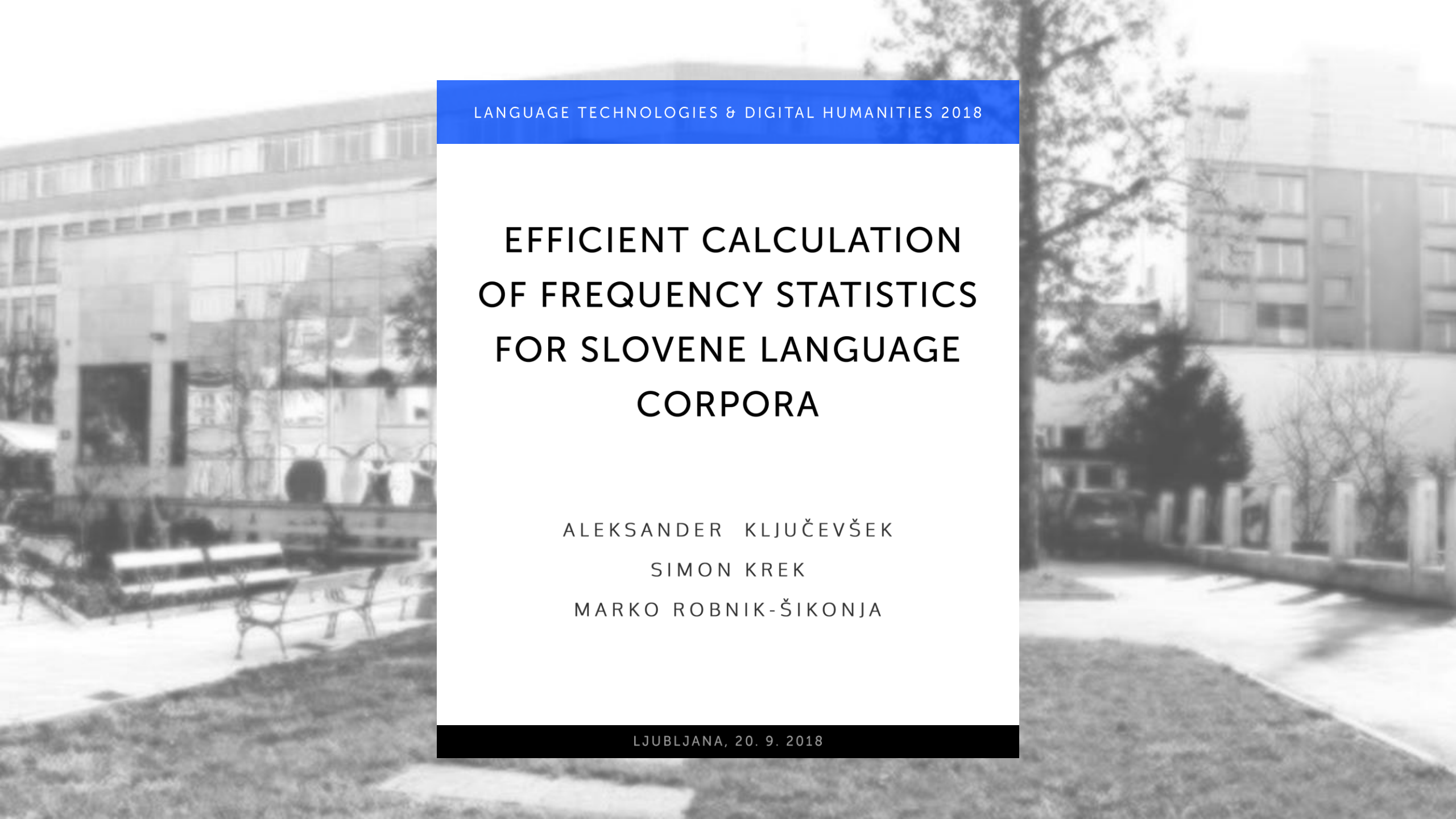


LANGUAGE TECHNOLOGIES & DIGITAL HUMANITIES 2018

EFFICIENT CALCULATION OF FREQUENCY STATISTICS FOR SLOVENE LANGUAGE CORPORA

ALEKSANDER KLJUČEVŠEK

SIMON KREK

MARKO ROBNIK-ŠIKONJA

LJUBLJANA, 20. 9. 2018

MOTIVATION

Language changes alongside us.

Statistical analysis gives us a peek into the fundamentals of language.

GOALS

Application for efficient calculation of frequency statistics for Slovene language corpora

- Efficiently analyse large text corpora
- Run on a single computer
- Utilize all resources at its disposal

CORPORA

GIGAFIDA

Reference corpus. Contains 1.187.002.502 words.

KRES

Balanced corpus. Sampled from the Gigafida corpus. Contains 99.831.145 words.

GOS

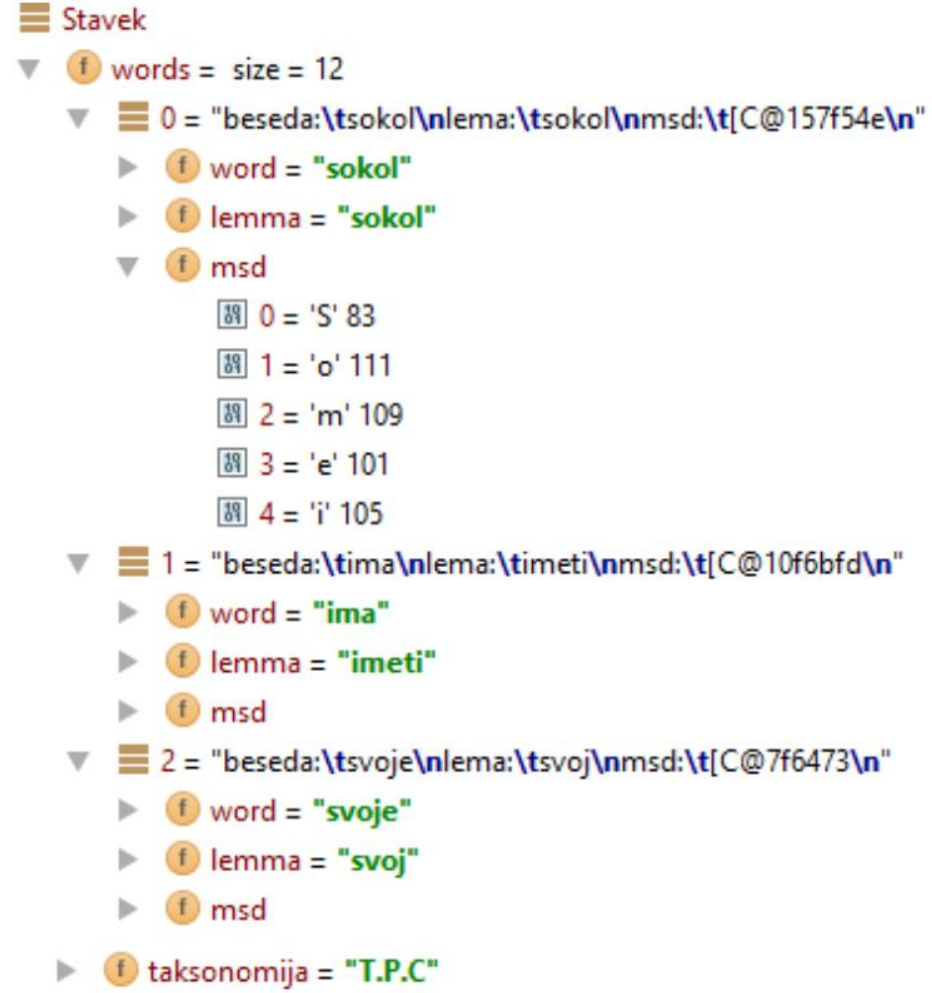
Corpus of spoken Slovene. Includes transcripts of approximately 120 hours of speech.

ŠOLAR

Corpus of school essays. A collection of texts written by pupils and students in Slovene primary and secondary schools.

DATA REPRESENTATION

```
<s>
   <w msd="Somei" lemma="sokol">Sokol</w>
   <S/>
   <w msd="Ggnste-n" lemma="imeti">ima</w>
   <S/>
   <w msd="Zp-set" lemma="svoj">svoje</w>
   <S/>
   <w msd="Soset" lemma="območje">območje</w>
   <S/>
   <w msd="Rsn" lemma="točno">točno</w>
   <S/>
   <w msd="Rsn" lemma="označeno">označeno</w>
   <S/>
   <w msd="Vp" lemma="in">in</w>
   <S/>
   <w msd="Rsn" lemma="lahko">lahko</w>
   <S/>
   <w msd="Ggnste" lemma="zajemati">zajema</w>
   <S/>
   <w msd="Rsr" lemma="več">več</w>
   <S/>
   <w msd="Kbg-mt" lemma="tisoč">tisoč</w>
   <S/>
   <w msd="Sommr" lemma="hektar">hektarjev</w>
   <c>.</c>
   <S/>
</s>
```



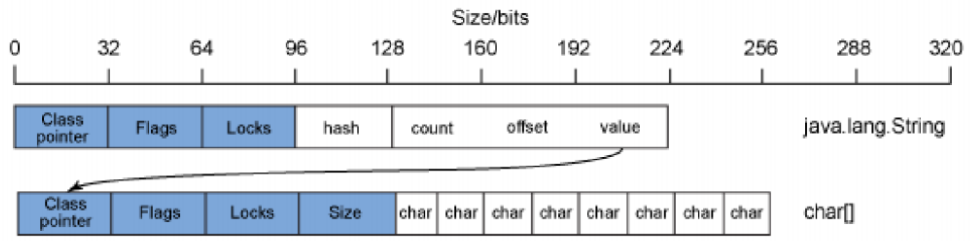
SIZE OF CORPORA

GIGAFIDA - 83,5 GB

Keeping the whole corpus in memory is not an option

OVERHEAD

1 GB of memory can hold at most 7.3 million words – 0.62% of all words contained in the corpus



SIZE OF CORPORA - SOLUTION

Algorithm Batch processing

```
1: while corpus contains unread sentences do
2:     subcorpus ← sentence
3:     if subcorpus.size ≥ limit then
4:         FORK-JOIN(subcorpus)
5:         subcorpus = ∅
6:     end if
7: end while
```

PARALLELIZATION

To calculate word frequencies:

- For each word first check if we already store it in our database
- If we do not we add it, otherwise we increment its count by 1

1,2 billion words = 1,2 billion such operations

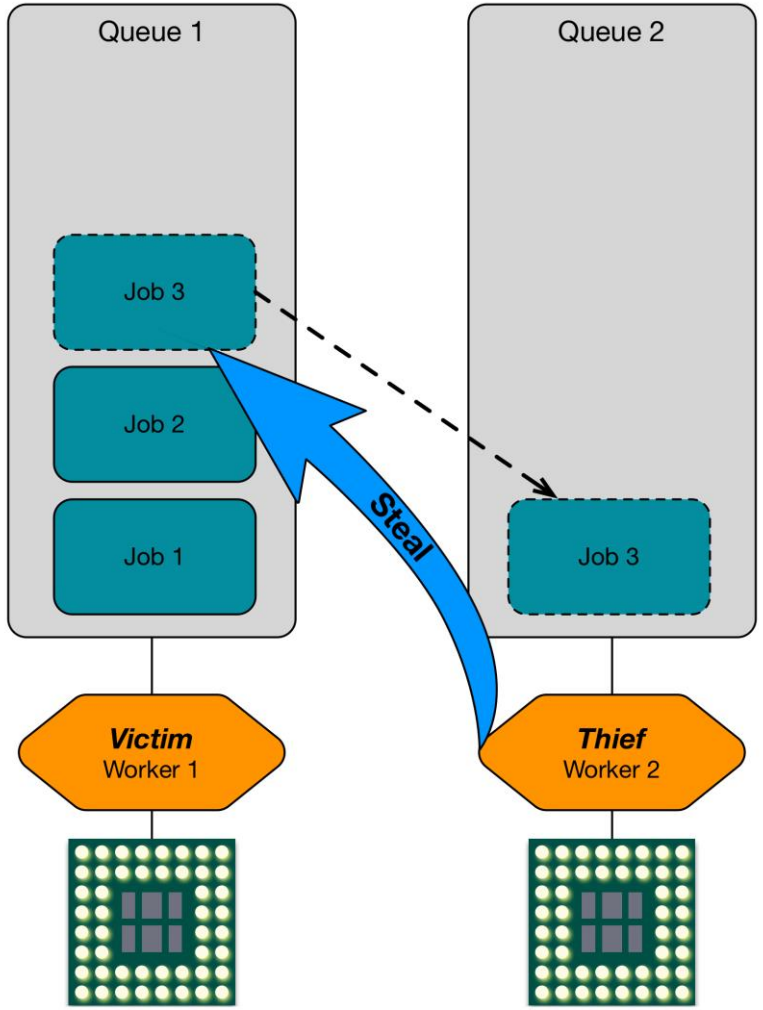
Modern CPUs consist of multiple cores and support multithreading

PARALLELIZATION - SOLUTION

Algorithm Fork-Join

```
1:  procedure SOLVE(problem)
2:      if problem is small enough then
3:          solve sequentially
4:      else
5:          split into subproblemA and subproblemB
6:          fork SOLVE(subproblemA)
7:          fork SOLVE(subproblemB)
8:          join solutions of subproblems A in B
9:      end if
10: end procedure
```

PARALLELIZATION - BONUS



CALCULATING STATISTICS

LetterCount

Ngrams

WordCount

WordLengthCount

CALCULATING STATISTICS

FREQUENCIES

Words, lemmas or letters, usually with added conditions such as taxonomy or morphosyntactic descriptions

QUERIES CAN BE SIMPLE

Frequencies of all words

OR MORE COMPLEX

Frequencies of bigrams corresponding to bigram „S*z** Gp-s***“ in news (N*f** Va-r***)

LETTER COUNT

	KRES		Gigafida	
	letter	%	letter	%
1.	a	10.12	a	10.01
2.	e	9.99	e	9.74
3.	o	9.07	o	9.03
4.	i	8.78	i	8.73
5.	n	6.74	n	6.69
6.	r	5.17	r	5.26
7.	s	4.57	t	4.47
8.	t	4.48	s	4.45
9.	l	4.46	l	4.36
10.	j	4.17	v	4.11

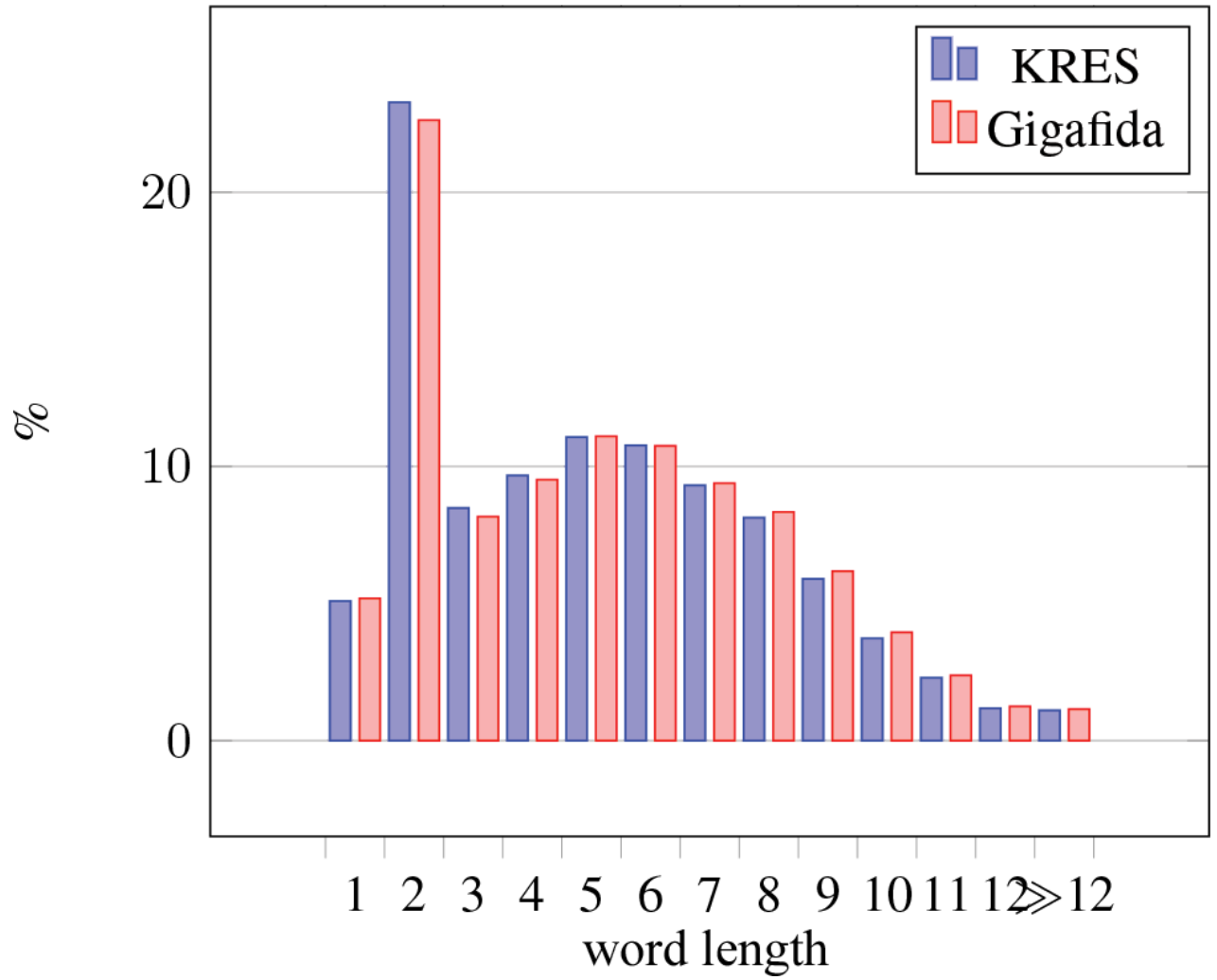
WORD COUNT

	KRES		Gigafida	
	lemma	%	lemma	%
1.	biti	7.60	biti	7.34
2.	in	2.84	v	2.63
3.	v	2.47	in	2.56
4.	se	1.87	se	1.59
5.	na	1.51	na	1.58
6.	z	1.39	z	1.33
7.	da	1.28	za	1.31
8.	on	1.24	da	1.23
9.	za	1.19	ki	1.02
10.	ta	1.06	ta	1.01

BIGRAMS
MORPHOSYNTACTIC DESCRIPTIONS

	KRES		Gigafida	
	bigram	%	bigram	%
1.	Slmei- Slmei-	0.75	Slmei- Slmei-	1.09
2.	Dm Sozem-	0.74	Dm Sozem-	0.74
3.	Vd Gp-ste-n	0.64	Vd Gp-ste-n	0.64
4.	Dm Somem-	0.62	Ppnzer- Sozer-	0.64
5.	Ppnzei- Sozei-	0.61	Dm Somem-	0.63
6.	Ppnzer- Sozer-	0.59	Ppnzei- Sozei-	0.63
7.	Rsn Rsn	0.55	Kag—- Kag—-	0.58
8.	Dt Sozet-	0.52	Ppnmeid Somei-	0.54
9.	L Rsn	0.50	L Rsn	0.51
10.	Kag—- Kag—-	0.50	Dt Sozet-	0.50

WORD LENGTH



SUMMARY

AN APPLICATION FOR EFFICIENT CALCULATION OF FREQUENCY STATISTICS

Optimized for multi-core and hyper-threaded CPUs. Batch processing of data. Computation time for a single statistic on a billion word corpus takes approximately 90 – 100 minutes.

ANALYSIS OF SEVERAL SLOVENE CORPORA ON MULTIPLE LEVELS

Strings, lemmas, parts of words and word-formation patterns.

FUTURE WORK

Auto-detection of corpus structure. Additional statistics. Internationalization and localization.