Korpusová lingvistika – 1
Úvod – korpus a korpusová lingvistika, základní pojmy

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Osnova

- Úvod – korpus a korpusová lingvistika, základní pojmy
- Vývoj korpusové lingvistiky
- Typy korpusů, české korpusy
- Budování korpusů
- Morfologické značkování
- Korpusové manažery
- Využívání korpusů
- Časopisy, konference, publikace

Korpusová lingvistika

- počítačová lingvistika (počítačové zpracování přirozeného jazyka, Natural Language Processing – NLP)
- korpusová lingvistika
- směr lingvistiky založený na empirii, observaci, zkoumání jazykového materiálu
- využití počítačové techniky a nástrojů
- 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.

Co je to korpus

- elektronický soubor textů (rozsáhlý)
- autentické texty, přirozený kontext
- jednotný formát
  - strojově čitelný, machine readable format/MRF
  - jednotné kódování
- označkovaná data
- reprezentativní vůči svému účelu

Jak korpus vypadá

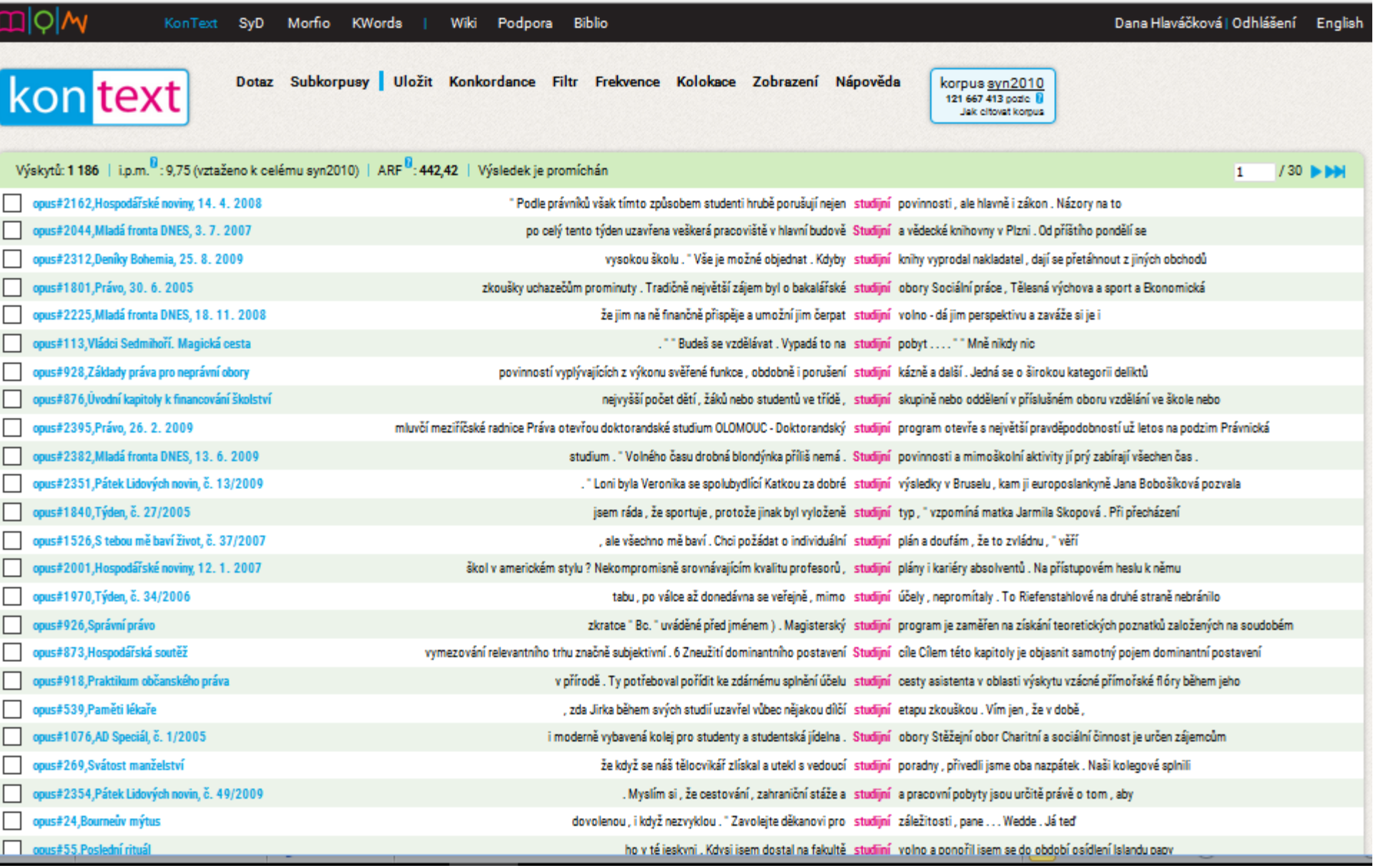
- vertikál, pozice (token, tokenizace)
- slovo (word) – řetězec znaků

ohraničený z obou stran mezerami

- 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>


KonText SyD Morfio KWords | Wiki Podpora Biblio
Dana Hlaváčková | Odhlášení English
kontext
Dotaz Subkorpusy | Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápověda
korpus syn2010
121 667 413 pozic
Jak citovat korpus
Výskytů: 1 186 | i.p.m.: 9,75 (vztaženo k celému syn2010) | ARF: 442,42 | Výsledek je promíchán
1 / 30
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
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
opus#2312,Deníky Bohemia, 25. 8. 2009 vysokou školu . " Vše je možné objednat . Kdyby studijní knihy vyprodal nakladatel , dají se přetáhnout z jiných obchodů
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
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
opus#113,Vládci Sedmihoří. Magická cesta ." " Budeš se vzdělávat . Vypadá to na studijní pobyt . . . . " " Mně nikdy nic
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
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
opus#2395,Právo, 26. 2. 2009 mluvčí mezifiřské radnice Práva otevřou doktorandské studium OLOMOUC - Doktorandský studijní program otevře s největší pravděpodobností už letos na podzim Právnická
opus#2382,Mladá fronta DNES, 13. 6. 2009 studium . " Volného času drobná blondýnka příliš nemá . Studijní povinnosti a mimoškolní aktivity jí prý zabírají všechen čas .
opus#2351,Pátek Lidových novin, č. 13/2009 ." Loni byla Veronika se spolubydlící Katkou za dobré studijní výsledky v Bruselu , kam ji europoslankyně Jana Bobošíková pozvala
opus#1840,Týden, č. 27/2005 jsem ráda , že sportuje , protože jinak byl vyloženě studijní typ , " vzpomíná matka Jarmila Skopová . Při přecházení
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
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
opus#1970,Týden, č. 34/2006 tabu , po válce až donedávna se veřejně , mimo studijní účely , nepromítaly . To Riefenstahlové na druhé straně nebránilo
opus#926,Správní právo zkratce " Bc. " uváděné před jménem) . Magisterský studijní program je zaměřen na získání teoretických poznatků založených na soudobém
opus#873,Hospodářská soutěž vymezování relevantního trhu značně subjektivní . 6 Zneužití dominantního postavení Studijní cíle Cílem této kapitoly je objasnit samotný pojem dominantní postavení
opus#918,Praktikum občanského práva v přírodě . Ty potřeboval pořídit ke zdárnému splnění účelu studijní cesty asistenta v oblasti výskytu vzácné přímořské flóry během jeho
opus#539,Paměti lékaře , zda Jirka během svých studií uzavřel vůbec nějakou dílčí studijní etapu zkouškou . Vím jen , že v době ,
opus#1076,AD Speciál, č. 1/2005 i moderně vybavená kolej pro studenty a studentská jídelna . Studijní obory Stěžejní obor Charitní a sociální činnost je určen zájemcům
opus#269,Svátost manželství že když se náš tělocvikář zlíbal a utekl s vedoucí studijní poradny , přivedli jsme oba nazpátek . Naši kolegové splnili
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
opus#24,Bourneův mýtus dovolenou , i když nezvyklou . " Zavolejte děkanovi pro studijní záležitosti , pane . . . Wedde . Já teď
opus#55,Poslední rituál ho v té jeskyni . Kdysi jsem dostal na fakultě studijní volno a ponořil jsem se do období osídlení Islandu pap

Obsah a rozsah korpusu

- korpusy psané a mluvené
- 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 pozic
  - počet slov
- opravdu velké korpusy (webové, 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 (předem stanoven)
  - otevřený korpus (plynule se zvětšuje)

Značkování korpusu

- značkování – zvyšuje informační hodnotu korpusu (vždy nutná dostupná interpretace značek)
- vnitřní značkování (vnitrotextové)
  - strukturní atributy (opus, doc, s)
  - morfologické značky

- vnější značkování, (vnětextové) na úrovni textu, metatextové informace (autor, název díla, rok vydání atd.)