# FREQUENCY (NOT) SACRED:
## THE HEADWORD LIST OF A CONTEMPORARY CZECH MONOLINGUAL DICTIONARY AND CORPORA

Ústav pro jazyk český
AV ČR, v. v. i.
(Czech Language Institute)

Jana Nová
Vít Michalec
Zdeňka Opavská
Renáta Neprašová

# Academic Dictionary of Contemporary Czech

- a monolingual dictionary
- started in 2012
- DWS Alexis
- a medium-sized dictionary (120–150,000 headwords)
- the vocabulary of contemporary Czech (since 1945)
- the headword list built on corpus data
- emphasis on the autonomy of lexemes (lexical units)

# Proportional distribution in Czech dictionaries

| | SSJČ number of headwords | % | SSČ number of headwords | % | ASSČ assumption |
|---|---|---|---|---|---|
| A | 3032 | 1.9 | 870 | 1.77 | 2124 |
| B | 5611 | 3.52 | 1445 | 2.94 | 3528 |
| C | 1573 | 0.99 | 433 | 0.88 | 1056 |
| Č | 1565 | 0.98 | 499 | 1.02 | 1224 |
| D | 6529 | 4.1 | 2119 | 4.32 | 5184 |
| Ď | 48 | 0.03 | 7 | 0.01 | 12 |
| E | 1682 | 1.06 | 466 | 0.95 | 1140 |
| F | 2292 | 1.44 | 568 | 1.16 | 1392 |
| G | 1178 | 0.74 | 293 | 0.6 | 720 |
| H | 4594 | 2.88 | 1307 | 2.66 | 3192 |
| CH | 1509 | 0.95 | 475 | 0.97 | 1164 |
| I | 1514 | 0.95 | 500 | 1.02 | 1224 |
| J | 1842 | 1.16 | 640 | 1.3 | 1560 |
| K | 10131 | 6.36 | 3086 | 6.29 | 7548 |
| L | 3761 | 2.36 | 1104 | 2.25 | 2700 |
| M | 6485 | 4.07 | 1830 | 3.73 | 4476 |
| N | 8833 | 5.54 | 3216 | 6.55 | 7860 |
| Ň | 28 | 0.02 | 3 | 0.01 | 12 |
| O | 8791 | 5.52 | 3026 | 6.16 | 7392 |
| P | 25415 | 15.95 | 7769 | 15.82 | 18984 |
| Q | 33 | 0.02 | 4 | 0.01 | 12 |
| R | 7570 | 4.75 | 2257 | 4.6 | 5520 |
| Ř | 621 | 0.39 | 178 | 0.36 | 432 |
| S | 14006 | 8.79 | 4397 | 8.96 | 10752 |
| Š | 3774 | 2.37 | 911 | 1.86 | 2232 |
| T | 6183 | 3.88 | 1665 | 3.39 | 4068 |
| Ť | 72 | 0.05 | 17 | 0.03 | 36 |
| U | 3555 | 2.23 | 1287 | 2.62 | 3144 |
| V | 12039 | 7.56 | 3959 | 8.06 | 9672 |
| W | 74 | 0.05 | 14 | 0.03 | 36 |
| X | 59 | 0.04 | 12 | 0.02 | 24 |
| Y | 27 | 0.02 | 6 | 0.01 | 12 |
| Z | 13419 | 8.42 | 4338 | 8.84 | 10608 |
| Ž | 1461 | 0.92 | 401 | 0.82 | 984 |
| celkem | 159306 | | 49102 | | 120000 |

← numbers of headwords in SSJČ and SSČ

Number of headwords in Alexis

| | assumption | reality |
|---|---|---|
| A | 2124 | 2682 |
| B | 3528 | 3580 |
| C | 1056 | 1197 |
| Č | 1224 | 1141 |

# Making of the headword list

## automatically generated headword list

**- set of reference corpora** – SYN2000, SYN2005, SYN2010
→ common nouns → absolute frequencies ≥ 5

## raw headword list – manual checking → elimination

- **non-words :** *a*; a/3, a+b, a§, ad8, abcd, adresa_registru*
- **word fragments:** *áclava, ádný, ákladní*
- **abbreviations:** *adm, admin, administr, adv., afp, akč*
- **typos:** *abces, abysi, acylpirin, adéla, adsorbce*
- **lemmatization mistakes:** *abstrakcinistů; acidozou, acidozy; admirála, akolyté*
- **foreign words:** from foreign contexts, from one or limited sources – *about, abdominis, absolute, academia, accent, accounting, actually, address*
- **possessive adjectives ending in *-ův*, *-in*:** *adeptův, agentův, albatrosův, afroditin*
- **specialized terms with low frequency:** *acetylkoenzym, adagietto, aglutinin*

| | generated headword list | raw headword list | % |
|---|---|---|---|
| A | 5625 | 3203 | 57 |
| B | 5949 | 3829 | 64 |
| C | 2981 | 1609 | 54 |
| Č | 1900 | 1481 | 78 |

# Reduction of / addition to the raw headword list I.

**REDUCTION**

**frequency criterion**

- frequency ≥ 5, but in one or limited sources → exclude

**criterion of distribution**

- only or usually in professional periodicals → exclude

**ADDITION**

**systemic criteria**

- **word families -** derivatives (relational adjectives + adverbs, property names, feminine nouns derived from masculines)

**control criterion**

- **words from the headword list of SSČ**

# Reduction of the raw headword list II.

**limited selection - systemic criteria**

- non-lexicalised action nouns: *-ní, -tí*
- verbal adjectives from passive participles: *-ný, -tý*
- verbal adjectives from transgressives: *-cí*

# Reduction of the raw headword list III.

**limited selection - usually higher frequency limits**

- compounds - prefix / suffix derivatives (*bílo-, celo-, černo-, čtyř-; bio-, eko-, euro-; anti-, bez-, ne-, -hodinový, -kilogramový, -fil, -fob*)
- chemical, physical, medical etc. terms not in common usage: *aldosteron, anizotropie, ascites*
- marginal slang words: *backupovat, botanizovat*
- marginal historicisms: *apelplac, ariston, beghard, cechovna*
- marginal archaisms: *arma, bábelský, bohověda*
- ephemeral expressions or nonce words: *véčkač, klausismus, kolébkoviště*
- marginal vulgarisms with metaphorical meaning: *bobr, číča*
- common nouns from proper names – especially from product names: *bugatti, cooper*
- words less known in the Czech environment – exotic foods, products etc.: *cachaca*

# Inclusion I.

**words from an allocated section of the headword list frequency in SYN (Araneum Bohemicum Maximum, electronic archive Newton Media a. s., Internet)**

| | | |
|---|---|---|
| článek | 39160 | |
| článkování | 15 | |
| článkovaný | 142 | |
| článkovat | 5 | SYN v6 17× (z toho 8 chyb); NW 36× (chyby), Ar 35× (chyby) |
| článkovitý | 10 | SYN v6 38×; NW 45×; Ar 139× |
| článkový | 107 | |
| člen | 101473 | |
| členění | 3019 | |
| členěný | 849 | |
| členicí | 6 | SYN v6 15×; NW 58×; Ar 37× |
| členící | 50 | |
| členit | 1726 | |
| členitě | 12 | SYN v6 74×; NW 86×; Ar 107× (v rámci hnízda) |
| členitele | 8 | - **členitel**; SYN v6 23× (pouze 2 odb. zdroje); NW 0×; Ar 3× |
| členitelný | 5 | SYN v6 12×; NW 7×; Ar 38× |
| členitost | 303 | |

# Inclusion II.

| členicí | 6 |
|---|---|

**INCLUDED**

not lemmatized correctly

**syn v6**

Query type: Lemma → Query: členicí 15 hits / Doc Ids: 13 items

Query type: Word part → Query: členicí 33 hits / Doc Ids: 22 items

kontext Query Corpora Save Concordance Filter Frequency Collocations View Help

**Corpus:** syn v6 | **Query**: .*členicí.* (33 hits)

Hits: **33** | i.p.m. ⓘ: **0.01** (! related to the whole "syn_v6") | ARF ⓘ: **12.81** | Result is sorted 1 / 1

Line selection: simple | Attributes:

| | | | | |
|---|---|---|---|---|
| ☐ | Gotika | o několik desetiletí později jeho kolega v Auxerre stěnu a | členicí | prvky . Novostavba chóru katedrály zahájená roku 1215 ( obr. |
| ☐ | Marketing & Media | nastavitelným stahováním , kapsy na zip včetně náprsní všité do | členicího | švu ) . Netkaná textilie je využita i v konstrukci |
| ☐ | Technický týdeník | . 20130404057 Irská společnost poptává výrobce plastů pro výrobu patentovaných | členicích | příček do odpadních košů umožňujících snadnější recyklaci . 20130404005 Ruský |
| ☐ | Auto Tip | v zadní části kabiny Zvýšená záď a maximum horizontálních horizonto | členicích | prvků - to je recept na moderní eleganci ci Stačí |
| ☐ | Juristická a lingvistická analýza právních textů | že nelze vyloučit i " homonymní " citační výskyt uvedených | členicích | signálů , tj. takový výskyt , který má po stránce |
| ☐ | Juristická a lingvistická analýza právních textů | 4.1 . Představuje vnitřní strukturaci formální věty . Jde o | členicí | prostředek složitějších formálních vět výčtového typu , jejichž struktura je |
| ☐ | Juristická a lingvistická analýza právních textů | . 1.2 . 4.3 . Písmeno se tedy uplatňuje jako | členicí | prostře - dek identifikace podvětných struktur v rámci formální věty |
| ☐ | Juristická a lingvistická analýza právních textů | . Z hlediska posuzování vlastností výrazových prostředků členění textu ( | členicích | signálů ) se zajímavě uplatňuje středník : je jednak signálem |
| ☐ | Juristická a lingvistická analýza právních textů | pro tyto své distinktivní vlastnos - ti uvažován jako významný | členicí | signál zvláště pro segmentaci formálních vět na podvěty , které |
| ☐ | Juristická a lingvistická analýza právních textů | zdá se , bez obtíží nahrazen tečkou nebo jiným verbálním | členicím | signálem podvětné strukturace . Tím by se formální kritéria formální |
| ☐ | Na co si ještě vzpomenu | zaujímal postoj , že dekor má v architektonickém celku významnou | členicí | funkci , a nazvat ho ozdobou je tedy úplně pochybené |
| ☐ | Hospodářské noviny | pravidla připouštějí rovněž psaní s dlouhou samohláskou - umísťovat ) | členicí | znaménka , tedy konkrétně čárky . Výše uvedený příklad je |
| ☐ | Mladá fronta DNES | siluetě rozšířené k dolnímu okraji . Opět jsou nutné podélné | členicí | švy . Pláště mohou být jednořadové nebo dvouřadové , ale |
| ☐ | Mladá fronta DNES | nebo dvouřadové , ale vždy s malými klopami a s | členicími | podélnými švy . Bez podložených a zdůrazněných ramen se tyto |
| ☐ | Hospodářské noviny | vyšitým monogramem . Nejslušivější saka mají prodlouženou délku se svislými | členicími | švy , kte - ré postavu opticky prodlužují . U |
| ☐ | Mladá fronta DNES | užší , aby působila přirozeně . Návrháři si hrají s | členicími | švy , štepy , krytým zapínáním , stahovacími pásky v |
| ☐ | Týden | v poslední kolekci Kláry Nademlýnské často vyskytuje , jsou zdůrazněné | členicí | švy . V tomto případě je sukně se všitou kožešinou |
| ☐ | Metro | jejíž kouzlo spočívá v maximálním prosklení a minimalizaci nosných i | členicích | rámů . POKRAČOVÁNÍ NA STRANĚ 06 Přírodní styl Zimní zahrada |
| ☐ | Lidové noviny | , s optickými iluzemi si totiž může pohrávat tvar zadního | členicího | švu ( pod páskem ) . Ostřejší trojúhelníkový tvar opět |
| ☐ | Právo | formálního hlediska kromě využívání velkých a malých liter , tzv. | členicích | znamének či odstavců příliš možností k experimentům , pokud si |

# Inclusion III.

| členitele | 8 |
|---|---|

**NOT INCLUDED**

lemmatization mistakes → **členitel**

**syn v6**

Query type: Lemma → Query: členitel 23 hits / Doc Ids: 2 items (professional sources)

kontext Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápověda

**Korpus:** syn v6 | **Dotaz:** členitel (23 výskytů)

Výskytů: **23** | i.p.m. : **0** (! vztaženo k celému "syn_v6") | ARF : **1,97** | Výsledek je setříděn 1 / 1

Výběr řádků: základní | Atributy:

| | | | | |
|---|---|---|---|---|
| Digitální fotografie | charakteristiky vyzařování daného zdroje F ( ) a příslušných trichromatických | členitelů | x ( ) , y ( ) a z ( |
| Digitální fotografie | ) pro zorný úhel pozorování do 4 o , resp. | členitelů | x 10 ( ) , y 10 ( ) a |
| Digitální fotografie | úhel větší než 4 o ) . Grafické znázornění trichromatických | členitelů | je na obrázku 5.4 . Křivka odpovídající členiteli y ( |
| Digitální fotografie | znázornění trichromatických členitelů je na obrázku 5.4 . Křivka odpovídající | členiteli | y ( ) odpovídá spektrální citlivosti lidského oka při fotopickém |
| Televizní technika | případě se tyto součinitele značí a nazývají se spektrální trojbarvé | členitele | . Při vyrovnání kolorimetru na bílé izoenergetické světlo jasu jsou |
| Televizní technika | rovné . Z této podmínky plyne měřítko pro poměrné velikosti | členitelů | na obr. 22 . Kdybychom měli dvě různá světla , |
| Televizní technika | některé literatuře , např. [ 1 ] jsou průběhy trojbarvých | členitelů | udávány v energetickém měřítku W/W , tzn. údaj ve wattech |
| Televizní technika | potřebné k vyrovnání jednoho wattu různé spektrální barvy . Průběhy | členitelů | se změní při použití měřítka v lumenech na watt ( |
| Televizní technika | barvy můžeme stanovit fotometricky nebo z průběhů jejich spektrálních trojbarvých | členitelů | vždy příslušné velikosti , tj. součinitele R , G , |
| Televizní technika | barvy mezi zelenou a modrou musíme však použít záporný trojbarvý | členitel | , a tím dostaneme i zápornou souřadnici - r . |
| Televizní technika | Y ) , ( Z ) stanovit normalizované spektrální trojbarvé | členitele | pro stejné výkony všech spektrálních barev ( obr. 31 ) |
| Televizní technika | vlnové délce nemají záporné hodnoty a poněvadž je spektrální trojbarvý | členitel | jediným zdrojem jasové informace , je jeho průběh shodný s |
| Televizní technika | pro bílé světlo C , můžeme stanovit průběhy spektrálních barvových | členitelů | pro televizní jednotky ( Re ) , ( Ge ) |
| Televizní technika | : Křivka citlivosti oka na obr. 8 odpovídá průběhu trojbarvého | členitele | a lze ji přiřadit měřítko v % jasových příspěvků od |
| Televizní technika | barev ; b ) tří základních barev Obr. 22 Trojbarevné | členitele | pro jednotlivé spektrální barvy izoenergetického světla při základních kolorimetrických barvách |
| Televizní technika | 34 Znázornění barev v polárních souřadnicích Obr. 36 Spektrální trojbarvé | členitele | televizních základních barev ( Re ) , ( Ge ) |
| Televizní technika | Souřadnice x , y a trojúhelník MKO Obr. 31 Trojbarvé | členitele | neskutečných barev ( X ) , ( Y ) , |
| Televizní technika | zařízení má přibližně stejnou spektrální citlivost , odpovídající průběhům trojbarvých | členitelů | pro základní barevná světla obrazovky . Zatím ještě nic nevíme |

## Our survey

**A**

- mostly foreign words / words of foreign origin
- many scientific terms
- prefixoid series (*aero-, agro-, akva-, anti-, audio-, auto-* …)

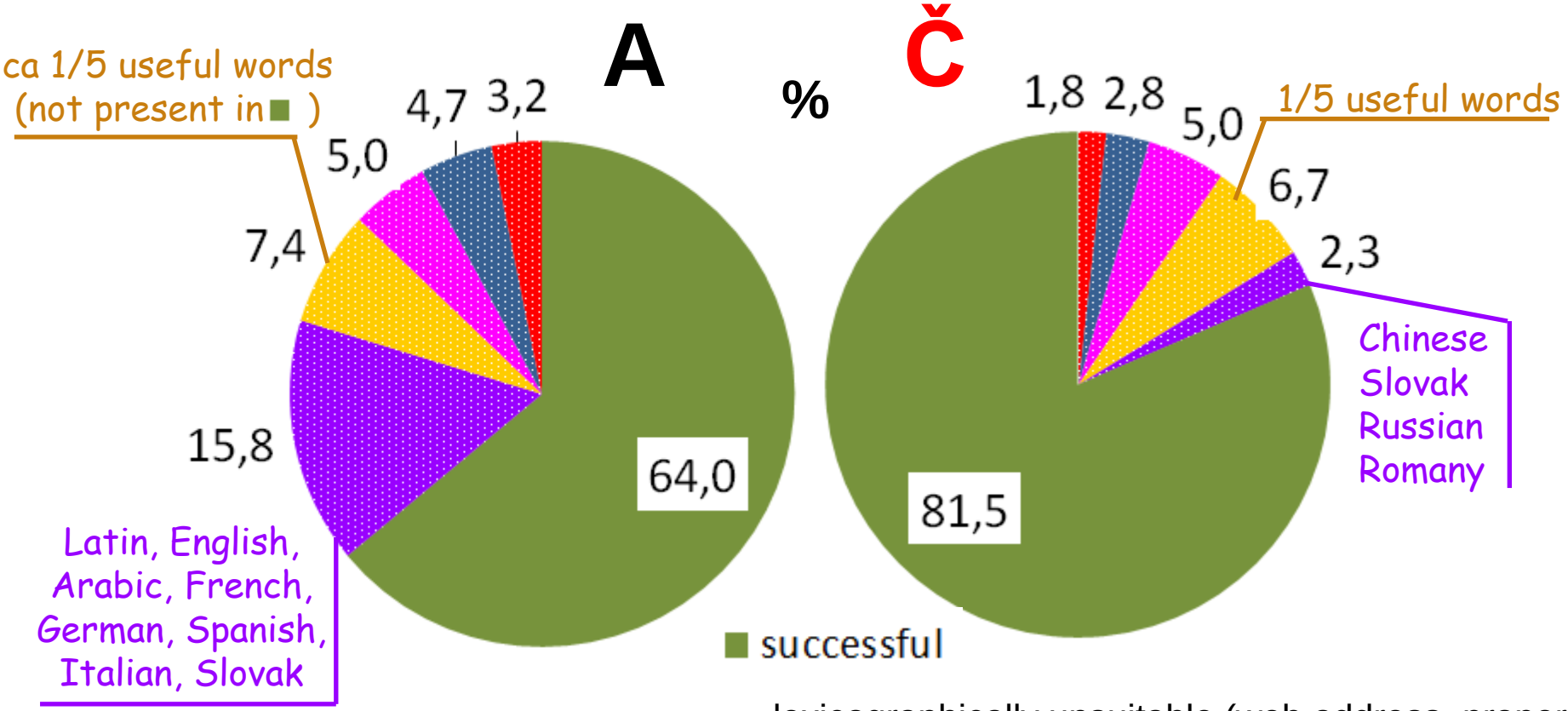
**Č**

- typical Czech letter/phoneme
- core lexicon words + expressive, colloquial words
- prefixoids, compounds (*černo-, červeno-, česko-, čtvrt-, čtyř-* …)

word list: absolute frequency $\geq$ 5 in 4 reference corpora together (SYN2000, SYN2005, SYN2010, SYN2015)
+ manual checking

A-AM: 2939 words – 1880 successful (64%) – 1059 unsuccessful
Č: 2169 words – 1767 successful (81.5%) – 402 unsuccessful

# Success rate in manual checking of the raw word list




unsuccessful:
- lexicographically unsuitable (web address, proper noun, nonce word…)
- non-word (abbreviation, formula…)
- text mistake (misspelling, fragment…)
- lemmatization mistake
- foreign language

# ARF compared to ABS: I. Numbers

- absolute frequency – ABS
- average reduced frequency – ARF:
  reduced for words occurring only in few sources
- 4 corpora (SYN2000-2015):

$$ARF = \frac{1}{v}\sum_{i=1}^{f}\min(d_i, v)$$

**A-AM**

| ABS | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| total | 2939 | 2605 | 2349 | 2145 |
| succ. (%) | 1880 (64%) | 1762 (67.7%) | 1662 (70.8%) | 1560 (72.7%) |

| ARF | 2 | 2.5 | 3 | 3.5 | 4 |
|---|---|---|---|---|---|
| Total | 4213 | 2737 | 2534 | 2098 | 1953 |
| succ. (%) | 2255 (53.5%) | 1815 (66.3%) | 1735 (68.5%) | 1538 (73.3%) | 1462 (74.9%) |

**Č**

| ABS | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| total | 2169 | 1983 | 1843 | 1731 |
| succ. (%) | 1767 (81.5%) | 1663 (83.9%) | 1569 (85.1%) | 1493 (86.3%) |

| ARF | 2.5 | 3 | 3.5 | 4 |
|---|---|---|---|---|
| Total | 2201 | 2087 | 1849 | 1725 |
| succ. (%) | 1820 (82.7%) | 1767 (84.7%) | 1608 (87%) | 1524 (88.7%) |

# ARF compared to ABS: I. Numbers

- ↑ threshold frequency = ↑ % success
- **ARF is not remarkably more efficient than ABS**
- the threshold frequency should be different for each letter

**A-AM**

| ABS | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| total | 2939 | 2605 | 2349 | 2145 |
| succ. (%) | 1880 (64%) | 1762 (67.7%) | 1662 (70.8%) | 1560 (72.7%) |

| ARF | 2 | 2.5 | 3 | 3.5 | 4 |
|---|---|---|---|---|---|
| Total | 4213 | 2737 | 2534 | 2098 | 1953 |
| succ. (%) | 2255 (53.[illegible]%) | 1815 (66.3%) | 1735 (68.5%) | 1538 (73.3%) | 1462 (74.9%) |

**2.1**
1904

82 more words to check
= 45 min. work

**č**

| ABS | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| total | 2169 | 1983 | 1843 | 1731 |
| succ. (%) | 1767 (81.5%) | 1663 (83.9%) | 1569 (85.1%) | 1493 (86.3%) |

| ARF | 2.5 | 3 | 3.5 | 4 |
|---|---|---|---|---|
| Total | 2201 | 2087 | 1849 | 1725 |
| succ. (%) | 1820 (82.7%) | 1767 (84.7%) | 1608 (87%) | 1524 (88.7%) |

# ARF compared to ABS: II. Words

What words do we get / lose using ARF / ABS? (4 corpora)

**A-AM**
ABS ≥ 5
ARF < 2.1
~ 150 words

**derivatives:** *abdukovat, agarově, aeronautka, amerikanistický…*
**terms:** *acidurie, adstrát, achondroplazie, aminofenol…*
☺ *ablaut, afrikáta, alonžový, ambuvak*

**Č**
ABS ≥ 5
ARF < 3
~ 100 words

**derivatives:** *čahounka, česárna, čirůvkový, čubrnění…*
**terms:** *čajot, čertkus, červor, čiplenka*

ABS < 5
ARF ≥ 2.1
~ 180 words

**derivatives:** *abatyšský, abstinentství, aktualitka, amébovitě…*
**[+ compounds]**
**terms:** *acesulfam, aldehydde-hydrogenáza, alkaptonurie…*
☺ *absolvenťák, alkáč, adžika, ajurvéda, ampersand*

ABS < 5
ARF ≥ 3
~ 100 words

**derivatives:** *čarodějnicky, čeledínský, číšnictví, čutorka…*
**[+ compounds]**
☺ *čatný, čočkostroj*

# ARF compared to ABS: II. Words

**terms + compounds:** very selective inclusion

**derivatives:** lexicographers are trained to check them within the word family

**A-AM**
ABS ≥ 5
ARF < 2.1
~ 150 words

**derivatives:** *abdukovat, agarově, aeronautka, amerikanistický…*
**terms:** *acidurie, adstrát, achondroplazie, aminofenol…*
☺ *ablaut, afrikáta, alonžový, ambuvak*

**Č**
ABS ≥ 5
ARF < 3
~ 100 words

**derivatives:** *čahounka, česárna, čirůvkový, čubrnění…*
**terms:** *čajot, čertkus, červor, čiplenka*

ABS < 5
ARF ≥ 2.1
~ 180 words

**derivatives:** *abatyšský, abstinentství, aktualitka, amébovitě…*
**[+ compounds]**
**terms:** *acesulfam, aldehydde-hydrogenáza, alkaptonurie…*
☺ *absolvenťák, alkáč, adžika, ajurvéda, ampersand*

ABS < 5
ARF ≥ 3
~ 100 words

**derivatives:** *čarodějnicky, čeledínský, číšnictví, čutorka…*
**[+ compounds]**
☺ *čatný, čočkostroj*

☺ **< 10 words per letter:** *does it matter?*

# Newer corpora: I. Making of the headword list

What do we get adding SYN2015 to SYN2000-2010?

1. A longer and dirtier list to check

**A-AM**, ABS ≥ 5
3 corp.: 2384 / 1656 succ. (69.5%)
4 corp.: 2939 / 1880 succ. (64%)

ARF ≥ 2.1
3 corp.: 2457 / 1669 succ. (67.9%)
4 corp.: 3004 / 1904 succ. (63.4%)

**Č**, ABS ≥ 5
3 corp.: 1912 / 1595 succ. (83.4%)
4 corp.: 2169 / 1767 succ. (81.5%)

ARF ≥ 3
3 corp.: 1773 / 1402 succ. (88.9%)
4 corp.: 2087 / 1767 succ. (84.7%)

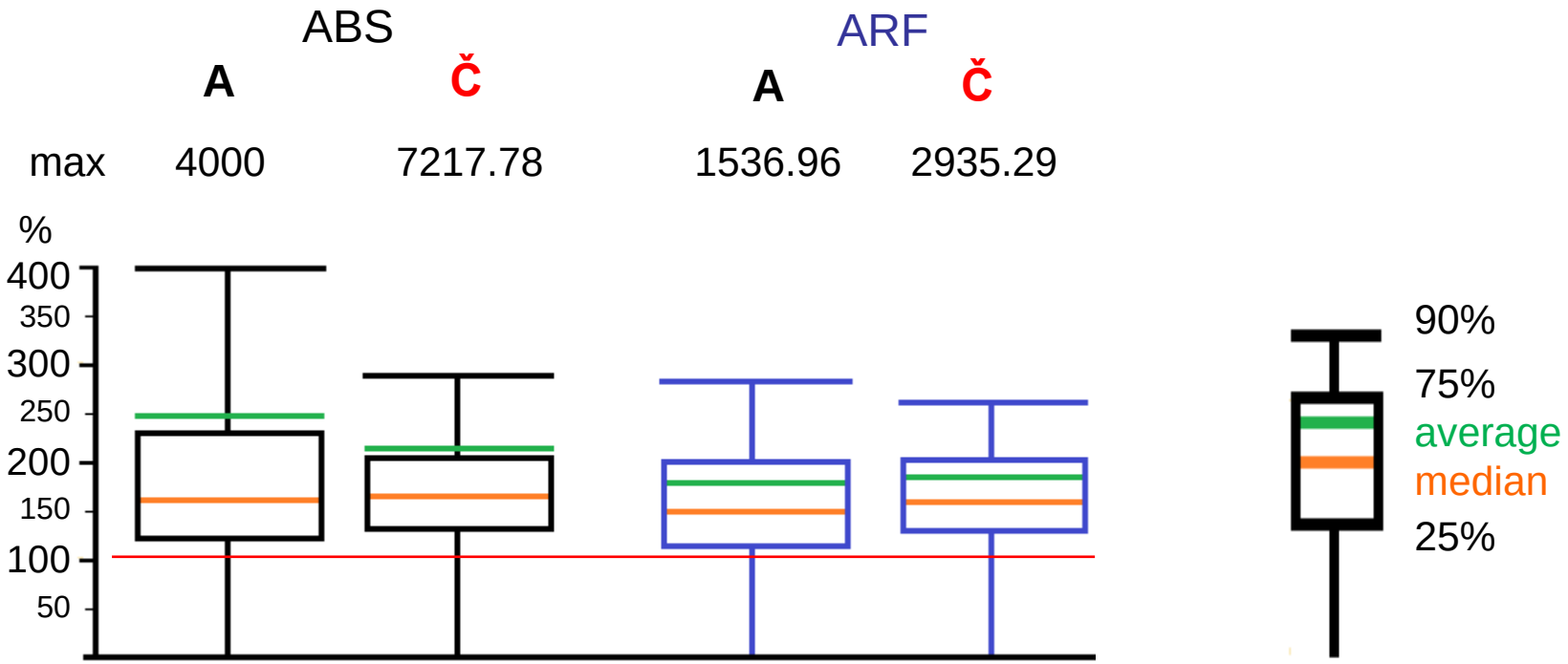
2. Some interesting words:

*absurdistán, açai, airsoft, ajvar, akvazorbing, all-inclusive, amaretto, ampír, amputář…*

*čajznout, čárkovat, černoprdelník, červivět, čikuli, čimčarovat, čórka, čůčo, čuránky…*

# Newer corpora: II. Including words in the dictionary

SYN v6 (4 bil.) instead of SYN v3 (2.2 bil.)?
If the frequency in SYN v3 = 100%, in SYN v6 it is…




more words to include: *aktivovatelný, akvírovat, alginát, čikuli, čuňácký, čurda*… (not plenty of them)

# Preliminary suggestions: Making of the raw headword list

1. Take SYN2000 + SYN2005 + SYN2010
2. Set absolute frequency ≥ 5 or ARF ≥ 2.5
3. Check this list carefully

4. Check briefly the following, search for word families absent in ↑:
   a) words with ABS < 5 and ARF ≥ 2.5
      (or ARF < 2.5 and ABS ≥ 5)
   b) words reaching the threshold ABS/ARF only when adding SYN2015 (SYN2020, 2025?)

## Including words in the dictionary

0. *All other criteria (number of sources, rules for terms, rules for prefixoid series…) must be satisfied.*
1. If the frequency in SYN v3 is too low, check SYN v6 (v7, v8…?)
2. Threshold frequency – **an open question**

Always use your lexicographical experience + sense.
*A computer can't make a good headword list*
*– it is not just the frequency that matters.*

**Thank you for your attention**