Spočítat Husa?

Petra Mutlová

Pavel Kosek

Radek Čech

Ján Mačutek

(14. 10. 2022, Kozojedy)

Orthographia Bohemica

- authorship

- comparison with texts
  - of similar character
  - from the same period

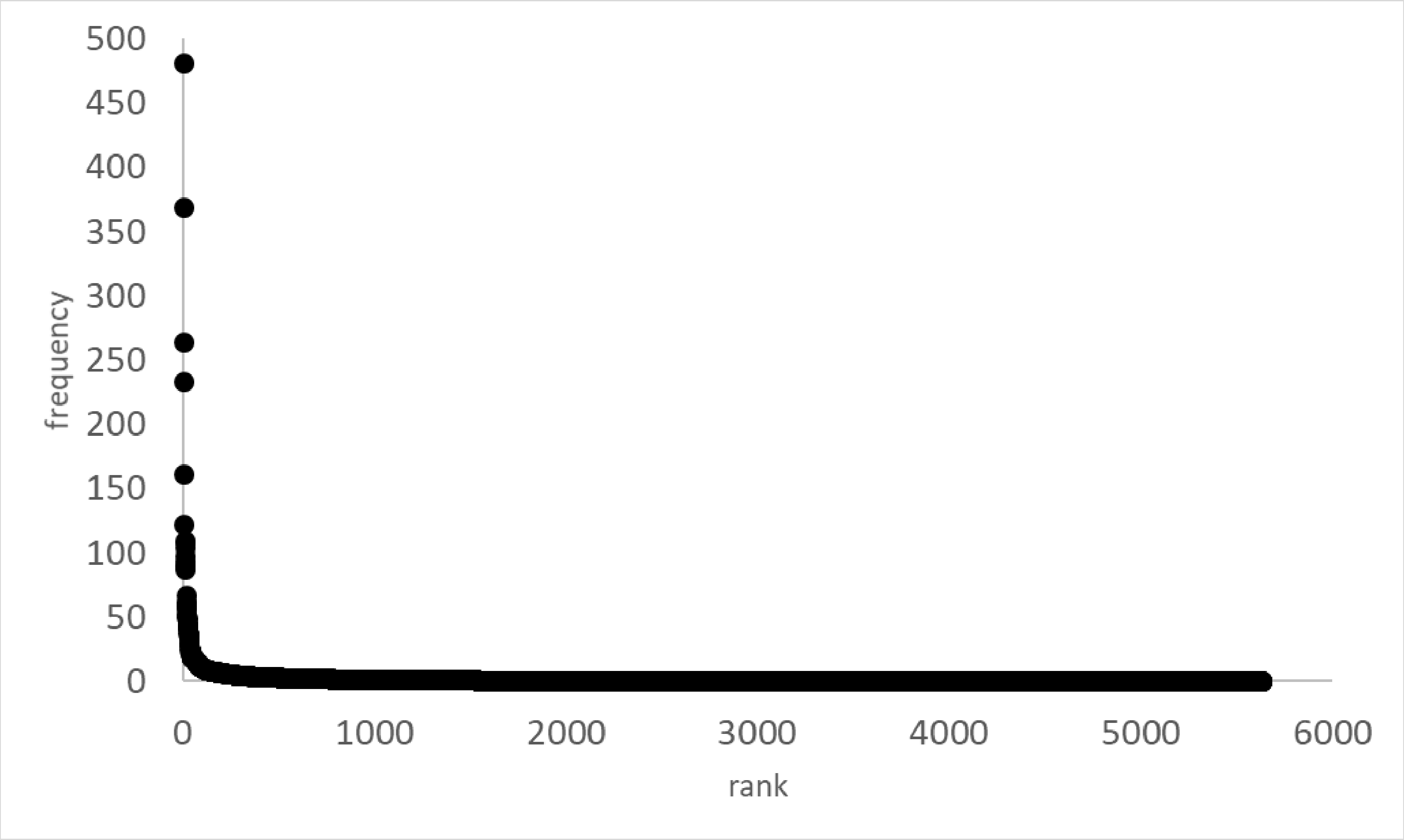
Orthographia Bohemica

- Jan Hus
    - De matrimonio, Responsiones ad Palec, Sacerdos, Sermo de pace
- Jakoubek ze Stříbra
    - Articulus, De cerimoniis, Defensio Decalogi, Magna cena
- Ondřej z Brodu
    - kvestie Utrum licitum, Planctus, rekomendace 1423, Tractus De orig. Hussitarum
- Petr z Pulky
    - Confutatio, Epistola, Sermo Ite, De congruitate
- Mikuláš z Drážďan
    - Apologia 1415, Contra Gallum, Nisi manducaveritis, Sermo 1416

Authorship & word frequencies

- lexical diversity / vocabulary richness
- proportion of hapax legomenon
- …
- **proportion of the most frequent words**

Most frequent words



J. Škvorecký: Eva byla nahá

Most frequent words

	Škvorecký: Eva byla nahá	
pořadí	slovo	f
1	a	481
2	se	369
3	na	264
4	v	234
5	jsem	161
6	s	122
7	z	110
8	američan	108
9	to	104
10	řekl	98
11	ale	94
12	do	91
13	že	89
14	řekla	87
15	dívka	67

Most frequent words

	Škvorecký: Eva byla nahá	
pořadí	slovo	f
1	a	481
2	se	369
3	na	264
4	v	234
5	jsem	161
6	s	122
7	z	110
8	američan	108
9	to	104
10	řekl	98
11	ale	94
12	do	91
13	že	89
14	řekla	87
15	dívka	67

Most frequent words

	Škvorecký: Eva byla nahá	
pořadí	slovo	f
1	a	481
2	se	369
3	na	264
4	v	234
5	jsem	161
6	s	122
7	z	110
8	američan	108
9	to	104
10	řekl	98
11	ale	94
12	do	91
13	že	89
14	řekla	87
15	dívka	67

	Hrabal: Perlička na dně	
pořadí	slovo	f
1	a	2239
2	se	1203
3	to	1037
4	na	879
5	ale	514
6	tak	504
7	do	467
8	si	459
9	jsem	456
10	v	446
11	že	440
12	je	432
13	já	363
14	když	296
15	jak	283

	Hašek: Osudy…I.	
pořadí	slovo	f
1	a	7045
2	se	6061
3	na	3927
4	že	3469
5	to	3075
6	v	2585
7	je	1801
8	do	1749
9	s	1667
10	si	1534
11	když	1387
12	z	1375
13	tak	1308
14	jsem	1286
15	švejk	1188

Most frequent words

	Škvorecký: Eva byla nahá	
pořadí	slovo	f_rel
1	a	0.037
2	se	0.028
3	na	0.020
4	v	0.018
5	jsem	0.012
6	s	0.009
7	z	0.008
8	američan	0.008
9	to	0.008
10	řekl	0.007
11	ale	0.007
12	do	0.007
13	že	0.007
14	řekla	0.007
15	dívka	0.005

	Hrabal: Perlička na dně	
pořadí	slovo	f_rel
1	a	0.054
2	se	0.029
3	to	0.025
4	na	0.021
5	ale	0.012
6	tak	0.012
7	do	0.011
8	si	0.011
9	jsem	0.011
10	v	0.011
11	že	0.011
12	je	0.010
13	já	0.009
14	když	0.007
15	jak	0.007

	Hašek: Osudy…I.	
pořadí	slovo	f_rel
1	a	0.035
2	se	0.030
3	na	0.020
4	že	0.017
5	to	0.015
6	v	0.013
7	je	0.009
8	do	0.009
9	s	0.008
10	si	0.008
11	když	0.007
12	z	0.007
13	tak	0.007
14	jsem	0.006
15	švejk	0.006

Most frequent words

	Škvorecký: Eva byla nahá	
pořadí	slovo	f_rel
1	a	0.037
2	se	0.028
3	na	0.020
4	v	0.018
5	jsem	0.012
6	s	0.009
7	z	0.008
8	američan	0.008
9	to	0.008
10	řekl	0.007
11	ale	0.007
12	do	0.007
13	že	0.007
14	řekla	0.007
15	dívka	0.005

	Hrabal: Perlička na dně	
pořadí	slovo	f_rel
1	a	0.054
2	se	0.029
3	to	0.025
4	na	0.021
5	ale	0.012
6	tak	0.012
7	do	0.011
8	si	0.011
9	jsem	0.011
10	v	0.011
11	že	0.011
12	je	0.010
13	já	0.009
14	když	0.007
15	jak	0.007

	Hašek: Osudy…I.	
pořadí	slovo	f_rel
1	a	0.035
2	se	0.030
3	na	0.020
4	že	0.017
5	to	0.015
6	v	0.013
7	je	0.009
8	do	0.009
9	s	0.008
10	si	0.008
11	když	0.007
12	z	0.007
13	tak	0.007
14	jsem	0.006
15	švejk	0.006

Distances between words / texts

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|$$

n … the number of MFW

A, B … texts for the comparison

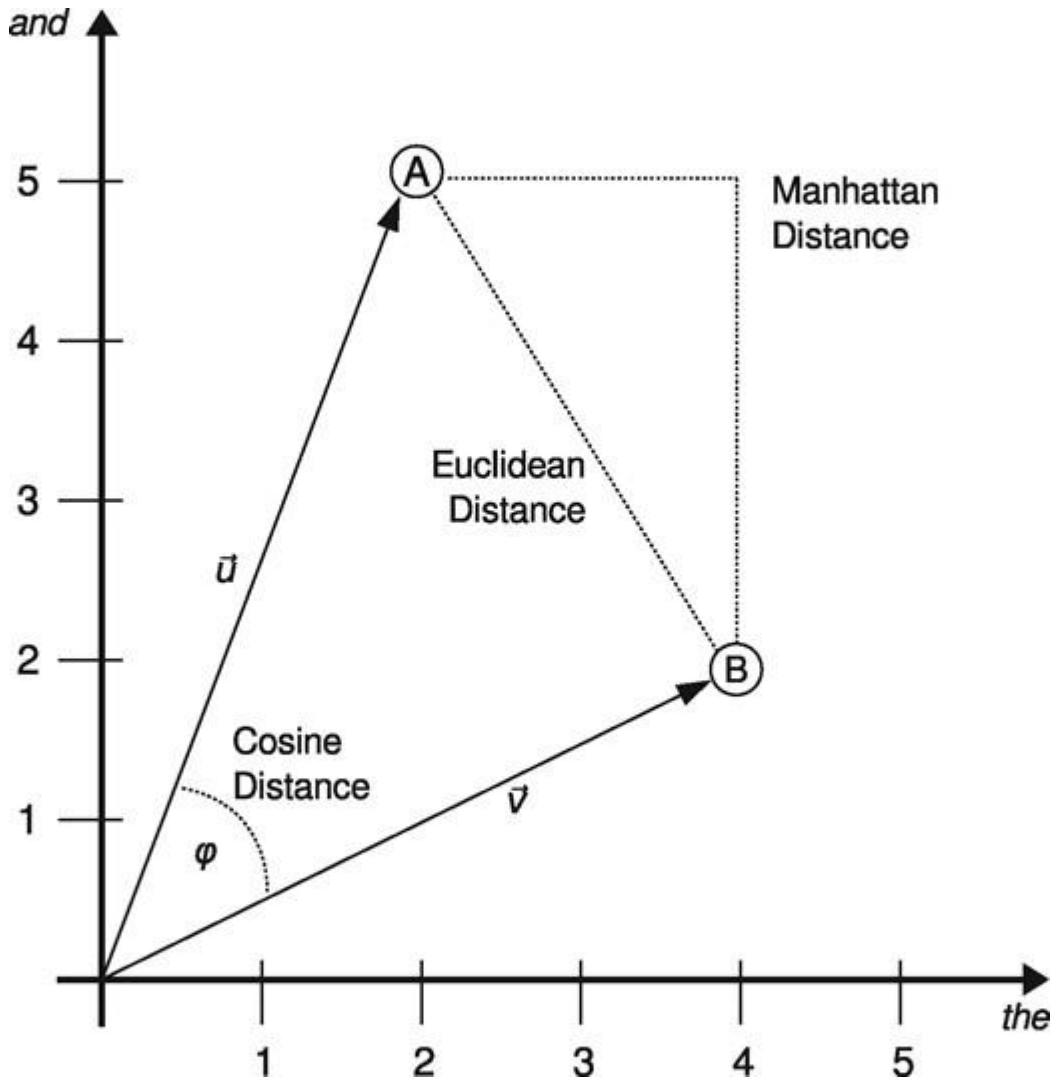
A_i … the frequency of a given word in the text A

B_i … the frequency of a given word in the text B

μ_i … the average frequency of a given word in corpus

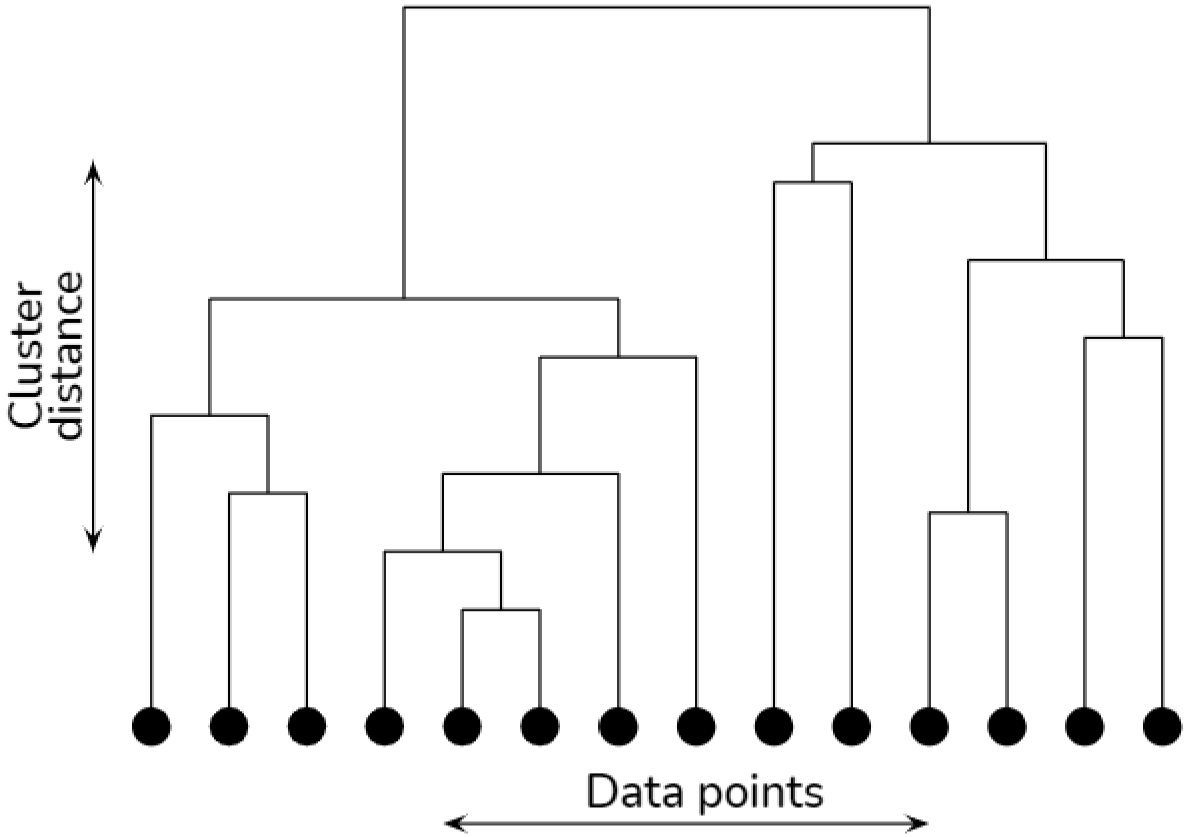
σ_i … the standard deviation of the frequency of a given word

Distances between words / texts



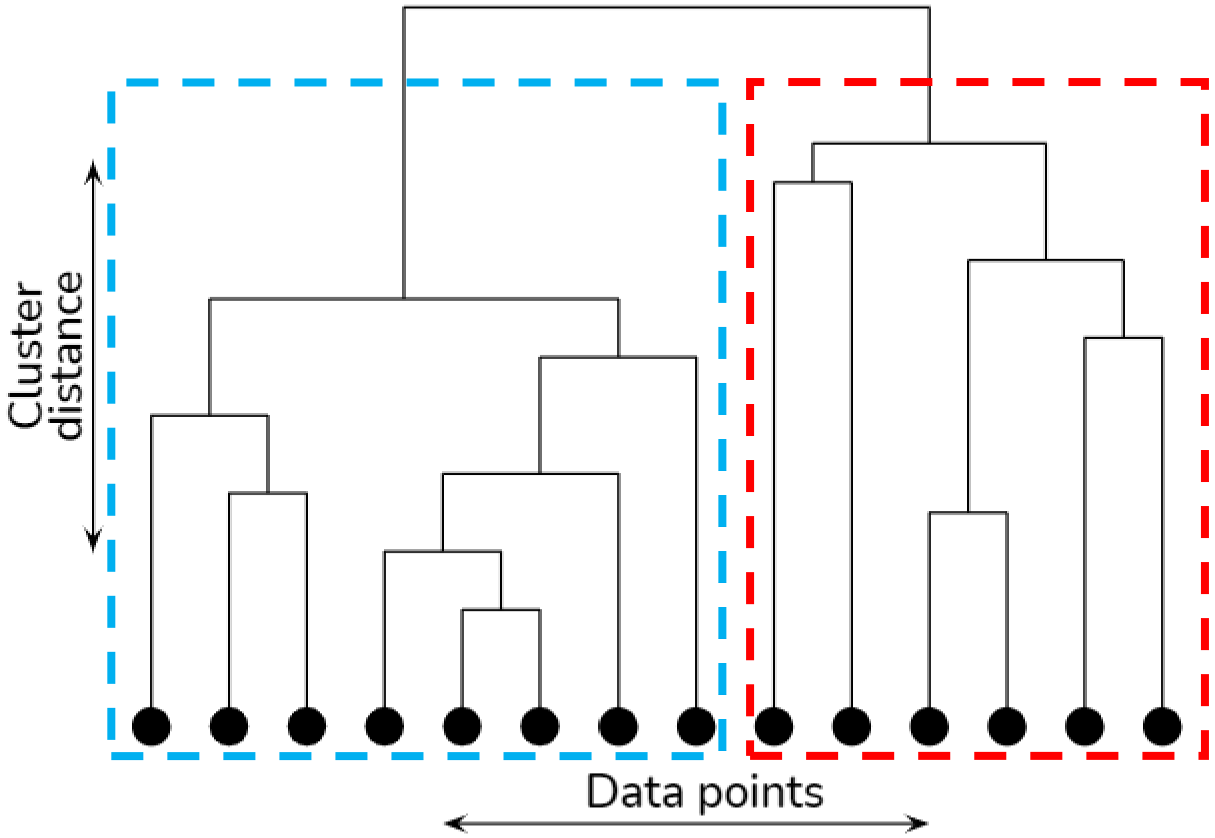
Evert et al. (2017)

Clustering



https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

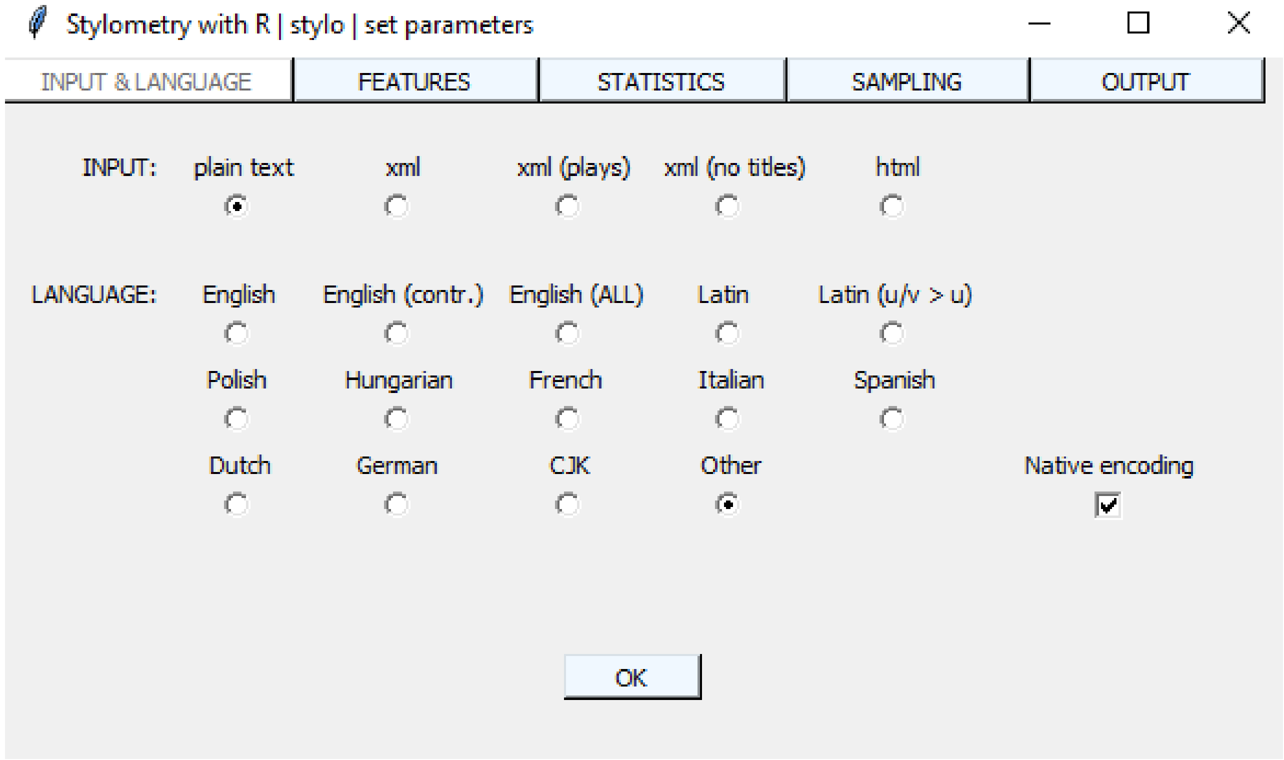
Clustering



https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

Stylo

- Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, *8*(1).

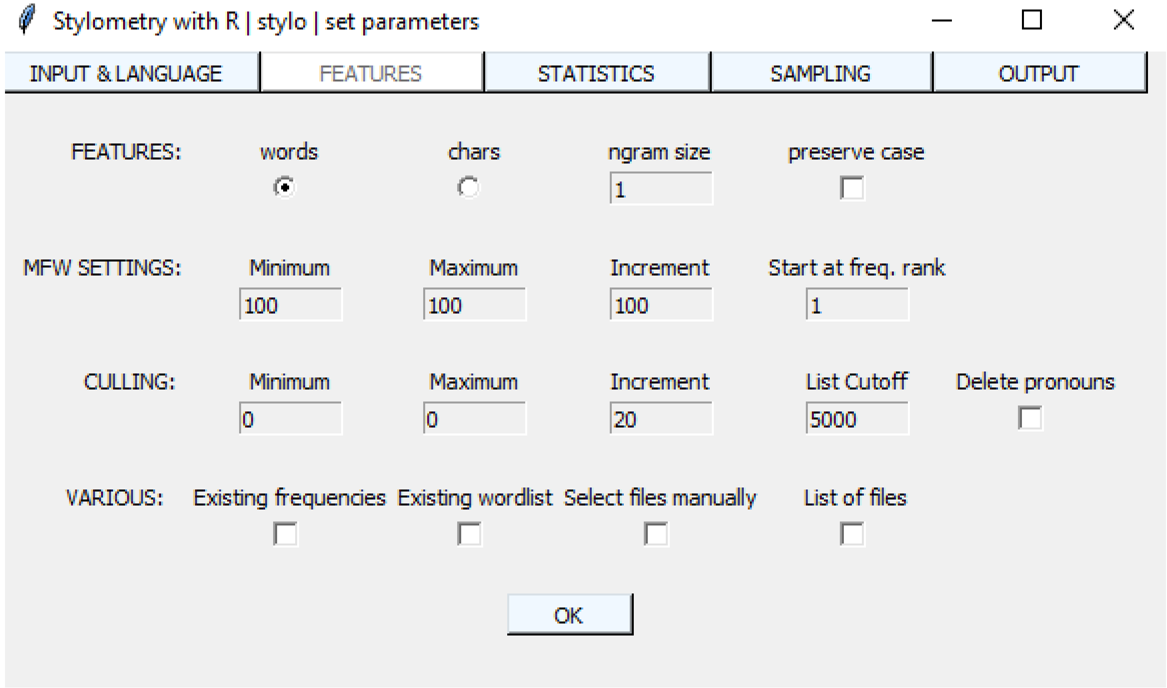


INPUT: ● plain text ○ xml ○ xml (plays) ○ xml (no titles) ○ html

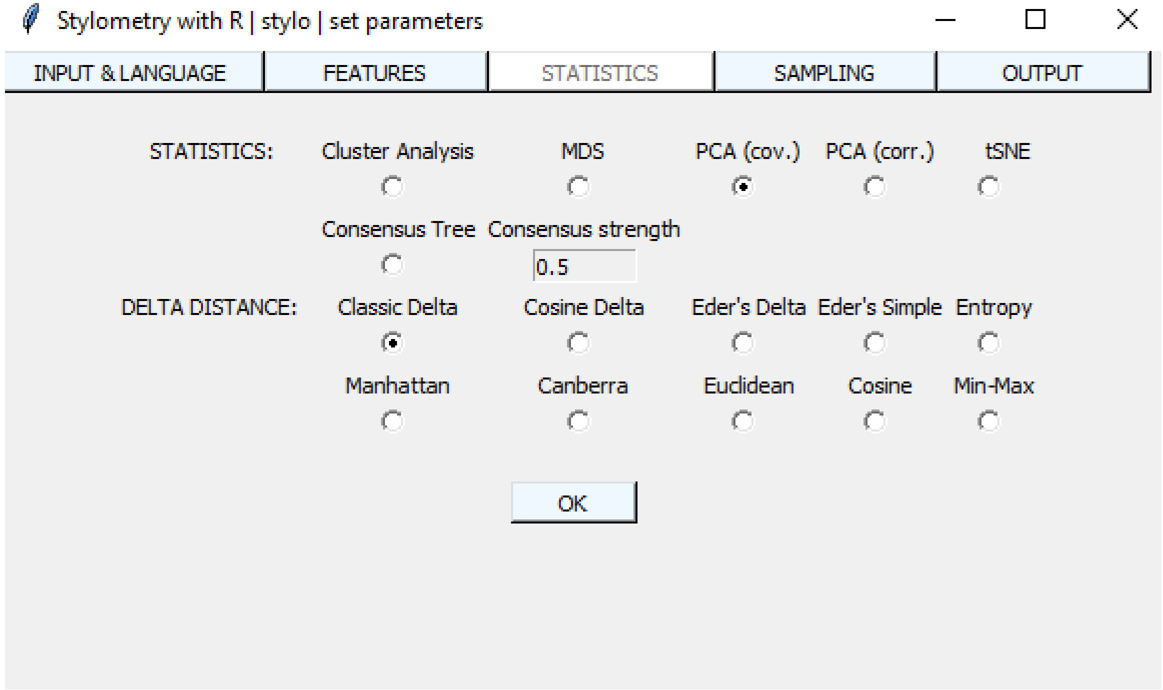
LANGUAGE:
○ English ○ English (contr.) ○ English (ALL) ○ Latin ○ Latin (u/v > u)
○ Polish ○ Hungarian ○ French ○ Italian ○ Spanish
○ Dutch ○ German ○ CJK ● Other

☑ Native encoding

Stylo



FEATURES:	words	chars	ngram size	preserve case	
	●	○	1	□	
MFW SETTINGS:	Minimum	Maximum	Increment	Start at freq. rank	
	100	100	100	1	
CULLING:	Minimum	Maximum	Increment	List Cutoff	Delete pronouns
	0	0	20	5000	□
VARIOUS:	Existing frequencies	Existing wordlist	Select files manually	List of files	
	□	□	□	□	



STATISTICS:	Cluster Analysis	MDS	PCA (cov.)	PCA (corr.)	tSNE
	○	○	●	○	○
	Consensus Tree	Consensus strength			
	○	0.5			
DELTA DISTANCE:	Classic Delta	Cosine Delta	Eder's Delta	Eder's Simple	Entropy
	●	○	○	○	○
	Manhattan	Canberra	Euclidean	Cosine	Min-Max
	○	○	○	○	○

Bible svatováclavská & comments

- Kosek, P., Čech, R. (2018). Stylové aspekty Bible svatováclavské – stylometrická analýza. In Zand, G., Newerkla, S.M. (eds.). Jezuitská kultura v českých zemích / Jesuitische Kultur in den böhmischen Ländern. Host, 195-209.

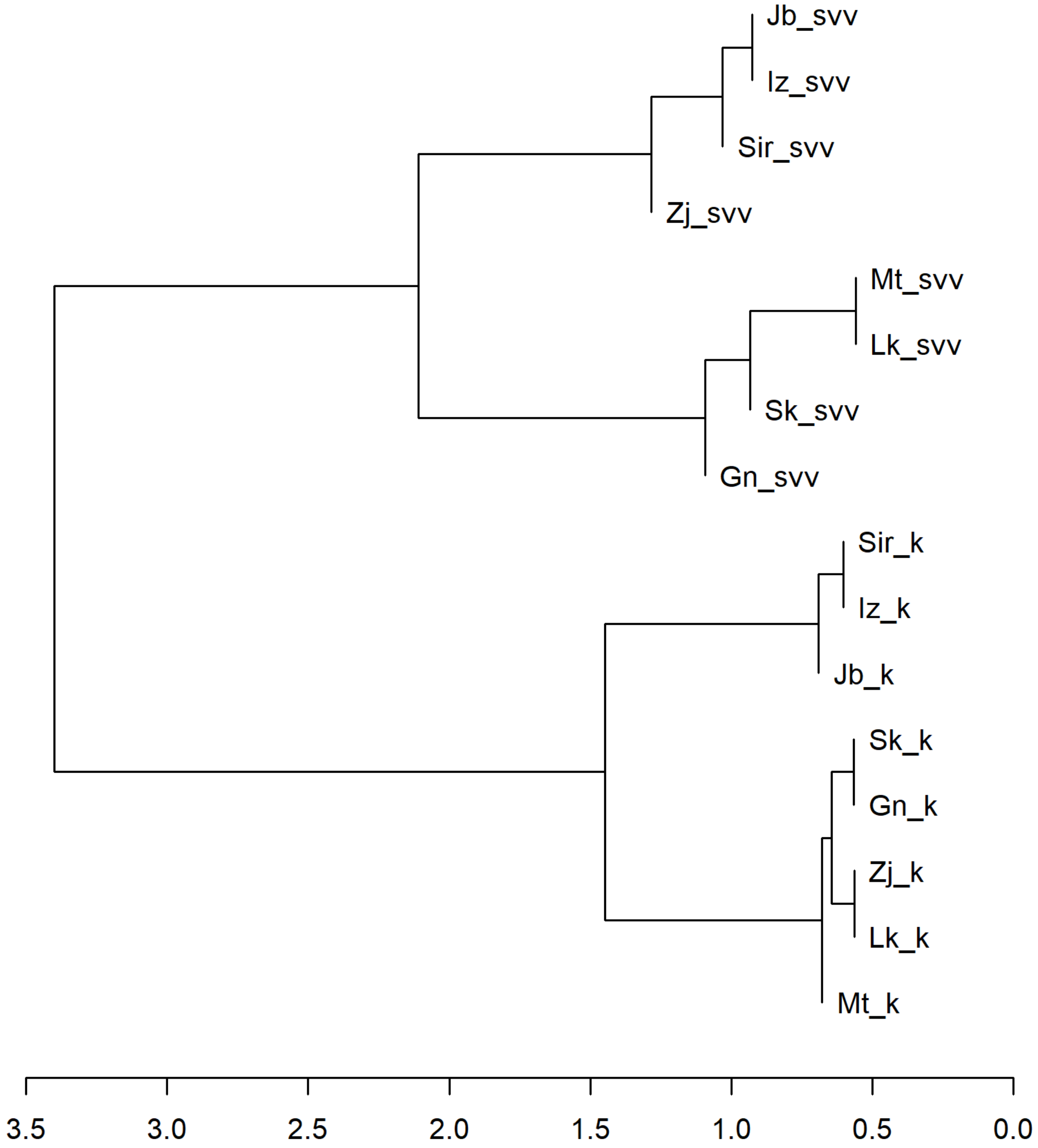
Bible svatováclavská & comments

- comments
    - Šteyer: New Testament + Genesis
    - Barner: Old Testament

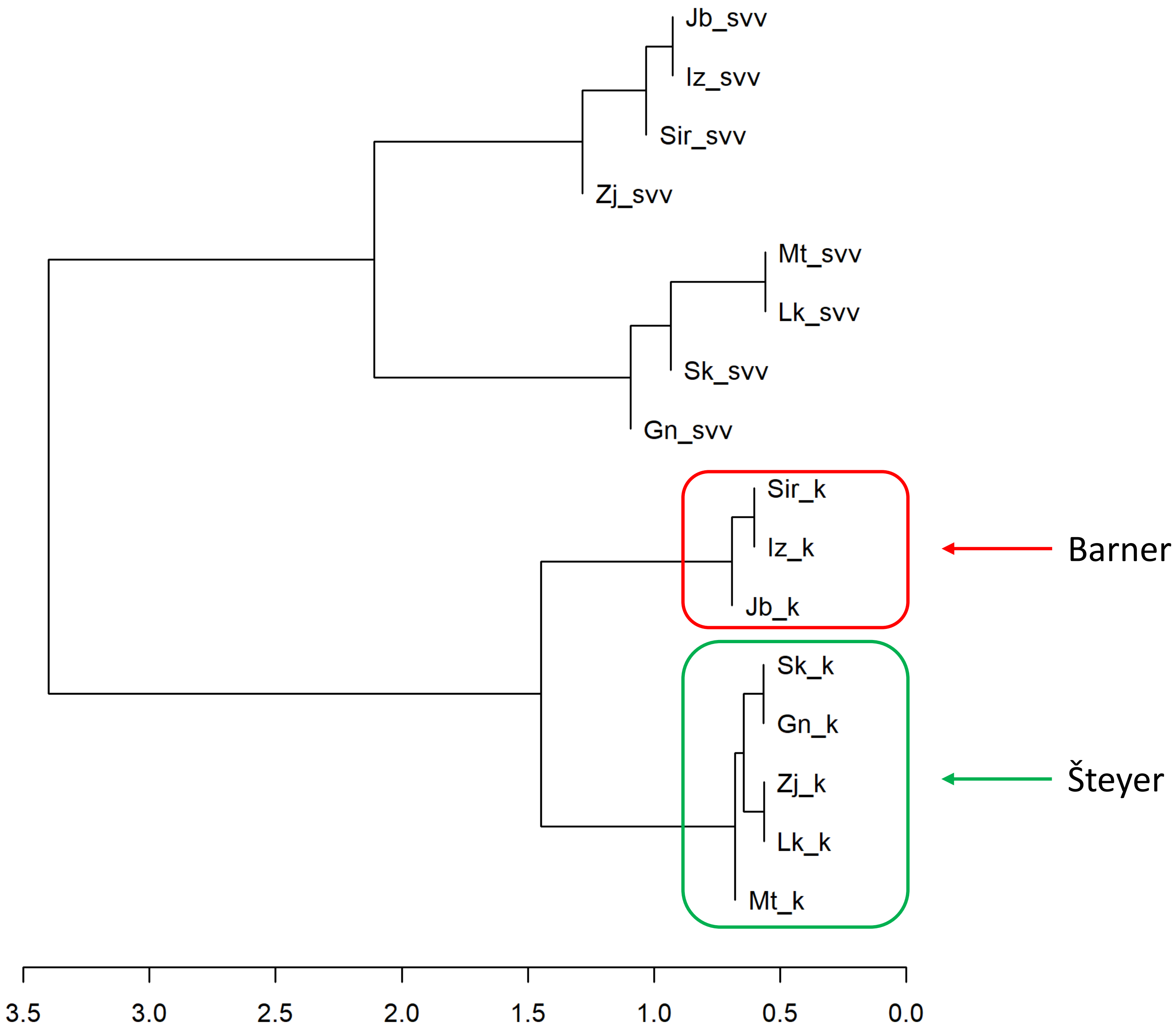
Bible svatováclavská & comments

- New Testament
  - Mt, Lk, Sk, Zj
- Old Testament
  - Gn, Jb, Iz, Sir

- 100 MFW
- culling = 0



Jb_svv
Iz_svv
Sir_svv
Zj_svv
Mt_svv
Lk_svv
Sk_svv
Gn_svv
Sir_k
Iz_k
Jb_k
Sk_k
Gn_k
Zj_k
Lk_k
Mt_k
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0



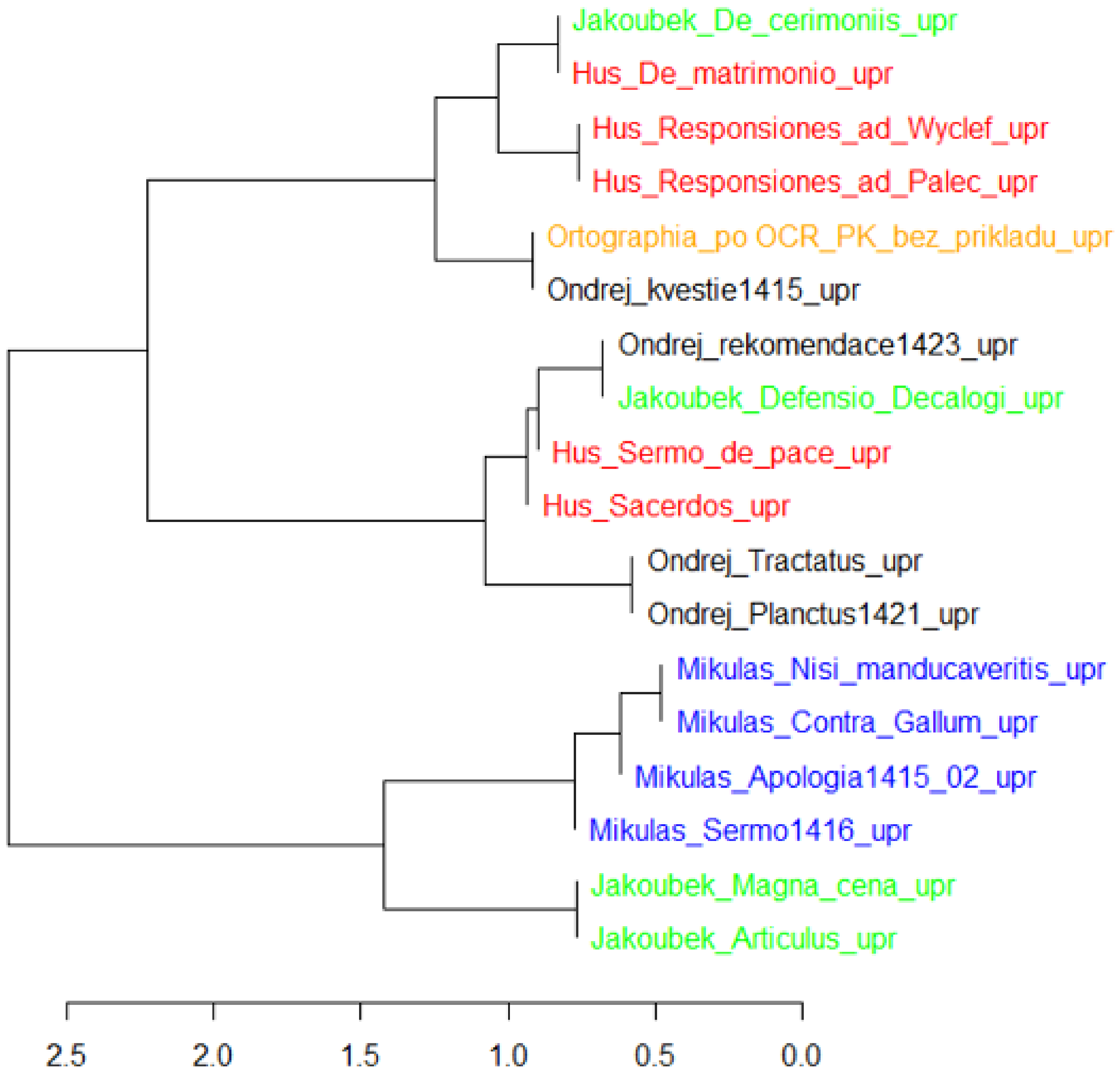
Barner

Šteyer

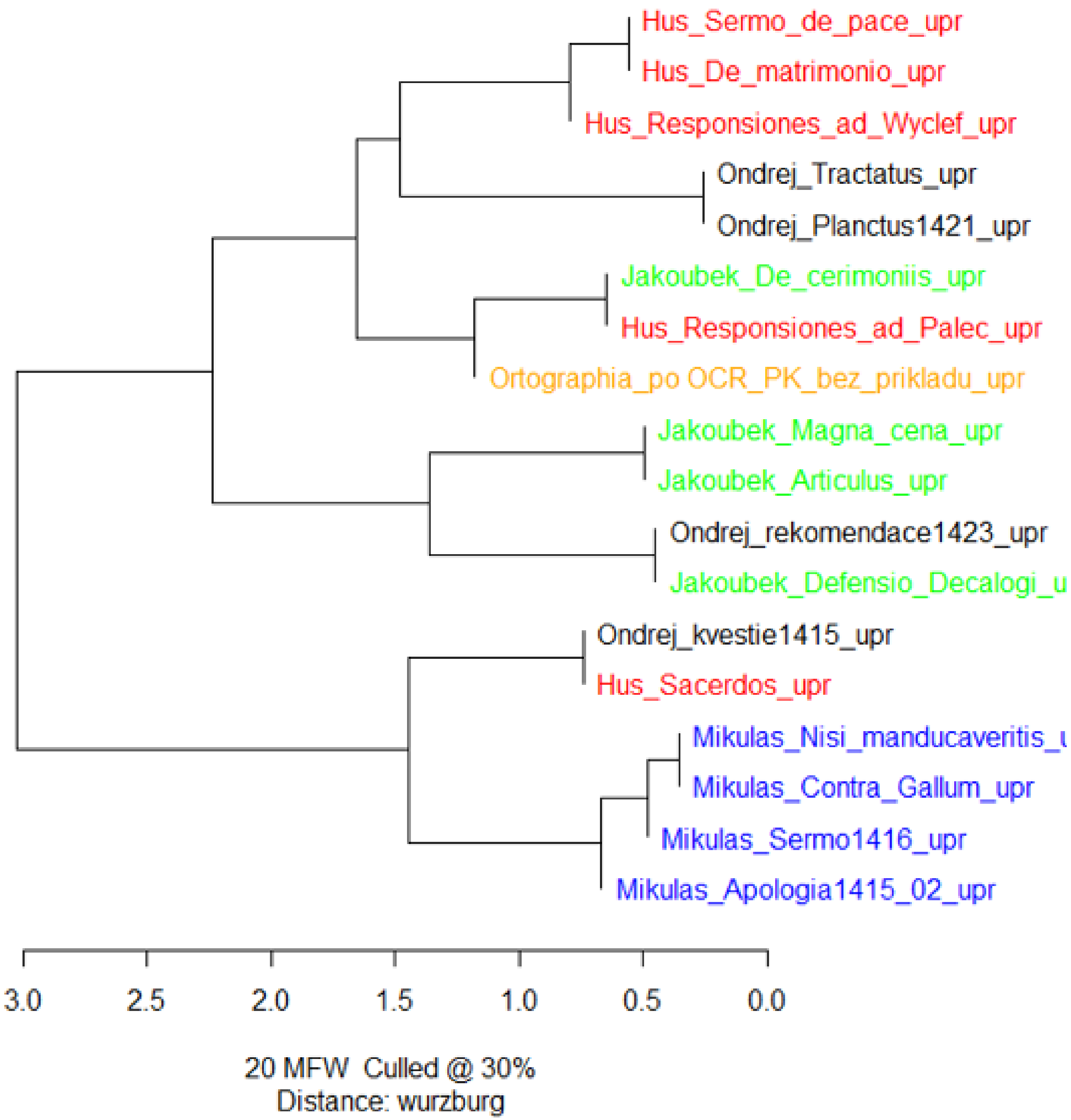
Orthographia Bohemica

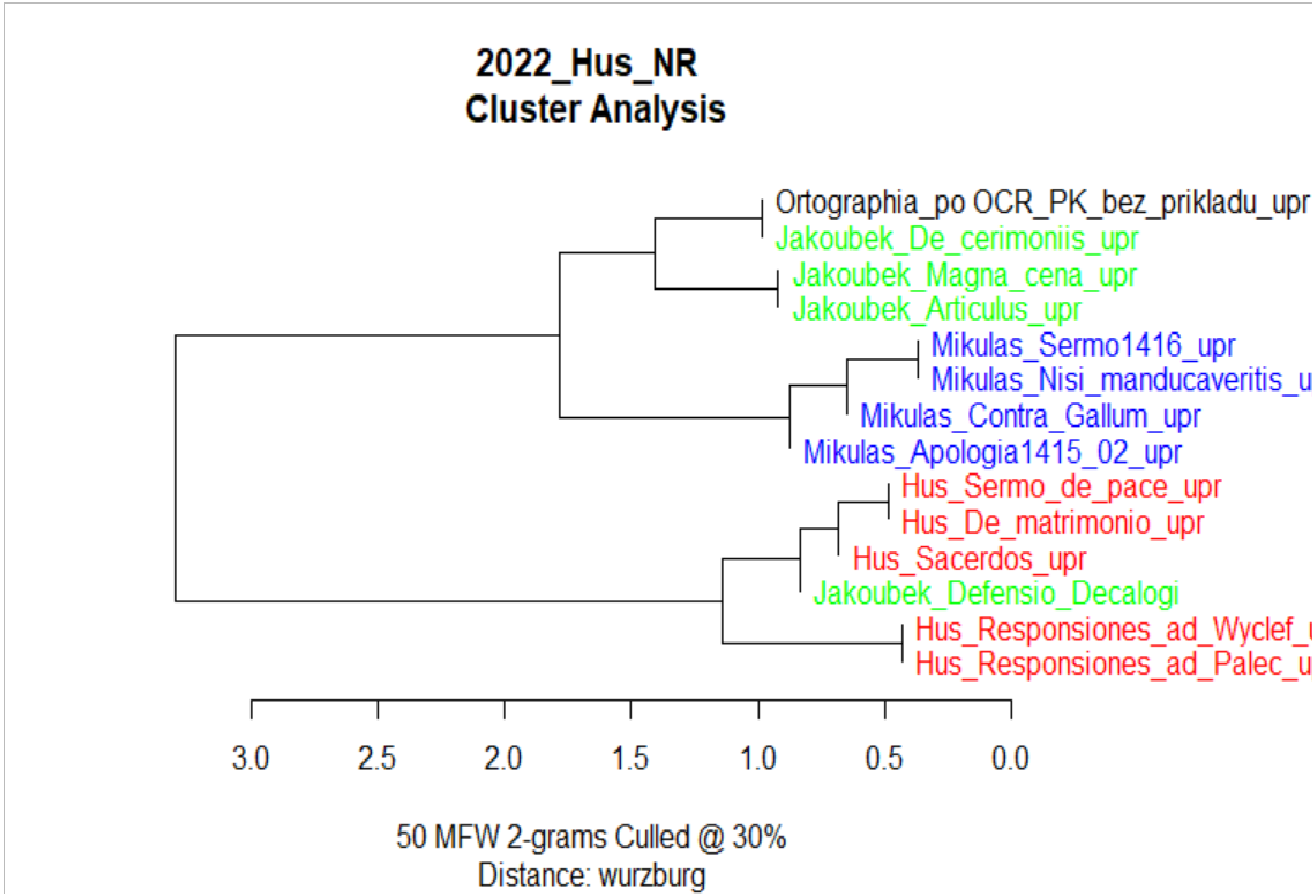
- preprocesing of texts
  - deleting numbers
  - non-word characters

  - Czech examples (Orthographia Bohemica)



100 MFW Culled @ 30%
Distance: wurzburg

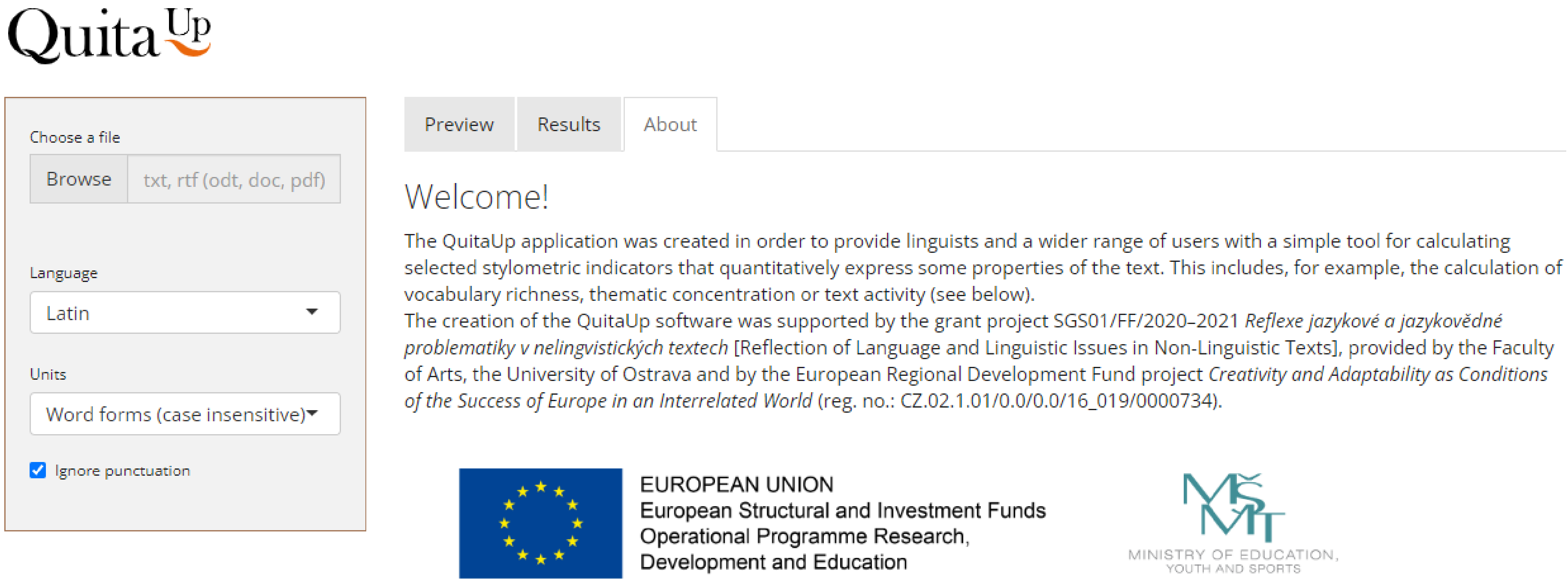

Hus_Sermo_de_pace_upr
Hus_De_matrimonio_upr
Hus_Responsiones_ad_Wyclef_upr
Ondrej_Tractatus_upr
Ondrej_Planctus1421_upr
Jakoubek_De_cerimoniis_upr
Hus_Responsiones_ad_Palec_upr
Ortographia_po OCR_PK_bez_prikladu_upr
Jakoubek_Magna_cena_upr
Jakoubek_Articulus_upr
Ondrej_rekomendace1423_upr
Jakoubek_Defensio_Decalogi_u
Ondrej_kvestie1415_upr
Hus_Sacerdos_upr
Mikulas_Nisi_manducaveritis_u
Mikulas_Contra_Gallum_upr
Mikulas_Sermo1416_upr
Mikulas_Apologia1415_02_upr
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 MFW Culled @ 30%
Distance: wurzburg


2022_Hus_NR
Cluster Analysis
Ortographia_po OCR_PK_bez_prikladu_upr
Jakoubek_De_cerimoniis_upr
Jakoubek_Magna_cena_upr
Jakoubek_Articulus_upr
Mikulas_Sermo1416_upr
Mikulas_Nisi_manducaveritis_u
Mikulas_Contra_Gallum_upr
Mikulas_Apologia1415_02_upr
Hus_Sermo_de_pace_upr
Hus_De_matrimonio_upr
Hus_Sacerdos_upr
Jakoubek_Defensio_Decalogi
Hus_Responsiones_ad_Wyclef_
Hus_Responsiones_ad_Palec_u
3.0
2.5
2.0
1.5
1.0
0.5
0.0
50 MFW 2-grams Culled @ 30%
Distance: wurzburg

Combination of measurements

- ratio of hapaxes to tokens
- verb distances
- text activity
  - proportion of verbs to the sum of verbs and adjectives
- average token length
- thematic concentration
- moving average TTR

Combination of measurements



Choose a file

txt, rtf (odt, doc, pdf)

Language: Latin

Units: Word forms (case insensitive)

☑ Ignore punctuation

Welcome!

The QuitaUp application was created in order to provide linguists and a wider range of users with a simple tool for calculating selected stylometric indicators that quantitatively express some properties of the text. This includes, for example, the calculation of vocabulary richness, thematic concentration or text activity (see below).

The creation of the QuitaUp software was supported by the grant project SGS01/FF/2020–2021 *Reflexe jazykové a jazykovědné problematiky v nelingvistických textech* [Reflection of Language and Linguistic Issues in Non-Linguistic Texts], provided by the Faculty of Arts, the University of Ostrava and by the European Regional Development Fund project *Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World* (reg. no.: CZ.02.1.01/0.0/0.0/16_019/0000734).

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

https://korpus.cz/quitaup/

Ortographia Bohemica

- paraphrases, non-attributed quotations …

- next attempt → analysis of 'authorial' parts

Ortographia Bohemica

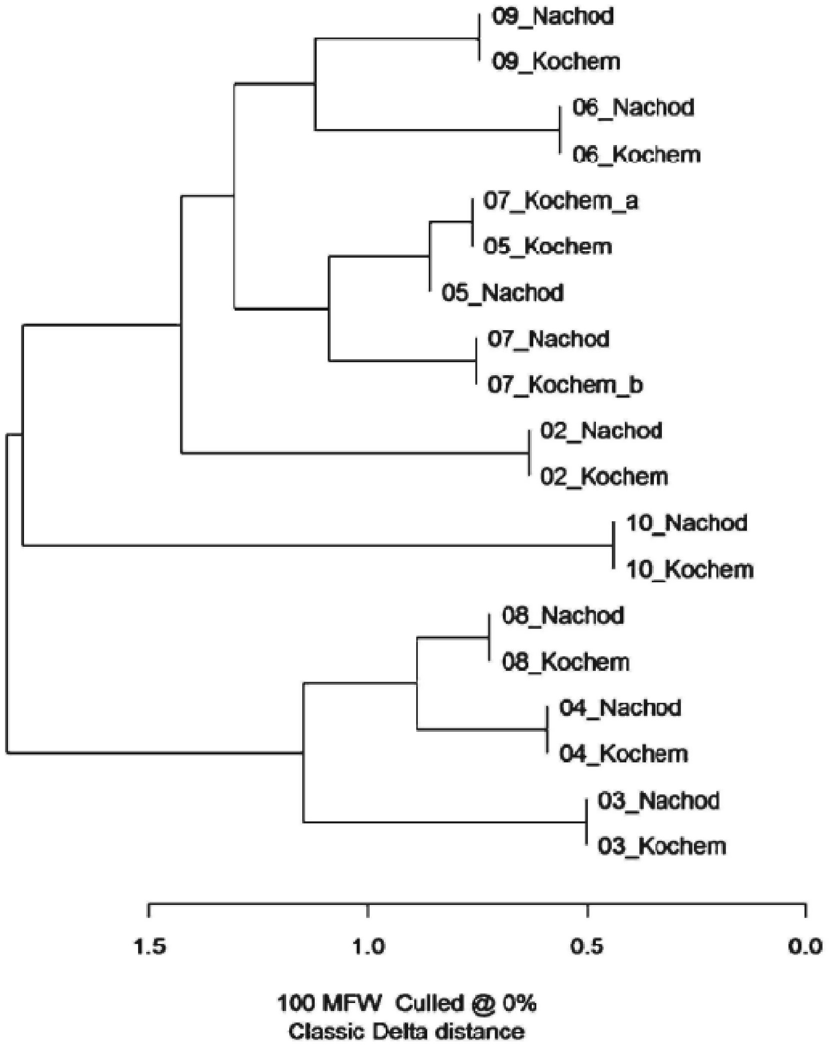
- paraphrases, non-attributed quotations …
- next attempt → analysis of 'authorial' parts
- Latin

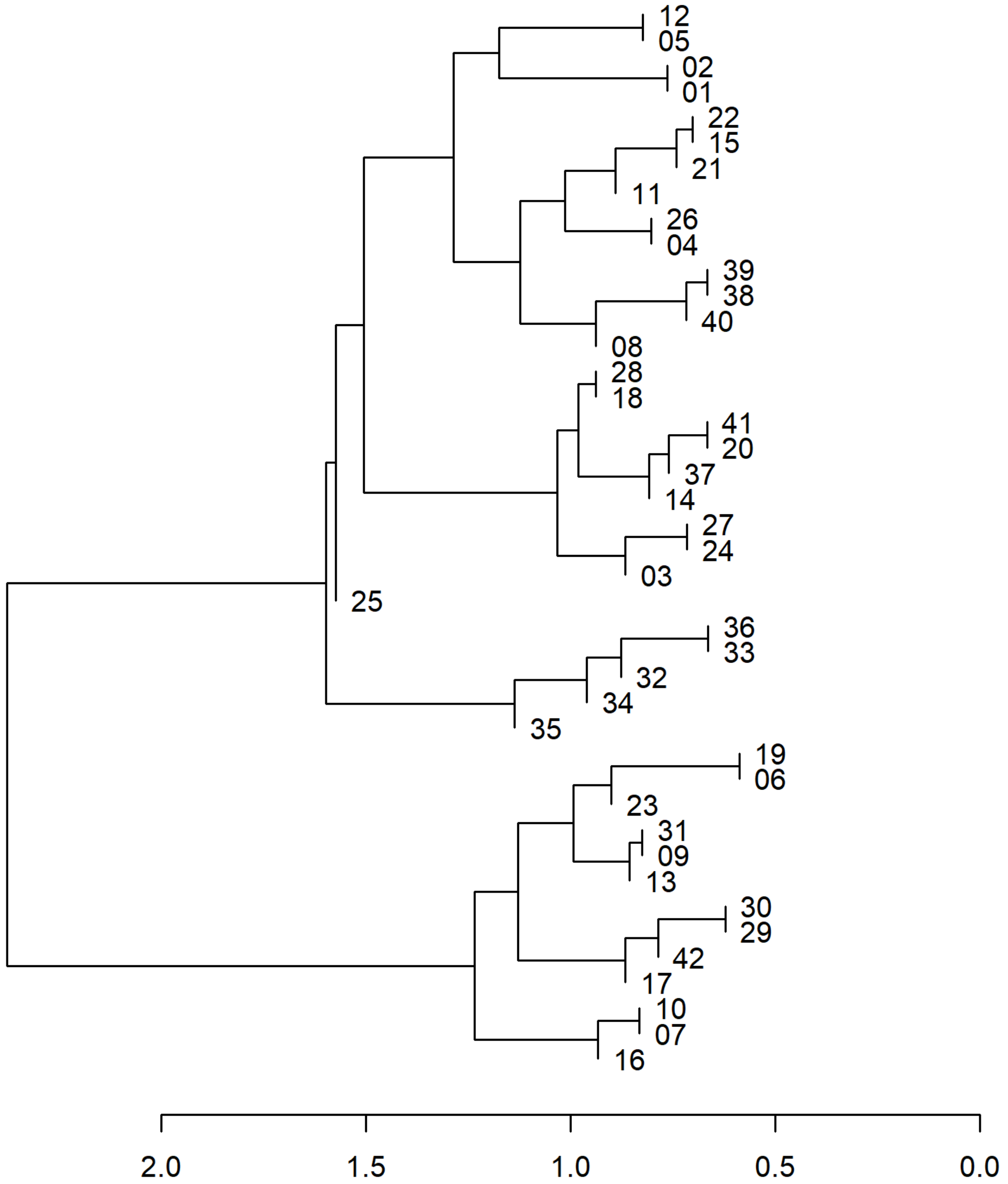
Thank you
for your attention!

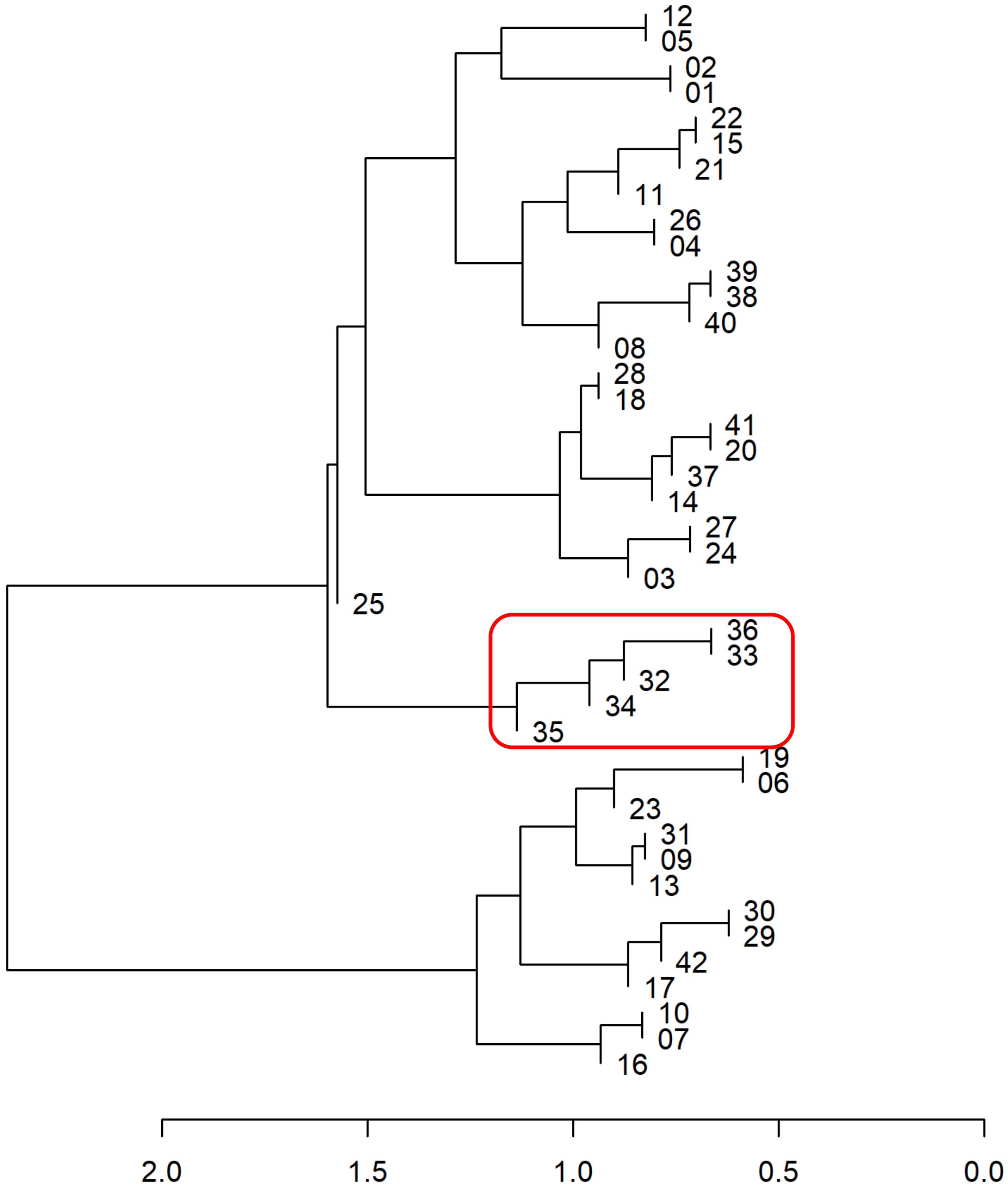
Baroque Prayers

- Kubát, M., Netolická, Š., Čech, R., Mačutek, J. (2021). Martin of Cochem's Golden Key of Heaven and its Czech Relatives: Quantitative Analysis of Baroque Prayers. Bohemistyka, 21, 283-294.

Baroque Prayers




12
05
02
01
22
15
21
11
26
04
39
38
40
08
28
18
41
20
37
14
27
24
03
25
36
33
32
34
35
19
06
23
31
09
13
30
29
42
17
10
07
16
2.0
1.5
1.0
0.5
0.0


12
05
02
01
22
15
21
11
26
04
39
38
40
08
28
18
41
20
37
14
27
24
03
25
36
33
32
34
35
19
06
23
31
09
13
30
29
42
17
10
07
16
2.0
1.5
1.0
0.5
0.0