CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

Comparing the Incomparable?

Rethinking n-grams for free word order languages

Lucie Lukešová (Chlumská) & David Lukeš

Faculty of Arts, Charles University (Prague)

OUTLINE

1. Using n-grams in contrastive studies
2. Major issues in n-gram extraction
3. An alternative to n-grams in free word order languages: n-choose-k-grams
4. Results: comparing methods

N-GRAMS IN CONTRASTIVE STUDIES

What is an n-gram?

- a sequence of n-words (tokens): n=3

Research shows that *children who read well do well .*

Research shows that children who read well do well .

Research shows that children who read well do well .

Research shows that children who read well do well .

Research shows that children who read well do well .

Research shows that children who read well do well .

Research shows that children who read well do well .

Research shows that children who read well do well .

- recurrent n-grams are interesting for linguistic analysis
  - they can reveal patterns, the syntagmatic nature of language and its grammatical, lexical and syntactic tendencies

Studies using n-grams

- First extensively used probably by Biber et al. (1999)
- Baker (2004): translated versus non-translated language
- Forchini and Murphy (2008): 4-grams in Italian and English
- Cortes (2008): 4-grams in English and Spanish
- Ebeling and Oksefjell Ebeling (2013): n-grams in English and Norwegian
- Granger (2014) and Granger & Lefer (2013): n-gram methodology in a comparison of English and French
- Čermáková & Chlumská (2017): English and Czech place expressions
- etc.

Issues in n-gram extraction

- General issues or **what to extract**?
  - suitable n-gram length?
  - minimum frequency of occurrence?
  - words, or lemmas?

- Further issues arise in cross-linguistic studies (cf. Granger 2014)
  - **length correspondence**
    - 4 – 4 *from side to side – ze strany na stranu*
    - 4 – 2 *he said to himself – řekl si*
    - 4 – 1 *for the first time – poprvé*
  - **word form variability** (*I am sure : jsem si jist/jistý/jistá*)
  - **free word order**

Czech v. English

- comparable corpora, the same frequency threshold...

	3-grams	4-grams	5-grams
Sample 1 (CZ)	150	41	25
Sample 2 (CZ)	103	9	7
Sample 3 (CZ)	170	21	9
Sample 4 (CZ)	119	19	6
Sample 5 (EN)	1036	360	169
Sample 6 (EN)	1198	454	190

(taken from Čermáková & Chlumská, 2017)

Free word order issue

A common feature in Czech (often connected to clitics):

myslel jsem si že (‘I thought that’)

jsem si myslel že (‘I thought that’)

Often combined with the issue of variable slots:

myslel jsem si nejdřív že

jsem si ale myslel že

jsem si totiž myslel že

etc.

AN ALTERNATIVE TO N-GRAMS

Challenges in automatic identification of recurring multi-word patterns

1. propensity of language for multi-word expressions
   - EN: *for the first time* × CZ: *poprvé*
   - no solution ☹ (shows limitations of "word" as cross-linguistic concept)
2. inflection
   - *research shows that* × *research showed that*
   - solution: lemmatization
3. variable slots
   - *once a ___ always a ___*
   - (partial) solution: skip-grams
4. free word order

Challenges in automatic identification of recurring multi-word patterns

1. propensity of language for multi-word expressions
   - EN: *for the first time* × CZ: *poprvé*
   - no solution ☹ (shows limitations of "word" as cross-linguistic concept)
2. inflection
   - *research shows that* × *research showed that*
   - solution: lemmatization
3. variable slots
   - *once a ___ always a ___*
   - (partial) solution: skip-grams
4. free word order

} *n-choose-k-grams* attempt to address both of these

An example

3-token window

↓

Research shows that children who read well do well .

Take account of all (unordered) combinations of 2 tokens within the window:

- { research, shows } (= { shows, research })
  - { shows, that } (= { that, shows })
  - { research, that } (= { that, research })

An example

3-token window

↓

Research shows that children who read well do well .

Take account of all (unordered) combinations of 2 tokens within the window :

- { shows, that } (= { that, shows })
- { that, children } (= { children, that })
- { shows, children } (= { children, shows })

An example

3-token window

↓

Research shows that children who read well do well .

Take account of all (unordered) combinations of 2 tokens within the window:

- { that, children } (= { children, that })
- { children, who } (= { who, children })
- { that, who } (= { who, that })

What to call the { … } entities?

- our pick: **3-choose-2-grams** – why?
- in combinatorics, "3 choose 2" is a shorthand for the number of different unordered combinations of 2 items that can be chosen from a set of 3

$$\text{“3 } choose \text{ 2”} = \binom{3}{2} = \frac{3 \times 2 \times 1}{2 \times 1} = 3$$

→ In each window of 3 tokens, 3 unordered combinations of 2 items can be considered.

n-choose-k-grams, version 1

In general:

1. Slide n-token window over each sentence in corpus.
2. Take account of all k-combinations of tokens ($k \leq n$) within the window.

Notice:

- **unordered** combinations → **free word order**
- when ***k*** **<** ***n*** → leaves room for gaps → **variable slots**

Caveat #1: Don't count twice

Research shows that children who read well do well .

3-choose-2-gram	frequency
→ { research, shows }	1
→ { shows, that }	1
→ { research, that }	1

Caveat #1: Don't count twice

Research shows that children who read well do well .

3-choose-2-gram	frequency
{ research, shows }	1
→ { shows, that }	2 (!)
{ research, that }	1
→ { that, children }	1
→ { shows, children }	1

Caveat #1: Don't count twice

Research shows that children who read well do well .

3-choose-2-gram	frequency
{ research, shows }	1
{ shows, that }	2 (!)
{ research, that }	1
→ { that, children }	2 (!)
{ shows, children }	1
→ { children, who }	1
→ { that, who }	1

Additional rule #1: Except for the first n-token window in each sentence, only k-combinations involving the most recently added token should be considered.

Caveat #2: Don't exclude sentences shorter than *n* but at least as long as *k*

- Task: Extract 3-choose-2-grams from *John sleeps*.

- Current answer: Can't slide a 3-token window over a 2-token sentence → abort.
- **Arguably a better answer:** We can still extract 2-combinations from a 2-token sentence → { john, sleeps }

Additional rule #2: If $n >$ length of sentence $\geq k$, bypass the sliding window step and extract k-combinations from the entire sentence.

n-choose-k-grams, version 2

1. Slide n-token window over each sentence in corpus.
2. Take account of all k-combinations of tokens ($k < n$) within the window.
3. Except for the first n-token window in each sentence, only k-combinations involving the most recently added token should be considered.
4. If $n >$ length of sentence $\geq k$, bypass the sliding window step and extract k-combinations from the entire sentence.

DATA

Test corpus

- contemporary written Czech
- texts from the scientific domain (both natural sciences and humanities) → formulaic language

documents	70
sentences	121,697
tokens	2,379,832
tokens (excl. punctuation)	2,023,724

RESULTS

Free word order

Observation: n-gram frequencies are generally much lower in Czech than in English for a variety of reasons, including free word order.

↓

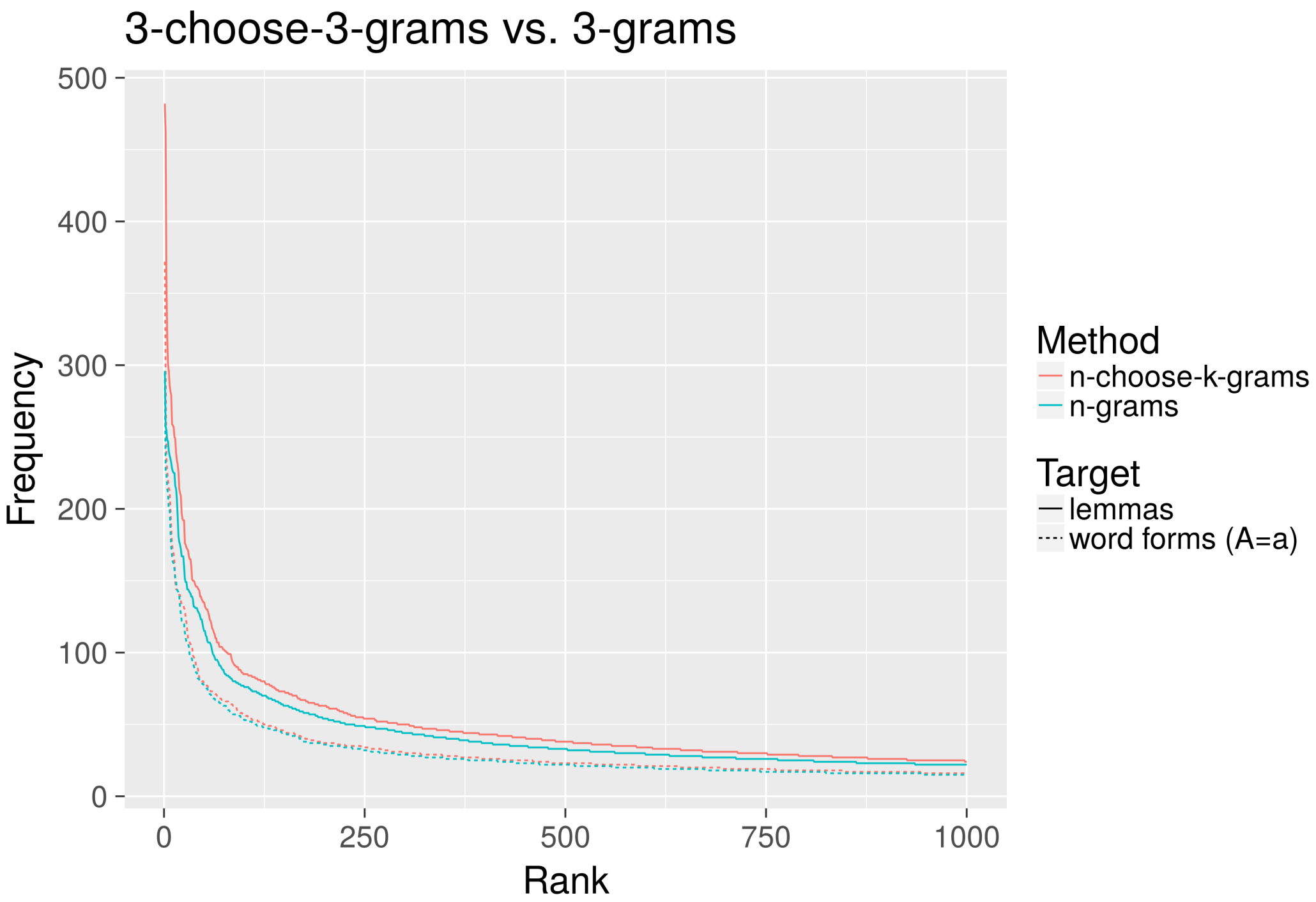
Question: If we found a way of looking past word order in Czech n-grams, would the observed frequencies increase?

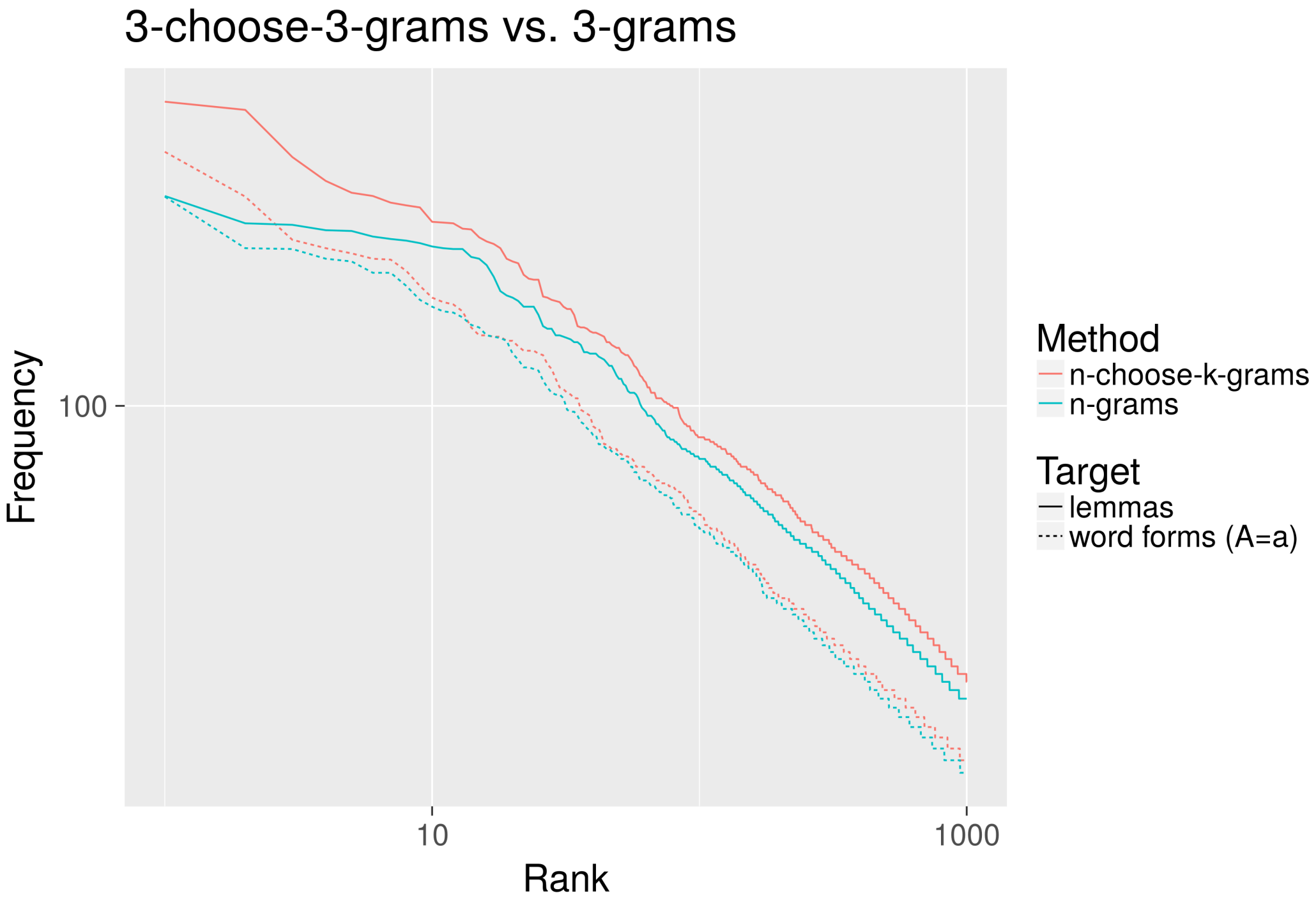
↓

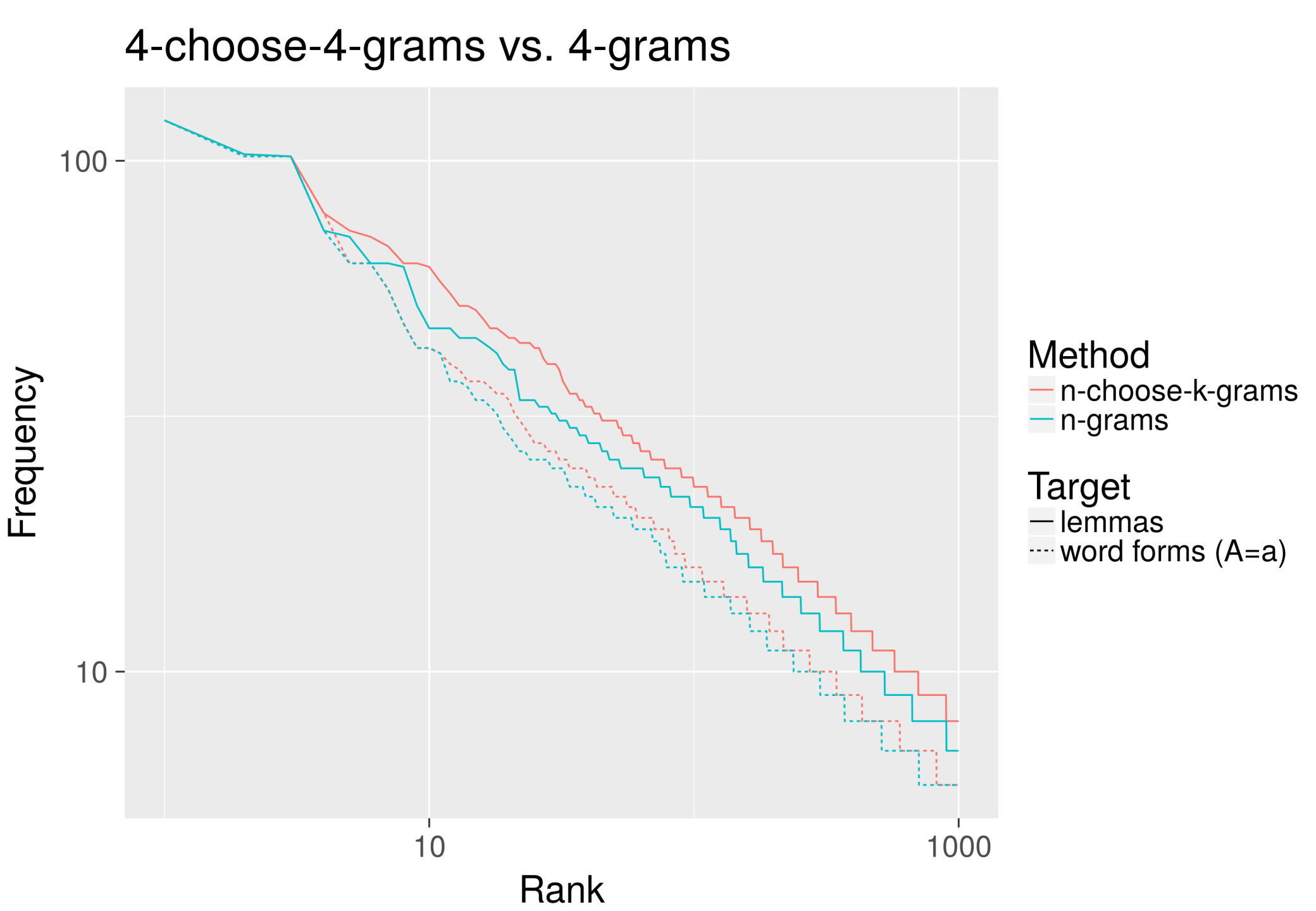
Solution: n-choose-k-grams ignore the ordering of constituents.

↓

Experiment: Compare Czech n-grams with Czech n-choose-k-grams where $n = k$. Do the latter yield higher frequencies?


3-choose-3-grams vs. 3-grams
Frequency
500
400
300
200
100
0
0
250
500
750
1000
Rank
Method
n-choose-k-grams
n-grams
Target
lemmas
word forms (A=a)


3-choose-3-grams vs. 3-grams
Frequency
100
10
1000
Rank
Method
n-choose-k-grams
n-grams
Target
lemmas
word forms (A=a)

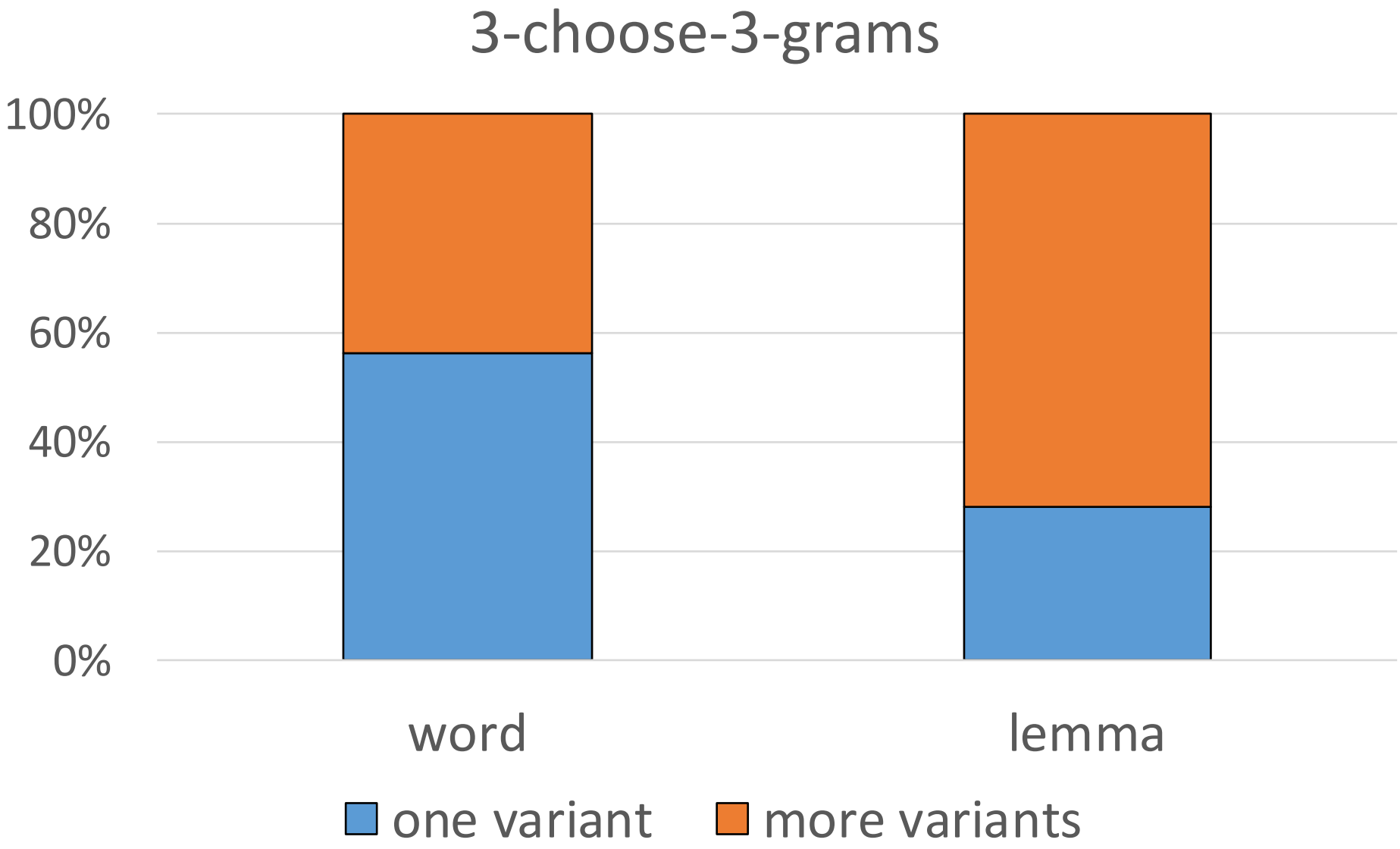

4-choose-4-grams vs. 4-grams
Frequency
100
10
10
1000
Rank
Method
n-choose-k-grams
n-grams
Target
lemmas
word forms (A=a)

One v. more variants

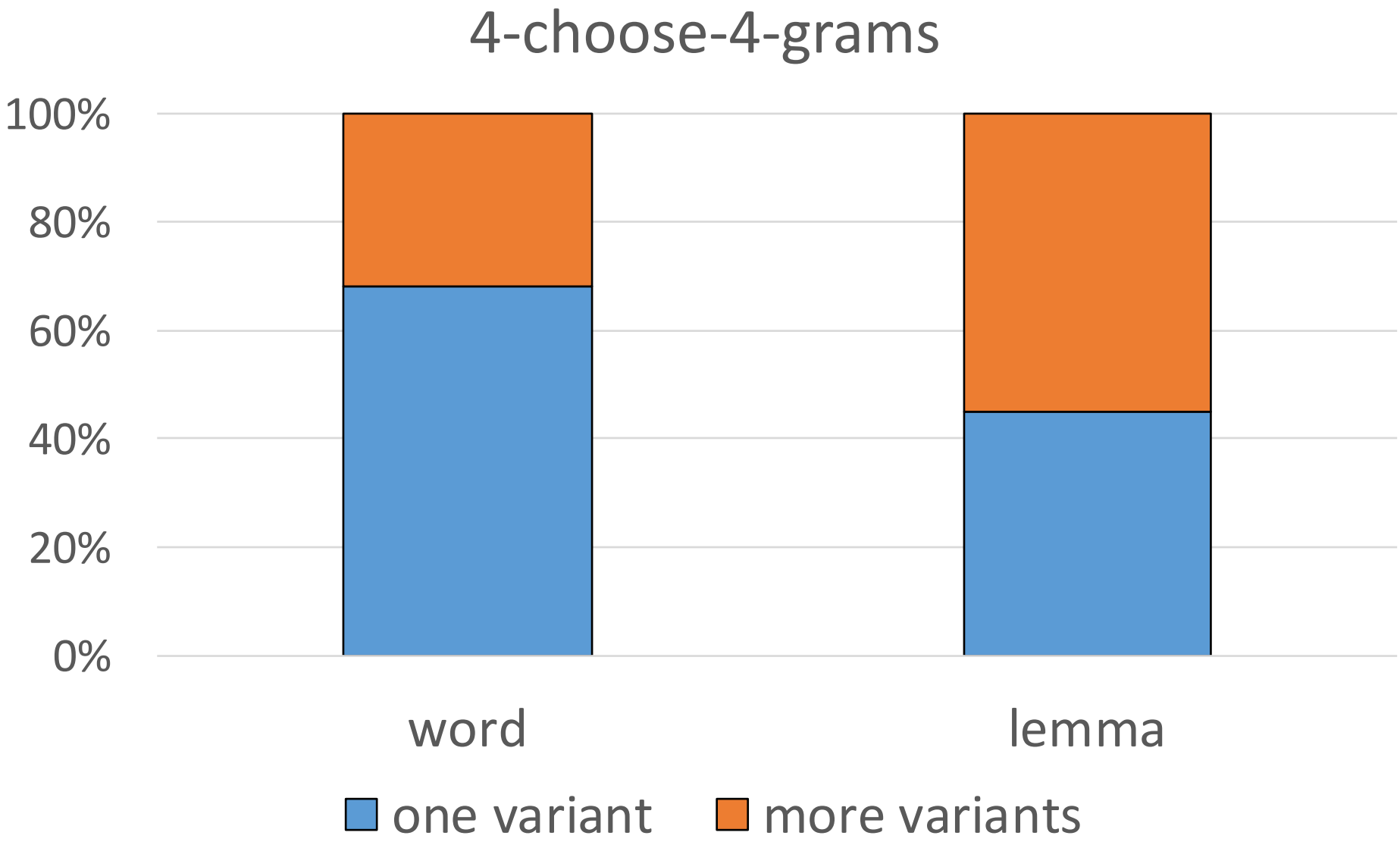
Example:

{ bez, na, ohledu }	>	*bez ohledu na*	>	only 1 variant
{ jednat, o, se }	>	*jednat se o*	>	2 variants
		se jednat o		
{ ale, je, to }	>	*ale je to*	>	5 variants!
		ale to je		
		to je ale		
		to ale je		
		je ale to		

Proportion of multiple variants



Proportion of multiple variants



Conclusions

We have probably run out of time by now… So quickly:

- n-choose-k-grams:
  - group word order variants of multi-word patterns under one entry → boosts frequency of some patterns
  - allow variable slots embedded within multi-word patterns (empirical details another time)

- not a silver bullet, of course!

Selected references

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167–193.

Biber, D., Conrad, S., Finegan, E., Leech, G. & Johansson, S. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Čermáková, A. & Chlumská, L. (2017). Expressing 'place' in children's literature: testing the limits of the n-gram method in contrastive linguistics. In T. Egan & H. Dirdal (Eds), *Cross-linguistic Correspondences: From lexis to genre*, pp. 75–95. Amsterdam: John Benjamins.

Cortes, V. (2008). A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish. *Corpora* 3 (1), 43–57.

Ebeling, J. & Oksefjell Ebeling, S. (2013). *Patterns in Contrast*. Amsterdam: John Benjamins.

Forchini, P., & Murphy, A. (2008). N-grams in comparable specialized corpora. Perspectives on phraseology, translation, and pedagogy. *International Journal of Corpus Linguistics*, 13(3), 351–367.

Granger, S. (2014). A Lexical Bundle Approach to Comparing Languages: Stems in English and French. *Languages in Contrast* 14 (1), 58–72.

Granger, S. & Lefer, M.-A. (2013). Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries. In K. Aijmer & B. Altenberg (Eds), *Advances in Corpus-based Contrastive Linguistics: Studies in honour of Stig Johansson*, pp. 157–176. Amsterdam: John Benjamins.

Thank you for your attention!

lucie.chlumska@korpus.cz
david.lukes@korpus.cz

Comparing n-choose-k-grams using entropy

- entropy ~ empirical freq. dist. over observed variants **(= uncertainty over variants)**
- entropy upper bound ~ uniform freq. dist. over all possible variants
- relative entropy = entropy / entropy upper bound

n-choose-k-gram: frequency	observed variants: frequency	relative entropy
{ na, od, rozdíl }: 296	*na rozdíl od*: 296	0
{ jednat, o, se }: 482	*se jednat o*: 247, *jednat se o*: 235	0.39
{ být, mít, ten }: 63	[showing only frequencies]: 17, 16, 13, 9, 6, 2	0.91