

Lemmatizace a morfologické značkování v korpusech soukromé korespondence

Na CD 1 jsou tři korpusey soukromé korespondence: **KSK-dopisy**, **KSK-e-maily**, **KSK-dopisy1**. Všechny tři byly lemmatizovány automatickým morfologickým analyzátozem **AJKA** <http://nlp.fi.muni.cz/projekty/ajka/> (počítačový program přiřadil každému izolovanému textovému slovu lemma/lemmata, tj. základní slovníkový tvar/slovníkové tvary). Automatický morfologický analyzátor AJKA byl původně vyvinut pro automatickou morfologickou analýzu a využíván pro automatickou lemmatizaci a značkování (tagování, tj. gramatickou morfologickou anotaci) korpusey spisovných textů. Pro korpusey s vysokou frekvencí substandardních tvarů (k nimž patří také korpusey soukromé korespondence) byla vytvořena modifikovaná verze, která automaticky rozpoznává i velké množství nespisovných tvarů.

Korpusey **KSK-dopisy** a **KSK-e-maily** jsou automaticky lemmatizovány modifikovanou verzí analyzátoru AJKA. Tvarům, jež AJKA nerozpoznala, byl místo lemmatu automaticky přiřazen tvar sám. Lemmatizace je tedy zatím pouze neúplná. Verze korpusey **KSK-dopisy** a **KSK-e-maily** uvedená na CD 1 není morfologicky označkována.

Korpus **KSK-dopisy1** zahrnuje 1 000 dopisů identických s dopisy první poloviny **KSK-dopisy**. Je automaticky lemmatizován a **morfologicky označován**, ručně disambiguován (95 %) a částečně ručně doznačován (1,6 % ze zbývajících 5 %).

Tokenizace

Všechny tři korpusey jsou automaticky tokenizovány, tj. rozděleny na jednotlivé pozice. Pozicí se rozumí samostatný řetězec znaků, s nímž pak pracují další automatické programy. Samostatné pozice jsou: textová slova (řetězce abecedních znaků mezi mezerami), ciferné výrazy, interpunkční znaky (nerozlišuje se tečka za větou a tečka, která je součástí zkratky, takže například zkratka „str.“ je rozdělena na dvě pozice a samostatně je označován řetězec „str“ a samostatně „.“) atd.

Základní vyhledávání v korpusu KSK-dopisy1

Vyhledávání podle lemmatu

(Vyhledávání podle lemmatu je v zásadě možné i v **KSK-dopisy** a v **KSK-e-maily**. Je však třeba počítat s tím, že některé substandardní formy vyhledávací program podle lemmatu nenajde; viz výše.)

Chceme-li vyhledat slovo nebo slovní spojení ve všech jeho jednoduchých tvarech, hledáme podle tzv. **lemmatu** (základního slovníkového tvaru). U slovesa je tímto tvarem infinitiv (jednoduchého slovesného tvaru, takže například ve větě „*Už jsem se nechala vyfotit ...*“ budou tři samostatné slovesné tvary, jimž budou odpovídat tři lemmata: tvaru *jsem* bude přiřazeno lemma *být*, tvaru *nechala* lemma *nechat* a tvaru *vyfotit* lemma *vyfotit*). U

podstatných jmen a bezrodých zájmen je lemmatem tvar 1. pádu, u přídavných jmen, adjektivně skloňovaných zájmen a číslovek je lemmatem tvar 1. pádu mužského rodu jednotného čísla. U neohebných slovních druhů je lemmatem tvar sám. Výjimkou jsou některé **substandardní varianty** (viz níže), které mají **lemma podle standardní varianty**.

Jde o tyto případy:

1) Varianty se substandardní koncovkou

Například: tvar *klukama* má lemma *kluk*, tvar *ject* má lemma *jet*, tvar *žijó* má lemma *žit*, tvar *bavěj* má lemma *bavit*, tvar *dobrej* má lemma *dobrý*, tvar *ktorejma* má lemma *který*.

2) Varianty se substandardní kmenotvornou příponou (u sloves)

Například: tvar *myslím* má lemma *myslet* nebo tvar *vidím* má lemma *vidět*.

3) Substandardní tvary zájmen

Například: tvar *ja* má lemma *já*, tvar *nama* má lemma *my*, tvar *teho* má lemma *ten*.

4) Substandardní tvary slovesa *být*

Například: tvary *su, seš, sou, sem, sme, bejt, ...* má lemma *být*.

5) Substandardní tvary kondicionálů *by, ... s* variantami *aby, ..., kdyby, ...*

Například: tvar *bysme* má lemma *by*

6) Tvary s protetickým *v-*

Například: tvar *vobšt'astňovat* má lemma *obšt'astňovat*, předložka *vod* má lemma *od*.

7) Varianty se substandardními pravopisnými jevy (chybami)

Například: tvar tvar *být* (*A nepiš už taťkovi že mě nemá být po hlavě a že budu blbej ...*) má lemma *bít*.

Lemmatem slovtvorně substandardních tvarů je pravidelně vytvořený tvar nominativu nebo infinitivu.

Například: tvar *dopendluju* má lemma *dopendlovat* nebo tvar *foťáku* má lemma *foťák*, tvar *kámoškou* má lemma *kámoška*, tvar *strejdou* má lemma *strejda*, tvar *bráchem* má lemma *brácha*.

Výsledkem hledání jsou všechny konkordance obsahující tvary hledaného slova (s malým i velkým počátečním písmenem). Dotaz zapíšeme do dotazového řádku v následujícím tvaru: **[lemma="být"]**.

Příklad: Do dotazového řádku napíšeme: **[lemma="být"]** a stiskneme *Enter*.

```
(...)  
došel . Doufám , že už < jste> zdraví . Já už celkem jo , ještě trochu  
rýmy , ale snad už to < bude> dobrý . Ale měla bych to někde zaťukat  
Samozřejmě , že < je> to blbost , na to se nepotřebuju ptát @  
vím už dávno , mimo to < jsme> to brali i loni v morfologii . Že je to  
neprožívá , protože < su> střízlivá , ale večer se ožeru a to pak  
na Lochotín a vydaly < sme> se pěšky přes sídliště k rodiným domkům  
vidět z obrázku , měly < ste> se přímo skvěle . Mimochodem ,  
" Přesto , že < seš> hroznej , mám Tě ráda " . Nevím , zdali  
nevím , protože ty < si> tam vlastně nechodila ) . S nimi jsme  
(...)
```

Příklad: Hledáme-li slovní spojení, zadáme dotaz:

[lemma="dobrý"] [lemma="den"]

```
(...)  
dopis čteš ráno , tak < dobrý den> a úspěšné , usměvavé odpoledne  
Zrovna dneska byl < dobrej den> . Koupil jsem si za slušnou  
(...)
```

Vyhledávání podle morfologické značky (tagu)

KSK-dopisy1 byl automaticky lemmatizován a označován modifikovanou verzí automatického morfologického analyzátoru **AJKA** <http://nlp.fi.muni.cz/projekty/ajka/>. Poté byl ručně disambiguován a ručně byla doplněna lemmata a značky většině tvarů nerozpoznaných automatickou analýzou. Označováno je 96,6 % tvarů.

V korpusu **KSK-dopisy1** můžeme vyhledávat podle morfologických značek (tagů).

Morfologický **tag** je složen z atributů a hodnot, které atributy aktuálně nabývají pro analyzovaný tvar (**word**). Všechny značky povinně obsahují atribut slovní druh.

Podle jednotlivých slovních druhů se pak ve značce objevují v daném pořadí další atributy:

podstatných jmen (rod, číslo, pád, fakultativní atributy – viz níže),

přídavných jmen (negace, rod, číslo, pád, stupeň, fakultativní atributy – viz níže),

zájmen (osoba – fakultativně u zájmen, která vyjadřují osobu, rod – fakultativně u zájmen, která vyjadřují rod, číslo, pád, fakultativní atributy – viz níže),

číslovek (rod – fakultativně u číslovek vyjadřujících rod, číslo – fakultativně u základních číslovek *jeden, dva, tři, čtyři* a u adjektivně skloňovaných číslovek, pád, fakultativní atributy – viz níže),

sloves (negace, vid, slovesný tvar, osoba – fakultativně podle slovesného tvaru, pokud ji tvar vyjadřuje, rod - fakultativně podle slovesného tvaru, pokud jej tvar vyjadřuje, číslo – fakultativně podle slovesného tvaru, pokud je tvar vyjadřuje, fakultativní atributy – viz níže),

příslovcí (negace, stupeň, fakultativní atributy – viz níže).

U dalších slovních druhů (**předložek, spojek, částic, citoslovcí**) se uvádějí pouze fakultativní atributy – viz níže.

Tvary *bych, bys, by, bychom, byste, abych, ...*, *kdybybych, ...* mají zvláštní značku, v níž se uvádí atribut slovesný tvar s hodnotou kondicionál, osoba, číslo a fakultativní atributy – viz níže.

Zkratky a interpunkce nemají u atributu slovní druh uveden slovní druh, ale značku, která říká, že jde o zkratku nebo interpunkci.

Fakultativní atributy

S-atribut

U většiny slovních druhů může stát fakultativní atribut signalizující **přítomnost volného morfému „-s“** zastupujícího pomocné sloveso „*být*“ ve tvarech 2. osoby singuláru přítomnosti. Například nejčastěji *ses, sis*, l-ové participium významového slovesa (*mělas mi říct*), tázačí zájmena (*cos mi napsala*), příslovce (*kdes zrovna poletovala*), spojky (*žes počkala*), ... atd.

Tento atribut pracovním nazýváme „-s“ atribut, označujeme jej **z** a nabývá hodnotu **S**.

Příklad: Tvar *muselas* má značku [tag="k5eAaImAgFnSzS"], tvar *žes* má značku [tag="k8zS"].

Atribut „stylistický příznak“

U všech slovních druhů je fakultativně uveden atribut **stylistický příznak (w)**. V tomto atributu jsou zachyceny varianty se substandardními koncovkami, varianty s protetickým *v-*, chybné užití zájmených tvarů (*mě/mně, jí/jí, ...*), některé nekodifikované slovtvorné inovace, pravopisné chyby. Zmíněné "anomálie" jsou signalizovány přítomností atributu **w**, který v těchto případech nabývá hodnoty **H**.

Příklad: Tvar *bráchem* má značku [tag="k1gMnSc7wH"], tvar *ktorej* má podle kontextu např. značku [tag="k3gMnSc1wH"], tvar *vo* má značku [tag="k7wH"], chybně napsaný tvar *jí* v kontextu „*ta jí poprosila*“ má značku [tag="k3p3gFnSc4wH"], chybně napsaný tvar *být* v kontextu „*... být po hlavě ...*“ má značku [tag="k5eAaImFwH"] atd.

Poznámka: Stylistický příznak **wH** nemají ve značce v **KSK-dopisy1** frekventované tvary citosloveného rázu (pozdravy) např. *ahojky, čauky, ahojda, ...* Stylistický příznak **wH** nemá dále řada slovtvorných inovací běžné mluvy. Slovníky spisovného jazyka některé z nich zaznamenávají, takže je automatický analyzátor AJKA byl schopen identifikovat. Jiné (především frekventované) byly doplněny do modifikované verze morfologického analyzátoru AJKA. Tyto tvary sice mají stylistický příznak v poznámce **wH**, ale uživateli se nezobrazuje. Jsme si vědomi inkonsistence tohoto prozatímního řešení. Stylistický příznak **wH** v podobě, která je popsána níže, mají tedy pouze ty substandardní slovtvorné inovace, které byly dodatečně označkovány ručně (*výkoňák, ...*).

Morfologické tagy – systém atribut/hodnota

Morfologická značka (**tag**) má **striktně stanovenou formu**. Je to **posloupnost** příslušných **atributů** a jejich **hodnot**. Pokud chceme vyhledávat pouze podle některého z atributů (například všechny tvary označené jako substantiva - [**tag="k1.*"**], všechny tvary mající značku signalizující množné číslo [**tag=".*nP.*"**], všechny tvary jmen v dativu [**tag=".*c3.*"**]) pak je třeba použít patričným způsobem **regulární výrazy**. Pro potřeby vyhledávání podle značek vystačíme se sekvencí „.*“, kde **tečka** „.“ představuje jeden libovolný znak a **hvězdička** „*“ představuje libovolný počet (0 a více) opakování předchozího znaku nebo výrazu.

Příklad: Dotaz [**tag="k5.*"**] čteme takto: vyhledej všechna slovesa (tvary, které mají hodnotu atributu slovní druh **k** vyplněnou **5**, což znamená sloveso), přičemž další kategorie ve značce jsou libovolné – „.*“ (nahrazené regulárním výrazem pro libovolné opakování 0-n znaků).

Tabulkové přehledy jednotlivých atributů

Atribut slovní druh - k

hodnota		atribut+hodnota
1	substantiva	k1
2	adjektiva	k2
3	zájmena	k3
4	číslovky	k4
5	slovesa	k5
6	příslovce	k6
7	předložky	k7
8	spojky	k8
9	částice	k9
0	citoslovce	k0
A	zkratky	kA
Y	tvary by, bych, bys, bychom, byste, aby, ..., kdyby, ..., (+ substandardní varianty)	kY
Z	interpunkce	kZ

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako předložky, do dotazového řádku napíšeme: [**tag="k7.*"**], kde **k** znamená atribut slovní druh nabývající hodnotu 7, tj. předložka, a stiskneme *Enter*.

Atribut jmenný rod - g

hodnota		atribut+hodnota
---------	--	-----------------

M	maskulinum životné	gM
I	maskulinum neživotné	gI
F	femininum	gF
N	neutrum	gN

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako substantiva rodu ženského, do dotazového řádku napíšeme: **[tag="k1gF.*"]** a stiskneme *Enter*.

Atribut číslo - n

hodnota		atribut+hodnota
S	singulár	nS
P	plurál	nP
D	duál	nD

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako substantiva rodu středního v singuláru, do dotazového řádku napíšeme: **[tag="k1gNnS.*"]** a stiskneme *Enter*.

Atribut pád - c

hodnota		atribut+hodnota
1	nominativ	c1
2	genitiv	c2
3	dativ	c3
4	akuzativ	c4
5	vokativ	c5
6	lokál	c6
7	instrumentál	c7

Příklad: Budeme-li chtít vyhledat všechny tvary substantiv rodu mužského životného v plurálu ve druhém pádě, do dotazového řádku napíšeme: **[tag="k1gMnPc2.*"]** a stiskneme *Enter*.

Atribut negace - e

hodnota		atribut+hodnota
A	adjektiva, slovesa a adverbia bez prefixu <i>ne-</i> signalizujícího negaci	eA
N	adjektiva, slovesa a adverbia s prefixem <i>ne-</i> signalizujícím negaci	eN

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako adjektiva, která nemají prefix *ne-* signalizující negaci, jsou rodu mužského neživotného v plurálu ve třetím pádě, do dotazového řádku napíšeme: **[tag="k2eAgInPc3.*"]** a stiskneme *Enter*.

Atribut stupeň - d

hodnota		atribut+hodnota
1	pozitiv	d1
2	komparativ	d2
3	superlativ	d3

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako adjektiva, která mají prefix *ne-* signalizující negaci, jsou rodu mužského neživotného v plurálu ve druhém pádě a pozitivu (prvním stupni), do dotazového řádku napíšeme: **[tag="k2eAgInPc3d1.*"]** a stiskneme *Enter*.

Atribut slovesný tvar - m

hodnota		atribut+hodnota
F	infinitiv	mF
I	indikativ (jednoduché tvary)	mI
R	imperativ	mR
A	l-ové participium	mA
N	n-/t-ové participium	mN
S	přechodník přítomný	mS
D	přechodník minulý	mD
B	tvary budu, budeš, bude, budeme, budete, budou	mB
C	tvary by, bych, bys, bychom, byste, aby, ..., kdyby, ..., (+ substandardní varianty)	mC

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako infinitiv, do dotazového řádku napíšeme: **[tag="k5.*mF.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako indikativ (jednoduché tvary), do dotazového řádku napíšeme: **[tag="k5.*mI.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako imperativ, do dotazového řádku napíšeme: **[tag="k5.*mR.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako l-ové participium, do dotazového řádku napíšeme: **[tag="k5.*mA.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako n-/t-ové participium, do dotazového řádku napíšeme: **[tag="k5.*mN.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako přechodník přítomný, do dotazového řádku napíšeme: **[tag="k5.*mS.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako přechodník minulý, do dotazového řádku napíšeme: **[tag="k5.*mD.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary *budu, budeš, ...,* do dotazového řádku napíšeme: **[tag="k5.*mB.*"]** a stiskneme *Enter*.

Příklad: Budeme-li chtít vyhledat všechny tvary *by, ..., aby, ..., kdyby, ...,* do dotazového řádku napíšeme: **[tag="kYmC.*"]** a stiskneme *Enter*.

Atribut osoba – p

hodnota		atribut+hodnota
1	první	p1
2	druhá	p2

Příklad: Budeme-li chtít vyhledat všechny tvary označované značkou signalizující gramatický význam osoby, a to 2. osoby, do dotazového řádku napíšeme: **[tag=".*p2.*"]** a stiskneme *Enter*.

Atribut vid – a

hodnota		atribut+hodnota
P	perfektivum	aP
I	imperfektivum	aI
B	obouvidové	aB

Příklad: Budeme-li chtít vyhledat všechny tvary označované jako imperfektivní slovesa, do dotazového řádku napíšeme: **[tag="k5.*aI.*"]** a stiskneme *Enter*.

Atribut „-s“ – z (fakultativní)

hodnota		atribut+hodnota
z	tvar s připojeným nesamostatným morfémem „-s“ signalizujícím 2. osobu (<i>ses, sis, byls, žes, kams, ...</i>)	zS

Příklad: Budeme-li chtít vyhledat všechny tvary označované značkou signalizující přítomnost nesamostatného morfému „-s“ nahrazujícího tvar 2. osoby pomocného slovesa *být*, do dotazového řádku napíšeme: **[tag=".*zS.*"]** a stiskneme *Enter*.

Atribut „stylistický příznak“ – w (fakultativní)

Poznámka: Tento atribut je fakultativní. Atribut **w** s hodnotou **H** je přiřazován všem substandardním variantám koncovek (*děčkama, blbej, bráchem, prosim, ...*), frekventovaným substandardním variantám kmenů (*bejt, su, ...*), pravopisným chybám (příčemž mnohdy nelze rozlišit překlep a pravopisnou chybu v koncovce), chybně užitým tvarům zájmen (*mě/mně, jí/jí*), některým variantám nekodifikovaných slovtvorných inovací (*výkoňák, ...*)

hodnota		atribut+hodnota
H	substandardní varianty koncovek a kmenů	wH

Příklad: Budeme-li chtít vyhledat všechny tvary označované atributem substandardní stylistický příznak, do dotazového řádku napíšeme: **[tag=".*wH"]** a stiskneme *Enter*.

Přehledné tabulky řazení atributů a jejich hodnot podle slovního druhu

Podívejme se ještě jednou, jak vypadají značky pro jednotlivé slovní druhy. V následujících tabulkách nalezneme řazení atributů podle atributu slovní druh, který je v systému značek *atribut/hodnota* závazný pro všechny označované tvary.

Následující tabulky mají uživateli ukázat závazné řazení atributů a jejich hodnot při tvorbě dotazu pro korpusový manažer **Bonito**. Postupujeme podle atributu slovní druh. Znak „|“ (čti „nebo“) se používá při konstrukci složitějších dotazů v korpusovém manažeru **Bonito** a umožňuje formulovat disjunkci (dotaz zahrnující alternativu). Zde jej používáme pro vyjádření paralelních možností.

Substantiva - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k1
g	rod	gM gI gF gN
n	číslo	nS nP nD
c	pád	c1 c2 c3 c4 c5 c6 c7
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných substantiv rodu ženského v singuláru ve třetím pádě, pak značka bude mít následující formu: **[tag=“k1gFnSc3.*“]**

Příklad: Budeme-li chtít hledat tvary označovaných substantiv rodu středního v plurálu v akuzativu, pak značka bude mít následující formu: **[tag=“k1gNnPc4.*“]**

Příklad: Budeme-li chtít hledat tvary označovaných substantiv, která mají v instrumentálu plurálu substandardní tvar, pak značka bude mít následující formu: **[tag=“k1.*nPc7.*wH“]**

Adjektiva - řazení atributů

U adjektiv a příslovcí se vyplňuje atribut **d** – stupeň s hodnotou **1** – pozitiv i u tvarů, které stupňovat nelze. Jsme si vědomi, že jde o kompromis. Atribut **e** – negace (přítomnost/nepřítomnost prefixu *ne-* vyjadřujícího negaci) se rovněž vyplňuje u všech adjektiv.

atribut		atribut+hodnota
k	slovní druh	k2
e	negace	eA eN
g	rod	gM gI gF gN
n	číslo	nS nP nD
c	pád	c1 c2 c3 c4 c5 c6 c7
d	stupeň	d1 d2 d3
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary ženského rodu singuláru označovaných adjektiv v šestém pádě, pak značka bude mít následující formu: **[tag=“k2.*gFnSc6.*“]**

Příklad: Budeme-li chtít hledat substandardní tvary označovaných adjektiv, v libovolném rodě v plurálu ve druhém pádě, pak značka bude mít následující formu: **[tag=“k2.*nPc2.*wH“]**

Zájmena - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k3
p (fakultativně) u zájmen vyjadřujících osobu	osoba	p1 p2 p3
g (fakultativně) u zájmen rodových	rod	gM gI gF gN
n	číslo	nS nP nD
c	pád	c1 c2 c3 c4 c5 c6 c7
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných zájmen, vyjadřujících druhou osobu, pak značka bude mít následující formu: [tag="k3p2.*"]

Příklad: Budeme-li chtít hledat tvary označovaných zájmen v ženském rodě, v plurálu v libovolném pádě, pak značka bude mít následující formu: [tag="k3.*gFnP.*"]

Číslovky - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k4
g (fakultativně)	rod	gM gI gF gN
n (fakultativně)	číslo	nS nP nD
c	pád	c1 c2 c3 c4 c5 c6 c7
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných číslovek v genitivu, pak značka bude mít následující formu: [tag="k4.*c2.*"]

Poznámka: Základní číslovky od 5 výše mají ve značce uvedeny pouze atributy slovní druh, pád a fakultativní atributy **z** a **w** ([tag="k4c.*"]).

Číslovky psané číslicemi jsou označovány pouze atributem slovního druhu, jejich značka má tedy formu [tag="k4"].

Slovesa- řazení atributů

Atribut **e** – negace vyznačuje přítomnost/nepřítomnost prefixu *ne-* vyjadřujícího negaci. Lemmatem tvaru s prefixem *ne-* je příslušný infinitiv bez prefixu *ne-*.

Poznámka: jedinou výjimkou jsou tvary slovesa *být* (viz níže).

atribut		atribut+hodnota
k	slovní druh	k5
e	negace	eA eN
a	vid	aP aI aB
m	slovesný tvar	mF mI mR mA mN mS mD mB
p (fakultativně)	osoba	p1 p2 p3
g (fakultativně)	rod	gM gI gF gN
n (fakultativně)	číslo	nS nP nD
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary sloves v imperativu ve 2. osobě, pak značka bude mít následující formu: [tag="k5.*mRp2.*"]

Příklad: Budeme-li chtít hledat tvary sloves v indikativu, ve 3. osobě plurálu, pak značka bude mít následující formu: [tag="k5.*mIp3.*nP.*"]

Příklad: Budeme-li chtít hledat substandardní tvary sloves v indikativu v 1. osobě singuláru, pak značka bude mít následující formu: [tag="k5.*mIp1.*nS.*wH"]

Příklad: Budeme-li chtít hledat tvary sloves přechodníku přítomného v singuláru, pak značka bude mít následující formu: [tag="k5.*mA.*nS.*"]

Příslovce - řazení atributů

U příslovců a adjektiv se vyplňuje atribut **d** – stupeň s hodnotou **1** – pozitiv i u tvarů, které stupňovat nelze. Jsme si vědomi, že jde o kompromis. Atribut **e** – negace (přítomnost/ne přítomnost prefixu *ne-* vyjadřujícího negaci) se rovněž vyplňuje u všech adjektiv.

atribut		atribut+hodnota
k	slovní druh	k6
e	negace	eA eN
d	stupeň	d1 d2 d3
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných příslovců s prefixem *ne-* signalizujícím negaci, pak značka bude mít následující formu: **[tag=“k6eN.*“]**

Předložky - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k7
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat substandardní tvary označovaných předložek, pak značka bude mít následující formu: **[tag=“k7wH“]**

Spojky - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k8
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných spojek s připojeným nesamostatným morfémem „-s“ nahrazujícím tvar 2. osoby pomocného slovesa *být*, pak značka bude mít následující formu: **[tag=“k8zS.*“]**

Částice - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k9
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných částic, pak značka bude mít následující formu:

[tag=“k9.*“]

Citoslovce - řazení atributů

atribut		atribut+hodnota
k	slovní druh	k0
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat tvary označovaných citoslovců, pak značka bude mít následující formu: **[tag=“k0.*“]**

Zkratky - řazení atributů

atribut		atribut+hodnota
k	slovní druh	kA

Příklad: Budeme-li chtít hledat tvary označované jako zkratky, pak značka bude mít následující formu: [tag="kA"]

Tvary „by, ...“ - řazení atributů

atribut		atribut+hodnota
k	slovní druh	kY
m	kondicionál	cM
p	osoba	p1 p2 p3
n	číslo	nS nP nD
z (fakultativně)	spojitelnost se „-s“	zS
w (fakultativně)	styl	wH

Příklad: Budeme-li chtít hledat označované tvary kondicionálu v 1. osobě plurálu, pak značka bude mít následující formu: [tag="kYmCp1nP.**"]

Poznámka: Tvary *by, aby, kdyby* mají po automatické morfologické analýze pouze dvě varianty značek [tag="kYmCp3nS.**"] nebo [tag="kYmCp3nP.**"]. Jsme si vědomi toho, že se jedná o chybu: v případech spojení *by sis, by ses, aby sis, aby ses, kdyby sis, kdyby ses* tvar *by* nevyjadřuje 3. osobu, nýbrž 2. osobu. V korpusu **KSK-dopisy1** zůstaly značky těchto tvarů prozatím neopraveny – uvádí se 3. osoba singuláru (celkem 72 výskytů).

Speciální značky

Vzhledem k tomu, že značkování korpusů s vysokou frekvencí substandardních jevů přineslo řadu problémů, z nichž některé se nám dosud nepodařilo uspokojivě vyřešit, snažili jsme se hledat prozatímní řešení. Jedním z nich je kategorizace a ruční značkování sporně řešitelných případů neoznačovaných automatickou analýzou. Pro tyto případy jsme vytvořili speciální značky. Jsou přiřazeny slovním tvarům, jež nebylo možno klasifikovat pomocí existujících značek, popřípadě jednotlivostem, které by sice bylo možné uspokojivě zařadit, nicméně by si to vyžádalo více času, než jsme měli k dispozici.

Speciální značky mají následující podobu: krátký text popisující důležitou charakteristiku tvaru uzavřený v úhlových závorkách <>.

Přehled speciálních značek:

Poznámka: Znak „&“ (čti: *a*) se používá při formulaci dotazů v korpusovém manažeru Bonito k vyjádření logické konjunkce. V následujícím textu jej používáme, abychom ukázali lemmatizaci a značkování speciálními značkami. Slova označovaná speciálními značkami mají lemma identické s tvarem slova samotného.

[tag="<graficka_chyba>"]

Tato značka byla ručně přiřazena následujícím případům:

neúplné slovo

Například: *ta* místo *tak* [lemma="ta" & tag="<graficka_chyba>"]

spojení více slov

Například: *AhojBlani* místo *Ahoj Blani* [lemma="AhojBlani" & tag="<graficka_chyba>"]

rozdělené slovo

Například: říkej me místo říkejme [lemma="říkej" & tag="<graficka_chyba>"] [lemma="me" & tag="<graficka_chyba>"]

neidentifikovatelné slovo

Například:... nelze vyjít ze brány knihovny ... - z kontextu není zřejmé, zda jde o překlep ze místo z nebo místo za [lemma="ze" & tag="<graficka_chyba>"]

[tag="<zkratka>"]

Tato značka byla ručně přiřazena případům zkratk, které nebyly označovány automatickou morfologickou analýzou značkou **kA**.

Poznámka: Automatická morfologická analýza pracuje záměrně jen s omezeným množstvím zkratk.

[tag="<anglicky>"]

[tag="<nemecky>"]

[tag="<francouzsky>"]

[tag="<jiny_jazyk>"]

Tyto značky jsou přiřazeny některým frekventovaněji užitým anglickým, francouzským, německým, slovenským, ruským aj. slovům v textech.

Poznámka: Delší úseky textů v cizích jazycích byly při přepisu dat odstraněny do "poznámky", takže se s nimi v běžném modu nepracuje. Záměrně však byla v textu ponechána jednotlivá cizojazyčná slova a slovní spojení. Makarónský způsob vyjádřování je totiž charakteristickým rysem v dopisech zejména mladých pisatelů.

Upozornění: Pokud si uživatel není jistý, jaké **lemma** případně značku (**tag**) mají slovní tvary, které chce vyhledávat, může se dotázat na jednoho ze zástupců skupiny, kterou chce vyhledat. Jakmile je vyhledán konkordanční seznam, z nabídky **Zobrazení** zvolí řádek **Atributy (Zobrazení > Atributy)**. Objeví se nabídka, v níž zvolí (kliknutím myši) atribut **lemma** a **tag**. U klíčových slov se objeví za znakem „/“ lomítko jejich /**lemma** a značka /**tag**.

Například: Uživatel si nebude jistý, jak vypadají značky u slovesných tvarů (jaké je řazení atributů ve značce). Zeptá se na lemma některého frekventovaného slovesa, do dotazového řádku zapíše např. [lemma="být"] a stiskne *Enter*. Na obrazovce se objeví konkordanční seznam, v pozici klíčového slova budou nejružnější tvary slovesa *být*. Uživatel zvolí nabídku **Zobrazení > Atributy**, kliknutím myši zatrhne **lemma** a **tag** a klikne na *Budiž*.

Z výsledku jednoduše vyčte pořadí atributů:

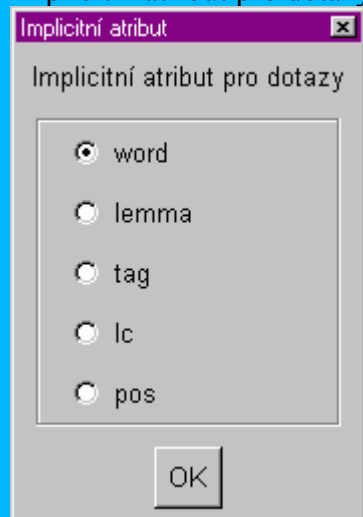
```
.....
Doufám , že už          < jste/být/k5eAaImIp2nP> zdraví . Já už celkem jo
rýmy , ale snad už to  < bude/být/k5eAaImBp3nS> dobrý . Ale měla bych to
Samozřejmě , že        < je/být/k5eAaImIp3nS> to blbost , na to se vím
už dávno , mimo to     < jsme/být/k5eAaImIp1nP> to brali i loni v
neprožívá , protože    < su/být/k5eAaImIp1nSwH> střízlivá , ale večer se
na Lochotín a vydaly   < sme/být/k5eAaImIp1nPwH> se pěšky přes sídliště k
vidět z obrázku , měly < ste/být/k5eAaImIp2nPwH> se přímo skvěle .
" Přesto , že          < seš/být/k5eAaImIp2nSwH> hroznej , mám Tě ráda
nevím , protože ty    < si/být/k5eAaImIp2nSwH> tam vlastně nechodila )
.....
```

Nastavení implicitního atributu

Po spuštění je program **Bonito** nastaven tak, že se v dotazovém řádku předpokládá dotaz na slovní tvar, případně posloupnost slovních tvarů. Znamená to, že je nastaven implicitní atribut **word**. Dotazy na ostatní atributy je nutné formulovat pomocí výše uvedených výrazů, které mají obecný tvar: **[jméno_atributu="hodnota_atributu"]**. Je ovšem možné zvolit jiný implicitní atribut podle toho, který typ dotazu klademe častěji. Zvolíme-li například atribut **tag** jako implicitní, nebudeme muset dotaz zapisovat formou **[tag="k2eAgFnPc1.*"]**, ale do dotazového řádku přímo napíšeme morfologickou značku: **k2eAgFnPc1.***. Implicitní atribut můžeme změnit pomocí položky **Korpus > Implicitní atribut**. Objeví se následující nabídka:

Implicitní atribut

Implicitní atribut pro dotazy



(Sada dostupných atributů se může lišit podle zvoleného korpusu.)

Implicitní atribut vybereme kliknutím myši do kolečka u zvoleného atributu a potvrdíme kliknutím na **OK**.

Při práci s korpusem **KSK-dopisy1** můžeme vybírat z těchto atributů:

- word** Tento atribut je nastaven jako implicitní vždy po spuštění programu. Do dotazového řádku zadáváme jednotlivé slovní tvary. Například: po zadání dotazu **kočky**, manažer vyhledá pouze texty s výskytem tvaru *kočky*. Při vyhledávání podle atributu **word** záleží na velikosti písmen.
- lemma** Nastavíme-li atribut lemma jako implicitní, budeme vyhledávat podle základního slovníkového tvaru (lemmatu). Do dotazového řádku pak zadáváme přímo lemmata. Například: do dotazového řádku napíšeme slovo **kočka**, manažer vyhledá výskyty tvarů odpovídajících tomuto lemmatu, tj.: *kočka, kočky, kočku, koček, kočkou, kočce* atd. Můžeme ovšem hledat i dvě (nebo více) lemmat vedle sebe. Zadáním dotazu **dravá kočka**, dostaneme výskyty: *dravá kočka, dravých koček* atd.

tag Pokud nastavíme tento atribut jako implicitní, do dotazového řádku budeme zapisovat přímo morfologickou značku nebo posloupnosti těchto značek. Například zadáním posloupnosti značek **k2.* k1gFnPc7.*** získáme výskyty všech adjektiv, za kterými stojí substantivum ženského rodu v instrumentálu: tj. *podobnými diskusemi, nebezpečnými dívkami, pravopisnými chybami, nějakými holkami* atd.

Označovaný korpus KSK-dopisy1 - rozsah a spolehlivost morfologického značkování

Morfologické značky (tagy) jsou výsledkem automatické lemmatizace (slovnímu tvaru v textu je automaticky přiřazen příslušný základní tvar – lemma) a automatické morfologické analýzy (danému slovnímu tvaru v textu jsou automaticky přiřazeny slovnědruhové a morfologické interpretace). Automatické morfologické analýze předchází tokenizace, tj. segmentace textu na jednotky, které v ideálním případě odpovídají textovým slovům, v podstatě jde však o zjednodušení lingvistického přístupu v tom smyslu, že slovní tvar se chápe jako řetězec znaků mezi mezerami, popř. jinými oddělovači, jimiž mohou být např. interpunkční znaménka. Automatická lemmatizace a automatická morfologická analýza přiřazují jednotkám textu (textovým slovům, token) všechny kontextově nezávislé interpretace. Morfologická analýza je obecně nejednoznačná. Nejednoznačnost je způsobena vysokou mírou homonymie způsobenou tvarovou homonymií uvnitř paradigmatu jednoho systémového slova, homonymií úplnou nebo částečnou (překrytí všech, či několika tvarů) dvou různých lexikálních jednotek, homonymií způsobenou funkčními i slovnědruhovými transpozicemi mezi jednotlivými (především neohebnými) slovními druhy. **Míra koncovkové homonymie uvnitř paradigmatu jednoho slova podstatně vzrostla zařazením automatické analýzy možných substandardních tvarů, které se v korpusu soukromé korespondence vyskytují poměrně frekventovaně.**

Automaticky označovaný korpus byl ručně disambiguován (z více interpretací lemmat a značek byla ručně vybrána jedna interpretace platná pro daný kontext).

Po automatické analýze a ruční disambiguaci zůstalo 5 % neoznačovaných popřípadě nedisambiguovaných tvarů. Z nich byla 1,6 % ručně doznačkována.

Jednalo se o tyto typy:

- 1) interpunkce (více teček, pomlček, ...)
- 2) ciferné výrazy (především data v dopisech)
- 3) pravopisné chyby (*zapomě, myslym, ...*)
- 4) substandardní tvary zájmen (*všeci, všici, ...*) a nesprávně použité tvary zájmen (*mě/mně, mně/mě, jí/ji, ji/jí*), které byly při ruční disambiguaci ponechány bez značek
- 5) grafické chyby (neúplné slovní tvary, spojení více slovních tvarů, slovní tvary rozdělené do více pozic, ...)
- 6) nejrůznější zkratky
- 7) cizí slova

8) substandardní tvary kondicionálů (*aby, by, kdyby + jsem, sem, jsi, si, jseš, je, sme, jsme, ste, jste,...*, *bys ses, ...*), které nebyly rozpoznány automatickou morfologickou analýzou, popř. disambiguovány při ruční disambiguaci

9) substandardní tvary slovesa *být*, které nebyly rozpoznány automatickou morfologickou analýzou, popř. disambiguovány při ruční disambiguaci

10) substandardní slovotvorné inovace (*výkoňák, ...*)

11) substandardní slovotvorné inovace pozdravů (*čauky, čauec, ahojda, ...*).

Korpus **KSK-dopisy1** je z 96,6 % lemmatizován a označován morfologickými tagy.

(Předkládaná verze je pracovní, není tedy ještě zcela spolehlivá. Momentálně probíhá její několikastupňová kontrola.)

Nedisambiguované označované tvary

Některé označované tvary byly záměrně ponechány nedisambiguované, takže je lze vyhledávat podle několika lemmat a jim odpovídajících značek. Jedná se především o frekventovaná slova patřící k více neohebným slovním druhům, u nichž ruční disambiguace narážela již při značkování korpusů spisovného jazyka na značné obtíže (disambiguátor se nebyl schopen rozhodnout, více disambiguátorů docházelo pravidelně k rozporným rozhodnutím). [Jejich seznam najdete zde.](#) [seznamnedisam.doc](#)

Do skupiny označovaných nedisambiguovaných tvarů patří rovněž případy často se vyskytujícího chybného užití tvaru zájmena *mě* v dativu nebo lokálu. Tvar je označován nejednoznačně dvěma možnými značkami [tag="k3p1nSc3wH"] a [tag="k3p1nSc6wH"]. Substandardnost tvaru naznačuje přítomnost atributu *w* s hodnotou *H*.

Nedisambiguovány zůstaly také tvary adjektiva *rád, ráda, rádo, rádi, rády*, u kterých nebyl disambiguován rod, číslo a pád, takže např. tvar *rád* má tagy [tag="k2gMnSc1"], [tag="k2gInSc1"], [tag="k2gInSc4"].

Poznámky:

Kompromisem je označování tvarů syntetického futura některých sloves pohybu tvořených prefixem *po-/pů-* připojeným ke tvarům indikativu přítomnosti (např. *půjdu, poletím, poběží, poteče, ...*). Tyto tvary se ve značce neliší od tvarů indikativu přítomnosti.

Například: Tvar *letím* a tvar *poletím* má stejnou značku: [tag="k5eAaImIp1nS"]

Nedokonalostí automatického morfologického analyzátoru AJKA je značkování tvarů slovesa *být*, které mají prefix *ne-* vyjadřující negaci. Analyzátor nabízí u těchto tvarů lemma *být* nebo lemma *nebýt*, ale značka je vždy [tag="k5eA.*"].

Například: Tvary *nebýt, nebyl, není, nejsem, nejsou, ...* mají lemma *být*, ale ve značce je hodnota atributu *e* *A* a nikoli *N*. Tato nedokonalost zůstala prozatím neopravena.

Vynechané atributy a hodnoty

V případě, že z kontextu nebylo možné jednoznačně disambiguovat hodnotu některého z atributů, byl při ručním doplnění značky tento atribut vynechán

Například: ... *tak mi připadalo* , že nemá *všech* pět pohromadě ... byla tvaru *všech* přiřazena značka [tag="k3nPc2"], která má vynechán atribut rod.

Neoznačkové tvary

Část slovních výskytů zůstala zatím neoznačkována (3,4 %, tedy 16 617 výskytů, 11 020 různých tvarů). K důvodům patřilo vysoké procento substandardních forem s nízkou frekvencí výskytu, jejichž ruční označování je velmi náročné. Jde například o neúplná slova, samostatná písmena, substandardní adaptace cizích slov atd. Objevilo se i množství nejednoznačně interpretovatelných jednotek. Sem patří například vlastní jména, u nichž nelze ani z kontextu určit rod (např. *Ahoj Rady ...*), dále neúplná slova, u nichž nelze jednoznačně rozhodnout, co bylo vynecháno (např. *Te sme se prostě jen tak ze srandy postrkovali...*) atd.

Úplný seznam neoznačkových tvarů najdete [zde_poslednifrekv.doc](#)