Week 2 – Preprocessing

Lyubomira Dimitrova

January 2018

1 Tool

I used TreeTagger (available at the institute), with the Bulgarian parameter files. My preprocessing steps are POS-tagging[1] and lemmatization.

2 Statistics

2.1 POS

The corpus consists of 552 different tags (every verb, for example, receives a tag denoting its tense, aspect, transitivity, finiteness, voice, form, person and number, e.g. Vpptf-o3s), but 11 groups could be formed by taking only the first letter of the tag: **N**ouns, **V**erbs, **P**ronouns, **A**djectives, A**D**verbs, Par**T**icles, **I**nterjections, Nu**M**erals, P**R**epositions, **C**onjunctions and the **H**ybrid tag for familial names and adjectives.

Using the statistics from the corpus analysis exercise, it is interesting to note that there are more tokens in the corpus than POS-tags (1923379 and 1893886, respectively), since the TreeTagger does its own tokenization, and thus correctly recognises some multiple-word conjunctions and adverbs like 'за да', 'като че ли'.

Group	Count
P	547054
V	380439
N	351545
R	176573
A	132303
C	111194
T	102583
D	75019
M	15230
I	1032
H	914

[1] `https://www.sketchengine.co.uk/bulgarian-treebank-part-of-speech-tagset/`

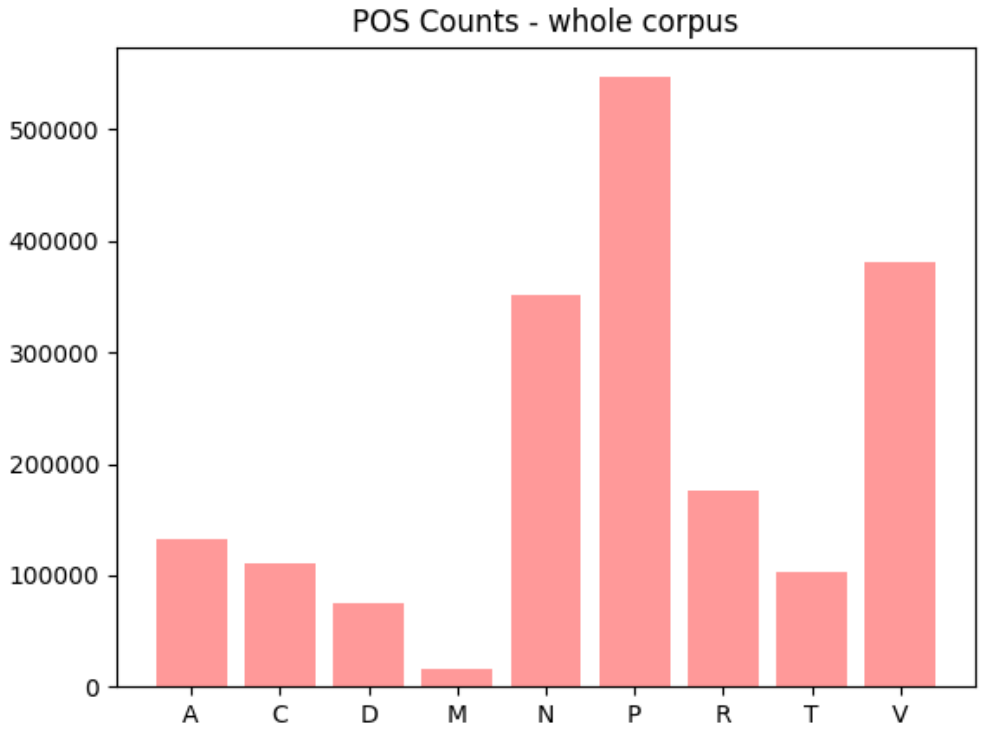
2.2 Lemmata

There are only 28482 distinct lemmata in the corpus. Some prominent ones are the '<unknown>'-tag, which covered 9% of all tokens, and the punctuation tags, the most frequent of which amounted to more than 14%. Short words (length 1 or 2 characters, excluding punctuation), which mostly include a mix of conjunctions, verbs, pronouns and particles, cover another 26% of all tokens (1893886).

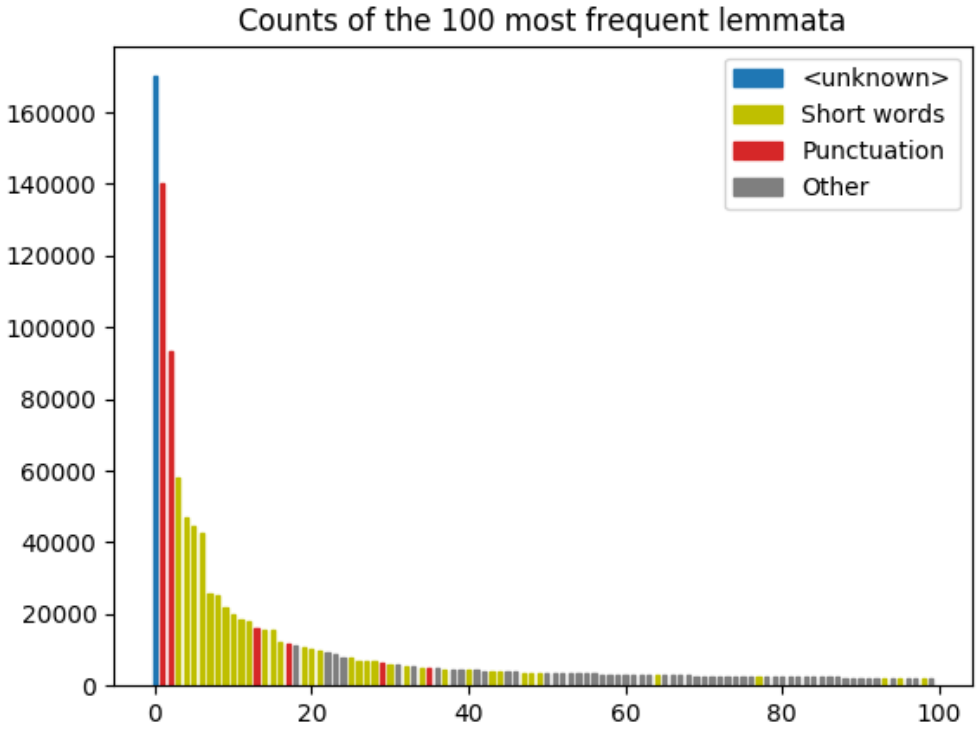
Distinct lemmata	# <unknown>	# Punctuation (.,!?;:)	Short words
28482	170067 (1 lemma)	272615 (6 lemmata)	493270 (107 lemmata)

It should be noted that 4% (114/28482) of all distinct lemmata account for almost 50% of all tokens.

3 Histograms



(The bars for **H** and **I** were not visible, so they were removed from the diagram.)


Counts of the 100 most frequent lemmata
<unknown>
Short words
Punctuation
Other
160000
140000
120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100