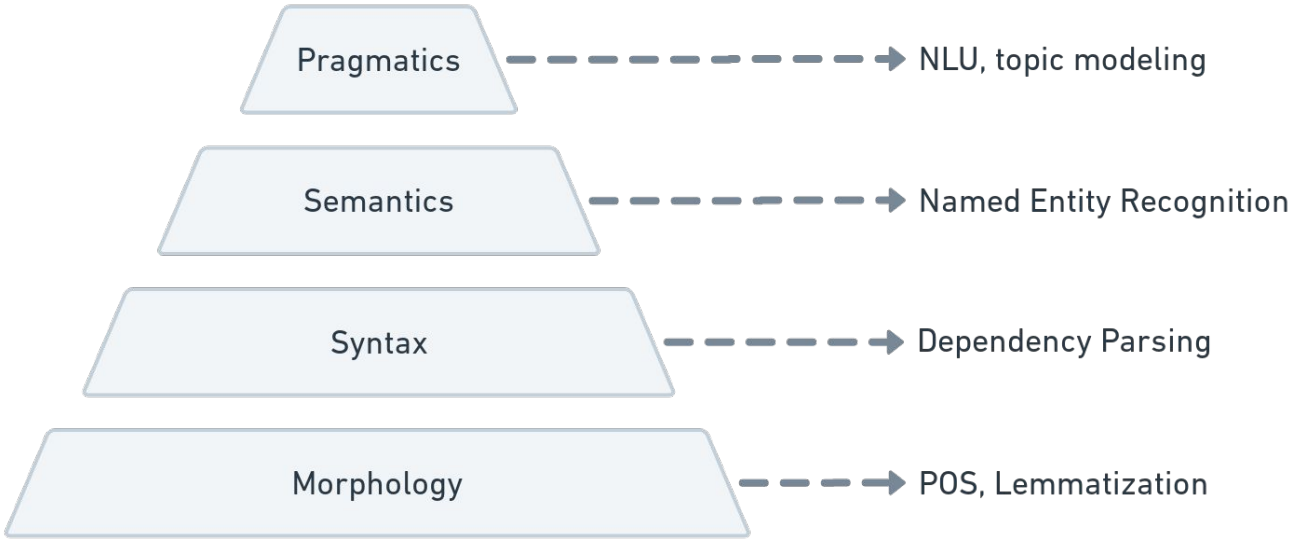
Text data preprocessing

Case Studies: NLP in Social Sciences

2022L-WB-NLP, 10.03.2022

MI EDUCATION

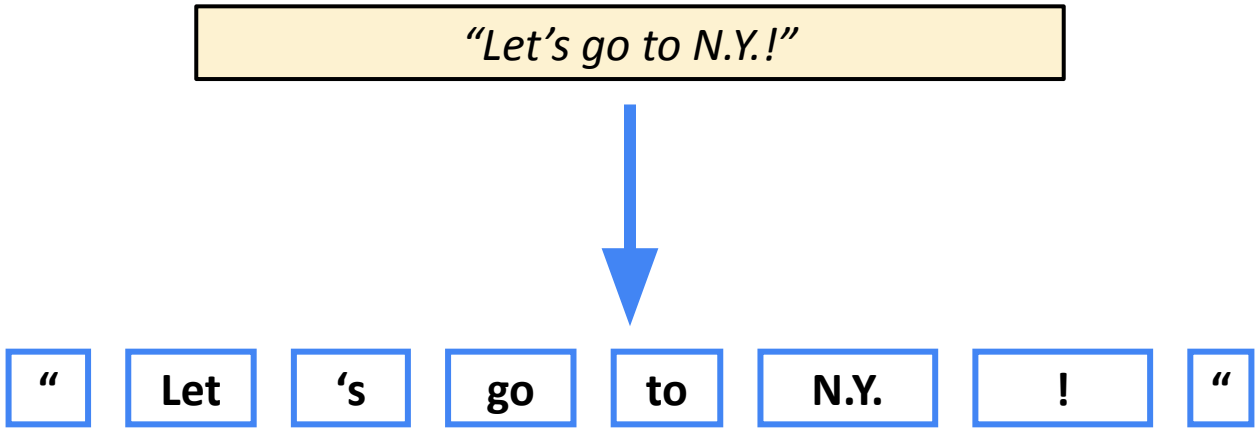
NLP tasks associated to each level of language understanding



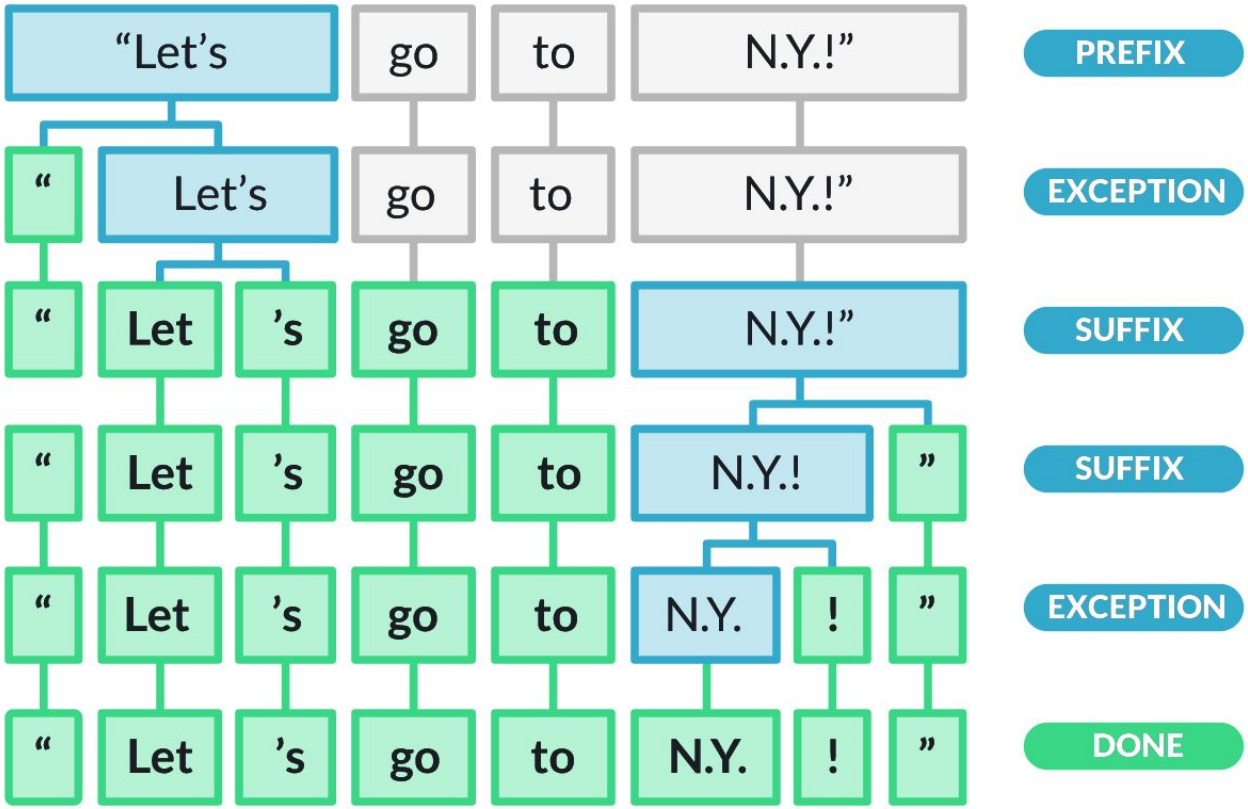
Exemplary preprocessing pipeline

1. Tokenization
2. Lemmatization
3. Stop words removal

Tokenization

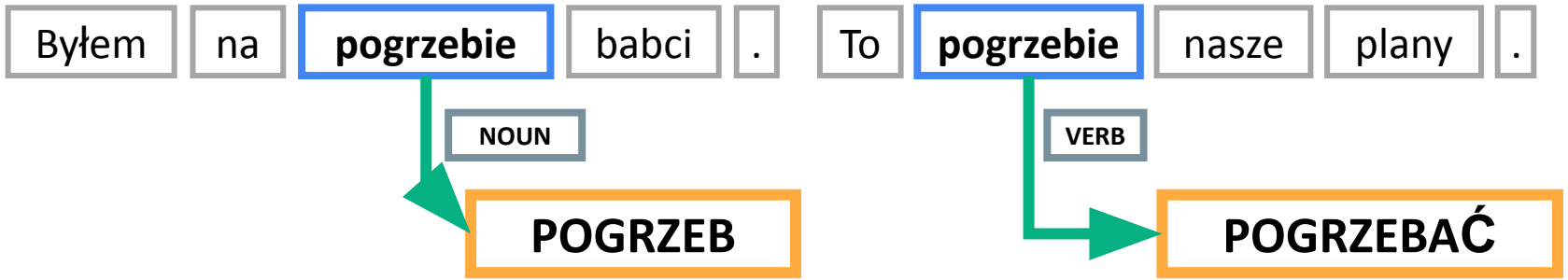


Tokenization algorithm

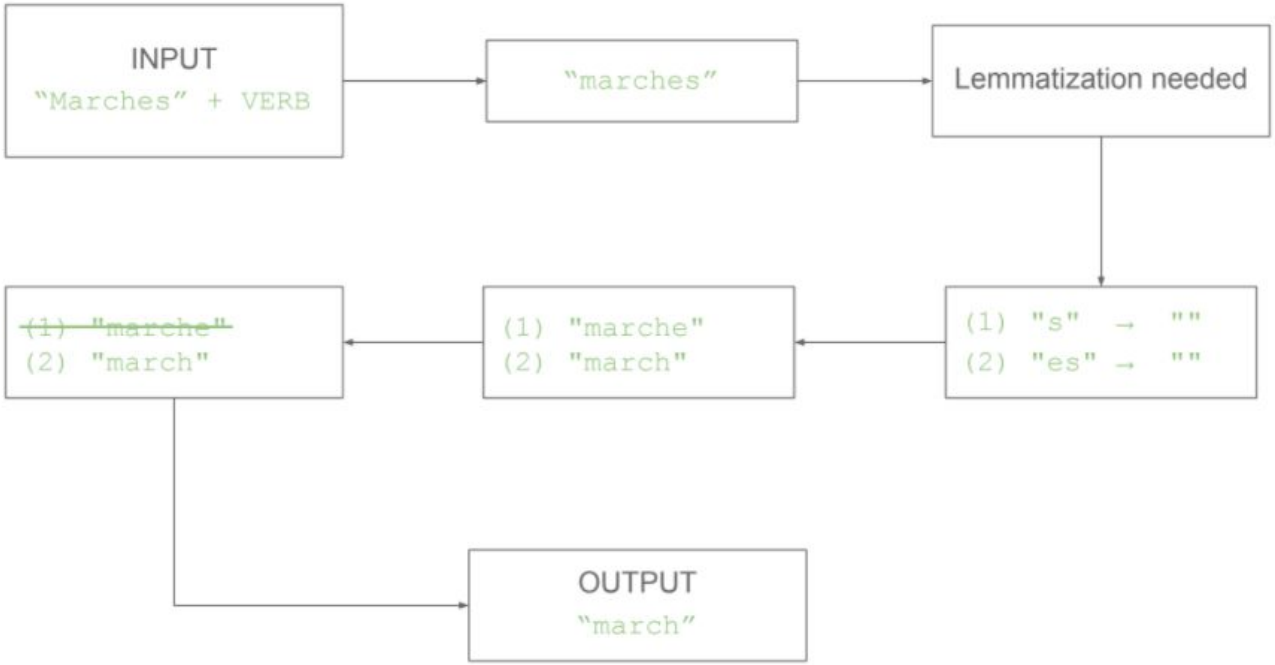


Lemmatization

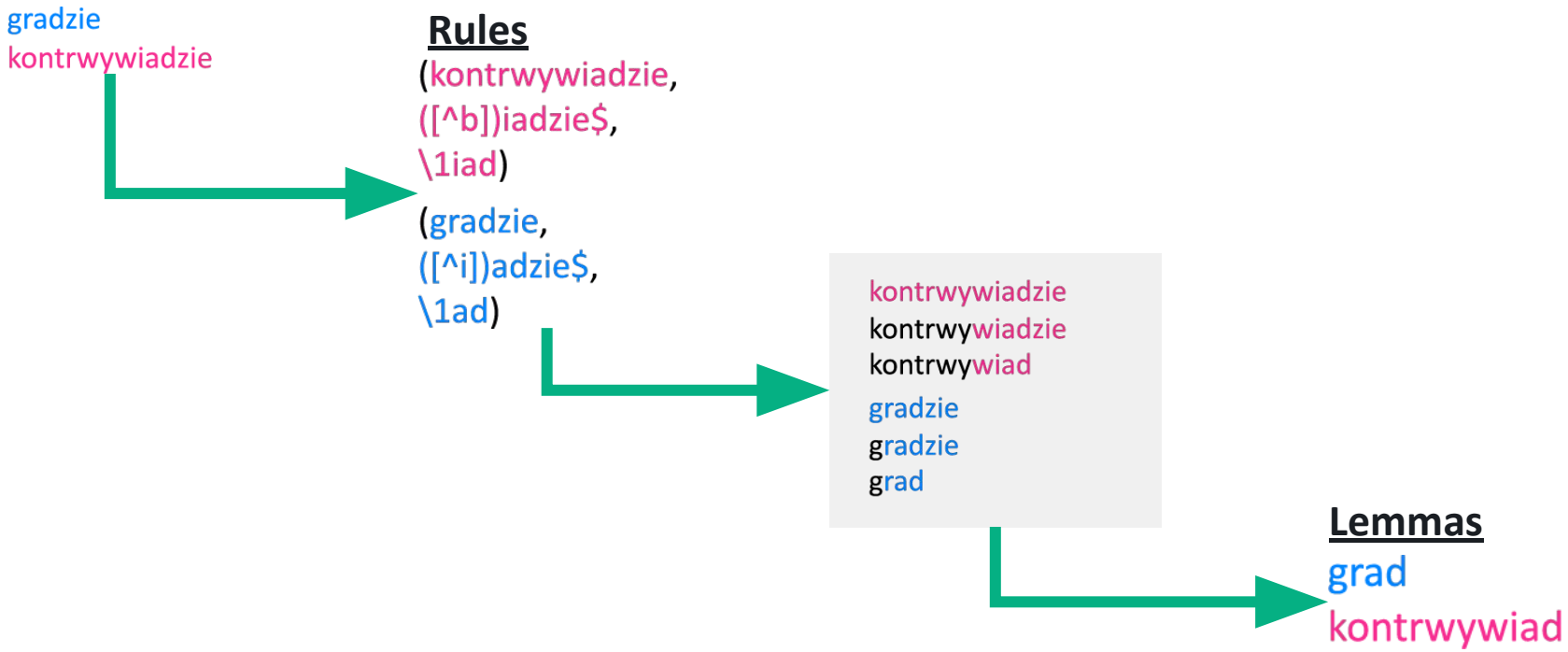
Converting word to its *lemma* (dictionary form)



Lemmatization algorithm



Lemmatization algorithm



Stop words removal

Stop words – words which in the context of the downstream task do not provide additional information.

Examples:

‘a’, ‘the’, ‘no’, ‘yes’

Domain stop words must be hand-crafted

Text representation for models: Bag of Words

1. "John","likes","to","watch","movies","Mary","likes","movies","too"
2. "Mary","also","likes","to","watch","football","games"

1. {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
2. {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};