Úvod do korpusové lingvistiky 1

KORPUS V MODERNÍM SLOVA SMYSLU A BUDOVÁNÍ KORPUSŮ

Není korpus jako korpus
https://cs.wikipedia.org/wiki/Korpus

uchovaným kouskem těsta jako záplatou . 5 . Zatímco se	**korpus**	peče , dejte rybu do kastrolu s mlékem a bobkovým
navrstvit na tvarohový koláč , do košíčků nebo na pusinkové	**korpusy**	či do jiných drobných sladkostí . Hodí se i jako
g mascarpone Pusťte se do toho : 1 Začneme přípravou	**korpusu**	. Sušenky vsypeme do pevného mikrotenového sáčku a paličkou na
nutné dál diskutovat , především i na základě materiálu z	**korpusů**	, ze záznamů hovoru , z literatury aj . Jen
až ke stropu . Zařízení kuchyně je kombinace lesklých bílých	**korpusů**	(Thermopal) , šedé pracovní desky (Fundermax s
pomůžu , až se vrátíš . “ „ Budu vyrábět	**korpus**	na dort , “ oznámila jsem mu a uvelebila se
radnice ve městě , neoklasicistický El Templete . Míjeli dlouhý	**korpus**	semináře svatých Karla a Ambrosia . Na vydlážděném Náměstí katedrály
ochuceného jogurtu a důkladně rozmícháme . Krém nalijeme na piškotové	**korpusy**	do formiček a necháme v chladu minimálně několik hodin ztuhnout
, že se ta fotka objeví v Perezu . Vyrábět	**korpus**	– to je prostě rozkoš . “ Zasmál se a
20 kHz . Jiným způsobem porovnávali spektrální sklon na velkém	**korpusu**	Kochanski , Grabeová , Coleman a Rosner (2005)
a cukrem dotuha a ochutíme likérem . Krém rozetřeme na	**korpus**	a stočíme . Hotovou roládu zabalíme do alobalu a uložíme
, přidáme moučkový cukr a lehce vyšleháme . 5 .	**Korpus**	prokrojíme . Na spodní díl nastříkáme 2 / 3 krému
jako pedagogicky zaměřený korpus . Lze ji zařadit mezi akviziční	**korpusy**	, které zachycují procesy a různé fáze osvojování určitého jazyka
podle pamětníků v průběhu 20 . století . Kříž s	**korpusem**	Ježíšova těla , dřevěná skulptura Vítězného Krista s praporcem a
je teď . KAPITOLA DVACÁTÁ Po čtyřech hodinách , pěti	**korpusech**	a šesti orgasmech jsem posadila svého Briťáka do jeho nového
Kakaová roláda s chilli Rozpis na 12 porcí : NA	**KORPUS**	: ! 4 vejce ! špetka soli ! 120 g
na pravoboku od něj . Bylo to podivuhodný . Celej	**korpus**	třídy zprava pokukoval po Karolíně . A ona pokukovala doleva

Myšlenka korpusu ve filologii a lingvistice

- V širším slova smyslu soubor textů
- Sbírka textů
- Korpus v moderním slova smyslu

Zdroje poznání fungování jazyka

- Introspekce (pozorování vlastní jazykové produkce za účelem zjistit, jak funguje jazyk);
- Elicitace (navození situace, v níž dochází k produkci textů, z nichž chceme získat jazykový materiál, který má osvětlit fungování jazyka);
- Pozorování textů vzniklých spontánně, když lidé mluví, píší;

Introspekce

2.3.3 Introspekce

Konečně se jako přijatelná forma poznávání jazyka jeví i **introspekce**, tedy sebepozorování, sledování toho, jak sami užíváme jazyka a jaké postoje k jeho různým složkám zaujímáme (jaké je naše vlastní jazykové povědomí). Při introspekci ovšem pochopitelně vždy hrozí nebezpečí subjektivního zkreslení, proto se na ni nelze spoléhat jako na jediné východisko poznatků o jazyce.

Elicitace

2.3.2 Elicitace

Dalším důležitým prostředkem pro získávání jazykových dat je tzv. **elicitace** (z latinského *ēlicere* ‚vylákat, vypátrat'). Elicitace v zásadě znamená cílené získávání informací od mluvčích či pisatelů (obecně respondentů). Při elicitaci se pracuje s rozmanitými anketami, dotazníky, řízenými rozhovory, psaním textů na stanovené téma, resp. za stanovených podmínek apod. Rozlišují se dva základní druhy elicitace. Uplatňuje-li se experimentální elicitace, vystupuje do popředí řízení projevu respondenta se zřetelem k formě sdělení (např. otázka: „Použili byste ve vyjádření oficiálního rázu tvar *cyklisti*, nebo pokládáte za náležitý jen tvar *cyklisté*"?). Naproti tomu klinická elicitace je volnější, umožňuje spontánnější vyjádření a vyzdvihuje spíše obsah sdělení (např. úkol: „Napište do novin čtenářský dopis, v němž pojednáte o postavení cyklistů ve velkoměstské dopravě"; zde pak máme naději, že ze vzniklého textu bude patrný i respondentův postoj k tvarům *cyklisté/cyklisti*).

Texty

2.3 Získávání jazykových dat

2.3.1 Texty

Fundamentální pramen pro získávání jazykových dat představují existující doklady užívání jazyka v komunikaci, tedy texty, a to jak texty psané, tak texty mluvené (jejich efektivní využití ovšem bylo možné, až když se objevily a rozšířily technické prostředky pro záznam mluveného vyjadřování). Předpokladem pro poznávání jazyka je tedy shromažďování textů, jejich popis a porovnávání a dále vyčleňování složek textů a třídění těchto složek.

Od excerpce ke korpusu

a) Tradiční postup představuje tzv. **excerpce**, tedy manuální (ruční) příprava excerpt (výpisků) z textů jako dokladů výskytu slov, slovních tvarů, syntaktických konstrukcí apod. Výpisky na excerpčních lístcích jsou tříděny a ukládány v kartotékách. Např. Ústav pro jazyk český AV ČR uchovává přibližně třináct milionů excerpčních lístků, které sloužily pro přípravu výkladových slovníků češtiny. Jako příklad zde na s. 19 uvádíme tři excerpční lístky, které byly podkladem pro zpracování hesla *sen* v *Příručním slovníku jazyka českého* (1935–1957).

b) Od sklonku dvacátého století se jako základní pramen pro získávání jazykových dat z textů prosazují **jazykové korpusy** a rozvíjí se korpusová lingvistika jako disciplína, která se zabývá principy vytváření jazykových korpusů i jejich praktickým sestavováním (první korpus angličtiny začal vznikat už v šedesátých letech). František Čermák definuje jazykový korpus takto:

Definice korpusu

Jazykový korpus je strukturovaný, unifikovaný a často i označkovaný (tagovaný) velmi rozsáhlý soubor jazykových dat, elektronicky uložený a zpracovávaný (obv. v podobě úhrnu jednotlivých textů), jejichž výběr chce být vzhledem k vytčenému cíli reprezentativní, často bývá i lemmatizovaný (Čermák, 2011: 108; zvýraznil Čermák).[7]

Definice korpusu v moderním slova smyslu

- vzorky (sampling) a reprezentativnost
- konečná velikost (omezený a vymezený rozsah)
- strojově čitelná forma (MRF)
- standardní reference

Reprezentativnost korpusu

- Texty mají reprezentovat jazyk, a to buď obecně v jeho různých podobách (psané/mluvené), nebo speciálně (např. žánrově vymezené korpusy, autorské korpusy, žákovské korpusy).
- Vzorky – z textů, z nichž se skládá korpus, se vybírá vzorek (reprezentativní část textu), nebo je text zařazen do korpusu jako celek.
- Vzorkování – proč a jak

Velikost korpusu

- Vymezený obsah i rozsah (kdy je možné/nutné upřednostni kvantitu a kdy je třeba brát v úvahu zejména kvalitu)
- Rozsah psaných a mluvených korpusů s ohledem na žánr
- Rozsah a obsah autorských korpusů (průnik korpusové lingvistiky a filologie)
- Rozsah a obsah specializovaných korpusů

Strojově čitelná a přístupná podoba

- Konverze textů existujících ve strojově čitelné podobě do jednotného formátu
- Převedení textů, které neexistují ve strojově čitelné podobě
- OCR metody
- Ruční přepis
- Budování pravidel pro ruční přepis jako metodologie
- Otázka transliterace a transkripce a dalších edičních strategií

Standardní reference

- Vnětextové značkování - metadata
- Vnitrotextové značkování – popis jazykových jednotek, z nichž je text složen
- struktury
- tokeny
- různé typy jazykových značek

Budování korpusu

- Specifikace cíle a účelu korpusu – proč korpus chceme
- specifikace cílové skupiny uživatelů – kdo ho bude používat?
- výběr a sběr jazykového materiálu (texty, nahrávky + přepisy, přepisy dat, které nejsou v el. podobě, přepisy dat související s verzemi rukopisných i starších tištěných památek)
- autorská práva – smlouvy s dodavateli textů, u mluvených korpusů informovaný souhlas mluvčích a anonymizace osobních
- zpracování textů – vertikál, tokenizace, jednotné kódování a jednotný formát (konverzní programy)
- vnitřní a vnější značkování (atributy, metadata, tagging)
- Zajištění kvalitních/kvalifikovaných anotátorů (programy, školení anotátoři)
- Zajištění nástrojů pro přístup a využití korpusů

Struktura korpusu

https://wiki.korpus.cz/doku.php/:pojmy:struktura_korpusu

- Korpus - jako soubor textů - je vnitřně strukturován do různých celků. Jednotlivé celky v rámci korpusu se nazývají **strukturní jednotky** (jako opus, dokument, věta), k nimž se vážou různé strukturní atributy (např. autor, název díla, rok vydání apod.). Zároveň je většina korpusů opatřena dodanou lingvistickou informací, která se týká jednotlivých slov (tj. pozičními atributy, jako třeba lemma, tag apod.).
- Za účelem zachycení takovéto mnohovrstevnaté struktury se užívají značkovací jazyky. Standardem v této oblasti je formát XML, často se ovšem používají i různé formy jazyka SGML

Vertikála – korpusy psaného jazyka

- Vertikála je interní formát sloužící pro zachycení struktury korpusu a textů v něm (spolu s jejich anotací).

Příklad


```
<opus autor="Doyle, Arthur Conan" nazev="Příběhy Sherlocka Holmese" nakladatel="Mladá fronta" mistovyd="Praha"
rokvyd="1971" isbnissn="" preklad="Henzl, V. - Zábrana, J. - Wolfová, Z." srclang="ENG" txtype_group="beletrie"
txtype="NOV" genre="CRM" med="B" id="pribshho">
<doc id="1">
...
<s id="10">
Když    když    J,--------------
školení školení NNNS4-----A-----
skončilo        skončit VpNS---3R-AA---P
,       ,       Z:--------------
přidělili       přidělit        VpMP---3R-AA---P
mne     já      PP-S4--1--------
k       k       RR--3-----------
Pátému  Pátý    NNMS3-----A-----
northumberlandskému     northumberlandský       AAIS3----1A-----
střeleckému     střelecký       AAIS3----1A-----
pluku   pluk    NNIS3-----A-----
jako    jako    J,--------------
pomocného       pomocný AAMS4----1A-----
chirurga        chirurg NNMS4-----A-----
.       .       Z:--------------
</s>
...
</doc>
...
</opus>
<opus>
...
</opus>
```

Hlavní zásady anotační praxe

- Anotační schéma by mělo vycházet z teoretických východisek, která by měla být jasně formulovaná a přístupná každému konečnému uživateli korpusu. Mnohé korpusy byly anotovány ručně (existence subjektivních interpretací zaviněných osobou anotátora ve sporných případech). Značkování by pak mělo být doplněno komentáři, z nichž by byl důvod příslušné volby patrný.

Co má uživatel korpusu vědět o anotaci, chce-li ji použít

- Mělo by být jasné JAK a KDO anotaci provedl (JAK – ručně x automaticky x poloautomaticky, s postkorekcí x bez korekce) (KDO – počítačový program, anotátor - člověk)
- Uživatel korpusu by si měl být vědom toho, že anotace nejsou nějakou nedotknutelnou neomylnou instancí. Anotace je pouze více či méně užitečným nástrojem. INTERPRETACE.
- Anotační schéma by mělo být založeno na široce schvalovaných a teoreticky nezatížených principech. Není na škodu i zjednodušující přístup.
- Žádné anotační schéma nemá právo být pokládáno za standardní. Je-li nějaké řešení uznávanější, děje se tak pouze z praktických důvodů.

Příklad otázek v testu

- V čem spočívají základní metody poznávání fungování jazyka?
- Jaký je rozdíl mezi sbírkou textů a korpusem v moderním slova smyslu?
- Je počet slov v korpusu objektivním měřítkem pro hodnocení jeho kvality?
- Proč nelze užívat texty na internetu týmž způsobem jakým se využívají jazykové korpusy?
- Co je to vertikála?
- Jaké jsou čtyři hlavní rysy korpusu v moderním slova smyslu.
- Jmenuj nějaké typy strukturních značek.
- Co je to OCR?
- Vysvětli rozdíl mezi transkripcí a transliterací.
- Proč nejsou mluvené korpusy přepsány fonetickou transkripcí?