

1 We observe today not a victory of party but a celebration of freedom -- symbolizing an end as well as a beginning -- signifying renewal as well as change .

2 For I have sworn before you and Almighty God the same solemn oath our forebears prescribed nearly a century and three quarters ago .

3 The world is very different now .

4 For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life .

Natural Language Processing

Rachele Sprugnoli
sprugnoli@fbk.eu



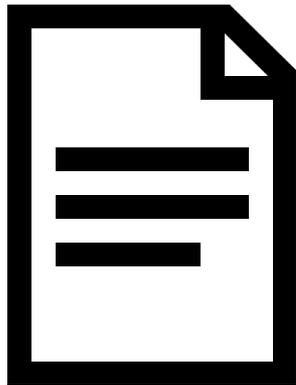
Natural Language Processing (NLP)

- **Goal:** create machines that understand natural languages
- Encompasses many disciplines:
 - linguistics
 - computer science
 - artificial intelligence
 - statistics
 - machine learning

Natural Language Processing

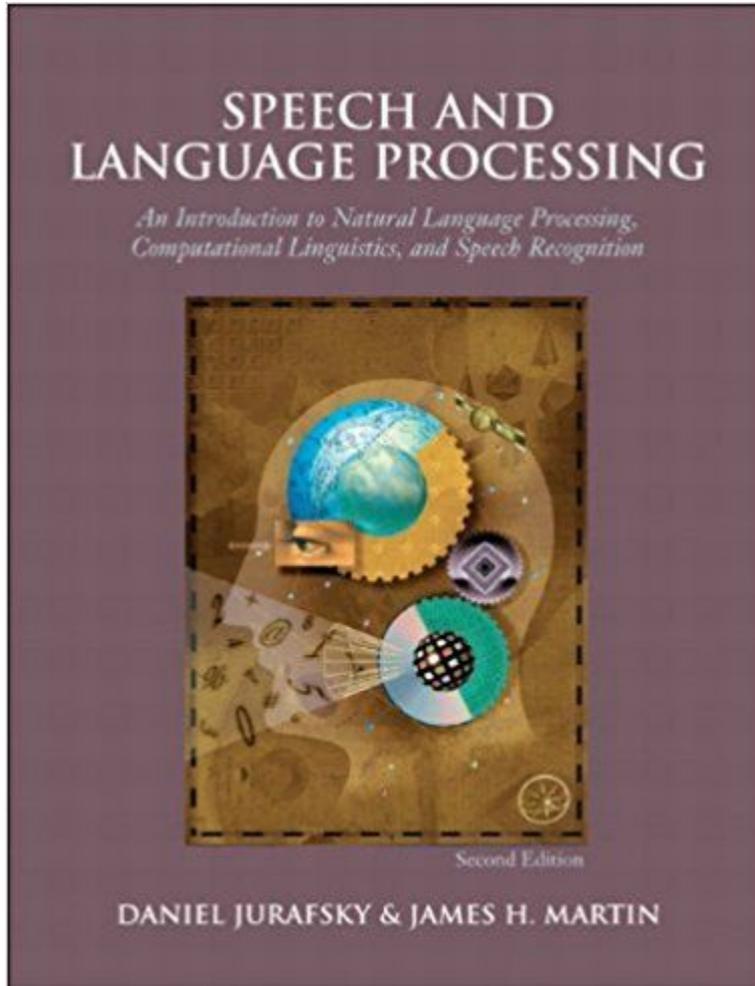


- Oral History and Technology:
<http://oralhistory.eu/>



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

What to read



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

Why NLP is hard but also fun?

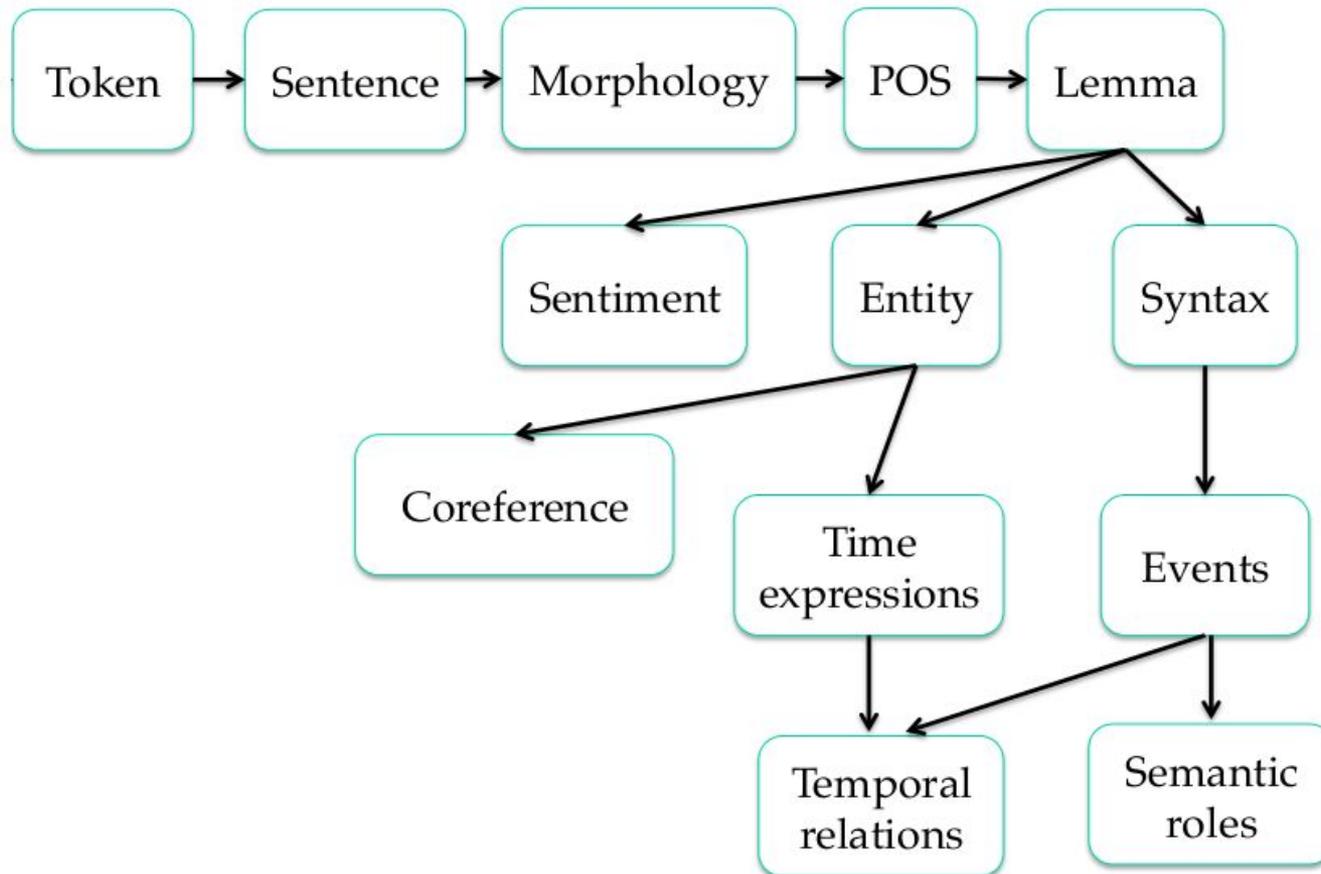
- Ambiguity: *Flying planes can be dangerous*
- Non-standard language: *#GreatJob @justinbieber!*
- Old language: *I have thee not, and yet I see thee still*
- Neologisms: *chilax*
- Idioms: *lose face*
- World knowledge: *Mary and Sue are sisters*
- Tricky named entities: *Let It Be was released in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

Where do people do research in NLP?

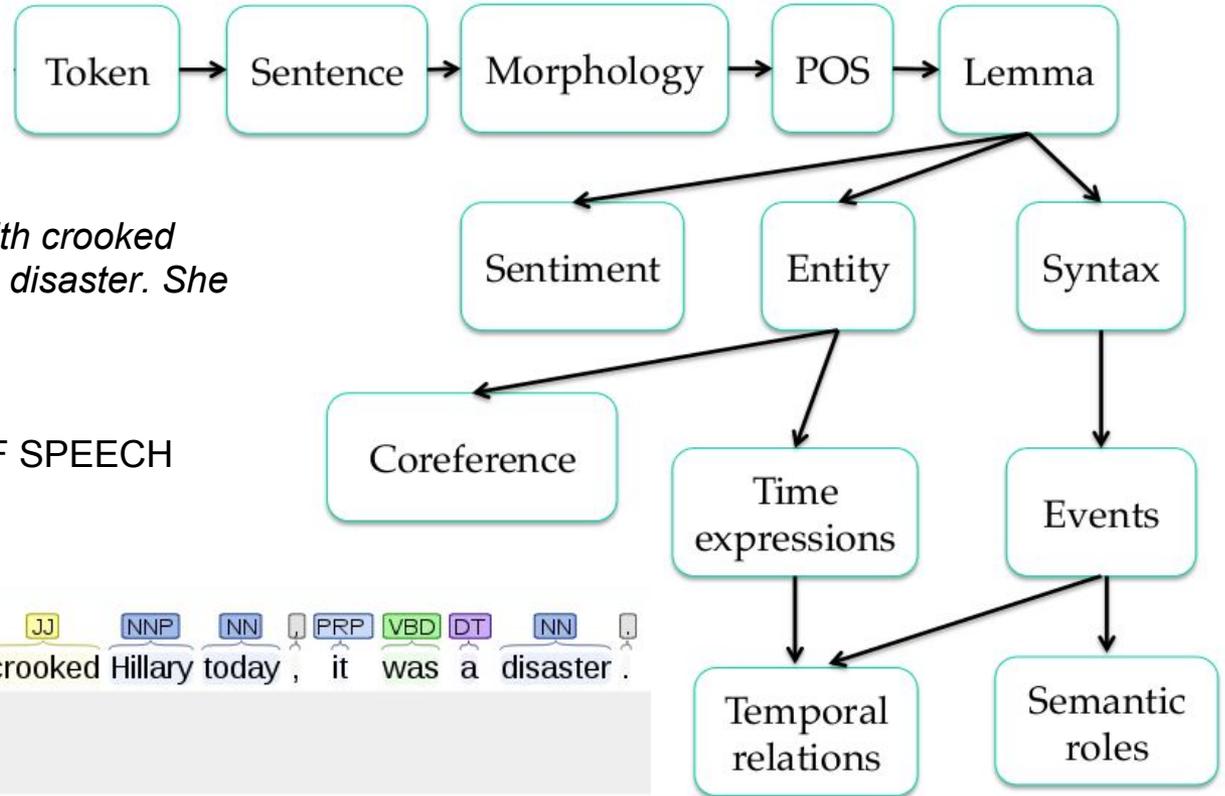


How to process the language?

- 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

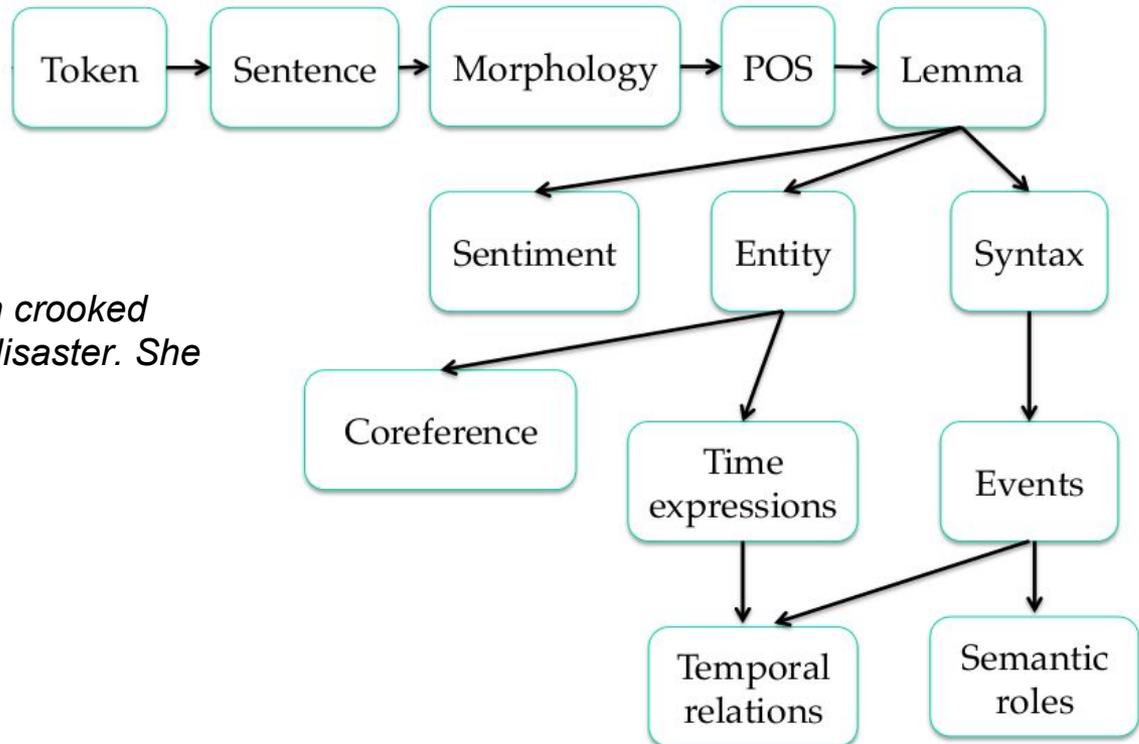
TOKEN - SENTENCE - PART OF SPEECH

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

2 DT NN .
A disaster .

3 PRP VBD DT NN .
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

MORPHOLOGY

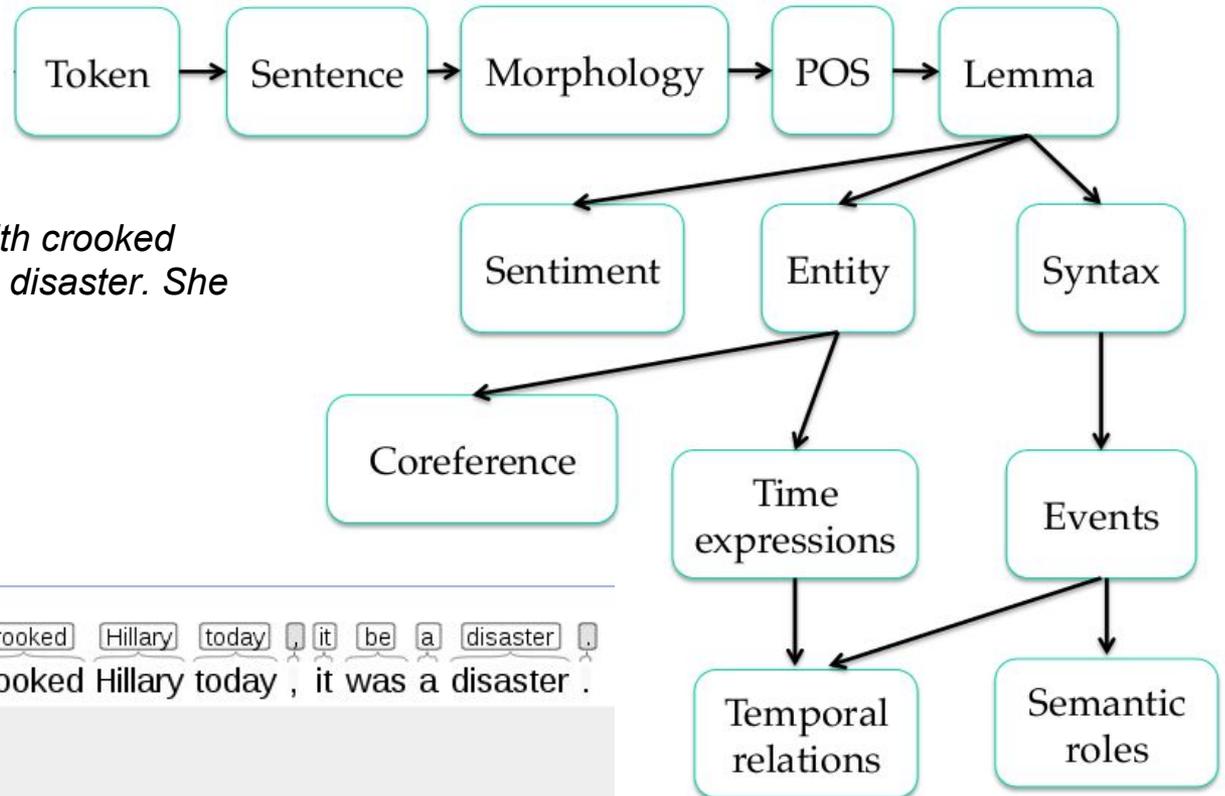
when+conj you+pron see+v+indic+pres+no3sing what+adj+zero happen+v+indic+past with+prep crooked+adj+zero NULL today+adv NULL
When you see what happened with crooked Hillary today ,

it+pron be+v+indic+past a+art disaster+n+sing .+punc
it was a disaster .

a+art disaster+n+sing .+punc
A disaster .

she+pron have+v+indic+past a+art disaster+n+sing .+punc
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

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 .

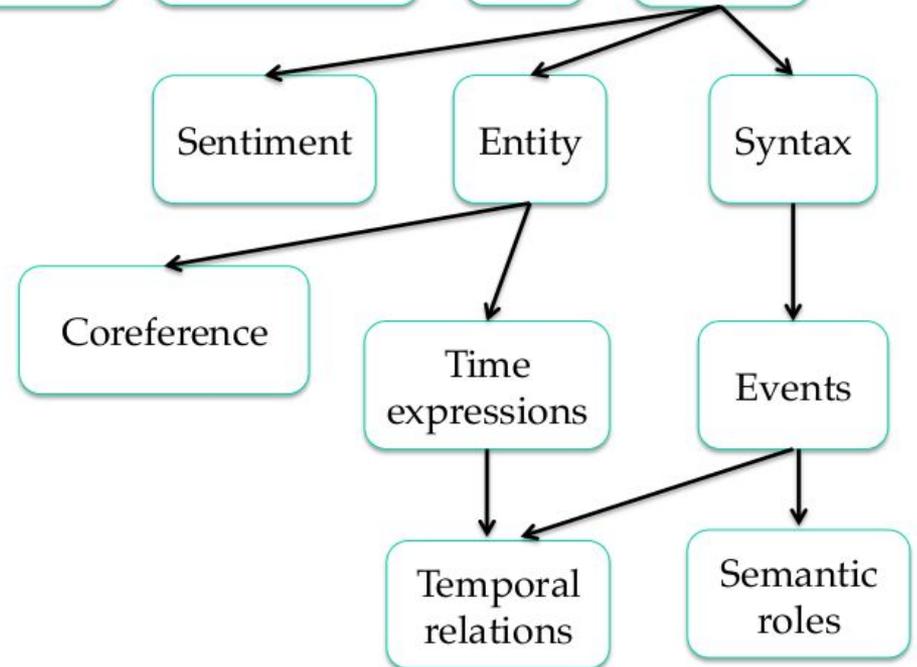
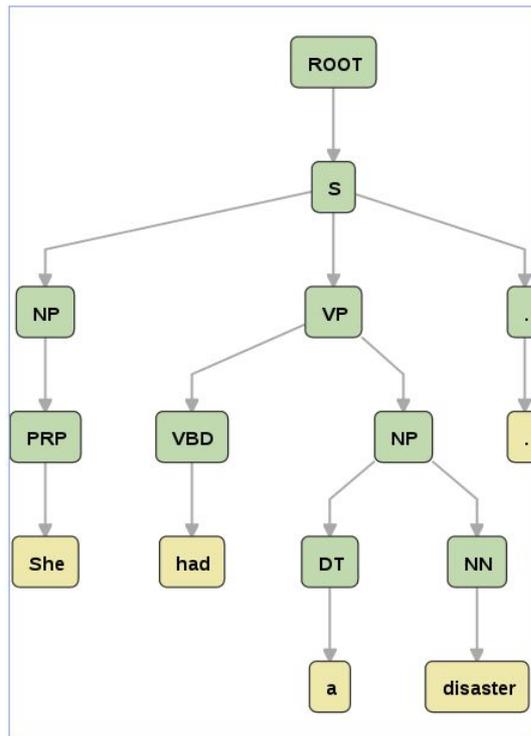
How to process the language?



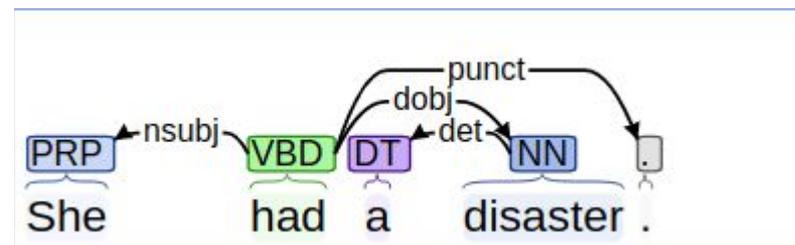
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
 Trump, 2016-08-05

SYNTAX
-
PARSING

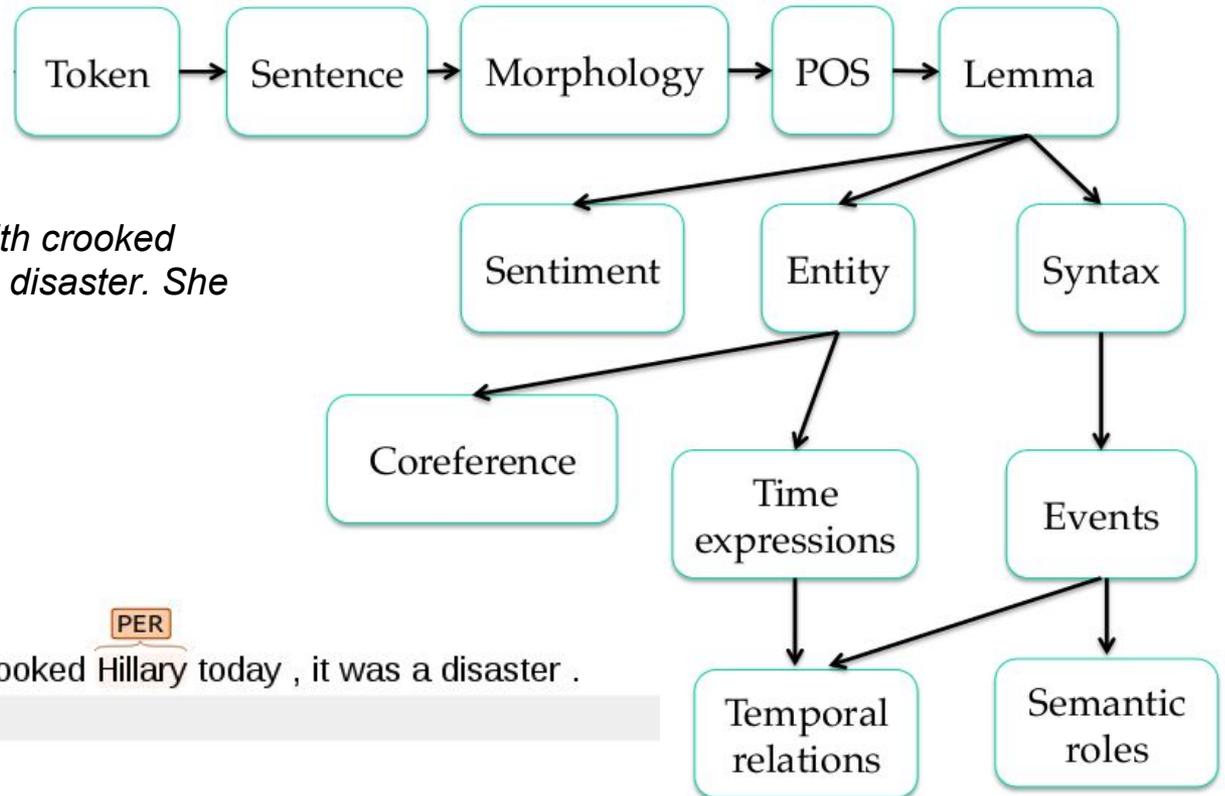
CONSTITUENCY PARSING



DEPENDENCY PARSING



How to process the language?

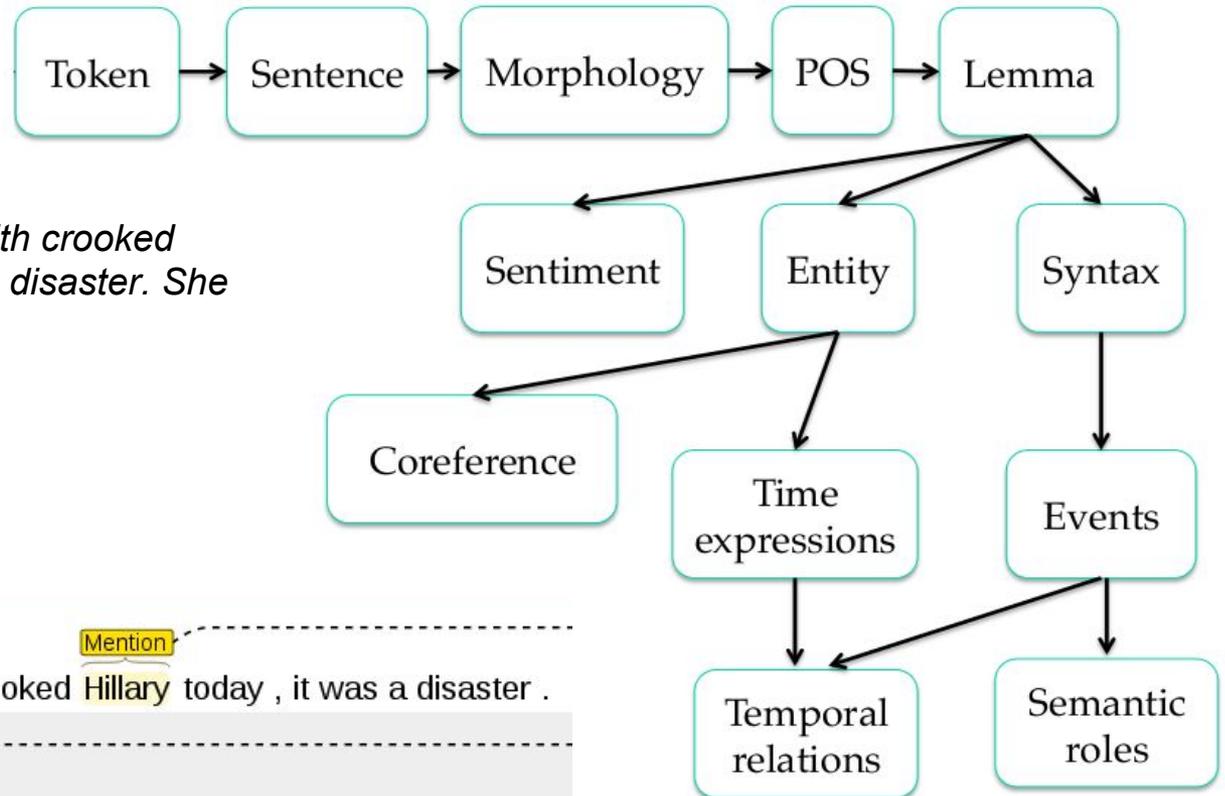


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

ENTITY

When you see what happened with crooked PER 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

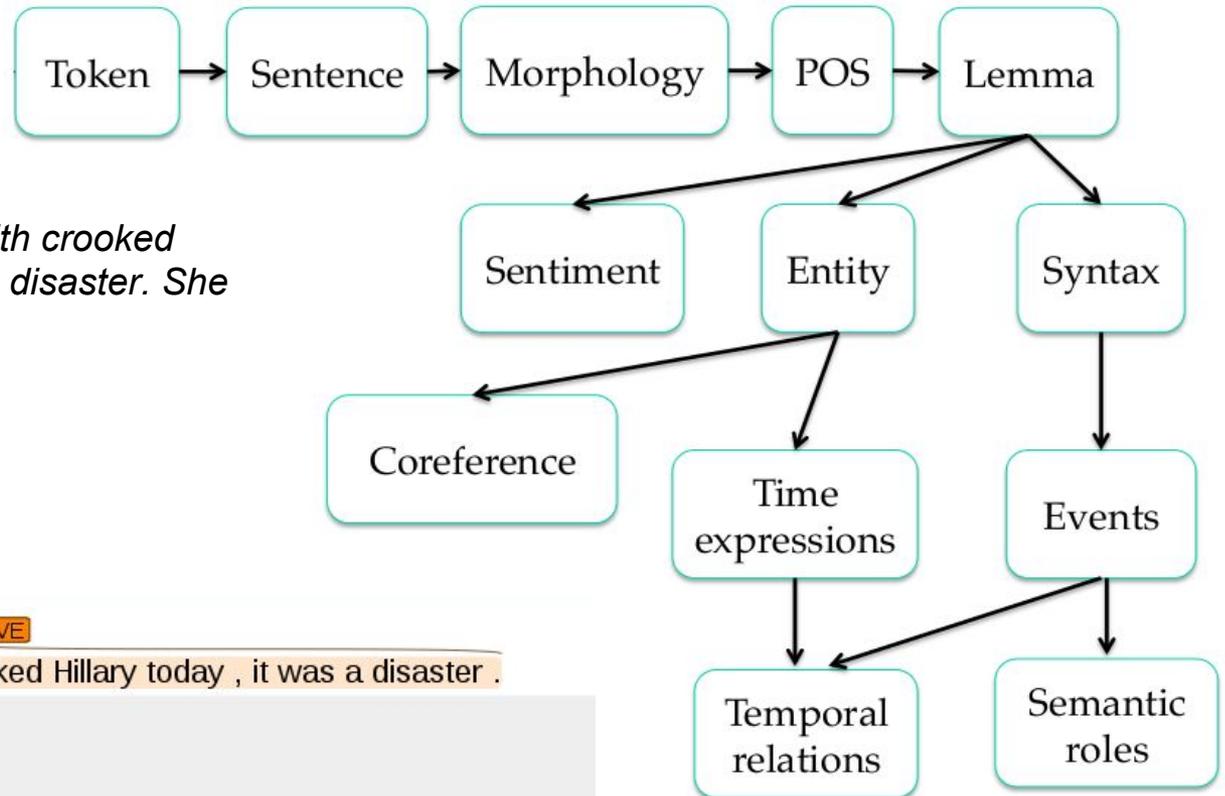
COREFERENCE

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

A disaster .

---coref--- 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

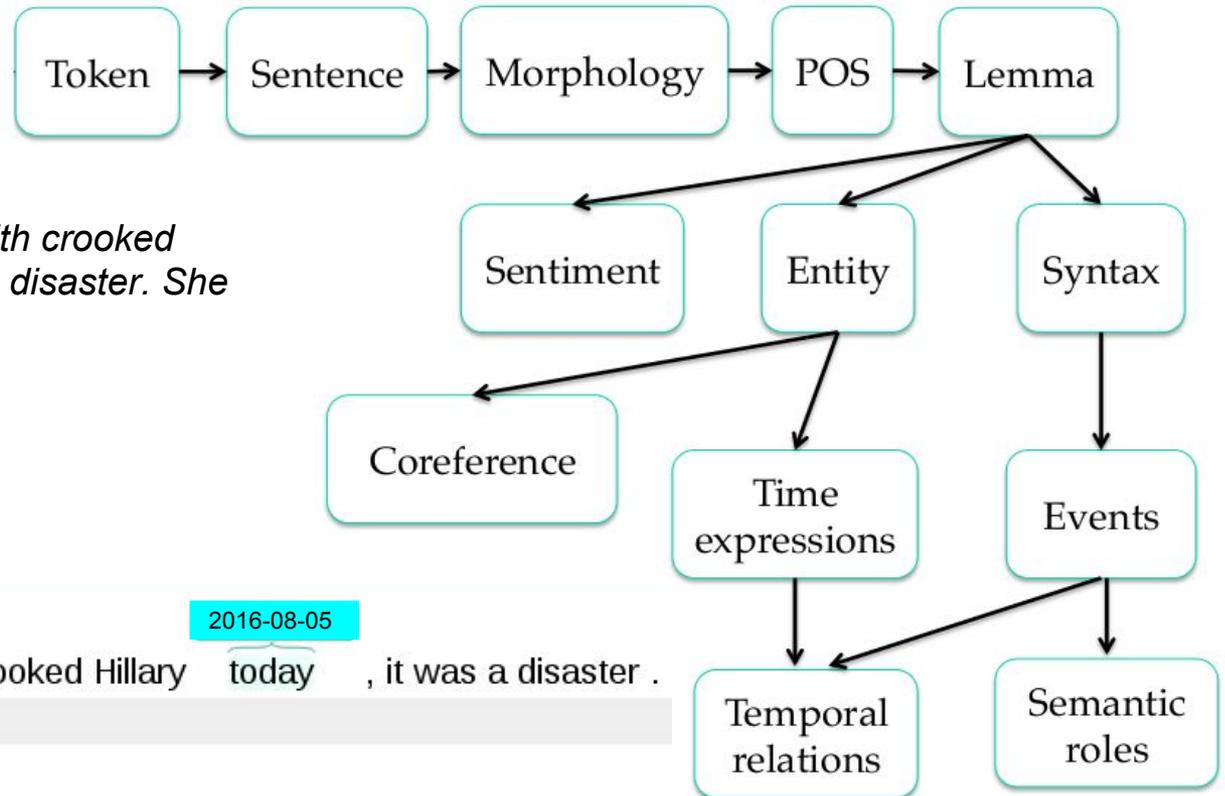
A disaster .

NEGATIVE

She had a disaster .

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

How to process the language?

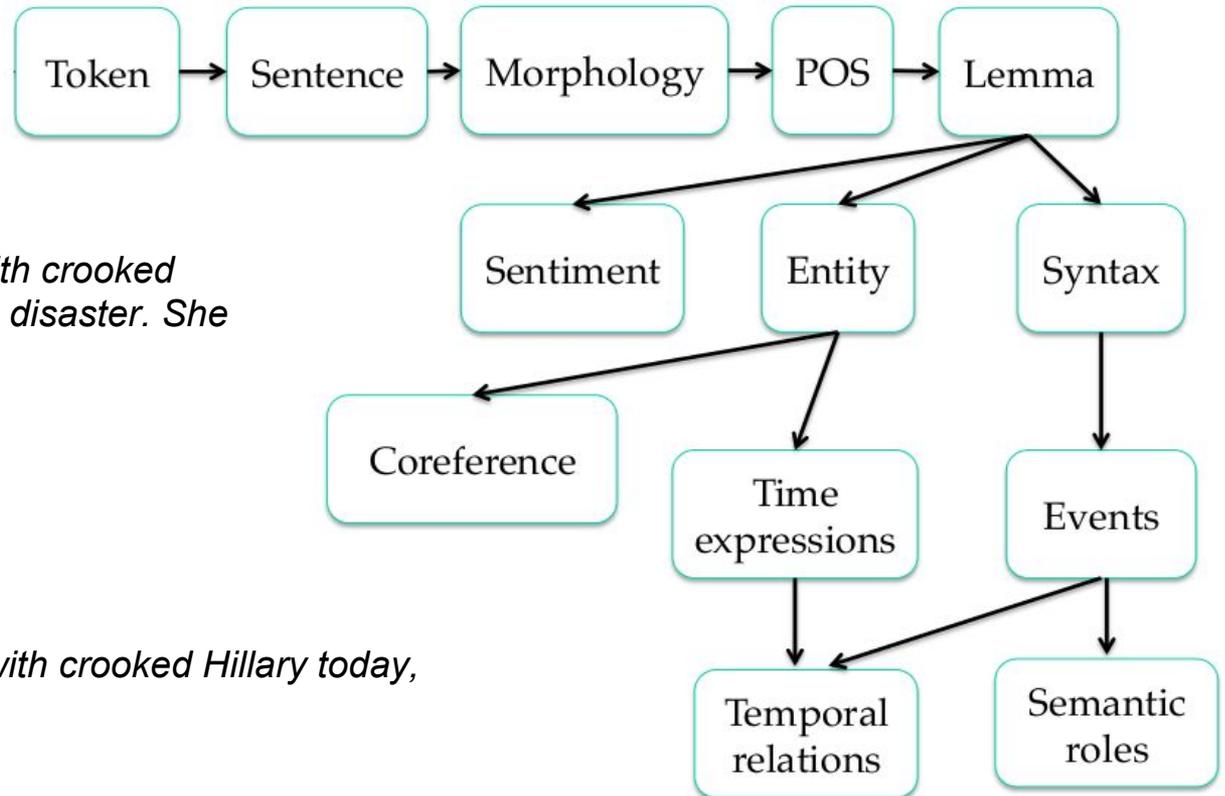


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TIME EXPRESSIONS

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

How to process the language?

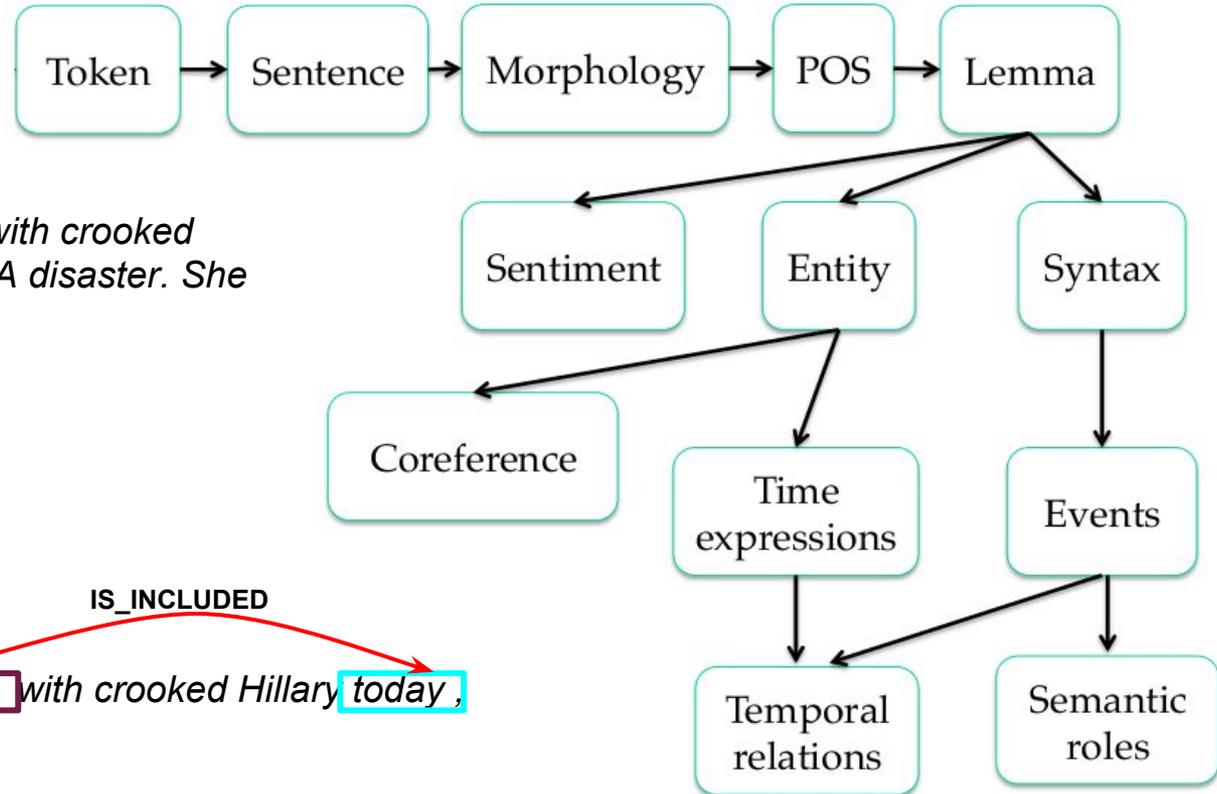


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

EVENTS

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

How to process the language?



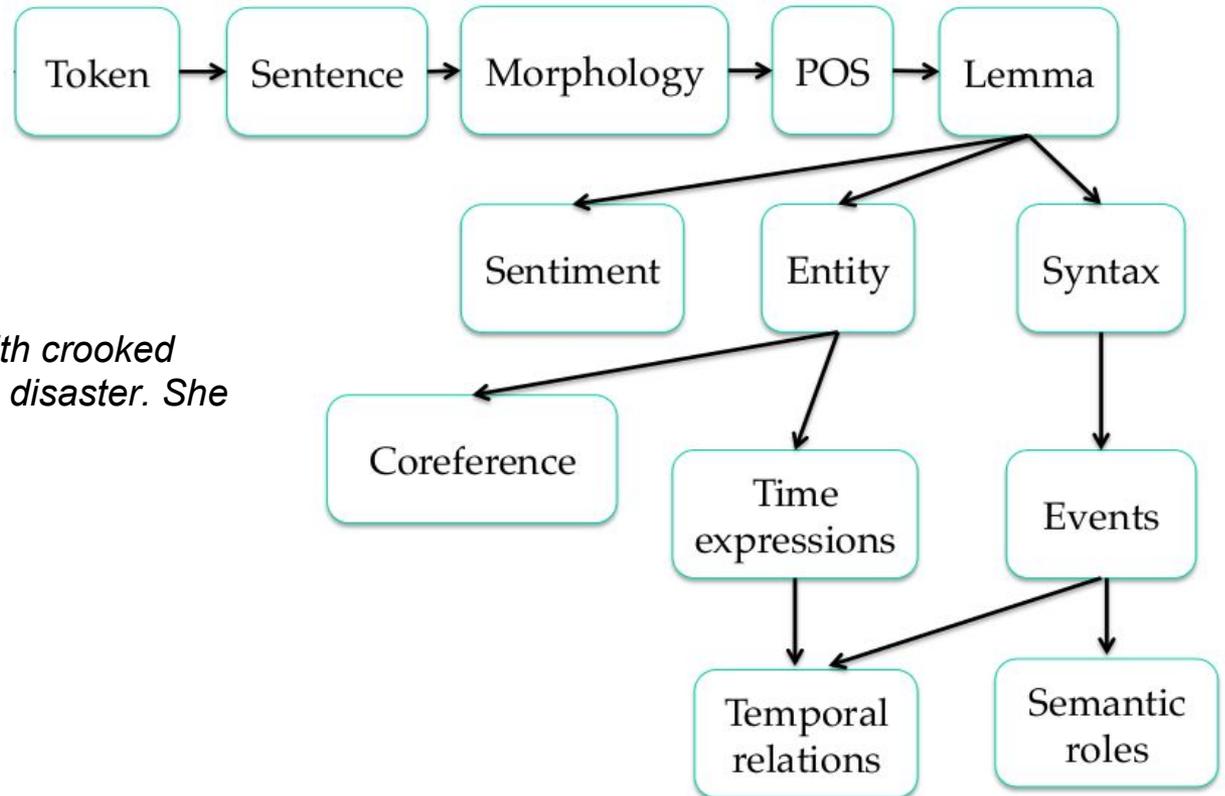
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

TEMPORAL RELATIONS

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

IS_INCLUDED

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

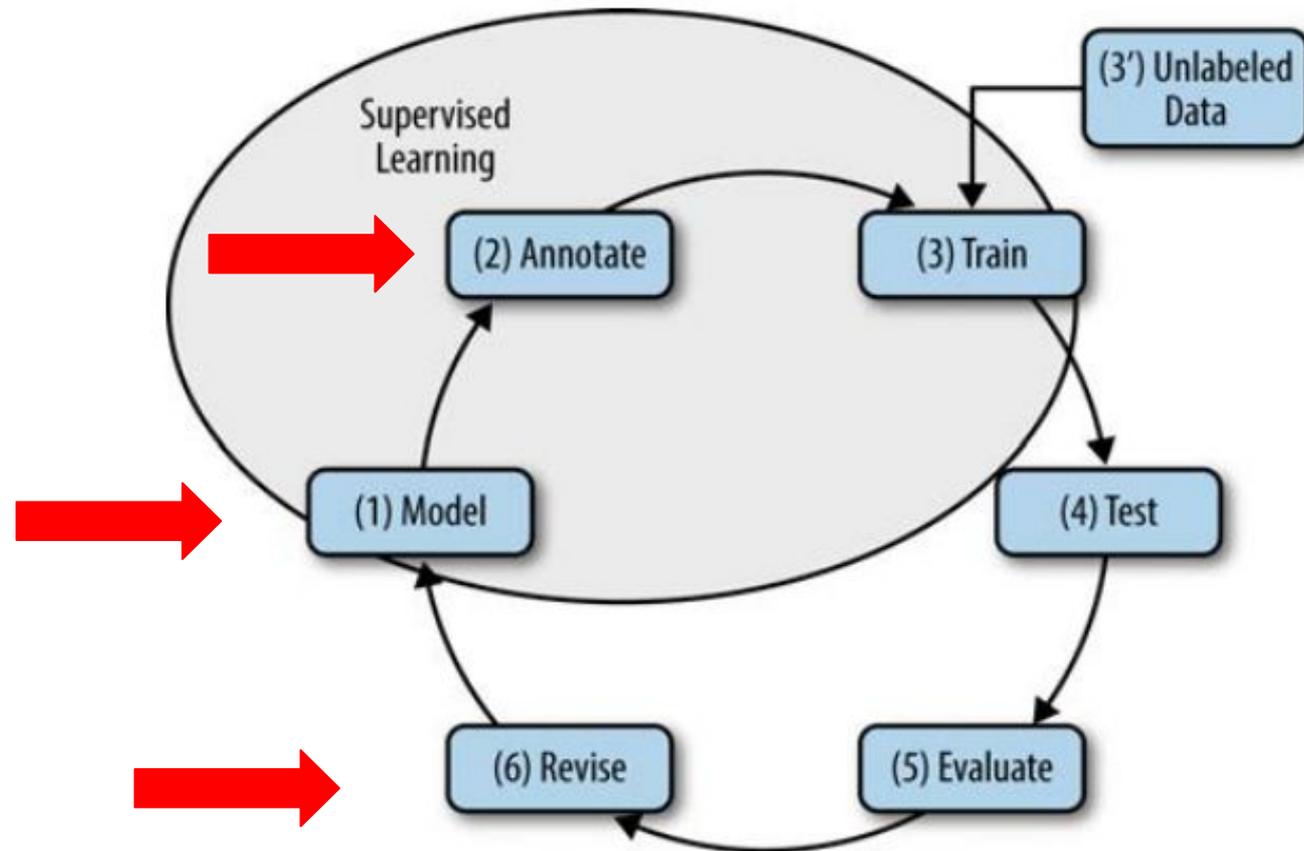
Possession

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

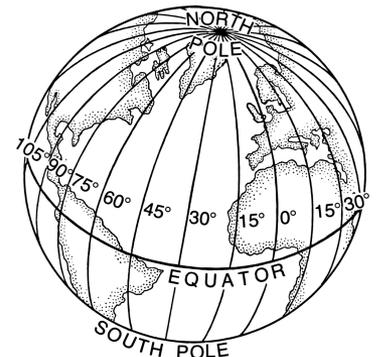
- Our approach: automatic extraction of motion trajectories from Wikipedia biographies

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

<p>But to fix ou must change Washington an quickly. Sadl other way. immigration anybody ever aren't known report on the talk about interests spen to cover the making an abso way it is. complicated subject, you fundamental immigration sy that it serve donors, poli powerful, powe</p>	<p>As secretary Clinton allo criminal alie because their to take them. They were too them back. Who would do this? this? A weak a would do thi described Hill most radical i United States summary of wh support sanctu Security, Med welfare for a by making them which will die immigrants.</p>	<p>Social Secu lifetime we immigrants k citizens. And being treated veterans. Reme going to all illegal immigr visa overstay release on the hey, go ahead It's called Expanding unconstitution including ins millions of i even more crim Obama's non- And she wants in Syrian ref country .</p>	<p>All Americans country, in wonderful, p immigrants are jobs and wag totally protec our nation are people livin everybody. An erased -- it lawful immigr if you look a the borders, are erased, borders, we r And that's r And I have endorsed by th 16,500. By IC First time anybody for pr</p>	<p>As I mentioned, Pueblo is filled with wonderful, hard-working immigrants. It's these hard-working immigrants who stand to lose the most from our open border immigration policy. Illegal immigration and broken Visa programs take jobs directly from Latino and Hispanic workers living here lawfully today -- you know that. They're taking your jobs. Illegal immigration also brings with it massive crime and massive drugs, including a terrible heroin problem right here in Colorado -- you have a big problem. So we're going to build the border wall and we are not -- what? We're going to build the wall and we're going to stop the drugs, the gangs, the violence from pouring into Colorado.</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

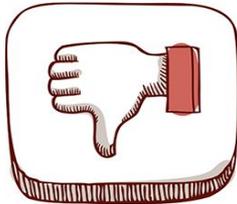
“That’s how topic modeling works in practice. You assign words to topics randomly and then just keep improving the model, to make your guess more internally consistent, until the model reaches an equilibrium that is as consistent as the collection allows.”

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>

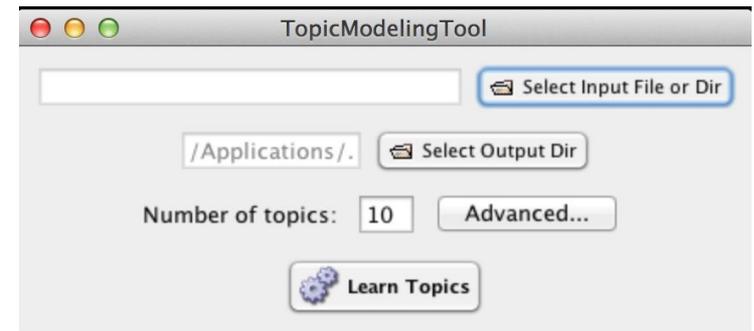
- MALLET: <http://mallet.cs.umass.edu/>

```
6      5      rings wilderness london park ring dust uranus numerous number no
ons narrow uranian particles dark discovered water found thylacinus launched
7      5      back gilbert thespis survived male pouch relative species relate
d theatre opera northern position shiloh states died markets alvida zaara
8      5      time including thylacine tasmanian tiger extinct record general
century debut devil marsupial australia return kings thernal omer wadia boyfrie
nd
9      5      battle union haves confederate kentucky army grant gen tennessee
war united confederates buell commonwealth day forced line men fighting

<900> LL/token: -9.16178
<910> LL/token: -9.15612
<920> LL/token: -9.1345
<930> LL/token: -9.13677
<940> LL/token: -9.10601

0      5      test cricket australian hill career record states mother gods en
ded innings scored batsman return held omer wadia online columns
1      5      including gunnhild united norway acting thespis american king to
p kehna headed opera creating rulers spent husband father details saga
2      5      system average equipartition theorem law energy kinetic independ
ent effects stars classical heat motion equilibrium thermal energies temperature
regular asia
3      5      zinta south role hindi actress film indian survived world grossi
ng naa ho female earned debut films narrow addition discovered
4      5      yard national wilderness life london parks years century standar
ds journalist found areas worked president government received society died acco
mplishments
```

- 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**.

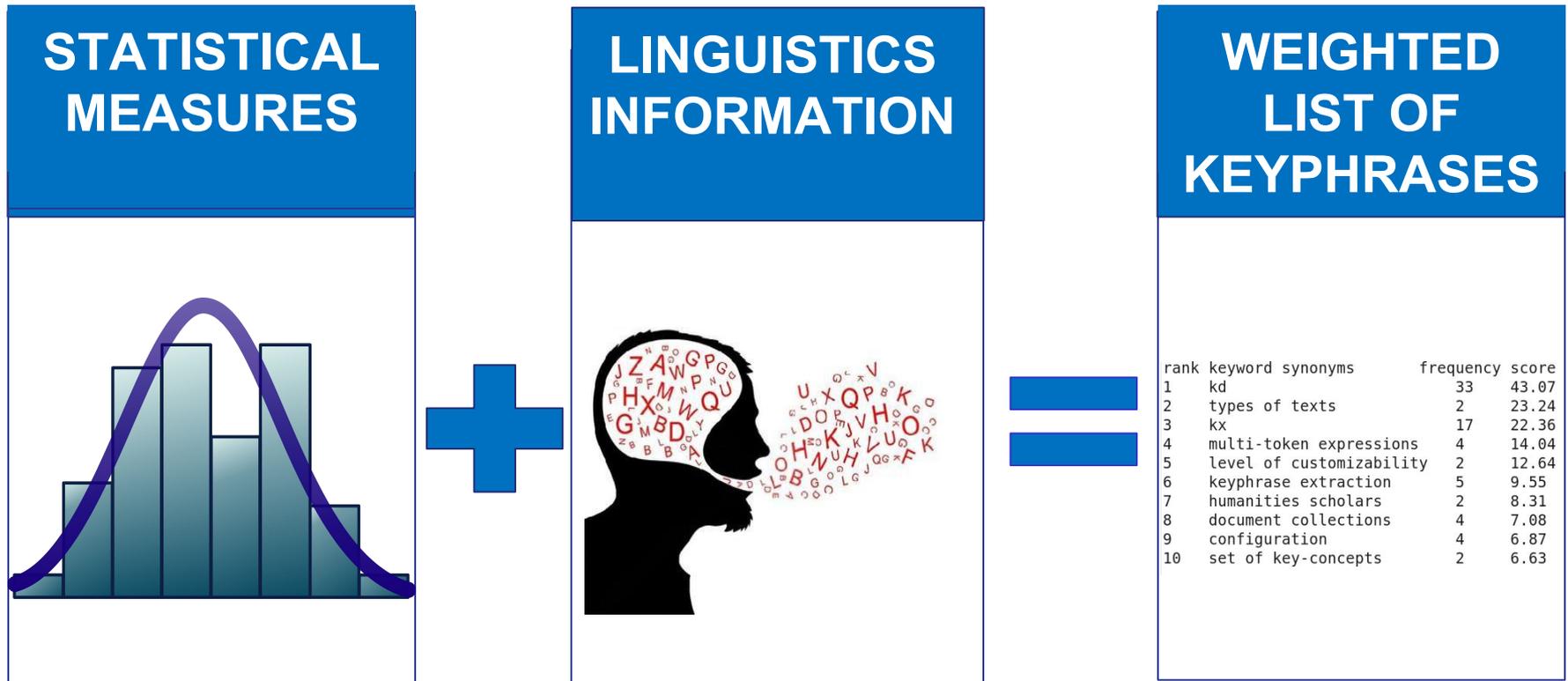
Both single words and multi-token expressions

Both single documents and whole corpora

Keyphrase Extraction: Tool

- KD = Keyphrase Digger

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



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	doobj
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
prepositional 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