

# ***Značkování kategorie slovního druhu v korpusech ČNK***

---

## ***Pozice 1 - Slovní druh***

Označuje hlavní slovní druh, víceméně podle obvyklého schématu známého z českých gramatik včetně školních. Přiřazení i těchto hlavních slovních druhů je však řízeno především potřebami konzistentnosti další analýzy přirozeného jazyka. Proto je možné, že v některých případech (zejména tehdy, kdy se gramatiky a slovníky v určení slovního druhu neshodují nebo uvádějí jiné rozdělení na významy slova nebo tam, kde ve slovníku najdeme slovnědruhové perly typu "zájmenné příslovce") nemusí být zařazení zcela "tradiční".

<b>N</b>	substantivum
<b>A</b>	adjektivum
<b>P</b>	pronomen
<b>C</b>	numerál
<b>V</b>	verbum
<b>D</b>	adverbium
<b>R</b>	prepozice
<b>J</b>	konjunkce
<b>T</b>	partikule
<b>I</b>	interjekce
<b>X</b>	neznámý, neurčený, neurčitelný slovní druh
<b>Z</b>	interpunkce, hranice věty

## ***Disambiguace slovnědruhově víceznačných jednotek v korpusech ČNK***

---

Rozdíl vidět v první pozici, která určuje SD.

### ***Pila***

- **Podstatné jméno** NNFS(P)..... (Na kameni ležela pila.)
- **Sloveso** VpFS..... (Celé dny jenom jedla a pila.)

### ***Lež***

- **Podstatné jméno** NNFS.... (Lež má krátké nohy.)
- **Sloveso** Vi-S.... (Nic nedělej, jenom lež!)

### ***Kolem***

- **Podstatné jméno:** NNNNS7 (Povšiml si nějaké ženy s kolem.)
- **Předložka:** RR—2 (Jel kolem ní)
- **Přísluvce:** Db----- (Šel kolem.)

### ***Hnát***

- **Podstatné jméno:** NNFS4....
- **Sloveso:** VB-S..... (Co zachráníš, když se budeš hnát jako šílená?)
- **ŠPATNĚ URČENO V KORPUSU:** Než se prostě nechat hnát proudem (určeno jako N); Jsi připraven mě hnát? (určeno jako N)

### ***Cestující***

- **Podstatné jméno:** NNFS4 (Vstup do metra vchodem pro cestující)
- **Přídavné jméno:** AGFS4 (Najít spolehlivou společnost cestující na jih)

Podobně slova **večer** (1,6), **co** (3,8,10), **bokem, časem, koukej** (5,10 – Koukej nezlobit), **ba**

## Tokenizace a značkování víceslovných jednotek v korpusech

Současný směr v korpusovce: zvýšená pozornost věnovaná syntagmatice (víceslovným jednotkám), vyvinuta nová metodologie a statistické postupy -> kolokační míry MI-score, T-score, které identifikují na základě frekvence slov ty dvojice, jejichž spoluvýskyt je statisticky významově častější, než bychom mohli očekávat na základě pouhé pravděpodobnosti.

*Tokenizace*: rozdělení na slova, rozčlenění textu na základní jednotky určené pro vyhledávání.

### Mi-score - vzájemná informace (mutual information)

Vychází z teorie informace, kde je pro jevy x a y definována takto:

$$I(x, y) = \log_2 \frac{p(x, y)}{p(x) \cdot p(y)}$$

kde  $P(x)$  je pravděpodobnost jevu x,  $P(y)$  pravděpodobnost jevu y a  $P(x, y)$  je pravděpodobnost, že jevy x a y nastanou současně. V našem případě rozumíme  $P(x)$  pravděpodobnost výskytu hledaného slova x, podobně  $P(y)$  pravděpodobnost výskytu slova y a  $P(x, y)$  pravděpodobnost výskytu slova y v kontextu slova x. Jednotlivé pravděpodobnosti můžeme tedy vyjádřit takto:

$$P(x) = \frac{f(x)}{N}$$

$$P(y) = \frac{f(y)}{N}$$

$$P(x, y) = \frac{f(x, y)}{N}$$

kde  $N$  je velikost korpusu (počet slov).

Po dosazení a úpravě dostaneme:

$$mi(x, y) = \log_2 \frac{N \cdot f(x, y)}{f(x) \cdot f(y)}$$

Nevýhodou vlastností mi-score je to, že je velmi ovlivňováno frekvencí jednotlivých slov. Nejvyšších hodnot totiž dosahují dvojice slov s nízkou frekvencí. Z tohoto důvodu umožňuje Bonito při výpočtu mi-score nastavit spodní hranici frekvence a pro slova s absolutní frekvencí pod touto hranicí se potom mi-score nepočítá.

### T-score - míra kontrastu

Vychází ze statistické metody testování hypotéz pomocí tzv. t-testu.

V případě kolokací testujeme, zda zjištěné počty výskytů jednotlivých slov a jejich dvojic odpovídají náhodnému rozložení slov v korpusu. Čím vyšší je hodnota t-score, tím méně je pravděpodobné, že jde o náhodné rozložení slov a naopak čím pravděpodobnější je, že jde o pevnější, ustálenější kombinace slov, tj. o kolokace.

Statistický vzorec pro náhodnou veličinu adaptujeme na rozložení slov v korpusu a jeho zjednodušením dostáváme pro výpočet t-score vztah:

$$T = \frac{\left( f(x, y) - \frac{f(x) \cdot f(y)}{N} \right)}{\sqrt{f(x, y)}}$$

## ***Disambiguace homonym lišících se kategorií rodu a životností***

---

### ***Pozice 3 - Jmenný rod***

- neurčuje se

**F** femininum (ženský rod)

**H** femininum nebo neutrum (tedy nikoli maskulinum)\*

**I** maskulinum inanimatum (rod mužský neživotný)

**M** maskulinum animatum (rod mužský životný)

**N** neutrum (střední rod)

**Q** femininum singuláru nebo neutrum plurálu (pouze u přídělných a jmenných adjektiv)\*

**T** masculinum inanimatum nebo femininum (jen plurál u přídělných a jmenných adjektiv)\*

**X** libovolný rod (F/M/I/N)

**Y** masculinum (animatum nebo inanimatum)\*

**Z** 'nikoli femininum' (tj. M/I/N; především u příslovcí)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

Homonyma dělíme na úplná, tj. taková, která přebírají všechny tvary slova (např. raketa) a na částečná, u nichž se například liší kategorie životnosti (kohoutek, rys, los, atlas = kniha, látka)

### **Rozlišuje se pouze u mužského rodu – životnost (M) a neživotnost (I)**

- Kohoutek
  - zvíře: NNM....
  - vodovodní: NNI....
- Rys
  - zvíře: NNM
  - ve tváři: NNI
- Los
  - zvíře: NNMS1
  - výherní: NNIS4

# Gramatická kategorie čísla v morfologickém značkování českých korpusů z hlediska desambiguace

---

## Pozice 4 - Číslo

- neurčuje se

**D** duál (pouze 7. pád feminin)

**P** plurál (množné číslo)

**S** singulár (jednotné číslo)

**W** pouze v kombinaci s jmenným rodem 'Q' (singulár pro feminina, plurál pro neutra)\*

**X** libovolné číslo (P/S/D)

\* Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

## SINGULARIA TANTUM

- **jména hromadná/kolektiva** – označují soubory počítatelných jednotlivin, větší množství předmětů chápaných jako celek, formanty: *-í (listí, kamení), -oví (olšovi), -stvo (panstvo), -ež (mládež), -ivo (topivo, učivo)* i pojmový obsah (*lid, hmyz*), spojují se s lexikálními prostředky vyjadřujícími množství a míru a s číslovkami druhovými, ne však s číslovkami základními (*plno mládeže, mnoho kamení, dvojí kamení, mnohé panstvo*)
  - v korpusech jak S tak P
- **jména látková** – označují materiál (*dřevo, měď, železo*), můžeme je najít i v plurálu, pak označují více druhů, s výrazem množství odměření určitého kvanta látky (*jedno pivo, dvě minerálky*)
  - v korpusech jak S tak P
- **jména abstraktní** – *chut', hmat, čich*, odvozeniny od adjektiv a sloves (*chytrost, chudoba, plavání*)
  - v korpusech jak S tak P
- **jména vlastní/propria** – označují jedinečný předmět, jev (*Říp, Praha*), vlastní jména osobní mít plurál mohou (*bratři Novákoví*)
  - v korpusech jak S tak P

## PLURALIA TANTUM

- **názvy „podvojných“ předmětů** – předměty složené z dvou částí (*nůžky, kleště, kalhoty, tepláky*)
  - v korpusech jak S tak P
- **názvy složitějších nástrojů** složených z více částí (*hodinky, housle, kamna*)
  - v korpusech jak S tak P
- **názvy časových období**, svátků složených z několika dní (*Hromnice, prázdniny, Vánoce*)
  - v korpusech jak S tak P
- **některá vlastní jména místní a pomístní** (*Krkonoš, Tatry, Uhry, Průhonice, Hradčany*)
  - v korpusech jak S tak P

## Značkování jmenných tvarů v korpusu

---

Hotov(i): ACYS---- (Jestli je s tím hotov)

**Funkční omezení:**

- v přísudku: byl (VpYS--) mlád (ACYS)
- v doplňku: vrátil (VpYS) se (P7-X4) zdrav (ACYS)

**Adjektivum rád:** jmenné tvary, v doplňku ACYS (Člověk si rád poslechne...)

**Maskulina:** Bos: NNMS1 (Byl bez košile a bos); ACYS (Ušel jste bos pěkný kus)

**Zachovali se ve frazéměch** – bylo jak nabíledni (Db-----), pln ACYS (pln života); **a příslovečných spřezkách** (doleva – Db----)

Tvary středního rodu: zvláštní druh příslovcí – predikativa

- smutno (Je mi pořád smutno; Db---); Aspoň mi tam nebude smutno (ACNS---)
- veselo (Dg--- a ACNS a taky NNNS1 – večer bývá veselo)
- možno (ACNS)
- nutno (ACNS)

## ***Druhy zájmen z hlediska morfologického značkování korpusů ČNK***

---

### ***Pozice 2 - Detailní určení slovního druhu***

- 0** předložka s připojeným "-ň" (něj), "proň", "naň", atd. (značkováno jako slovní druh: zájmeno - 'P')
- 1** vztažné přivlastňovací zájmeno "jehož", "jejíž", ...
- 4** vztažné nebo tázací zájmeno s adjektivním skloňováním (obou typů: "jaký", "který", "čí", ...)
- 5** zájmeno "on" ve tvarech po předložce (tj. "n-": "něj", "něho", ...)
- 6** reflexivní zájmeno "se" v dlouhých tvarech ("sebe", "sobě", "sebou")
- 7** reflexivní zájmeno "se", "si" pouze v těchto tvarech, a dále "ses", "sis"
- 8** přivlastňovací zájmeno "svůj"
- 9** vztažné zájmeno "jenž", "již", ... po předložce ("n-": "něhož", "níž", ...)
- D** zájmeno ukazovací ("ten", "onen", ...)
- E** vztažné zájmeno "což"
- H** krátké tvary osobních zájmen ("mě", "mi", "ti", "mu", ...)
- J** vztažné zájmeno "jenž" ("již", ...), bez předložky
- K** zájmeno tázací nebo vztažné "kdo", vč. tvarů s "-ž" a "-s"
- L** zájmeno neurčité "všechn", "sám"
- O** samostatně stojící zájmena "svůj", "nesvůj", "tentam"
- P** osobní zájmena (vč. tvaru "tys")
- Q** zájmeno tázací/vztažné "co", "copak", "cožpak"
- S** zájmeno přivlastňovací "můj", "tvůj", "jeho" (vč. plurálu)
- W** zájmena záporná ("nic", "nikdo", "nijaký", "žádný", ...)
- Y** zájmeno "co" spojené s předložkou ("oč", "nač", "zač")
- Z** zájmeno neurčité ("nějaký", "některý", "číkoli", "cosi", ...)

## ***Druhy číslovek z hlediska morfologického značkování korpusů ČNK***

---

### ***Pozice 2 - Detailní určení slovního druhu***

- =** číslo psané číslicemi (značkováno jako slovní druh: číslovka - 'C')
- ?** číslovka "kolik"
- }** číslovka psaná římskými číslicemi
- 3** zkratka jako číslovka
- a** číslovka neurčitá ("mnoho", "málo", "tolik", "několik", "kdovíkolik", ...)
- d** číslovka druhová, adjektivní skloňování ("jedny", "dvoji", "desaterý", ...)
- h** číslovky druhové "jedny" a "nejedny"
- j** číslovka druhová  $\geq 4$ , substantivní postavení ("čtvero", "desatero", ...)
- k** číslovka druhová  $\geq 4$ , adjektivní postavení, krátký tvar ("čtvery", ...)
- l** číslovky základní 1-4, "půl", ...; sto a tisíc v nesubstantivním skloňování
- n** číslovky základní  $\geq 5$
- o** číslovky násobné neurčité ("-krát": "mnohokrát", "tolikrát", ...)
- r** číslovky řadové
- u** číslovka tázací násobná "kolikrát"
- v** číslovky násobné ("-krát": "pětkrát", "poprvé" ...)

- w** číslovky neurčité s adjektivním skloňováním ("nejeden", "tolikátý", "několikátý" ...)  
**y** zlomky zakončené na "-ina" (značkováno jako slovní druh: číslovka - 'C')  
**z** číslovka tázací řadová "kolikátý"

## ***značkování gramatické kategorie osoby slovesných tvarů v korpusech ČNK***

---

### ***Pozice 8 - Osoba***

- neurčuje se
- 1** 1. osoba
- 2** 2. osoba
- 3** 3. osoba
- X** libovolná osoba (1/2/3)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

### ***Pozice 4 - Číslo***

- neurčuje se
- D** duál (pouze 7. pád feminin)
- P** plurál (množné číslo)
- S** singulár (jednotné číslo)
- W** pouze v kombinaci s jmenným rodem 'Q' (singulár pro feminina, plurál pro neutra)\*
- X** libovolné číslo (P/S/D)

\* Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

## ***vid a čas, značkování v korpusech ČNK***

---

### ***Pozice 16 - Vid***

Tato pozice byla k původní sadě doplněna Miroslavem Spoustou na základě slovníku morfologické analýzy. Tato pozice není k dispozici v korpusech [SYN2000](#) a [ORWELL](#).

- P** perfektivum (dokonavé sloveso)
- I** imperfektivum (nedokonavé sloveso)
- B** obouvidé sloveso

### ***Pozice 9 - Čas***

- neurčuje se
- F** futurum (budoucí čas)
- H** minulost nebo přítomnost (P/R)\*
- P** prézens (přítomný čas)
- R** minulý čas
- X** libovolný čas (F/R/P)\*

\* Tato značka je k dispozici pouze v korpusech: SYN2006PUB, SYN2005, SYN2000, ORWELL.

## ***Sovesný rod, způsob, značkování v korpusech ČNK***

---

### ***Pozice 12 - Aktivum/pasívum***

- neurčuje se

**A** aktivum nebo 'nikoli pasívum'

**P** Pasívum

### ***Pozice 2***

**c** kondicionál slovesa být ("by", "bych", "bys", "bychom", "byste")

**i** slovesný tvar rozkazovacího způsobu

## ***Disambiguace neohebných slovních druhů v případě slovnědruhově homonymie***

---

### **PREPOZICE:**

Díky:

- podstatné jméno: S díky odmítl (NNIP7)
- předložka: Díky své neinformovanosti (R--3)

### **KONJUNKCE:**

Co:

- zájmeno
- spojka (Ten co (J---) nedával pozor)

## ***Morfologické varianty a dublety z hlediska jejich distribuce a kodifikace, značkování variant a dublet a jejich desambiguace***

---

### ***Pozice 15 - Varianta, stylový příznak apod.***

- neurčuje se ("základní" tvar pro kategorie v pozicích 1-14)

**1** varianta, víceméně rovnocenná ("méně častá")

**2** řídka, archaická nebo knižní varianta

**3** velmi archaický tvar, též hovorový

**4** velmi archaický nebo knižní tvar, pouze spisovný (ve své době)

**5** hovorový tvar, ale v zásadě tolerovaný ve veřejných projevech

**6** hovorový tvar (koncovka standardní obecné češtiny)

**7** hovorový tvar (koncovka standardní obecné češtiny), varianta k '6'

**8** Zkratky

**9** speciální použití (tvary zájmen po předložkách apod.)

Píšu: neurčeno; Píši: VB-S---1P-AA—1I

Tisknul: neurčeno; tiskl: 1

### **Dubletní tvary u substantiv:**

z popelu: neurčeno; z popela: 1; z popele: 2

3.p: loktu - neurčeno a lokti: 1

## DETAILNĚ K POZICI 2, KTERÁ PODROBNĚ URČUJE SLOVNÍ DRUH

- ! zkratka jako adverbium
- \* slovo "krát" (slovní druh: spojka)
- , spojka podřadící (vč. "aby" a "kdyby" ve všech tvarech)
- . zkratka jako adjektivum
- : interpunkce všeobecně
- ; zkratka jako substantivum
- = číslo psané číslicemi (značkováno jako slovní druh: číslovka - 'C')
- ? číslovka "kolik"
- ^ spojka souřadící
- } číslovka psaná římskými číslicemi
- ~ zkratka jako sloveso
- @ slovní tvar, který nebyl morfologickou analýzou rozpoznán (značkováno jako slovní druh: neznámý - 'X')
- 0 předložka s připojeným "-ň" (něj), "proň", "naň", atd. (značkováno jako slovní druh: zájmeno - 'P')
- 1 vztažné přivlastňovací zájmeno "jehož", "jejíž", ...
- 2 slovo před pomlčkou
- 3 zkratka jako číslovka
- 4 vztažné nebo tázací zájmeno s adjektivním skloňováním (obou typů: "jaký", "který", "čí", ...)
- 5 zájmeno "on" ve tvarech po předložce (tj. "n-": "něj", "něho", ...)
- 6 reflexivní zájmeno "se" v dlouhých tvarech ("sebe", "sobě", "sebou")
- 7 reflexivní zájmeno "se", "si" pouze v těchto tvarech, a dále "ses", "sis"
- 8 přivlastňovací zájmeno "svůj"
- 9 vztažné zájmeno "jenž", "již", ... po předložce ("n-": "něhož", "níž", ...)
- A adjektivum obyčejné
- B sloveso, tvar přítomného nebo budoucího času
- C adjektivum, jmenný tvar
- D zájmeno ukazovací ("ten", "onen", ...)
- E vztažné zájmeno "což"
- F součást předložky, která nikdy nestojí samostatně ("nehledě", "vzhledem", ...)
- G přídavné jméno odvozené od slovesného tvaru přítomného přechodníku
- H krátké tvary osobních zájmen ("mě", "mi", "ti", "mu", ...)
- I citoslovce (značkováno jako slovní druh: citoslovce - 'I')
- J vztažné zájmeno "jenž" ("již", ...), bez předložky
- K zájmeno tázací nebo vztažné "kdo", vč. tvarů s "-ž" a "-s"
- L zájmeno neurčité "všechn", "sám"
- M přídavné jméno odvozené od slovesného tvaru minulého přechodníku
- N substantivum, obyčejné
- O samostatně stojící zájmena "svůj", "nesvůj", "tentam"
- P osobní zájmena (vč. tvaru "tys")
- Q zájmeno tázací/vztažné "co", "copak", "cožpak"
- R předložka, obyčejná
- S zájmeno přivlastňovací "můj", "tvůj", "jeho" (vč. plurálu)
- T částice (slovní druh 'T')
- U adjektivum přivlastňovací (na "-ův" i "-in")
- V předložka vokalizovaná ("ve", "pode", "ku", ...)
- W zájmena záporná ("nic", "nikdo", "nijaký", "žádný", ...)
- X slovní tvar, který byl rozpoznán, ale značka (ve slovníku) chybí
- Y zájmeno "co" spojené s předložkou ("oč", "nač", "zač")
- Z zájmeno neurčité ("nějaký", "některý", "číkoli", "cosi", ...)
- a číslovka neurčitá ("mnoho", "málo", "tolik", "několik", "kdovíkolik", ...)
- b příslovce (bez určení stupně a negace; "pozadu", "naplocho", ...)
- c kondicionál slovesa být ("by", "bych", "bys", "bychom", "byste")

<b>d</b>	číslovka druhová, adjektivní skloňování ("jedny", "dvoji", "desaterý", ...)
<b>e</b>	slovesný tvar přechodníku přítomného ("-e", "-íc", "-íce")
<b>f</b>	slovesný tvar: infinitiv
<b>g</b>	příslovce (s určením stupně a negace; "velký", "zajímavý", ...)
<b>h</b>	číslovky druhové "jedny" a "nejedny"
<b>i</b>	slovesný tvar rozkazovacího způsobu
<b>j</b>	číslovka druhová $\geq 4$ , substantivní postavení ("čtvero", "desatero", ...)
<b>k</b>	číslovka druhová $\geq 4$ , adjektivní postavení, krátký tvar ("čtvery", ...)
<b>l</b>	číslovky základní 1-4, "půl", ...; sto a tisíc v nesubstantivním skloňování
<b>m</b>	slovesný tvar přechodníku minulého, příp. (zastarale) přechodník přítomný dokonavý
<b>n</b>	číslovky základní $\geq 5$
<b>o</b>	číslovky násobné neurčité ("-krát": "mnohokrát", "tolikrát", ...)
<b>p</b>	slovesné tvary minulého aktivního přičestí (včetně přidaného "-s")
<b>q</b>	archaické slovesné tvary minulého aktivního přičestí (zakončení "-t")
<b>r</b>	číslovky řadové
<b>s</b>	slovesné tvary pasivního přičestí (vč. přidaného "-s")
<b>t</b>	archaické slovesné tvary přítomného a budoucího času (zakončení "-t")
<b>u</b>	číslovka tázací násobná "kolikrát"
<b>v</b>	číslovky násobné ("-krát": "pětkrát", "poprvé" ...)
<b>w</b>	číslovky neurčité s adjektivním skloňováním ("nejeden", "tolikátý", "několikátý" ...)
<b>x</b>	zkratka, slovní druh neurčen/neznámý
<b>y</b>	zlomky zakončené na "-ina" (značkováno jako slovní druh: číslovka - 'C')
<b>z</b>	číslovka tázací řadová "kolikátý"