CJBB105 – 5
Korpusové manažery

Mgr. Dana Hlaváčková, Ph.D.

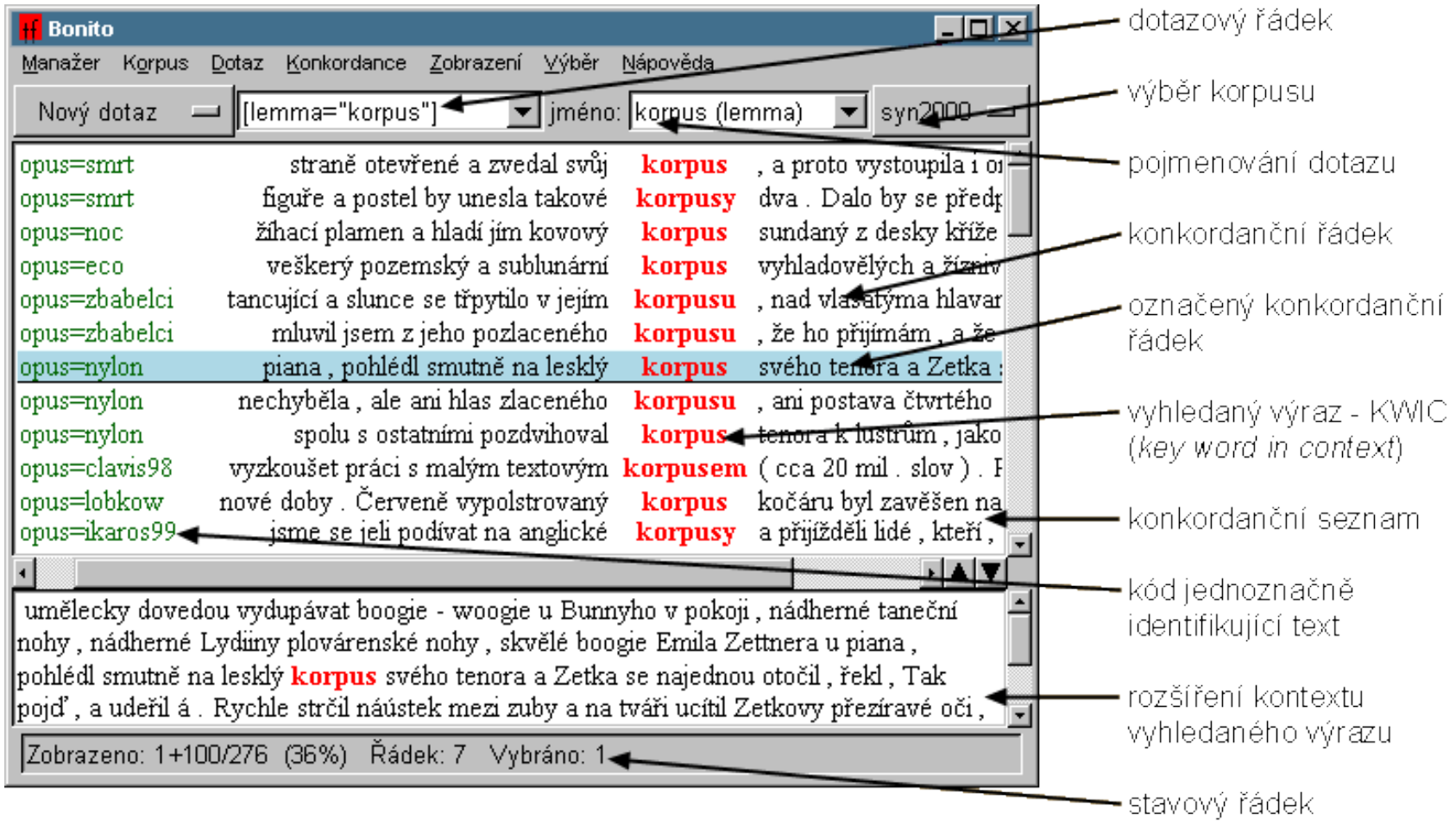
CJBB105

Korpusové manažery

- zpracování textů do korpusové podoby
- **prohlížení** korpusových dat a **práce s nimi**
- budování korpusů
- navazující aplikace spojené s korpusovým zpracováním dat
- desktopová aplikace, webová stránka, webové rozhraní
- často omezený přístup (pouze ukázky), nutná registrace, příp. stažení a instalace

Korpusové manažery

- 1995 – cesta do Velké Británie po centrech korpusové lingvistiky – Pala, Čermák, Petkevič, Schmiedtová
- Oxford University Press, University of Oxford – **Patrick Hanks**
- School od English, Birmingham City University – **John Sinclaire**
- Lancaster University – **Geoffrey Leech**
- – příprava korpusového manažeru – Pavel Rychlý – CQP (Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, prof. Ulrich Heid, autoři CQP Schulze a Christ)
- – **Manatee Bonito** – Pavel Rychlý, server-klient, dizertační práce (r. 2000)


Bonito
Manažer Korpus Dotaz Konkordance Zobrazení Výběr Nápověda
Nový dotaz
[lemma="korpus"]
jméno:
korpus (lemma)
syn2000
opus=smrt straně otevřené a zvedal svůj korpus , a proto vystoupila i oi
opus=smrt figuře a postel by unesla takové korpusy dva . Dalo by se předp
opus=noc žíhací plamen a hladí jím kovový korpus sundaný z desky kříže
opus=eco veškerý pozemský a sublunární korpus vyhladovělých a žízniv
opus=zbabelci tancující a slunce se třpytilo v jejím korpusu , nad vlasatýma hlavar
opus=zbabelci mluvil jsem z jeho pozlaceného korpusu , že ho přijímám , a že
opus=nylon piana , pohlédl smutně na lesklý korpus svého tenora a Zetka
opus=nylon nechyběla , ale ani hlas zlaceného korpusu , ani postava čtvrtého
opus=nylon spolu s ostatními pozdvihoval korpus tenora k lustrům , jako
opus=clavis98 vyzkoušet práci s malým textovým korpusem (cca 20 mil . slov) . F
opus=lobkow nové doby . Červeně vypolstrovaný korpus kočáru byl zavěšen na
opus=ikaros99 jsme se jeli podívat na anglické korpusy a přijížděli lidé , kteří ,
umělecky dovedou vydupávat boogie - woogie u Bunnyho v pokoji , nádherné taneční nohy , nádherné Lydiiny plovárenské nohy , skvělé boogie Emila Zettnera u piana , pohlédl smutně na lesklý korpus svého tenora a Zetka se najednou otočil , řekl , Tak pojď , a udeřil á . Rychle strčil náústek mezi zuby a na tváři ucítil Zetkovy přezíravé oči ,
Zobrazeno: 1+100/276 (36%) Řádek: 7 Vybráno: 1
dotazový řádek
výběr korpusu
pojmenování dotazu
konkordanční řádek
označený konkordanční řádek
vyhledaný výraz - KWIC (*key word in context*)
konkordanční seznam
kód jednoznačně identifikující text
rozšíření kontextu vyhledaného výrazu
stavový řádek

Korpusové manažery

- jádro – Manatee (server), korpusové zpracování textů (Pavel Rychlý, FI MU)
- Manatee + Bonito, Bonito2, Sketch Engine, NoSketch Engine
- uživatelské rohraní (Bonito), webové rozhraní
  - Sketch Engine – MU (CZPJ FI MU + Lexical Computing, Ltd.), Brno
  - KonText – ÚČNK, Praha, využívá Manatee a vychází z NoSketch Engine (Tomáš Machálek)

Možnosti zobrazení

- vybraný korpus, počet nalezených **výskytů**
  - **i.p.m.** – *instances per million* (počet výskytů na milion pozic)
  - **ARF** – *average reduced frequency* (průměrná redukovaná frekvence vzhledem k rozložení tvaru v korpusu)
- zobrazení ve formě konkordance (**KWIC**) nebo **věty**
- atributy – word, lemma, tag, lc, část tagu
- strukturní značky – hranice vět, dokumentů ad.
- reference – metainformace o textech
- šířka kontextu, počet konkordancí na stránku
- popis dotazu (konkordance)

Možnosti hledání

- konkrétní tvar slova (*slovo, slovní tvar, word*)
- lemma – nalezeny všechny tvary slova vyskytující se v korpusu
- fráze – spojení dvou a více slov s výskytem těsně vedle sebe
  - možná specifikace kontextu
- tag
  - konstrukce značky (KT)
- znak (SKE), podřetězec (KT)
- CQL (Corpus Query Language), CQL editor (SKE)
  - [word=„ježkem“]
- specifikace dle **kontextu**
- specifikace dle **metainformací**
- **regulární výrazy** – znaky umožňující efektivnější hledání v korpusech

Třídění výsledků

- náhodný vzorek, promíchání výsledků
- třídění kontextu a KWIC (podle abecedy)
  - podle atributů
  - víceúrovňové a retrográdní
- filtrování konkordancí
  - pozitivní a negativní filtry
  - pouze 1. výskyt v dokumentu (odfiltruje vše kromě 1. výskytu v dokumentu, SKE)

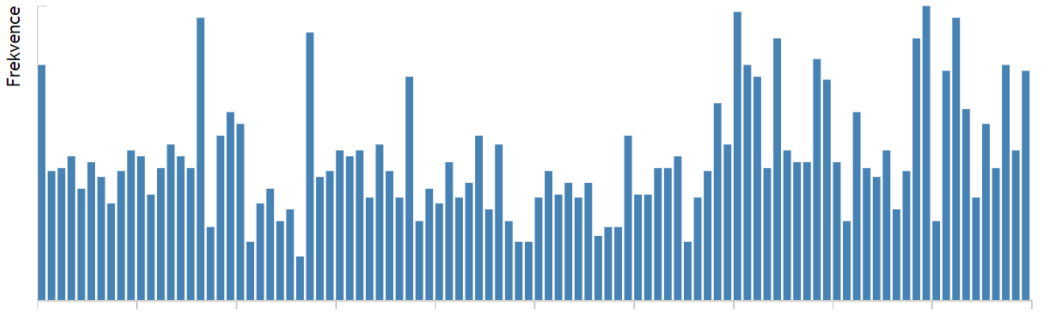
Frekvenční distribuce

- **frekvenční údaje** – číselné i grafické znázornění
  - KWIC (lemmata, slovní tvary)
  - tagy
  - typy dokumentů
  - víceúrovňové
- vizualizace frekvenčního rozložení přes celý korpus (SKE)

Domů
Hledání
Seznam slov
Word sketch
Tezaurus
Sketch rozdíl
Info o korpusu
Mé úlohy
Uživatelská příručka

konkordance
Frekvence
Značky (tags)
Slovní tvary
ID dokumentů
Typy textů

Frekvenční rozložení přes pozice konkordance



Granularita: 100 Překresli

Kolokace

- výpočet kandidátů na kolokace (ustálená slovní spojení)
  - frekvence spojení (dvou a více jednotek) – vysoká
  - frekvence spojení s ostatními jednotkami – nízká
    - vztaženo k velikosti korpusu
  - kolokační paradigma, monokolokabilita (*stroužek česneku, tratoliště krve*)
- asociační míry
- MI-score
  - pravděpodobnost současného výskytu dvou slov (mutual information)
- T-score
  - zapojeno rozložení spojení přes celý korpus, nenáhodný jev
- Dice, Log-Dice
  - nepočítají s velikostí korpusu

Další funkce

- vytvoření subkorpusu
  - podle metainformací o textech (KT)
  - z aktuálních konkordancí (SKE)
- seznam slov
  - podle frekvence
  - uživatel definuje kritéria
- uložení výsledků v různých formátech

KonText – externí funkce

- SyD
  - korpusový průzkum variant slov
  - synchronní i diachronní korpusy
  - psaný i mluvený jazyk
- KWords
  - generování klíčových slov
  - porovnání výskytů s referenčním korpusem
- Morfio
  - vyhledání seznamů slov (až n-tic) na základě slovotvorných charakteristik

Sketch Engine

- Tezaurus – podobná slova, míra podobnosti na základě kontextů, vizualizace
  - hra Uhádni to slovo (*https://nlp.fi.muni.cz/projekty/uhadni_to_slovo/*)
- Word Sketch – slovní profily, tagging
  - tabulky zachycují okolí zadaného lemmatu podle určitých kategorií
- Sketch Diff – porovnání slovních profilů dvou lemmat
- tvorba korpusů a subkorpusů
- SkELL – příkladové věty
  - angličtina https://skell.sketchengine.co.uk/run.cgi/skell
  - čeština https://cskell.sketchengine.co.uk/run.cgi/skell


mouka
šťáva
jogurt
omáčka
smetana
vejce
tvaroh
máslo
lžička
lopata
kapka
troška
rýže
sýr
paprika
ocet
talíř
porce
cukr
vidlička
šálek
gram
cibule
sklenice
žloutek
čokoláda
hrnek
kelímek
příbor
trocha
vývar
sáček
miska
naběračka
mrkev
sklenička
hrst
směs
koření
plátek

KonText a Sketch Engine

- https://kontext.korpus.cz
- registrace: https://www.korpus.cz/signup
- https://ske.fi.muni.cz
- přihlášení (studenti a zaměstnanci MU):
- UČO + sekundární heslo
- mimo MU: zdarma NoSketch Engine
- https://nlp.fi.muni.cz/trac/noske
- https://the.sketchengine.co.uk (možnost 30 dní zdarma)