Frequency Wordlists

Dictionary Express

František Kovařík

Riddle

100 000 most frq. headwords ~ 1984–2024

Riddle

100 000 most frq. headwords ~ 1984–2024

2nd 100 000 ~ 2024–???

Riddle

100 000 most frq. headwords ~ 1984–2024

2nd 100 000 ~ 2024– ~~2064~~

Frequency wordlist

verb (25,676 items)

	Word	Lemma	Tag	Frequency
1	is	be	VBZ	730,929
2	are	be	VBP	379,918
3	be	be	VB	344,024
4	was	be	VBD	278,396
5	has	have	VHZ	210,972

Frequency types

Absolute Frequency (+ relative F)

Document Frequency

ARF

ALDF

Dictionary Express

Words from corpora

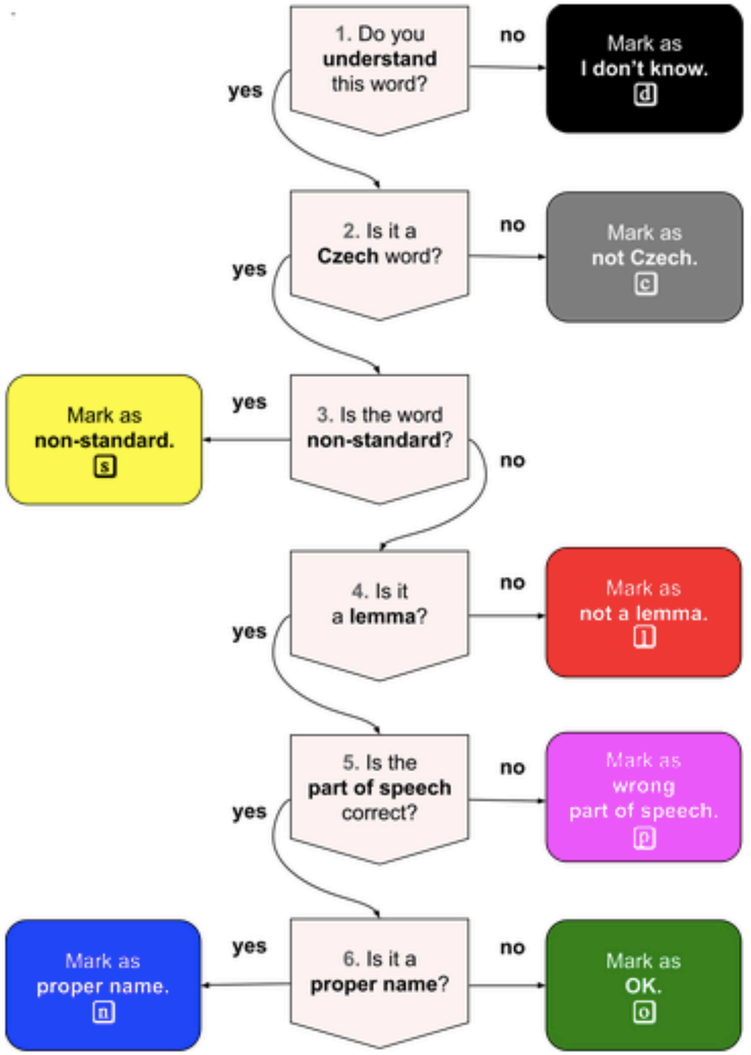
+ opinions of people.

CZECH HEADWORDS 119 (VERONIKAS)

total 1000 entries

search | starts like this

1. **kábl** *noun* — non-standard
2. **káblík** *noun* — non-standard
3. **Kábrt** *noun* — proper name
4. **káča** *noun* — OK
5. **káčátko** *noun* — OK
6. **kádinka** *noun* — OK
7. **kádrovák** *noun* — non-standard
8. **kádrování** *noun* — OK
9. **kádrovat** *verb* — OK
10. **káhirský** *adjective* — OK
11. **kách** *adjective* — I don't know



CZECH REVISION 8 (MARTINR)

total 1000 entries

search | starts like this

NEW + | ID 21621 | SAVE | CLONE | DELETE

1. **Kačena** *noun* ✓
2. **Kaláb** *noun* ✓
3. **kalafuna** *noun* ✓
4. **Kalas** *noun* ✓
5. **Kalaš** *noun* ✓
6. **kale** *adverb* ✓
7. **kalhot** *noun* ✓
8. **kalina** *noun* ✓
9. **Kalina** *noun* ✓
10. **kaliště** *noun* ✓
11. **Kalivoda** *noun* ✓

Kalas NOUN

(no consensus) opinions: I don't know | proper name

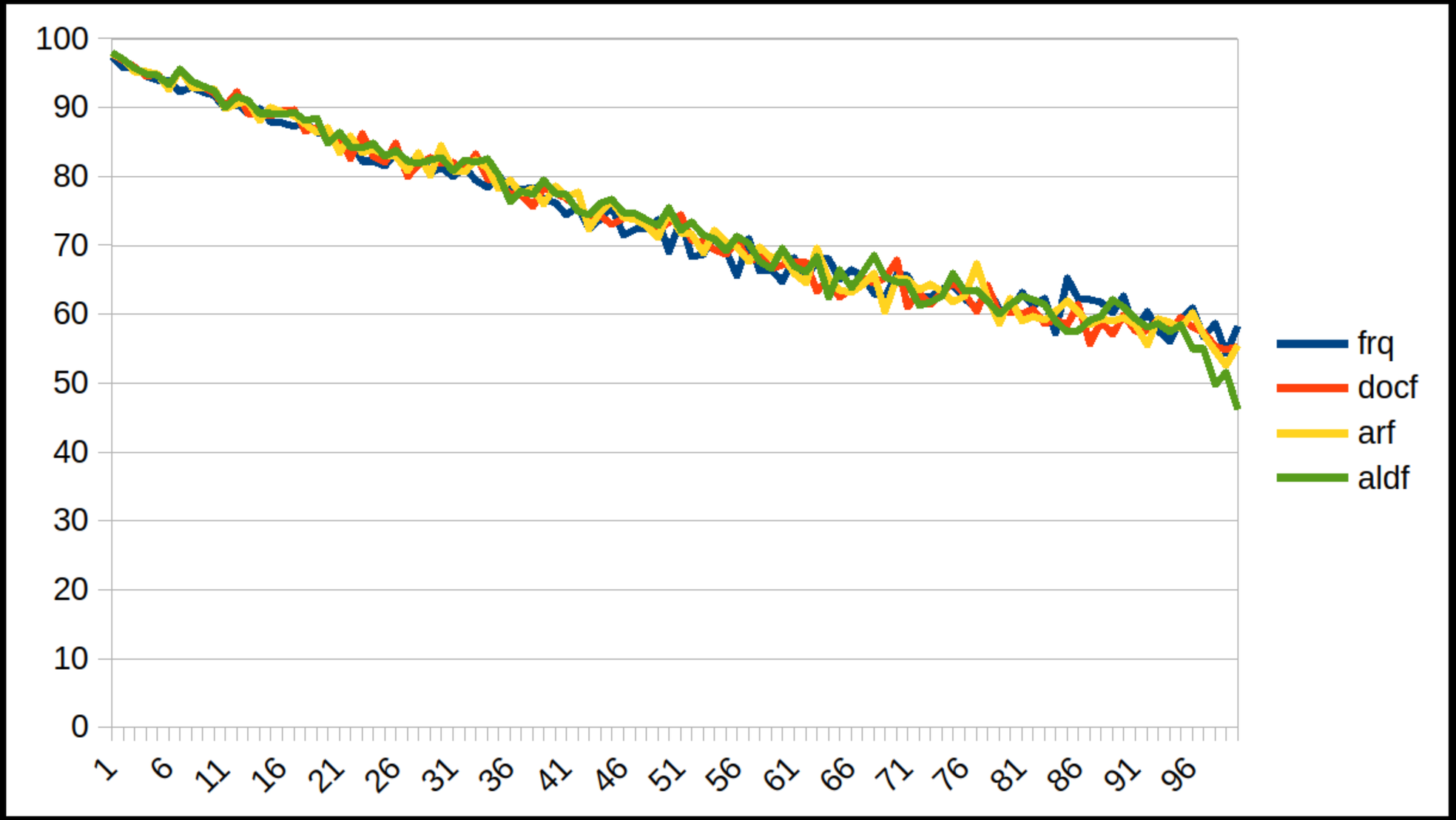
- ○ CORRECT HEADWORD SHOULD BE:
- ○ I DO NOT UNDERSTAND THIS WORD
- ○ THIS IS NOT A CZECH WORD
- ● THIS HEADWORD IS CORRECT

proper name? ✓

EXAMPLES

1. **Kalas** poskytne cyklistické i volnočasové oblečení pro všechny jezdce a realizační tým.
2. **Kalas** sice u gólu chyboval, ale špatnej určitě neni.
3. Kde bude působit v příští sezoně **Kalas**?
4. Pak cyklo dres **Kalas**, jeden velikost 158 a druhý dres velikost 164.
5. Na počátku značky **Kalas** nebyl podnikatelský záměr, ale láska k cyklistice.
6. Chelsea získala **Kalase** v létě 2010 z Olomouce.
7. Když nebudu padat a tloustnout, tak mi takový dres od **Kalasu** musí vydržet minimálně 5 se

Frequent x “trouble-less“



Czech DExpress

100 000 corpus lemposes

=> cca 79 700 dictionary entries.

Frequency difference: a 100 000 DocF wordlist

AF: 4962

ARF: 1722

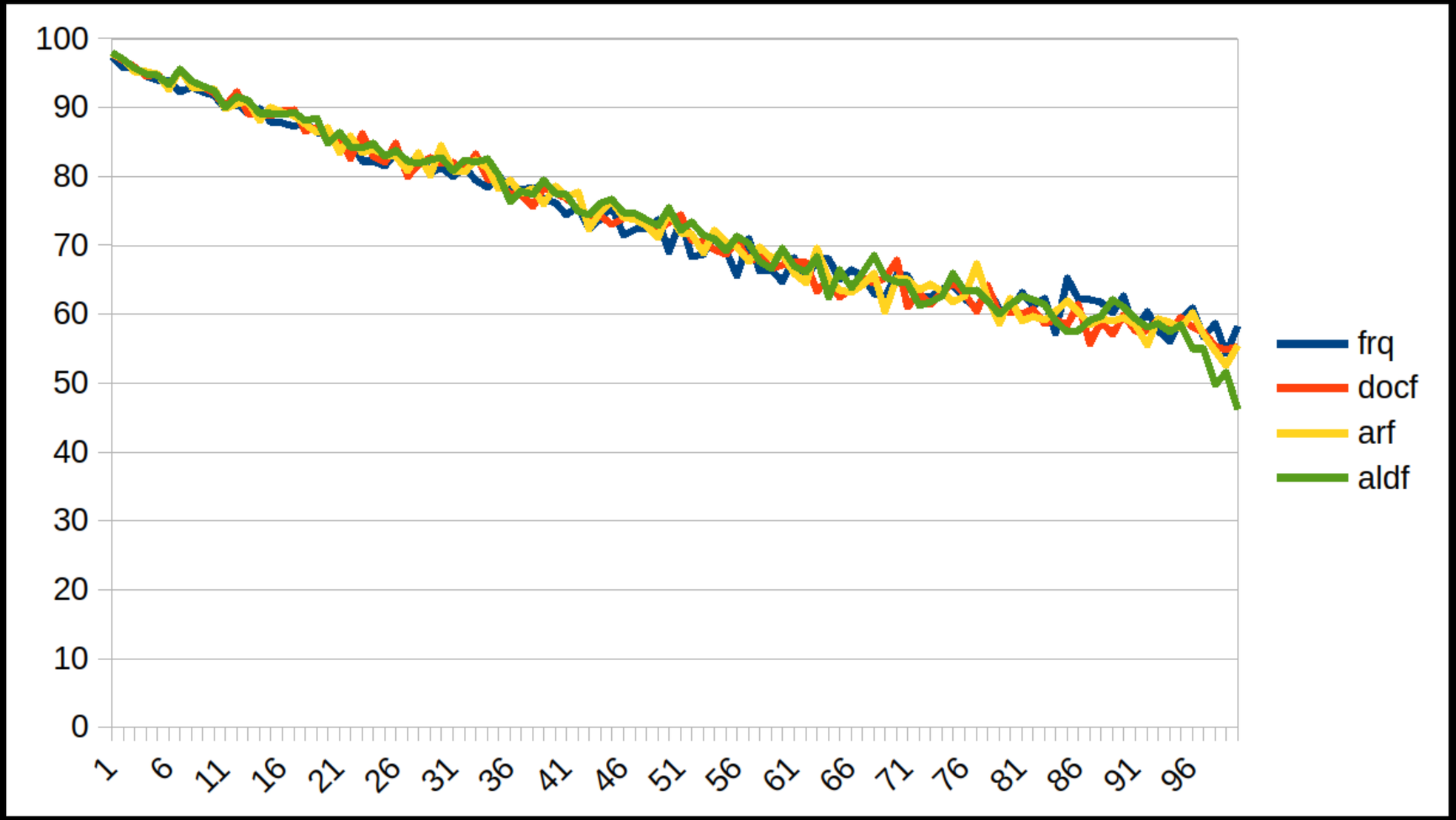
ALDF: 1927

For 100 000 entries?

Cca 150 000 corpus lemposes

(assumption)

Frequent x “trouble-less“



Which frequency?

	80 000	20 000
frq	76.78%	59.70%
docf	77.07%	58.13%
arf	77.29%	57.24%
aldf	77.62%	55.92%

Comparing 2 frequency wordlists

missing accepted

present accepted

missing rejected

present rejected

without annotation

Present accepted – absolute frequency

Vareni.cz, *Echo24*, Skyscanner, Ulož.to, ČSFD.cz

… lacks less frequent Czech words.

Present accepted – ARF and ALDF

ARF – more URLs and company names.

Present accepted – ARF and ALDF

ARF – more URLs and company names.

ALDF has the most „normal“ Czech words.

Present accepted – ARF and ALDF

ARF – more URLs and company names.

ALDF has the most „normal“ Czech words.

=> ALDF could be the best freq. type.

Incomplete research

… we still need to examine document frequency.

Wrapping up

Not many differences

in 100k headwords.

Another 100k could be

more interesting.