Digital Humanities Across Borders

Class 9: Part-of-speech tagging

So far, we've all been in the same boat

MALLET
machine learning for language toolkit

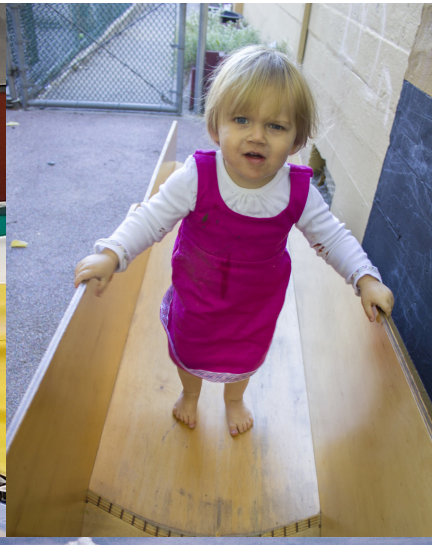
VOYANT
see through your text

Counting different kinds of words

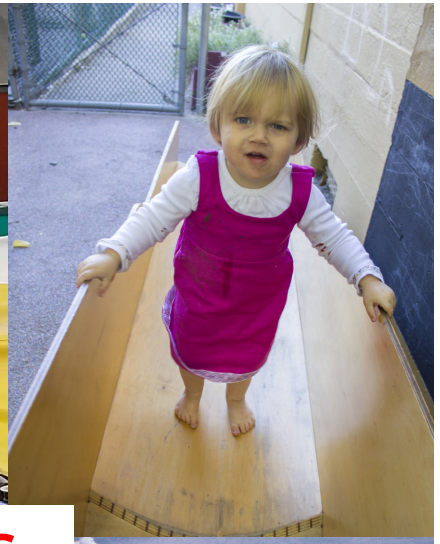
Counting different kinds of words

To each language its own boat(s)

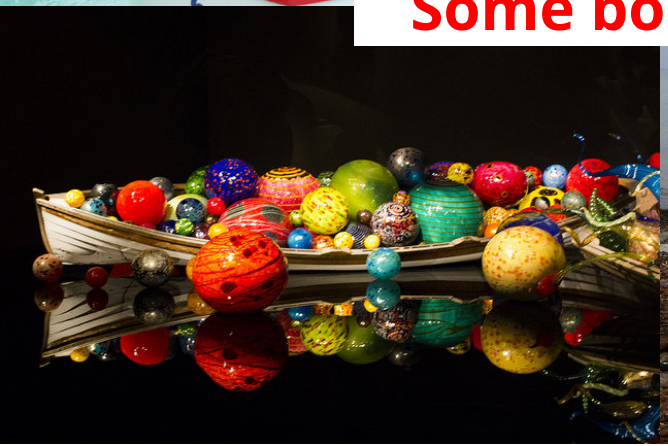





To each language its own boat(s)



Some boats work better than others



Enter the treebank

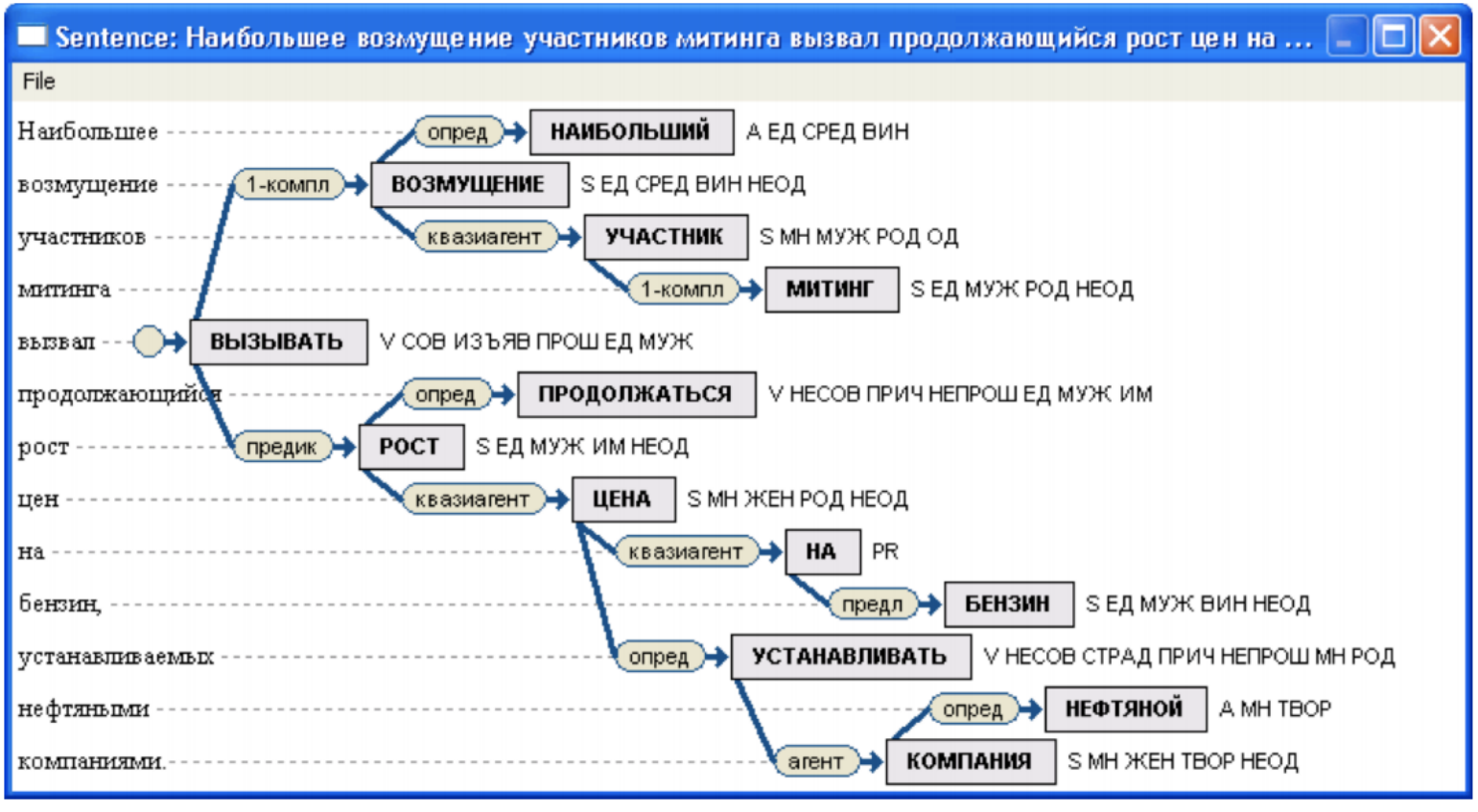


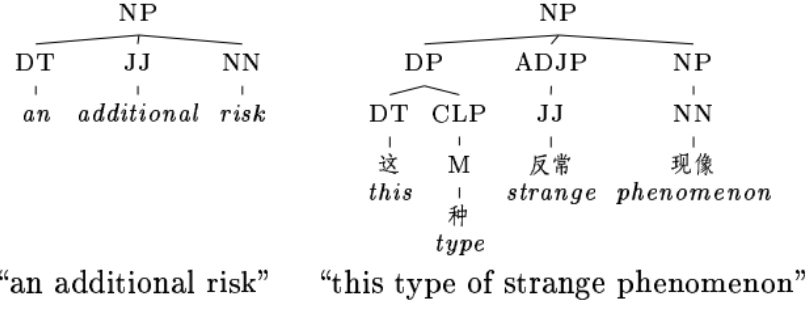
Figure 1: A syntactically annotated sentence from the SYNTAGRUS treebank.

From "Parsing the SYNTAGRUS Treebank of Russian"

Penn Treebank

- 40,000 sentences of Wall Street Journal newspaper text annotated with phrase-structure trees
- Automatic parsing + manual correction
- 3 years of work
- "10 years of studying the Wall Street Journal"

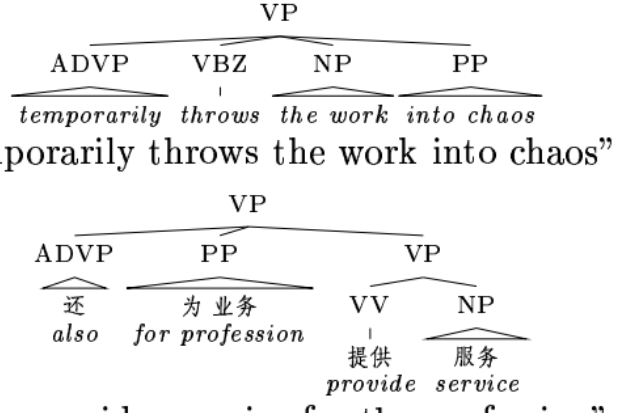
Based on work done using English (WSJ)



"an additional risk" "this type of strange phenomenon"

Figure 1: Noun modification in English and Chinese Treebanks

From "Is it harder to parse Chinese, or the Chinese Treebank?"



"temporarily throws the work into chaos"

"also provides service for the profession"

How POS tagging works

How POS tagging works

- Language-dependendent segmentation and tokenization rules (identifying words and sentences)

How POS tagging works

- Language-dependendent segmentation and tokenization rules (identifying words and sentences)
- Look up each token in a dictionary to discover the lemma and possible parts-of-speech

How POS tagging works

- Language-dependendent segmentation and tokenization rules (identifying words and sentences)
- Look up each token in a dictionary to discover the lemma and possible parts-of-speech
- Disambiguate (or don't, depending on your POS tagger! Looking at you, MyStem...)

News-centric text

Text with code-switching is hard

You can create your own models, using the properties you care about!

Let's try it out!