MUNI

Corpora

PA153 – Natural Language Processing

Pavel Rychlý

Corpora

Corpora

Corpus

- collection of natural language texts
- text written by users
- not generated by machines
- big size

Corpus content

- language: Czech, English, usually only one
- real authentic usage by humans
- written/spoken
- could be domain specific
  - FI web
  - Shakepeara plays
  - old language
- explore at SketchEnginej

Corpus sizes

- the bigger the better
- are often limited by the text source
  - Shakespeare will never write more
- first coprus: 1 million words
  - too small for more interesting results
  - sentence/word length, most frequent words
- now commonly billions of words
  - average reading speed is 125–225 words per minute
  - 200 * 60 * 18 = 216,000 words per day (18 hours)
  - 79 million per year (365 days)

Corpus sizes

- we also work with eighty billion word corpora
  - roughly 1000 years of reading at 18 hours a day
- ChatGPT (2023)
  - trained on 300 billion words (web, books, wikipedia, …)
  - mostly English
  - many non-English texts

Creating corpora: data sources

- document databases (doc, pdf, …)
- datasets (XML)
- news feeds (RSS)
- web

Creating corpora: web

- downloading pages from the web
- usually the largest source
- readily available, for any language
- crawler (SpiderLing)
  - crawls pages, follows links
  - tracks language, yield (how much text from downloaded data)
  - parallel downloads from multiple servers
  - decent handling (doesn't overload)
- removal of headers, footers, menus, ads, …
- up to several billion words per week

Creating corpora: filtering

- language detection (delete/separate)
- unwanted content detection
- duplicate removal

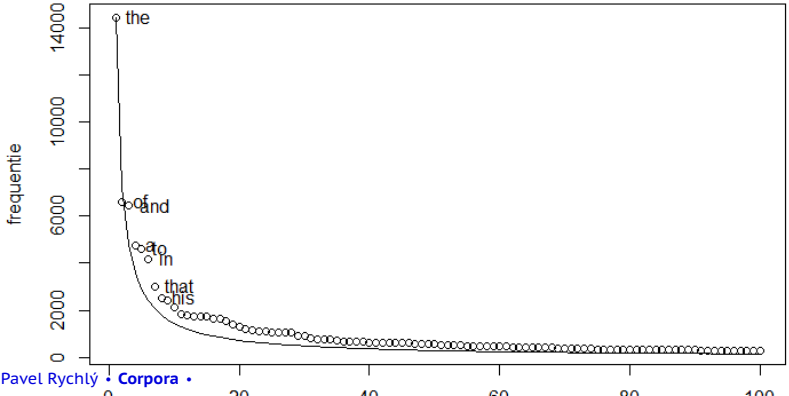
Unwanted content

- types: spam, generated content, noise, machine translation
- detection
  - depends on the angle of view
  - copywriting doesn't matter for learning the language, it matters for getting for information
- often only visible from the result
  - need to identify source/reason
  - repeat processing

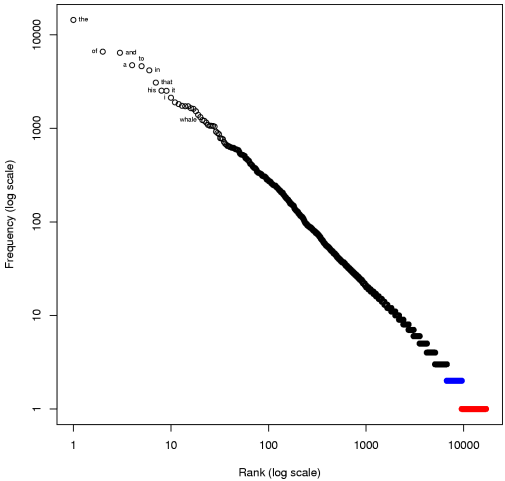
Zipf's Law

Zipf's Law

- rank-frequency plot
- rank × frequency = constant
- highly skewed distribution



Zipf's Law


the
of
and
to
a
in
that
I
it
whale
Frequency (log scale)
10000
1000
100
10
1
1
10
100
1000
10000
Rank (log scale)

Morphology

Tokenization

- splitting text into tokens (positions)
- token = basic unit of the corpus
- mostly word, number, punctuation
- sometimes multi-word: *New York*, *out of*
- sometimes parts of words: *don't* = do + n ' t

Tagging

- morphological
  - basic forms
  - word types (noun, verb, …)
  - grammatical categories (gender, number, case, …)

Universal Dependencies

- collection of treebanks
- annotation guidelines
- https://universaldependencies.org/
- current version: 2.14, 283 treebanks, 161 languages
- new version every 6 months

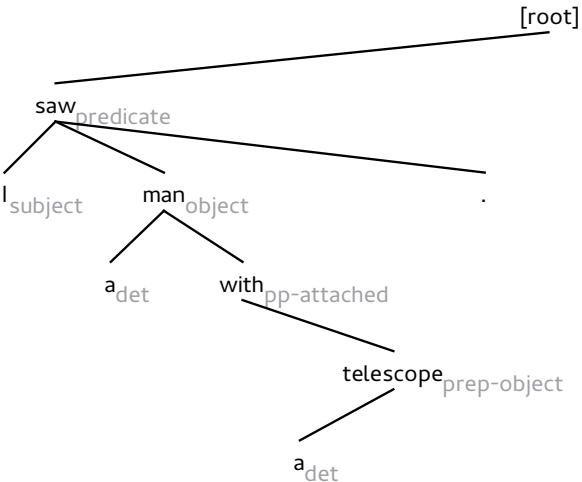
Universal Dependencies

```
# newpar id = vesm9211-001-p7
# sent_id = vesm9211-001-p7s1
# text = Všechny tři světy si vzájemně trvale povídají a ovlivňují s
# orig_file_sentence vesm9211_001#8
Všechny    DET    Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|PronTy
tři        NUM    Case=Nom|Number=Plur|NumForm=Word|NumType=Card|NumVa
světy      NOUN   Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|Polari
si         PRON   Case=Dat|PronType=Prs|Reflex=Yes|Variant=Short
vzájemně   ADV    Degree=Pos|Polarity=Pos
trvale     ADV    Degree=Pos|Polarity=Pos
povídají   VERB   Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polarity=Po
a          CCONJ    _
ovlivňují  VERB   Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polarity=Po
se         PRON   Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short
```

Parsing

- syntax
  - nominal phrases
  - word dependencies (modifier, subject, …)
- syntactic trees
- tree-banks

Dependency tree



Phase-structure tree