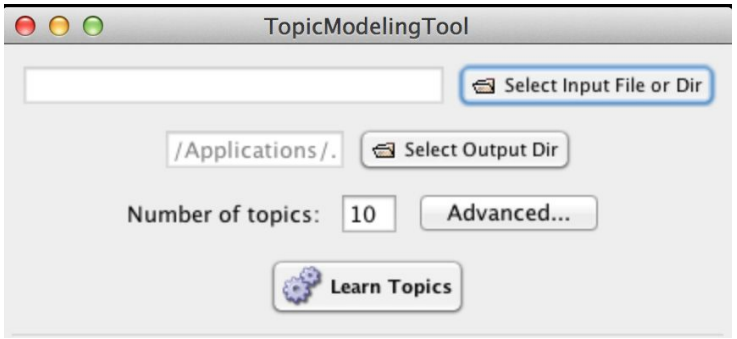
Topic Modeling : Tools

- Online Demo: https://mimno.infosci.cornell.edu/jsLDA/jslda.html

- MALLET: http://mallet.cs.umass.edu/



- Topic-modeling-tool:



Topic Modeling Online Demo

1. Download Promessi_Sposi.txt and stopwords-it.txt from the shared Google folder
2. Open Promessi_Sposi.txt with a text editor and look how it is formatted
3. Open stopwords-it.txt and check what it is inside the file
4. Open a browser (no Explorer) https://mimno.infosci.cornell.edu/jsLDA/jslda.html
5. Click on “Choose File” for the Document Upload option in the grey menu and upload Promessi_Sposi.txt
6. Click on “Choose File” for the Stoplist Upload option in the grey menu and upload stopwords-it.txt
7. Click on Load
8. Try to run several iterations (Click on Run 50 iterations)
9. Try to clean the Vocabulary

Topic-modeling-tool - 1

1. Download the Clinton-Trump Corpus.zip
2. Open the topic-modeling-tool
3. Windows: double click on the jar file
4. Mac: open the Terminal, go to the folder containing the tool, write
   java -jar TopicModelingTool.jar
5. Click on Select Input File or Dir and choose the Trump folder
6. Change the number of topics in **30**
7. Click on Learn Topics
8. Two new folders appeared in the folder containing the tool: output_csv and output_html
9. Rename them adding "Trump" to the file name: eg. output_csv_Trump
10. Do the same process for the Clinton folder

Topic-modeling-tool - 2

11. Open all_topics.html using a browser and check the topics
12. Select two topics in common between Trump and Clinton:
    e.g. **immigration / work / terrorism**

    Please note that topics may differ from person to person!

13. Click on the list of keywords associated to those topics; another file opens up

Keyphrase Extraction

- Keyphrases (or key-concepts) = n-grams capturing the main concepts of documents

But to fix our **immigration system**, we must change our **leadership in Washington** and we must change it quickly. Sadly, sadly there is no other way. The truth is our **immigration system** is worse than anybody ever realized. But the facts aren't known because the **media** won't report on them. The **politicians** won't talk about them and the special interests spend a **lot of money** trying to cover them up because they are making an absolute **fortune**. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the **immigration system** in our **country** is that it serves the needs of **wealthy donors**, **political activists** and powerful, **powerful politicians**.

Both single words and multi-token expressions

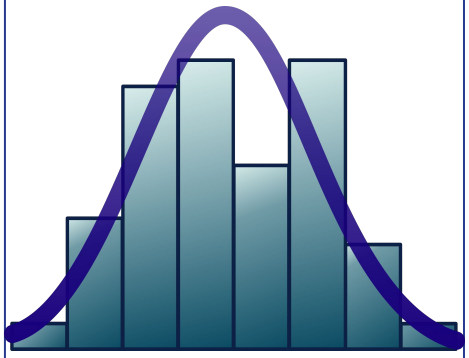
Both single documents and whole corpora

Keyphrase Extraction: Tool

- KD = Keyphrase Digger

http://celct.fbk.eu:8080/KD_KeyDigger/

STATISTICAL MEASURES



+

LINGUISTICS INFORMATION

=

WEIGHTED LIST OF KEYPHRASES

rank	keyword synonyms	frequency	score
1	kd	33	43.07
2	types of texts	2	23.24
3	kx	17	22.36
4	multi-token expressions	4	14.04
5	level of customizability	2	12.64
6	keyphrase extraction	5	9.55
7	humanities scholars	2	8.31
8	document collections	4	7.08
9	configuration	4	6.87
10	set of key-concepts	2	6.63

KD

1. Open all_topics.html obtained on the Trump folder
2. Find the topic connected to **immigration**
3. Click on the list of keywords and check the name of the most relevant document: *Trump_2016-08-31.txt*
4. Open this file with a text editor, copy the text and go to the online demo and paste it: http://celct.fbk.eu:8080/KD_KeyDigger/
5. Run the demo and check the results: how to deal with “country - countries”? http://textanalysisonline.com/spacy-word-lemmatize
6. Download the results of KD in tsv format by clicking on the “Download Data” link under the word cloud
7. Do the same process for Clinton’s most relevant file for the immigration topic: *Clinton_2016-08-05.txt*

CoreNLP

- Integrated NLP toolkit with a broad range of analysis tools
  - made by Java-based modules for the solution of a range of basic NLP tasks
  - languages: **English**, Arabic, Chinese, German, French, Spanish
    - Italian: http://tint.fbk.eu/
  - command line access
  - online demo: http://corenlp.run/

https://stanfordnlp.github.io/CoreNLP/

CoreNLP: (Some) Annotators

ANNOTATOR	FUNCTION
tokenize	subdivides a text into individual tokens, i.e. words, punctuation marks etc.
ssplit	segments a text into sentences
pos	assigns word class labels to each token according to a model and annotation scheme
lemma	provides the lemma or base form for each token
ner	identifies tokens that are proper nouns as members of specific classes such as Person, Organization etc.
parse	analyses and annotates the syntactic structure of each sentence in the text
relation	finds relations between two entities
quote	picks out quotes from a text

PoS Tagset

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Dependency Tagset

Description	UD tag
clausal modifier of noun	acl
adverbial clause modifier	advcl
adverbial complement	advmod
adverbial modifier	advmod
adjectival modifier	amod
appositional modifier	appos
auxiliary	aux
passive auxiliary	auxpass
coordinating conjunction	cc
obligatory clausal complement	ccomp
optional clausal complement	ccomp
classifier	nmod
compound	compound
conjunct	conj
clausal subject	csubj
clausal passive subject	csubjpass
unspecified dependency	dep
determiner	det
discourse element	discourse
dislocated elements	dislocated
direct object	dobj
expletive	expl

foreign word	foreign
general adjunct	advmod
goes with	goeswith
honorific	name
list	list
marker	mark
multi-word expression	mwe
name	name
obligatory nomino-adjectival complement	nmod
optional nomino-adjectival complement	nmod
nomino-adjectival modifier	nmod
obligatory nominal complement	nmod
optional nominal complement	nmod
nominal modifier	nmod
nominal subject	nsubj
passive nominal subject	nsubjpass
numeric modifier	nummod
parataxis	parataxis
obligatory prepositional complement	case
optional prepositional complement	case
object of preposition	nmod
predicate	root
preositional modifier	case

Constituency Tagset

ADJP	Adjective Phrase
ADVP	Adverb Phrase
CONJP	Conjunction Phrase
FRAG	Fragment
INTJ	Interjection
LST	List marker
NX	head of the NP
NP	Noun Phrase
PP	Prepositional Phrase
PRN	Parenthetical
PRT	Particle
QP	Quantifier Phrase
RRC	Reduced Relative Clause
VP	Verb Phrase
WH...	Phrases containing wh-words

CoreNLP: Tutorial

- Copy the file *Daredevil_Joins_Tanks.txt* in the CoreNLP folder
- Launch the command prompt
  - Windows: *All Programs -> Accessories -> Command Prompt*
  - Mac: open the *Terminal* window
- Go in the CoreNL folder using the `cd` command
- Type the following command:

```
java -Xmx2g -cp "./*" edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators "tokenize,ssplit,pos,lemma,ner,parse,relation,coref,quote" -ner.useSUTime false -coref.algorithm neural -outputFormat text -file Daredevil_Joins_Tanks.txt
```

- Try different output format: conll / text / xml
- Analyze the output

THANK YOU!

Email: sprugnoli@fbk.eu
Web Site: http://dh.fbk.eu
Twitter: https://twitter.com/DH_FBK

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES