

Kvantitativní analýza textů nerodilých mluvčích češtiny

Michaela Hanušková & Michaela Nogolová



OSTRAVSKÁ UNIVERZITA
FILOZOFICKÁ FAKULTA

Workshop: Žákovský korpus nerodilých mluvčích češtiny
25.-26. 1. 2023

Lexikální analýza

- Lexikální bohatství (MATTR)
 - 6073 textů
 - 55 slov +
- Do budoucna
 - Morfologické bohatství
 - Délka tokenů a typů
 - Klouzavé průměry - entropie, repeat rate
- V zahraničí
 - Lexical sophistication
 - Problém: referenční korpusy, seznamy slov

Problémy

- Jazykové úrovně a C2
- Nevyváženost jazyků

Syntaktická analýza

- Průměrná délka věty
 - V počtu slov (ASL)
 - V počtu klauzí
- Průměrná délka klauze
 - V počtu slov
 - V počtu lineárních dependenčních segmentů
- Průměrná dependenční vzdálenost (MDD)

Jazykový materiál

- 5 721 textů z žákovského korpusu CzeSL-SGT
 - A1–C1
 - 55 slov
 - 5 vět

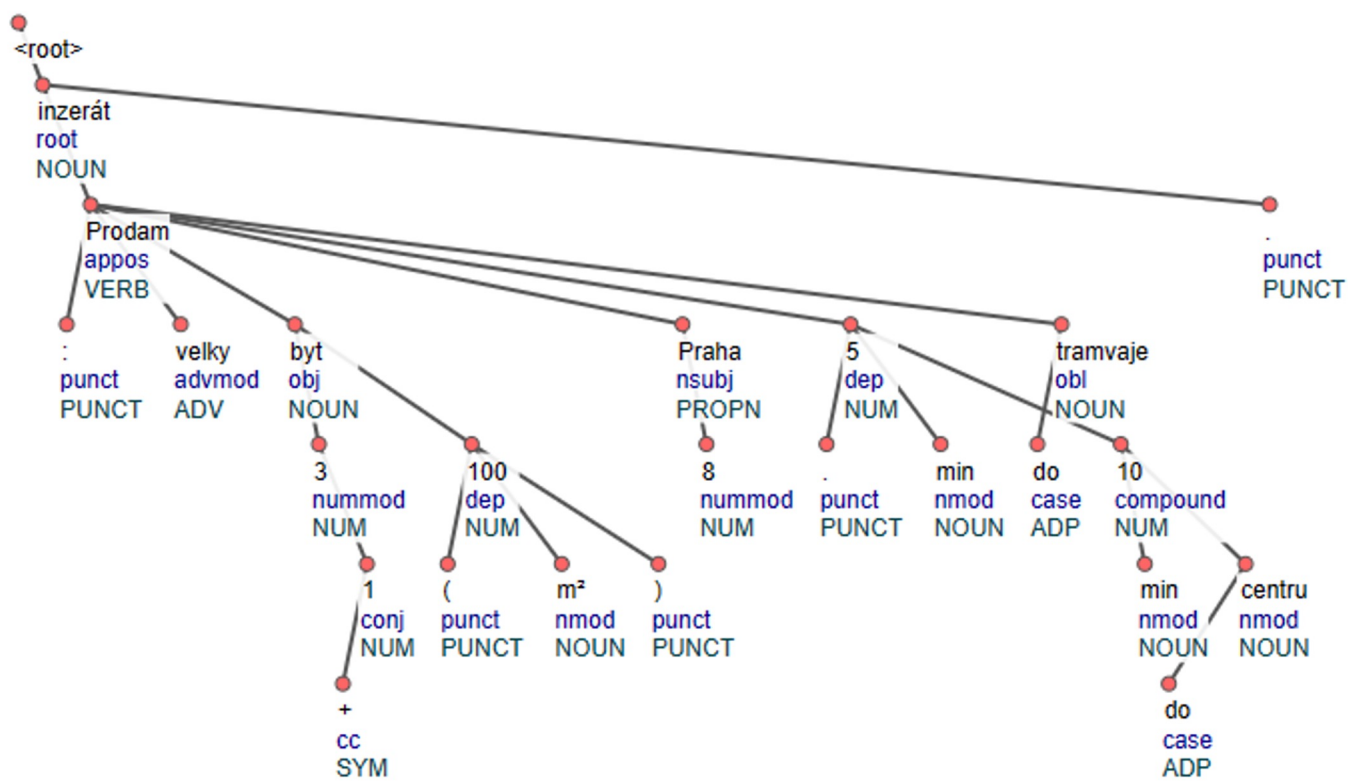
Jazykový materiál – problémy

- Texty se špatně užitými predikáty.

Jazykový materiál – problémy

- Texty se špatně užitými predikáty

inzerát : Prodam velky byt 3 + 1 (100 m²) Praha 8. 5 min do tramvaje 10 min do centru. částečno zařízeny (3. patro) hezký světlý byt vedle parku. cena : 4000000 Kč
Tel : QQQ idealny pro mladou rodiny



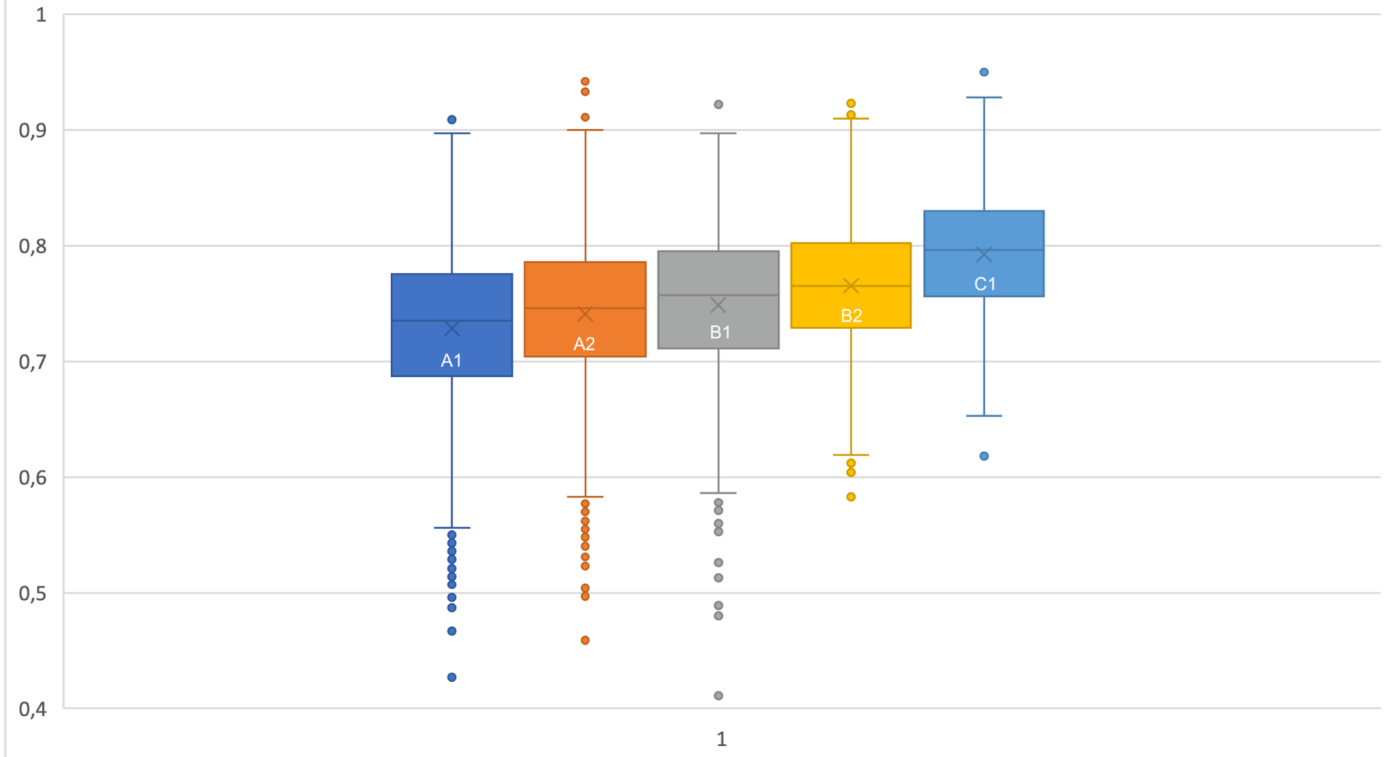
Jazykový materiál – problémy

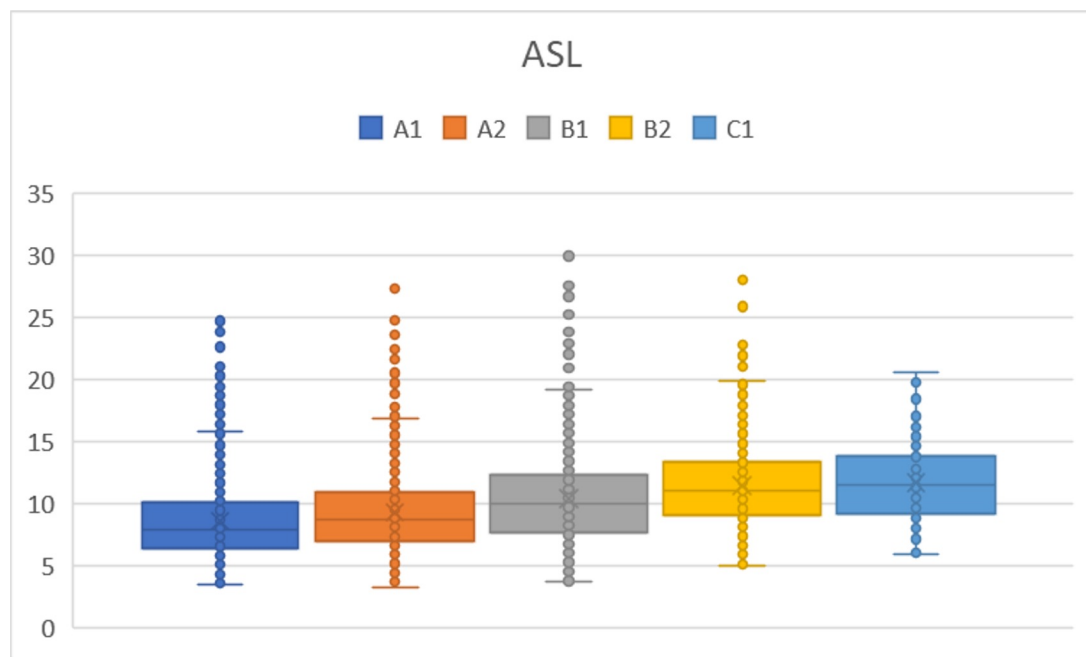
- C2 – pouze jeden text
- Texty o kratší délce než 55 slov.

Jazykový materiál – klady

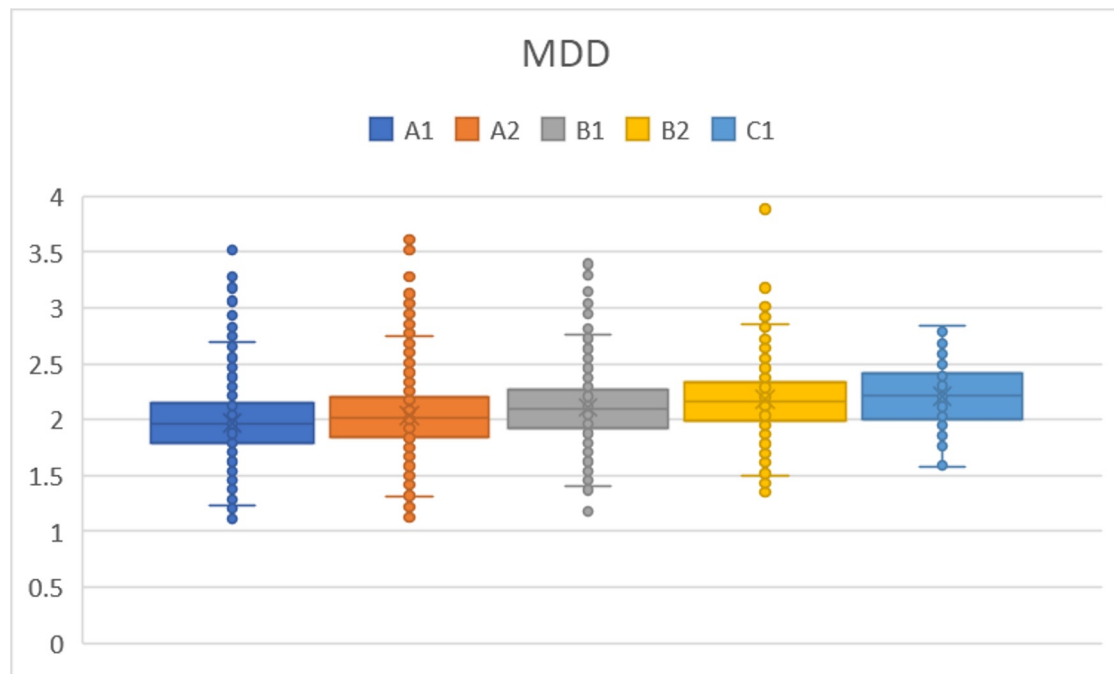
- Dostatečné množství textů na úrovních A1–C1
- Výsledky ukazují, že velká většina textů je zařazena do správné úrovně

MATTR mix (+55 slov)





ASL	A1	A2	B1	B2
A2	<0.001			
B1	<0.001	<0.001		
B2	<0.001	<0.001	<0.001	
C1	<0.001	<0.001	<0.001	0.336



MDD	A1	A2	B1	B2
A2	<0.001			
B1	<0.001	<0.001		
B2	<0.001	<0.001	<0.001	
C1	<0.001	<0.001	<0.001	0.191

Budoucí plán

- Shluková analýza