Sketch Engine pre-processing pipelines: towards on-the-fly tokenization of user queries

Matúš Kostka, Marek Medveď
xkostka4@fi.muni.cz, marek.medved@sketchengine.eu

Lexical Computing

June 30, 2023

Pipeline

SkE pipeline

Input

‘eLex’ conferences aim to explore innovative developments in the field of lexicography.

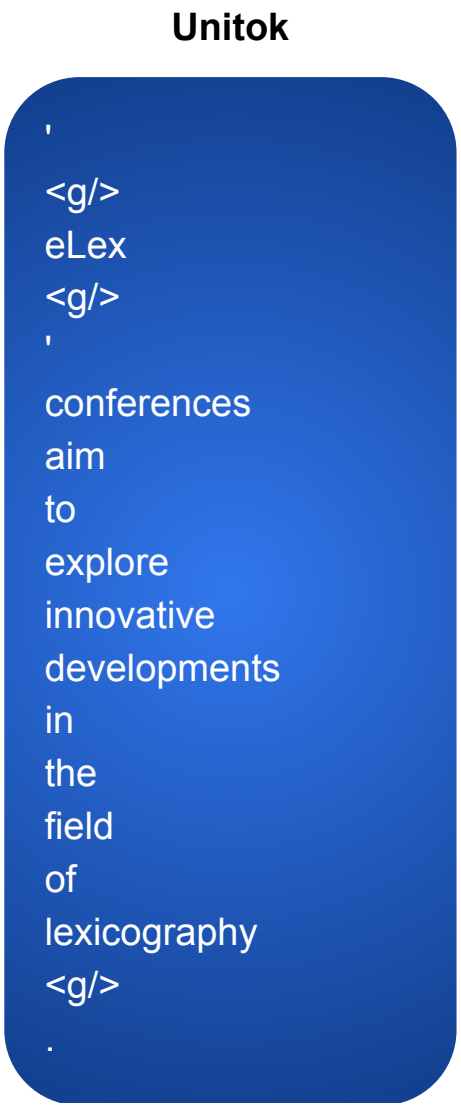
Input sentence containing single quotes ‘ and ’.

Uninorm

'eLex' conferences aim to explore innovative developments in the field of lexicography.

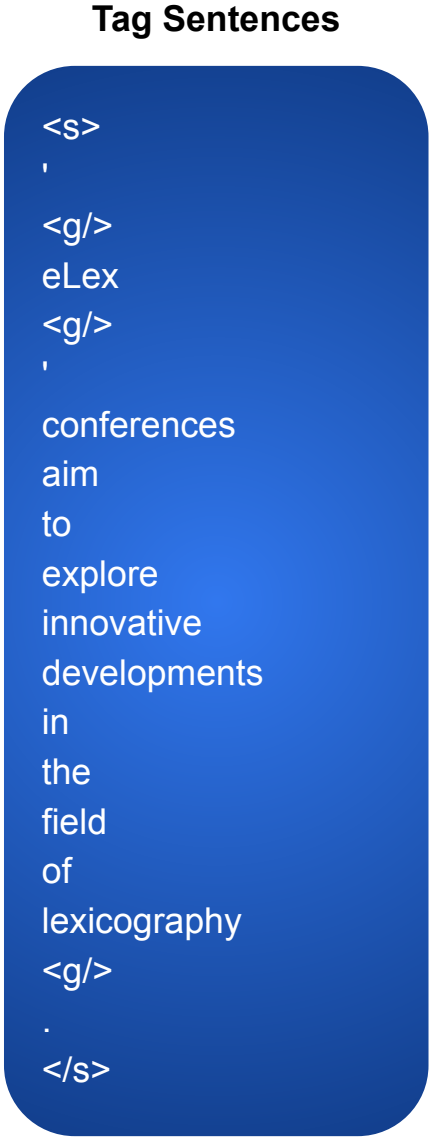
Uninorm provides normalization of text and converts the content into NFKC normalization form. The ‘ and ’ becomes ''. Unifies search and improves stability.

SkE pipeline



Unitok splits the input text into units suitable for further computational processing. It is an important data preparation step allowing us to perform more advanced tasks.

SkE pipeline



Tag Sentences identifies the boundaries of individual sentences inside the text.

SkE pipeline

Lemmatizer assigns a lemma (base form of the word) to each word form in a corpus.

SkE pipeline

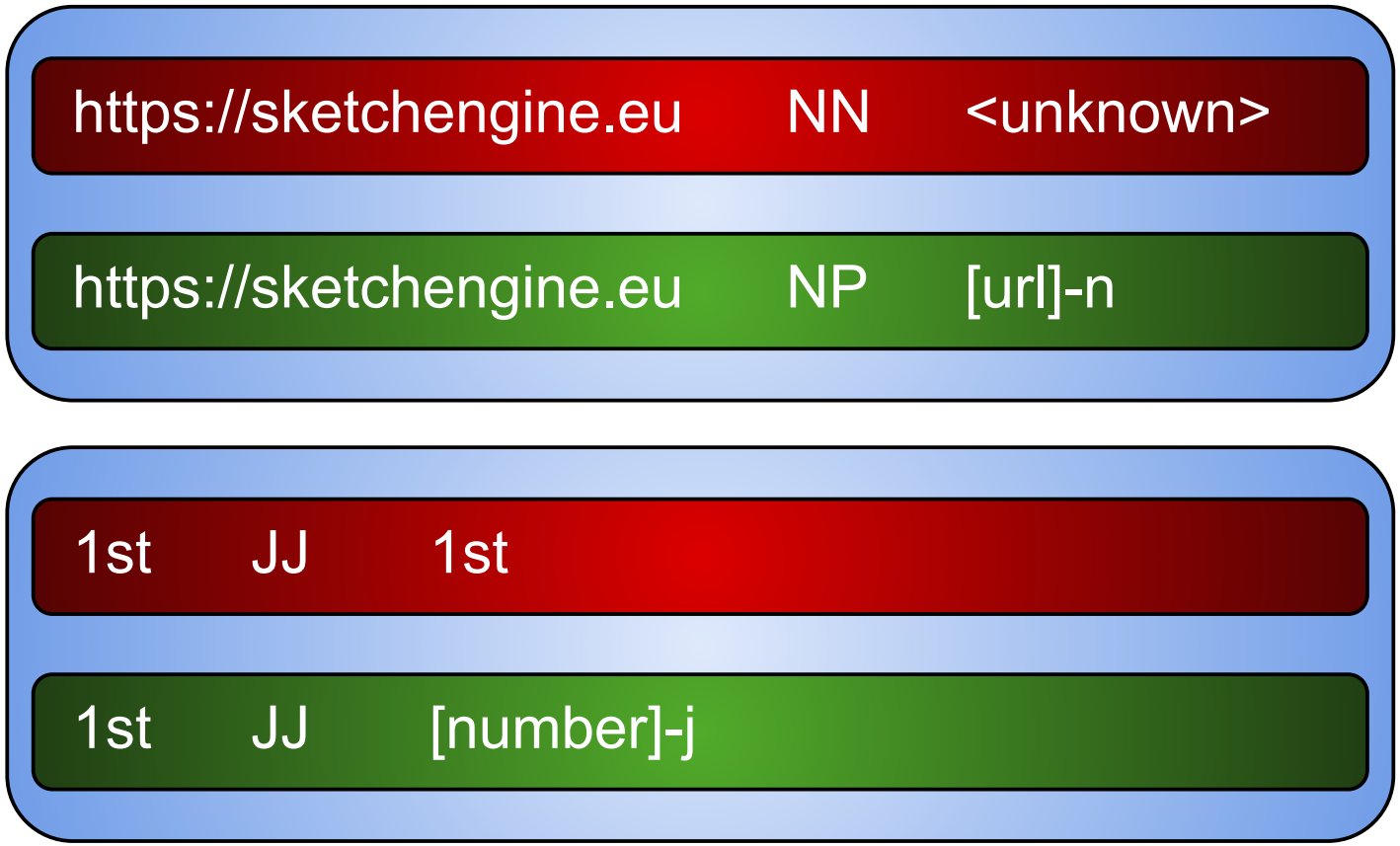
Tag

```
<s>
'       POS     '
<g/>
eLex    NP      Elex
<g/>
'       POS     '
conferences             NNS     conference
aim     VVP     aim
to      TO      to
explore         VV      explore
innovative      JJ      innovative
developments            NNS     development
in      IN      in
the     DT      the
field   NN      field
of      IN      of
lexicography            NN      lexicography
<g/>
.       SENT            .
</s>
```

Tagger assigns special labels to each token in the corpus to indicate part of speech and other grammatical categories.

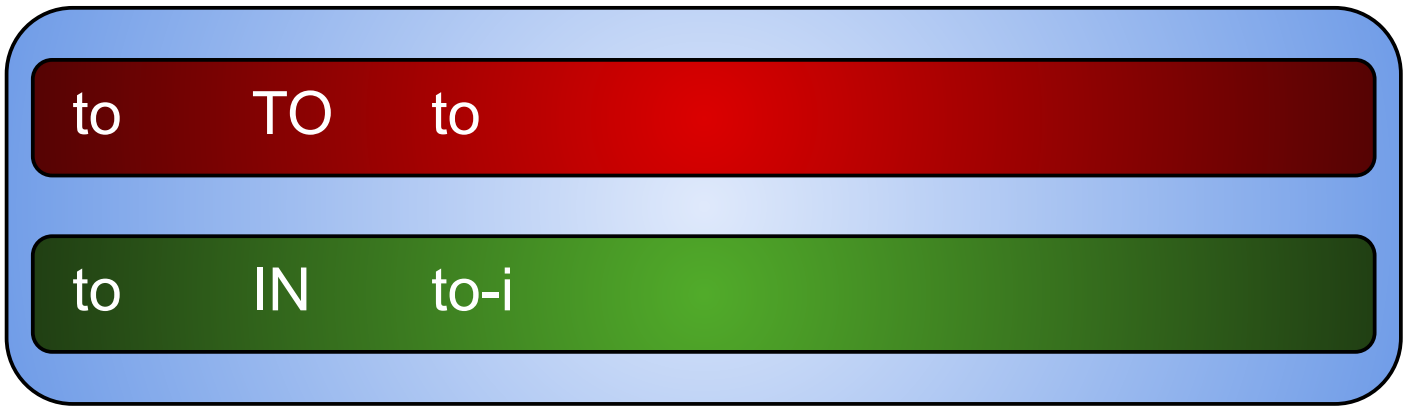
SkE pipeline

- **Post-processing**: simple tag and lemma modifications



SkE pipeline

- **Post-processing**: assign tag ’TO’ only if followed by an adverb or a verb, otherwise assign tag ’IN’

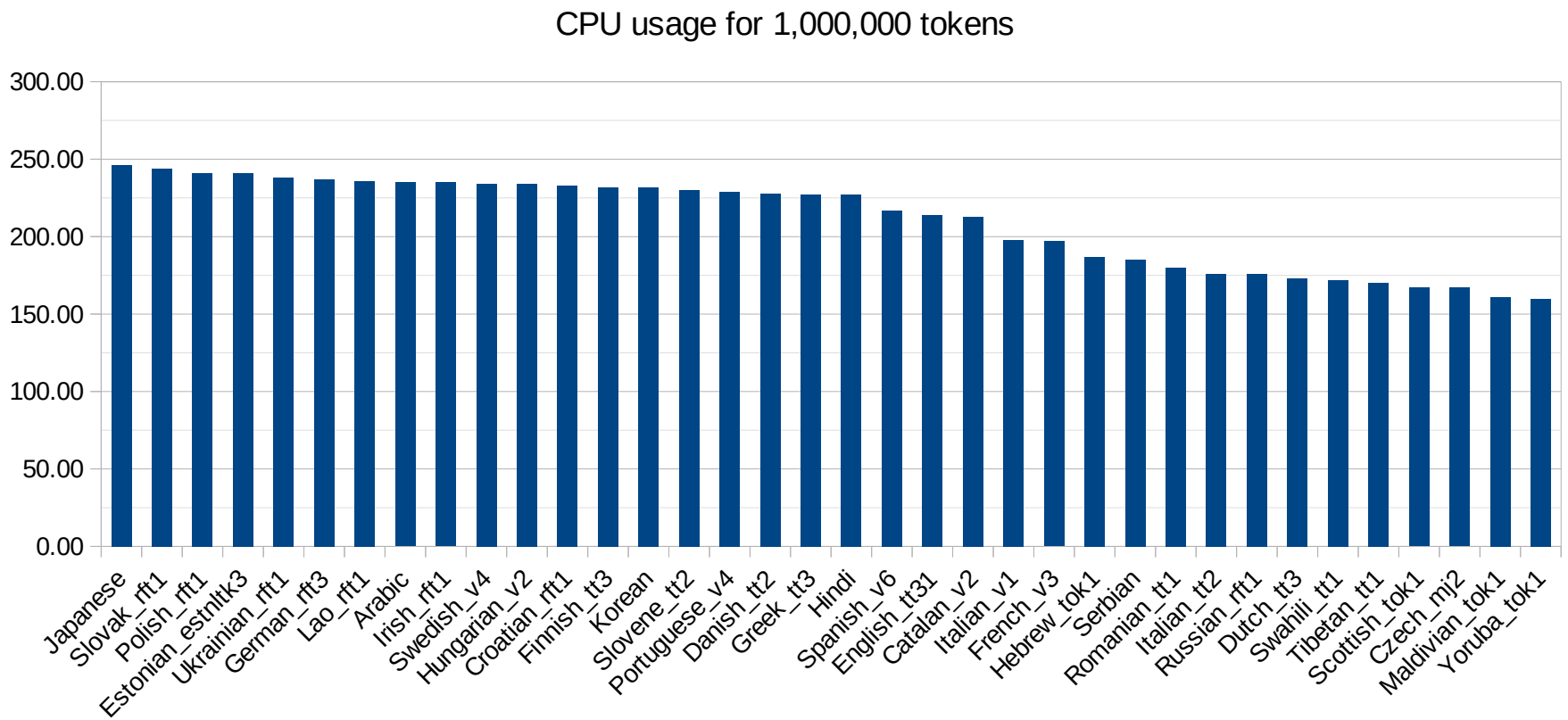


Results

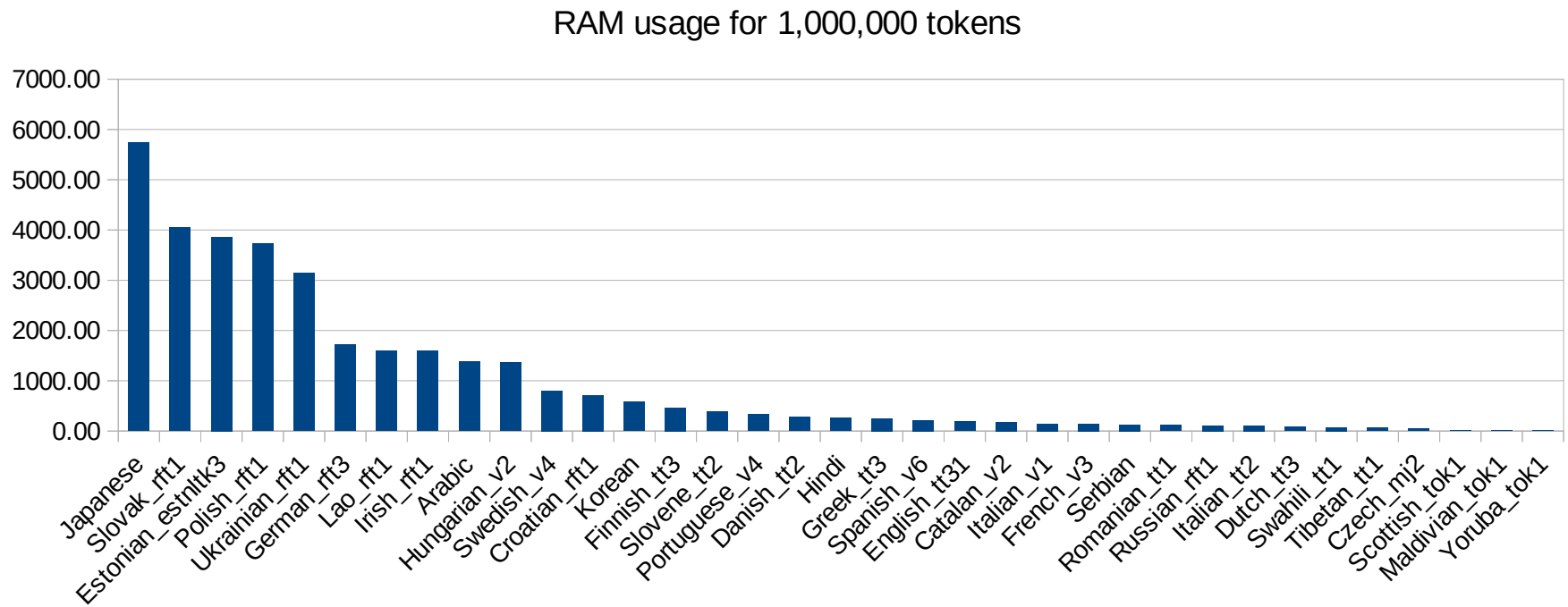
Hardware setup

- RAM: 252 GB
- CPU: 32 threads

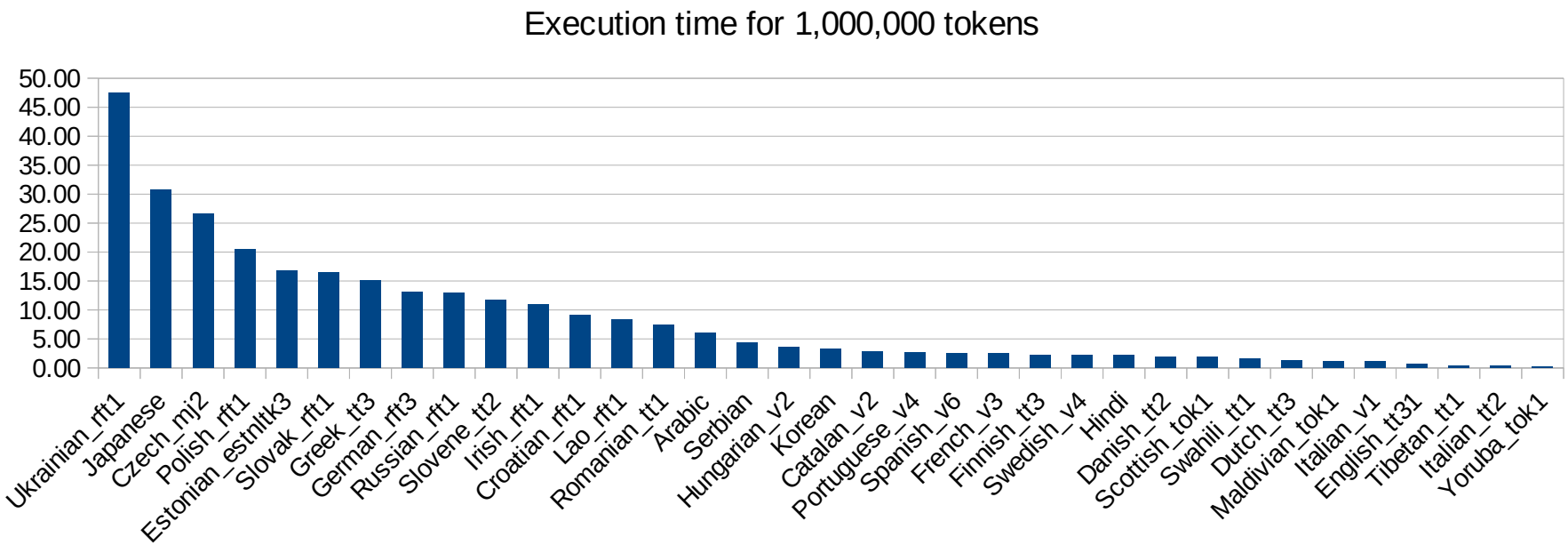
CPU usage


CPU usage for 1,000,000 tokens
300.00
250.00
200.00
150.00
100.00
50.00
0.00
Japanese
Slovak_rft1
Polish_rft1
Estonian_estnltk3
Ukrainian_rft1
German_rft3
Lao_rft1
Arabic
Irish_rft1
Swedish_v4
Hungarian_v2
Croatian_rft1
Finnish_tt3
Korean
Slovene_tt2
Portuguese_v4
Danish_tt2
Greek_tt3
Hindi
Spanish_v6
English_tt31
Catalan_v2
Italian_v1
French_v3
Hebrew_tok1
Serbian
Romanian_tt1
Italian_tt2
Russian_rft1
Dutch_tt3
Swahili_tt1
Tibetan_tt1
Scottish_tok1
Czech_mj2
Maldivian_tok1
Yoruba_tok1

RAM usage


RAM usage for 1,000,000 tokens
7000.00
6000.00
5000.00
4000.00
3000.00
2000.00
1000.00
0.00
Japanese
Slovak_rft1
Estonian_estnltk3
Polish_rft1
Ukrainian_rft1
German_rft3
Lao_rft1
Irish_rft1
Arabic
Hungarian_v2
Swedish_v4
Croatian_rft1
Korean
Finnish_tt3
Slovene_tt2
Portuguese_v4
Danish_tt2
Hindi
Greek_tt3
Spanish_v6
English_tt31
Catalan_v2
Italian_v1
French_v3
Serbian
Romanian_tt1
Russian_rft1
Italian_tt2
Dutch_tt3
Swahili_tt1
Tibetan_tt1
Czech_mj2
Scottish_tok1
Maldivian_tok1
Yoruba_tok1

Execution time

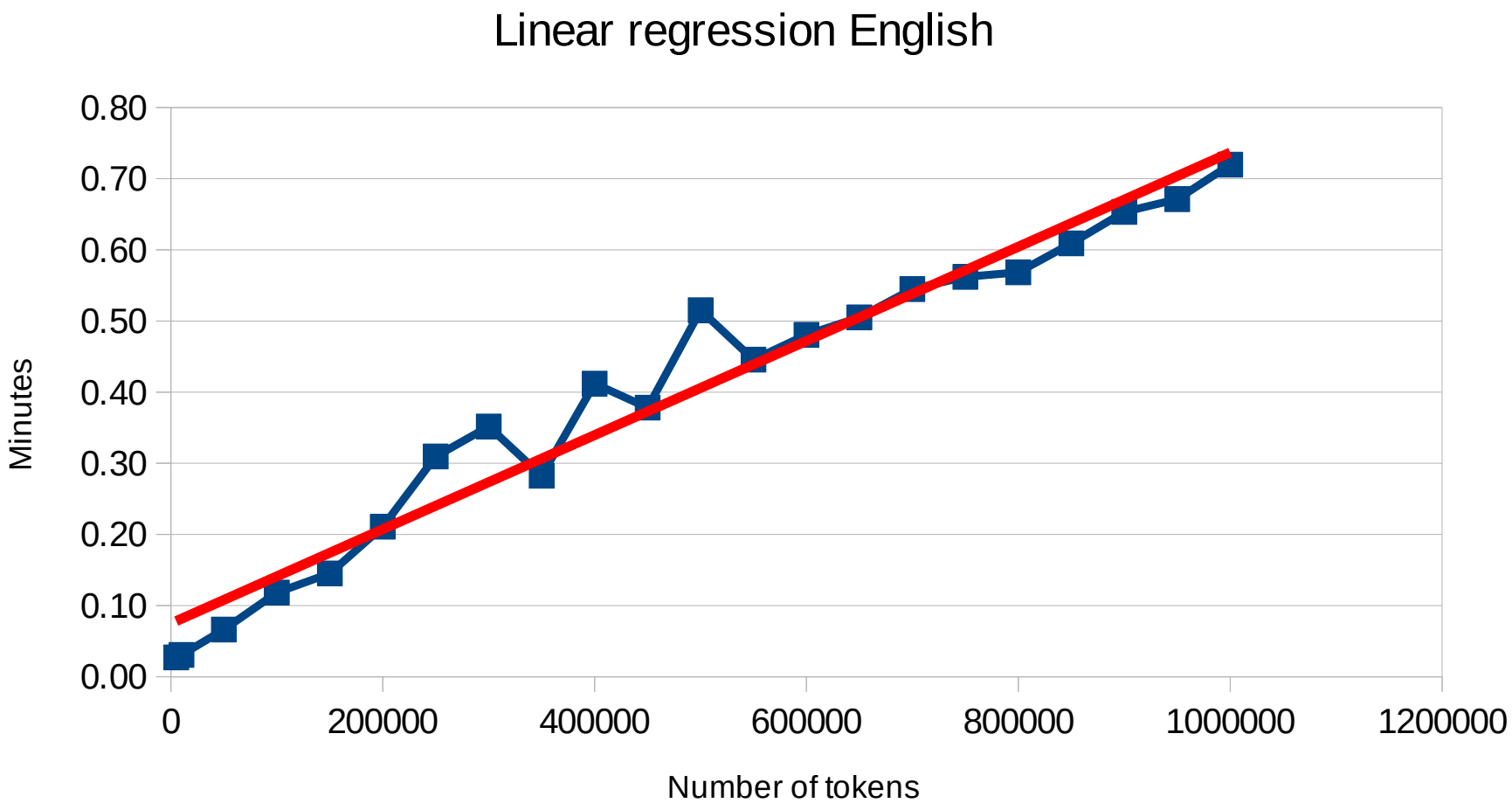


Statistics

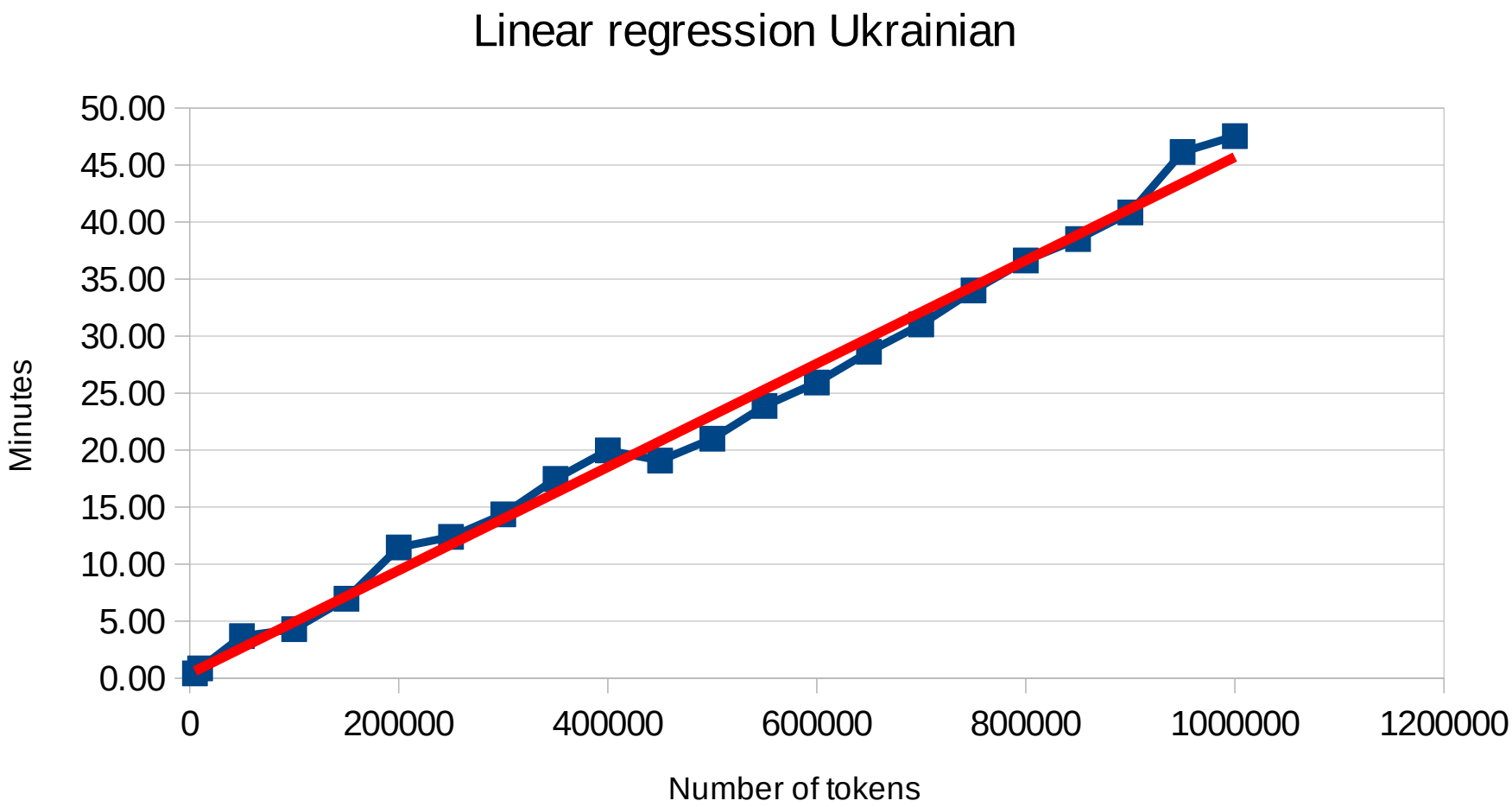
Table: SM: Overall stats 1,000,000 tokens

	Min value	Max value	Average	Median
Execution time (sec)	11.77	2851.95	475.74	171.24
CPU usage (%)	57	224	111	105
RAM usage (GB)	0.007	5.600	0.866	0.243

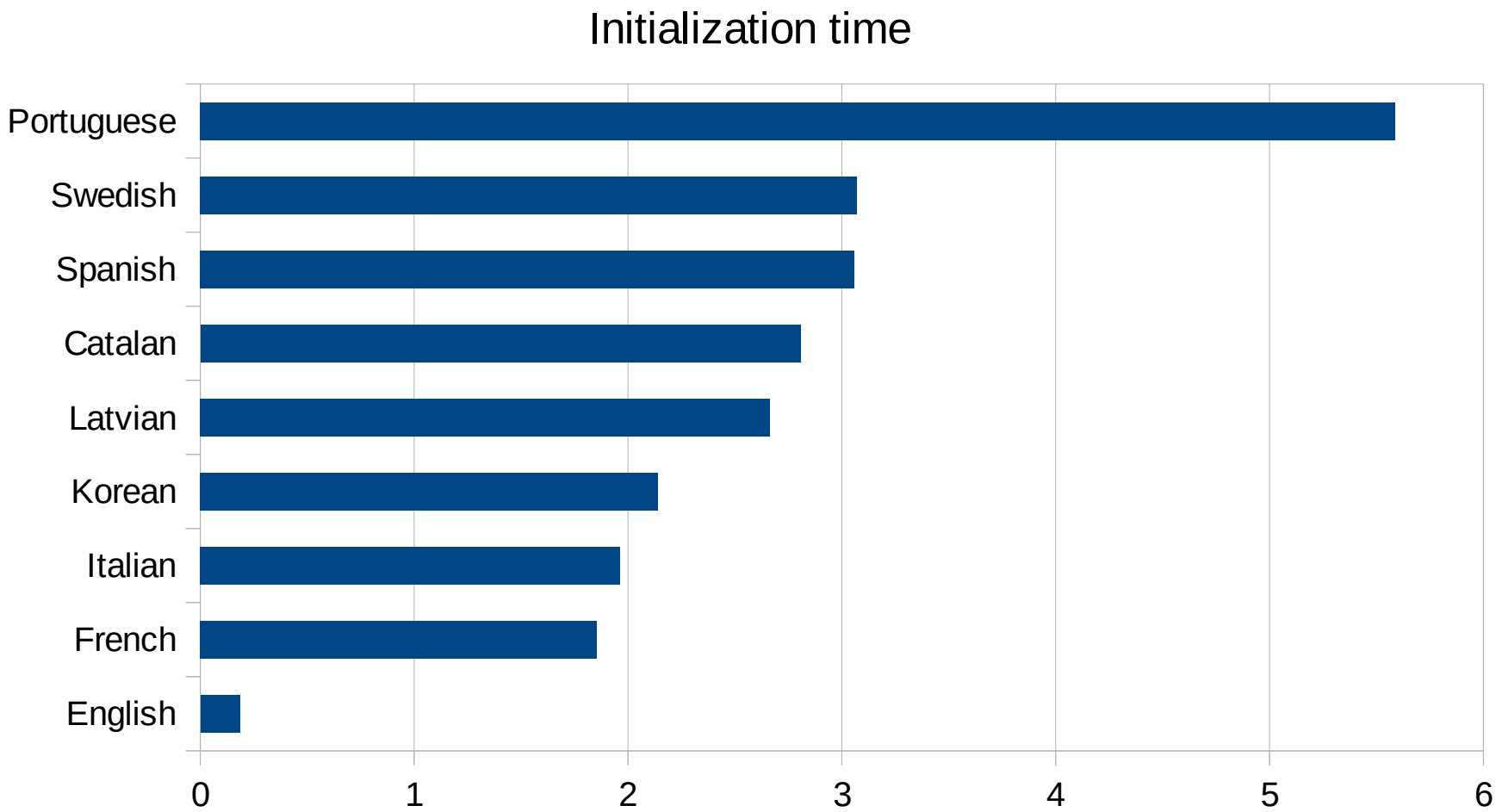
English


Linear regression English
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Minutes
0
200000
400000
600000
800000
1000000
1200000
Number of tokens

Ukrainian


Linear regression Ukrainian
Minutes
50.00
45.00
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
0
200000
400000
600000
800000
1000000
1200000
Number of tokens

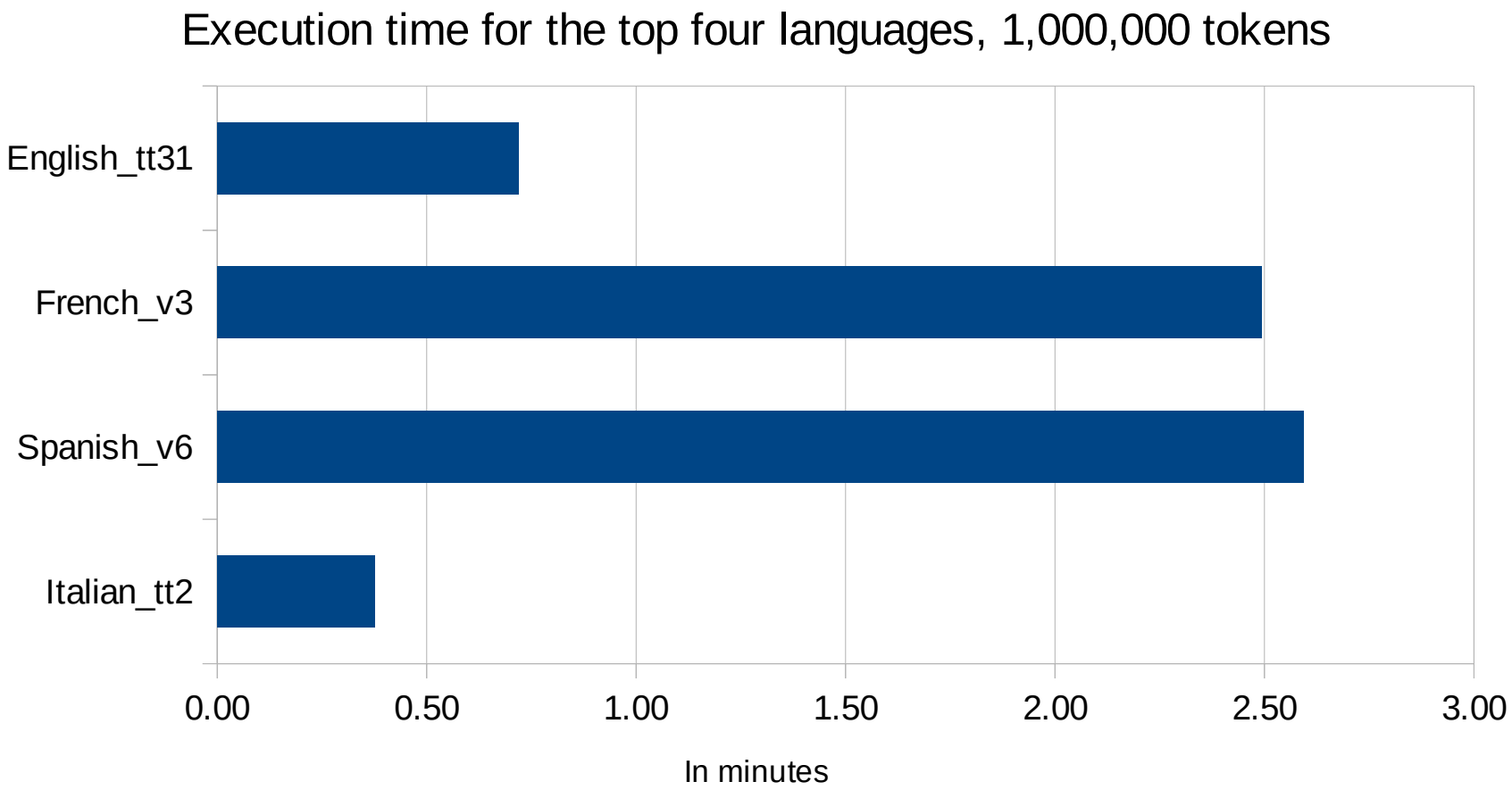
Initialization time



Conclusions

- the less performing pipeline inside SkE is the **Tagalog**
- RAM and CPU heavy usage pipelines are **Thai** (thai_sw1) and **Hebrew** (yap_he_v1)
- languages with **different alphabets** as Latin are usually slower
- **12%** of pipelines can be used in the multi-threaded setup
- **positive fact** is that all pipelines are in a linear relationship with the number of tokens
- **state-of-the-art** pipeline performance for the most used languages inside SkE

Top 4 used languages in SkE



Future work

- improve the time complexity
- implement parallel processing for more pipelines
- improve pipelines stability


SKETCH
ENGiNE