



ČESKÝ NÁRODNÍ  
KORPUS

# Rozhraní KonText a akviziční korpusy

Lucie Chlumská  
ÚČNK, FF UK



# NAUČÍTE SE...

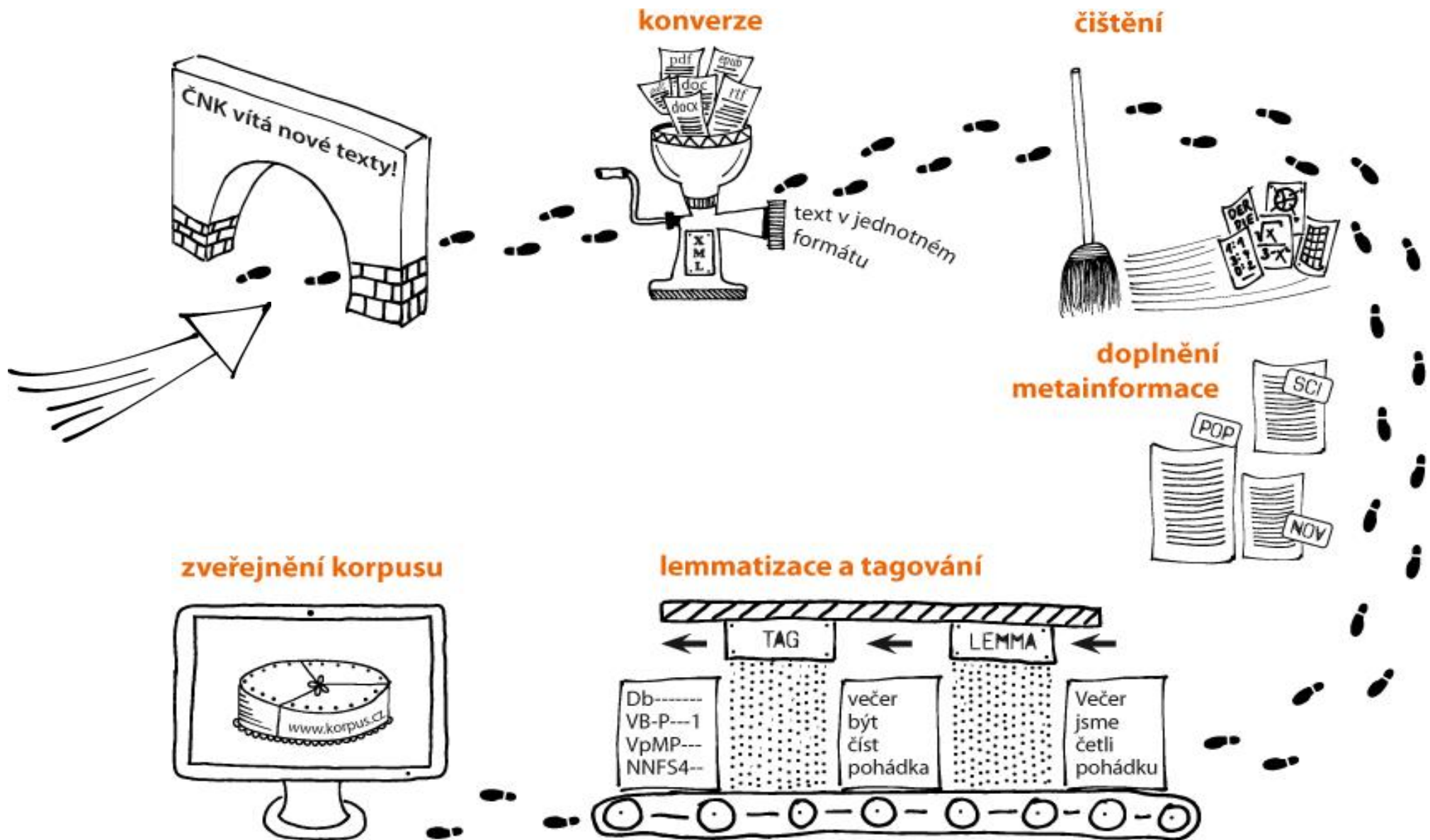
1. jak vypadá **lemmatizovaný a značkový korpus** a k čemu nám to při vyhledávání je
  - jak zkombinovat různé atributy při vyhledávání v **CQL**
2. jaký **druh informace** nám korpus může zprostředkovat
  - jakou anotací disponují akviziční korpusy CZESL, SCHOLA a SKRIPT
  - kde najdu informace o korpusu a jednotlivých textech
3. jak si vytvořit **subkorpus podle vlastních kritérií**
  - jak si ověřím, co mám ve svém korpusu (Seznam slov)



# JAK KORPUS VE SKUTEČNOSTI VYPADÁ?



# Zpracování textů v ČNK





# Korpus v podobě „vertikály“ v CZESL-sgt

```
<doc t_id="UJA2_PH_003" t_date="2010-12-21" t_medium="manuscript" t_limit_minutes="45" t_aid="none"
t_exam="yes|interim" t_limit_words="25" t_title="E-mail kamarádce/kamarádovi" t_topic_type="general"
t_activity="" t_topic_assigned="specified" t_genre_assigned="specified"
t_genre_predominant="informative" t_words_count="30" t_words_range="-50" s_id="UJA2_PH" s_sex="m"
s_age="17" s_age_cat="16-" s_L1="vi" s_L1_group="nIE" s_other_langs="" s_cz_SER="A1"
s_cz_in_family="" s_years_in_CzR="" s_study_cz="university"
s_study_cz_mesice="" s_study_cz_hrs_week="15-" s_textbook="NCSS" s_bilingual="no">
```

```
<s id="1">
```

```
mám mít VB-S---1P-AA--- mám mít VB-S---1P-AA---
dobře dobře Dg-----1A---- dobře dobře Dg-----1A----
. . Z:----- . . Z:-----
```

```
</s>
```

```
<s id="2">
```

```
V v RR--4----- V v RR--4-----
neděli neděle NNFS4-----A---- neděli neděle NNFS4-----A----
díval díval X@----- díval dívat VpYS---XR-AA--- S Quant0
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
na na RR--6----- na na RR--6-----
televizi televize NNFS6-----A---- televizi televize NNFS6-----A----
a a J^----- a a J^-----
uklízěl uklízěl X@----- uklízěl uklízet VpYS---XR-AA--- S Quant0|Caron1
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
. . Z:----- . . Z:-----
```

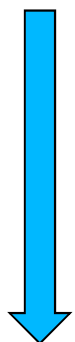
```
</s>
```

```
<s id="3">
```

```
Ano ano TT----- Ano ano TT-----
přijdu přijít VB-S---1P-AA--- přijdu přijít VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
tebe ty PP-S2--2----- tebe ty PP-S2--2-----
do do RR--2----- do do RR--2-----
kina kino NNNS2-----A---- kina kino NNNS2-----A----
```



# Korpus v podobě „vertikály“ v CZESL-sgt



word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@-----	Ten	ten	PDYS1-----	S	Quant1
pes	pes	NNMS1-----A----	pes	pes	NNMS1-----A----		
míluje	míluje	X@-----	miluje	milovat	VB-S---3P-AA---	S	Quant1
svého	svého	X@-----	svého	svůj	P8MS4-----	S	Voiced
kamarada	kamarada	X@-----	kamaráda	kamarád	NNMS4-----A----	S	Quant0
-	-	Z:-----	-	-	Z:-----		
člověka	člověk	NNMS2-----A----	člověka	člověk	NNMS4-----A----		
.	.	Z:-----	.	.	Z:-----		

Tabulka 1: Ukázka anotace jedné věty

- **word** – původní slovní tvar
- **lemma** – lemma původního tvaru, pokud tvar není rozpoznán, je lemma totožné s původním tvarem
- **tag** – slovnědruhov a morfologická značka původního tvaru, pokud tvar není rozpoznán, je uvedena značka pro neznámé slovo: X@-----
- **word1** – opravený tvar, pokud byl původní tvar vyhodnocen jako správný, je zde uveden tvar původní
- **lemma1** – lemma určené na základě slovního tvaru a jeho kontextu v opraveném textu
- **tag1** – slovnědruhov a morfologická značka, určená na základě slovního tvaru a jeho kontextu v opraveném textu
- **gs** – údaj o tom, zda případná chyba byla vyhodnocena jako pravopisná (S) nebo gramatická (G); gramatická chyba se obvykle vyznačuje tím, že původní slovní tvar byl rozpoznán
- **err** – typ chyby, určený na základě porovnání původního a opraveného tvaru, podrobný popis viz <http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1.html>.

# hledám podle typu chyby

word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@-----	Ten	ten	PDYS1-----	S	Quant1
pes	pes	NNMS1-----A----	pes	pes	NNMS1-----A----		
míluje	míluje	X@-----	miluje	milovat	VB-S---3P-AA---	S	Quant1
svého	svého	X@-----	svého	svůj	P8MS4-----	S	Voiced
kamarada	kamarada	X@-----	kamaráda	kamarád	NNMS4-----A----	S	Quant0
-	-	Z:-----	-	-	Z:-----		
člověka	člověk	NNMS2-----A----	člověka	člověk	NNMS4-----A----		
.	.	Z:-----	.	.	Z:-----		

Tabulka 1: Ukázka anotace jedné věty

Příklad: [err="Quant1"]

hledám všechna slova, kde pisatel udělal chybu v délce (prodloužil hlásku)



# hledám podle chybného a správného tvaru

word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@-----	Ten	ten	PDYS1-----	S	Quant1
pes	pes	NNMS1-----A-----	pes	pes	NNMS1-----A-----		
míluje	míluje	X@-----	miluje	milovat	VB-S---3P-AA---	S	Quant1
svého	svého	X@-----	svého	svůj	P8MS4-----	S	Voiced
kamarada	kamarada	X@-----	kamaráda	kamarád	NNMS4-----A-----	S	Quant0
-	-	Z:-----	-	-	Z:-----		
člověka	člověk	NNMS2-----A-----	člověka	člověk	NNMS4-----A-----		
.	.	Z:-----	.	.	Z:-----		

Tabulka 1: Ukázka anotace jedné věty

Příklad: [**word**=".+é.+" & **word1**=".+e.+" & **err**="Quant1"]

hledám **chybně zapsaná** slova s dlouhým **é**,  
kde má být správně krátké **e** (*tén* > *ten*)

# hledám podle chybného a správného tvaru

word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@-----	Ten	ten	PDYS1-----	S	Quant1
pes	pes	NNMS1-----A----	pes	pes	NNMS1-----A----		
míluje	míluje	X@-----	miluje	milovat	VB-S---3P-AA---	S	Quant1
svého	svého	X@-----	svého	svůj	P8MS4-----	S	Voiced
kamarada	kamarada	X@-----	kamaráda	kamarád	NNMS4-----A----	S	Quant0
-	-	Z:-----	-	-	Z:-----		
člověka	člověk	NNMS2-----A----	člověka	člověk	NNMS4-----A----		
.	.	Z:-----	.	.	Z:-----		

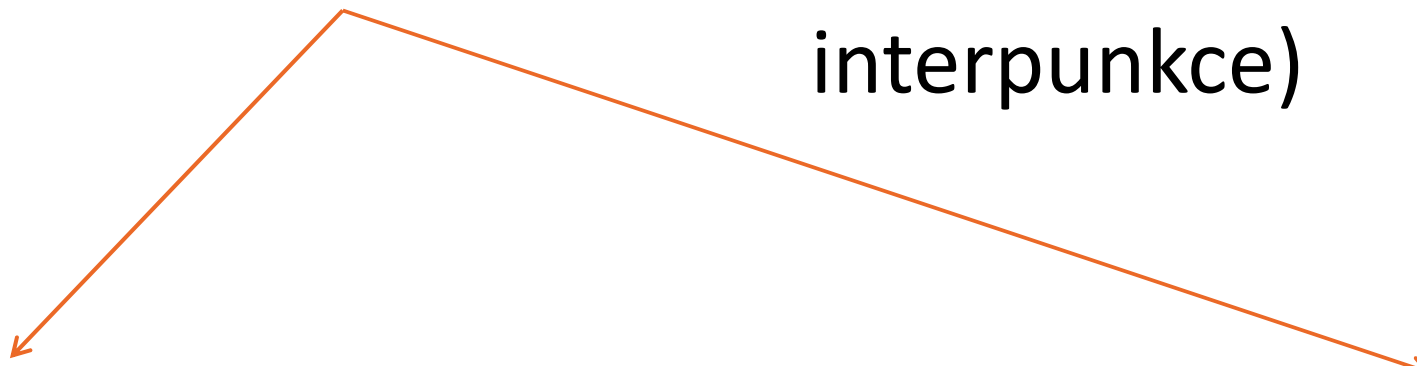
Tabulka 1: Ukázka anotace jedné věty

Příklad: [**word**="s.+" & **word1**="z.+" ]

hledám slova začínající na **z**, ale  
chybně napsaná se **s** na začátku

# dotaz na shodu podmětu s přísudkem

Příklad: hledáme všechna feminina v 1. os. pl. ve větě se slovesem v minulém čase, které chybně končí na „i“ (s vyloučením muž. živ., spojek a interpunkce)



```
<s>[tag!="(Z.* | N.M.1.* | J.*)"]*[tag="N.FP1.*"] [tag!="(Z.* | N.M.1.* | J.*)"]*[tag="V..P....R.*" & word=".+i"]  
within <s/>
```



# Využití regulárních výrazů

nejpoužívanější:

- .**\*** libovolný počet libovolných znaků ( $s.\text{*}$  > *s, se, sám, sirka...*)
- .**+** minimálně jeden libovolný znak ( $s.\text{+}$  > *se, sám, sirka...*)
- ?** předchozí znak tam může nebo nemusí být

Příklady:

- |                        |  |
|------------------------|--|
| .....                  | všechna pětispísmenná slova                  |
| . <b>+</b> nést        | všechny předponové varianty od <i>nést</i>   |
| mega. <b>+</b>         | všechna slova začínající na <i>mega-</i>     |
| . <b>*</b> a. <b>*</b> | slova, která někde obsahují písmeno <i>a</i> |
| v <b>?</b> okno        | slovo <i>okno</i> nebo <i>vokno</i>          |





# Využití regulárních výrazů

kde najdu další informace?

[http://wiki.korpus.cz/doku.php/kurz:regularni\\_vyrazy](http://wiki.korpus.cz/doku.php/kurz:regularni_vyrazy)

[http://wiki.korpus.cz/doku.php/pojmy:regularni\\_vyrazy](http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy)



 JAKÝ DRUH INFORMACE V AK NAJDEME?



# Zásadní pravidlo: poznej svůj korpus!

- každý korpus může být **jinak anotovaný**: jiné značky a jiné informace
- před vyhledáváním je třeba zjistit o korpusu co nejvíc
- kde zjistím, co v mém korpusu je a není a jak je značkován?

<http://wiki.korpus.cz/doku.php/cnk:uvod>

wiki.korpus.cz > Korpusy ČNK > SKRIPT...



# Jak si ověřit značkování přímo v KonTextu

- pokud si nejsme jistí, jak je nějaký jev značkován, je dobré si jej **vyhledat na příkladu konkrétního slova** a ověřit si to

Příklad: jak se značí chyba v délce?

- najdu nějaké chybně zapsané slovo, např. *tén* > zobrazím si atribut **err** ve funkci Frekvence > Vlastní nebo Zobrazení > Korpusová nastavení





# Jak si sumarizovat vyhledané výsledky

- po vyhledání vidíme výsledek v podobě konkordance, kterou lze procházet, třídit a dělat z ní vzorky
1. pokud chceme souhrn toho, co jsme našli, potřebujeme funkci **Frekvence** (4 „zrychlené“ volby a volba Vlastní)
  2. pokud rovnou chceme souhrn informací a nechceme nic konkrétního hledat, můžeme rovnou využít funkci **Seznam slov**

Příklad: jaké jsou ty nejčastější typy chyb v korpusu CZESL-sgt?

- Dotaz > Seznam slov > Hledat podle atributu: **err**





# JAK SI VYTVOŘIT SUBKORPUS?



# Tvorba subkorpusu

The screenshot shows the KonText web application interface. At the top, there is a navigation bar with icons and links for 'KonText', 'SyD', 'Morfo', 'KWords', 'Treq', 'SkE', 'Wiki', 'Podpora', and 'Biblio'. Below this, the 'kon text' logo is visible on the left, and a menu of options is on the right: 'Dotaz', 'Korpusy', 'Uložit', 'Konkordance', 'Filtr', 'Frekvence', 'Koklokace', 'Zobrazení', and 'Nápověda'. The 'Korpusy' option is highlighted, and a dropdown menu is open, showing 'Dostupné korpusy...', 'Mé subkorpusy...', and 'Vytvořit nový subkorpus...'. Below the menu, the current corpus is identified as 'syn2015'. The main search area is titled 'Hledat v korpusu' and contains several input fields: 'Korpus:' with a dropdown set to 'syn2015' and a sub-corpus dropdown set to '--celý korpus--'; 'Typ dotazu:' with a dropdown set to 'Lemma'; a 'Lemma:' input field with a dropdown menu open showing 'Hláskova' and 'přetřesení hlásky'; and 'Slovní druh:' with a dropdown set to 'nespecifikováno'. There are two checkboxes: 'Specifikovat kontext' and 'Specifikovat dotaz podle metainformací', both of which are currently unchecked. A blue 'Hledat' button is located at the bottom left of the search area.

# Informace o textech v CZESL-sgt

<http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-cs.pdf>





Máte jakékoli dotazy?

Sem s nimi!

Děkuji Vám za pozornost!

[lucie.chlumska@korpus.cz](mailto:lucie.chlumska@korpus.cz)

