

Korpusová lingvistika – 1

Úvod – korpus a korpusová lingvistika,
základní pojmy

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Organizace

- **CJBB105 Korpusová lingvistika – přednáška**
- **PRZA009 Korpusová lingvistika**
- *Počítačová lingvistika, Český jazyk a literatura*
- *Překladatelství románských a germánských jazyků*

- přednáška, částečně praktické ukázky
- prezentace z přednášek ve studijních materiálech IS
- **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
- Typy korpusů, české korpusy (ČNK)
- Budování korpusů, reprezentativnost
- Korpusové manažery
- Morfologické a syntaktické značkování
- Využívání korpusů, korpusové aplikace
- Časopisy, konference, publikace
- *Praktická část*

Doporučená literatura

- Čermák, F. *Korpus a korpusová lingvistika*. Praha: Nakladatelství Karolinum, 2017.
- *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: Routledge, 1998 (hardback 2016).
- McEnery, T. and Wilson, A. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 1996.
- McEnery, T. and Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.
- <http://korpus.cz> – Český národní korpus
- <http://wiki.korpus.cz> – výklad termínů
- NESČ <https://www.czechency.org/>

Instituce v ČR

- **Ústav Českého národního korpusu FF UK**
- Ústav teoretické a počítačové lingvistiky 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

Korpusová lingvistika

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

Korpusová lingvistika

- **užší vymezení – vztah k NLP**
- lingvistika, matematika a informatika
- **počítačová lingvistika**
 - počítačové zpracování přirozeného jazyka (*Natural Language Processing*)
 - korpusová lingvistika je podmnožinou
- **korpusová lingvistika** – stojí samostatně
- **vymezení v rámci lingvistiky**
- samostatný obor
 - přístup **corpus-driven**, výzkum korpusem řízený
 - reformulování introspekci 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 označovaný, 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** (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

- elektronický soubor textů (rozsáhlý)
- autentické texty, slova v přirozeném kontextu
– **konkordance**
- sjednocené texty
 - strojově čitelný formát, machine readable format/MRF
 - jednotné kódování
- **označovaná data** (přidané informace)
- reprezentativní vůči svému účelu

Jak korpus vypadá uvnitř

- **vertikál** (vertikální text)
- **token** (tokenizace)
 - řetězec znaků ohraničený z obou stran mezerami
- **type** (word, slovní tvar n. lemma)
- **token – type**
 - **token-type ratio, type/token**
 - vysoké číslo = bohatost slovníku
 - nízké číslo = velké opakování slov
- pro uživatele – **korpusové manažery**
- **konkordance, KWIC** (key word in context)

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>

Výskytů: **1 186** | i.p.m.⁰: 9,75 (vztaženo k celému syn2010) | ARF⁰: **442,42** | Výsledek je promíchán

1 / 30

<input type="checkbox"/>	opus#2162,Hospodářské noviny, 14. 4. 2008	" Podle právníků však tímto způsobem studenti hrubě porušují nejen studijní povinnosti , ale hlavně i zákon . Názory na to
<input type="checkbox"/>	opus#2044,Mladá fronta DNES, 3. 7. 2007	po celý tento týden uzavřena veškerá pracoviště v hlavní budově Studijní a vědecké knihovny v Plzni . Od příštího pondělí se
<input type="checkbox"/>	opus#2312,Deníky Bohemia, 25. 8. 2009	vysokou školu . " Vše je možné objednat . Kdyby studijní knihy vyprodal nakladatel , dají se přetáhnout z jiných obchodů
<input type="checkbox"/>	opus#1801,Právo, 30. 6. 2005	zkoušky uchazečům prominuty . Tradičně největší zájem byl o bakalářské studijní obory Sociální práce , Tělesná výchova a sport a Ekonomická
<input type="checkbox"/>	opus#2225,Mladá fronta DNES, 18. 11. 2008	že jim na ně finančně přispěje a umožní jim čerpat studijní volno - dá jim perspektivu a zaváže si je i
<input type="checkbox"/>	opus#113,Vládoi Sedmihoří. Magická cesta	. " " Budeš se vzdělávat . Vypadá to na studijní pobyt " " Mně nikdy nic
<input type="checkbox"/>	opus#928,Základy práva pro neprávní obory	povinností vyplývajících z výkonu svěřené funkce , obdobně i porušení studijní kázně a další . Jedná se o širokou kategorii deliktů
<input type="checkbox"/>	opus#876,Úvodní kapitoly k financování školství	nejvyšší počet dětí , žáků nebo studentů ve třídě , studijní skupině nebo oddělení v příslušném oboru vzdělání ve škole nebo
<input type="checkbox"/>	opus#2395,Právo, 26. 2. 2009	mluví meziférické radnice Práva otevřou doktorandské studium OLOMOUC - Doktorandský studijní program otevře s největší pravděpodobností už letos na podzim Právnická
<input type="checkbox"/>	opus#2382,Mladá fronta DNES, 13. 6. 2009	studium . " Volného času drobná blondýnka příliš nemá . Studijní povinnosti a mimoškolní aktivity jí prý zabírají všechny čas .
<input type="checkbox"/>	opus#2351,Pátek Lidových novin, č. 13/2009	. " Loni byla Veronika se spolubydlící Katkou za dobré studijní výsledky v Bruselu , kam ji europoslankyně Jana Bobořková pozvala
<input type="checkbox"/>	opus#1840,Týden, č. 27/2005	jsem ráda , že sportuje , protože jinak byl vyložen studijní typ , " vzpomíná matka Jarmila Skopová . Při přecházení
<input type="checkbox"/>	opus#1526,S tebou mě baví život, č. 37/2007	, ale všechno mě baví . Chci požádat o individuální studijní plán a doufám , že to zvládnu , " věří
<input type="checkbox"/>	opus#2001,Hospodářské noviny, 12. 1. 2007	škol v americkém stylu ? Nekompromisně srovnávajícím kvalitu profesorů , studijní plány i kariéry absolventů . Na přístupovém heslu k němu
<input type="checkbox"/>	opus#1970,Týden, č. 34/2006	tabu , po válce až donedávna se veřejně , mimo studijní účely , nepromítaly . To Riefenstahlové na druhé straně nebránilo
<input type="checkbox"/>	opus#926,Správní právo	zkratce " Bo. " uváděné před jménem) . Magisterský studijní program je zaměřen na získání teoretických poznatků založených na soudobém
<input type="checkbox"/>	opus#873,Hospodářská soutěž	vymezování relevantního trhu značně subjektivní . 6 Zneužití dominantního postavení Studijní cíle Cílem této kapitoly je objasnit samotný pojem dominantní postavení
<input type="checkbox"/>	opus#918,Praktikum občanského práva	v přírodě . Ty potřeboval pořídit ke zdárnému splnění účelu studijní cesty asistenta v oblasti výskytu vzácné přimořské flóry během jeho
<input type="checkbox"/>	opus#539,Paměti lékaře	, zda Jirka během svých studií uzavřel vůbec nějakou dílčí studijní etapu zkouškou . Vím jen , že v době ,
<input type="checkbox"/>	opus#1076,AD Speciál, č. 1/2005	i moderně vybavené kolej pro studenty a studentská jídelna . Studijní obory Stěžejní obor Charitní a sociální činnost je určen zájemoům
<input type="checkbox"/>	opus#269,Svatost manželství	že když se náš tělocvikář zláskal a utekl s vedoucí studijní poradny , přivedli jsme oba nazpátek . Naši kolegové splhli
<input type="checkbox"/>	opus#2354,Pátek Lidových novin, č. 49/2009	. Myslím si , že cestování , zahraniční stáže a studijní a pracovní pobyty jsou určitě právě o tom , aby
<input type="checkbox"/>	opus#24,Bourneiv mýtus	dovolenou , i když nezvyklou . " Zavolejte děkanovi pro studijní záležitosti , pane . . . Wedde . Já teď
<input type="checkbox"/>	opus#55,Poslední rituál	ho v té ieskvní . Kdysi isem dostal na fakultě studijní volno a ponořil isem se do období osídlení Islandu azv

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 (verbtage)***)
 - 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 texty v jednotném formátu**
 - 2. značkování**
 - 3. zobrazení konkordancí v korpusových manažerech**
 - 4. vymezený obsah a rozsah**