

CJBB105 – 4

Korpusové manažery

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Korpusové manažery

- umožňují **zpracování textů** do korpusové podoby (tokenizace, vertikál, značkování)
- **prohlížení** korpusových dat a **práci s nimi** (řada fcí)
- **budování korpusů**
 - pouze některé manažery, v ČR Sketch Engine
- poskytují navazující **aplikace** spojené s korpusovým zpracováním dat
- **vývoj**:
 - desktopová aplikace – starší typ, nutná instalace do počítače
 - webová stránka – jednoduchý typ, v některých zemích, omezené fce
 - **webové rozhraní** – nejčastější současný typ, pro všechny prohlížeče a operační systémy, uživatelská přívětivost
- často omezený přístup (pouze ukázky), nutná registrace, příp. i stažení a instalace (v některých zemích)

Historie českého manažeru

- 1995 – cesta do Velké Británie po **centrech korpusové lingvistiky**
 - Karel Pala (FI MU), František Čermák (ÚČNK), Vladimír Petkevič (ÚTKL), Věra Schmiedtová (ÚČNK)
- cílem bylo načerpat informace, zkušenosti a získat korpusový manažer
- Oxford University Press, University of Oxford – **Patrick Hanks**
- School of English, Birmingham City University – **John Sinclair**
- Lancaster University – **Geoffrey Leech**
- byly navázány kontakty, ale nikdo manažer neposkytl
- příprava vlastního českého korpusového manažeru – **Pavel Rychlý** (FI MU)
 - převzal CQP (Corpus Query Processor, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, prof. Ulrich Heid, autoři CQP Schulze a Christ)
- na jeho základě vytvořil **Manatee Bonito** (server-klient), dizertační práce (r. 2000)
- na něm jsou postaveny současné manažery **Sketch Engine** a **KonText**

Původní manažer **Bonito**, desktopová aplikace, 90. léta 20. st. až cca 2014 (od 2003 začíná vývoj Sketch Engine) používal se v celé ČR i pro ČNK, manažer byl velmi oblíbený, lingvisté ho neradi opouštěli a přecházeli na webové rozhraní

The screenshot shows the Bonito application window with the following components and annotations:

- dotazový řádek**: The search bar containing the query "[lemma="korpus"]".
- výběr korpusu**: The corpus selection dropdown menu showing "syn2000".
- pojmenování dotazu**: The dropdown menu for the lemma, showing "jméno: korpus (lemma)".
- konkordanční řádek**: The list of concordance results, each starting with a corpus code (e.g., "opus=smrt").
- označený konkordanční řádek**: The selected row, "opus=nylon", which is highlighted in blue.
- vyhledaný výraz - KWIC (key word in context)**: The search term "korpus" is highlighted in red within the selected row's text.
- konkordanční seznam**: The list of concordance results, including the selected row.
- kód jednoznačně identifikující text**: The corpus code "opus=nylon" in the selected row.
- rozšíření kontextu vyhledaného výrazu**: The expanded context of the selected row, showing the surrounding text: "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 náustek mezi zuby a na tváři ucítil Zetkovy přezíravé oči,".
- stavový řádek**: The status bar at the bottom showing "Zobrazeno: 1+100/276 (36%) Řádek: 7 Vybráno: 1".

Korpusové manažery – vývoj

- jádro – **Manatee** (server), korpusové zpracování textů, **Bonito** – klient (Pavel Rychlý, FI MU)
 - Manatee + **Bonito**
 - **Bonito2** – první webové rozhraní, používalo se krátce, přešel pod něj i ČNK
 - **Sketch Engine** (placená verze, pro MU zdarma), **NoSketch Engine** (zdarma bez některých fcí) – krátkou dobu užíval i ČNK
- současná webová rozhraní
 - **Sketch Engine** – MU (CZPJ FI MU + Lexical Computing, Ltd.), Brno
 - **KonText** – ÚČNK, Praha, využívá Manatee a vychází z NoSketch Engine (Tomáš Machálek)

Prohlížení korpusu a fce manažeru KonText – částečná náhrada za praktické části předmětu

- Sketch Engine
- <https://www.sketchengine.eu/>
- doporučuji zaregistrovat se k ČNK, abyste mohli využívat všechny fce rozhraní <https://korpus.cz/signup>
- KonText <https://kontext.korpus.cz/>
- manuál ke KonTextu najdete zde <https://wiki.korpus.cz/doku.php/manualy:kontext:index>

Možnosti hledání – Dotaz (co a jak je možné v korpusu hledat)

- používám zkratky pro KonText (KT) a Sketch Engine (SKE) pro upozornění na rozdíly mezi manažery, jinak fce platí pro oba
- konkrétní **tvar** slova (*slovo, slovní tvar, word*)
- **lemma** – nalezeny všechny tvary zadaného slova vyskytující se v korpusu
- **fráze** – spojení dvou a více slov s výskytem těsně vedle sebe
 - možná specifikace kontextu, slova fráze nejsou těsně vedle sebe
- **znak** (SKE)
- **CQL** (Corpus Query Language) – dotazovací jazyk
 - konstrukce značky (KT), nápověda pro uživatele
 - CQL builder (SKE)
 - [word=„ježkem“] – formální podoba dotazu v CQL
- specifikace dle **kontextu** – možnost vyhledávat podle kontextu zadaného tvaru
- specifikace dle **metainformací** – možnost vyhledávat podle metadat o textech
- **regulární výrazy** – znaky umožňující efektivnější hledání v korpusech, viz https://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy

Možnosti zobrazení

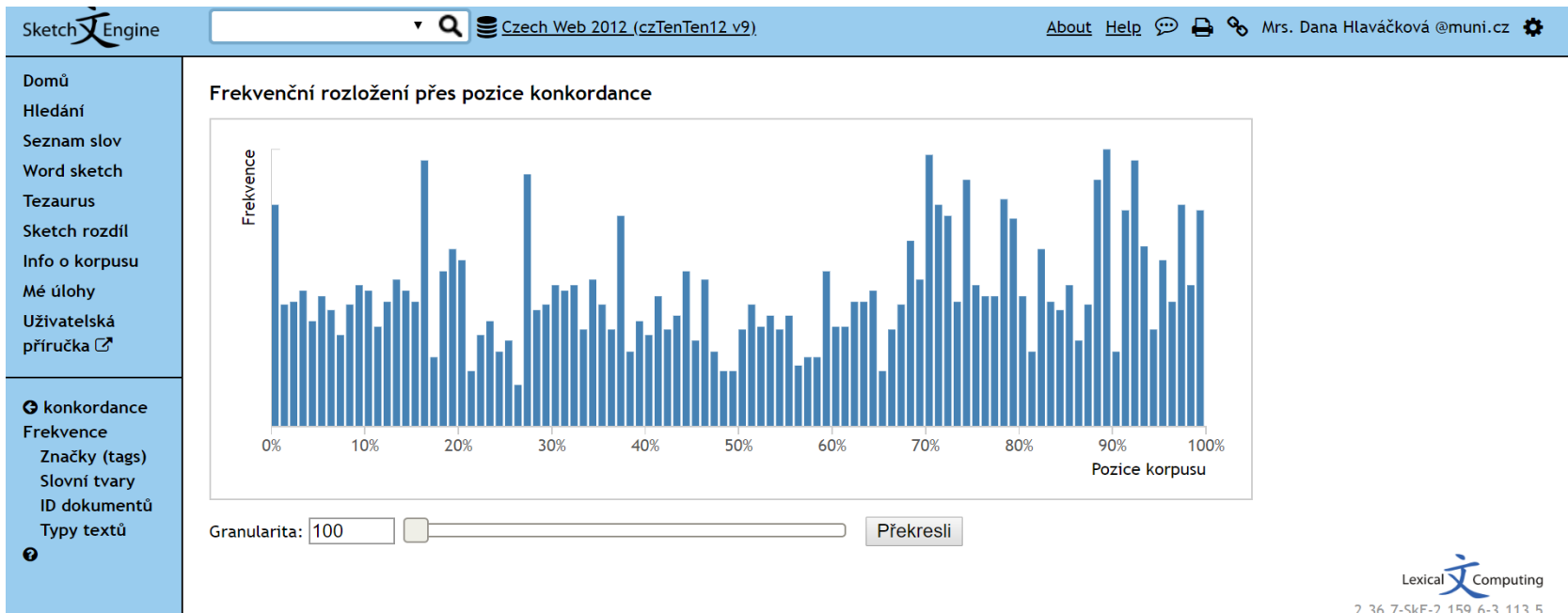
- uživatel vidí **vybraný korpus**, počet nalezených **výskytů**
 - **i.p.m.** – *instances per million* (počet výskytů na milion pozic, důležitý údaj při srovnávání výsledků z různě velkých korpusů)
 - **ARF** – *average reduced frequency* (průměrná redukováná frekvence vzhledem k rozložení tvaru v korpusu, odstraňuje problém, kdy je nějaký tvar frekventovaný např. jen v jednom typu dokumentů) **KT**
 - **procento v korpusu SKE**
- Funkce **Zobrazení**
- zobrazení ve formě konkordance (**KWIC**) nebo **věty** (možnost přepínání)
- **poziční atributy** – word, lemma, tag, lc (lowercase), část tagu
- **strukturní značky** – hranice vět, dokumentů ad.
- **reference** – metainformace o textech
- možnost nastavit šířku kontextu, počet konkordancí na stránku
- možnost zobrazit popis dotazu (konkordance)

Třídění výsledků

- možnost vygenerovat **náhodný vzorek**
- možnost **promíchání výsledků**
- **třídění** kontextu a KWIC (podle abecedy)
 - podle atributů
 - víceúrovňové a retrográdní
- **filtrování** konkordancí
 - pozitivní a negativní filtry (uživatelé definuje, co chce ve výsledcích nechat, nebo co chce odstranit)
 - 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)



Frekvenční seznam tvarů lemmatu „kočka“ v korpusu SYN2015

WaG KonText SyD Morfio KWords Treq | Wiki Podpora Biblio |

kon text Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápověda

Korpus: syn2015 | Dotaz: kočka (8 287 výskytů)

Frekvenční seznam

1 / 1

Minimální frekvence: [Použít](#)

Celkem: 11 položek (1 stránka)

	Filter	word	Freq	
1	p / n	kočky	2575	<div style="width: 100%;"></div>
2	p / n	kočka	2495	<div style="width: 97%;"></div>
3	p / n	kočku	1152	<div style="width: 45%;"></div>
4	p / n	koček	976	<div style="width: 38%;"></div>
5	p / n	kočce	356	<div style="width: 14%;"></div>
6	p / n	kočkou	267	<div style="width: 11%;"></div>
7	p / n	kočkami	183	<div style="width: 7%;"></div>
8	p / n	kočkám	166	<div style="width: 7%;"></div>
9	p / n	kočkách	76	<div style="width: 3%;"></div>
10	p / n	kočko	32	<div style="width: 1%;"></div>
11	p / n	kočkama	9	<div style="width: 0%;"></div>

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
 - můžeme sledovat kolokační paradigma, např. monokolokabilita (*stroužek česneku, tratoliště krve* – *stroužek* a *tratoliště* se nepojí s jinými slovy)
- **asociační míry** (číselné hodnoty, které dokládají kolokabilitu slov)
- **MI-score**
 - pravděpodobnost současného výskytu dvou slov (mutual information)
- **T-score**
 - zapojeno rozložení spojení slov 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** <https://wiki.korpus.cz/doku.php/manualy:syd?redirect=1>
 - korpusový průzkum variant slov
 - synchronní i diachronní korpusy
 - psaný i mluvený jazyk
- **Kwords** <https://wiki.korpus.cz/doku.php/manualy:kwords?redirect=1>
 - generování klíčových slov
 - porovnání výskytů s referenčním korpusem
- **Morfio** <https://wiki.korpus.cz/doku.php/manualy:morfio?redirect=1>
 - vyhledání seznamů slov (až n-tic) na základě slovotvorných charakteristik
- **Treq** <https://treq.korpus.cz/>
 - databáze překladových ekvivalentů
- **Slovo v kostce** <https://www.korpus.cz/slovo-v-kostce/> (vyzkoušejte)

Sketch Engine

- <https://www.sketchengine.eu/>
- LOG IN – Institutional Login – Masarykova univerzita – UČO + primární heslo
- OVLÁDACÍ PANEL
 - **Konkordance** (Concordance) – hledání v korpusu
 - v Profil – Nastavení – možnost přepnout do češtiny
 - funkce jsou stejné jako v KonTextu (s drobnými rozdíly)
- manuál, vysvětlení termínů (Glossary) v angličtině <https://www.sketchengine.eu/guide/>

Sketch Engine – externí funkce

- **Tezaurus** – podobná slova, míra podobnosti na základě kontextů, vizualizace
 - hra Uhádni to slovo (podle kterého synonyma je vytvořen wordcloud, https://nlp.fi.muni.cz/projekty/uhadni_to_slovo/)
- **Word Sketch** – slovní profily, na základě morfol. značkování
 - tabulky zachycují okolí zadaného lemmatu podle určitých kategorií
- **Sketch Diff** – porovnání slovních profilů dvou lemmat
- tvorba korpusů a subkorpusů a další aplikace
- **SkELL** – generování příkladových vět z korpusu
 - <https://skell.sketchengine.eu/>