Měřit bibli.

Stylometrická diachronní analýza

Radek Čech

(Ostravská univerzita)

Pavel Kosek

(Masarykova univerzita)

Ján Mačutek

(Matematický ústav, Slovenská akadémia vied; Univerzita Konštantína Filozofa v Nitre)

Michaela Nogolová

(Ostravská univerzita)

Měřit bibli?

Stylometrická diachronní analýza

Radek Čech

(Ostravská univerzita)

Pavel Kosek

(Masarykova univerzita)

Ján Mačutek

(Matematický ústav, Slovenská akadémia vied; Univerzita Konštantína Filozofa v Nitre)

Michaela Nogolová

(Ostravská univerzita)

Poděkování

- Pavel Kosek: Gramatika a lexikon češtiny (MUNI/A/1181/2020)
- Ján Mačutek: VEGA 2/0096/21

Diachronní (ne)stabilita

- vliv jazykových změn na podobu textu
- různá dynamika
  - lexikální rovina
  - morfologická rovina
  - fonologická rovina

Diachronní (ne)stabilita

[1] Zěvenie Jezukristovo, ješto jest dal i zěvil bóh otec, aby je prohlásil sluhám svým o těch věcech, ješto sě mají státi skoro, vzkázav skrzě svého anjela služě svému Janovi,

(Bible olomoucká, 1417)

[1] Zjevení Ježíše Krista, které mu dal Bůh, aby svým otrokům ukázal, co se má brzy stát. On to prostřednictvím svého anděla naznačil svému otroku Janovi.

(Český studijní překlad Bible, 2009)

Diachronní (ne)stabilita

[1] Zěvenie Jezukristovo, ješto jest dal i zěvil bóh otec, aby je prohlásil sluhám svým o těch věcech, ješto sě mají státi skoro, vzkázav skrzě svého anjela služě svému Janovi,

(Bible olomoucká, 1417)

[1] Zjevení Ježíše Krista, které mu dal Bůh, aby svým otrokům ukázal, co se má brzy stát. On to prostřednictvím svého anděla naznačil svému otroku Janovi.

(Český studijní překlad Bible, 2009)

Diachronní (ne)stabilita

[1] Zěvenie Jezukristovo, ješto jest dal i zěvil bóh otec, aby je prohlásil sluhám svým o těch věcech, ješto sě mají státi skoro, vzkázav skrzě svého anjela služě svému Janovi,

29 tokenů

(Bible olomoucká, 1417)

[1] Zjevení Ježíše Krista, které mu dal Bůh, aby svým otrokům ukázal, co se má brzy stát. On to prostřednictvím svého anděla naznačil svému otroku Janovi.

25 tokenů

(Český studijní překlad Bible, 2009)

Cíl analýzy

- modelování vývojové dynamiky
- zachycení velikosti rozdílů u vybraných textových charakteristik

- návaznost na předchozí výzkum
  - Čech, R., Mačutek, J. Kosek, P. (2021). Czech translations of the Gospel of Matthew from the diachronic point of view – plus ça change.... *Jazykovedný časopis*, 72, s. 656-666.
  - Čech, R., Kosek, P., Mačutek, J., Nogolová, M. (Ne)stabilita českého biblického překladu. Diachronní stylometrická analýza. (v recenzním řízení)

Cíl analýzy

- zachycení velikosti rozdílů u vybraných textových charakteristik
  - proporce identických slovních forem (PIF)
  - délka slova (typu) (N)
  - proporce hapax legomen (PHL)
    - klouzavý průměr TTR (MATTR)
    - entropie (H)
  - délka textu (N)
  - analýza nejfrekventovanějších slov (MFW)

Jazykový materiál

- Evangelium sv. Matouše
- Zjevení sv. Jana
- Job

Jazykový materiál

- Bible drážďanská (BiblDrážď), (70. léta 14. stol.)
- Bible olomoucká (BiblOl), (1417)
- Bible mlynářčina (BiblMlyn), (3. čtvrtina 15. st.)
- Bible benátská (BiblBen), (1506)
- Bible Melantrichova (BiblMel), (1570)
- Bible kralická (BiblKral), (1593/1594)
- Bible svatováclavská (BiblSvat), (1677)
- Nový zákon (BiblSýk), (1909)
- Hejčlův překlad (BiblHejcl) (1917)
- Nový zákon (BiblPetrů), (1969/1992)
- Český ekumenický překlad (BiblEkum), (1995)
- Český studijní překlad Bible (BiblČSP), (2009)
- Bible 21 (Bibl21), (2014)

Proporce identických slovních forem (PIF)

- počet slovních forem, které se vyskytují v obou textech, vydělený počtem slov ve zvoleném výchozím textu

Proporce identických slovních forem (PIF)

- počet slovních forem, které se vyskytují v obou textech, vydělený počtem slov ve zvoleném výchozím textu

$$PIF = \frac{|V_i \cap V_j|}{|V_i|}$$

V_i … množina slovních tvarů (types) v prvním textu

V_j … množina slovních tvarů (types) v druhém textu

Proporce identických slovních forem (PIF)

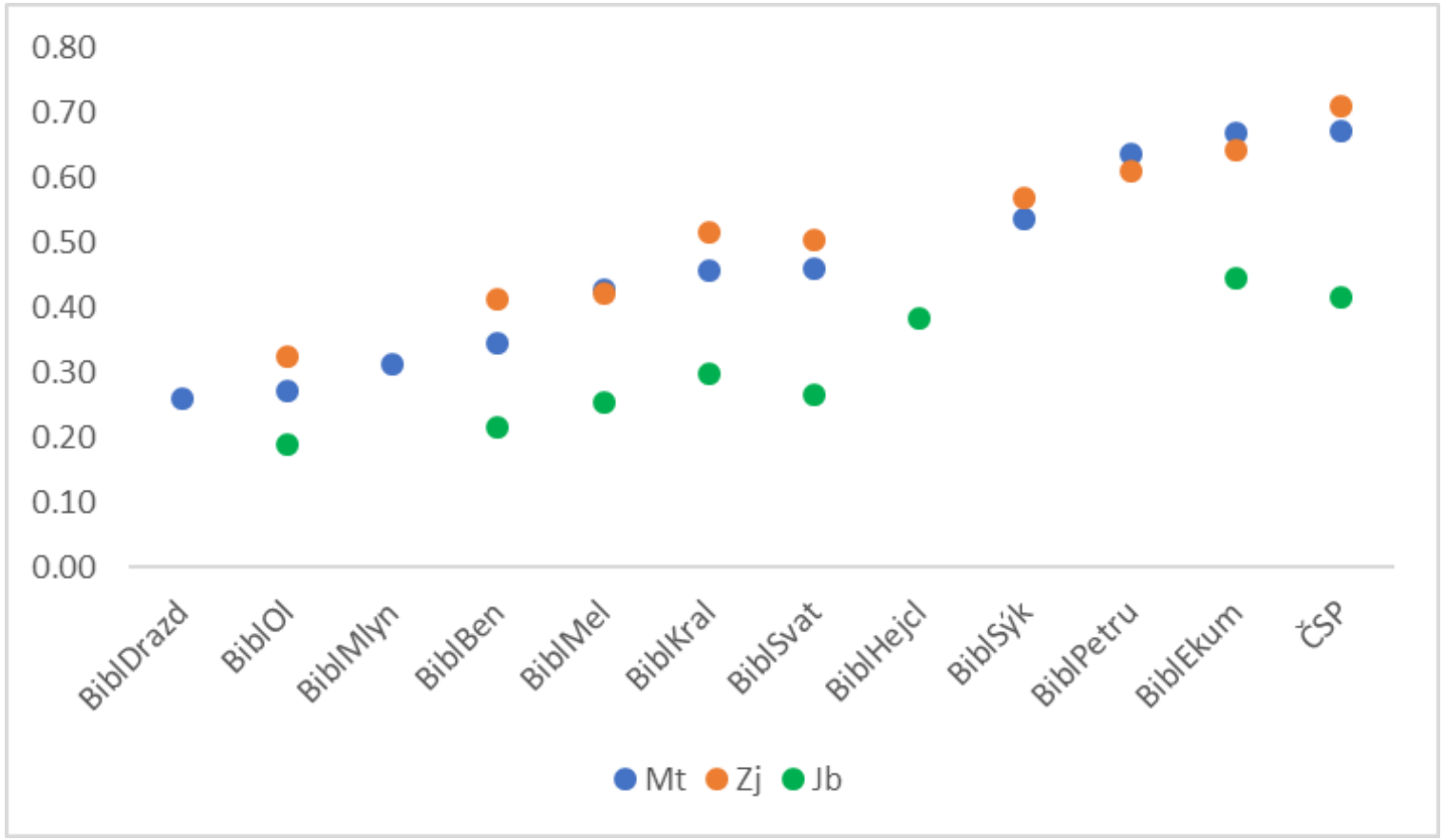
$$PIF = \frac{|V_i \cap V_j|}{|V_i|}$$

V_i … množina slovních tvarů (typů) v prvním textu

V_j … množina slovních tvarů (typů) v druhém textu

$$PIF = \frac{|V_{Bibl21} \cap V_{BiblDrážď}|}{|V_{Bibl21}|} = \frac{1075}{4150} = 0{,}259$$

Proporce identických slovních forem (PIF) vzhledem k Bibl21

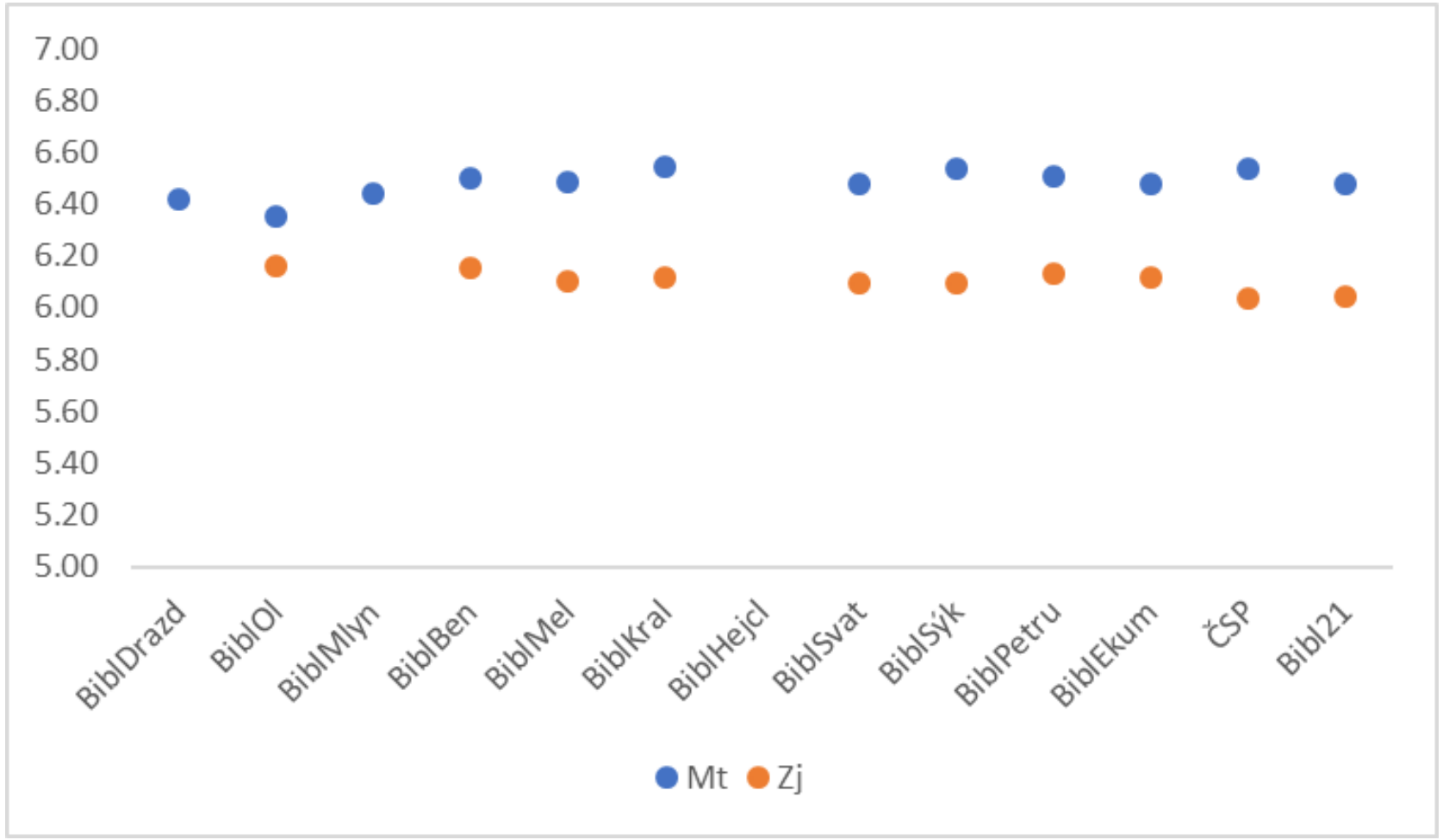


- Kendallův korelační koeficient: $\tau_{Mt} \geq 0{,}99$, $\tau_{Zj} = 0{,}94$, $\tau_{Job} = 0{,}86$ p-hodnota < 0,001

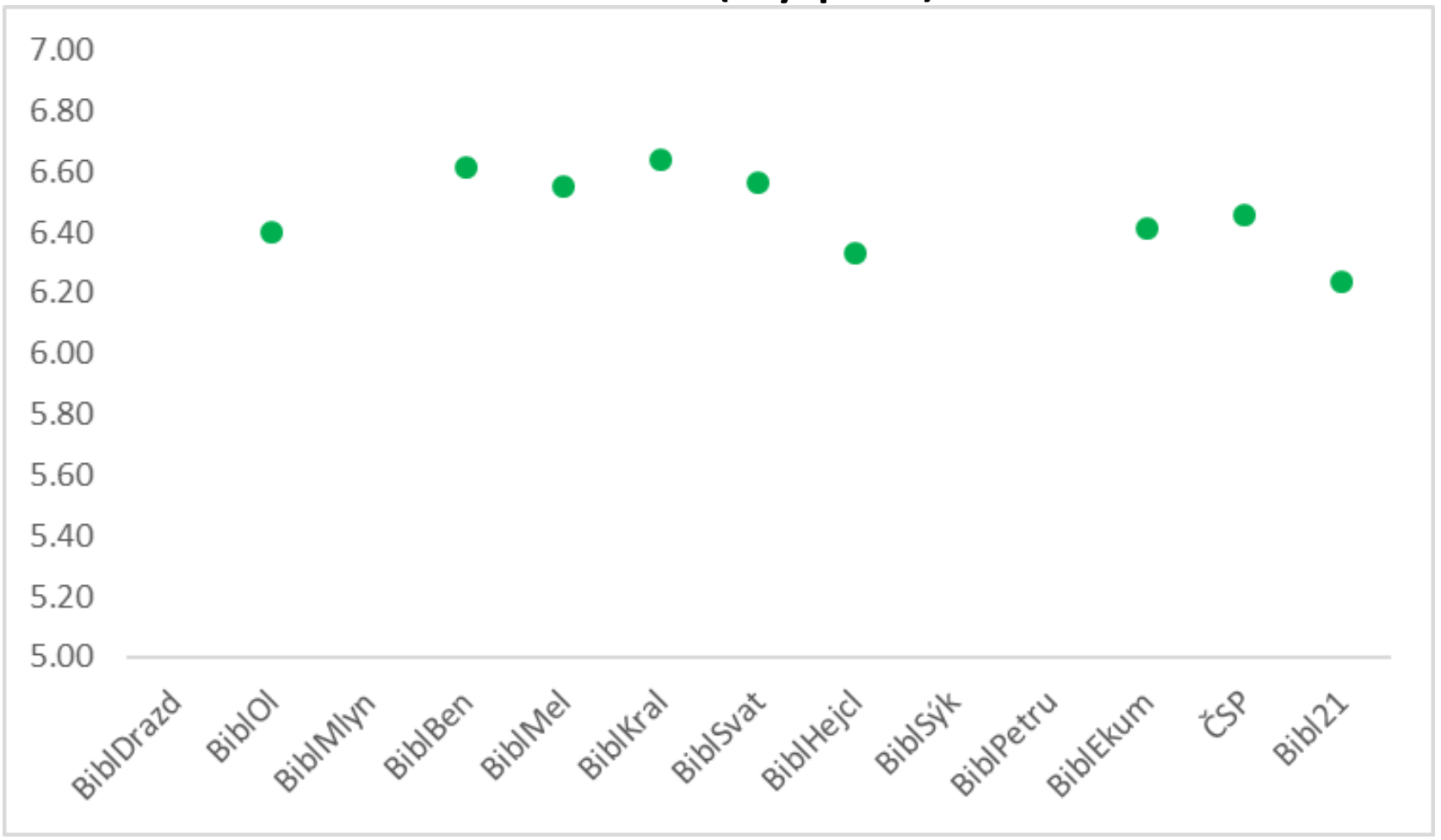
Průměrná délka slova (typu)

- délka měřena v počtu písmen
- aritmetický průměr délek slovních tvarů (typů)

Průměrná délka slova (typu)



Průměrná délka slova (typu)



- $\tau_{Mt} \geq -0{,}33$, p-value = 0,26, interval PDV: <6,24; 6,64>

Průměrná délka slova (typu)

- osmisetletý vývoj češtiny
- různé
  - předlohy
  - překladové jazyky
  - překladatelské strategie
- průměrná délka slova stabilní ve všech překladech

Proporce hapax legomen (PHL)

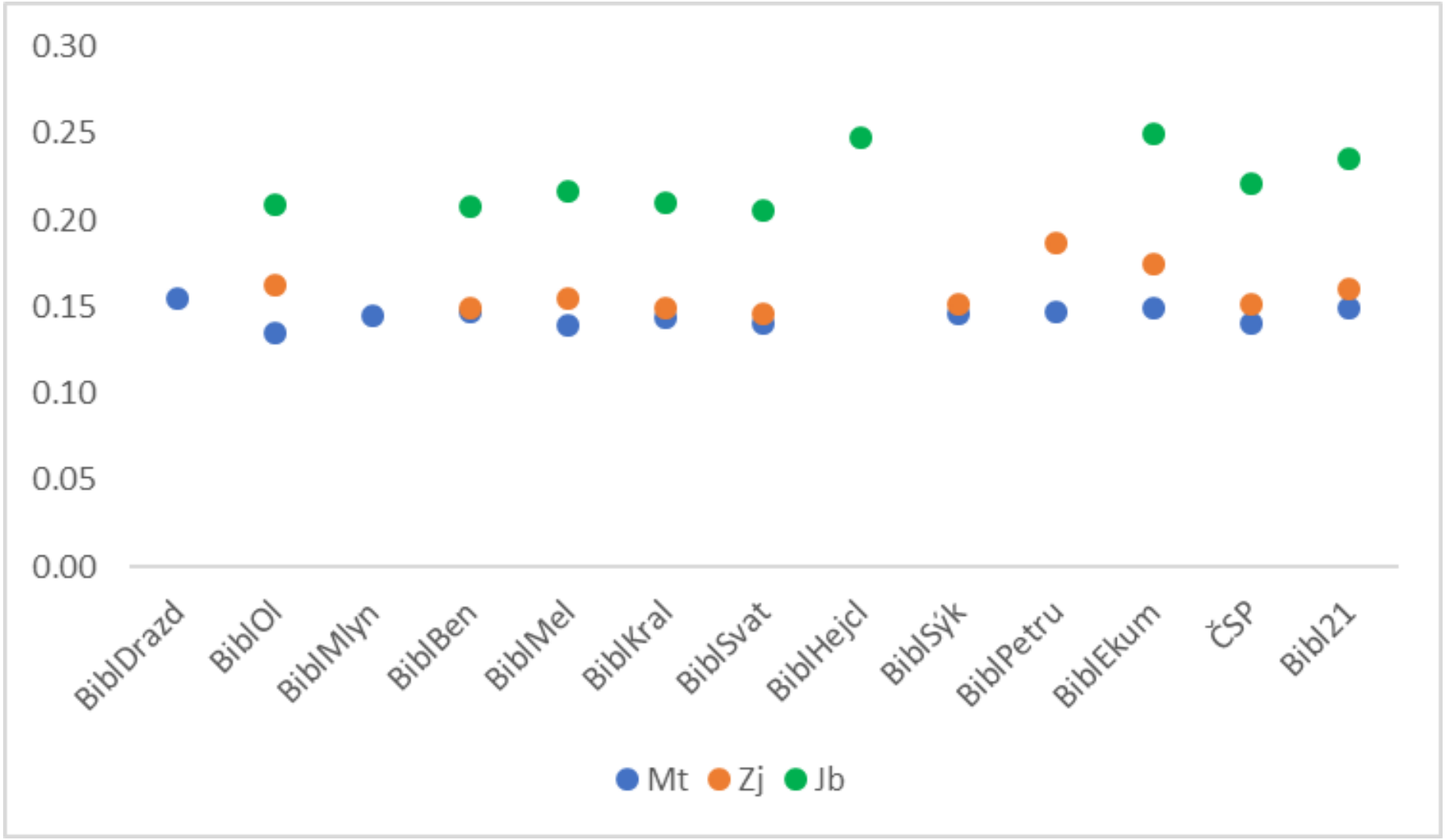
$$PHL = \frac{N_{HL}}{N}$$

N_{HL} … počet hapax legomen

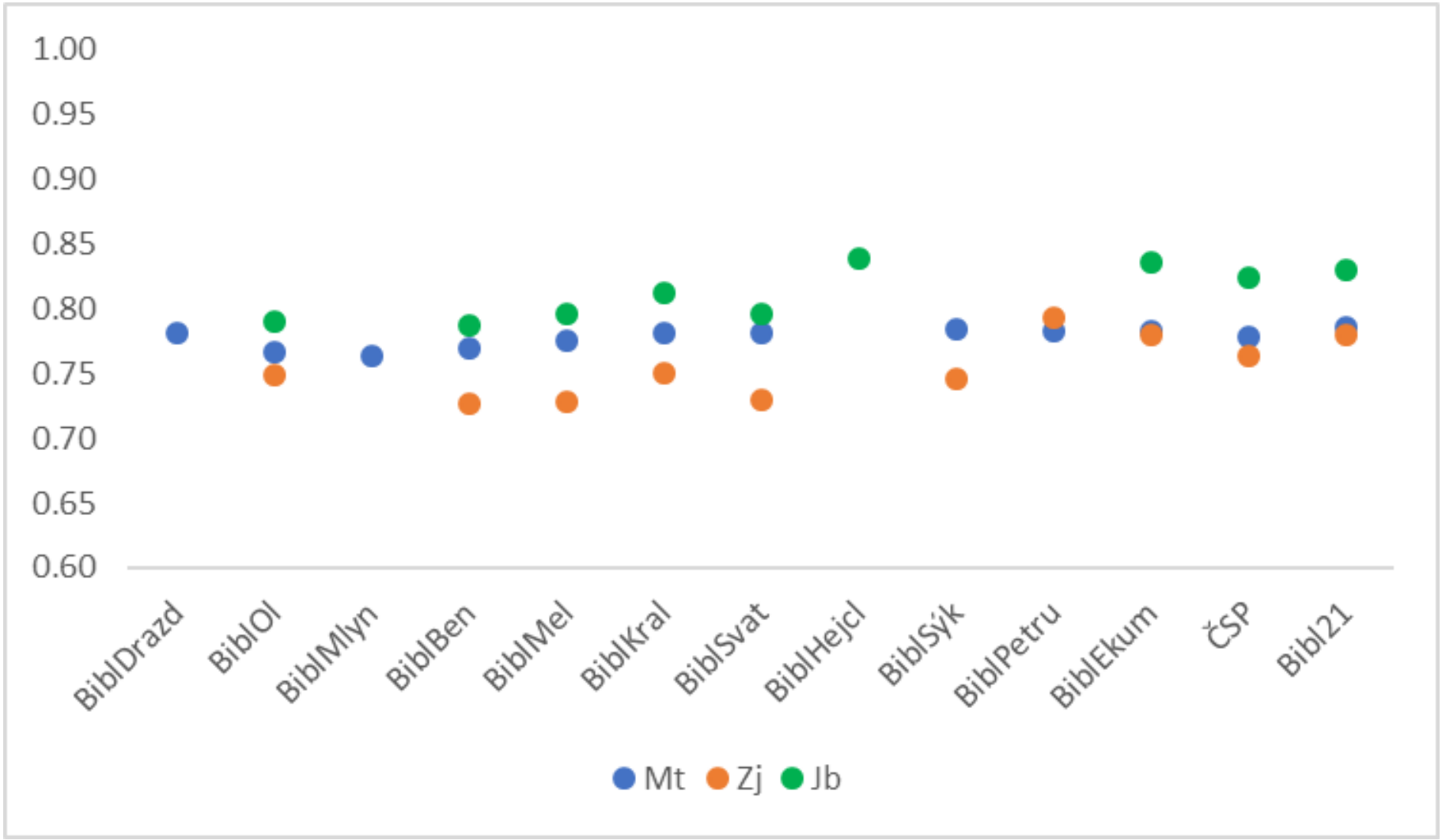
N … počet slov (tokenů) v textu

- vyjadřuje míru lexikální diverzity (slovního bohatství) textu

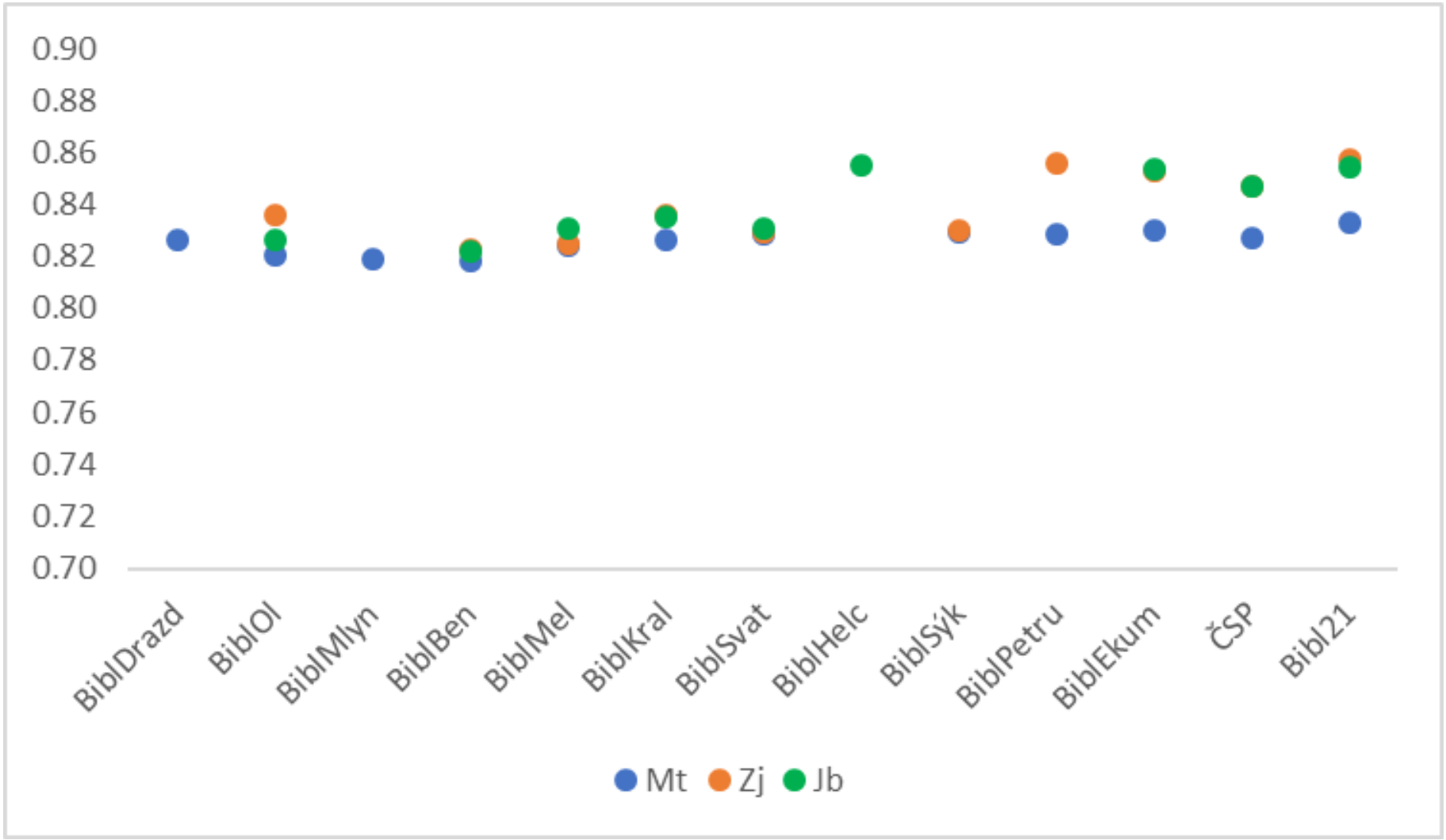
Proporce hapax legomen (PHL)



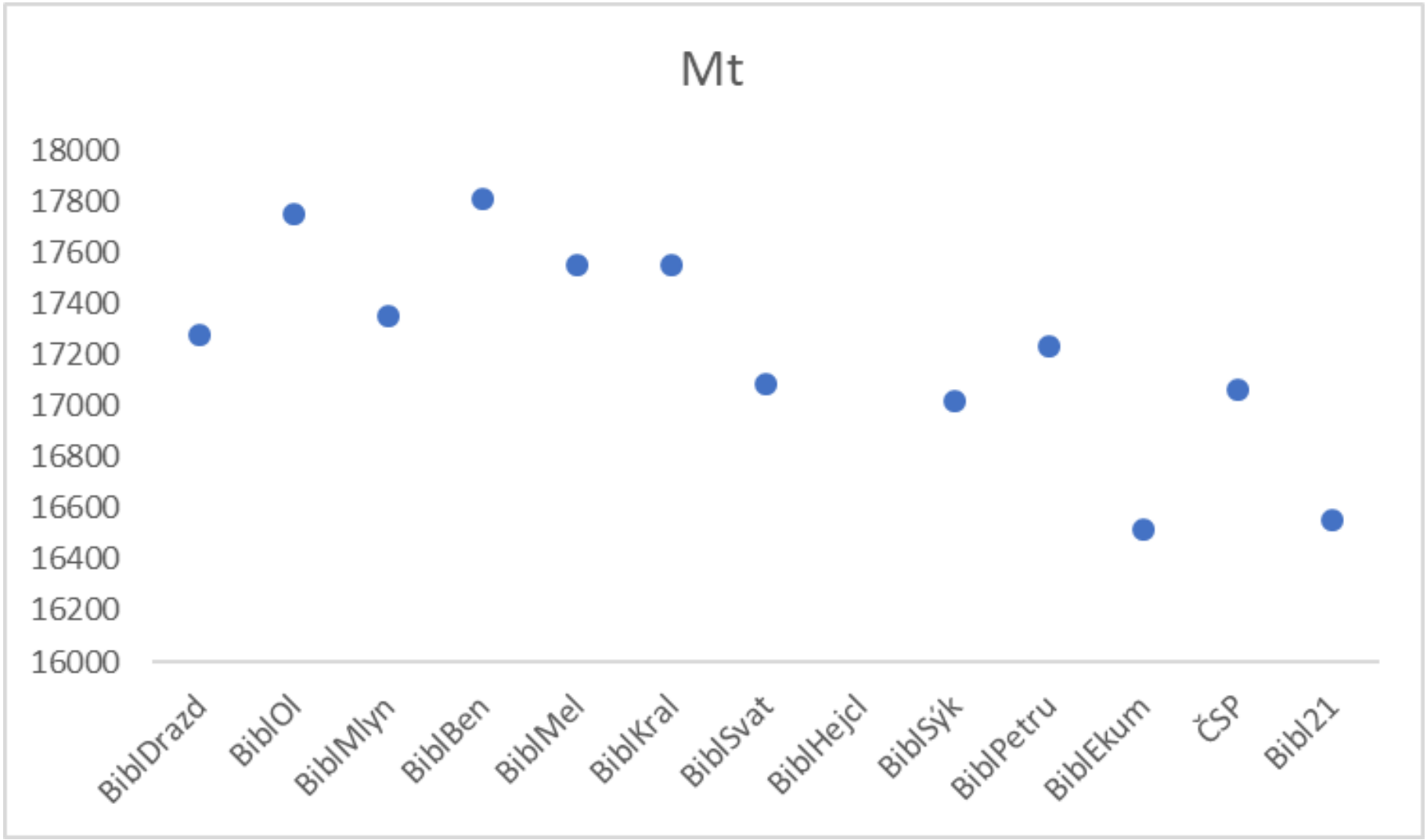
MATTR


1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
BiblDrazd
BiblOl
BiblMlyn
BiblBen
BiblMel
BiblKral
BiblSvat
BiblHejcl
BiblSýk
BiblPetru
BiblEkum
ČSP
Bibl21
Mt
Zj
Jb

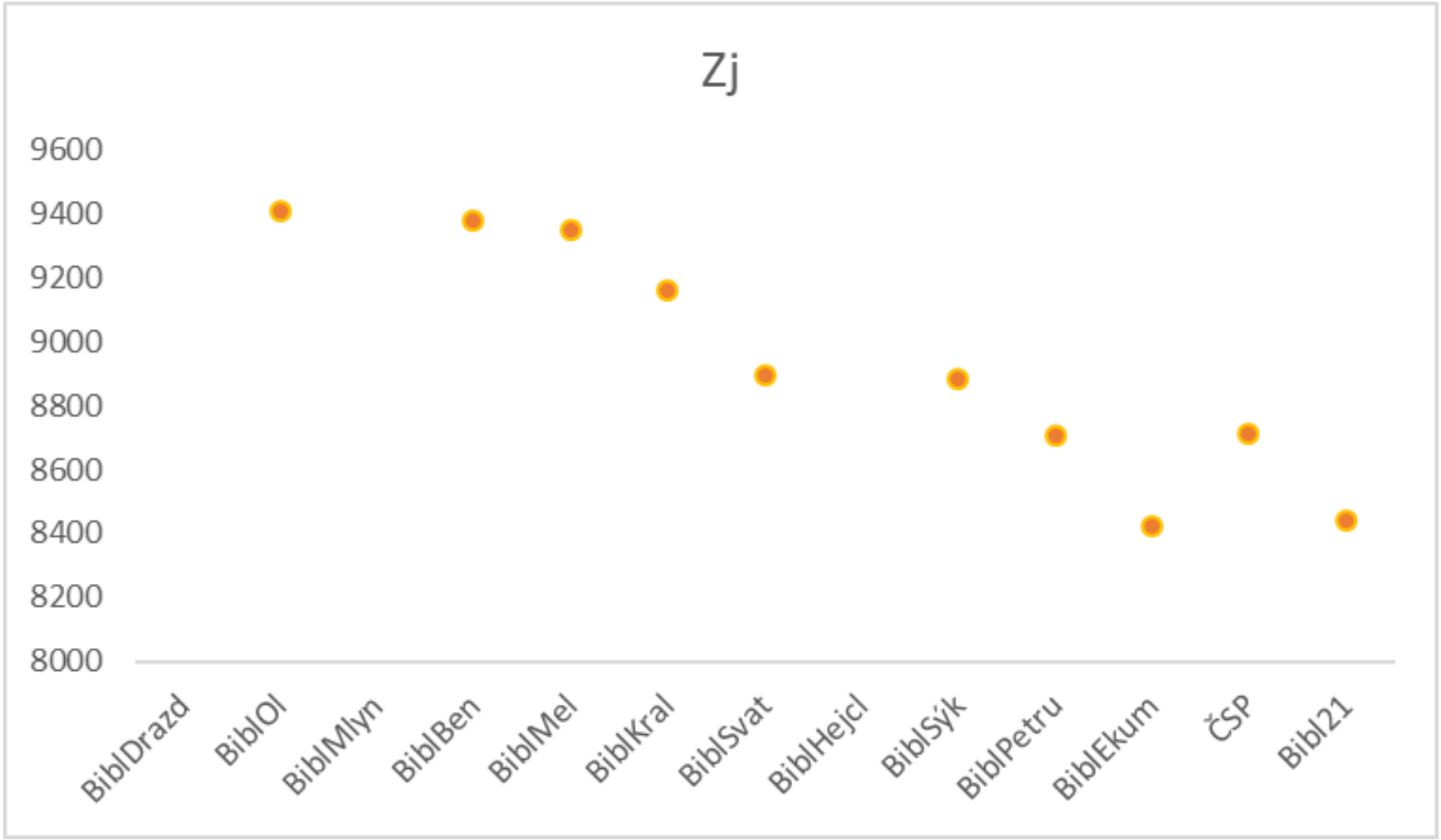
Relativní entropie



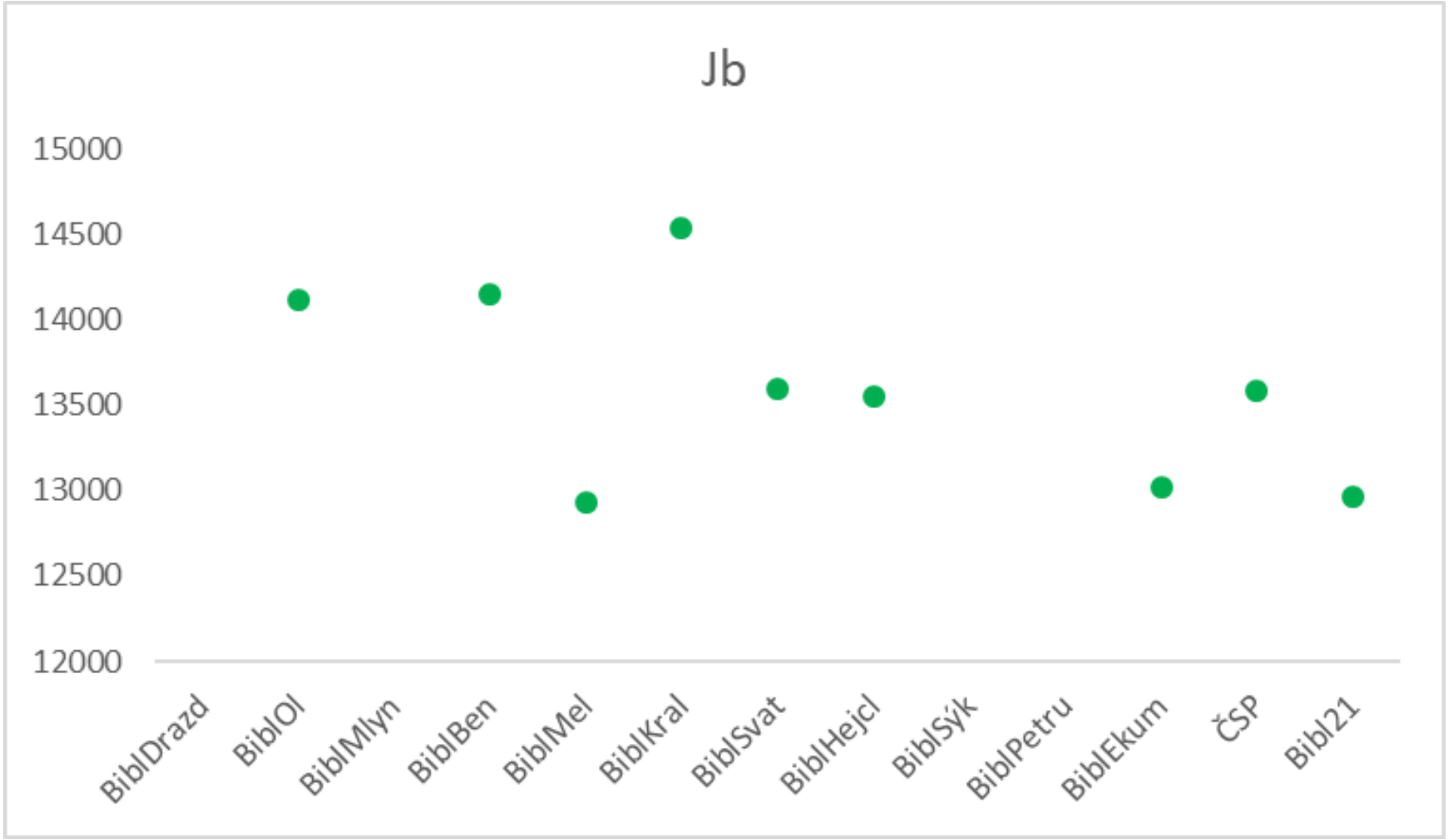
Délka textu – Evangelium sv. Matouše
(v počtu tokenů)



Délka textu – Zjevení
(v počtu tokenů)



Délka textu – Job
(v počtu tokenů)



Délka textu

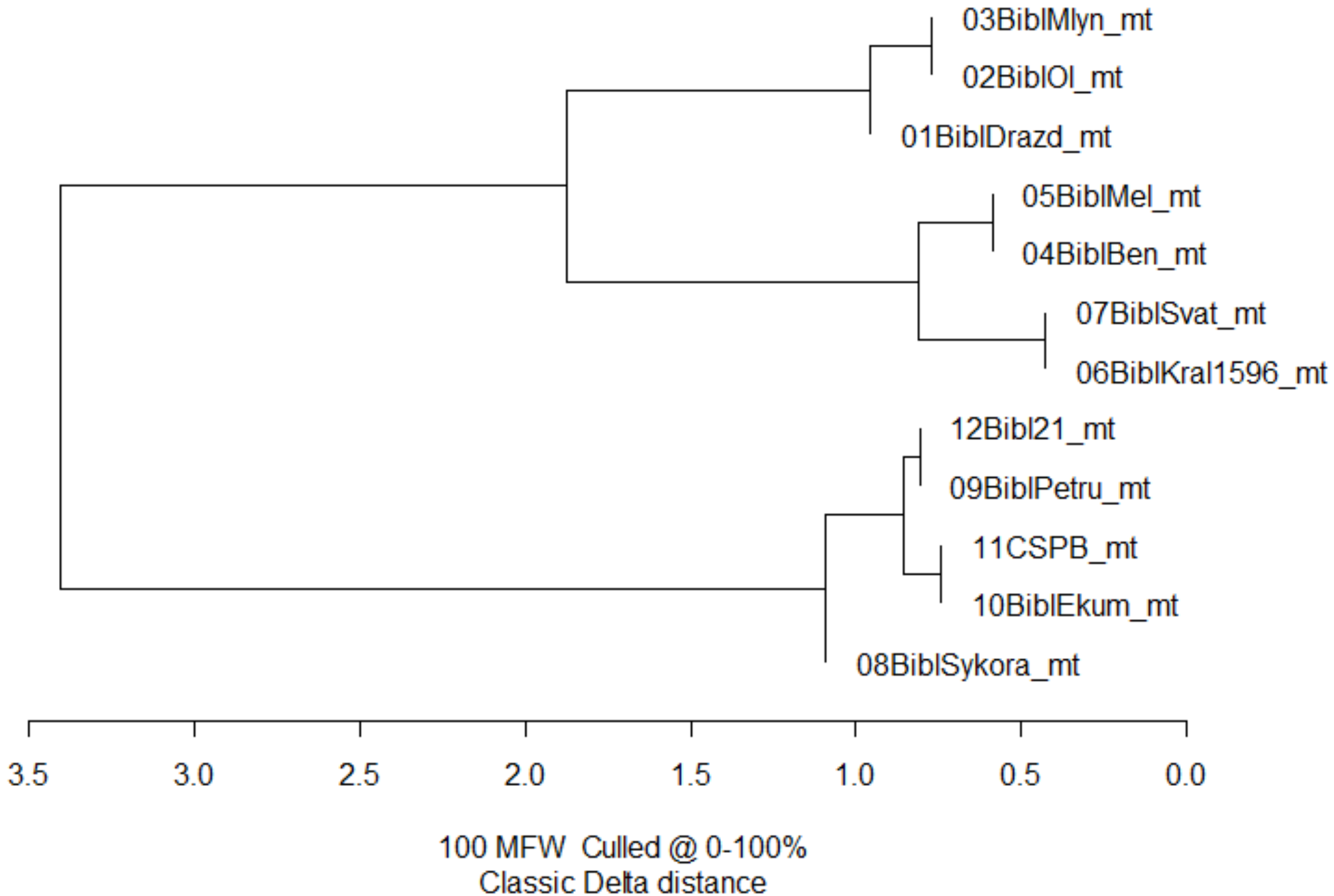
- Kendallův korelační koeficient

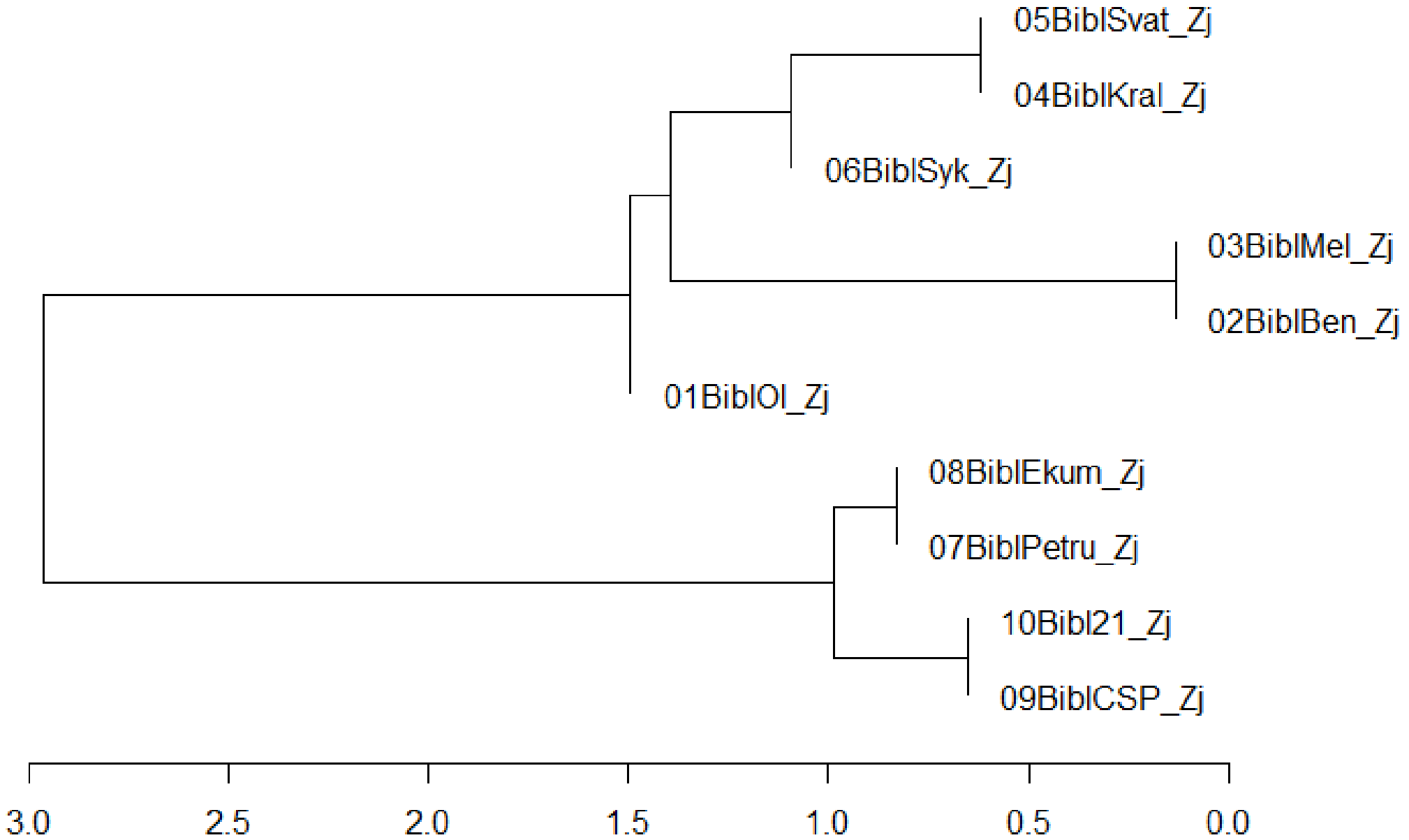
- τ_{Mt} = -0,55 p-hodnota = 0,014
- τ_{Zj} = -0,87 p-hodnota < 0,001
- τ_{Job} = -0,33 p-hodnota < 0,18

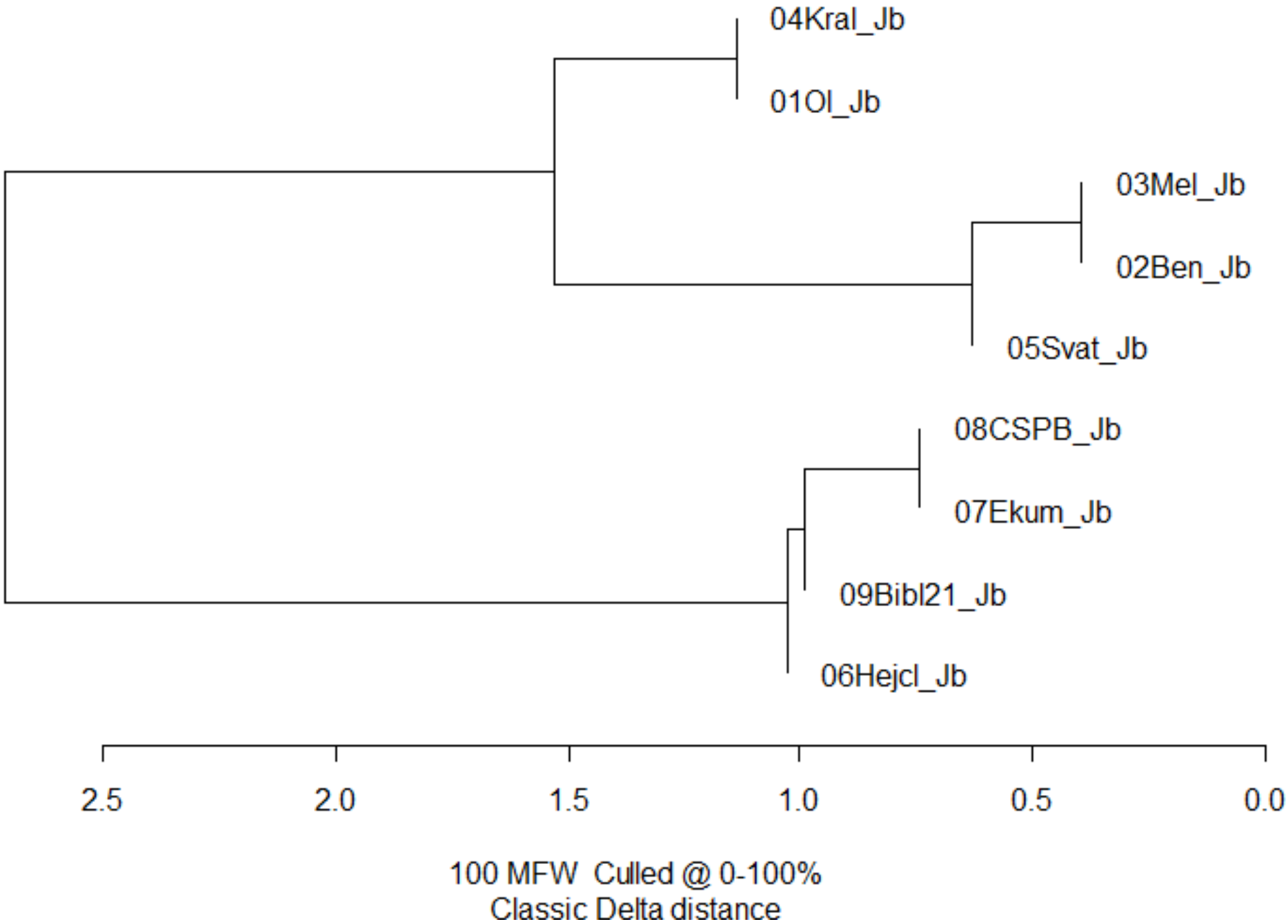
Analýza nejfrekventovanějších slov (MFW)

- Eder, Maciej, Jan Rybicki, and Mike Kestemont. "Stylometry with R: a package for computational text analysis." The R Journal 8.1 (2016).

- porovnání relativních frekvencí nejfrekventovanějších slov
- naše analýza: 100 nejfrekventovanějších slov
- Classic Delta Distances


03BiblMlyn_mt
02BiblOl_mt
01BiblDrazd_mt
05BiblMel_mt
04BiblBen_mt
07BiblSvat_mt
06BiblKral1596_mt
12Bibl21_mt
09BiblPetru_mt
11CSPB_mt
10BiblEkum_mt
08BiblSykora_mt
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0-100%
Classic Delta distance


05BiblSvat_Zj
04BiblKral_Zj
06BiblSyk_Zj
03BiblMel_Zj
02BiblBen_Zj
01BiblOl_Zj
08BiblEkum_Zj
07BiblPetru_Zj
10Bibl21_Zj
09BiblCSP_Zj
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0-100%
Classic Delta distance


04Kral_Jb
01Ol_Jb
03Mel_Jb
02Ben_Jb
05Svat_Jb
08CSPB_Jb
07Ekum_Jb
09Bibl21_Jb
06Hejcl_Jb
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0-100%
Classic Delta distance

Nástroje

- QuitaUp
  - https://korpus.cz/quitaup/
- Stylo
  - http://maciejeder.org/projects/stylo/

Závěr

- čím je text starší, tím méně slovních forem je shodných vzhlede k nejnovějšímu překladu
- překvapivá stabilita průměrné délky slov v průběhu 8 stol.
- relativně stabilní lexikální diverzita (měřena počtem hapax legomen)
- vztah mezi délkou textu a stářím překladu
  - tendence: čím je překlad mladší, tím je kratší
- MWF jako metoda měření blízkosti/vzdálenost mezi texty

Měřit bibli?

Měřit bibli.

Měřit bibli !!!

Děkujeme za pozornost!