

N-gram Frequency Lists for Reference Corpora of Slovenian

Kaja Dobrovoljc

Jožef Stefan Institute

Centre for Language Resources and Tools UL



Univerza v Ljubljani
Fakulteta za računalništvo
in informatiko



Institut
"Jožef Stefan"
Ljubljana, Slovenija

Motivation

- **frequency lists** a fundamental tool in corpus linguistics
 - investigations into general/specialized vocabulary, glossaries, language modelling etc.
- shift from words to **multi-word expressions**
 - most work on collocations (association-based approaches)
- increase of research on **formulaic language**
 - lexical bundles, strings ... n-grams (frequency-based approaches)
- practical **limitations**
 - software/hardware performance, data restrictions

Corpora

- **GOS** reference corpus of **spoken Slovenian**
 - spontaneous speech in everyday situations (1M tokens)
- **IMP** reference corpus of **historical Slovenian**
 - digitized Slovenian texts from the period 1584-1919 (17M tokens)
- **KRES** reference corpus of modern **written Slovenian**
 - books and periodicals from 1990-2011 (120M tokens)
- **JANES** reference corpus of **user-generated Slovenian**
 - tweets, blogs, forums, news comments and wiki user pages (253M tokens)

N-gram extraction

- set of python scripts
- parameters:
 - token type (e.g. lemma)
 - the size of n (e.g. 3 words)
 - ignoring punctuation (e.g. *tako da* vs. *tako, da*)
- 3 types of frequency lists:
 - **regular** (all n-grams sorted by frequency)
 - **filtered** (n-grams above a given corpus/text frequency threshold)
 - **adjusted**

Adjusted frequency list

- difficult to compare n-grams of different lengths
 - substrings always more frequent than parent strings
 - e.g. *glede na* ($f = 309$) > *glede na to* ($f = 178$)
 - although realistically ***glede na to* (178) > *glede na* (131)**
- statistical reduction of substrings
 - O'Donnell 2011: selective reduction of counts in a **pre-indexed corpus**
 - n-gram counted only if not part of a longer relevant n-gram
 - relevancy defined by min. freq. of occurrence
 - a **joint list of all n-grams** with a more telling list of types, tokens and rankings

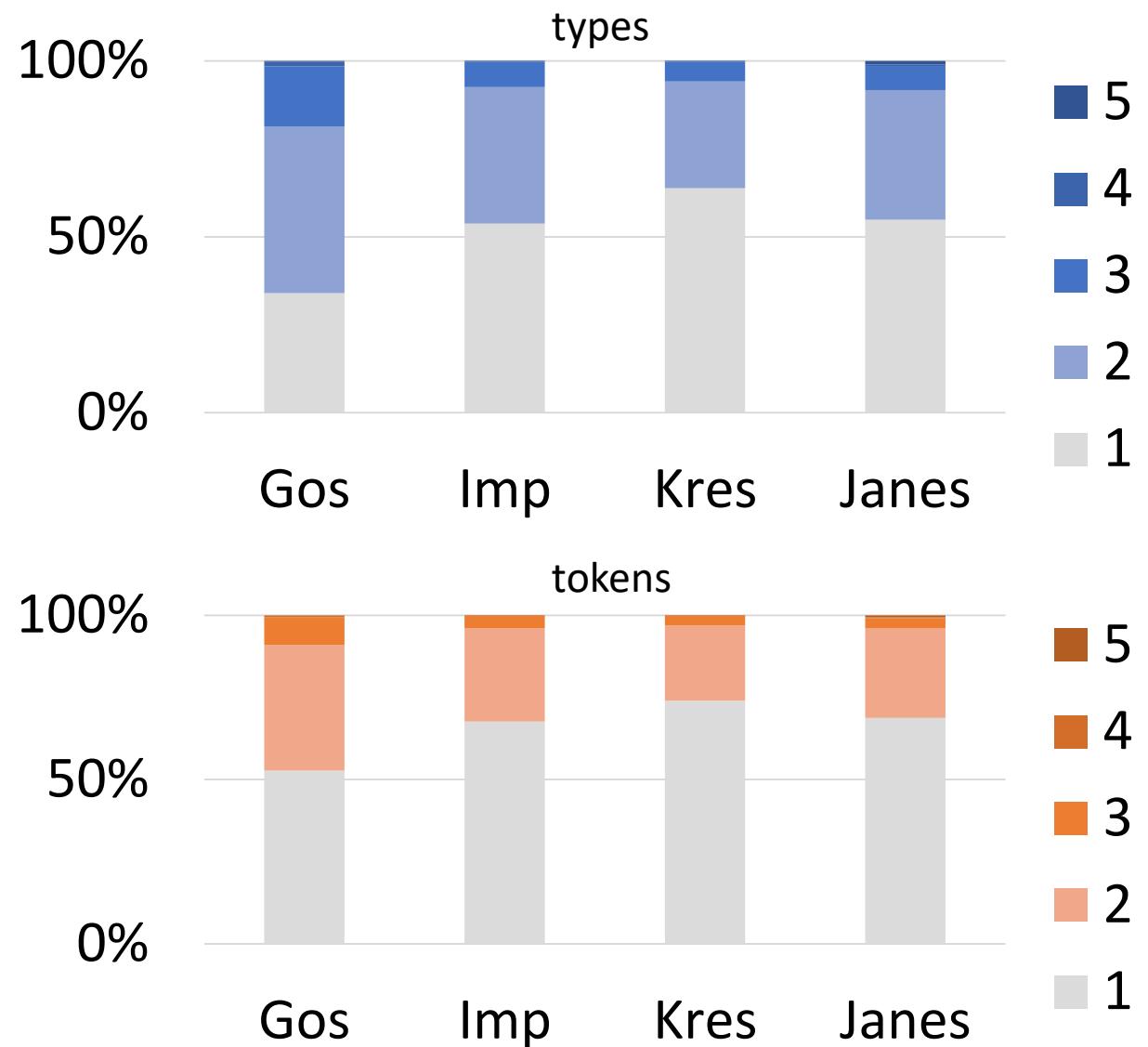
Results

- an adjusted frequency list of **normalized** n-grams of 1-5 tokens with a freq. $\geq 10/\text{mio.}$, occurring in at least 2 texts ("core vocabulary")

No. of words	GOS	IMP	KRES	JANES
1	6.371	8.460	10.270	8.914
2	8.860	6.087	4.885	5.984
3	3.199	1.131	901	1.110
4	244	43	32	68
5	47	6	11	171
SUM	18.721	15.727	16.099	16.247

Structure and frequency

- **large no. of multi-word units > formulaicity of human communication**
- **esp. in spontaneous speech**, in which multi-word strings prevail
- not just bigrams!



Overlap of core vocabulary

	% of overlapping n-grams				% of unique
	in GOS	in IMP	in KRES	in JANES	
GOS		34,0	42,5	48,2	43,8
IMP	40,5		49,1	43,4	41,8
KRES	49,5	48,0		60,8	24,9
JANES	55,6	42,0	60,1		25,6

- a large number of unique n-grams in all corpora
 - some differences in annotation principles (e.g. [@per] [URL])
 - **distinct vocabulary**

filled pauses; expressions of agreement, vagueness; non-standard vocabulary

dated vocabulary and syntactic constructions

legislative jargon, periodical metatext

non-lexical tokens, English (meta)text, online inquiries

GOS

1 *eee, eem, tlele, nnn, tipo, čao, majčeno, tukajle, šestdeset, devetdeset*

2 *in eee, eee eee, mhm mhm, eee v, ne eee, pa eee, eee in, eee ja, eee ne, pa pol*

3 *ja ja ja, ne ne ne, ja ne vem, na neki način, ne to je, mhm mhm mhm, eee to je, ne tako da, eee ne vem, eee tako da*

IMP

je., zavoljo, ondi, baron, lice, urno, zmerom, rekoč, dasi, čebele

ako se, je zopet, ako bi, ako je, dejal je, n. pr., in kakor, ne bil, ter je, moj bog

se je bil, kakor bi se, i. t. d., da bi ne, bi se bil, se je bila, na vse strani, mu je bil, mu je bila, se je bilo

KRES

mag., členu, dodamo, določbe, priprava, varstva, odločbe, 1999, organa, odstavka

z dne, d. d., tega zakona, s področja, v postopku, v obdobju, osebnih podatkov, zaradi česar, foto reuters, za opravljanje

d. o. o., člena tega zakona, iz prejšnjega odstavka, pri tem pa, v republiki Sloveniji, člena zakona o, državna revizijska komisija, po vsem svetu, v nasprotju s, v sodelovanju z

JANES

:), ;), :d, :p, :-), #link, :)), slo., :(, 😊

v slo., v lj., p. s., for the, on the, to the, this is, to be, is a, is the

hvala za odgovor, tole je pa, še malo pa, ha ha ha, na to temo, me zanima če, zanima me če, je možno da, všeč mi je, in lep pozdrav

Conclusion

- collection of n-gram frequency lists for selected corpora
 - regular, filtered and adjusted
 - 1-5 normalized tokens
 - **CLARIN.SI repository (CC-BY-SA)**
- first quantitative comparison
 - similar structure and frequency of core vocabulary
 - a substantial amount of formulaic sequences
 - only partial overlap of n-grams

Future work

- open-source tool (NSSS, CLARIN.SI)
 - computationally efficient
 - TEI-friendly
 - GUI
 - new functionalities
- new lists
- linguistic analysis
 - categorization by structure and function

Thanks!