

Základy využití korpusu v praxi **cjbb75**

Středa: 10.50-12.20 G13

28. 2. Dú: získat přístupová práva (**uživatelské jméno/heslo**), přečíst si charakteristiky dostupných korpusů ÚČNK, znát odpovědi na otázky v hdt.

7. 3. Kontext – **příkazový řádek** – vyhledávání podle atributů (**word, lemma, tag, lc**), **cql, regulární výrazy, KWIC, konkordanční seznam a frekvenční seznam** jako zdroj poznání o jazykových vlastnostech vyhledaných jednotek (způsob psaní/pravopis: přejatá slova /*mailovat/mejlovat/majlovat/*, aplikace pravidel – varianty /ismus/izmus/, varianty pravopisu zdobnělin vlastních jmen – Terinka/Terynka).

https://wiki.korpus.cz/doku.php/manualy:kontext:novy_dotaz

Typ dotazu

Typy dotazů a jejich využití


Typ dotazu	Nač je vhodný	Jak funguje	Co umí	Příklady
Základní dotaz	pro první seznámení s korpusem	Vyhledá vložený výraz jako slovní tvar bez ohledu na velikost písmen: jde-li zároveň o základní slovníkový tvar (lemma), vyhledají se také všechny jeho tvary.	bez regulárních výrazů (RE), case-insensitive (nerozlišuje malá a velká písmena)	černý kočka > černá kočka, černou kočku, černých koček... černá kočka > černá kočka
Lemma	pro analýzu celého paradigmatu/lexému	Vyhledá všechny tvary přiřazené k danému lemmatu.	RE (možnost použít regulární výrazy), case-sensitive (rozlišuje malá a velká písmena), možnost upřesnit slovní druh	černý > černý, černému, černá, černé, černými... kočka > kočka, kočku, koček, kočkám...
Fráze	pro vícetvarovou kombinaci v daném tvaru	Vyhledá přesně zadanou frázi.	RE, case-sensitive	černý pes > černý pes český pes > český pes černého psa > černého psa
Slovní tvar	pro analýzu jednoho konkrétního tvaru	Vyhledá přesně zadaný tvar.	RE, case-in/sensitive (možnost volby Shoda velikosti)	jakkoli > jakkoli jakkoliv > jakkoliv jakkoli.* > jakkoli, jakkoliv, Jakkoli, JAKKOLIV...
Podřetězec	pro vyhledání řetězce znaků kdekoli ve slově	Vyhledá po sobě následující znaky v rámci jednoho slova.	RE, case-sensitive	pra > praděda, praxe, doprava, lepra... křá > pookřát, křáp, Jiskřákovi...
CQL	pro vyhledání všeho, co lze pomocí korpusového manažera vyhledat	CQL je Corpus Query Language, korpusový dotazovací jazyk (do něhož si rozhraní KonText samo interně převádí všechny předchozí typy dotazů).	RE, case-sensitive, CQL syntax	[lemma="kočka"] > kočka, kočku, koček, kočkám... [word="černá"] > černá [lemma="číst"] [tag="N.*"] > číst levity, četli článek, nečtete noviny...

Cvičení

Jak česky píšeme?

mail

? mailovat, meilovat, majlovat, mejlovat ?

Korpus:	<input type="text" value="syn"/> 
Typ dotazu:	<input type="text" value="Lemma"/>
Lemma:	<input type="text" value="mail.*"/>

Celkem: 69 (2 str.)

	Filtr	lemma	Frekvence
1.	p/n	mail	18 839
2.	p/n	mailový	3 773
3.	p/n	mailovat	1 357
4.	p/n	mailbox	138
5.	p/n	mailing	129
6.	p/n	mailování	117
7.	p/n	mailer	102
8.	p/n	mailto	41
9.	p/n	mailingový	18
10.	p/n	maillot	13
11.	p/n	mailserveru	11
12.	p/n	mailerový	11
13.	p/n	mail-artu	9
14.	p/n	mailovací	8
15.	p/n	mailově	6

Složitější formulace dotazu

[lemma="m(ai|ei|aj|ej)l.*"]

https://wiki.korpus.cz/doku.php/manualy:kontext:frekvencni_distribuce

Menu: Frekvence

Pod označením *Frekvence* se v menu rozhraní KonText nachází funkce pro vytváření **frekvenčních distribucí**. Pomocí ní je možné získat přehled **typů** (např. různých slov) ve výsledku hledání spolu s jejich frekvencí. Vyhledáme-li např. všechna substantiva v gen. pl., můžeme pomocí této funkce zjistit, jaké **tvary** se v tomto pádě a čísle vyskytují a jak často. Stejně tak je ale možné frekvenční distribuci použít na zjišťování frekvencí jednotek předcházejících či následujících, počítání **lemmat** v **konkordanci**, nebo pro zjištění distribuce hledaného jevu napříč různými texty či jejich skupinami (podle **genre**, **txtype** apod.).

Frekvenční distribuce umožňuje jednak vlastní (obecné) nastavení a jednak **rychlé volby** (ty jsou přístupné v druhé úrovni menu):

1. **Lemmata** - vyhodnotí dotaz (**KWIC**) a vypíše všechna různá lemmata (atribut **lemma**) spolu s jejich frekvencí¹⁾
2. **Slovní tvary** - vyhodnotí dotaz (**KWIC**) a vypíše všechny různé tvary (atribut **word**) spolu s jejich frekvencí
3. **Dokumenty** - vyhodnotí celou **konkordanci** a vypíše názvy textů (**strukturní atribut nazev**), v nichž se hledaný jev vyskytuje, spolu s frekvencí tohoto jevu v jednotlivých textech
4. **Typy textů** - vyhodnotí celou **konkordanci** a vypíše přehled strukturních atributů²⁾, které se vztahují k typu textu (**strukturní atributy txtype_group, txtype, med, srclang**), spolu s jejich frekvencí (význam jednotlivých zkratk je k dispozici v **seznamu zkratk a kódů**)

Podobnou funkcionalitu, která se ovšem většinou vztahuje na celý korpus (nikoli pouze na jednu konkrétní konkordanci), poskytují také funkce **Nový dotaz** → **Seznam slov**.

Obsah

- Menu: Frekvence
- Vlastní nastavení frekvenční distribuce
- Frekvenční distribuce podle pozičních atributů
- Frekvenční distribuce podle metainformací
- Frekvenční seznam (výpis)

Celkem: 154 (4 str.)

	Filtr	lemma	Frekvence
1.	p/n	mail	18 839
2.	p/n	mailový	3 773
3.	p/n	mailovat	1 357
4.	p/n	majlant	785
5.	p/n	mej1	249
6.	p/n	mejlit	161
7.	p/n	mailbox	138
8.	p/n	mailing	129
9.	p/n	mailování	117
10.	p/n	mailer	102
11.	p/n	mejlka	59
12.	p/n	majland	58
13.	p/n	mailto	41
14.	p/n	mejlovat	22

Můžeme sledovat frekvence pravopisných variant. Pracujeme s korpusem psaných textů, které navíc prošly jazykovou kontrolou. Úzus, který není jednotný, se může stát zdrojem pro úvahy o kodifikaci, která má odrážet úzus.

Dvojice typu *Terinka/Terynka*

lemma

ter[iy]nka

popř. oblasti . Soutěž byla završena slavnostním menu (domácí	terinka /terinka/X@-----	, Cassoulette ze šneků s bramborovým pyré , konfitované stehýnko
Rozhodně nebudou litovat hosté , kteří nevynechají předkrm . Zaječří	terinka /terinka/X@-----	servírovaná s hruškovým pyré nebo variace domácích zvěřinových paštik na
minut + 1 hodina marinování + pečení 20 minut Zeleninová	terinka /terinka/X@-----	s estragonem NA 1/2L FORMU 40 g karotky , na
395 Kč . Můj tip je ze studených předkrmů jehněčí	terinka /terinka/X@-----	s olivami (195 Kč) , z teplých předkrmů
objeví například Apiciova jatýrka , což je podle jídelního lístku	terinka /terinka/X@-----	z kachních jater marinovaných v brandy , či Mořská pěna
různých pochutin . Návštěvníci tak mohli vidět jídla s názvy	terinka /terinka/X@-----	z mořského d'asa a toro z tuňáka , variace jehněčího
Jan Březina . A menu bylo opravdu vybrané : Králíčí	terinka /terinka/X@-----	s dýňovým semínkem jako předkrm a pak silný hovězí vývar
achní prsíčka s vlašským ořechem marinovaným v titinovém cukru ,	terinka /terinka/X@-----	z husích jater a uzených prsou v listovém těstě s

Není rozpoznáno (X), tvar na -ynka není doložen.

Je opravdu takový pravopis zdobnělin vlastních jmen možný?

Položme obecněji zadaný dotaz. Všechna podobná slova – užití regulárních výrazů.

Dotaz: Vyhledej všechna lemmata taková, že pro ně platí, že začínají libovolným znakem, za nímž následuje libovolný počet opakování libovolného znaku a končí na řetězec *inka* nebo *ynka*.

Hledat v korpusu

Korpus:

syn



Typ dotazu:

Lemma

Lemma:

.*[iy]nka

Slovní druh: nespecifikováno

.*[iy]nka


, Eltonem Johnem nebo Mary J. Blidge , a dvě	novinky /novinka/NNFP1-----A-----	. Reprezentativní výběr krátké čekání na nové album , které
ustašovského koncentračního tábora v Jasenovaci	Dinka /Dinka/NNMS1-----A-----	Šakiče (76) , byla 1 . února v
ednice Jičín - Městská knihovna přišla se zajímavou	novinkou /novinka/NNFS7-----A-----	. Na začátku prázdnin vyzvala čtenáře , aby ze svých
Vlašim - Praha , čeká od června podstatná změna .	Linka /linka/NNFS1-----A-----	by měla být prodloužena do Pacova . Po tomto datu
základem života . Tekutina čirá , s jíškou , se	zeleninkou /zeleninka/NNFS7-----A-----	Zastánci rybích polévek se dělí na dva nesmířitelné tábory :
rozdíly jsme se stali drahou destinací , " tvrdí Pavel	Hlinka /Hlinka/NNMS1-----A-----	. Česká ubytovací zařízení hledala nový způsob , jak přilákat
jejich dcera Chelsea s Marií a jejím novým přítelem	Fanyinky /fanyinka/NNFP4-----A-----	ječely před rytířem ! Princ FELIPE (28) ,
valuji odpadky První a poslední hudební nástroj Milé	maminky /maminka/NNFP5-----A-----	. Je zde další varování . Tentokrát se jedná o

Celkem: 2284 (46 str.)

	Filtr	lemma	Frekvence
1.	p/n	linka	276 255
2.	p/n	maminka	199 302
3.	p/n	novinka	194 274
4.	p/n	skupinka	52 575
5.	p/n	Hlinka	32 266

ve druhé třetině . " Rozhodující moment byla neproměnná šance	Barinky /Barinka/NNMS2-----A-----	. Mohli jsme jít do vedení 3:1 , " podotkl
. Hrál napravo . Nejdříve na led chodili Suchánek s	Barinkou /Barinka/NNMS7-----A-----	, potom Jindra se Slavíkem , znovu Suchánek s Barinkou
. Vyrovnat se podařilo až na začátku 8 . minuty	Barinkovi /Barinka/NNMS3-----A-----	, jehož vyběhl ke skórování nezištnou přihrávkou Erat . Zpět
podal dvě trestní oznámení . Kvůli podvodu při přestupu Petera	Barinky /Barinka/NNMS2-----A-----	do Třince , ale také kvůli dluhům vůči hráčům .
Nosek , Malinský - Rachůnek , Král , Melenovský -	Barinka /Barinka/NNMS1-----A-----	, Kovář , Pištek - Okál , M. Čech ,
sil , série byla velice vyrovnaná , " připomněl .	Barinka /Barinka/NNMS1-----A-----	patří mezi kandidáty slovenské reprezentace . Jenže ho trápí bolest

Přestože *Barinka* není hypokoristikum od Bára, může nám nalezení tohoto dokladu pomoci. Jak? Omezíme se na případy podobné, tedy na lemmata u nichž před *[iy]* předchází tzv. tvrdá souhláska *[rkh]*

Korpus: 

Typ dotazu:

Lemma:

.*[rkh][iy]nka

Broskve , 1 kg 39,90 CELKEM 304,50 ANDY Česká Lípa	Nektarinky /nektarinka/NNFP1-----A-----	, 1 kg 38,- Víno modré , 1 kg 53,-
v první a druhé útočné formaci elitního kanonýra Balaštica s	Barinkou /Barinka/NNMS7-----A-----	. Cítili , že je potřeba něco změnit . A
ina totiž marně hledala ztracenou fenku labradorského retrívra	Terinku /Terinka/NNFS4-----A-----	, která se zaběhla . Psí slečnu našla hodná paní
vstříc , " řekla &url; Bočanová . Od září má	Marinka /Marinka/NNMS1-----A-----	v lavici " tetu " - asistentku , která jí
obit vizor . Konečně dospěl k proraženému vstupu vyřiznutému	Katrínko /Katrínka/NNFS5-----A-----	. Osadníci už tam byli ve velké síle a splétali
do Brna uvolnil dalšího brankáře Kubeše . Z obrany hrají	Barinka /Barinka/NNMS1-----A-----	, Hanzl a Žilůva za juniory , ale pokud nastoupí
- já asi nijak . Nejlépe , myslím , obstála	Terinka /Terinka/NNFS1-----A-----	, její čistý , čistý hlásek . Začala mi hned

To už vypadá lépe. Ale najdeme vůbec nějaké relevantní doklady?

Celkem: 227 (5 str.)

	Filtr	lemma	Frekvence
1.	p/n	mandarinka	3 658
2.	p/n	Barinka	2 608
3.	p/n	Kuchynka	920
4.	p/n	nektarinka	817
5.	p/n	Wawrinka	657
6.	p/n	Márinka	592
7.	p/n	Darinka	370
8.	p/n	Drinka	220
9.	p/n	Terinka	214
10.	p/n	Sárinka	182
11.	p/n	Blahynka	165
12.	p/n	efedrinka	136
13.	p/n	Katrínka	133
14.	p/n	Machynka	118

a 15 . minutě likvidoval Lilko střely útočného hrotu domácích	Terynka /Terynka/X@ -----	. Poté hosté převzali iniciativu
výsledku . V 57 . minutě se dočkali a brankou	Terynka /Terynka/X@ -----	snížili na rozdíl jediného gólu .
, ale Bartůněk nastřelil pouze Soldána . Souš zahrozila hlavičkou	Terynka /Terynka/X@ -----	deset minut před poločasem .
třídy SK SIAD Bílina . Hosty porazil 3:0 dvěma góly	Terynka /Terynka/X@ -----	a brankou Schoře po poločas
frajer , víme ? " -- Manželka Antonie čili	Terynka /Terynka/X@ -----	jen poňukává . Celá tahle pat
míči se tak dostal Wagner , jenž pro svého kolegu	Terynka /Terynka/X@ -----	vymyslel příhrávku na míru . Te
dřevěným zábradlím , které odděluje divácká místa , sedí jeho	Terynka /Terynka/X@ -----	, celá v slzách , jimiž hodlá doj
, Tak poď , paní ! " -- Poslušná	Terynka /Terynka/X@ -----	vyskočí , celá šťastná se zavě
se znaky . Temperamentní a vhodný na zahradu . "	Terynka /Terynka/X@ -----	3 roky , fenka " Nekastrovaná t
první hrubka děčínské obrany a také první trest z kopačky	Terynka /Terynka/X@ -----	- 1:0 . Vyrovnat mohl Ženišek ,
: Eliška , Terynka a Jindřich Dvorec - Eliška ,	Terynka /Terynka/X@ -----	a Jindřich . Jmény postav z po
je šupa , že ho má zrovna Baník ! (Terynka /Terynka/X@ -----) - Tak skvělý fotbal jsem na B
Jílek , starosta Jankova i Holašovic Tygřici : Eliška ,	Terynka /Terynka/X@ -----	a Jindřich Dvorec - Eliška , Ter
Záleží ale také na tom , kdy půjdou Eliška ,	Terynka /Terynka/X@ -----	a Jindřich do světa , " dodal Ar
bude muset zkusit najít náhradu za elitního střelce soutěže Aleše	Terynka /Terynka/X@ -----	, o kterého má zájem rezerva f
nný byl zejména odchod našeho loňského nejlepšího střelce Aleše	Terynka /Terynka/X@ -----	. Soupiska tak byla doplněna c
Souš spoléhala . Ve třinácté minutě šli hosté po akci	Terynka /Terynka/X@ -----	s Berešem do vedení a domác

shromážďovala kolem sebe dívky , mezi nimiž si oblíbila zvláště	Márynku /Márynka/NNFS4 ----A----	Kropáčkovou , s níž si ještě po letech dopisovala .
bylinek . Jak zpracovávat ovčí vlnu ukázala lesní mateřská školka	Márynka /Márynka/NNFS1 ----A----	. Druhý den konference se učitelé všech stupňů pod vedením
Havelka (oba Maratonstav Úpice) 32:33 , 6 .	Márynka /Márynka/NNFS1 ----A----	(Ski Jilemnice) 32:40 , 8 . Novák Tomáš
" V srpnu už chceme být doma , abychom naši	Márynku /Márynka/NNFS4 ----A----	připravili na školu . Nejhorší bude ji naučit chodit brzo
M . D. Rettigové , adresovaný do Rychnova nad Kněžnou	Márynce /Márynka/NNFS3 ----A----	Kropáčkové (Litomyšl 12 . 8 . 1836) ,
zana Nováčková . Za největší ozdoby své sbírky malířka považuje	Márynku /Márynka/NNFS4 ----A----	. " Je to následovnice té , kterou vyšlechtil přímý
proti proudu a technika přejela do obce Hostomice k rybníku	Márynku /Márynka/NNFS1 ----A----	. Jednotka za pomoci člunu po proudu řeky zjišťovala ,
1 . Pala , 2 . Havelka , 3 .	Márynku /Márynka/NNFS1 ----A----	, muži do 49 let : 1 . Urbánek ,
(3,5 - 24 - 3,5) : 1 .	Márynku /Márynka/NNFS1 ----A----	(Ski Jilemnice) 1:09:25 . Po třech závodech je

Co nám říká IJP a co lze nalézt v korpusech

Domácké podoby rodných jmen (Klárinka, Lukí/Luky)

Užíváním domáckých podob vlastních jmen vyjadřujeme určitý citový vztah k nositeli jména, proto se při jejich tvoření uplatňují hlavně různé citově zabarvené jazykové prostředky, např. zdrobněliny. V běžném hovoru také často slyšíme oslovení jako Šárí, Štěpi, Tomí, Víki apod. Jsou to tvary domáckých jmen vzniklé původně jako zkrácená podoba **zdrobnělím utvořených příponou -ínka, -ička** nebo **-ík, -iček**: Šárinka – Šárí, Stázička/Stázi(n)ka – Stázi, Tomík/Tomíček – Tomí. Potřeba tvořit citově zabarvené podoby jmen je však natolik silná, že tvary na **-í** pronikly a nadále pronikají také ke jménům, která zdrobněliny s **-í/-i** v příponě obvykle netvoří, např. Věra – (Věruška, Věrunka) – Věri, Lukáš – (Lukášek) – Lukí.

Domácké podoby jmen se uplatňují téměř výhradně v soukromých, důvěrných projevech, a to především v mluvené podobě. V písmu se objevují jen v omezené míře – v soukromé komunikaci, v neformálních textech či v beletrii, kde signalizují vztahy mezi jednotlivými postavami. Při řešení jejich pravopisných podob lze vycházet ze způsobu tvoření nebo z hláskové stavby tvaru jména a dojit tak k různým podobám.

Vlivem analogie s typy, které mají **-i** ve zdrobnělích příponě (Stázička/Stázi(n)ka – Stázi, Tomík/Tomíček – Tomí), obvykle píšeme měkké **i** u všech domáckých podob zakončených na **-í**: Lukí, Víki, Šárí, Věri apod. Právě živé a frekventované zdrobněliny s **-i** v příponě jsou pro psaní Lukí, Šárí atd. silnou oporou, stejně jako přípony zveličující až hanlivé: Stázi(n)ka, Šárí(n)ka. Psaní s **-í** tedy považujeme za základní, a to bez ohledu na způsob tvoření a v některých případech i navzdory tomu, že koncovému **-i** předchází pravopisně tvrdá souhláska.

Předchází-li koncovému **-i** tvrdá souhláska, nemůžeme označit za chybné ani podoby s **-y**: Věri/Věry, Lukí/Luky atd. Předchází-li koncovému **-i** souhláska **d, t, n**, dává se přednost psaní **-i**, nebo **-y** podle výslovnosti předchozí souhlásky: Hanička – Hani [-ň-]. Vlasta – Vlastí [-t-], ale Rudolf – Rudy, Antonín – Tony. Méně obvyklé (ale přípustné) je psát též Rudí [rudý].

jedním z větších kousků dýňové úrody vyfocen i rodinný pejsek	Sárynka /Sárynka/X@ -----	. O bohatou úrodu se Manovi podělili i se sousedy
. Mám ráda televizi a koukám ráda na pohádky .	Sárynka /Sárynka/X@ -----	Kovářová Mně dají rodiče na zadek . Pak můžu koukat

Filtr konkordancí

Filtr: pozitivní negativní

Vybraný token: první poslední

Rozsah hledání: od 0 do 0 včetně KWIC

Typ dotazu: Lemma

klávesnice

Lemma: .*ynka

P-filtr

.*ynka

Celkem: 41 (1 str.)

	Filtr	lemma	Frekvence
1.	p/n	Kuchynka	920
2.	p/n	Blahynka	165
3.	p/n	Machynka	118
4.	p/n	Terynka	17
5.	p/n	Marynka	9
6.	p/n	Márynka	9
7.	p/n	marynka	7
8.	p/n	Štrynka	6
9.	p/n	Kachynka	5
10.	p/n	Horynka	5
11.	p/n	Merynka	4
12.	p/n	Barynka	4
13.	p/n	Térynka	3
14.	p/n	Terrynka	3
15.	p/n	Nohynka	3
16.	p/n	Hrynka	3
17.	p/n	Surynka	2
18.	p/n	rynka	2
19.	p/n	Korynka	2
20.	p/n	Sárynka	2
21.	p/n	Cherrynka	2
22.	p/n	Bechynka	2
23.	p/n	Jackynka	1
24.	p/n	Bukrynka	1
25.	p/n	sukynka	1

Závěr: Je patrné, že analýza korpusových dat může přinést některé poznatky odporující běžným poučkám a pravidlům. Analýza deminutivního sufixu *-inka* užívaného pro tvoření podob vlastních jmen prokázala rozkolísanost úzu, který je v rozporu s pravidly uvedenými v IJP.

Naučili jsme se:

Používat některé **regulární výrazy** v jazyce *cql*. (.*, [], |)

Používat **statistiky a zobrazení**.

Používat **filtry**.

Ukázali jsme, **jak lze používat jazykový korpus k hledání odpovědí na otázky týkající se grafické podoby jazyka**.

Doplňující četba:

OSOLSOBĚ, Klára a Jana MACHALOVÁ. Hypokoristika z rodných jmen v Korpusu soukromé korespondence. In Hladká, Zdeňka a kol.. *Soukromá korespondence jako lingvistický pramen*. Vyd. 1. Brno: Masarykova univerzita, 2013. s. 33-59.

14. 3. Dú:

1. Jaká jsou lemmata/tvary adaptovaných slov a jaké je frekvenční rozložení variantních způsobů jejich grafické realizace (slovesa utvořená od anglických základů: *použí(va)t Skype= ????*, *použí(va)t google = ???*, *použí(va)t twitter = ???*).

Uvědomte si, že nově přejatá slova nebudou ve slovníku automatické analýzy, a je tedy třeba na ní moc nespolehat.

2. V IJP se uvádí, že:

Časté chyby – výrazy nesprávně považované za spřežky

Jelikož se v následujících případech velmi často chybuje, připomínáme, že jen zvlášť píšeme zpodstatnělá přídavná jména, která se užívají pouze či převážně s předložkou *na* a pojí se se 4. p. Jsou to např. spojení: *na shledanou*, *na viděnou*, *na slyšenou*, *na rozloučenou*, *na odchodnou*, *na uvítanou*, *na zotavenou*, *na posílněnou*, *na pováženou*, *na uváženou*, *na rozmyšlenou*, *na vysvětlenou*, *na srozuměnou*, *na upřesněnou*, *na vybranou*, (*jako*) *na zavolanou*.

Dalším slovním spojením, které je často mylně chápáno, a proto psáno jako spřažený (jednoslovný) výraz, je *na rozdíl*. Také v tomto případě je nutné psát uvedené spojení jako slova dvě, respektive sekundární předložku *na rozdíl od* jako slova tři.

Zjistěte, jaký je stav dodržování uvedené pravopisné normy v korpusech řady SYN a popište, jak jsou nesprávně (dohromady) zapsané doklady interpretovány na rovině lemmatizace a tagování.

