

Využití korpusů

při výuce češtiny jako cizího jazyka

PLIN022

2.

ÚKOL

Praktický úkol – frekvenční analýza

Zipfův zákon – souvislost mezi frekvencí výskytu slova a jeho pořadím

Mluvnice současné češtiny (MSC; 2010): Když seřadíme různá slova v určitém korpusu od nejfrekventovanějšího po nejméně časté a přiřadíme ke každému slovu číslo označující jeho pořadí, můžeme si všimnout, že pořadí slova krát jeho frekvence je víceméně konstantní.

→ v jazyce existuje **málo slov s vysokou frekvencí** (gramatická slova)

→ **většina lexikonu slova s nízkou frekvencí**

(MSC 2010, s.78):

- nejfrekventovanějších **100 slov** pokrývá téměř **40 % textu**
- **1000 slov** tvoří **62 % textu**
- se znalostí **3000 slov** jsme schopni rozumět více než **75 % textu**

Praktický úkol – frekvenční analýza

- 1) Napište **slova** (3–5), která mají podle vás **největší frekvenci**.
- 2) Napište slova (3–5) od slovních druhů:
 - substantiva
 - adjektiva
 - slovesa
 - předložky
 - spojky
- 3) Srovnání s frekvenčním seznamem z korpusu SYN2015.
Co vás překvapilo?

Tabulka frekvenční distribuce slovní zásoby

KonText, SYN2015

typ dotazu **lemma** = základní (slovníkový) tvar, není ovlivněn různou frekvencí výskytů (různých pád. tvarů)

vyhledávací řádek: **symbol** .* = vyhledání všech slov o délce alespoň jednoho písmena

omezení: pouze beletrie a publicistika (oborová literatura ne)

The screenshot shows the KonText search interface for the SYN2015 corpus. The search bar contains the query ".*". The search is filtered by document type (doc.txttype_group) to include "FIC: beletrie" and "NMG: publicistika", while "NFC: oborová literatura" is excluded. The search type is set to "lemma".

Search filters and results:

doc.txttype_group (info)	Count
<input checked="" type="checkbox"/> FIC: beletrie	41 619 587
<input type="checkbox"/> NFC: oborová literatura	39 384 709
<input checked="" type="checkbox"/> NMG: publicistika	39 744 419

Search parameters and options:

- syn2015 (selected corpus)
- Pokročilý dotaz (Advanced search)
- Shoda velikosti písmen (Case sensitivity)
- Povolit regulární výrazy (Allow regular expressions)
- Výchozí atribut: lemma (Default attribute)
- Omezit hledání (Restrict search)
- Uložit jako koncept subkorpusu (Save as subcorpus concept)
- Uložit seznam dokumentů (Save document list)
- Zúžit výběr (Narrow selection)
- Krok zpět (Back step)
- Zrušit výběr (Cancel selection)

Pořadí	Vše	Substantiva	Adjektiva	Slovesa	Předložky	Spojky
1	být	rok	velký	být	v	a
2	se	člověk	celý	mít	na	že
3	a	den	další	moci	s	ale
4	v	ruka	dobrý	chtít	z	i
5	ten	dítě	nový	řici	do	jako
6	na	život	jiný	muset	o	když
7	on	místo	malý	vědět	k	aby
8	že	hlava	český	jít	za	nebo
9	s	doba	poslední	stát	po	než
10	z	oko	starý	dát	pro	ani
11	mít	muž	rád	říkat	od	protože
12	do	žena	vysoký	vidět	u	však