STUDIJNÍ OPORA K DISCIPLÍNĚ KORPUSOVÁ LINGVISTIKA Katedra českého jazyka a literatury Pedagogické fakulty Univerzity Palackého	
Počet kreditů:	1
Typ předmětu:	volitelný
Způsob zakončení:	zápočet
Garant předmětu:	Mgr. Jana Sladová, Ph.D.
Autor studijní opory, kontakt:	Mgr. Jana Kusá, Ph.D., e-mail: jana.kusa@upol.cz
Cíle předmětu:	Cílem předmětu je představit možnosti analýzy gramatiky přirozeného jazyka prostřednictvím jazykových korpusů. Aplikace jednotlivých metod bude prezentována na materiálu Českého národního korpusu.
Obsah předmětu:	Základy korpusové lingvistiky. Struktura Českého národního korpusu. Vyhledávání pomocí tzv. regulárních výrazů, morfologických značek a lemmat. Pravopis. Morfologie.

OBSAH STUDIJNÍ OPORY

1. KORPUS A KORPUSOVÁ LINGVISTIKA

Korpus: soubor dokladů autentického užití přirozeného jazyka – materiálová základna, která slouží lingvistické analýze a popisu, a to jak z hlediska jazyka psaného, tak mluveného. V současné době je toto označení užíváno pro elektronicky uložený a uchovávaný soubor textů.

Korpusová lingvistika: lingvistická disciplína zkoumající jazyk pomocí elektronických jazykových korpusů; zabývá se výstavbou a zpracováním těchto korpusů, rozvíjí metodologii oboru.

Výhody využití korpusu při lingvistické analýze

- minimalizuje časovou náročnost a pracnost ručního zpracování lingvistického materiálu;
- umožňuje počítačové zpracování jazyka;
- pomocí vhodného programu je možné během několika sekund nalézt velké množství konkrétních dokladů užití jednoho slovního tvaru a jeho frekvenci výskytu (analýza umožní určit jevy centrální a periferní);
- umožňuje získat doklady přímo z autentických textů;
- zaručuje spolehlivost a přesnost měření;
- díky efektivnímu počítačovému zpracování poukazuje na stabilizované rysy a jevy jazyka, které lze abstrahovat a zobecnit.

1.1 Typy korpusů

A. Podle množství obsažených jazyků:

- **korpusy jednoho jazyka**: např. Český národní korpus;
- **korpusy paralelní**: obsahují stejné texty ve dvou i více jazycích.

B. Podle časového záběru:

- **diachronní**: zpracovávají jazyk v průběhu delšího časového období;
- **synchronní**: zabývají se časovým obdobím tak krátkým/dlouhým, že v něm není třeba přihlížet k vývojovým změnám.

C. Podle účelu sestavení:

- **všeobecné (základní)**: nebyly sestaveny za účelem blíže specifikovatelného výzkumu, neupřednostňují žádnou oblast lingvistiky, rovinu jazyka či stylovou příslušnost; většinou slouží ke studiu slovní zásoby nebo gramatiky;[1]

[1] Všeobecné korpusy bývají projektovány tak, aby byly co nejvíce vyvážené (při jejich sestavování se berou v úvahu různé typy žánrů, různá média přenosu či stupně formálnosti – komunikace veřejná či soukromá).

- **specializované:** sestavené ke konkrétnímu výzkumu; zkoumají např. rozvoj jazyka u dětí, dialektové variety, chyby studentů určitého jazyka apod.

D. Podle formy:

- **korpusy mluveného jazyka** (pro náročnost zpracování jsou většinou méně rozsáhlé);
- **korpusy psaného jazyka.**

1.2 Zpracování korpusu

- text, který do korpusu vstupuje, je převáděn do jednotného formátu a získává identifikační hlavičku (technické informace o konverzi);

- k nejběžnějšímu softwarovému vybavení patří **vyhledávací program**: vyhledává všechny slovní výskyty, které odpovídají zadanému dotazu, a zobrazí je spolu s jejich kontextem – tzv. KWIC formát (hledané slovo je zobrazeno ve středu obrazovky a kontext se rozbíhá v jedné řádce na obě strany); program zároveň poskytuje informaci o počtu nalezených dokladů;

- **značkování textů**: **vnější anotace** (typ textu, rok vzniku, autor apod. – informace jsou uváděny na okraji konkordačního[2] řádku dokumentujícího výskyt hledaného slova), **vnitřní anotace** (lingvistické informace – morfologické značkování a lemmatizace; strukturní informace – o členění textů na kapitoly, odstavce, věty, slova);
  - **lemmatizátor**: program, který každému slovnímu výskytu ve všech textech korpusu přiřadí odpovídající lemma (základní podoba lexému, která se uvádí jako reprezentativní ve slovnících – u substantiv např. nominativ singuláru);
  - pomocí tzv. **tagerů** jsou texty opatřeny značkami gramatických kategorií (jsou otagovány);

- pro odstranění gramatických kategorií, které neodpovídají danému slovnímu tvaru (z důvodu tvarové homonymie), jsou využívány tzv. **disambiguátory**;

- z hlediska syntaktického text značkují tzv. **parsery**.

2. ČESKÝ NÁRODNÍ KORPUS[3]

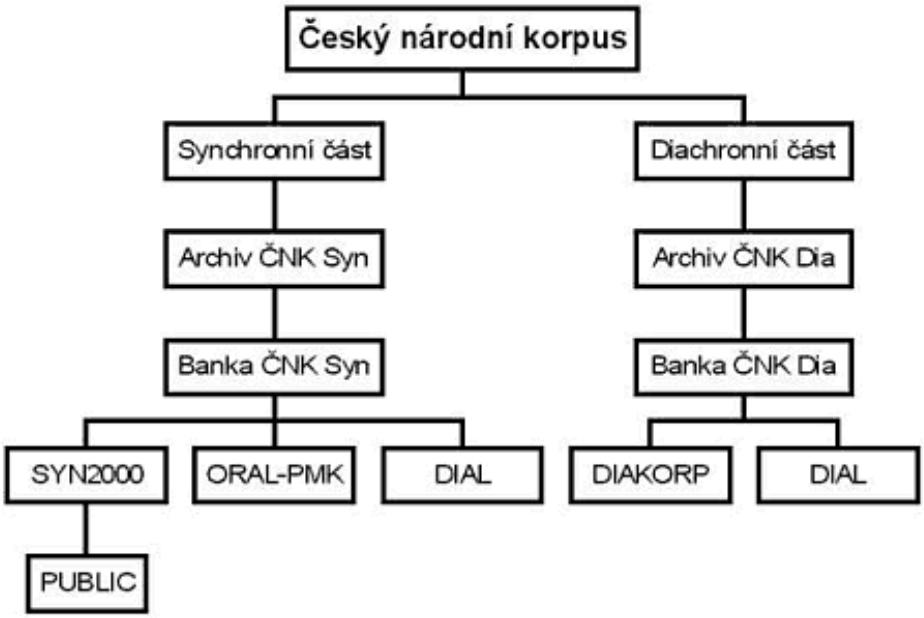
- v roce 1996 vznikl na Filozofické fakultě Univerzity Karlovy v Praze **Ústav Českého národního korpusu (ÚČNK)**, který je veden **prof. dr. Františkem Čermákem** a vytváří ČNK;

- ČNK je kontinuální projekt, jehož jednotlivé konkrétní korpusy **mapují a monitorují různé podoby českého jazyka** tak, aby uživatelům zpřístupnily co nejbohatší zdroj jazykových dat a příslušné nástroje k jejich využívání;

- označení ČNK v sobě zahrnuje několik složek (specializovaných korpusů), vytvářených z elektronických textů různé povahy, zaměření a rozsahu.

[2] Konkordace: výpis všech řádků s výskytem zkoumaného jevu v kontextu.
[3] Dále jen ČNK.

SCHÉMA ZÁKLADNÍHO ČLENĚNÍ ČNK[4]



2.1 Synchronní korpus

- velikostí dominuje **synchronní psaný korpus** (SYN2000), ze kterého vychází korpus PUBLIC (veřejně dostupný na internetu); SYN2000 zahrnuje novinové a časopisecké texty, texty krásné literatury a odborné texty, které vznikly v letech 1990 až 1999 (texty, které leží za touto časovou hranicí, jsou součástí korpusu diachronního);

- na SYN2000 navazuje **SYN2005** (obsahuje texty z let 2000 až 2004) a **SYN2010** (texty z let 2005 až 2009);

- součástí korpusu SYN jsou dva **korpusy publicistických textů** – **SYN2006PUB** a **SYN2009PUB**, zahrnující publicistické texty vzniklé od roku 1989 do roku 2007;

- **další součásti synchronního korpusu**: **FSC2000** (upravený SYN2000, referenční zdroj *Frekvenčního slovníku češtiny*); **CZE SL-PLAIN** (nereferenční žákovský korpus češtiny nerodilých mluvčích); **KSK-DOPISY** (přepisy ručně psané korespondence z let 1990 až 2004), **LINK** (nereferenční korpus sestavený z odborných lingvistických textů), **ORWELL** (ručně označkovaný korpus Orwellova románu 1984), **SKRIPT2012** (korpus školních písemných prací);

- **synchronní mluvený korpus** (ORAL) zaznamenává autentickou mluvenou češtinu – v současné době obsahuje tyto verze: **ORAL2006** (korpus neformální mluvené češtiny), **ORAL2008** (sociolingvisticky vyvážený korpus neformální mluvené češtiny), **SCHOLA2010** (korpus vyučovacích hodin), **PMK** (Pražský mluvený korpus), **BMK** (Brněnský mluvený korpus).

- součástí ČNK jsou rovněž **nářeční korpusy** obojího typu (DIAL).

[4] ČERMÁK, F.; SCHMIEDTOVÁ, V. *Český národní korpus – základní charakteristika a širší souvislosti* [online]. 2004 [cit. 2013-09-30]. Dostupné z: http://knihovna.nkp.cz/nkkr0403/0403152.html. Schéma zachycuje členění ČNK v roce 2000.

2.2 Diachronní korpus

- **diachronní psaný korpus** (DIAKORP) má ve srovnání se synchronním korpusem menší rozsah – tvoří materiálovou základnu pro výzkum vývoje českého jazyka od prvních dochovaných souvislejších záznamů (2. polovina 13. století) zhruba do poloviny 20. století.

2.3 Další součásti Českého národního korpusu[5]

- nekorpusovou složku tvoří **archivy**, v nichž jsou uloženy získané texty ve výchozí podobě (Archiv ČNK Syn a ČNK Dia), a **banky**, v nichž se ukládají všechny texty v konečném formátu (Banka ČNK Syn a ČNK Dia).

3. PŘÍSTUP K ČESKÉMU NÁRODNÍMU KORPUSU

Český národní korpus je dostupný na internetových stránkách Ústavu Českého národního korpusu:

http://ucnk.ff.cuni.cz/

3.1 Veřejný přístup do korpusu SYN2010

- k využití dat z veřejného korpusu není nutná registrace;

- vyhledávání pomocí WWW rozhraní je jednoduché: slovo (nebo třeba jen příponu nebo předponu) napíšeme do vstupního pole a stiskneme tlačítko *Hledej* – po chvíli se objeví výsledek vyhledávání ve formě konkordačního řádku (zobrazí se prvních 50 konkordačních řádků – s uvedením údaje o celkovém počtu výskytů);

- při vyhledávání je možné využít i takzvaných **regulárních výrazů**:

  - **tečka** (.) - představuje jakýkoli znak,
  - **hvězdička** (*) - představuje libovolný počet (tj. 0-n) opakování předchozího znaku nebo výrazu,
  - **plus** (+) - představuje 1 nebo více opakování předchozího znaku nebo výrazu,
  - **otazník** (?) - představuje žádný nebo jeden výskyt předchozího znaku nebo výrazu,
  - **interval {n, k}**- představuje n až k opakování, je-li k vynecháno, odpovídá intervalu nejméně n opakování, pokud má interval tvar {n}, odpovídá mu přesně n opakování.

Pokud chcete například vyhledat slova, která začínají na *les*, do dotazového řádku napište: **les.*** Tento dotaz vyhledá tvary *lesy*, *lesního*,*lesů*..., ale i např. *lesklé*, *leskly*...

Podobně dotaz **.*tel** vyhledá slova zakončená na *tel*, tj. *spasitel*, *spisovatel*, *ředitel*, *obyvatel* atd.

[5] Kromě korpusů zaměřených na český jazyk jsou dostupné korpusy cizojazyčné (např. korpus horní a dolní lužické srbštiny, webový korpus němčiny, francouzštiny apod.).

Další příklady:[6]

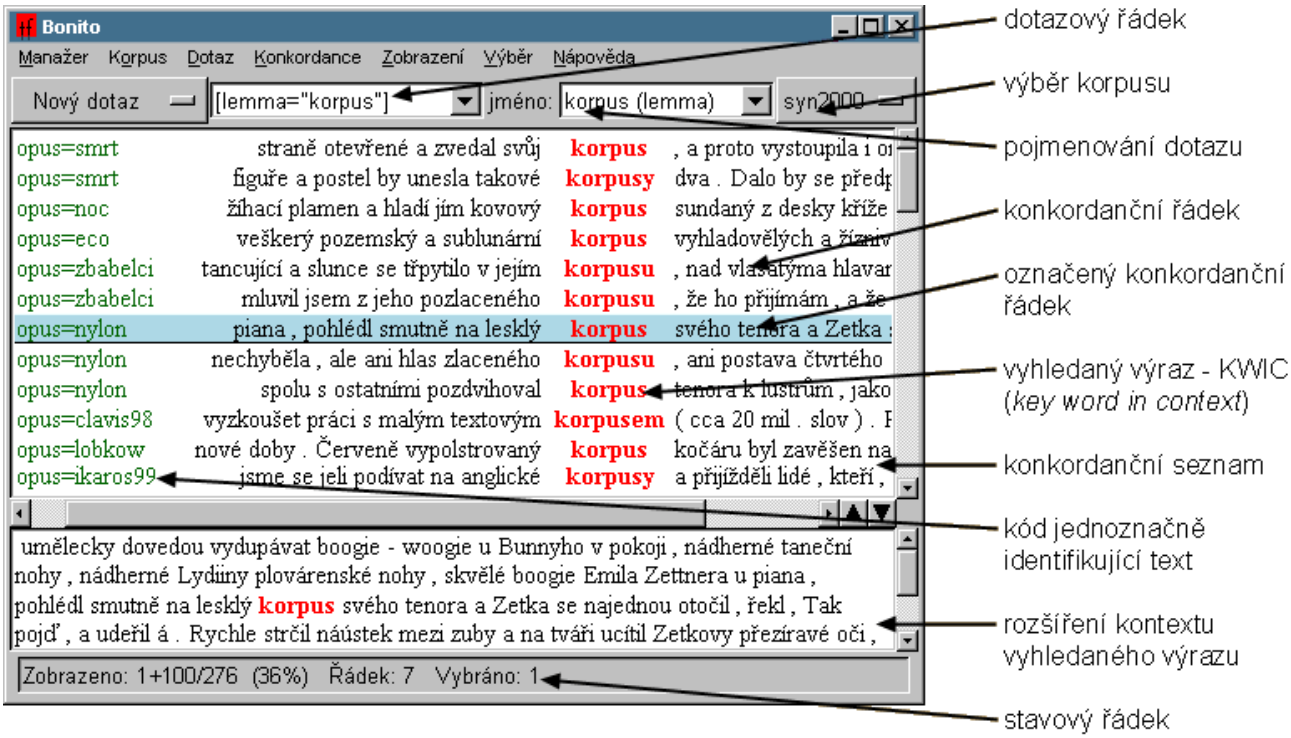
Příklad dotazu	zadání dotazu
slovo "**kdy**" s malým nebo velkým počátečním písmenem	**[kK]dy**
všechny tvary slova "**kočka**"	**koče?[kc].***
infinitivy předponových sloves od "**nést**"	**.+nést**
různě dlouhé varianty citoslovce "**ratata**"	**ra(ta)+**
pravopisnou dubletu: "**diskuze**" psané i se s	**disku(s|z)e** nebo**disku[sz]e**
všechny morfologické varianty slova "**smích**" (s vyloučením tvarů odvozených od slov "**Smíchov**" a "**smíchat**")	**[Ss]mích[^oaá].*|[Ss]mích**

3.2 Plný přístup

- plnohodnotný přístup ke všem korpusům ÚČNK mají pouze registrovaní uživatelé (registrace uživatele proběhne na základě souhlasu s *Prohlášením uživatele korpusů ČNK* – je dostupné na internetových stránkách ÚČNK – ve formátech určených k tisku nebo ve formě elektronického formuláře);

- uživatelé pracují s korpusy pomocí speciálního korpusového manažeru **Bonito** (v tradiční nebo webové verzi).[7]

KORPUSOVÝ MANAŽER BONITO[8]



[6] KOCEK, J.; KOPŘIVOVÁ, M. *Manuál korpusového manažeru Bonito* [online]. 2004 [cit. 2013-09-30]
[7] Manuál korpusového manažeru Bonito je dostupný na http://ucnk.ff.cuni.cz/bonito/kontext.php.
[8] KOCEK, J.; KOPŘIVOVÁ, M. *Manuál korpusového manažeru Bonito* [online]. 2004 [cit. 2013-09-30]

4 POŽADAVKY K UDĚLENÍ ZÁPOČTU

1. Samostudium – studijní opora, příp. další doporučená literatura.

2. Registrace pro plný přístup k ČNK.

3. Instalace korpusového manažeru BONITO.

4. Prostudování dílčích kapitol manuálu korpusového manažeru Bonito:[9]

- Začínáme…
- Základní vyhledávání
- Seznam slov
- Popis morfologických značek
- Vytvořit tag

5. Seminární projekt – zpracování lingvistických jevů korpusovými metodami.

A. PRAVOPIS: vyhledávání lingvistického jevu zadáním slova do dotazového řádku (s využitím regulárních výrazů):

- student si zvolí **tři pravopisné jevy** (např. z oblasti psaní velkých písmen – univerzita/Univerzita, pravopisu slov přejatých – jazz/džez, pravopisných variant – odesilatel/odesílatel);
- vytvoří vhodný dotaz s využitím regulárních výrazů;
- využije vyhledávání všech tvarů zadaného slova pomocí atributu „lemma“;
- zjistí počet výskytů a rozhodne, který jev je pravopisně správný (výsledek své analýzy ověří ve *Slovníku spisovné češtiny*);
- u pravopisných variant zjistí jejich frekvenci přes Konkordance/Statistiky/Frekvenční distribuce (při zadávání Frekvenční distribuce musí zadat atribut lemma);
- **při zpracování úlohy student:** popíše zkoumaný lingvistický jev, uvede příslušný dotaz v podobě, kterou využil při vyhledávání, prezentuje několik příkladů nalezených v korpusu (případně vloží sken vyhledávacího okna) a slovně lingvistický jev vyhodnotí – popíše dané výskyty, uvede příklady, zachytí frekvenci, formuluje závěry své analýzy.

B. MORFOLOGIE: vyhledávání pomocí morfologických značek:

- při vyhledávání pomocí morfologických značek využije student atribut „tag“;
- pracuje pomocí grafického vytváření dotazu (viz kapitola *Vytvořit tag* v manuálu korpusového manažeru Bonito);
- student si zvolí **dva problematické morfologické jevy** (např. tvar slovesa být v 1. osobě množného čísla u podmiňovacího způsobu, tvary 6. pádu substantiva kámen, kolísání rodu u substantiva esej aj.);
- vytvoří vhodný dotaz s využitím morfologických značek;
- zjistí četnost výskytu jednotlivých variant a rozhodne o gramaticky správné variantě (výsledek své analýzy ověří ve *Slovníku spisovné češtiny*);
- **při zpracování úlohy student:** popíše zkoumaný lingvistický jev, uvede příslušný dotaz v podobě, kterou využil při vyhledávání, prezentuje několik příkladů nalezených v korpusu (případně vloží sken vyhledávacího okna) a slovně lingvistický jev vyhodnotí – popíše dané výskyty, uvede příklady, zachytí frekvenci, formuluje závěry své analýzy.

[9] Manuál korpusového manažeru Bonito je dostupný na http://ucnk.ff.cuni.cz/bonito/kontext.php.

5 POUŽITÉ A DOPORUČENÉ INFORMAČNÍ ZDROJE

BLATNÁ, R.; ČERMÁK, F. (eds.). *Jak využívat Český národní korpus*. Praha: Nakladatelství Lidové noviny, 2005.

ČERMÁK, F.; BLATNÁ, R. *Korpusová lingvistika: stav a modelové přístupy.* Praha: Nakladatelství Lidové noviny, 2006.

ČERMAK, F., KLÍMOVÁ, J.; PETKEVIČ, V. (eds.). *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000.

ČERMÁK, F.; SCHMIEDTOVÁ, V. *Český národní korpus – základní charakteristika a širší souvislosti* [online]. 2004 [cit. 2013-09-30]. Dostupné z: http://knihovna.nkp.cz/nkkr0403/0403152.html.

KOCEK, J; KOPŘIVOVÁ, M.; KUČERA, K. (eds.). *Český národní korpus*. *Úvod a příručka uživatele*. Praha: Filozofická fakulta UK, Ústav Českého národního korpusu, 2000.

KOCEK, J.; KOPŘIVOVÁ, M. *Manuál korpusového manažeru Bonito* [online]. 2004 [cit. 2013-09-30]

ŠULC, M. *Korpusová lingvistika*. *První vstup*. Praha: Karolinum, 1999.

http://ucnk.ff.cuni.cz/