



NLP for History

Methods and tools for textual analysis in the historical domain

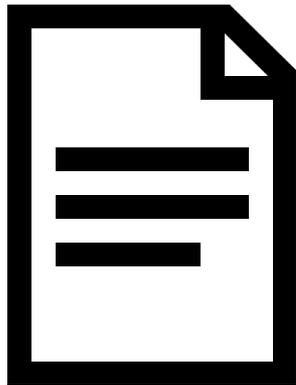
Rachele Sprugnoli



Natural Language Processing

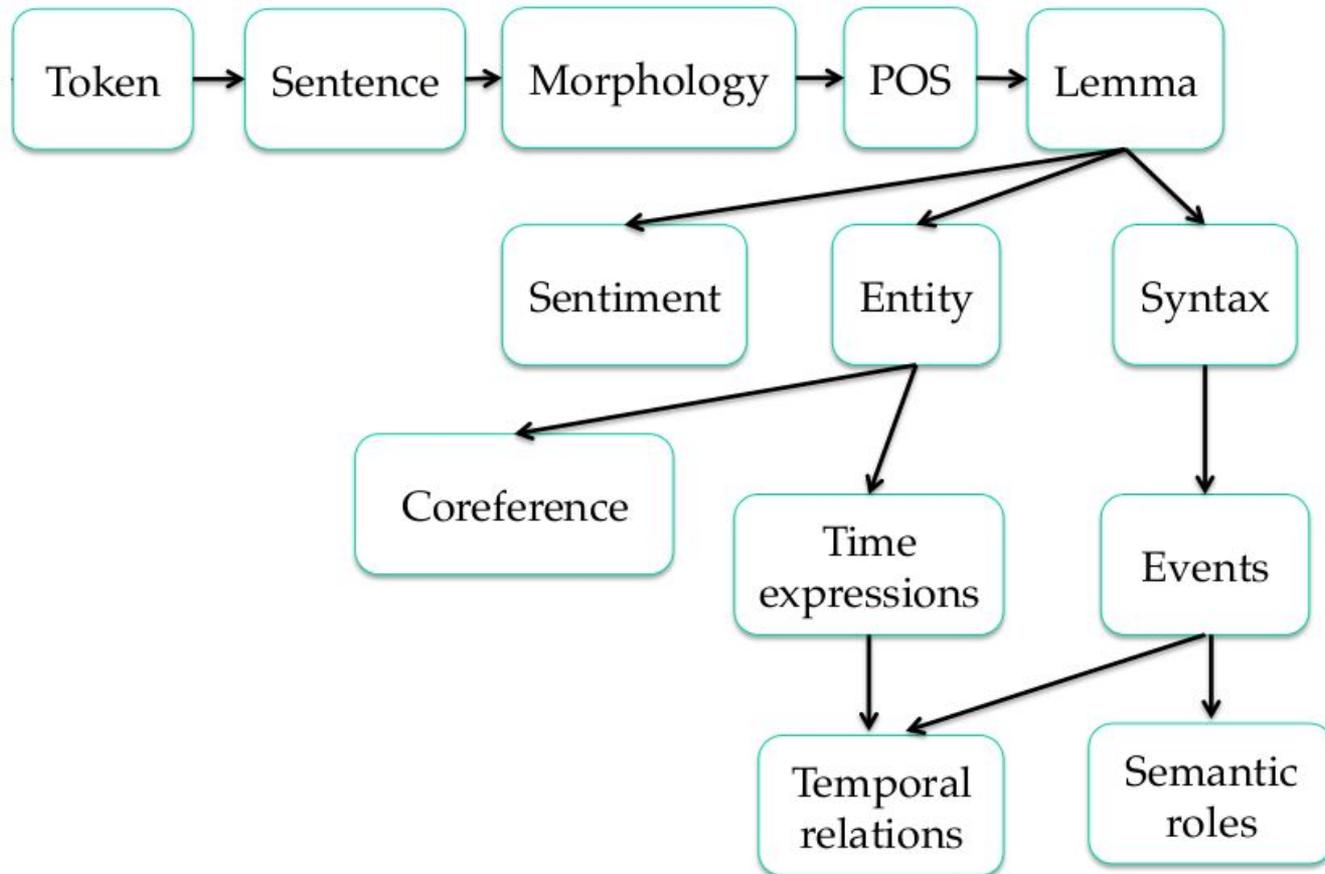


- Oral History and Technology:
<http://oralhistory.eu/>

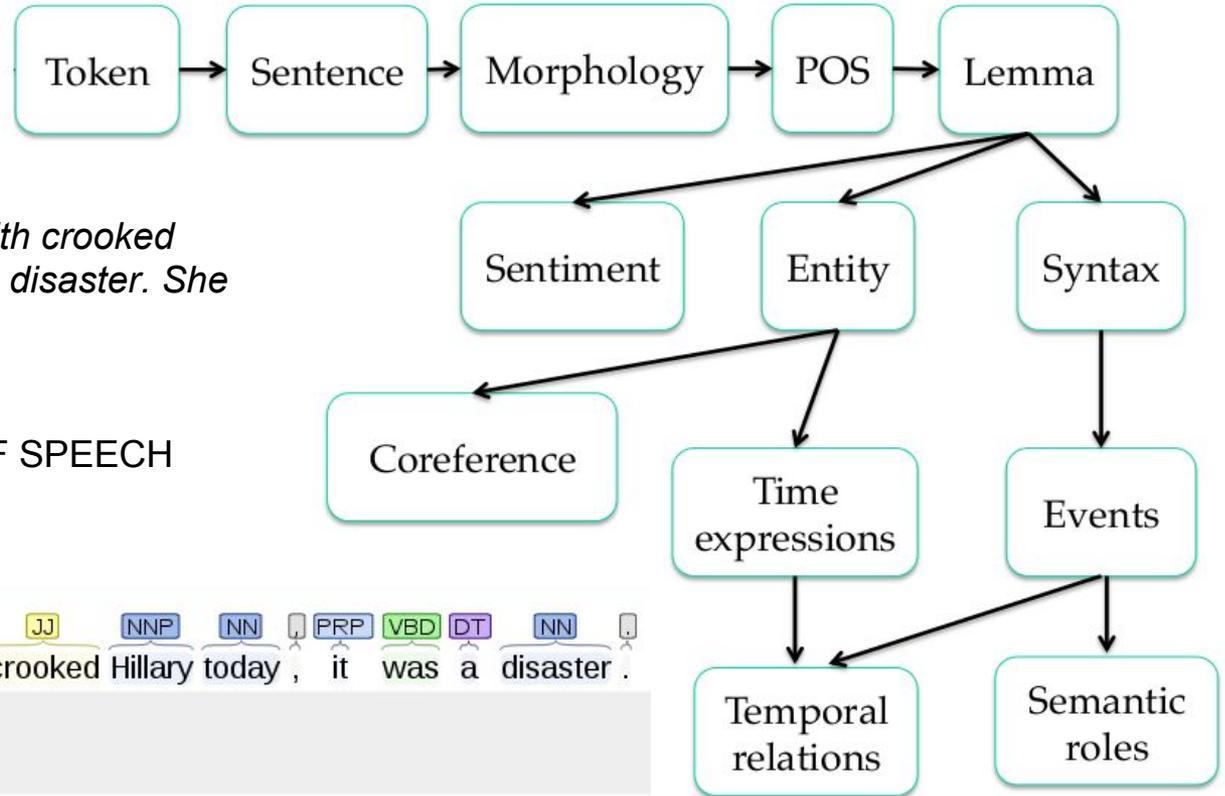


- “Natural Language Processing for Historical Texts” by Michael Piotrowski

How to process the language?



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

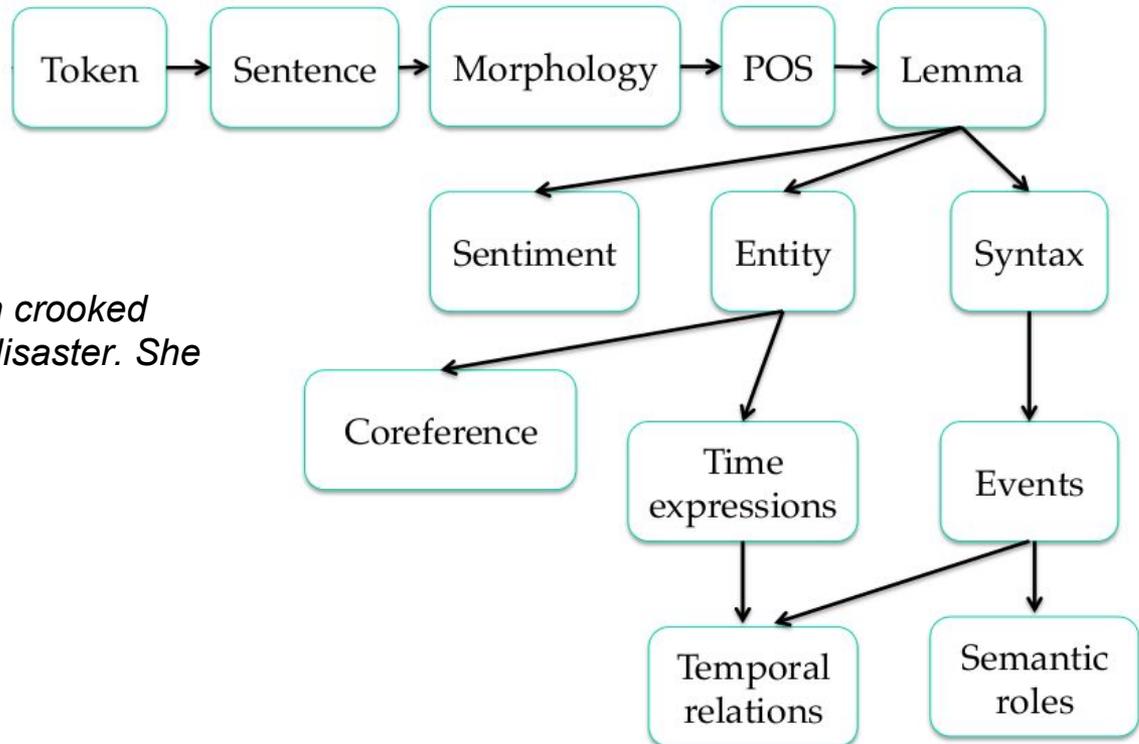
TOKEN - SENTENCE - PART OF SPEECH

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

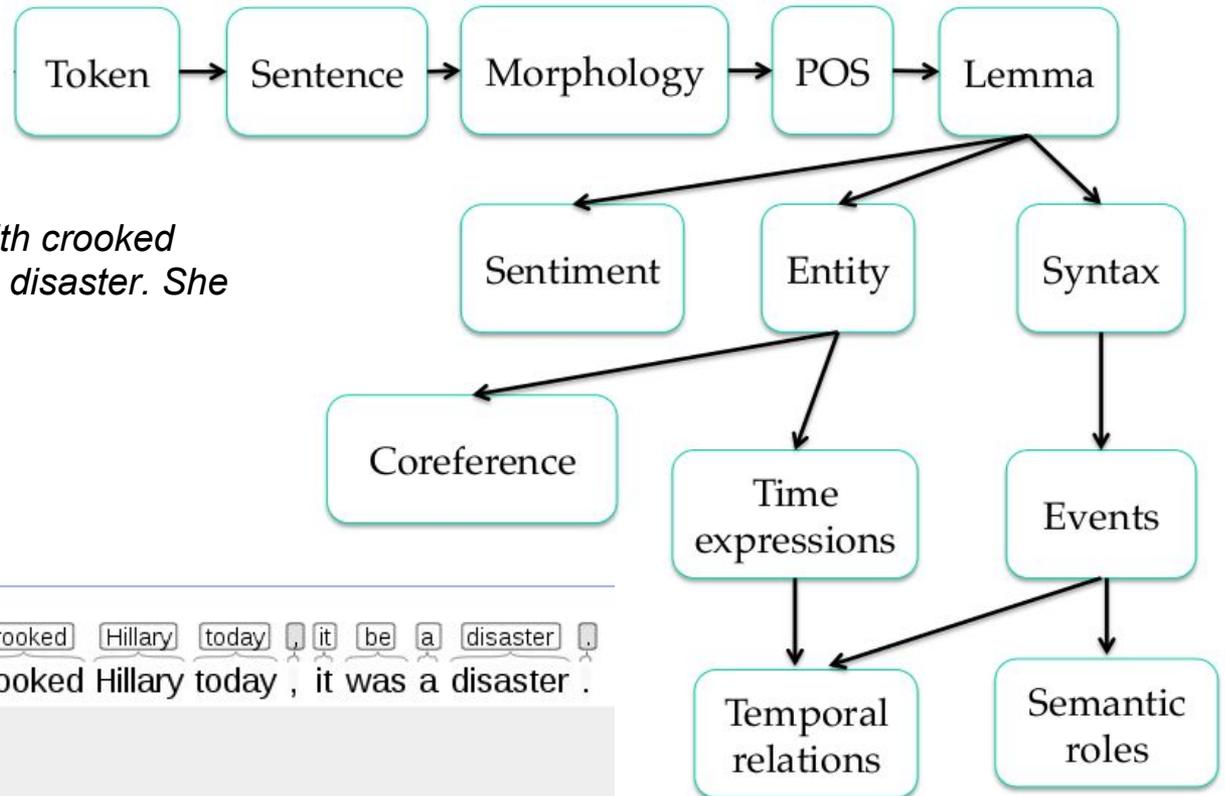
when+conj you+pron see+v+indic+pres+no3sing what+adj+zero happen+v+indic+past with+prep crooked+adj+zero NULL today+adv NULL
When you see what happened with crooked Hillary today ,

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 .

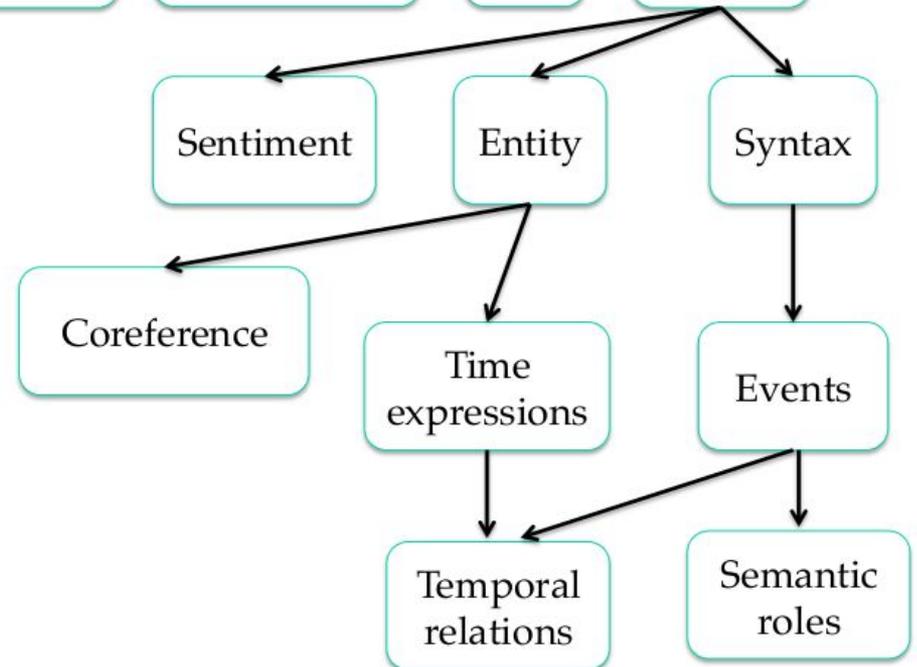
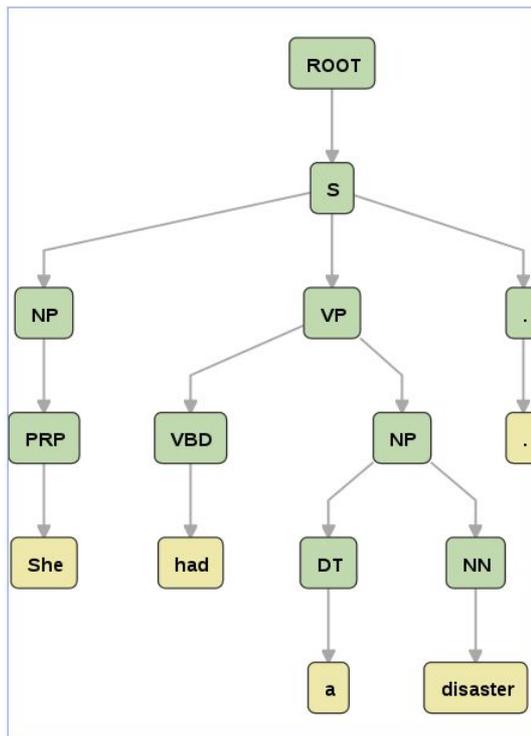
she have a disaster .
She had a disaster .

How to process the language?

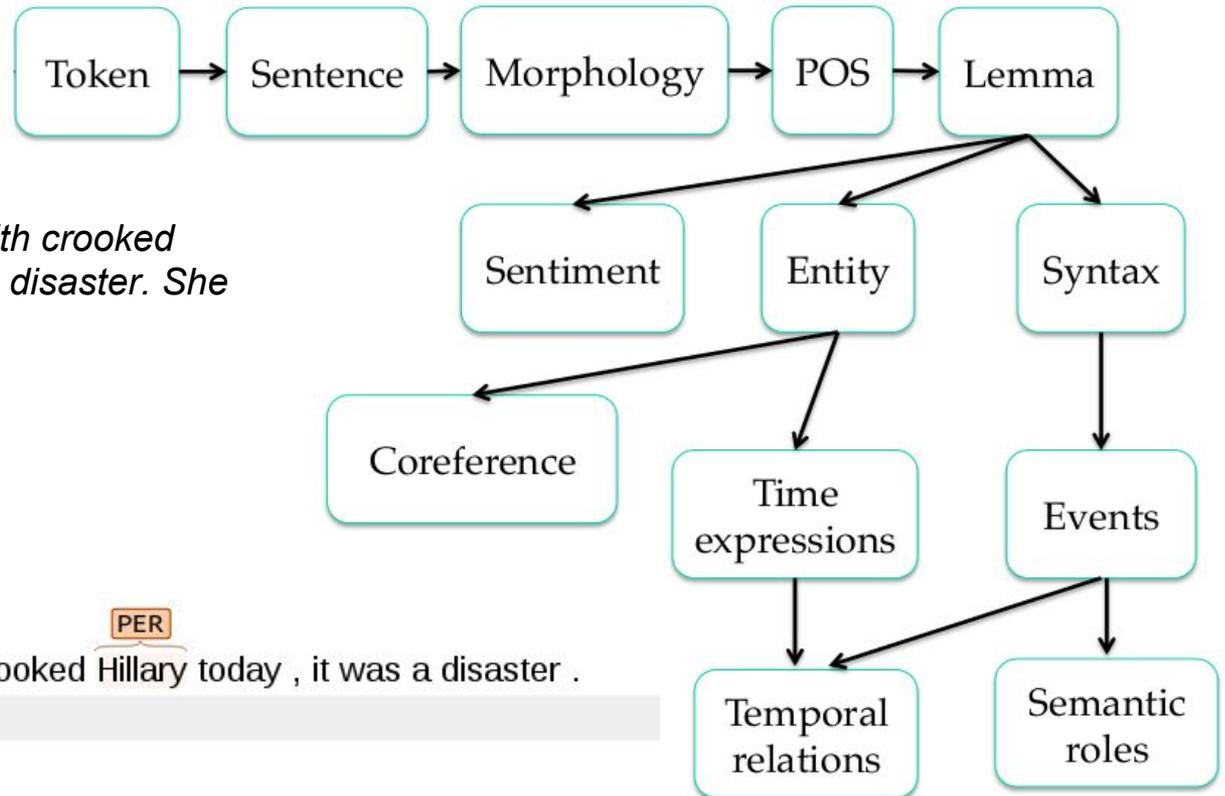


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

SYNTAX



How to process the language?

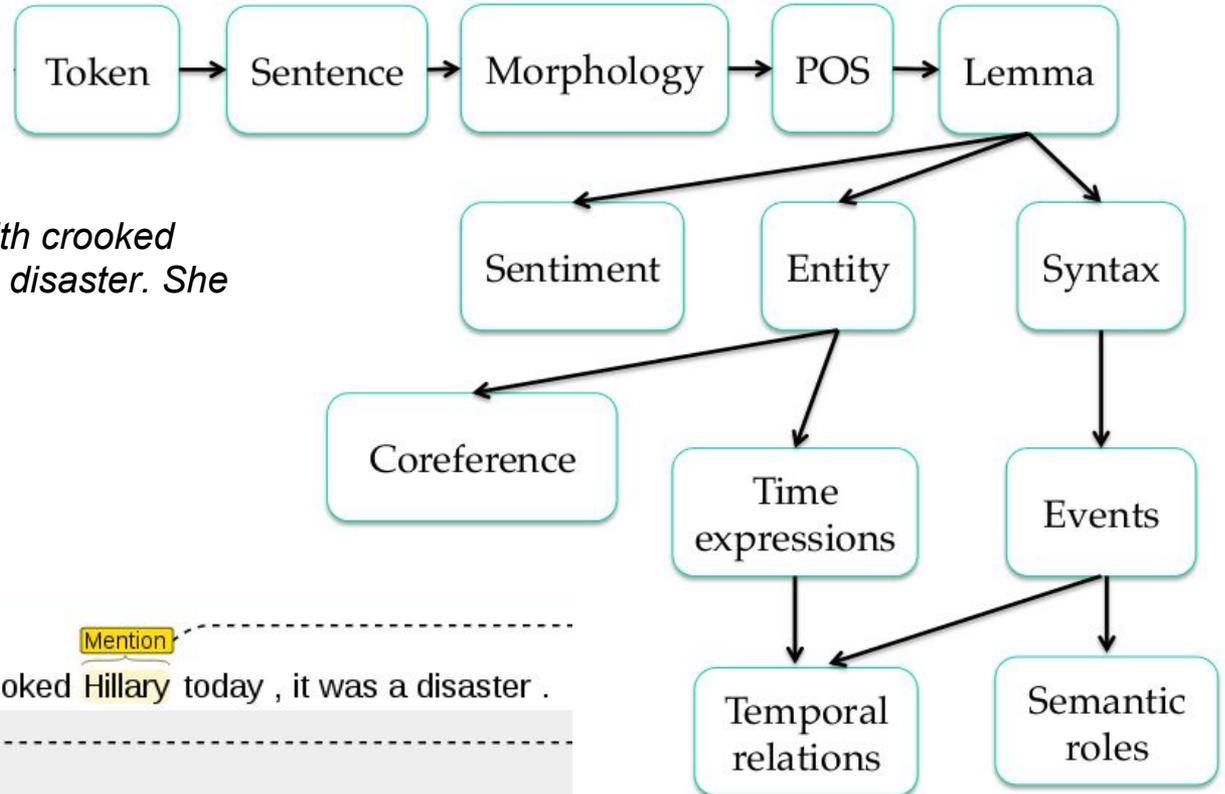


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

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

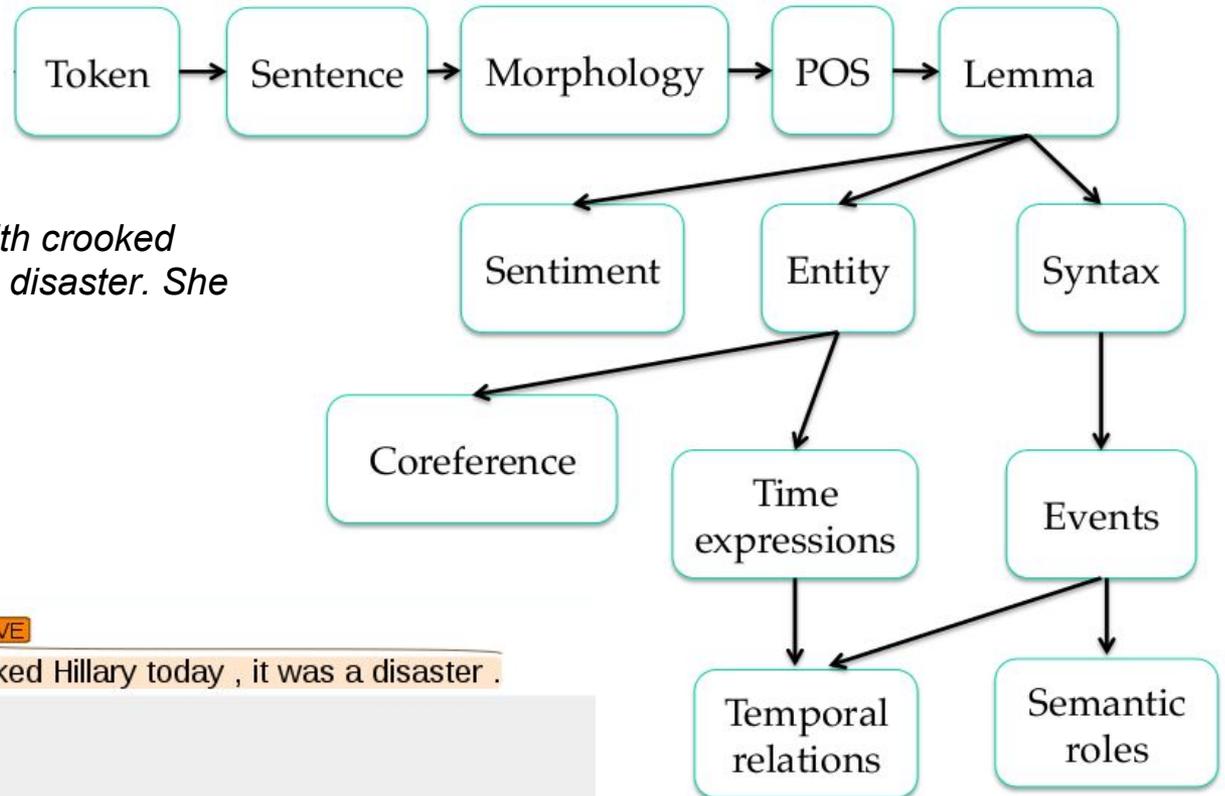
COREFERENCE

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 .

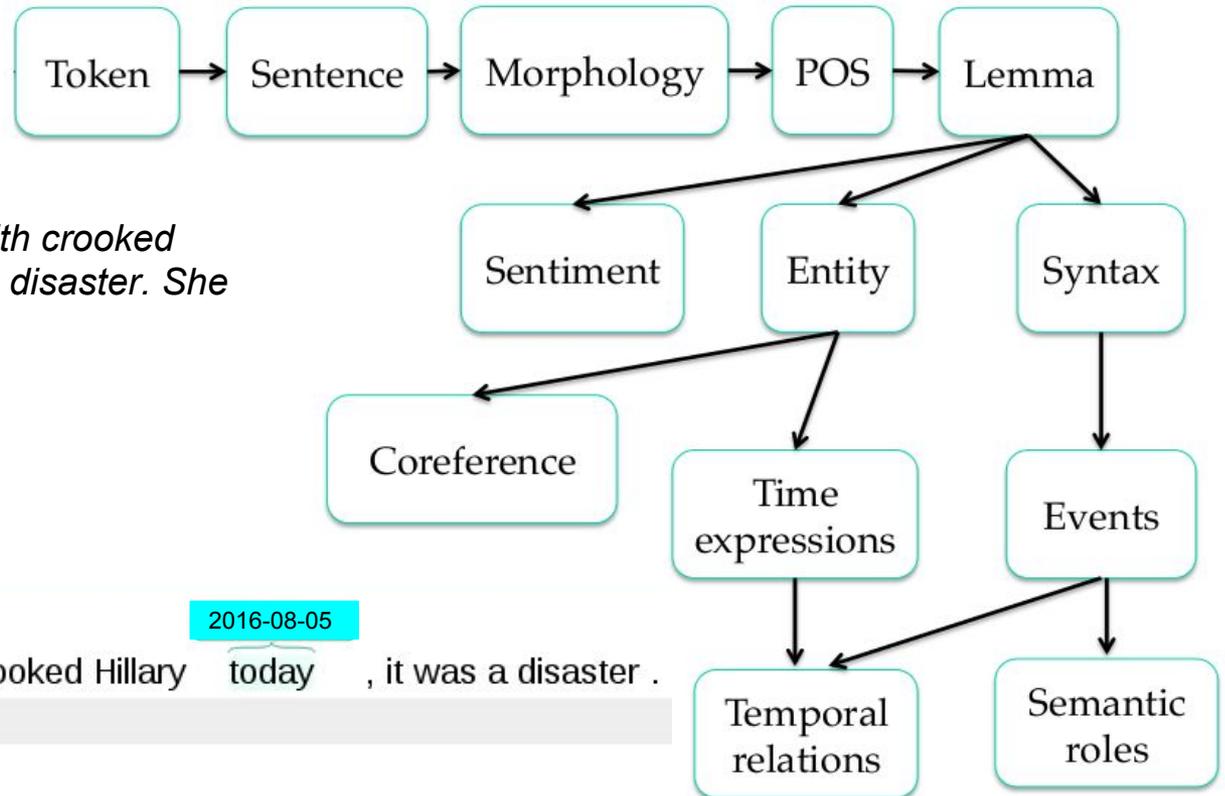
VERY NEGATIVE

A disaster .

NEGATIVE

She had a disaster .

How to process the language?

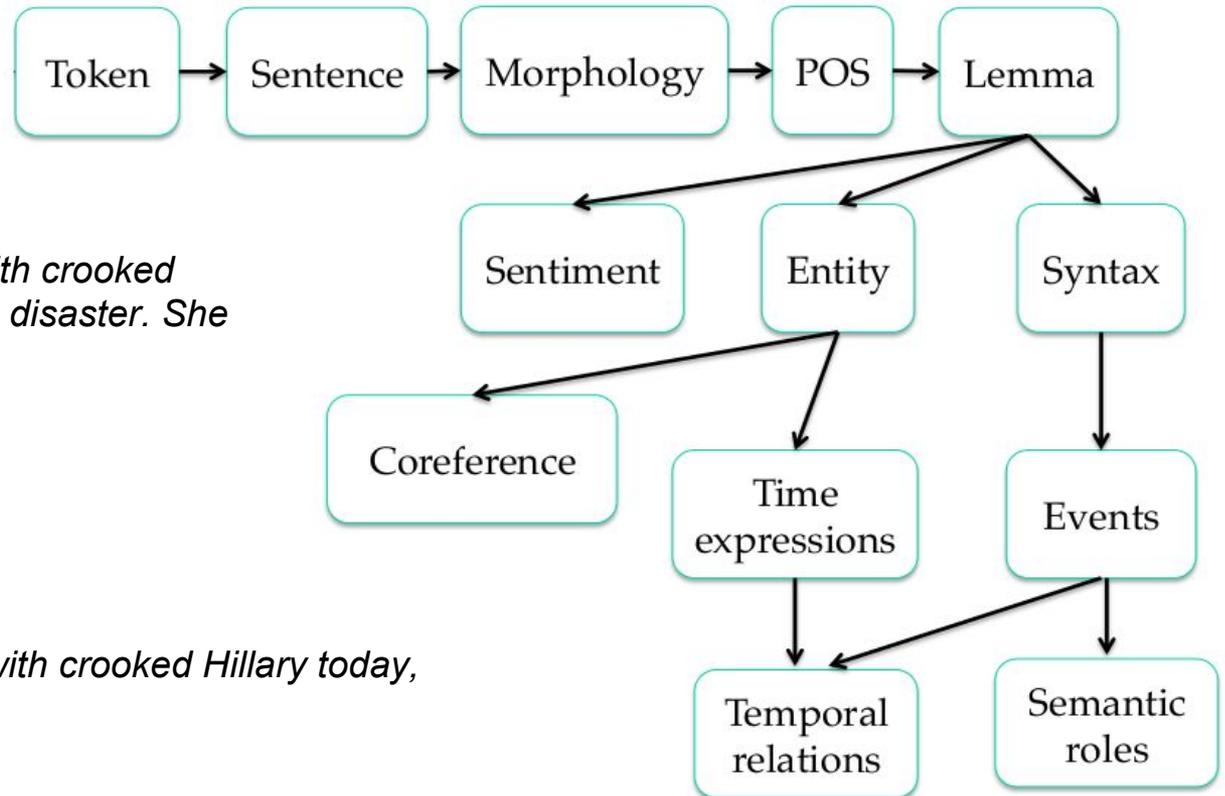


When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.
Trump, 2016-08-05

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?



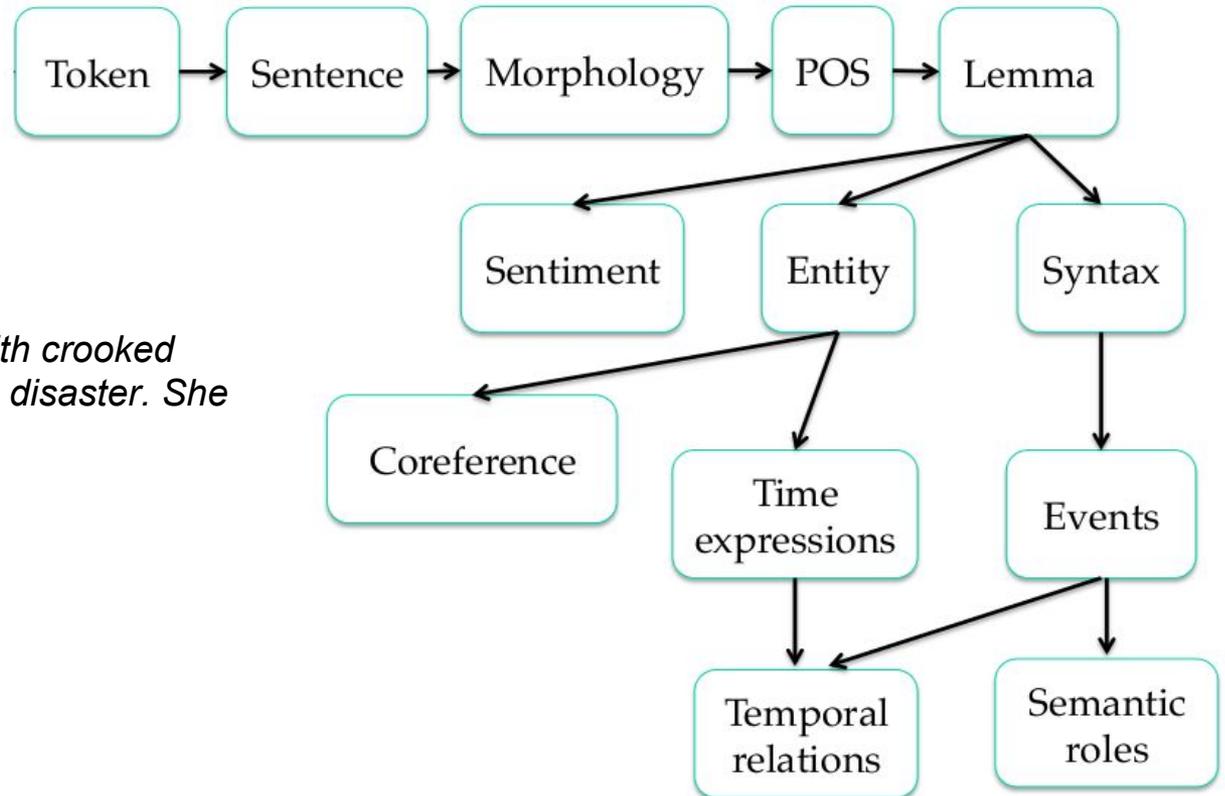
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

EVENTS

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

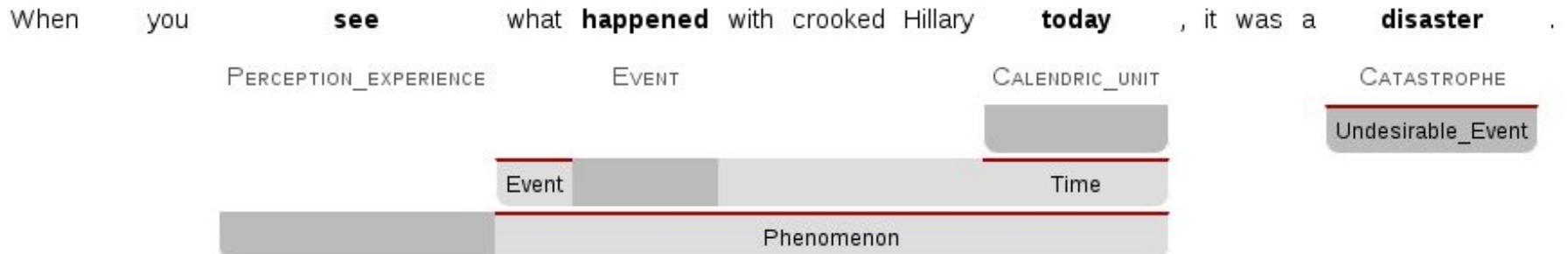
How to process the language?



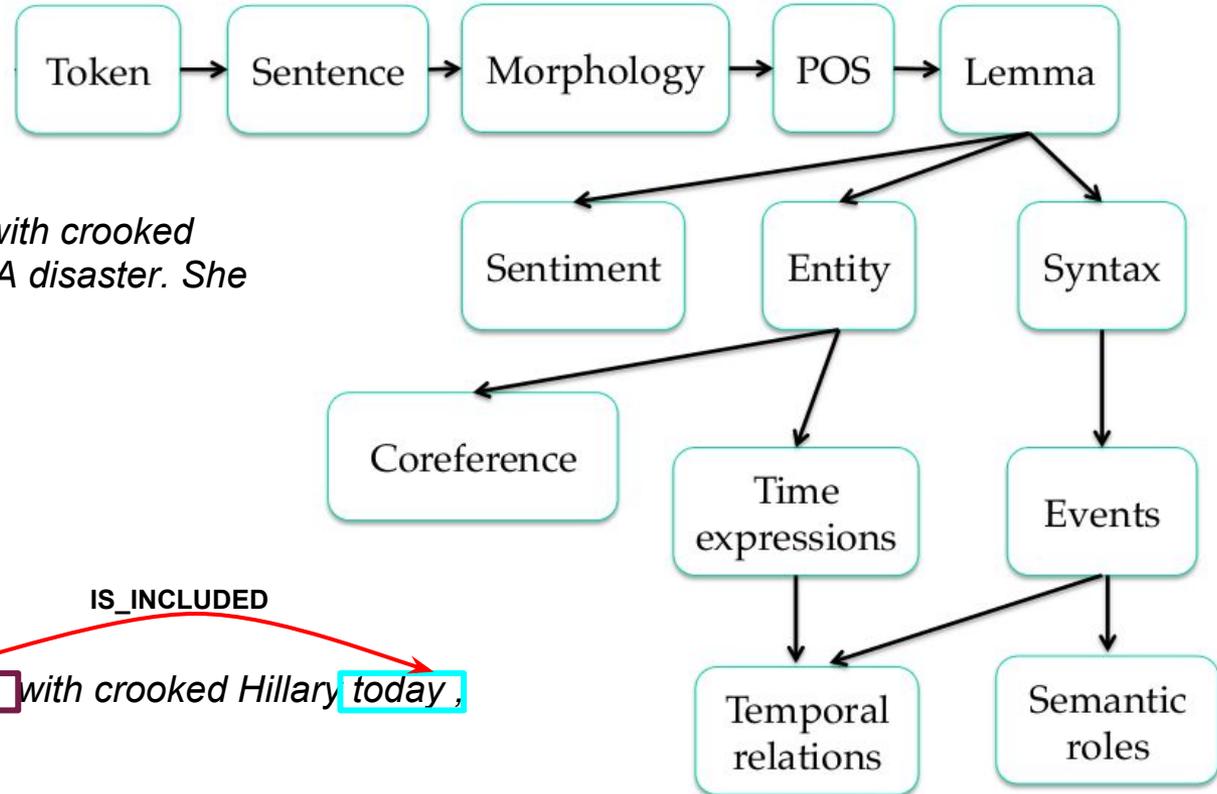
When you see what happened with crooked Hillary today, it was a disaster. A disaster. She had a disaster.

Trump, 2016-08-05

SEMANTIC ROLES



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

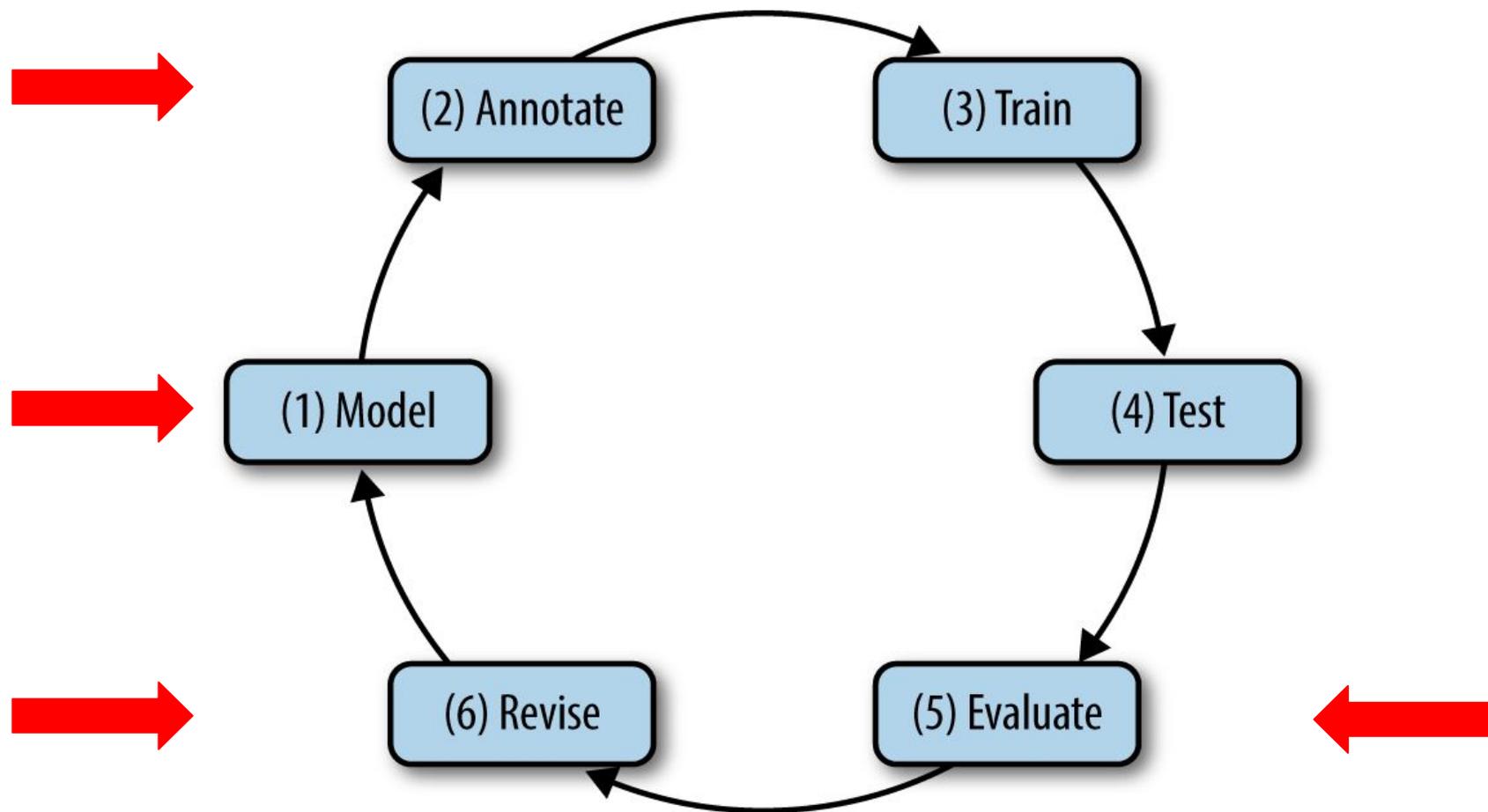
*When you see what **happened** with crooked Hillary **today**, it was a disaster.*

IS_INCLUDED

Mission



The MATTER cycle



From model to annotated data to automatic systems

(Pustejovsky and Stubbs, 2012)

Applications

- Digital History applications:
 1. ALCIDE
 2. RAMBLE ON / LOD NAVIGATOR
 3. MARTIN for HISTORY

1) The ALCIDE Platform

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



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

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

- Moretti, G., Sprugnoli, R., Menini, S., & Tonelli, S. (2016). ALCIDE: Extracting and visualising content from large document collections to support humanities studies. *Knowledge-Based Systems*, 111, 100-112.
- Sprugnoli, R., Moretti, G., Tonelli, S., & Menini, S. (2016). Fifty years of European history through the Lens of Computational Linguistics: the De Gasperi Project. *Italian Journal of Computational Linguistics*.
- Cau, M., & Largaiolli, M. (2017). La piattaforma ALCIDE per l'analisi del discorso politico. Un progetto di ricerca transdisciplinare. *Storicamente*, 12.

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

- Movements of notable individuals can affect social organization and culture
- Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., & Lepri, B. (2017). RAMBLE ON: Tracing Movements of Popular Historical Figures. *EACL 2017*, 77.

2) RAMBLE ON

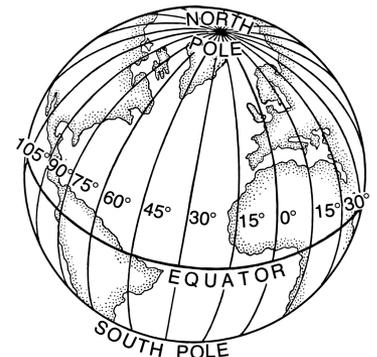
- Our approach: automatic extraction of motion trajectories from Wikipedia biographies

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>



2) RAMBLE ON

- LOD NAVIGATOR

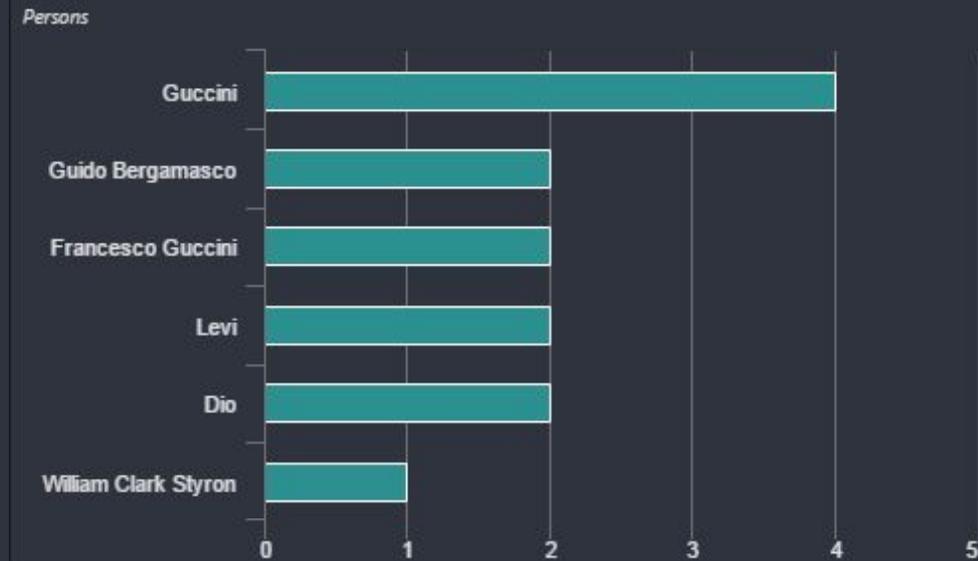


<https://youtu.be/OxGob8INbtM>

3) MARTIN

- *Monitoring and Analysing Real-time Tweets in Italian Natural language*

#giornatadellamemoria



- Sprugnoli, R., Moretti, G., Tonelli, S.. (2017). Twitter Data Exploration for Italian History. AIUCD 2017.

Hands-on Session

- Topic modeling
- Key-concept extraction
- Domain identification

+

- Data visualization for dissemination

Corpus: goo.gl/yQqvVA

Data Visualization

- From words to numbers to visualizations
 - D3.js: <https://d3js.org/>
Examples: networks in ALCIDE
[other](#) dynamic and interactive graphs
 - RAW Graphs: <http://rawgraphs.io/>
Examples: with topic modeling, keyphrase extraction,
domain identification

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 our must change Washington and quickly. Sadly other way. immigration s anybody ever aren't known report on their talk about interests spent 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 in United States summary of wh support sanctu Security, Med welfare for al by making them which will die immigrants.</p>	<p>Social Secu lifetime we immigrants b citizens. And being treated veterans. Rem 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 refu country .</p>	<p>All Americans country, in wonderful, p immigrants are jobs and wag totally protec our nation are people livin everybody. Ar erased -- it lawful immigr if you look a the borders, a are erased, borders, we r And that's r And I have t 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>
--	---	--	---	---

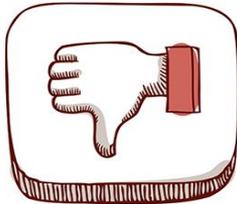
“That’s how topic modeling works in practice. You assign words to topics randomly and then just keep improving the model, to make your guess more internally consistent, until the model reaches an equilibrium that is as consistent as the collection allows.”

Ted Underwood, 2012

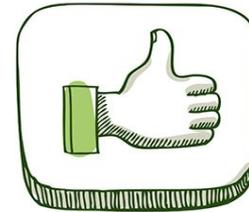
Topic Modeling : Pros and Cons

“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

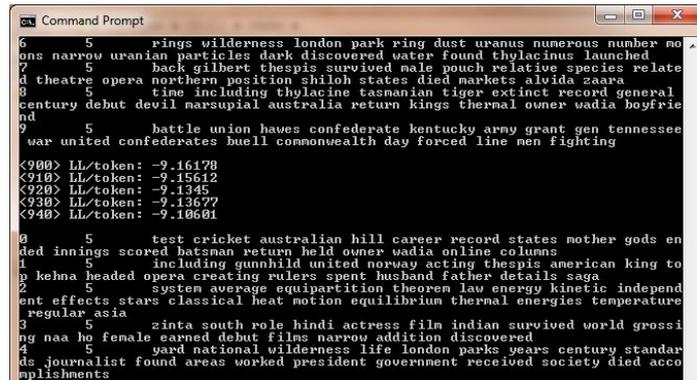


- Good starting point to explore data
- Generates new way of looking to big amount of texts

Topic Modeling : Tools

- MALLET:

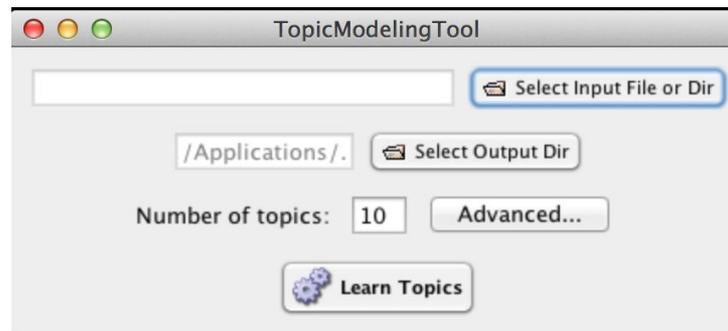
<http://mallet.cs.umass.edu/>



```
Command Prompt
6 rings wilderness london park ring dust uranus numerous number mo
ons narrow uranian particles dark discovered water found thylacinus launched
7 back gilbert thespis survived male pouch relative species relate
d theatre opera northern position shiloh states died markets alvida zaara
8 time including thylacine tasmanian tiger extinct record general
century debut devil marsupial australia return kings thermal owner wadia boyfrie
nd
9 battle union haves confederate kentucky army grant gen tennesse
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 test cricket australian hill career record states mother gods en
ded innings scored batsman return held owner wadia online columns
1 including gunnchild united norway acting thespis american king to
p kehna headed opera creating rulers spent husband father details saga
2 system average equipartition theorem law energy kinetic independ
ent effects stars classical heat motion equilibrium thermal energies temperature
regular asia
3 zinta south role hindi actress film indian survived world grossi
ng naa ho fenale earned debut films narrow addition discovered
4 yard national wilderness life london parks years century standar
de journalist found areas worked president government received society died acco
mplishments
```

- Topic-modeling-tool:

<https://code.google.com/archive/p/topic-modeling-tool/>



Let's try it!

Topic-modeling-tool

- Run the tool on Trump folder and then on Clinton folder (be careful not to overwrite the output folders) obtaining 30 topics for each folder
- Open all_topics.html using a browser and check the topics
- Select two topics in common between Trump and Clinton: e.g. **immigration / work / terrorism**
- Click on the list of keywords associated to those topics; another file opens up
- Copy the content of this file and paste it in a spreadsheet
- Repeat with the other topics both for Trump and Clinton

Topic-modeling-tool + RAW

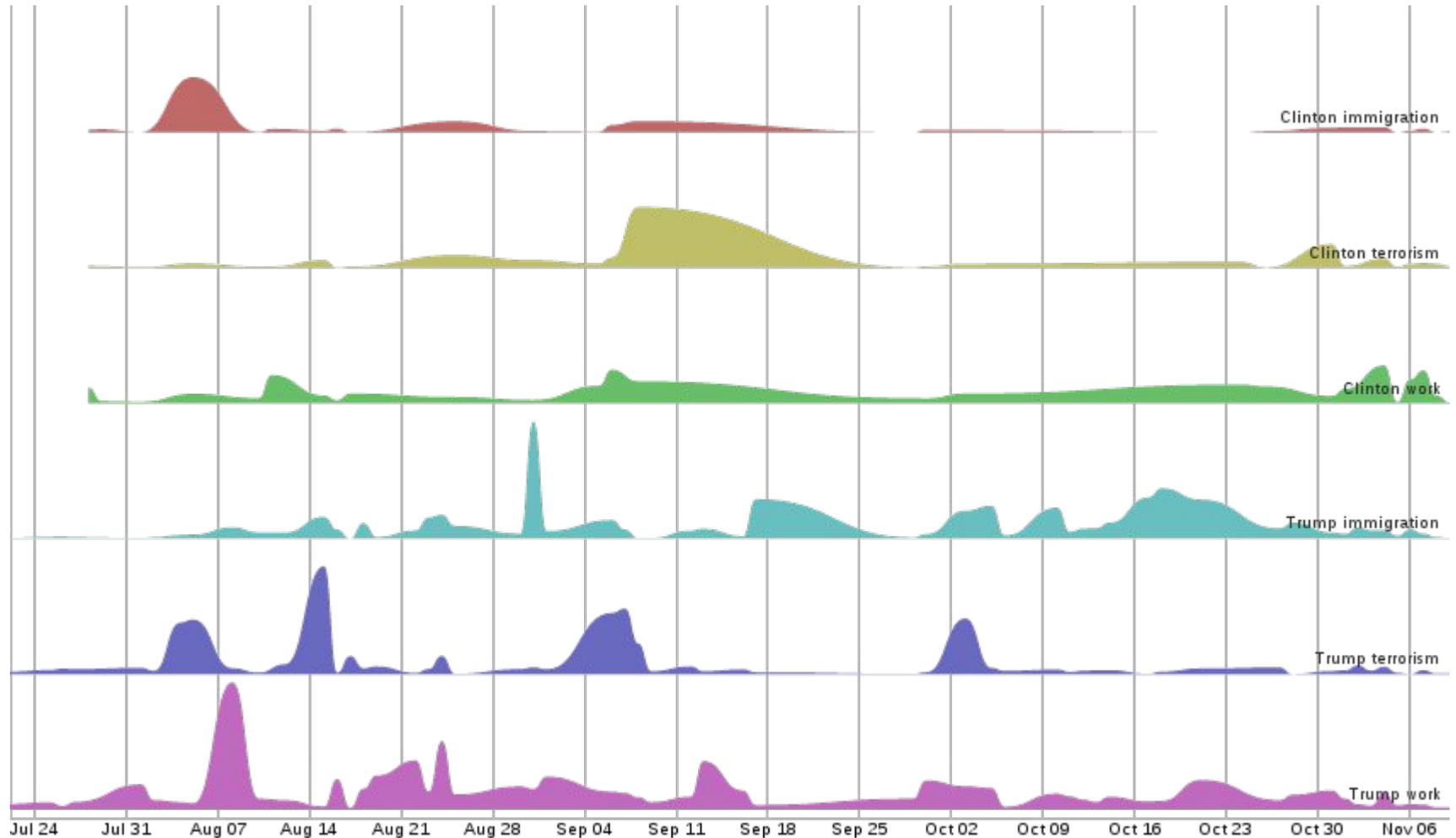
- Manipulate the content of the spreadsheet so to obtain something like this:

	A	B	C	D	E	F
1	speaker	topicID	rank	#words	filename	date
2	Trump	terrorism	2.	621	Trump_2016-08-15.txt	2016-08-15
3	Trump	terrorism	3.	376	Trump_2016-09-07-A.txt	2016-09-07
4	Trump	terrorism	4.	314	Trump_2016-08-05.txt	2016-08-05
5	Trump	terrorism	5.	300	Trump_2016-09-06-A.txt	2016-09-06

- Go to RAW: <http://app.rawgraphs.io/>
- Copy the content of the spreadsheet, select the **area chart** and map the dimensions as follows:

The screenshot shows the 'Map your Dimensions' interface in the RAW application. On the left, there is a list of available dimensions: 'topicID string', 'rank number', '#words number', 'filename string', and 'date date'. On the right, three dimension cards are visible: 'Group', 'Date', and 'Size'. The 'Group' card contains 'topicID string', the 'Date' card contains 'date date', and the 'Size' card contains '#words number'. Each card has a green asterisk icon in the top right corner and a close button (x) in the bottom right corner.

Topic-modeling-tool + RAW



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**.

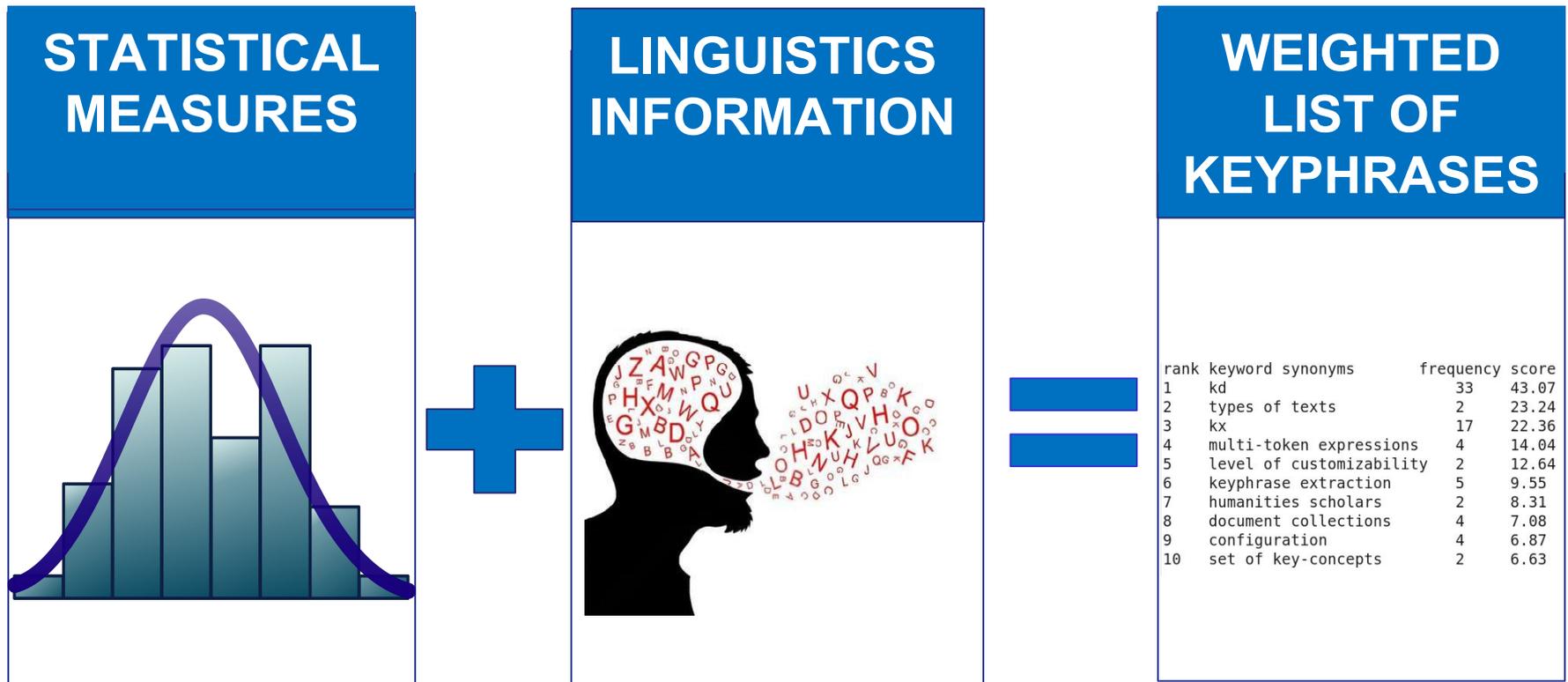
Both single words and multi-token expressions

Both single documents and whole corpora

Keyphrase Extraction: Tool

- KD = Keyphrase Digger

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



KD

- Open the most relevant document for the immigration topic in Trump: *Trump_2016-08-31.txt*
- Run the demo and check the results: how to deal with “country - countries”? <http://textanalysisonline.com/spacy-word-lemmatize>
- Download the results of KD in tsv format by clicking on the “Download Data” link under the word cloud
- Do the same process for Clinton’s most relevant file for the immigration topic: *Clinton_2016-08-05.txt*

KD+RAW

- Copy the content of each tsv file in a spreadsheet and create a file like the following:

	A	B	C	D	E
1	Speaker	Topic	Keywords	Freq	Weight
2	Trump	immigration	country	54	6.7323276399452325
3	Trump	immigration	hillary clinton	19	4.737563896963101
4	Trump	immigration	people	34	4.2388729584840235

- Copy the content of the spreadsheet and paste it in RAW
- Select the **Alluvial Diagram** chart with the following dimensions:

Map your Dimensions

- Speaker string →
- Topic string →
- Keywords string →
- Freq number →
- Weight number →

Steps

Drag numbers, strings, dates here

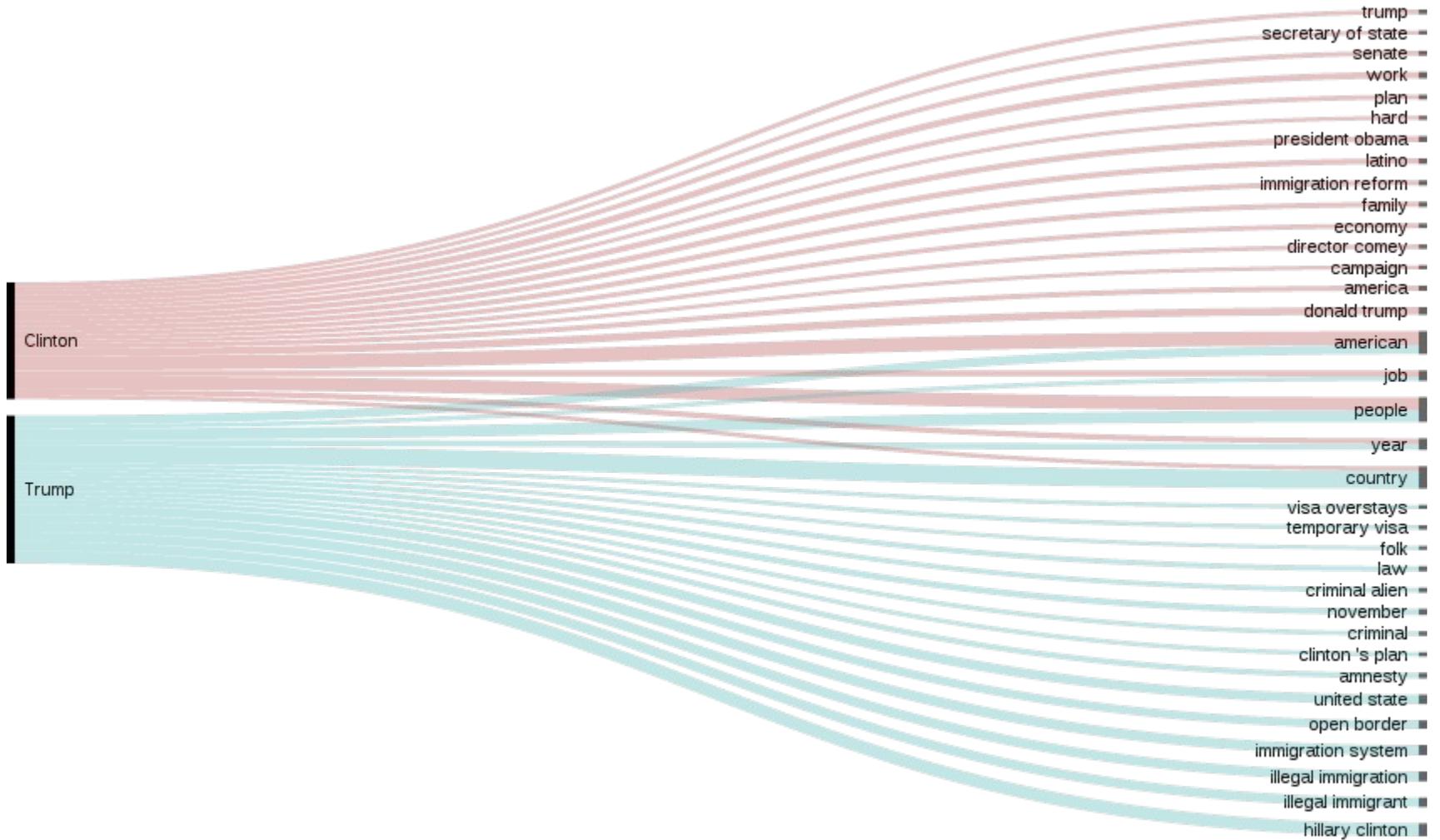
- Speaker string ×
- Keywords string ×

Size

Drag numbers here

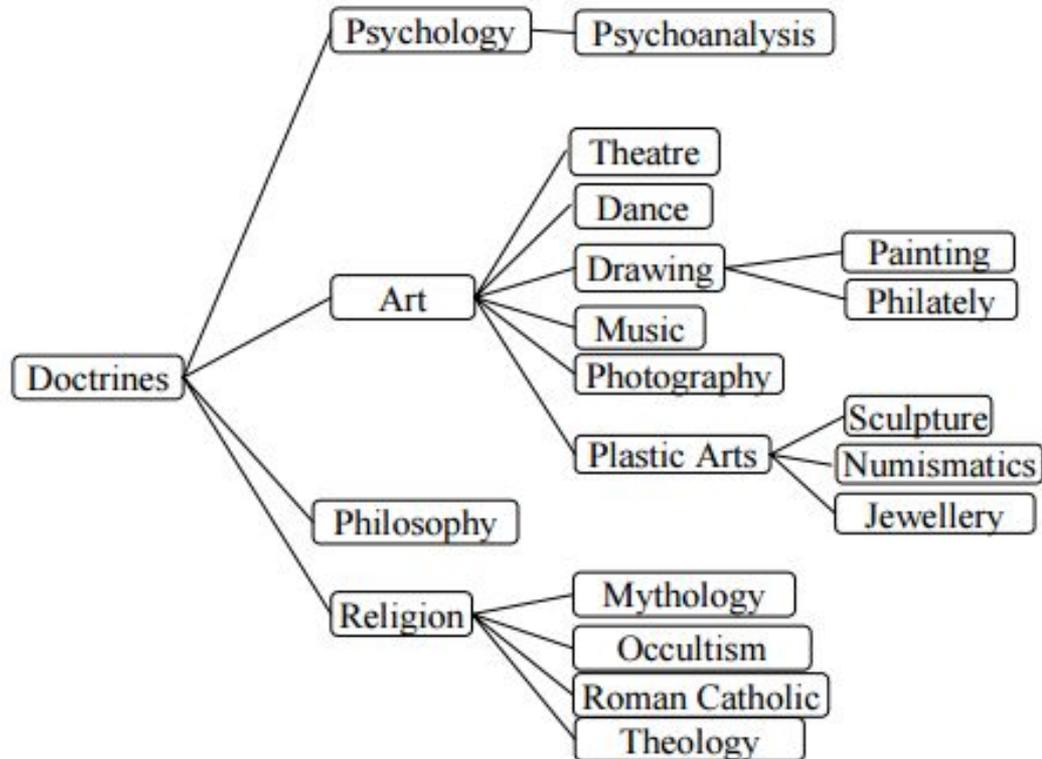
- Weight number ×

KD + RAW



Domain Identification

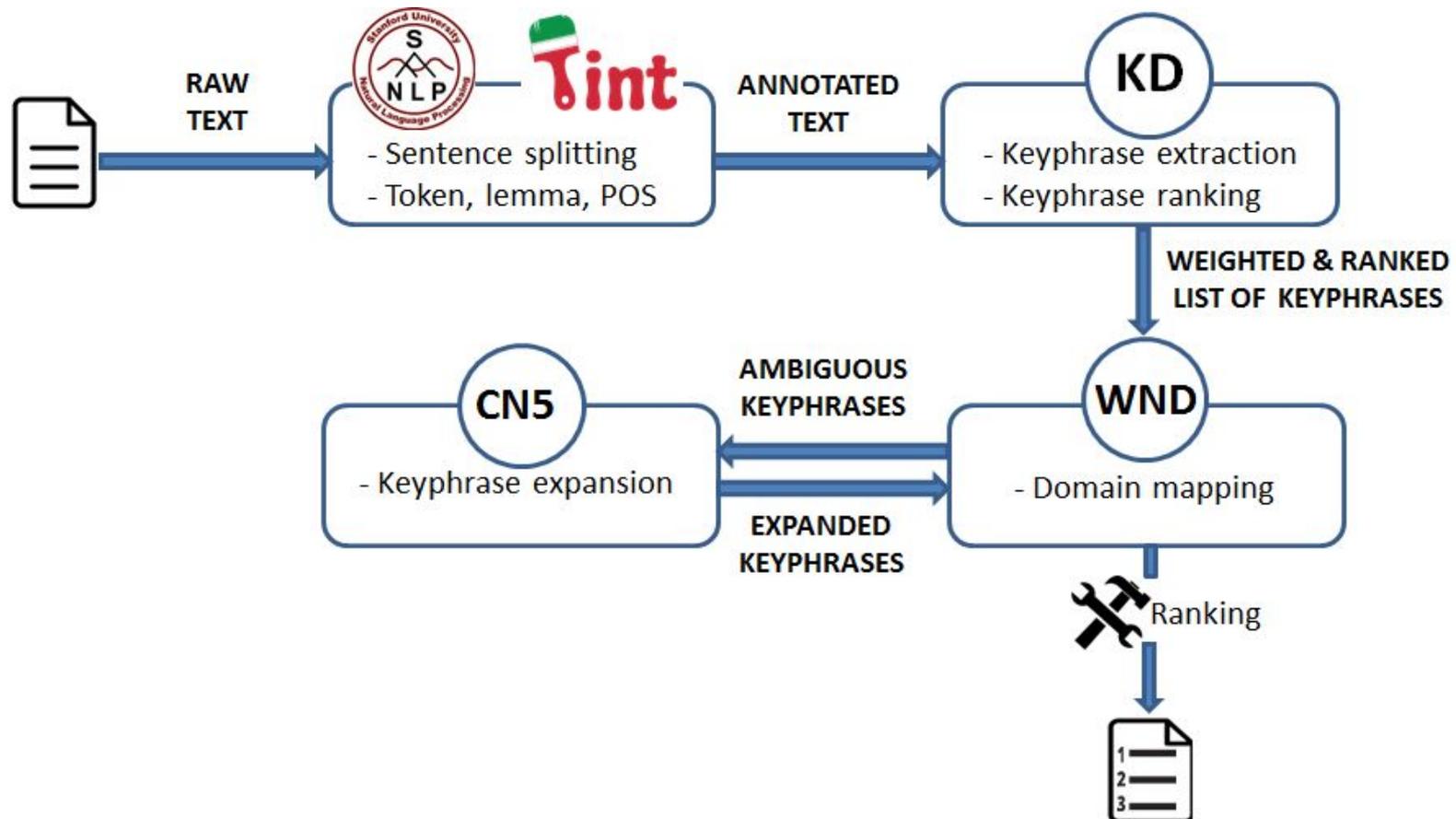
- Create clusters of keyphrases labelled by domain



Domain Identification: Tool

- L-KD: Labelled Keyphrase Digger

http://dhlab.fbk.eu:8080/L_KD/



L-KD

- Open the most relevant document for the immigration topic in Trump: *Trump_2016-08-31.txt*
- Copy the text and paste it in the L-KD online demo: http://celct.fbk.eu:8080/L_KD/
- Copy the output and paste it in a spreadsheet creating a file like the following:

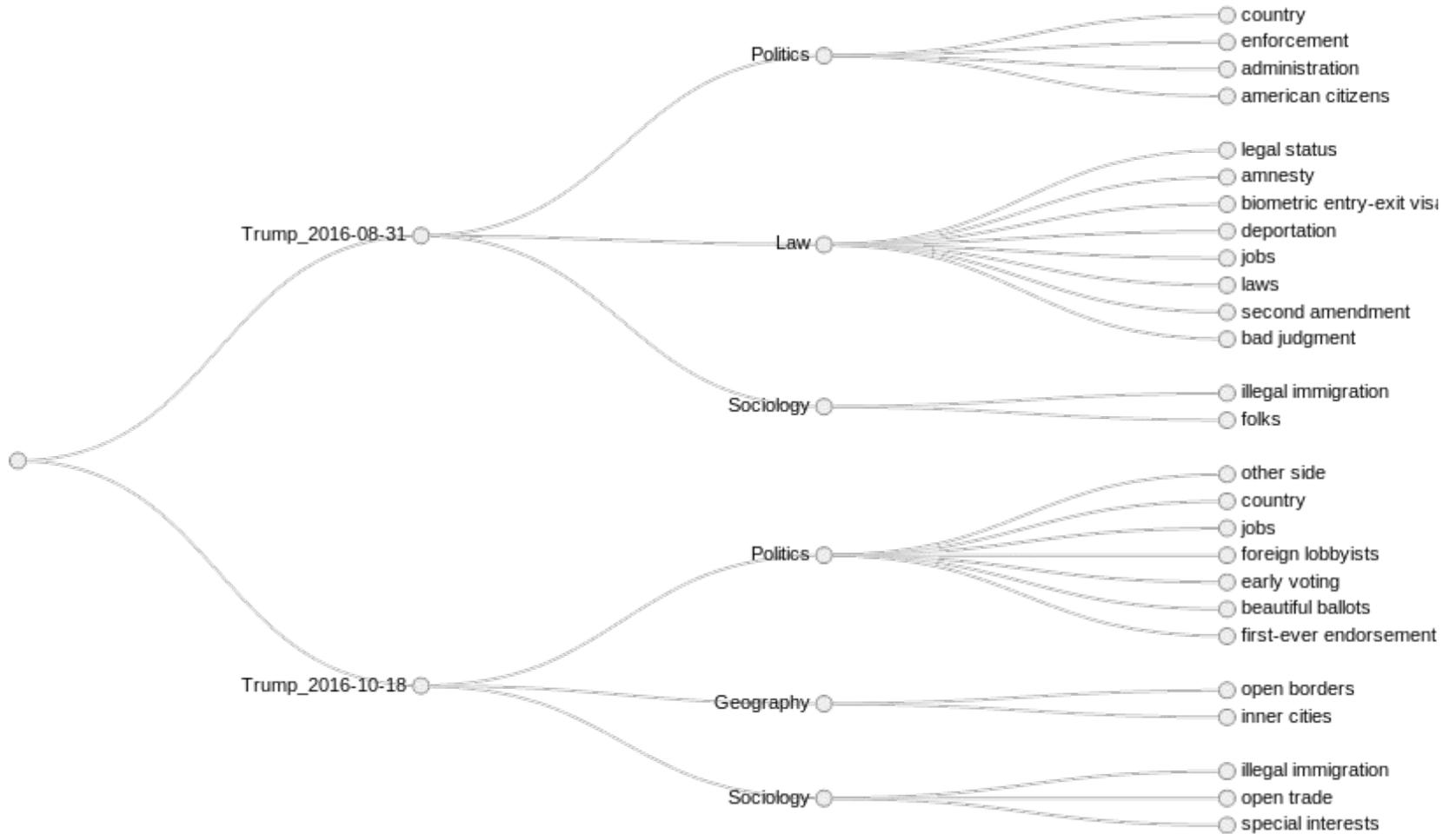
	A	B	C
1	filename	domain	keyphrase
2	Trump_2016-08-31	Politics	country
3	Trump_2016-08-31	Politics	enforcement
4	Trump_2016-08-31	Politics	administration
5	Trump_2016-08-31	Politics	american citizens
6	Trump_2016-08-31	Law	legal status
7	Trump_2016-08-31	Law	amnesty

L-KD + RAW

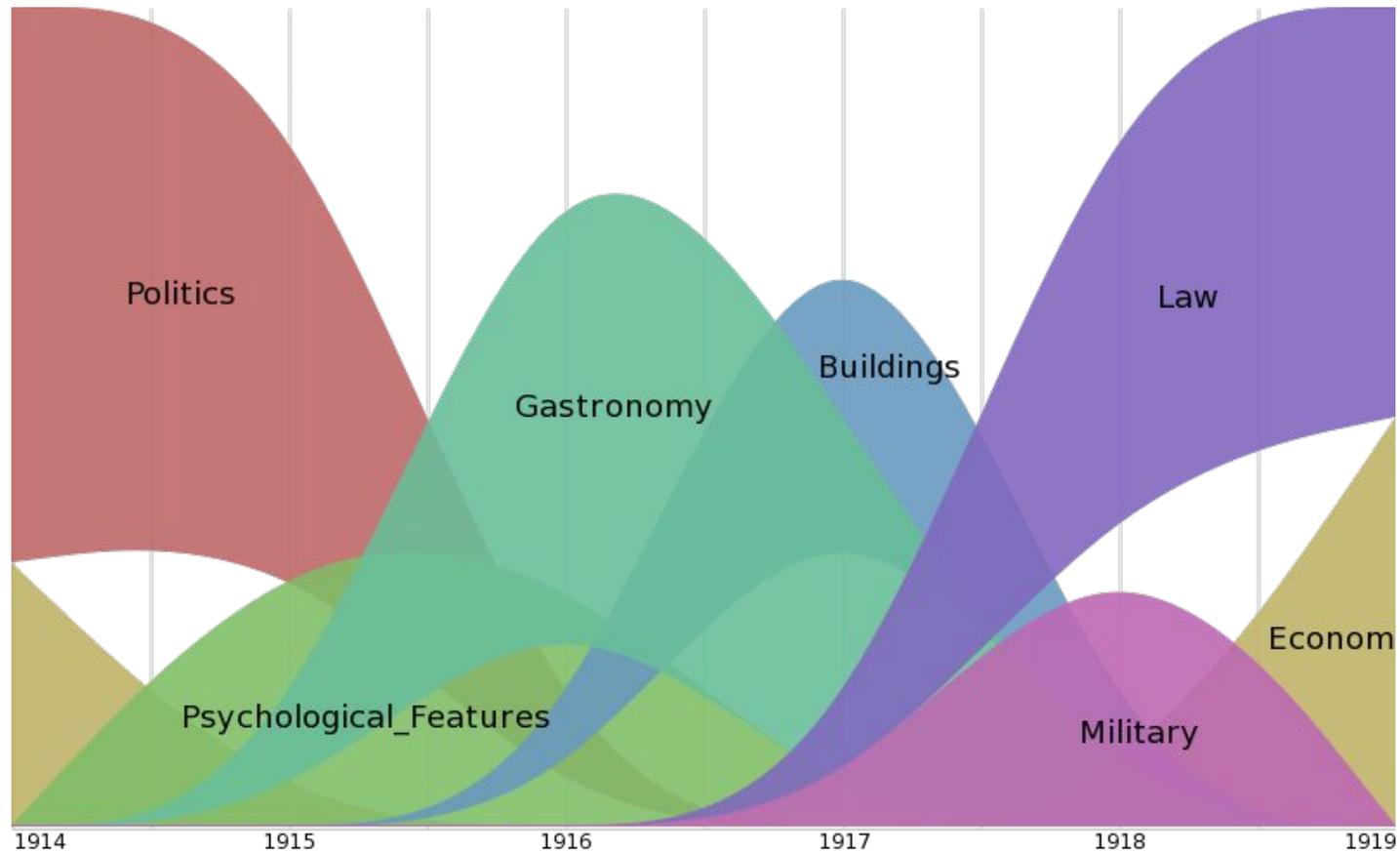
- Repeat for other files, eg. *Trump_2016_10_18.txt*
- Copy the content of the spreadsheet in RAW, select the Cluster Dendrogram with the following dimensions:

The screenshot shows a user interface for mapping dimensions. On the left, under the heading "Map your Dimensions", there are three teal buttons with white text and right-pointing arrows: "filename string", "domain string", and "keyphrase string". On the right, a white panel titled "Hierarchy" contains the instruction "Drag numbers, strings, dates here" and three teal buttons with white text and 'x' icons: "filename string", "domain string", and "keyphrase string".

L-KD + RAW



Domain Identification: Use Case



Top domains in De Gasperi's documents from 1914 to 1918

Conclusions

- In “traditional” NLP research, final users are not in the loop. In Digital Humanities, final users (scholars/historians) play a central role
- Friendly suggestions:
 - Be curious
 - Stay updated
 - Look for interdisciplinary collaborations

**Stay hungry, stay foolish.
-Steve Jobs**



THANK YOU!

Email: sprugnoli@fbk.eu

Web Site: <http://dh.fbk.eu>

Twitter: https://twitter.com/DH_FBK