Estonian verb forms

Ahto Kiil
LTAT.02.002

Goal

- to analyse the Estonian language corpora's verb forms' distribution
- to prove or reject the hypothesis that changes in verb distribution are statistically significant

Long-term goal: to study verb distributions inside a lemma in different corpora and detect anomalies – possible change of word classes (recategorization)

Corpora

- the morphologically disambiguated corpus is manually tokenized (524 335 words, 98 512 verbs, weight = 0.1841)[1]
- the balanced corpus of fiction, journalistic and science texts (9 377 947 words, 1 864 620 verbs, weight = 0.1984) [2]
- the fiction corpus (original Estonian texts + translations, 16 444 403 words, 3 897 201 verbs, weight = 0.2369) [2]
- the Estonian Wikipedia corpus 2021 (8 618 382 words, 1 371 876 verbs, weight = 0.1593) [2]

Corpus sample (verb tokens in blue)

<s>						
Mustamäe	H.sg.g prop_sg_gen nmod	Mustamäe-h	sg_g	Mustamäe	Mustamäe	0 1 2
ühiselamutel	ühiselamu ühis elamu	S ühis_elamu	pl_ad tel 2	rootühiselamutel 0	S.pl.ad com_pl_ad root	ühiselamu-s pl_ad
on	V.b mod_indic_pres_ps3_sg_ps_af 2	olema-v cop	b ühiselamutel	ole ühiselamu	ole fin S	0 Intr pl_ad 3 root
hooneregistri	S.sg.g com_sg_gen	hooneregister-s	sg_g	hoone register	hoone_register	0

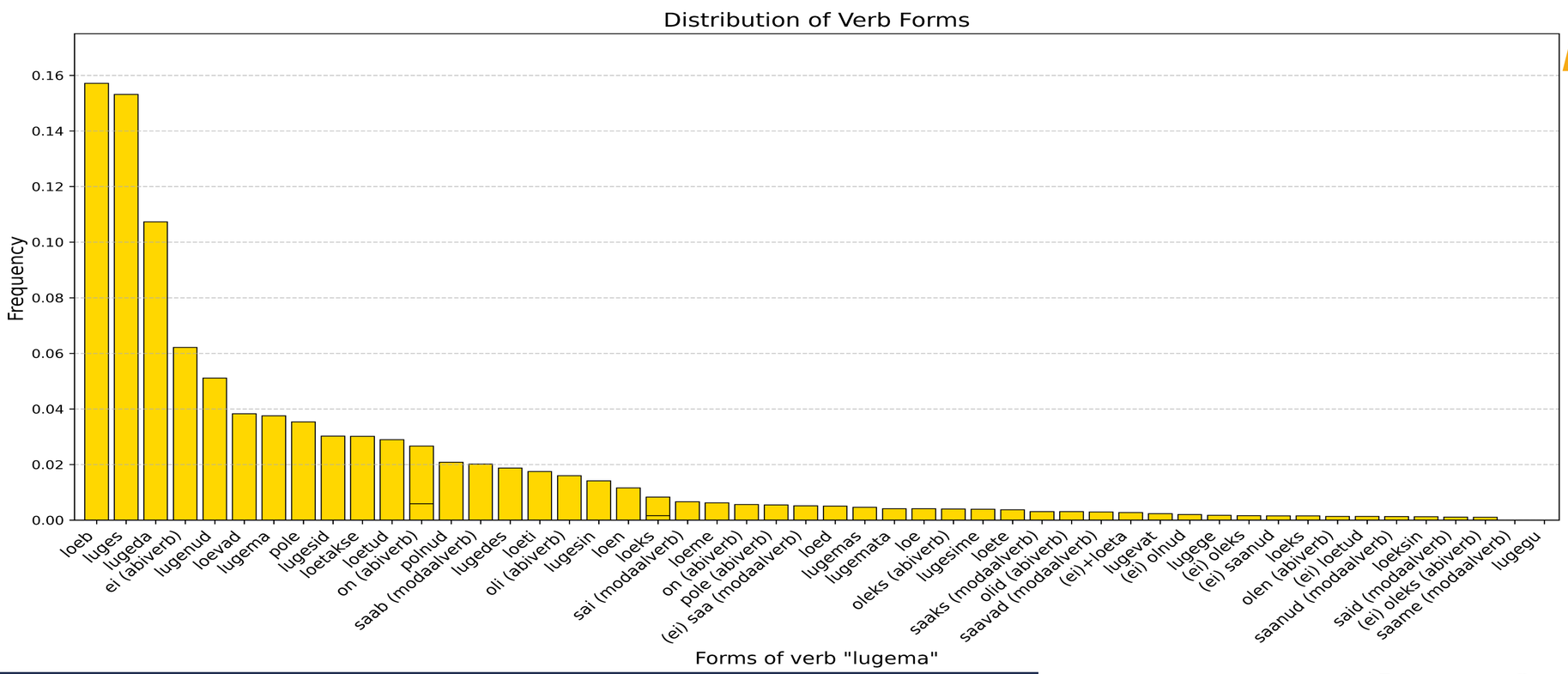
Process

- clean the corpus from metadata, formatting tags, interpunctuation, numeric values
- count all words
- separate all words tokenized as 'verbs'
- count all different tokens
- calculate weight of words in corpus
- calculate relative frequencies of verb forms (tokens)

Top 3 verb forms in Estonian

1. present tense third person singular active affirmative - ***loeb***
2. past tense third person singular active affirmative - ***luges***
3. infinitive - ***lugeda***

Distribution of verb forms in morph-corp



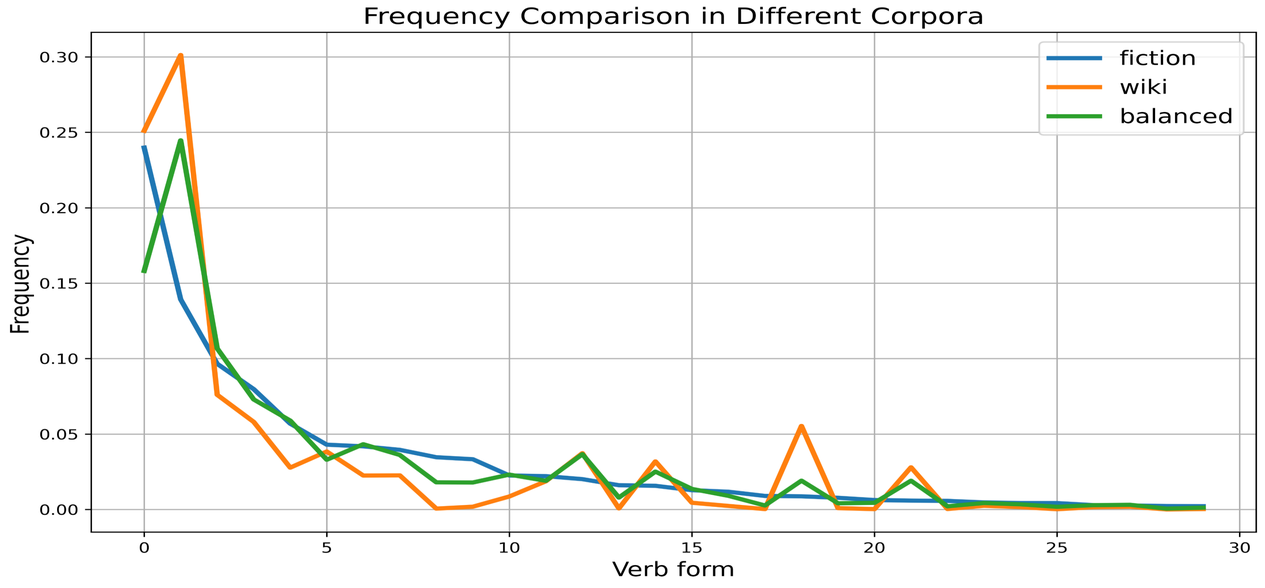
Statistical significance

We consider morph-corp as H0: weight of verbs = 0.184066
Hypothesis testing with H1, H2, H3 (balanced, fiction, wiki corpora with weight = [0.1984, 0.2369, 0.1593])

```
For alpha = 0.05:
Comparison of H1 against H0:
Z-score: 110.13728634705134
P-value: 0.00e+00
Reject the null hypothesis (H0)
Comparison of H2 against H0:
Z-score: 504.72493003529473
P-value: 0.00e+00
Reject the null hypothesis (H0)
Comparison of H3 against H0:
Z-score: -199.69160272155264
P-value: 0.00e+00
Reject the null hypothesis (H0)
```

Conclusions

- the analysed corpora's verb weights differ in a statistically significant level at alpha = 0.05 (also at higher values of α = 0.1, 0.2, ..., 0.95, ...)
- a standard weight and distribution of verb forms cannot be calculated (based on the four analysed corpora)
- distribution of verb forms follows closely the Zipf's law[3] (after the second most frequent verb)
- the contextual approach to verb distribution calculation is recommended (novels vs. fiction corpus)
- analyses of larger corpora should be performed



UNIVERSITY OF TARTU
1632

References

1. *https://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=en*

2. *https://entu.keeleressursid.ee/shared/9942/G3qonxL3kavZ1NGJ79jk7eIamFRkoLZBMkHCc8jVgudimdUDeSZC5XKrn6Uowxhj*

3. *https://en.wikipedia.org/wiki/Zipf%27s_law*

Ahto Kiil
ahto.kiil@ut.ee
ahtokiil.github.io

Scan the QR code to visit the GitHub repo