Natural Language Processing

Aims, tools, consequences

Marco Petolicchio

October 2020

A first overview to NLP

- **NLP** is an acronym for Natural Language Processing
- With the use of computational methods, we are able to retrieve quantitative information about language use, patterns...
- NLP is used either at a morphologycal, syntactical, phonological, semantical level, and also for comparing different languages

- We start with a text, or a collection of texts, as the input for the machine
- Then we elaborate the input in order to parse some information, e.g. as we want to retrieve a syntactical analysis
- We can make the machine to recognize all the Part-of-Speech for each word (POS-tagging)

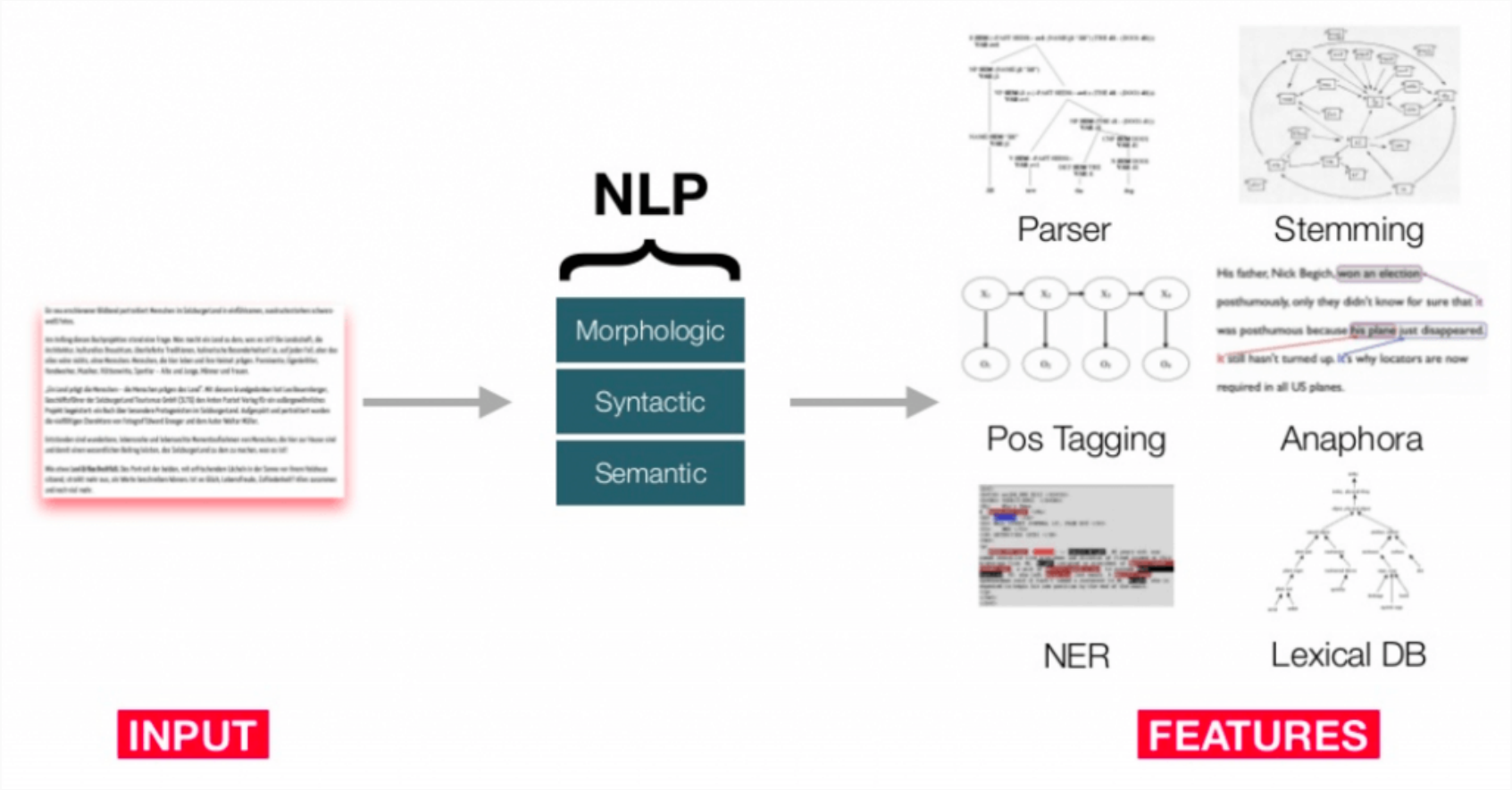

Figure 1: NLP Pipeline (from `wordlift.io`)

- The use of NLP is a powerful tool in the field of Human-Machine interaction, as in automated translation, summary creation from texts, assistive technologies, speech analysis, automated caption generation, recognition of text from scan and images (OCR), and so on
- But how can a machine work this manner?

Corpora

- The first thing we should be aware of is that the machine works using a set of information as a model
- A text should be compared to other texts in some respects, in order to retrieve quantitative information

- We can define a **corpus** as a collection of texts, with a certain grade of homogeneity, that we can use for a certain kind of generalization about language use and structures

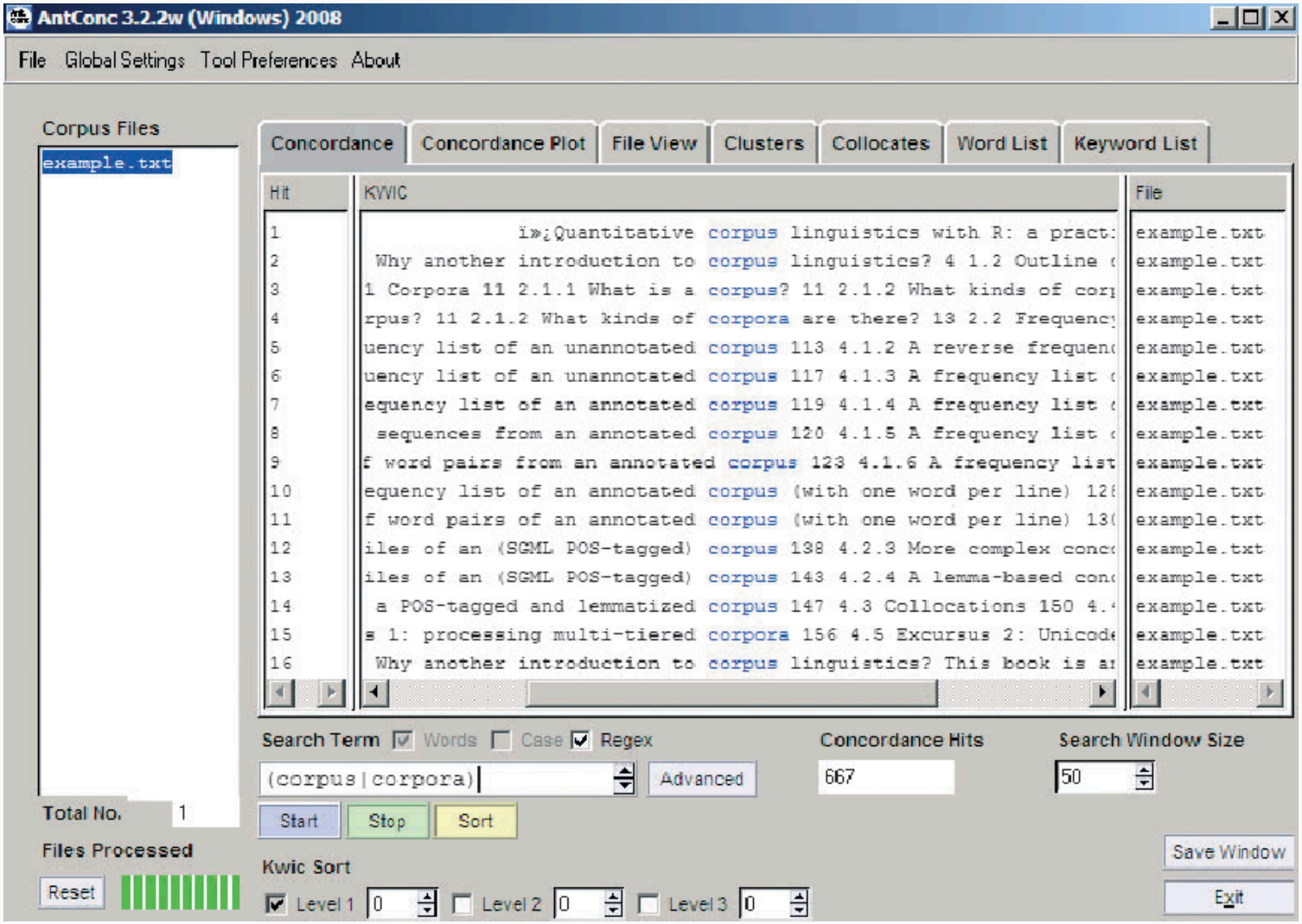

Figure 2: Example of concordances in a corpus

Critics on Corpus Linguistics

Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list.
(Chomsky 1962; 159)

A good question now can arise:

*what we define with the term **text**?*

Types of texts

As we said, a corpus is a collection of texts.

- Texts do not mean that they are solely related to written language, because we can have audio transcript as well
- Texts are collections of words (which are collections of letters (XD) ?)
- Barely said, words are **tokens**

Markup

- In order to provide a better input for our software, it could be useful to add some information to the texts. In informatics, we can add some pieces of information to our work, using *metadata* and *markup languages*.
- For example, if we want to annotate a certain text, specifying where is the subject in each sentence, we can add this information and use it later. We can use for doing it a markup language, which has a clear structure and encodes all these information in a formal way. Examples of markup languages are HTML (the language for the majority of the web pages), XML (which is used in RSS feeds), and from XML we have different dialects as XML-TEI (for encoding literary texts, manuscrips, etc..), MML (for encoding MIDI files for music), KML (for geospatial information), etc...

- In computational linguistics the quite-default markup language is the CONLL-U language, which provides syntactic, morphological, dependency information.

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	NER
# newdoc url = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html									
# newdoc s3 = s3://aws-publicdatasets/common-crawl/crawl-data/CC-MAIN-2016-07/segments...									
...									
# sent_id = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html#60									
# text = The American Museum of Natural History was established in New York in 1869.									
0	The	the	DT	DT	_	2	det	2:det	O
1	American	American	NNP	NNP	_	2	nn	2:nn	B-Organization
2	Museum	Museum	NNP	NNP	_	7	nsubjpass	7:nsubjpass	I-Organization
3	of	of	IN	IN	_	2	prep	_	I-Organization
4	Natural	Natural	NNP	NNP	_	5	nn	5:nn	I-Organization
5	History	History	NNP	NNP	_	3	pobj	2:prep_of	I-Organization
6	was	be	VBD	VBD	_	7	auxpass	7:auxpass	O
7	established	establish	VBN	VBN	_	7	ROOT	7:ROOT	O
8	in	in	IN	IN	_	7	prep	_	O
9	New	New	NNP	NNP	_	10	nn	10:nn	B-Location
10	York	York	NNP	NNP	_	8	pobj	7:prep_in	I-Location
11	in	in	IN	IN	_	7	prep	_	O
12	1869	1869	CD	CD	_	11	pobj	7:prep_in	O
13	.	.	.	.	_	7	punct	7:punct	O
...									

Figure 3: Example of CONLL-U

Model

As we have seen, a corpus can be annotated in order to result in a better **model** for the machine, an input that permits to the various algurithms the possibility to perform different analyses on the text that we want to process.

There are different technologies that we can use for doing it ==>

Artificial Intelligence, Machine Learning, and other things

NLP and Computational Linguistics are two faces of the same medal: one is problem-oriented and it’s the technology where the theoretical approach grounds on.

With the use of Artificial Intelligence, we can elaborate a bit our understanding of linguistic facts.

Machine learning (and Deep Learning) is the application of the algurithm based on the assumptions made upon the model we have as input.

We continue it later