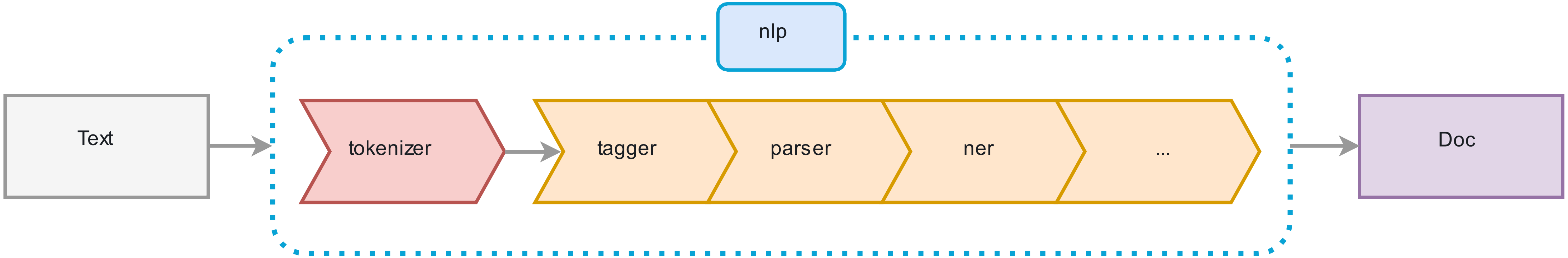
Integrating Polish Language Tools and Resources in **spaCy**

Ryszard Tuora, Łukasz Kobyliński

IPI PAN

Wrocław, October 2019

spaCy



spaCy is a **general-purpose**, **open-source** library for NLP in Python.

It is aimed at **ease of access** and use in **production**.

Last month alone, it was downloaded **738,043** times.

A standard pipeline consists of a **tagger**, **parser** and **NER** components.

Newest version (2.2) was released on Octobert 1st.

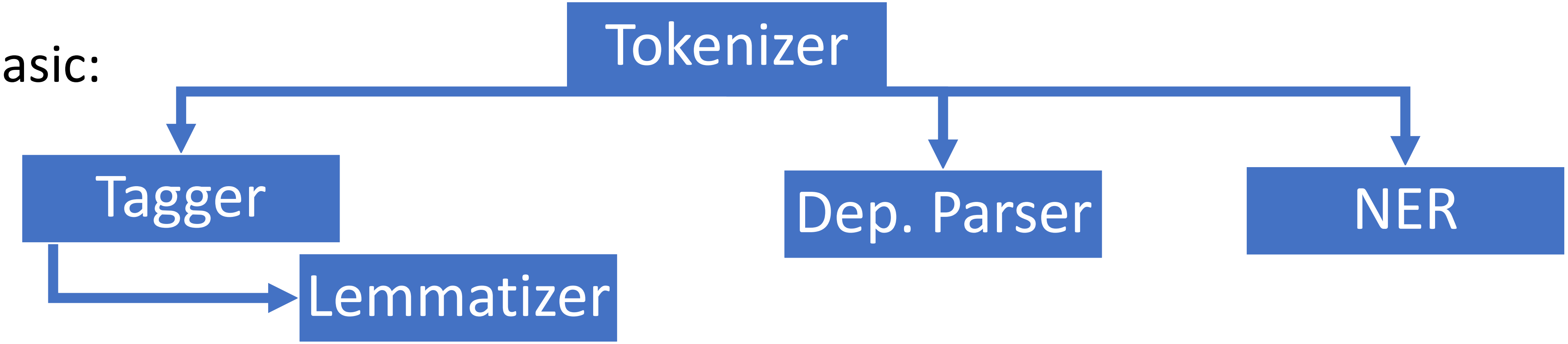
spaCy for Polish

- 23 models for 11 languages
- No official models for Polish
- Existing NLP resources, and solutions to integrate
- No other pipelines integrating NLP for Polish

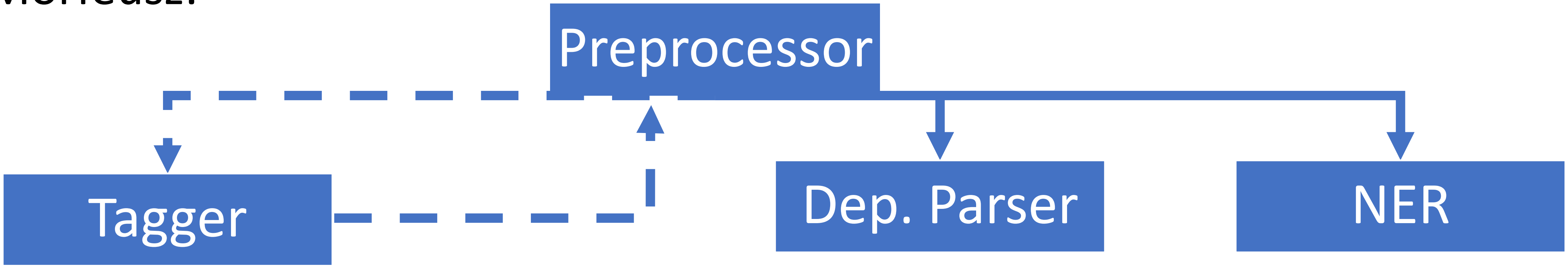
Let’s create **spaCy**-PL!

2 versions of spaCy-PL

Basic:



Morfeusz:



Using spaCy-PL

```
>>> import pandas # for visualization
>>> import spacy
>>> nlp = spacy.load("pl_spacy_model")

# Tokenization, Tagging, Lemmatization and Dependency Parsing
>>> sent1 = "Granice mojego języka oznaczają granice mojego świata" # ~Wittgenstein
>>> parse1 = nlp(sent1)
>>> attribs = ['orth_', 'lemma_', 'tag_', 'pos_', 'dep_', 'head']
>>> table = [{att:tok.__getattribute__(att) for att in attribs} for tok in parse1]
>>> df = pandas.DataFrame(table)
>>> print(df[attribs])

       orth_    lemma_   tag_  pos_      dep_       head
0    Granice   granica  SUBST  NOUN     nsubj  oznaczają
1     mojego       mój    ADJ   ADJ  det:poss     języka
2     języka     język  SUBST  NOUN  nmod:arg    Granice
3  oznaczają  oznaczać    FIN  VERB      ROOT  oznaczają
4    granice   granica  SUBST  NOUN       obj  oznaczają
5     mojego       mój    ADJ   ADJ  det:poss     świata
6     świata     świat  SUBST  NOUN      nmod    granice
```

Plans for the future:

- Further optimization
- Additional components (e.g. a chunker, sentiment analysis component)
- Models of different sizes (e.g. a 20 MB model for quick and easy tasks)
- Better integration with Morfeusz
- Tell us!

Thank you for your attention!