IB047

Četnosti a kolokace

Pavel Rychlý

pary@fi.muni.cz

8. dubna 2022

Obsah korpusu

Jak zjistíme, co je v korpusu?

- přečteme ho – nepraktické
- struktura korpusu: atributy, struktury, metainformace
- seznam typických (častých) slov
  - pro každý atribut (slovo, lemma, značka)
  - pro atributy struktur (autor, datum, ...)
- základní informaci dávají i velikosti seznamů
- text/type ratio = počet tokenů / počet různých slov

	tokenů	slov	text/type ratio
Susanne	150 000	16 000	9,1
BNC	112 000 000	774 000	145
EnTenTen	13 000 000 000	28 000 000	465

Obsah korpusu

Poznáte o jaký korpus jde?

	řádků	slov	bajtů
Kniha 1	3.715	37.703	223.415
Kniha 2	1.601	16.859	91.031

Seznamy slov

- seznamy nejčastějších prvků (slov) dávají dobrou charakteristiku obsahu
- `sort |uniq -c |sort -rn`
- některá slova dávají informaci o jazyce, jiná o typu jazyka, jiná o obsahu

Kniha 1	**Kniha 2**
the, and, of, to, you, his, in, said, that, I, will, him, your, he, a, my, was, with, s, for, me, He, is, , , it, them, be, The, all, , have, from, , on, her, , , , are, their, were, they, which, , t, up, , had, there	the, I, to, a, of, is, that, , you, he, and, said, was, , in, it, not, me, my, have, And, are, one, for, But, his, be, The, It, at, all, with, on, will, as, very, had, this, him, He, from, they, , so, them, no, You, do, would, like

Seznamy slov

- seznamy nejčastějších prvků (slov) dávají dobrou charakteristiku obsahu
- `sort |uniq -c |sort -rn`
- některá slova dávají informaci o jazyce, jiná o typu jazyka, jiná o obsahu

Kniha 1	**Kniha 2**
the, and, of, to, you, his, in, said, that, I, will, him, your, he, a, my, was, with, s, for, me, He, is, **father**, , it, them, be, The, all, **land**, have, from, , on, her, , **son**, , are, their, were, they, which, **sons**, t, up, , had, there	the, I, to, a, of, is, that, **little**, you, he, and, said, was, , in, it, not, me, my, have, And, are, one, for, But, his, be, The, It, at, all, with, on, will, as, very, had, this, him, He, from, they, **planet**, so, them, no, You, do, would, like

Seznamy slov

- seznamy nejčastějších prvků (slov) dávají dobrou charakteristiku obsahu
- `sort |uniq -c |sort -rn`
- některá slova dávají informaci o jazyce, jiná o typu jazyka, jiná o obsahu

Kniha 1	**Kniha 2**
the, and, of, to, you, his, in, said, that, I, will, him, your, he, a, my, was, with, s, for, me, He, is, **father**, **God**, it, them, be, The, all, **land**, have, from, **Jacob**, on, her, **Yahweh**, **son**, **Joseph**, are, their, were, they, which, **sons**, t, up, **Abraham**, had, there	the, I, to, a, of, is, that, **little**, you, he, and, said, was, **prince**, in, it, not, me, my, have, And, are, one, for, But, his, be, The, It, at, all, with, on, will, as, very, had, this, him, He, from, they, **planet**, so, them, no, You, do, would, like

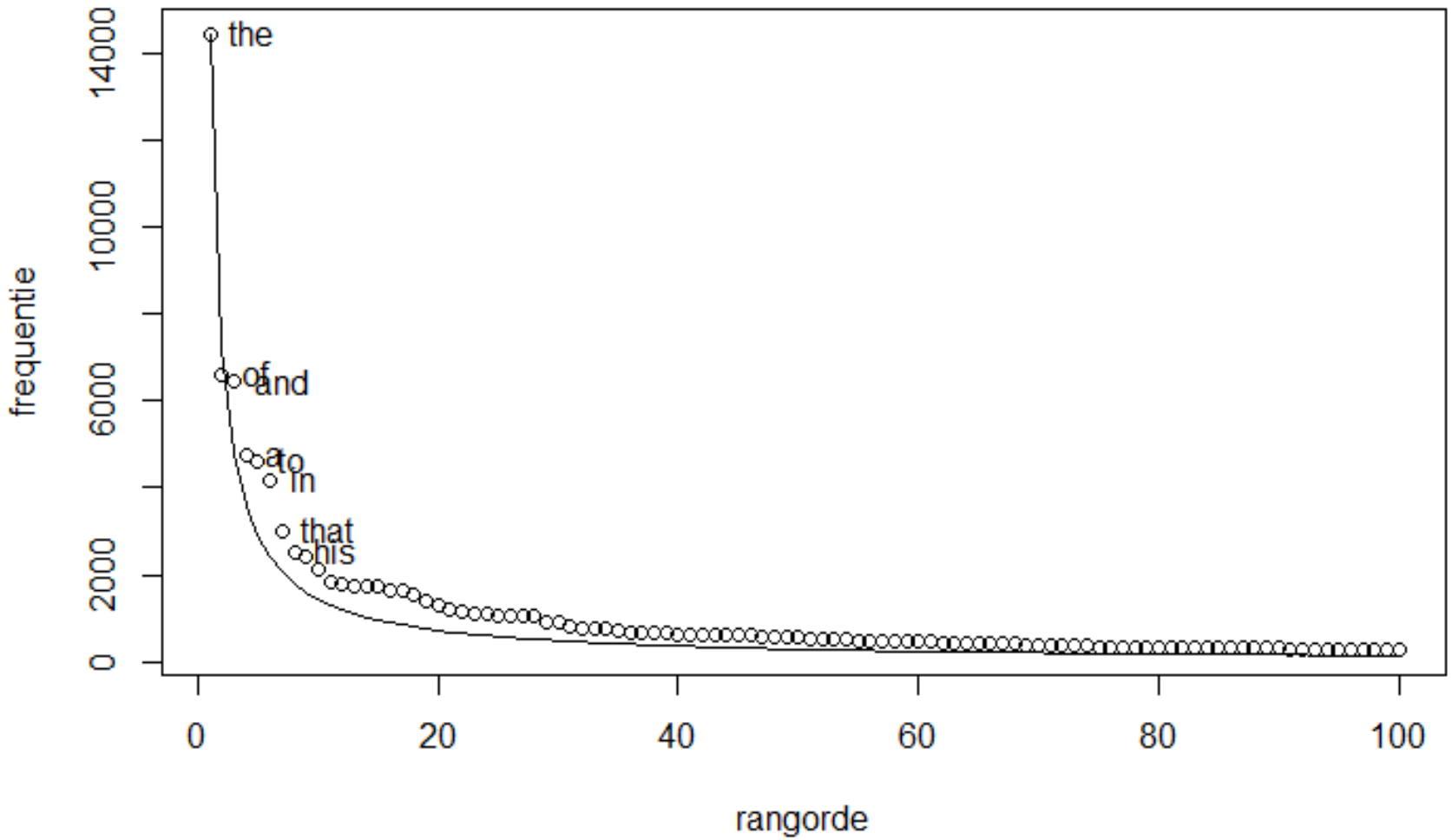
Zipfův zákon

Jaké je rozložení slov v korpusu?

- proč jsou nejčastější slova dobrou charakteristikou?
- nejčastější pokrývají velkou část textu
- $f * r = C$
  součin četnosti a pořadí v seznamu slov (dle četnosti) je zhruba konstantní
- slova, slovní spojení
- vlastní jména, velikosti měst

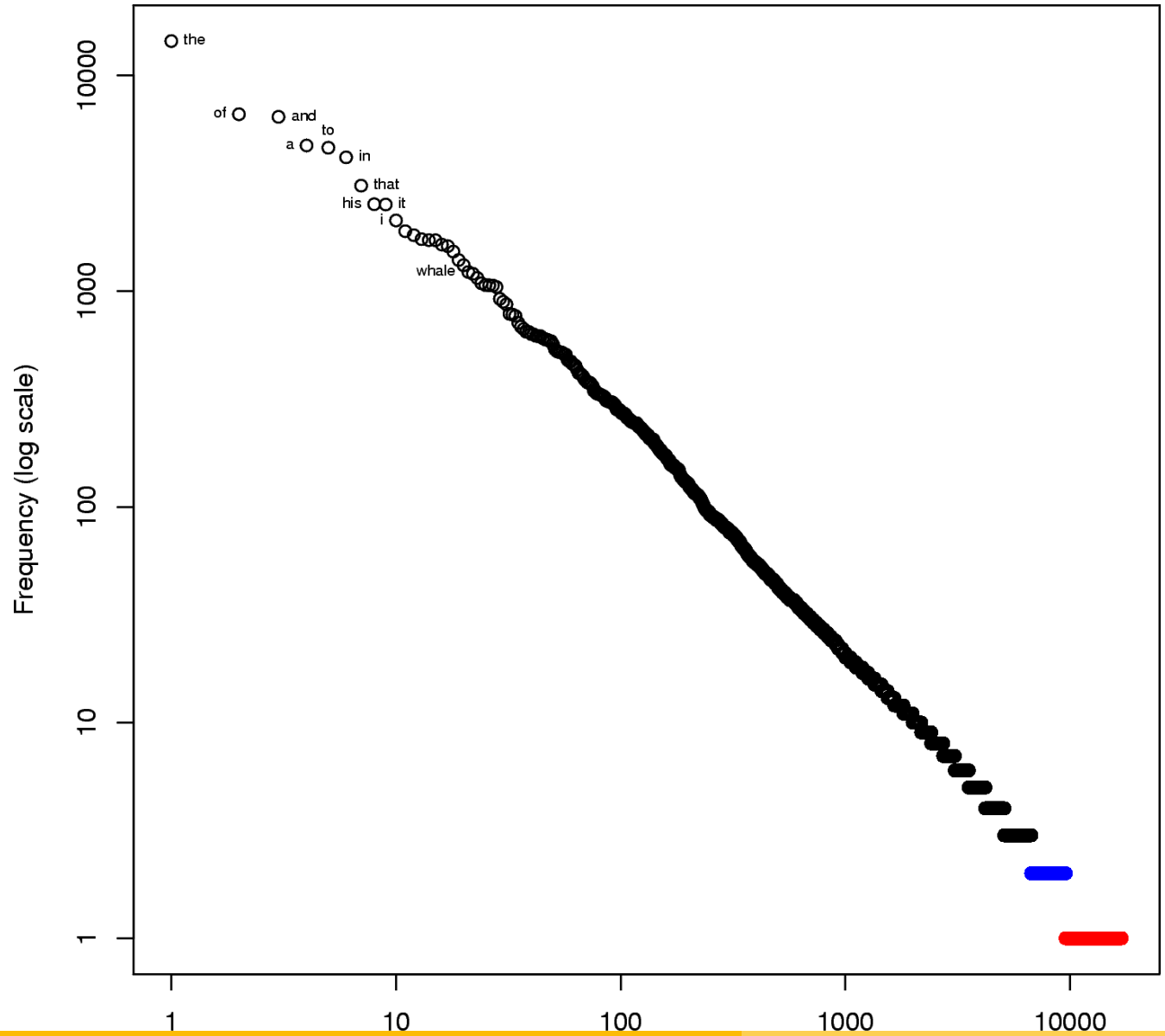
Zipfův zákon

- graf: pořadí – četnost
- $f * r = C$



Zipfův zákon

Velká část slov je s četností 1 = *hapax legomena*



Redukovaná četnost

- pravděpodobnost výskytu slova vs. četnost slova v korpusu
- některá slova jsou pouze v jenom dokumentu, ale mnohokrát
- redukované četnosti normalizují výskyty
- $RF <= F$
- $RF >= 1$
- dokumentová četnost (range)
  = počet dokumentů obsahujících dané slovo

ARF: average reduced frequency

- četnost slova f
- korpus rozdělíme na f částí
- počítáme počet částí, které dané slovo obsahují
  min = 1, max = f
- rozdělení posuneme o jeden token, znovu spočteme počet částí
- ARF = průměr počtu pokrytých částí ze všech posunutí

ALDF: Average Logarithmic Distance Frequency

- četnost slova f, velikost korpusu N
- d_i počet tokenů mezi jednotlivými výskyty slova
- $ALD = \frac{1}{N} * \sum_{i=1}^{f} d_i \log_2 d_i$
- $ALDF = 2^{ALD}$
- rovnoměrné rozložení: $ALDF = f$
- výpočet lze distribuovat, snadné skládání částí

Kolokace

Jaká slova se vyskytují v kontextech daného výrazu?

- záchodové XXX
- tratoliště XXX
- očitý XXX
- polehčující XXX

Kolokace

Jaká slova se vyskytují v kontextech daného výrazu?

- záchodové XXX
- tratoliště XXX
- očitý XXX
- polehčující XXX

Jak můžeme kolokace najít?

- četnosti
- relativní četnosti
- skóre – asociační míry
- filtrování

Asociační míry

Počítáme na základě kontingenční tabulky.

	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

expected frequencies

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed frequencies

O_{ij} – pozorované hodnoty (observed) E_{ij} – očekávané hodnoty (expected)

Asociační míry

- T-score: $T = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} = \frac{f_{xy} - \frac{f_x f_y}{N}}{\sqrt{f_{xy}}}$
- MI-score: $MI = \log_2 \frac{O_{11}}{E_{11}} = \log_2 \frac{f_{xy} N}{f_x f_y}$
- Log-likelihood:
  $LL = -\log_2 \frac{L(O_{11}, C_1, r) L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1) L(O_{12}, C_2, r_2)}$
  $L(k, n, r) = r^k (1 - r)^{n-k}$
  $r = \frac{R_1}{N}; r_1 = \frac{O_{11}}{C_1}; r_2 = \frac{O_{12}}{C_2}$

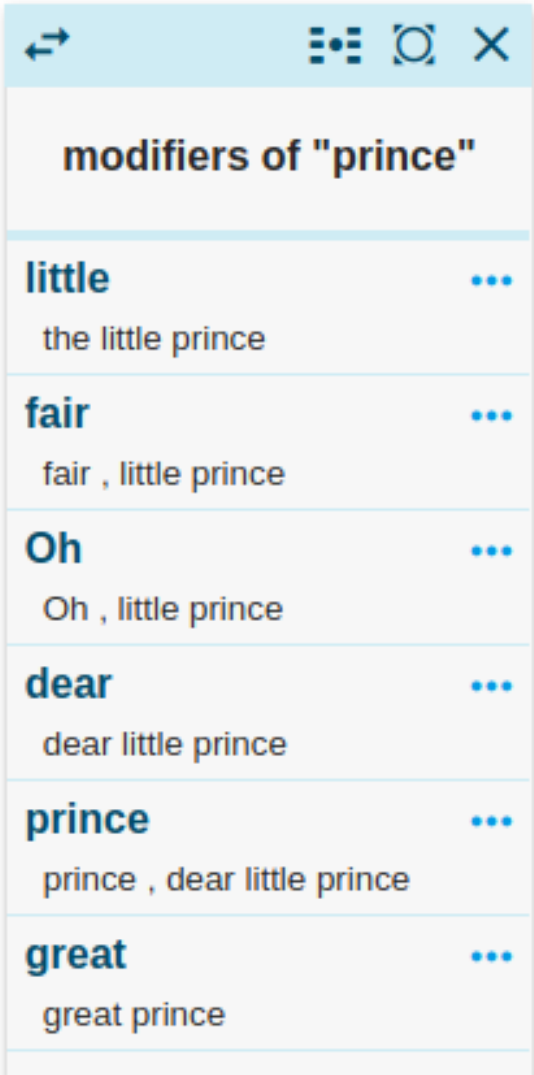
Asociační míry

- Minimum sensitivity: $MS = \min\{\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\} = \min\{\frac{f_{xy}}{f_x}, \frac{f_{xy}}{f_y}\}$
  – minimum z relativních četností
- Dice: $D = \frac{2O_{11}}{R_1+C_1} = \frac{2f_{xy}}{f_x+f_y}$
- logDice: $lD = 14 + \log_2 D = 15 + log_2 f_{xy} - log_2(f_x + f_y)$

Filtrování

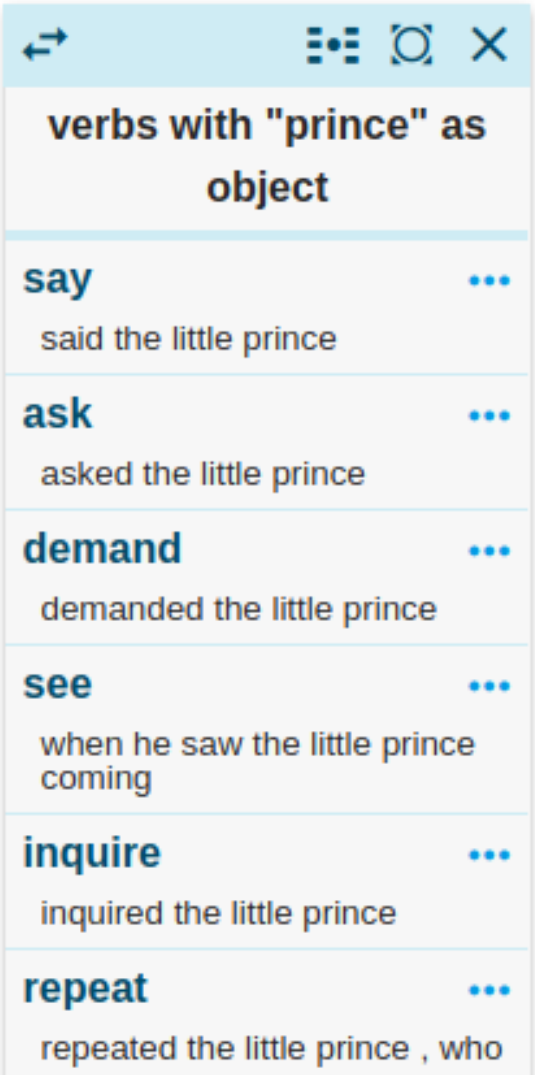
- vybíráme jen ty kolokace, které splňují podmínku na značkách
- ADJ NN
- NN NN
- word sketches – jednostránkový souhrn chování slov

Kolokace slova Prince



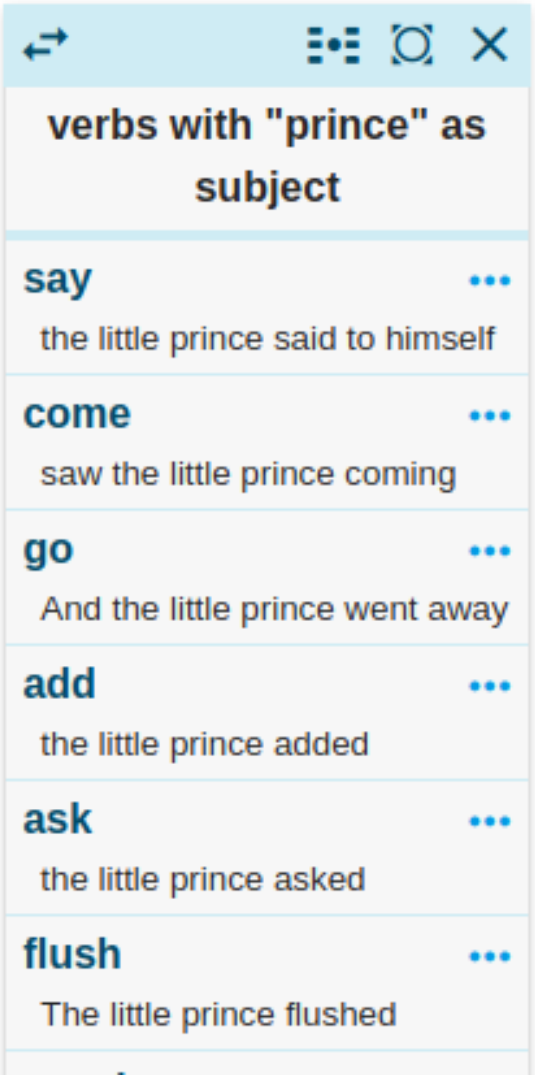
modifiers of "prince"

little	the little prince
fair	fair , little prince
Oh	Oh , little prince
dear	dear little prince
prince	prince , dear little prince
great	great prince



verbs with "prince" as object

say	said the little prince
ask	asked the little prince
demand	demanded the little prince
see	when he saw the little prince coming
inquire	inquired the little prince
repeat	repeated the little prince , who



verbs with "prince" as subject

say	the little prince said to himself
come	saw the little prince coming
go	And the little prince went away
add	the little prince added
ask	the little prince asked
flush	The little prince flushed

Word Sketches

Jak se vytváří

- Velký vyvážený korpus
- Vyhledáme závislé prvky (subjects, objects, heads, modifiers etc)
- Gramatické relace definované formou dotazů v CQL
- Seznam kolokací pro každou gramatickou relaci
- Statistika pro třídění každého seznamu

Klíčová slova

- porovnání relativních četností oproti referenčním datům
- Simple Math:
  $s = \frac{f_1 + N}{f_{ref} + N}$

Klíčová slova

- porovnání relativních četností oproti referenčním datům
- Simple Math:
  $s = \frac{f_1 + N}{f_{ref} + N}$

Co je v korpusu/textu?
Klíčová slova dávají typická (častá) slova, ale ne obyčejná.