Popis morfologických značek – poziční systém

Jan Hajič

Ústav formální a aplikované lingvistiky MFF UK

Morfologická analýza a syntéza

Morfologické značky jsou součástí výsledku (výstupem) morfologické analýzy, která pracuje s izolovanými slovními tvary, tedy bez ohledu na jejich kontext. Druhou částí výsledku je tzv. **lemma**, které identifikuje příslušnou abstraktní lexikální jednotku, někdy i včetně jejího významu, ve smyslu jednoznačné identifikace slovníkového hesla. V opačném směru, tj. pro syntézu slovních tvarů, je značka spolu s lemmatem vstupem pro proceduru tvorby (opět izolovaného) slovního tvaru. Morfologická analýza je obecně nejednoznačná; slovní tvary, brány izolovaně a bez ohledu na kontext, pochopitelně nemohou být v mnoha případech jednoznačně určeny, a to jak z hlediska lemmatu, tak z hlediska morfologické značky.

Struktura značky

Každá značka je řetězcem 15 znaků (16. pozice je dostupná pouze v některých korpusech). Značka je konstruována tak, aby každá pozice odpovídala jedné morfologické kategorii podle víceméně tradičního lingvistického pojetí. Každé hodnotě v dané kategorii odpovídá jeden znak, převážně písmeno velké abecedy (např. 'P' pro plurál, neboli množné číslo), výjimečně i jiný znak (např. 'f' pro infinitiv, nebo ',' pro podřadicí spojky). Hodnota, která nedává smysl (např. pád u sloves), je reprezentována znakem '-' (pomlčka).
Tradiční lingvistické detailní rozdělení není ovšem vždy respektováno (z nejrůznějších důvodů). Například tvary minulého příčestí sloves (aktivního i pasívního) nejsou rozlišeny z hlediska rodu (ve spojení s gramatickým číslem) pro tvary končící na *-l*, *-ly* ani *-la*. Podobně zkratky a nesklonná substantiva nedávají na výstupu morfologické analýzy 14 značek, jak by bylo možno očekávat, ale jsou anotovány (v technickém smyslu) jednoznačně značkou, kde je pro číslo a pád uveden znak 'X', používaný převážně pro tento typ nejednoznačnosti (či spíše neurčitosti).

Popis jednotlivých pozic značky

Pozice jsou v závislosti na konkrétním korpusu číslovány od 1 do 15 (16. pozice dostupná pouze v některých korpusech). V nadpisech jsou na konci v závorce uvedeny zkratky pro jednotlivé pozice, používané v jiných programech (jen pro informaci).

Pozice 1 - Slovní druh (POS)

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".

A - adjektivum (přídavné jméno)
C - numerál (číslovka, nebo číselný výraz s číslicemi)
D - adverbium (příslovce)
I - interjekce (citoslovce)
J - konjunkce (spojka)
N - substantivum (podstatné jméno)
P - pronomen (zájmeno)
R - prepozice (předložka)
T - partikule (částice)
V - verbum (sloveso)
X - neznámý, neurčený, neurčitelný slovní druh
Z - interpunkce, hranice věty

Pozice 2 - Detailní určení slovního druhu (SUBPOS)

Detailní slovní druh slouží především k určení dalších relevantních morfologických kategorií, které jsou uvedeny na dalších pozicích (ne vždy však jednoznačně). Ze znaku použitého pro detailní určení slovního druhu je možné jednoznačně vyvodit hlavní slovní druh (pozice 1).

! - zkratka jako adverbium
- hranice věty (jen u "virtuálního" slova "###")
* - slovo "krát" (slovní druh: spojka)
, - spojka podřadicí (vč. "aby" a "kdyby" ve všech tvarech)
. - zkratka jako adjektivum
: - interpunkce všeobecně (ne však "virtuální" slovo ### jako hranice věty)
; - zkratka jako substantivum
= - číslo psané číslicemi (značkováno jako slovní druh: číslovka - 'C')
? - číslovka "kolik"
^ - spojka souřadicí
} - 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 - reflexívní zájmeno "se" v dlouhých tvarech ("sebe", "sobě", "sebou")
7 - reflexívní 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 "všechen", "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")
d - číslovka druhová, adjektivní skloňování ("jedny", "dvojí", "desaterý", ...)
e - slovesný tvar přechodníku přítomného ("-e", "-íc", "-íce")
f - slovesný tvar: infinitiv
g - příslovce (s určením stupně a negace; "velký", "zajímavý", ...)
h - číslovky druhové "jedny" a "nejedny"
i - slovesný tvar rozkazovacího způsobu
j - číslovka druhová >= 4, substantivní postavení ("čtvero", "desatero", ...)
k - číslovka druhová >= 4, adjektivní postavení, krátký tvar ("čtvery", ...)
l - číslovky základní 1-4, "půl", ...; sto a tisíc v nesubstantivním skloňování
m - slovesný tvar přechodníku minulého, příp. (zastarale) přechodník přítomný dokonavý
n - číslovky základní >= 5
o - číslovky násobné neurčité ("-krát": "mnohokrát", "tolikrát", ...)
p - slovesné tvary minulého aktivního příčestí (včetně přidaného "-s")
q - archaické slovesné tvary minulého aktivního příčestí (zakončení "-ť")
r - číslovky řadové
s - slovesné tvary pasívního příčestí (vč. přidaného "-s")
t - archaické slovesné tvary přítomného a budoucího času (zakončení "-ť")
u - číslovka tázací násobná "kolikrát"
v - číslovky násobné ("-krát": "pětkrát", "poprvé" ...)
w - číslovky neurčité s adjektivním skloňováním ("nejeden", "tolikátý", "několikátý" ...)

x - zkratka, slovní druh neurčen/neznámý
y - zlomky zakončené na "-ina" (značkováno jako slovní druh: číslovka - 'C')
z - číslovka tázací řadová "kolikátý"

Pozice 3 - Jmenný rod (GENDER)

- - 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 příčestí a jmenných adjektiv)
T - masculinum inanimatum nebo femininum (jen plurál u příčestí 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)

Pozice 4 - Číslo (NUMBER)

- - 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)

Pozice 5 - Pád (CASE)

- - neurčuje se
1 - nominativ (1. pád)
2 - genitiv (2. pád)
3 - dativ (3. pád)
4 - akuzativ (4. pád)
5 - vokativ (5. pád)
6 - lokativ (6. pád)
7 - instrumentál (7. pád)
X - libovolný pád (1/2/3/4/5/6/7)

Pozice 6 - Přivlastňovací rod (POSSGENDER)

Rody mužský neživotný a střední se nikdy nevyskytují samostatně. 'M'
se může vyskytnout jen u přivlastňovacích adjektiv (ne u příslovcí).

- - neurčuje se
F - femininum (ženský rod)
M - maskulinum animatum (rod mužský životný)
X - libovolný rod (F/M/I/N)
Z - 'nikoli femininum' (tj. M/I/N; u přivlastňovacích adjektiv)

Pozice 7 - Přivlastňovací číslo (POSSNUMBER)

- - neurčuje se
P - plurál (množné číslo)
S - singulár (jednotné číslo)

Pozice 8 - Osoba (PERSON)

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

Pozice 9 - Čas (TENSE)

- - 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)

Pozice 10 - Stupeň (GRADE)

- - neurčuje se
1 - 1. stupeň
2 - 2. stupeň
3 - 3. stupeň

Pozice 11 - Negace (NEGATION)

- - neurčuje se
A - afirmativ (bez negativní předpony "ne-")
N - negace (tvar s negativní předponou "ne-")

Pozice 12 - Aktivum/pasívum (VOICE)

- - neurčuje se
A - aktivum nebo 'nikoli pasívum'
P - pasívum

Pozice 13 - Nepoužito (RESERVE1)

- - neurčuje se

Pozice 14 - Nepoužito (RESERVE2)

- - neurčuje se

Pozice 15 - Varianta, stylový příznak apod. (VAR)

- - neurčuje se ("základní" tvar pro kategorie v pozicích 1-14)
1 - varianta, víceméně rovnocenná ("méně častá")
2 - řídká, 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.)

Pozice 16 - Vid (ASPECT)

Tato pozice byla k původní sadě doplněna Miroslavem Spoustou na základě slovníku morfologické analýzy. Je dostupná pouze v korpusech SYN2005 a SYN2006PUB.

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