KORPUSOVÁ LINGVISTIKA

ÚVOD A ZÁKLADNÍ POJMY

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Organizace

- **CJBB105 Korpusová lingvistika – přednáška**
- **PRZA009 Korpusová lingvistika**
- *Počítačová lingvistika, Český jazyk a literatura*
- *Překladatelství moderních evropských jazyků*
- *Digitální lingvistika (FI)*

- přednáška, částečně praktické ukázky
- prezentace ve studijních materiálech v ISu
- **zakončení** – test v ISu (volné odpovědi)

- CJBB75 Základy využití korpusů (pro praxi)
- CJBB84 Morfologie a korpus
- PLIN032 Gramatika a korpus
- Korpusový workshop v Praze (+ online)

Osnova

- Úvod – korpus a korpusová lingvistika, základní pojmy
- Vývoj korpusové lingvistiky
- Typologie korpusů, české korpusy (ČNK)
- Budování korpusů, reprezentativnost
- Korpusové manažery
- Morfologické a syntaktické značkování
- Využívání korpusů, korpusové aplikace
- Časopisy, konference, publikace, organizace
- *Praktická část*

Doporučená literatura

- Stefanowitsch, A. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7). Berlin: Language Science Press, 2020.
- Čermák, F. *Korpus a korpusová lingvistika*. Praha: Nakladatelství Karolinum, 2017.
- McEnery, T. and Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.
- *Studie z korpusové lingvistiky*. Čermák, F., Klímová, J. a Petkevič, V. (eds.). Praha: Karolinum, 2000.
- Kennedy, G. *An Introduction to Corpus Linguistics.* London, New York: Routlege, 1998 (hardback 2016).
- McEnery, T. and Wilson, A. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 1996.

Doporučené odkazy

- http://korpus.cz – Český národní korpus
- http://wiki.korpus.cz – výklad termínů
- NESČ https://www.czechency.org/ – výklad termínů

- **Ústav Českého národního korpusu FF UK**
- Ústav formální a aplikované lingvistiky MFF UK
- Ústav pro jazyk český AV ČR
- Centrum zpracování přirozeného jazyka FI MU
- Ústav českého jazyka FF MU
- Lexical Computing CZ, s. r. o.

Pracoviště v ČR

Evropská centra

- **Centre for English Corpus Linguistics**, UCL Lovaň
- **Centre for Corpus Research**, University of Birmingham
- **Programmbereich Korpuslinguistik** (Corpus Linguistics Programme Area), Leibniz-Institut für Deutsche Sprache (IDS), Mannheim

Korpusová lingvistika

- vědní obor **– vymezení v systému věd**
- průnik **humanitních** (lingvistika) a **přírodních** (matematika, informatika) věd
  - studium **přirozeného** jazyka s využitím **metod a nástrojů** přírodních věd
- dostatečné množství **autentických** jazykových dat
- empirie, observace (x introspekce)
- objektivita a evidence
- **opakovaný experiment**
- **hardware a software/webové rozhraní**

Korpusová lingvistika

- **vymezení v kontextu NLP**
- lingvistika, matematika a informatika
- počítačová lingvistika / počítačové zpracování přirozeného jazyka (***N****atural* ***L****anguage* ***P****rocessing*)
  - korpusová lingvistika je **podmnožinou**
- korpusová lingvistika – stojí **samostatně**
- **vymezení v kontextu lingvistiky**
- **samostatný obor**
  - přístup **corpus-driven**, výzkum korpusem řízený
  - reformulování introspekcí stanovené hypotézy
- **metodologie** pro všechny části lingvistiky
  - přístup **corpus-based**, výzkum korpusem ověřovaný
  - exemplifikace hypotézy, hledání dokladů
- **poskytuje zdroj jazykových dat**

Co je to korpus

Jazykový korpus (z lat. *corpus* „tělo, těleso“) je **rozsáhlý** soubor **autentických textů** (psaných nebo mluvených) převedený do **elektronické podoby** v jednotném formátu tak, aby v něm bylo možné jednoduše **vyhledávat** jazykové jevy, zejména slova a slovní spojení. Korpus zobrazuje jazykové jevy v jejich **přirozeném kontextu**, a umožňuje tak vytvářet na reálných datech podložený jazykový výzkum v rozsahu, který byl dříve nemyslitelný.

http://wiki.korpus.cz/doku.php/pojmy:korpus

Co je to korpus

Rozsáhlý soubor elektronicky uložených jazykových dat, obvykle označkovaný, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž je také považován za reprezentativní.

Čermák, F. Jazykový korpus: Prostředek a zdroj poznání. In *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 15–38.

Elektronický text v korpusu

- **lineární řetězec znaků, mezi kterými se vyskytují mezery** (znak, řetězec, mezera)
  - psaný a mluvený
- jednotný **kód** (*Unicode – UTF-8*) a **formát** (*txt*)
- **autentičnost** – data se neupravují, korpus je **deskriptivní**
  - „Korpusová data jsou posvátná.“ (F. Čermák)
- etický kodex
- autorská práva

Co je to korpus – shrnutí

- rozsáhlý elektronický soubor textů
- **autentické** texty, výskyt slova v přirozeném kontextu – **konkordance**
- **sjednocené texty**
  - strojově čitelný formát, machine readable format/MRF
  - jednotné kódování a formát
- **označkovaná data** (přidané informace)

Korpusové zpracování textu

- *Pro představu, jakým přívětivým místem byl Americký park v minulosti, uvádíme několik historických fotografií.*

- **vertikál** (vertikální text, v souboru s příponou .vert)
- **token** (**tokenizace**)
  - řetězec znaků ohraničený z obou stran mezerami
- **type** (word, slovní tvar)
- **strukturní značky**
  - **s** = sentence (věta)
  - **g** = glue (spojení následujícího znaku s předchozím tokenem)

2	<s>
3	Pro
4	představu
5	<g/>
6	,
7	jakým
8	přívětivým
9	místem
10	byl
11	Americký
12	park
13	v
14	minulosti
15	<g/>
16	,
17	uvádíme
18	několik
19	historických
20	fotografií
21	<g/>
22	.
23	</s>

Korpusové zpracování textu

- **token – type**
  - **token-type ratio (TTR)**
  - vysoké číslo = bohatost slovníku
  - nízké číslo = velké opakování slov
- **velikost korpusu**
  - počet slov (type)
  - počet tokenů (vyšší číslo)
- pro uživatele – **korpusové manažery**
- **konkordance**, **KWIC** (Key Word in Context)

*náš rozmazlený **kocour** zase spal v posteli*

" Podle právníků však tímto způsobem studenti hrubě porušují nejen	studijní	povinnosti , ale hlavně i zákon . Názory na to
po celý tento týden uzavřena veškerá pracoviště v hlavní budově	Studijní	a vědecké knihovny v Plzni . Od příštího pondělí se
vysokou školu . " Vše je možné objednat . Kdyby	studijní	knihy vyprodal nakladatel , dají se přetáhnout z jiných obchodů
zkoušky uchazečům prominuty . Tradičně největší zájem byl o bakalářské	studijní	obory Sociální práce , Tělesná výchova a sport a Ekonomická
že jim na ně finančně přispěje a umožní jim čerpat	studijní	volno - dá jim perspektivu a zaváže si je i
. " " Budeš se vzdělávat . Vypadá to na	studijní	pobyt . . . . " " Mně nikdy nic
povinností vyplývajících z výkonu svěřené funkce , obdobně i porušení	studijní	kázně a další . Jedná se o širokou kategorii deliktů
nejvyšší počet dětí , žáků nebo studentů ve třídě ,	studijní	skupině nebo oddělení v příslušném oboru vzdělání ve škole nebo
mluvčí meziříčské radnice Práva otevřou doktorandské studium OLOMOUC - Doktorandský	studijní	program otevře s největší pravděpodobností už letos na podzim Právnická
studium . " Volného času drobná blondýnka příliš nemá .	Studijní	povinnosti a mimoškolní aktivity jí prý zabírají všechen čas .
. " Loni byla Veronika se spolubydlící Katkou za dobré	studijní	výsledky v Bruselu , kam ji europoslankyně Jana Bobošíková pozvala
jsem ráda , že sportuje , protože jinak byl vyloženě	studijní	typ , " vzpomíná matka Jarmila Skopová . Při přecházení
, ale všechno mě baví . Chci požádat o individuální	studijní	plán a doufám , že to zvládnu , " věří
škol v americkém stylu ? Nekompromisně srovnávajícím kvalitu profesorů ,	studijní	plány i kariéry absolventů . Na přístupovém heslu k němu
tabu , po válce až donedávna se veřejně , mimo	studijní	účely , nepromítaly . To Riefenstahlové na druhé straně nebránilo
zkratce " Bc. " uváděné před jménem) . Magisterský	studijní	program je zaměřen na získání teoretických poznatků založených na soudobém
vymezování relevantního trhu značně subjektivní . 6 Zneužití dominantního postavení	Studijní	cíle Cílem této kapitoly je objasnit samotný pojem dominantní postavení
v přírodě . Ty potřeboval pořídit ke zdárnému splnění účelu	studijní	cesty asistenta v oblasti výskytu vzácné přímořské flóry během jeho
, zda Jirka během svých studií uzavřel vůbec nějakou dílčí	studijní	etapu zkouškou . Vím jen , že v době ,
i moderně vybavená kolej pro studenty a studentská jídelna .	Studijní	obory Stěžejní obor Charitní a sociální činnost je určen zájemcům
že když se náš tělocvikář zlískal a utekl s vedoucí	studijní	poradny , přivedli jsme oba nazpátek . Naši kolegové splnili
. Myslím si , že cestování , zahraniční stáže a	studijní	a pracovní pobyty jsou určitě právě o tom , aby
dovolenou , i když nezvyklou . " Zavolejte děkanovi pro	studijní	záležitosti , pane . . . Wedde . Já teď
ho v té jeskyni . Kdysi jsem dostal na fakultě	studijní	volno a ponořil jsem se do období osídlení Islandu pod

Obsah a rozsah korpusu

- **typ komunikace** – korpusy psané, mluvené, multimodální
- obsah – typy textů
  - **beletrie, odborné texty, publicistické texty**
  - texty z internetu
  - soukromá korespondence
  - přepisy mluvených nahrávek
  - texty zahraničních studentů češtiny (žákovské korpusy)
- vyváženost (poměr kategorií)

Obsah a rozsah korpusu

- **rozsah** – velikost korpusu
  - počet tokenů
  - počet slovních tvarů (type, word)
- **opravdu velké korpusy** (webové i klasické korpusy – několik miliard pozic)
  - frekvenční studie
- **malé specializované korpusy** (stovky tisíc pozic, jednotky milionů)

Obsah a rozsah korpusu

- **celé texty**
- **vzorky** (sampling) – vybraná část textu
- **rozsah**
  - vymezený rozsah, uzavřený (předem stanoven) - **referenční**
  - otevřený/monitorovací korpus (plynule se zvětšuje) – **nereferenční**
  - korpus, který se pravidelně obnovuje a zvětšuje – **verzovaný** (verze se číslují)

Značkování korpusu

- **značkování** – zvyšuje informační hodnotu korpusu (vždy nutná dostupná interpretace značek = tagset)
- **vnitřní značkování** (vnitrotextové)
  - strukturní atributy (*doc, text, p, s*)
  - poziční atributy (***word***, ***lemma*** *(sublemma)*, ***tag*** *(verbtag)*)
  - morfologické značky
- **vnější značkování**, (vnětextové)
  - na úrovni textu, **metatextové** informace (*autor, název díla, rok vydání atd.*)

Hlavní rysy korpusu

- *aneb čím se korpus liší od webu nebo elektronického archivu*

1. **elektronické autentické texty v jednotném formátu**
2. **značkování**
3. **zobrazení konkordancí v korpusových manažerech**
4. **vymezený obsah a rozsah**