

**PLIN063\_3**

# Algoritmický popis morfologie

osolsobe@phil.muni.cz

# Vybrané problémy tokenizace

- definice slova jako lingvistický problém
- technické problémy a technická řešení
- interpretace víceslovných jednotek (MWE)
- analytické tvary
- agregáty, zkratky
- problém chybného grafického zápisu
- problémy konverze textu

# TOKEN (NESČ)

- Nejmenší jednotka textu, většinou grafické slovo, resp. jedna jeho realizace ([↗type-token](#)). V korpusové lingvistice je v některých případech jedno grafické slovo rozděleno na dvě slova (např. *mohu -li*), často je také z praktických důvodů (pro snadné vyhledávání) oddělována interpunkce od předcházejícího či následujícího slova (3 tokeny: *řekl , že*). O jednotlivých **t.** v korpusu se také mluví jako o pozicích. – Velikost [↗korpusu](#) se udává v **t.n.** také v [↗textových slovech](#). Rozčlenění textu na **t.** je výsledkem procesu [↗tokenizace](#). Viz také [↗tokenizér](#).

# TOKENIZACE (NESČ)

- V korpusové lingvistice automatický proces, který člení text složený z písmen, interpunkčních znamének a mezer na jednotlivé izolované [↗tokeny](#), tj. na slovní tvary a interpunkční znaménka pro účely dalšího (obvykle počítačového) zpracování; viz [👉 Baker & Hardie ad. \(2006\)](#); [👉 Jelínek & Petkevič \(2011\)](#). Při t. se typicky odděluje interpunkce od následujícího či předcházejícího slova, v některých případech se rozděluje slovo skládající se z více slov spojených spojovníkem, identifikují se zkratky končící či nekončící tečkou. T. je obvykle první fází zpracování textu, typicky bezprostředně předchází procesu [↗větné segmentace](#) a [↗morfologické analýzy](#). T. provádí počítačový program zvaný [↗tokenizér](#).

# TOKENIZÉR (NESČ)

- Počítačový program provádějící [↗tokenizaci](#). T. je buď:  
(a) stochastický (statistický), který tokenizuje text na základě [↗strojového učení](#) (zvláště při zpracování více jaz.), n. (b) založený na pravidlech tokenizace platných pro zpracovávaný jaz. Viz také [↗token](#).

# definice slova jako lingvistický problém

- Textové slovo
- Gramatická jednotka (*ptát se, řekl by sis*)
- Slovo tvorná jednotka (*na měkko*)
- Morfosyntax (*dveře se zabouchly*)

# technické problémy a technická řešení

- Jednoslovná morfologie závislá na tokenizaci
- Definice slova

# interpretace víceslovných jednotek (MWE)

- Vlastní jména
- Databáze víceslovných jednotek (frazémy, idiomy, kolokace)
- Cizojazyčné kratší úseky textů



# Vlastní jména uložená v morfologickém slovníku

CQL



[Vložit tag](#) | [Vložit 'within'](#) | [Klávesnice](#) | [Předchozí dotazy](#)

```
[tag="N.*" & lemma="[:upper:].*"]
```

kdy na jeho duel s	<b>Eintrachtem</b>	Frankfurt přišlo v roce 1960
, které měli Dušek a	<b>Dalekorej</b>	a spoustu dalších náznaků na
moje jméno , pomyslela si	<b>Malin</b>	. Navenek každopádně nedal nic
10 438 km čtverečních . Na	<b>Oahu</b>	pak leží hlavní město Honolulu
jedinců na tomto zimovišti u	<b>Štěpánova</b>	rok od roku roste .
Evropa po roce 1250 ,	<b>Holandsko</b>	v 16 . století ,
HELENA POD ZÁMKEM Návrhářka Frida	<b>Giannini</b>	čerpá z florentinské řemeslné tradice
Syngenta ve spolupráci s Agro	<b>Žlunice</b>	ve Žlunicích , okres Jičín
zemřel kapitán Lübbe a od	<b>Lisabonu</b>	převzal pro zbytek cesty velení
povolení , " poukázal .	<b>Krnov</b>	chystá výstavbu nového skate-bikeparku KRNOV
přepravily letecké společnosti Aeroflot a	<b>ČSA</b>	celkem 48 945 cestujících , což
Vím já ? " řekl	<b>Carl</b>	a olízl si rty .
, kterou popáté zorganizovalo domažlické	<b>SOU</b>	. Zvítězil tým ze Střední
Watfordu nastřílel 22 branek .	<b>Matěj</b>	Vydra GETTY Zprávy krátce Miami
případě Československa , Polska ,	<b>Maďarska</b>	, Rumunska , Bulharska či
" zhodnotil starosta Litoměřic Ladislav	<b>Chlupáč</b>	( ODS ) výsledek náročných

# Databáze víceslovných jednotek (frazémy, idiomy, kolokace)

- Zlepšení desambiguace
- FRANTA (FRazémová ANotace a Textová Analýza) Program pracuje tak, že v korpusu vyhledává a označuje frazémy a ustálené kolokace (dále budeme používat zastřešující pojem *víceslovné jednotky*) z předem daného slovníku (seznamu), který v současné době vychází především ze *Slovníku české frazeologie a idiomatiky* a obsahuje okolo 40 000 položek.
- [https://wiki.korpus.cz/doku.php/kurz:hledani\\_frazemy](https://wiki.korpus.cz/doku.php/kurz:hledani_frazemy)

# Příklady vyhledávání frazémů v korpusu

## SYN2015

**([https://wiki.korpus.cz/doku.php/kurz:hledani\\_frazemy](https://wiki.korpus.cz/doku.php/kurz:hledani_frazemy))**

- [Hledání podle kolokačního lemmatu](#)
- [Hledání podle části frazému](#)
- [Hledání přirovnání s konkrétním tvarem slova](#)
- [Hledání slovesné fráze s konkrétním substantivem](#)
- [Hledání nominální fráze s konkrétním adjektivem](#)
- [Hledání všech frazeologických užití daného paradigmatu](#)
- [Hledání podle slovnědruhového vzoru](#)

# Hledání podle kolokačního lemmatu

- Úkol: Najděte všechny výskyty konkrétního frazému (např. obsahujícího substantivum označující část těla: *přijít věci na kloub*).
- V „neofrazémovaném“ korpusu bychom zadali asi tyto dotazy:
- **[lemma="přijít"] [word=".\*"]{0,5} [lemma="na"] [lemma="kloub"]**  
**[lemma="na"] [lemma="kloub"] [word=".\*"]{0,5} [lemma="přijít"]**
- Ve frazeologicky označovaném korpusu nalezneme všechny výskyty tohoto frazému dotazem
- **[col\_lemma="přijít\_na\_kloub" & col\_type=".\*H"]**
- Chceme-li nalézt všechna rozšíření nebo varianty tohoto frazému, můžeme zadat dotaz
- **[col\_lemma=".\*na\_kloub" & col\_type=".\*H"]**
- **[lemma="kloub" & prep="na" & e\_lemma="přijít"]**

# Všimněte si značkování

nazval tuto inscenaci „ work	<b>in</b> /in/XX----- <b>progress</b> /progress/X@-----	“, čímž narážel na
, LLC . First appeared	<b>in</b> /in/XX----- <b>THE</b> /the/XX-----	STRAND MAGAZINE . Reprinted by
zavřískla : Dios mío ,	<b>qué</b> /qué/X@----- <b>me</b> /me/XX-----	has hecho ? – a
v Brestu , Lorientu a	<b>Saint</b> /Saint/XX----- <b>Brieuc</b> /Brieuc/X@-----	. V Bretani v Breizhu
překlada znamená Město andělů .	<b>Ofi</b> /Ofi/X@----- <b>ciální</b> /ciální/X@-----	jméno thajské metropole pak představuje
Andrewa Rawnsleyho nazvaná The End	<b>of</b> /of/XX----- <b>the</b> /the/XX-----	Party . Britský list The
National Science Foundation : Foraging	<b>behavior</b> /behavior/X@----- <b>and</b> /and/XX-----	demography of Pygoscelis penguins .
na dobře prostudovaném " Misteriu	<b>Méki</b> /Méki/X@----- <b>Pańskieij</b> /Pańskieij/X@-----	" v malopolské Kalwarii Zebrzydowské
in time a change ,	<b>doesn't</b> /doesn't/XX----- <b>matter</b> /matter/X@-----	your party you believe .
místopředsedou lotyšského Svazu zemědělců (	<b>Zemnieku</b> /Zemnieku/X@----- <b>savienība</b> /savienība/XX-----	) Ādolfsem Klīvem , pozvaná
vůbec . Co víte o	<b>Nižněkam</b> /Nižněkam/X@----- <b>sku</b> /sku/X@-----	kromě toho , že se
nové knize What has nature	<b>ever</b> /ever/X@----- <b>done</b> /done/XX-----	for us ? ( Co
angličtině vždy písmenem L. Producentka	<b>Ilene</b> /Ilene/X@----- <b>Chaiken</b> /Chaiken/X@-----	se snažila , aby alespoň
Ale řečeno mezi námi ,	<b>Shmuel</b> /Shmuel/X@----- <b>Rodensky</b> /Rodensky/X@-----	je lepší . . .

# agregáty

- Agregát popisuje slovní tvar, který zastupuje dva nebo více slovních tvarů (složek agregátu) a většinou mu není možné přiřadit jednoduše slovní druh. Většinou lze agregát ve větě původními slovními tvary nahradit, aniž by se změnil smysl věty.
- Příklady: *naň = na něj...* agregát *naň* má dvě složky: *na* a *něj*, *byls = byl jsi*
- Agregátem není slovo, které vzniklo jedním z tradičních způsobů slovo tvorby – skládáním. Takové slovo má svůj vlastní význam a již se jednoznačně přiřadilo k některému slovnímu druhu. Agregáty tedy například nejsou slova *černobílý*, *novotvar*, *spolupořádat*, přestože také vznikla složením různých slov. Mají totiž vlastnosti jediného slovního druhu (v uvedených příkladech po řadě adjektivum, substantivum, sloveso), a to jak morfologické, tak i syntaktické.

# Zájmenné agregáty

- **Zájmenné agregáty** jsou slovní tvary vzniklé spojením předložky a substantivního zájmena *co* nebo substantivního určitého (osobního) zájmena *on*. Podle toho rozlišujeme dva typy, které jsou tvořeny uzavřenou množinou tvarů. Můžeme je tedy zadat výčtem.
- Za X v posledních řádcích seznamu je možné dosadit některé ze slov, která se podílejí na tvorbě neurčitých zájmen, číslovek a příslovcí, totiž *kdoví, bůhví*, a další.

# Seznam agregátů s druhou složkou zájmenem *co*

- *zač* = *za co*, lemma: {*za*, *co*}
- *nač* = *na co*, lemma: {*na*, *co*}
- *oč* = *o co*, lemma: {*o*, *co*}
- *več* = *v co*, lemma: {*v*, *co*}
- *začpak* = *za copak*, lemma: {*za*, *copak*}
- *načpak* = *na copak*, lemma: {*na*, *copak*}
- *očpak* = *o copak*, lemma: {*o*, *copak*}
- *?večpak* = *v copak*, lemma: {*v*, *copak*}
- *Xzač* = *za Xco*, lemma: {*za*, *Xco*}
- *Xnač* = *na Xco*, lemma: {*na*, *Xco*}
- *Xoč* = *o Xco*, lemma: {*o*, *Xco*}
- *?Xveč* = *v Xco*, lemma: {*v*, *Xco*}



# Seznam agregátů s druhou složkou zájmenem *on*

- *zaň* = *za něho*, lemma: {*za*, *on*}
- *naň* = *na něho*, lemma: {*na*, *on*}
- *oň* = *o něho*, lemma: {*o*, *on*}
- *proň* = *pro něho*, lemma: {*pro*, *on*}
- *veň* = *v něho*, lemma: {*v*, *on*}
- *doň* = *do něho*, lemma: {*do*, *on*}

# Problém

	ze školky a zarotí :	veň /veň/x@-----	jsme přezpívali všechny naše nej
	další dvě trefy . Úrojsme	veň /veň/x@-----	zápasu nebyla valná , "
	kolem metanu a zá ro	veň /veň/x@-----	proměřit jeho emise v zemské
	, ale i může rozhodovat	veň /veň/x@-----	s o tom , na
	která je nyní použita zárottest	veň /veň/x@-----	jako volič převodovky . O
	byl zároveň lékařem závodní preventivní	veň /veň/x@-----	péče . Pro zaměstnavatele je
	ho řadil na úroveň Zidaneho	veň /veň/x@-----	a dalších . " d
ny	veliké obecno , neboť možno	veň /veň/x@-----	vtěsnati celé veškerenstvo , všechna
	až do předsíně . Zár	veň /veň/x@-----	zmizí umakartové jádro , které
	prostředí a zá ro -	veň /veň/x@-----	to vyjadřuje záměr zadavatele .
	George Bushe nemyslitelné . Gorezáro	veň /veň/x@-----	zasahuje i do výběru poradců
	ovit slova , musím zároveň	veň /veň/x@-----	se zpěvem zvládnout to šílené
	1 Jan Bürgermeister . Záro	veň /veň/x@-----	po otevření tako vých středisek

# Slovesné agregáty (se jmény)

- *Z latinys měl reparát loni.*
- *Věrnýs jak kůň, jak býk všaks vášnivý.*
- *Salome, podobnas úponku (z písně Karla Kryla)*
- *Copaks to musel řešit zrovna takhle?*
- *Tos řekl ty, já ne.*
- *Všechno, o čems mluvil...*
- *Koliks jich koupila?*
- *Kolikráts to viděl?*

# Slovesné agregáty (s neohebnými slovními druhy)

- *Včeras* měl narozeniny.
- *Posledněs* říkala, že...
- *A ještěs* mě nikdy neodměnil.
- *Nikdys* nechtěla vařit.
- *Jaks* k tomu došla?
- *To mi neříkej. Snads* ji neznala?
- *Vždyťs* povídal, že sis ji sám vymyslel.
- *A určitěs* to ztratila?
- *Možnás* jim radši vůbec neměl dávat!
- ... *kdyžs* teda říkal,...
- ... *neměls* už čas se zase stejnou cestou vrátit, *nebos* na to zapomněl.
- *Nevím, jestlis* ho vůbec znal.

# K slovesným agregátům se slovesy a kondicionálovým agregátům

- ... *má milá ženo, **bylas** tak statečná...*
- ***Koupils** ho Ireně.*
- spisovné: *bych, bys, bychom, byste*
- nespisovné: *bysem, byjsem, bysi, byjsi, bysme, byjsme, byjste*
- spisovné: *abych, abys, bychom, byste*
- nespisovné: *abysem, abyjsem, abysi, abyjsi, abysme, abyjsme, abyjste*
- spisovné: *kdybych, kdybys, kdybychom, kdybyste*
- nespisovné: *(k)dybysem, (k)dybyjsem, (k)dybysi, (k)dybyjsi, (k)dybysme, (k)dybyjsme, (k)dybyjste, dybych, dybys, dybychom, dybyste*

# Zkratky

- V NovaMorf se za zkratkový agregát považuje pouze takový vybraný jednoslovný výraz, který zastupuje několik slov, ve své základní podobě je psán malými písmeny a zároveň buď (i) zajišťuje textovou návaznost (např. *např.*, *atd.*, *tj.*, *apod.*, *tzn.*, *aj.* aj.), nebo (ii) zastupuje frekventovaná spojení apelativ (čp.).
- Zkratkový agregát B (B podle aBreviace) se skládá vždy ze zkratky a dále z alespoň jedné další složky, již může, ale nemusí být zkratka (viz např. *atd.* se dvěma zkratkami *t*, *d* a jednou nezkratkou *a*). Každé ze složek agregátu je přiřazena hodnota AGR=B.

# Afixový segment (POS=S)

[https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attrs=lemma&attrs=tag&attr\\_vmode=visible-kwic&base\\_viewattr=word&refs=%3Ddoc.title&q=~228M8mgkMUCY](https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attrs=lemma&attrs=tag&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.title&q=~228M8mgkMUCY)

- Nový slovní druh **S** zahrnuje dva typy slovních částí: prefixoidy a sufixoidy.
- Jako **prefixoidy** se značkují první části složených slov, které se obvykle nevyskytují samostatně, ale v daném textu jsou odděleny (nejčastěji spojovníkem a mezerou) od následujícího slova, např. *troj - až pětipodlažní*, původně psáno jako *troj- až pětipodlažní*). Značku **S** dostávají také vymezené slovní části, které byly od kompozita odtrženy procesem tokenizace (např. *makro - úroveň*, původně psáno jako *makro-úroveň*; viz popis procesu [tokenizace](#)). U prefixoidů se značkují pouze první dvě pozice. Na 2. pozici se vždy uvádí hodnota **2**.
- Jako **sufixoidy** se značkují pouze striktně vymezené koncové části substantiv a adjektiv po lomítku (a případně po spojovníku), jimiž pisatel značí morfologickou alternativu: např. *učitel / ka*, *pověřený / - á*. U slovesných příčestí (např. *viděl / a*, *unaven / -a*) budou kvůli chybě ve zpracování sufixoidy anotovány až od korpusu SYNv10. V tagu sufixoidy přebírají na dalších pozicích všechny značky od morfologické třídy, kterou zastupují, např. *ka* z výše uvedeného příkladu bude mít tag SNFS1-----A-----.

# problém chybného grafického zápisu

- Mezera navíc
- Čím kratší řetězec, tím větší možnost náhodného vzniku existujícího tvaru
- Komplikace pro rozšíření slovníku



# Příklad

na kterou následně reaguje společnost	<b>OPEK</b> /Opek/NNIS1-----A-----	, která se zabývá distribucí
pana hostinskýho někdo právě pěkně	<b>vypek</b> /vypéci/VpIS---3R-AA--6P	. Ostatně ? " Ostatně
jsi všechno tohle uvařil a	<b>upek</b> /upéci/VpIS---3R-AA--6P	, tak si vážně budeme
byl . Říká mu "	<b>opek</b> /opéci/VmIS-----A----P	a doma má v džbánečku
: Starej , cos to	<b>upek</b> /upéci/VpIS---3R-AA--6P	, samý brus ! .
akvizicí polské lokální kurýrní společnosti	<b>Opek</b> /Opek/NNIS1-----A-----	, která má roční tržby
, aby je nákej krajan	<b>vypek</b> /vypéci/VpMS---3R-AA--6P	. Káčo , ozdob vlas
. " " Už jsem	<b>upek</b> /upéci/VpMS---1R-AA--6P	plech těch čokoládovejch sušenek .
naší atletiky šestasedmdesátiletý EMIL ZÁT	<b>OPEK</b> /Opek/NNIS4-----A-----	. V roce 1952 v

# problémy konverze textu

- Konverze textu do jednotného formátu
- Čištění a deduplikace textů z webu
- Konverze elektronických textů (bez dalších oprav)
- OCR
- Přepisy
- Ve výsledku jsou v textu chyby

# Chyby v textu

- Rozdělená a spojená slova
- Překlepy a chyby proti kodifikaci
- Záměrná užití nekorektních tvarů
- Nedostatky v užití nástroje elektronického zpracování textu
- Nedostatky způsobené nedostatečnou kontrolou při posteditaci OCR textu
- Chyby proti pravidlům přepisu