Most Frequent Words as a Tool for Authorship Recognition

Radek Čech

Ján Mačutek

Acknowledgment

- supported by the project GA ČR GA17-02545S, 2017–2019

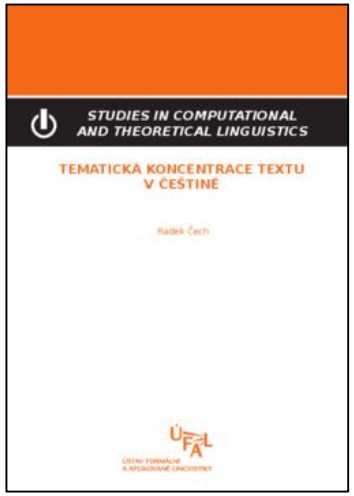
Pavel Kosek (MU), Petra Mutlová (MU), Miroslav Kubát (OU)

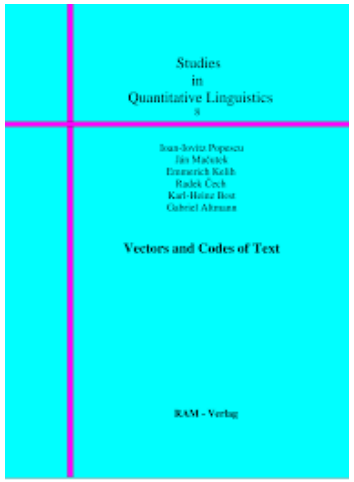
Background

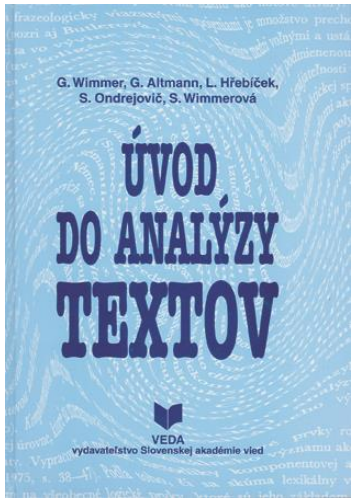
- theory of text
  - quantitative properties
  - regularities
  - mechanisms
  - laws

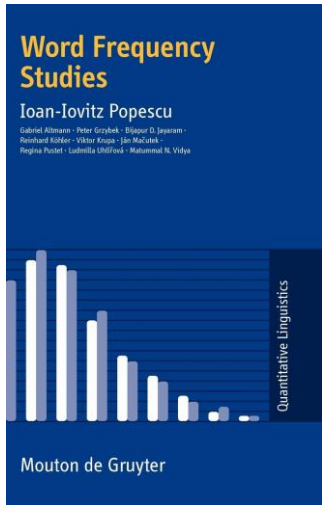
Background

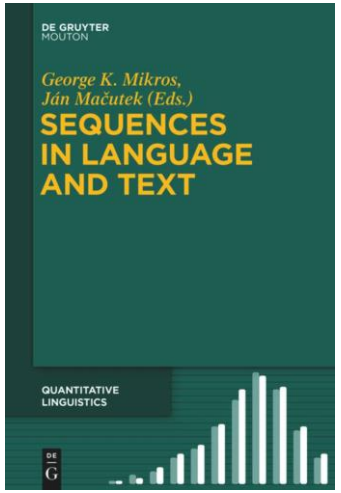
- theory of text
  - quantitative properties
  - regularities
  - mechanisms
  - laws







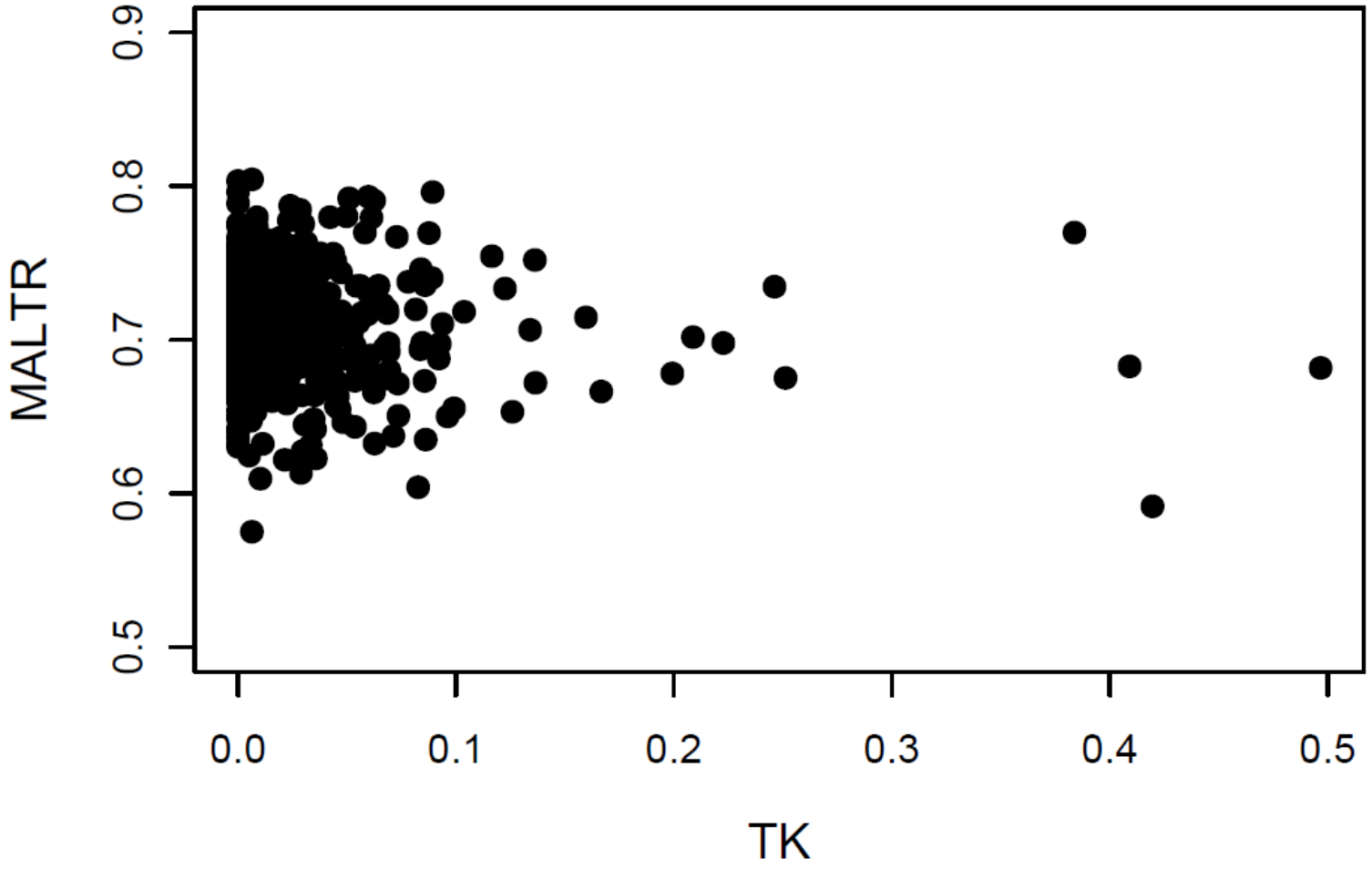




Theory of text (and language)

- relationships among particular properties (stochastic)
  - lexical diversity – text length
  - lexical diversity – thematic concentration
  - entropy – verb distances
  - distances between hapax legomena – lexical diversity
  - distribution of clause/sentence lengths – distances between hapax legomena
  - autosemantic text structure – thematic concentration

Theory of text (and language)

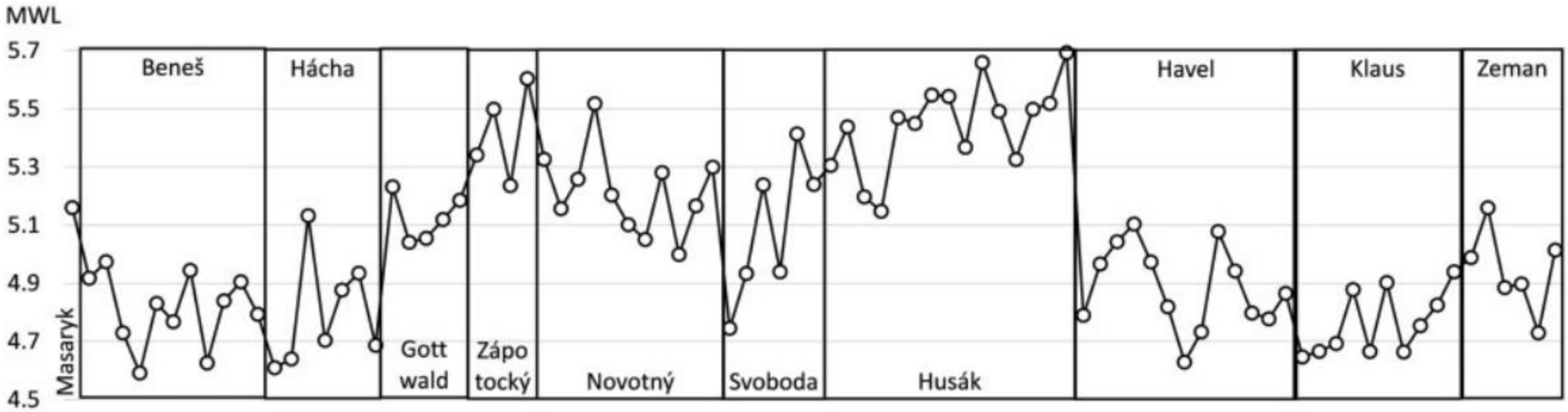


Theory of text (and language)

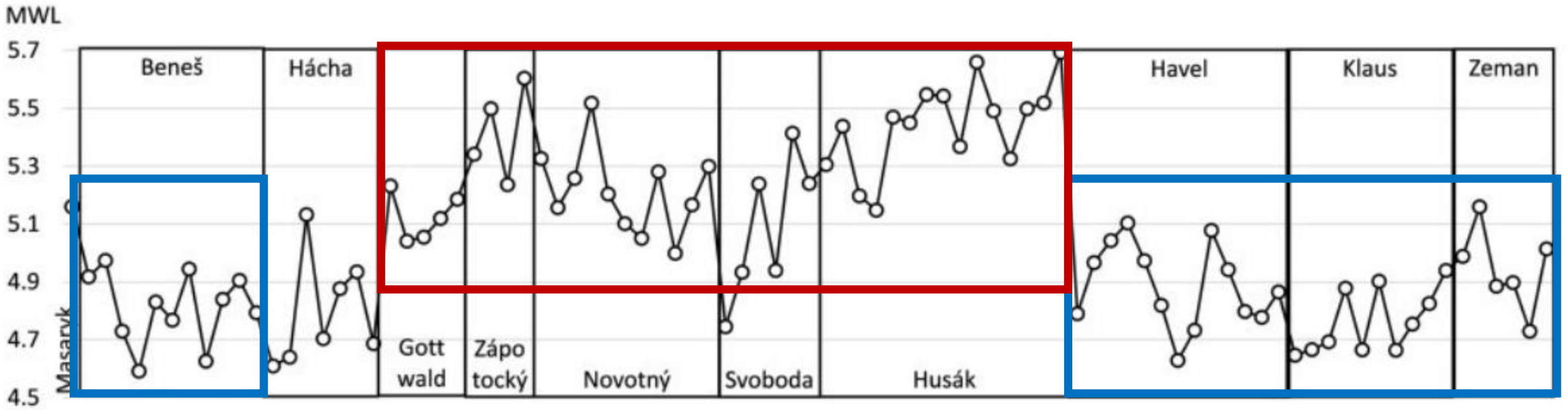
- relationships between particular properties and pragmatics
- stylometry

    - text type
    - genre
    - period
    - gender
    - political background / ideology
    - authorship

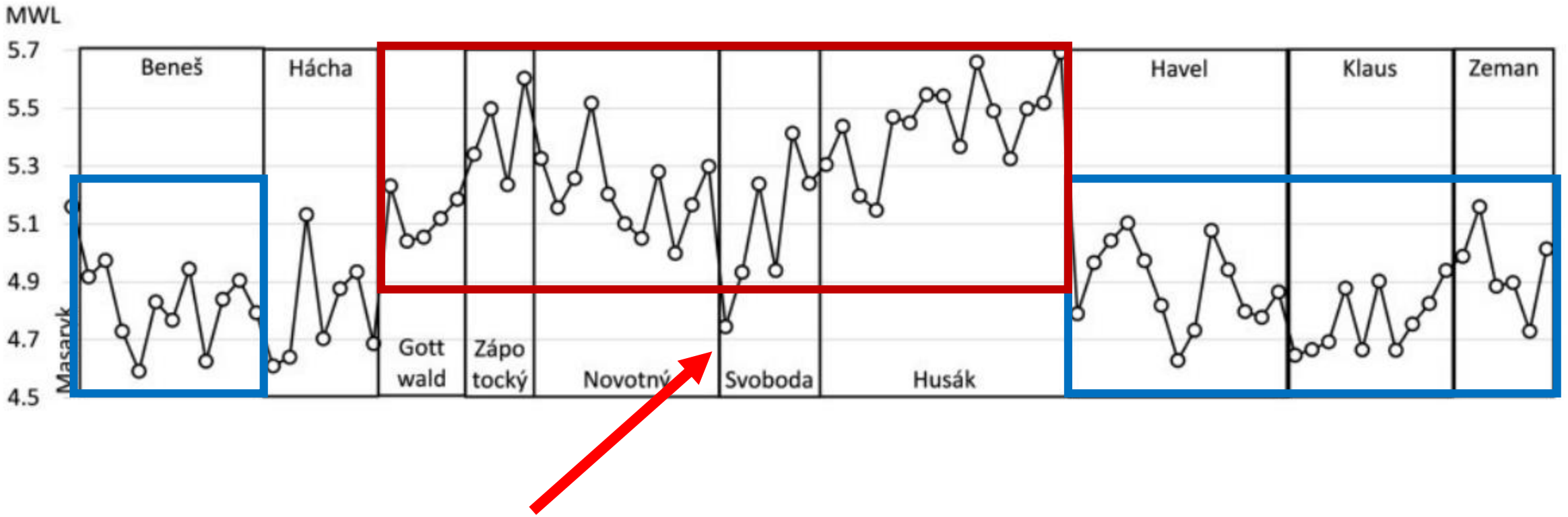
Mean word length



Mean word length



Mean word length

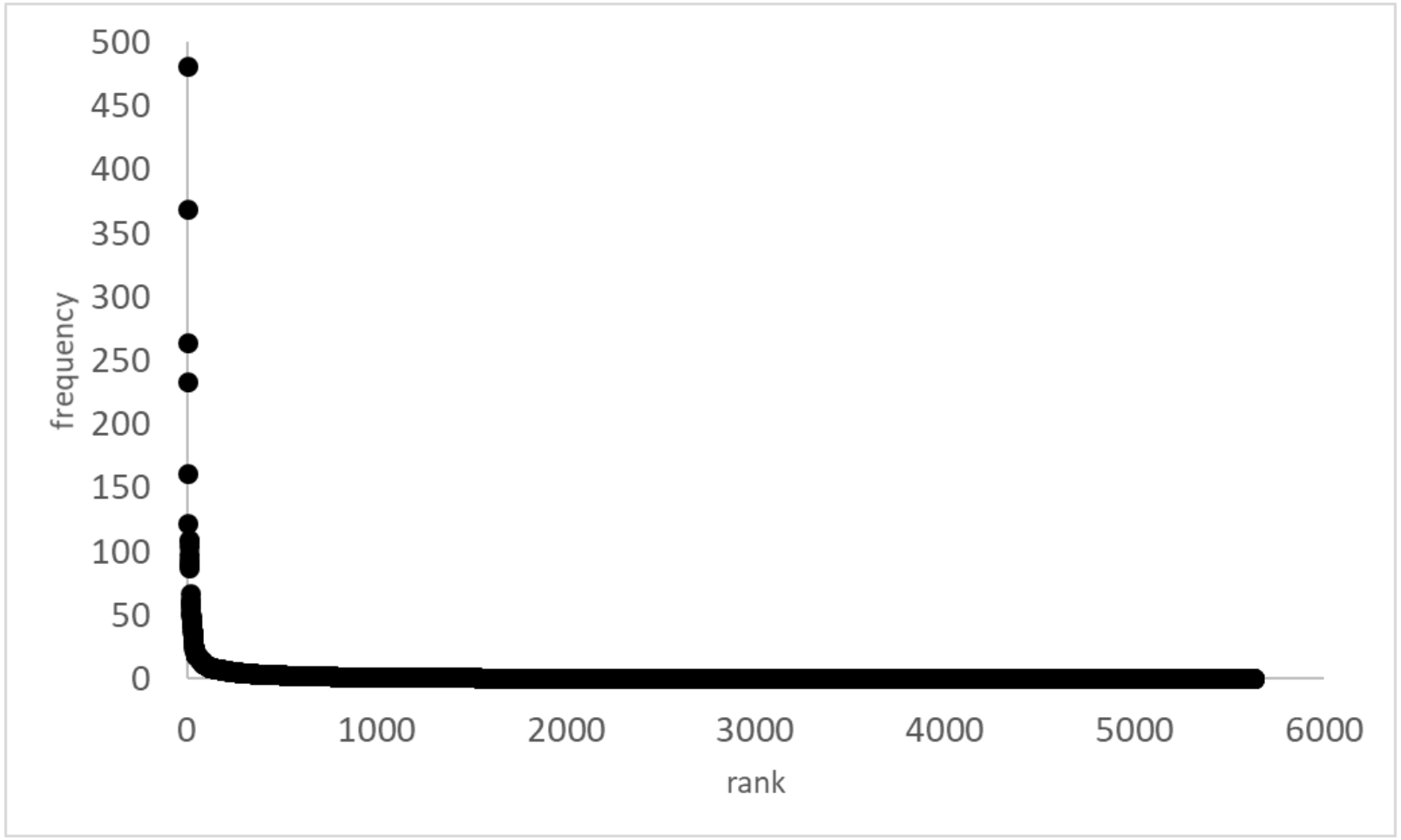


From theory to the application

Word frequencies and authorship

- 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

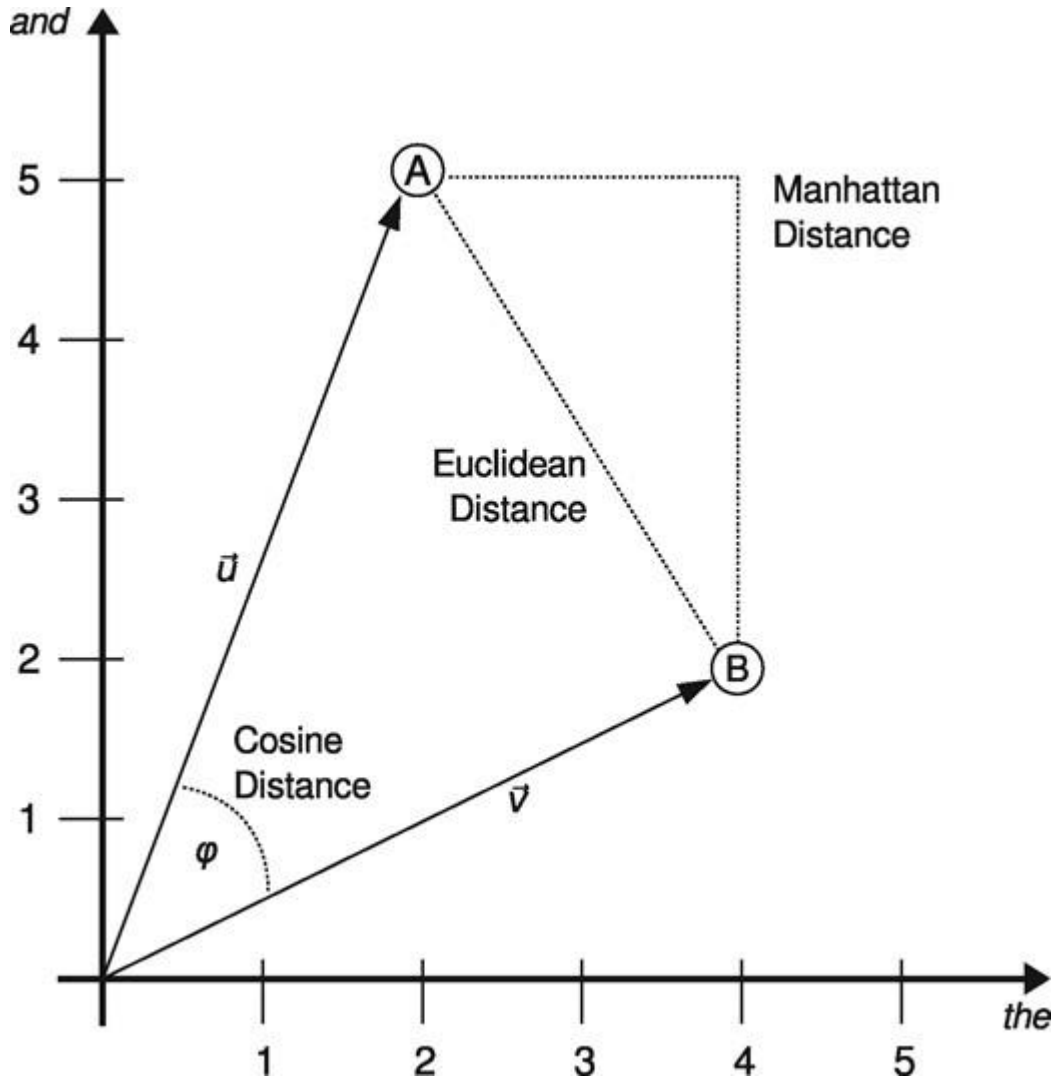
A_i … the frequency of a given word in the text A

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

μ_i … the average frequency of a given word in corpus

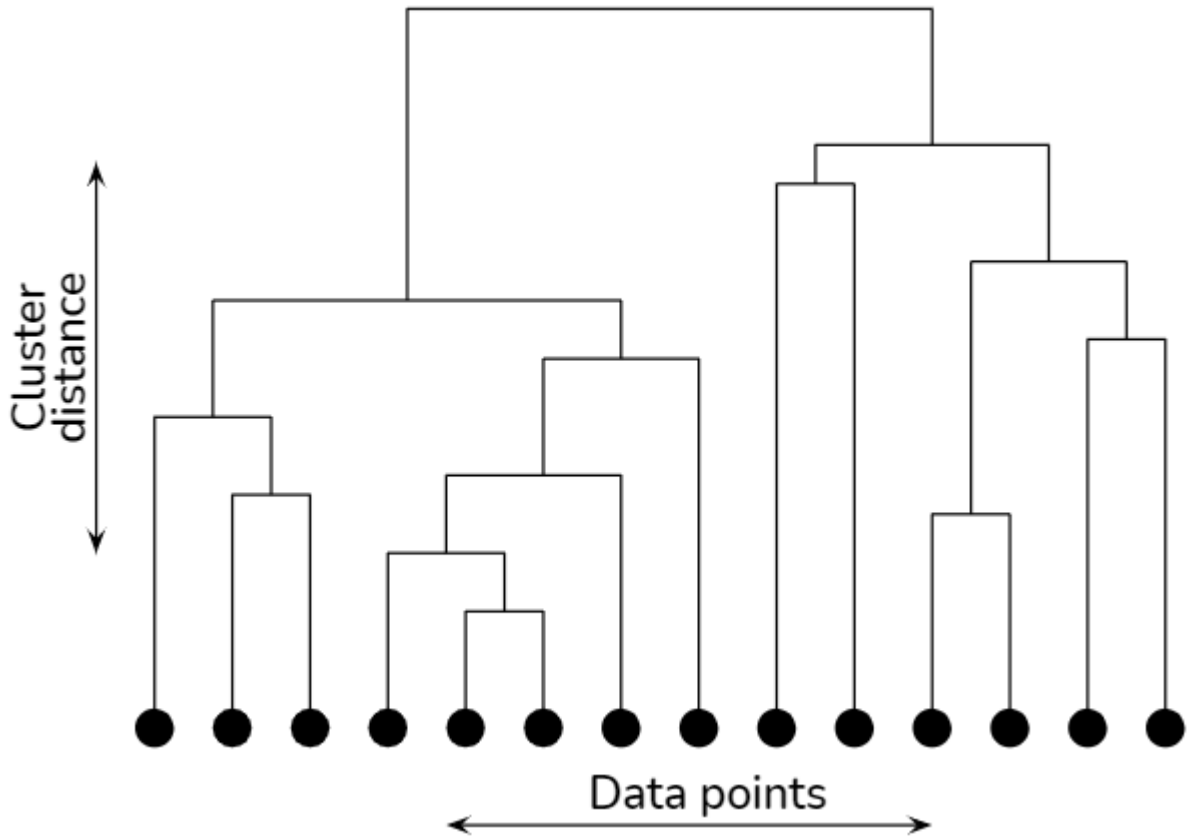
σ_i … the standard deviation of the frequency of a given word

Distances between words / texts



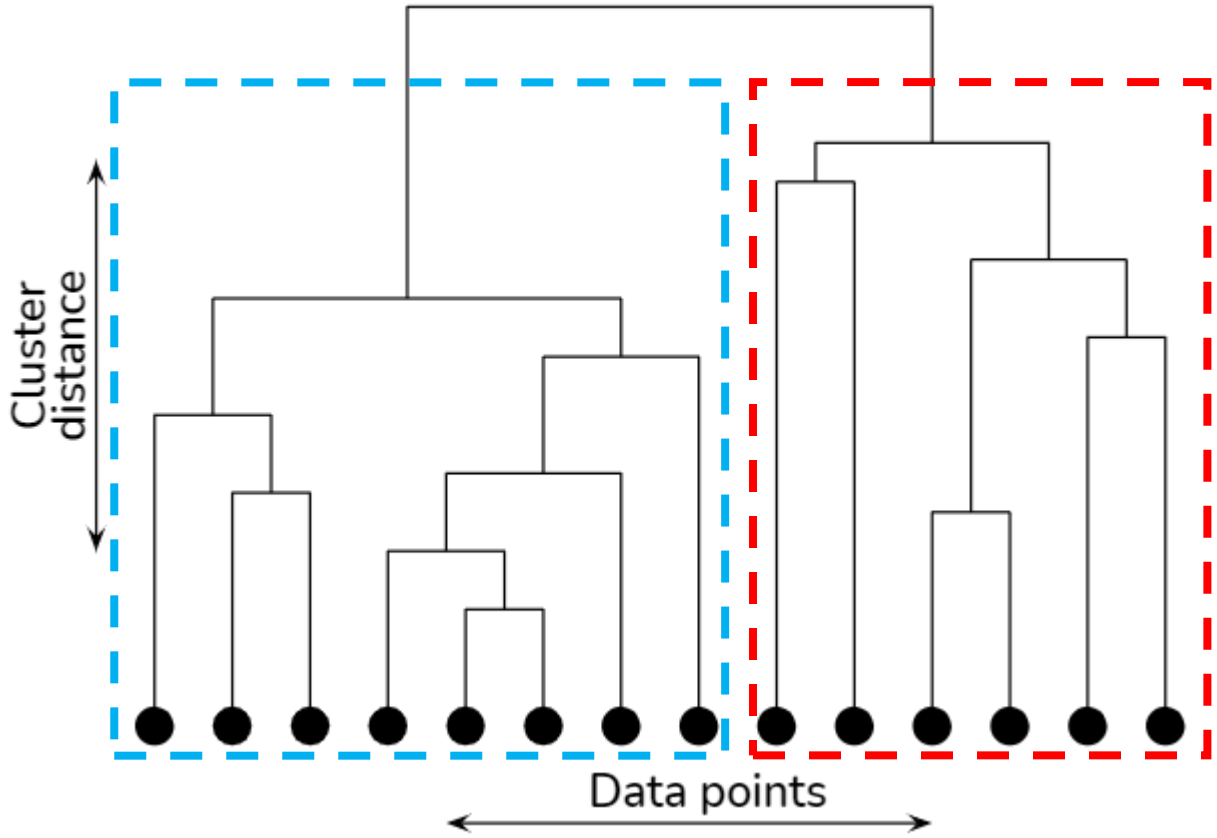
Evert et al. (2017)

Clustering



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

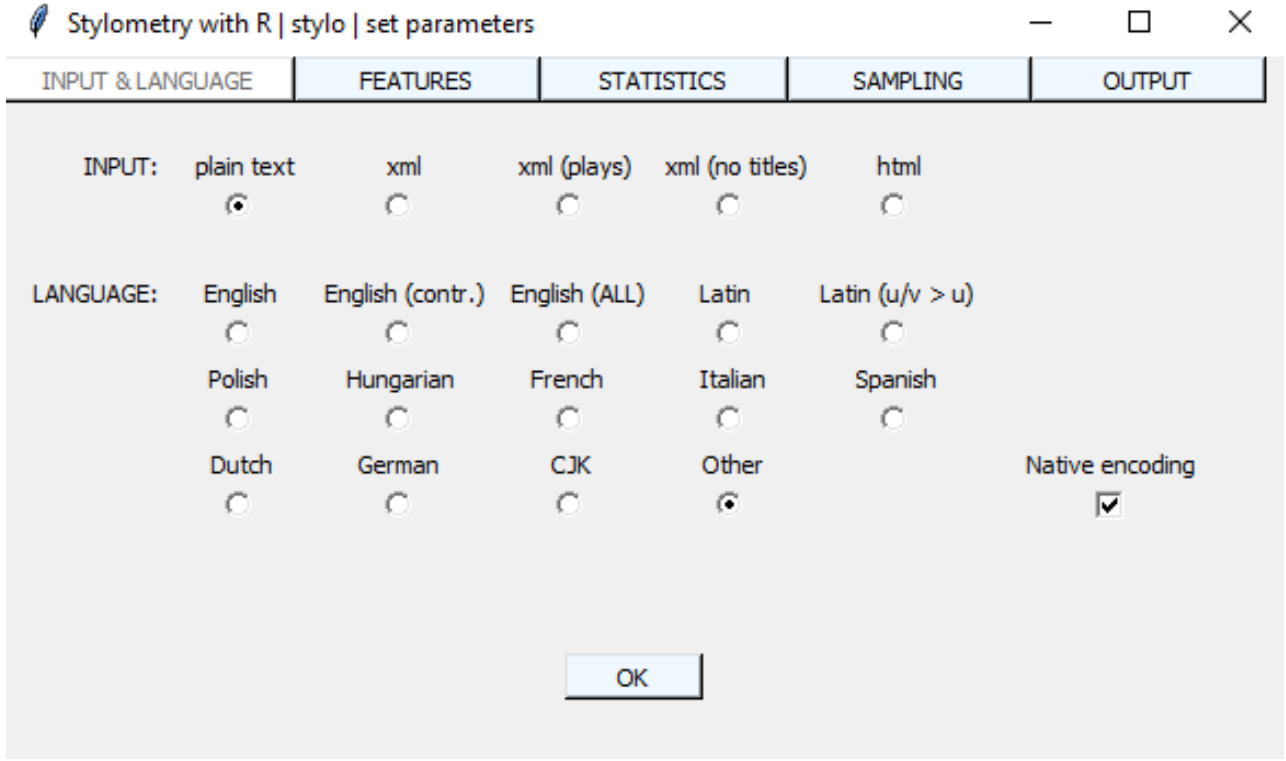
Clustering



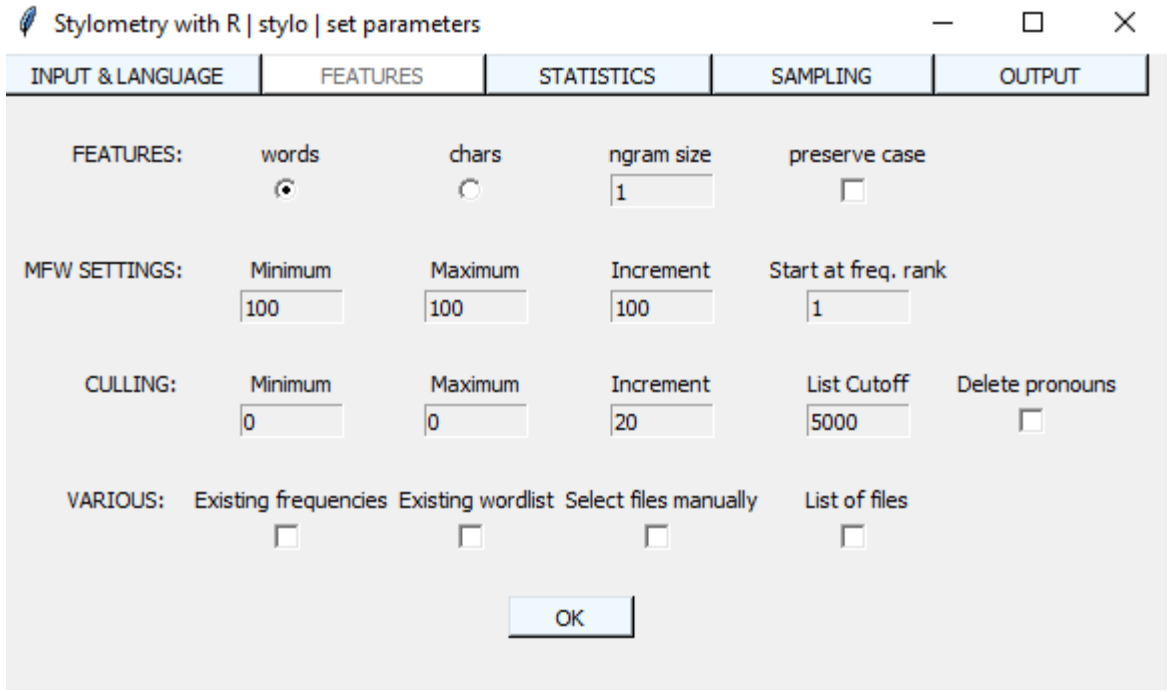
https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

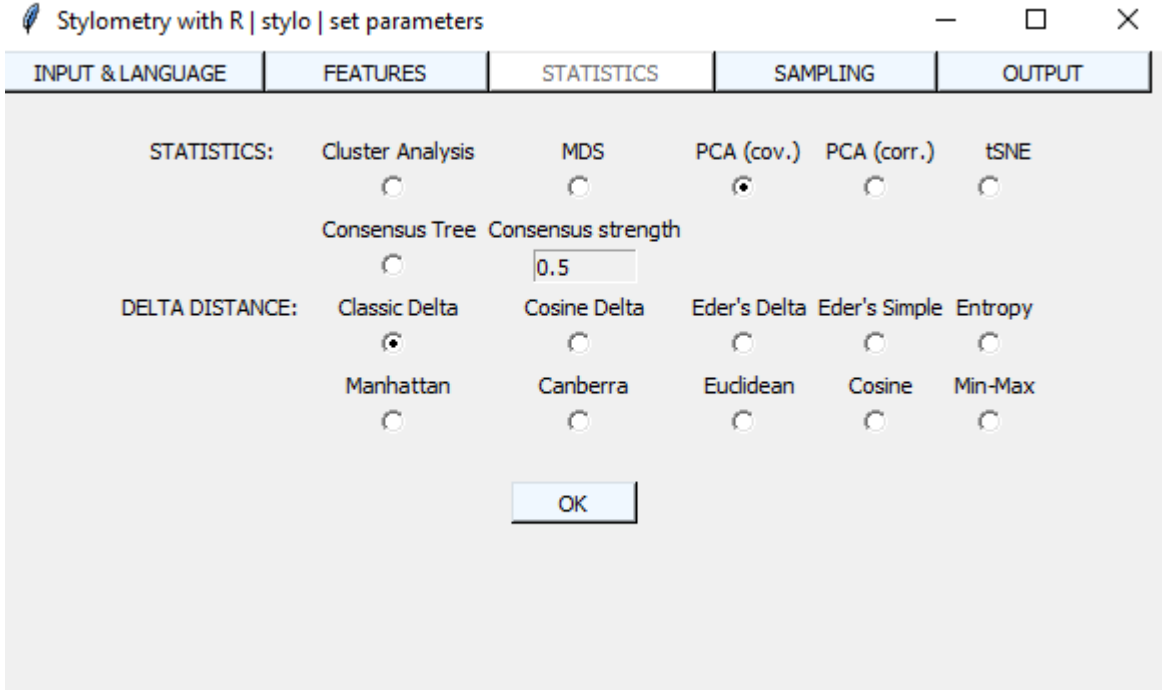
Stylo

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



Stylo


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
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
OK

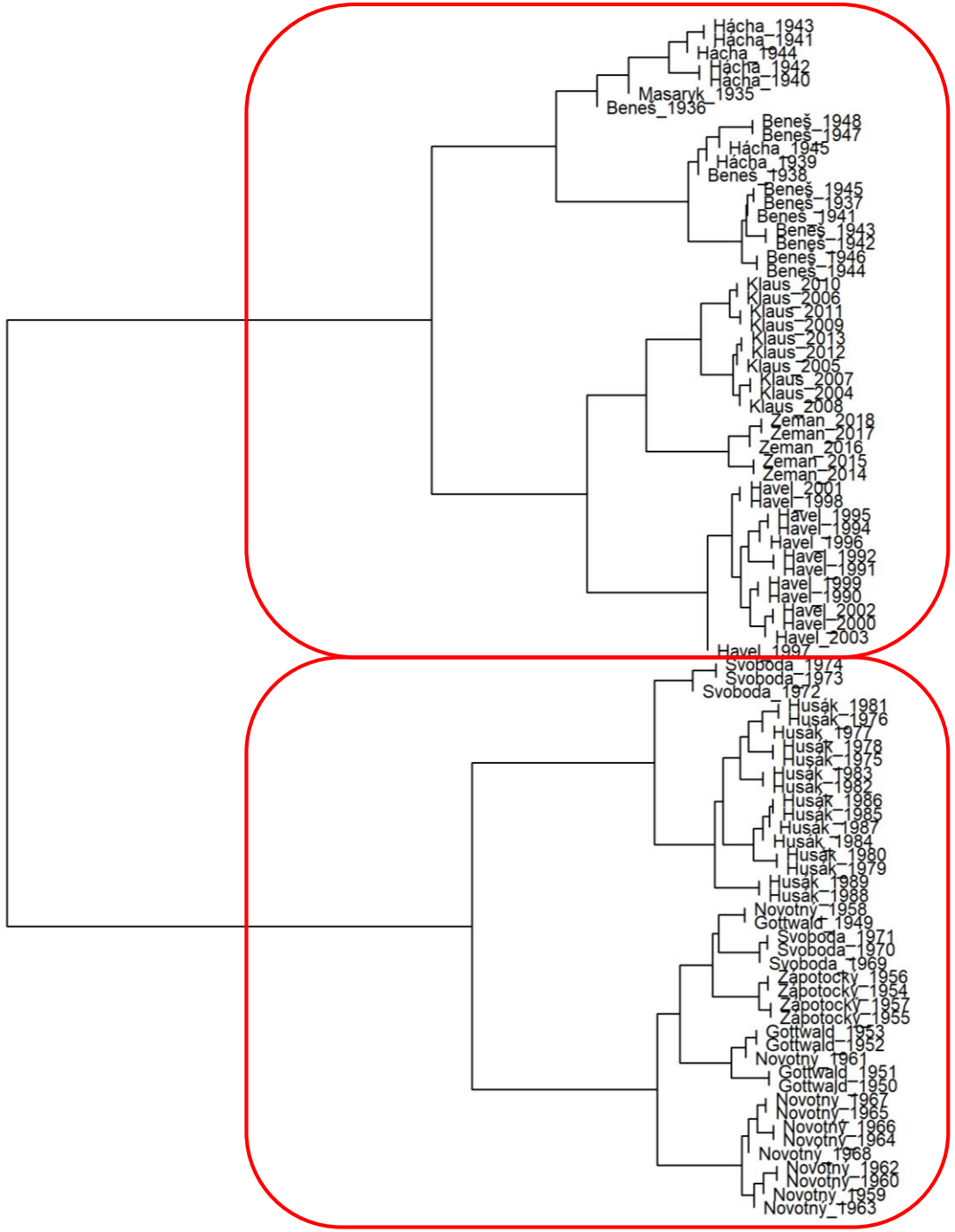

Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
STATISTICS:
Cluster Analysis
MDS
PCA (cov.)
PCA (corr.)
tSNE
Consensus Tree
Consensus strength
0.5
DELTA DISTANCE:
Classic Delta
Cosine Delta
Eder's Delta
Eder's Simple
Entropy
Manhattan
Canberra
Euclidean
Cosine
Min-Max
OK

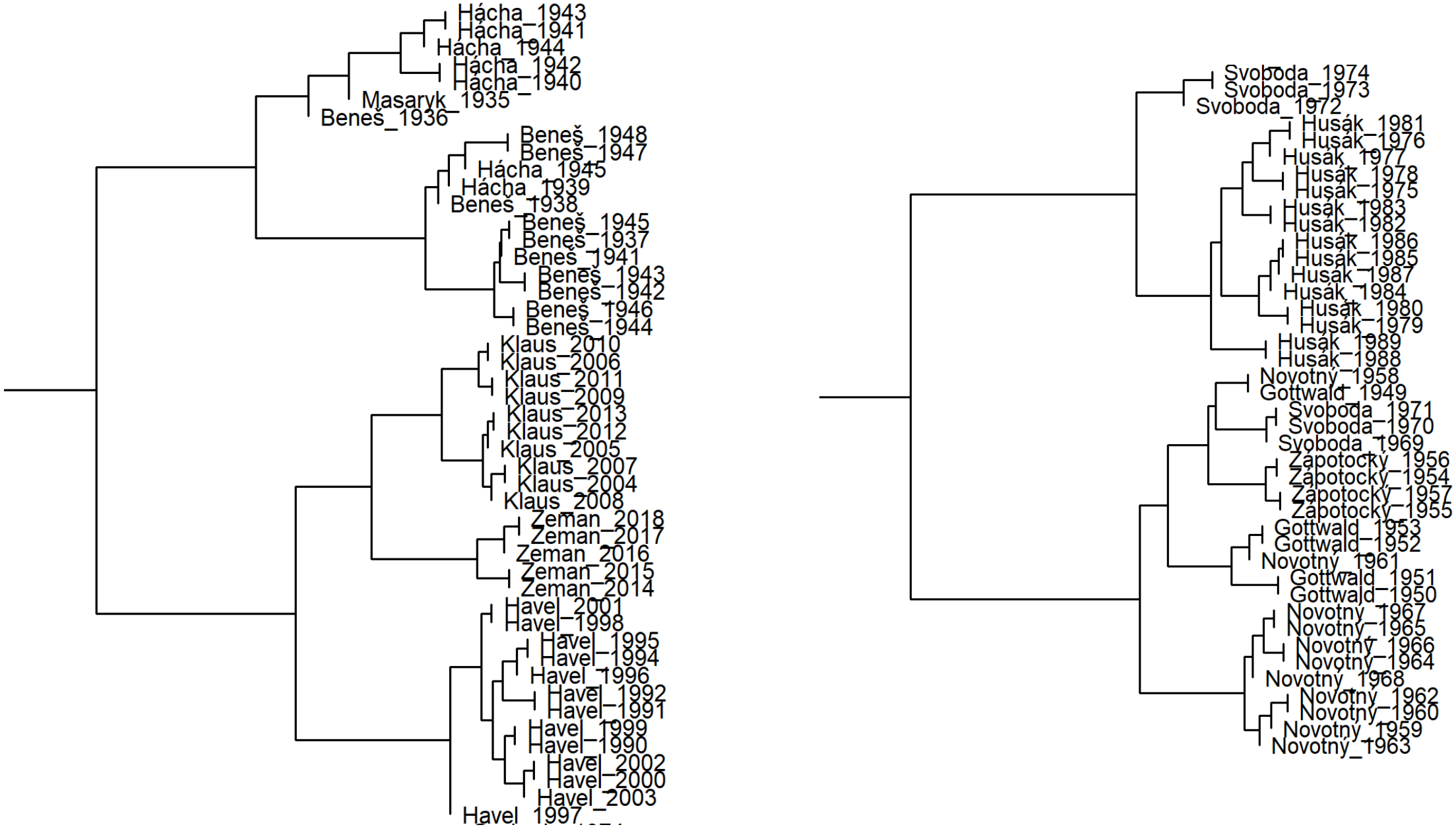
Prezidentské projevy

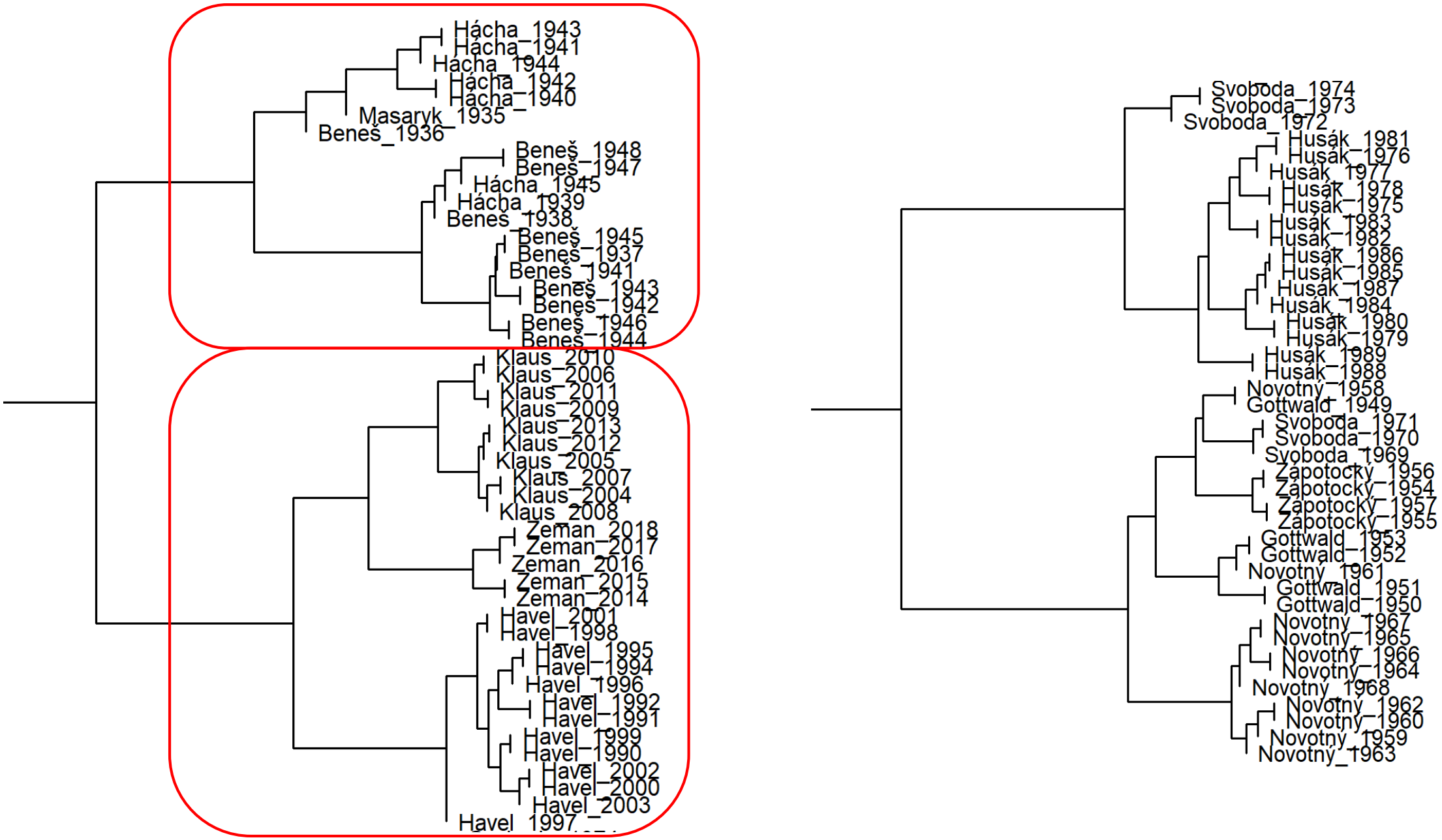
- Kubát, M., Mačutek, J., Čech, R. (2021). Communists spoke differently. An analysis of Czechoslovak and Czech annual presidential speeches. Digital Scholarship in the Humanities, 36, 138-152.

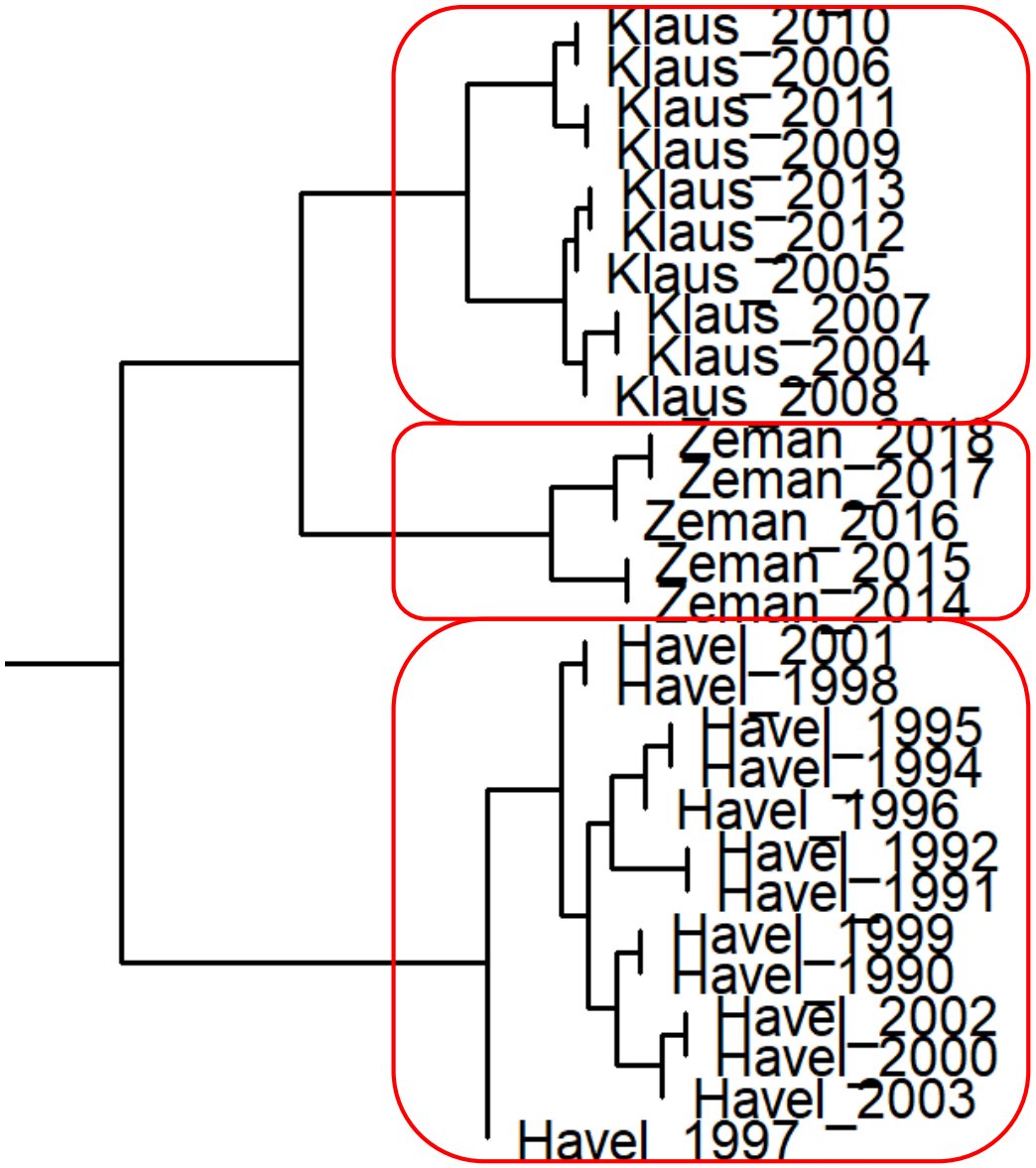
- presedential speeches: 1935–2018

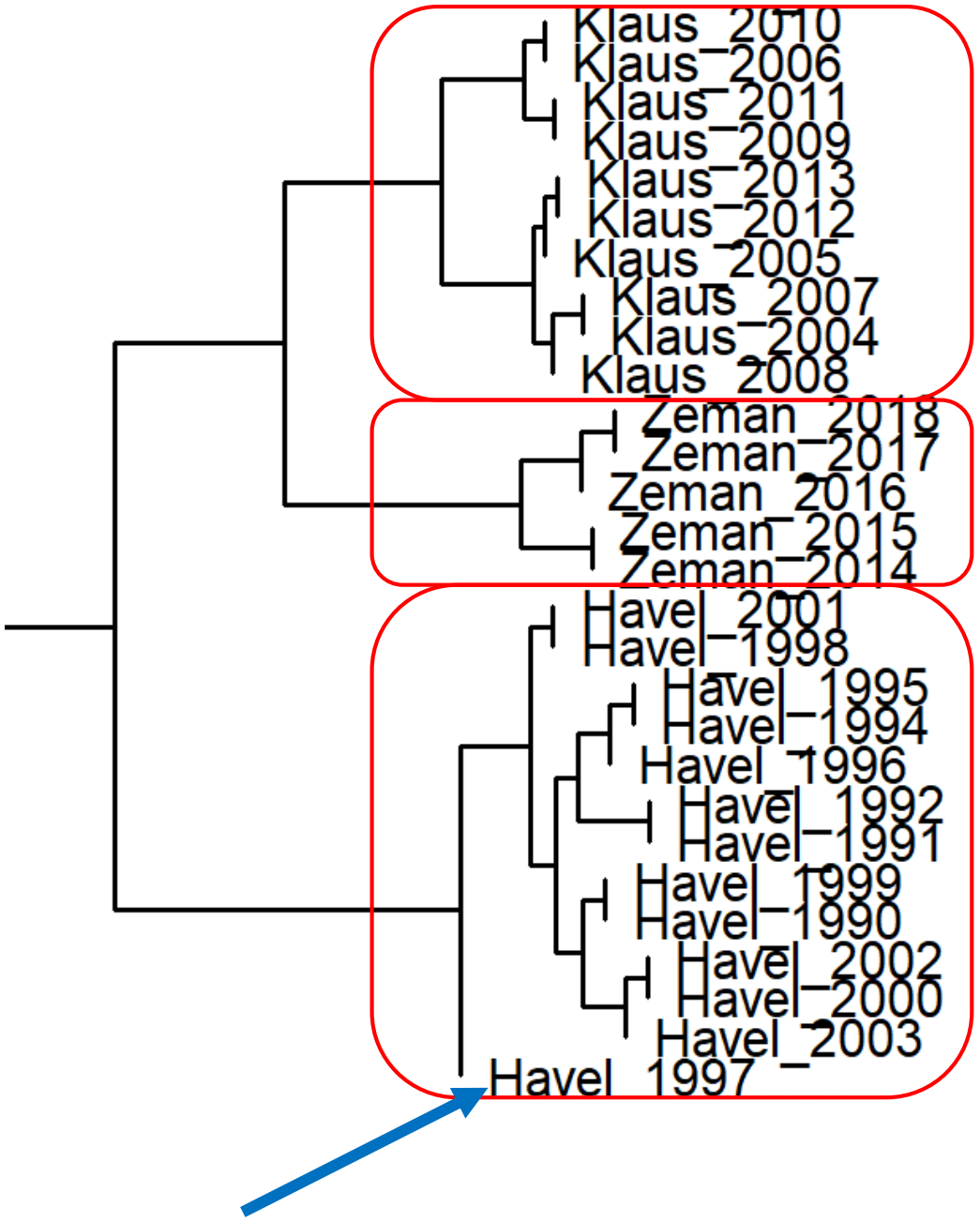
- 100 MFW, culling = 60 %

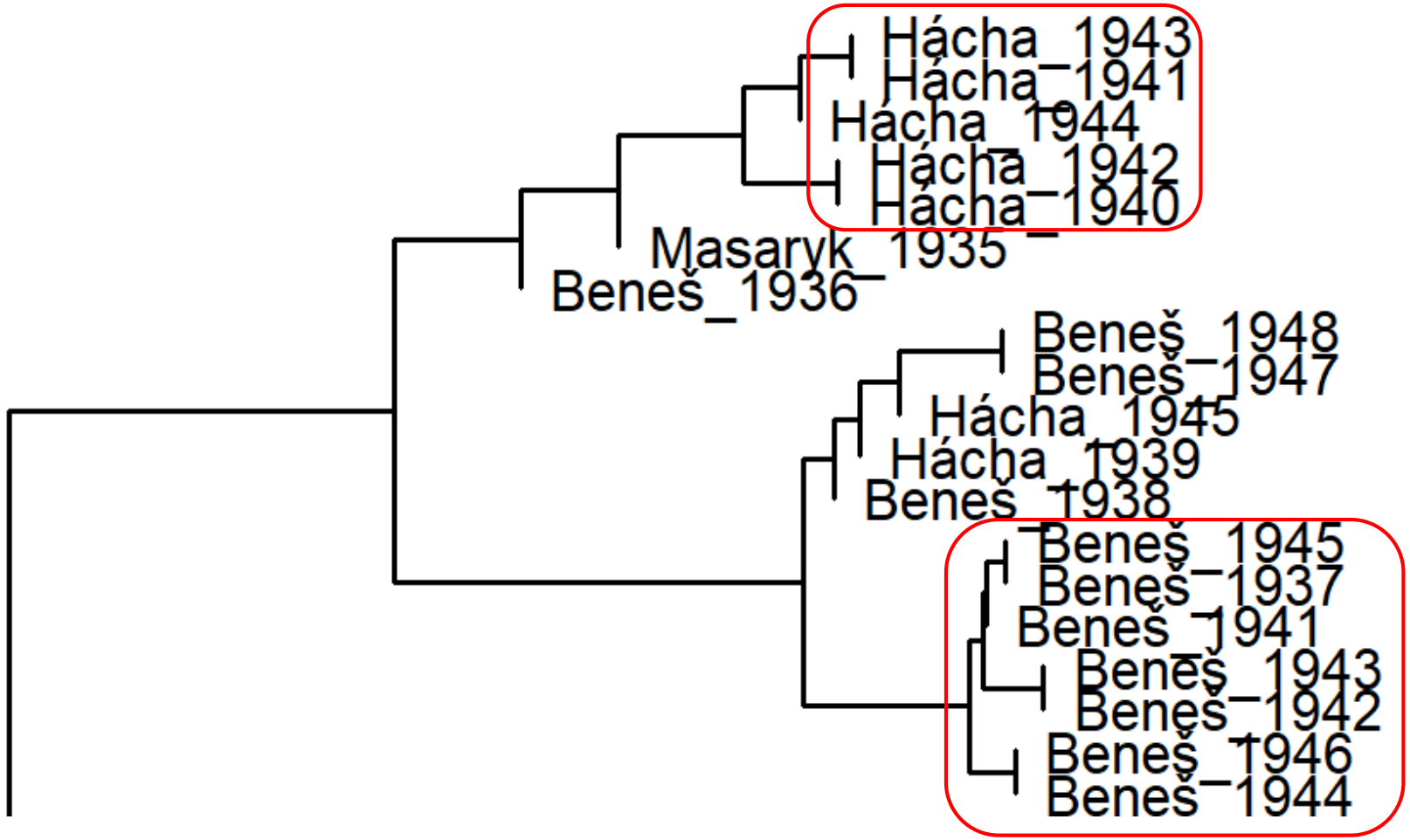

Hácha_1943
Hácha_1941
Hácha_1944
Hácha_1942
Hácha_1940
Masaryk_1935
Beneš_1936
Beneš_1948
Beneš_1947
Hácha_1945
Hácha_1939
Beneš_1938
Beneš_1945
Beneš_1937
Beneš_1941
Beneš_1943
Beneš_1942
Beneš_1946
Beneš_1944
Klaus_2010
Klaus_2006
Klaus_2011
Klaus_2009
Klaus_2013
Klaus_2012
Klaus_2005
Klaus_2007
Klaus_2004
Klaus_2008
Zeman_2018
Zeman_2017
Zeman_2016
Zeman_2015
Zeman_2014
Havel_2001
Havel_1998
Havel_1995
Havel_1994
Havel_1996
Havel_1992
Havel_1991
Havel_1999
Havel_1990
Havel_2002
Havel_2000
Havel_2003
Havel_1997
Svoboda_1974
Svoboda_1973
Svoboda_1972
Husák_1981
Husák_1976
Husák_1977
Husák_1978
Husák_1975
Husák_1983
Husák_1982
Husák_1986
Husák_1985
Husák_1987
Husák_1984
Husák_1980
Husák_1979
Husák_1989
Husák_1988
Novotný_1958
Gottwald_1949
Svoboda_1971
Svoboda_1970
Svoboda_1969
Zápotocký_1956
Zápotocký_1954
Zápotocký_1957
Zápotocký_1955
Gottwald_1953
Gottwald_1952
Novotný_1961
Gottwald_1951
Gottwald_1950
Novotný_1967
Novotný_1965
Novotný_1966
Novotný_1964
Novotný_1968
Novotný_1962
Novotný_1960
Novotný_1959
Novotný_1963

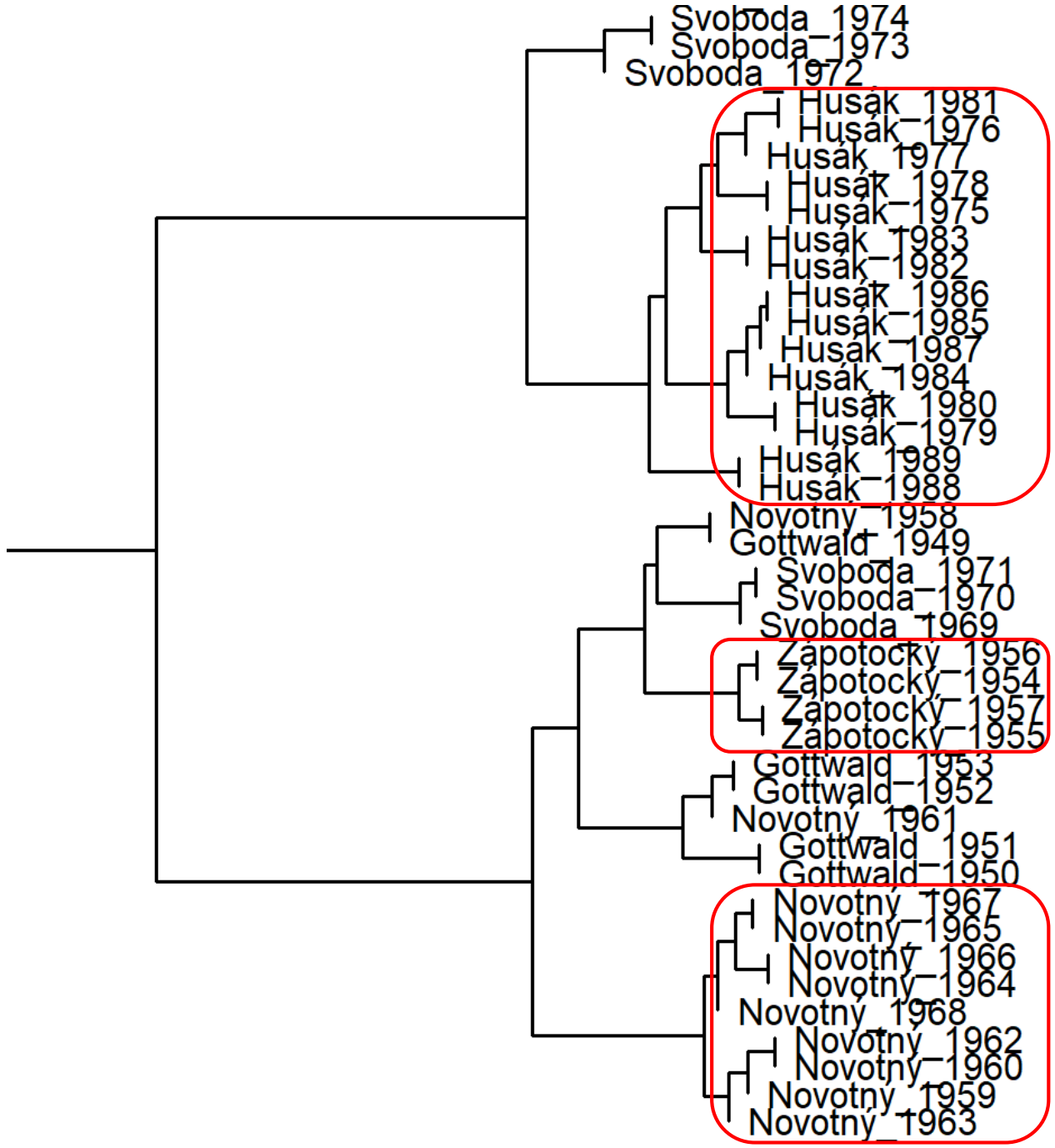

Hácha_1943
Hácha_1941
Hácha_1944
Hácha_1942
Hácha_1940
Masaryk_1935
Beneš_1936
Beneš_1948
Beneš_1947
Hácha_1945
Hácha_1939
Beneš_1938
Beneš_1945
Beneš_1937
Beneš_1941
Beneš_1943
Beneš_1942
Beneš_1946
Beneš_1944
Klaus_2010
Klaus_2006
Klaus_2011
Klaus_2009
Klaus_2013
Klaus_2012
Klaus_2005
Klaus_2007
Klaus_2004
Klaus_2008
Zeman_2018
Zeman_2017
Zeman_2016
Zeman_2015
Zeman_2014
Havel_2001
Havel_1998
Havel_1995
Havel_1994
Havel_1996
Havel_1992
Havel_1991
Havel_1999
Havel_1990
Havel_2002
Havel_2000
Havel_2003
Havel_1997
Svoboda_1974
Svoboda_1973
Svoboda_1972
Husák_1981
Husák_1976
Husák_1977
Husák_1978
Husák_1975
Husák_1983
Husák_1982
Husák_1986
Husák_1985
Husák_1987
Husák_1984
Husák_1980
Husák_1979
Husák_1989
Husák_1988
Novotný_1958
Gottwald_1949
Svoboda_1971
Svoboda_1970
Svoboda_1969
Zápotocký_1956
Zápotocký_1954
Zápotocký_1957
Zápotocký_1955
Gottwald_1953
Gottwald_1952
Novotný_1961
Gottwald_1951
Gottwald_1950
Novotný_1967
Novotný_1965
Novotný_1966
Novotný_1964
Novotný_1968
Novotný_1962
Novotný_1960
Novotný_1959
Novotný_1963

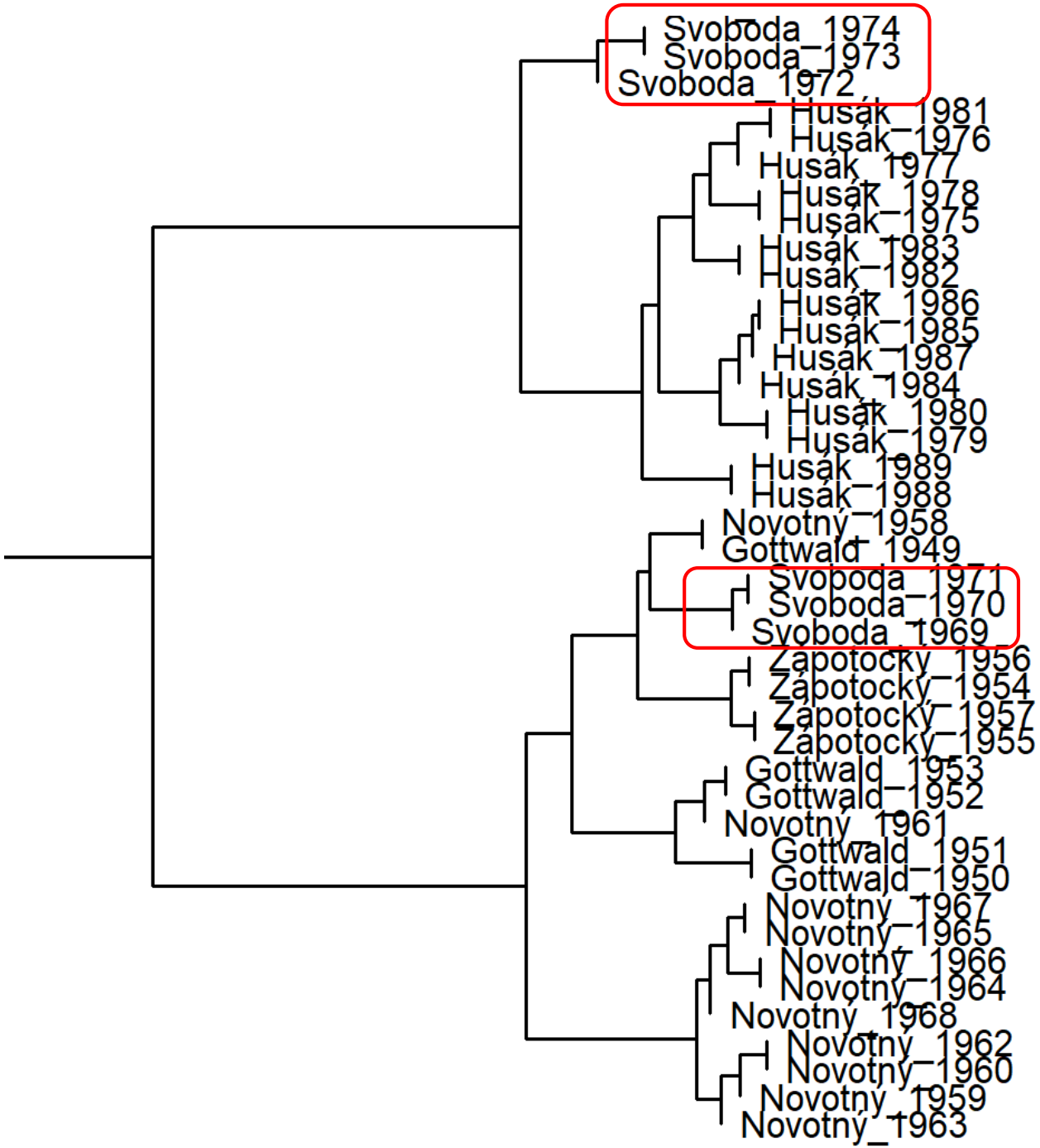

Hácha_1943
Hácha_1941
Hácha_1944
Hácha_1942
Hácha_1940
Masaryk_1935
Beneš_1936
Beneš_1948
Beneš_1947
Hácha_1945
Hácha_1939
Beneš_1938
Beneš_1945
Beneš_1937
Beneš_1941
Beneš_1943
Beneš_1942
Beneš_1946
Beneš_1944
Klaus_2010
Klaus_2006
Klaus_2011
Klaus_2009
Klaus_2013
Klaus_2012
Klaus_2005
Klaus_2007
Klaus_2004
Klaus_2008
Zeman_2018
Zeman_2017
Zeman_2016
Zeman_2015
Zeman_2014
Havel_2001
Havel_1998
Havel_1995
Havel_1994
Havel_1996
Havel_1992
Havel_1991
Havel_1999
Havel_1990
Havel_2002
Havel_2000
Havel_2003
Havel_1997
Svoboda_1974
Svoboda_1973
Svoboda_1972
Husák_1981
Husák_1976
Husák_1977
Husák_1978
Husák_1975
Husák_1983
Husák_1982
Husák_1986
Husák_1985
Husák_1987
Husák_1984
Husák_1980
Husák_1979
Husák_1989
Husák_1988
Novotný_1958
Gottwald_1949
Svoboda_1971
Svoboda_1970
Svoboda_1969
Zápotocký_1956
Zápotocký_1954
Zápotocký_1957
Zápotocký_1955
Gottwald_1953
Gottwald_1952
Novotný_1961
Gottwald_1951
Gottwald_1950
Novotný_1967
Novotný_1965
Novotný_1966
Novotný_1964
Novotný_1968
Novotný_1962
Novotný_1960
Novotný_1959
Novotný_1963


Klaus_2010
Klaus_2006
Klaus_2011
Klaus_2009
Klaus_2013
Klaus_2012
Klaus_2005
Klaus_2007
Klaus_2004
Klaus_2008
Zeman_2018
Zeman_2017
Zeman_2016
Zeman_2015
Zeman_2014
Havel_2001
Havel_1998
Havel_1995
Havel_1994
Havel_1996
Havel_1992
Havel_1991
Havel_1999
Havel_1990
Havel_2002
Havel_2000
Havel_2003
Havel_1997


Klaus_2010
Klaus_2006
Klaus_2011
Klaus_2009
Klaus_2013
Klaus_2012
Klaus_2005
Klaus_2007
Klaus_2004
Klaus_2008
Zeman_2018
Zeman_2017
Zeman_2016
Zeman_2015
Zeman_2014
Havel_2001
Havel_1998
Havel_1995
Havel_1994
Havel_1996
Havel_1992
Havel_1991
Havel_1999
Havel_1990
Havel_2002
Havel_2000
Havel_2003
Havel_1997


Hácha_1943
Hácha_1941
Hácha_1944
Hácha_1942
Hácha_1940
Masaryk_1935
Beneš_1936
Beneš_1948
Beneš_1947
Hácha_1945
Hácha_1939
Beneš_1938
Beneš_1945
Beneš_1937
Beneš_1941
Beneš_1943
Beneš_1942
Beneš_1946
Beneš_1944


Svoboda_1974
Svoboda_1973
Svoboda_1972
Husák_1981
Husák_1976
Husák_1977
Husák_1978
Husák_1975
Husák_1983
Husák_1982
Husák_1986
Husák_1985
Husák_1987
Husák_1984
Husák_1980
Husák_1979
Husák_1989
Husák_1988
Novotný_1958
Gottwald_1949
Svoboda_1971
Svoboda_1970
Svoboda_1969
Zápotocký_1956
Zápotocký_1954
Zápotocký_1957
Zápotocký_1955
Gottwald_1953
Gottwald_1952
Novotný_1961
Gottwald_1951
Gottwald_1950
Novotný_1967
Novotný_1965
Novotný_1966
Novotný_1964
Novotný_1968
Novotný_1962
Novotný_1960
Novotný_1959
Novotný_1963


Svoboda_1974
Svoboda_1973
Svoboda_1972
Husák_1981
Husák_1976
Husák_1977
Husák_1978
Husák_1975
Husák_1983
Husák_1982
Husák_1986
Husák_1985
Husák_1987
Husák_1984
Husák_1980
Husák_1979
Husák_1989
Husák_1988
Novotný_1958
Gottwald_1949
Svoboda_1971
Svoboda_1970
Svoboda_1969
Zápotocký_1956
Zápotocký_1954
Zápotocký_1957
Zápotocký_1955
Gottwald_1953
Gottwald_1952
Novotný_1961
Gottwald_1951
Gottwald_1950
Novotný_1967
Novotný_1965
Novotný_1966
Novotný_1964
Novotný_1968
Novotný_1962
Novotný_1960
Novotný_1959
Novotný_1963

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

- New Testament
  - Konstanc → Šteyer
- Old Testament
  - Šteyer → Barner
    - Job – divide

- „[…] (*domníváme se, že by se stylistickým rozdílem textu dalo zjistit, odkud překládal už jen Šteyer*) […]“ Vintr (1997)

Bible svatováclavská & comments

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

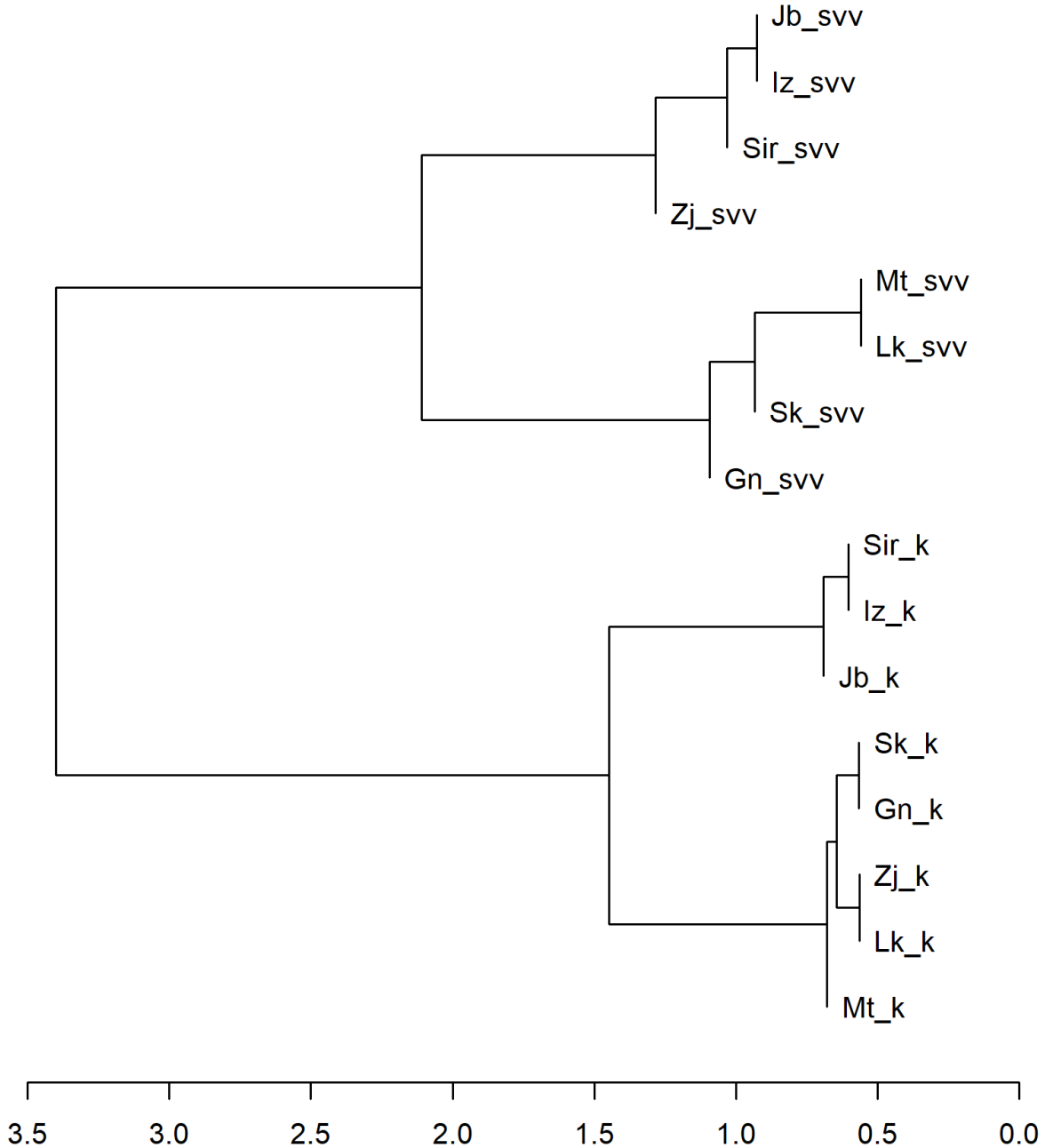
Bible svatováclavská & comments

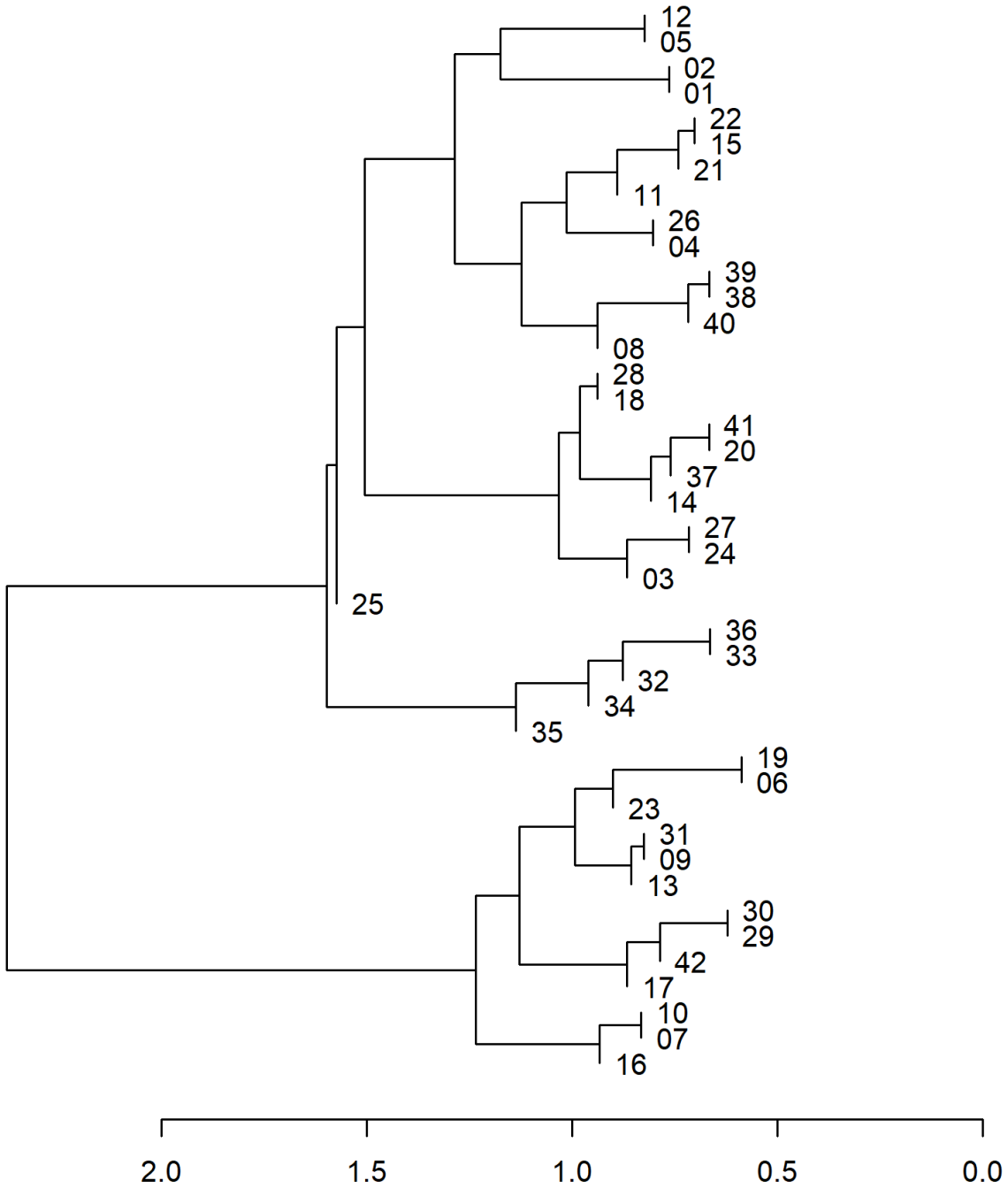
- translation
- sacred text
- editing

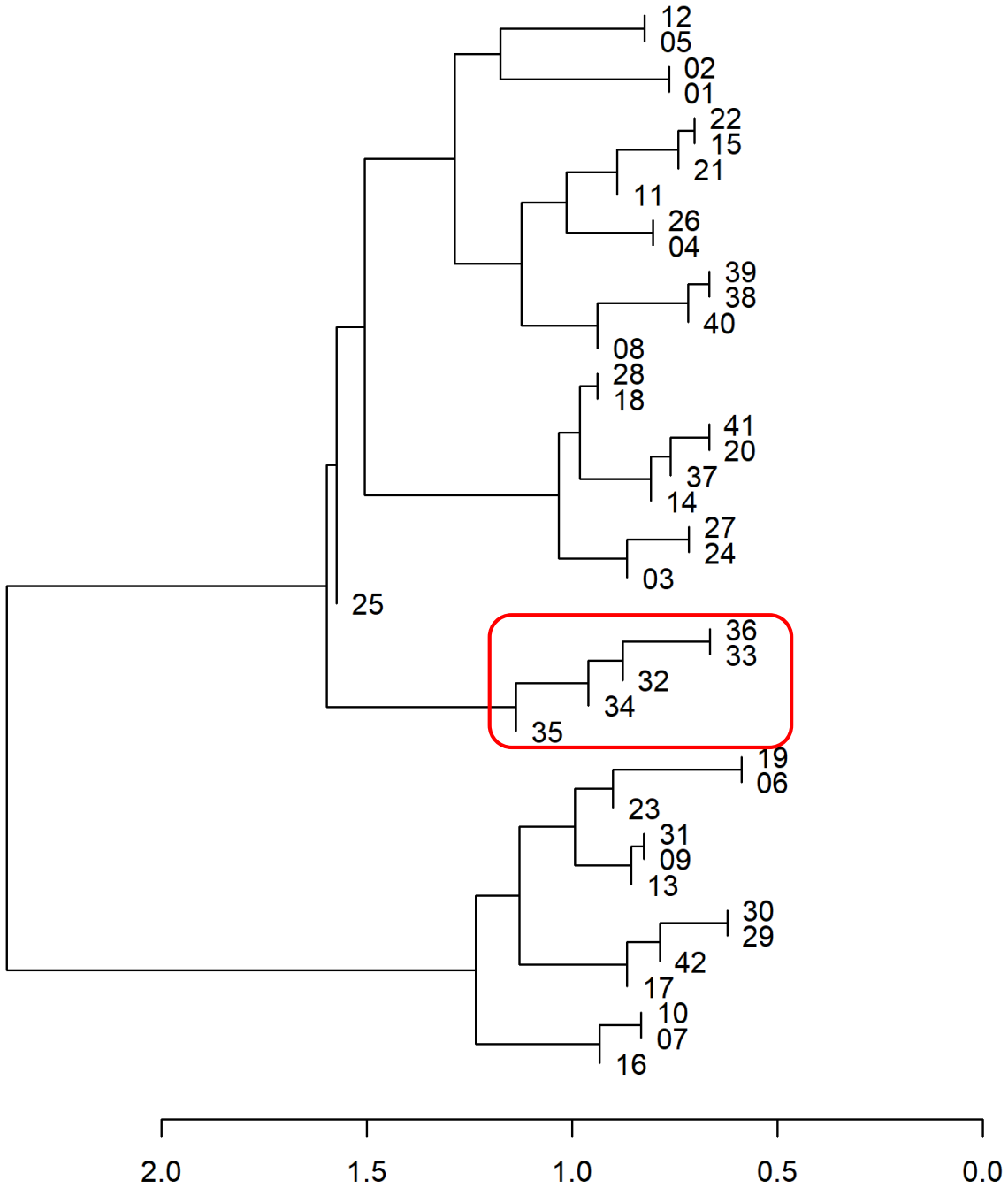
Bible svatováclavská & comments

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

- 100 MFW
- culling = 0

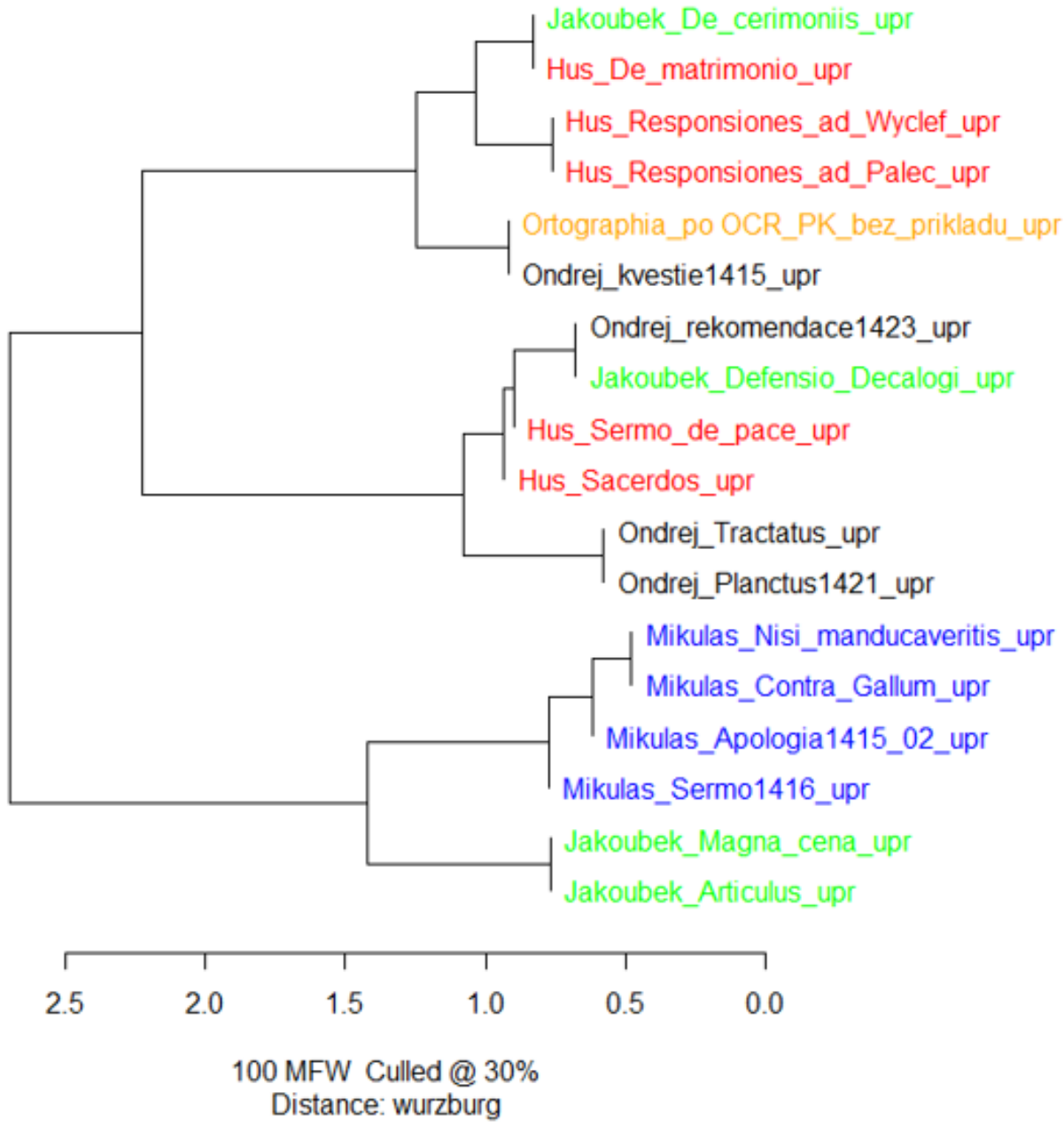

Jb_svv
Iz_svv
Sir_svv
Zj_svv
Mt_svv
Lk_svv
Sk_svv
Gn_svv
Sir_k
Iz_k
Jb_k
Sk_k
Gn_k
Zj_k
Lk_k
Mt_k
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

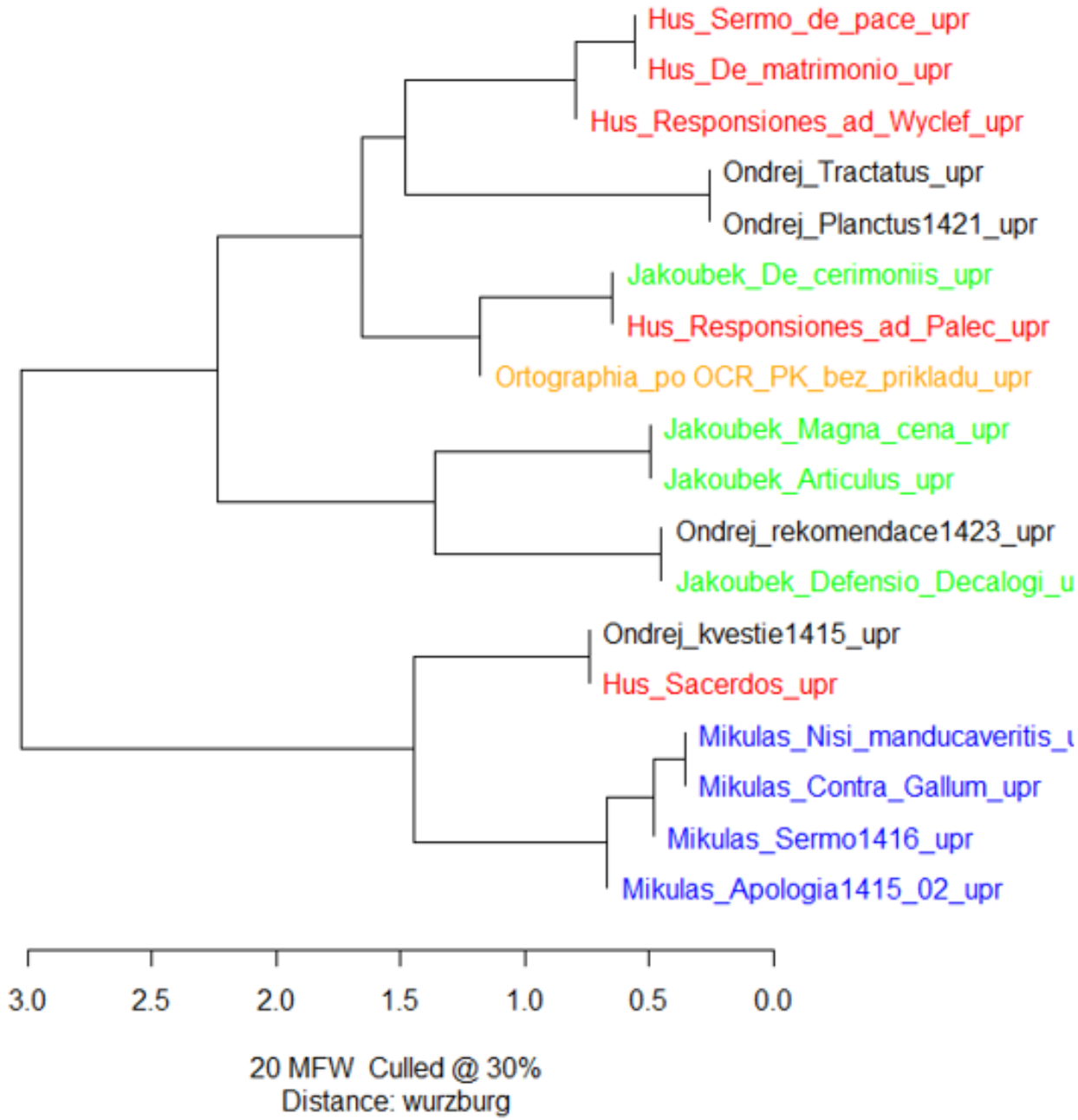

12
05
02
01
22
15
21
11
26
04
39
38
40
08
28
18
41
20
37
14
27
24
03
25
36
33
32
34
35
19
06
23
31
09
13
30
29
42
17
10
07
16
2.0
1.5
1.0
0.5
0.0


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

Ortographia Bohemica

- Jan Hus
  - De matrimonio, Responsiones ad Palec, Sacerdos, Sermo de pace
- Jakoubek ze Stříbra
  - Articulus, De cerimoniis, Defensio Decalogi, Magna cena
- Mikuláš z Drážďan
  - Apologia 1415, Contra Gallum, Nisi manducaveritis, Sermo 1416,
- Ondřej z Brodu
  - Kvestie 1415, Planctus 1421, Recomendace 1423, Tractatus,
- Petr z Pulky
  - Confutatio, Epistola, Sermo ite, De congruitate


Jakoubek_De_cerimoniis_upr
Hus_De_matrimonio_upr
Hus_Responsiones_ad_Wyclef_upr
Hus_Responsiones_ad_Palec_upr
Ortographia_po OCR_PK_bez_prikladu_upr
Ondrej_kvestie1415_upr
Ondrej_rekomendace1423_upr
Jakoubek_Defensio_Decalogi_upr
Hus_Sermo_de_pace_upr
Hus_Sacerdos_upr
Ondrej_Tractatus_upr
Ondrej_Planctus1421_upr
Mikulas_Nisi_manducaveritis_upr
Mikulas_Contra_Gallum_upr
Mikulas_Apologia1415_02_upr
Mikulas_Sermo1416_upr
Jakoubek_Magna_cena_upr
Jakoubek_Articulus_upr
2.5
2.0
1.5
1.0
0.5
0.0
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

Ortographia Bohemica

- paraphrases, non-attributed quotations → analysis of 'authorial' parts

- Latin

Syntactic functions

- Čech, R., Mačutek, J., Kubát, M., Koščová, M. (2022). Does an author leave a syntactic footprint? In: Misuraca, M., Scepi, G., Spano, M. Proceedings of the 16th International Conference on statistical analysis of textual data. Naples: Vadistat Press, 221-228.

- presidential speeches: 1935–2018

Syntactic functions

- syntactic annotation (UD pipe)

- syntactic relations
  - UD → 37 universal syntactic relations
  - subject, object, indirect object, nominal modifier, adverbial modifier etc.

Syntactic functions

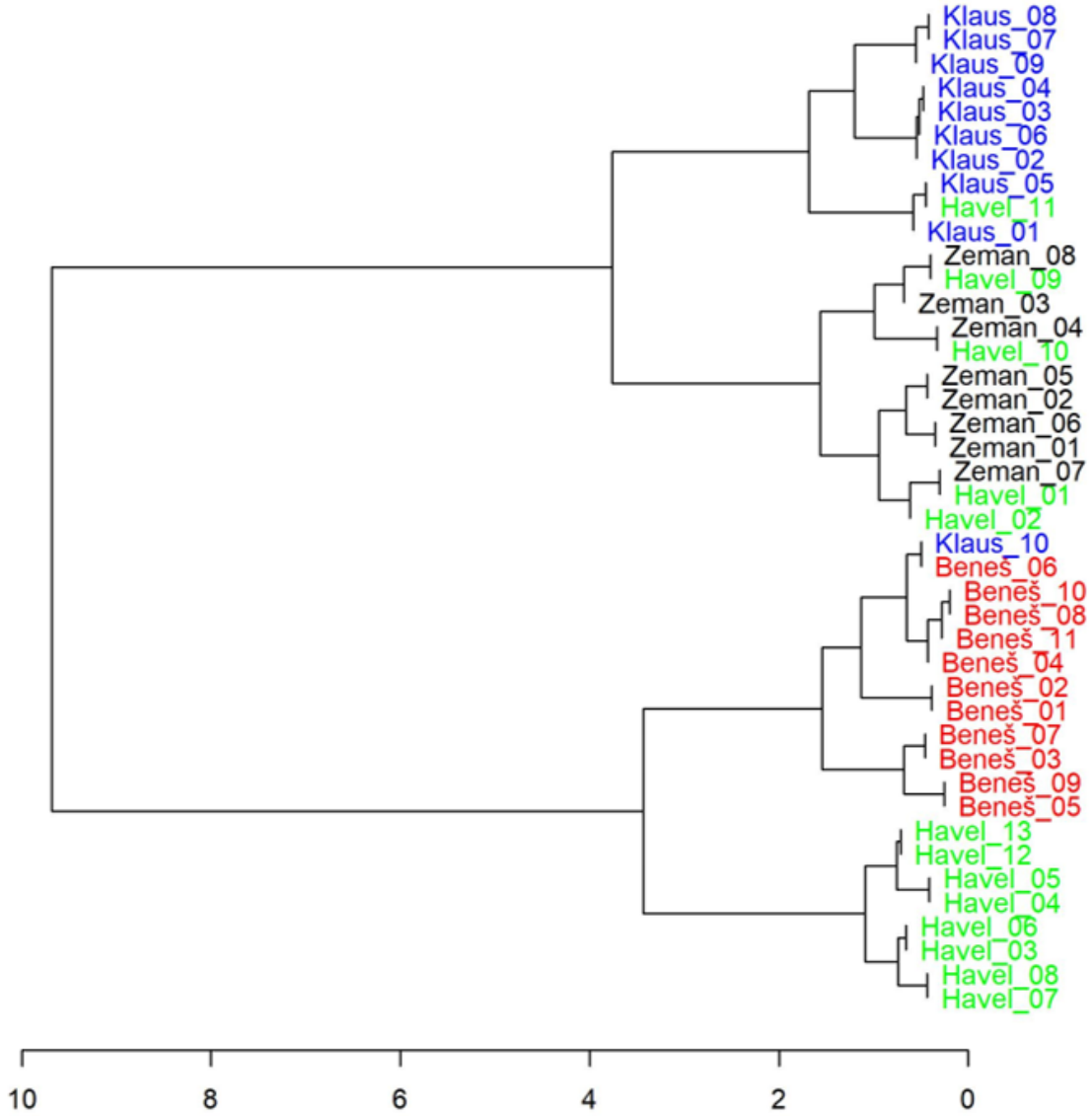
- Milí spoluobčané, když jsem vás před pěti lety v tento den poprvé oslovil, řekl jsem, že do našich srdcí se opět vrátila naděje

- amod nsubj mark aux obj case nummod obl case det obl obl advcl root aux mark case det obl expl:pv advmod ccomp obj

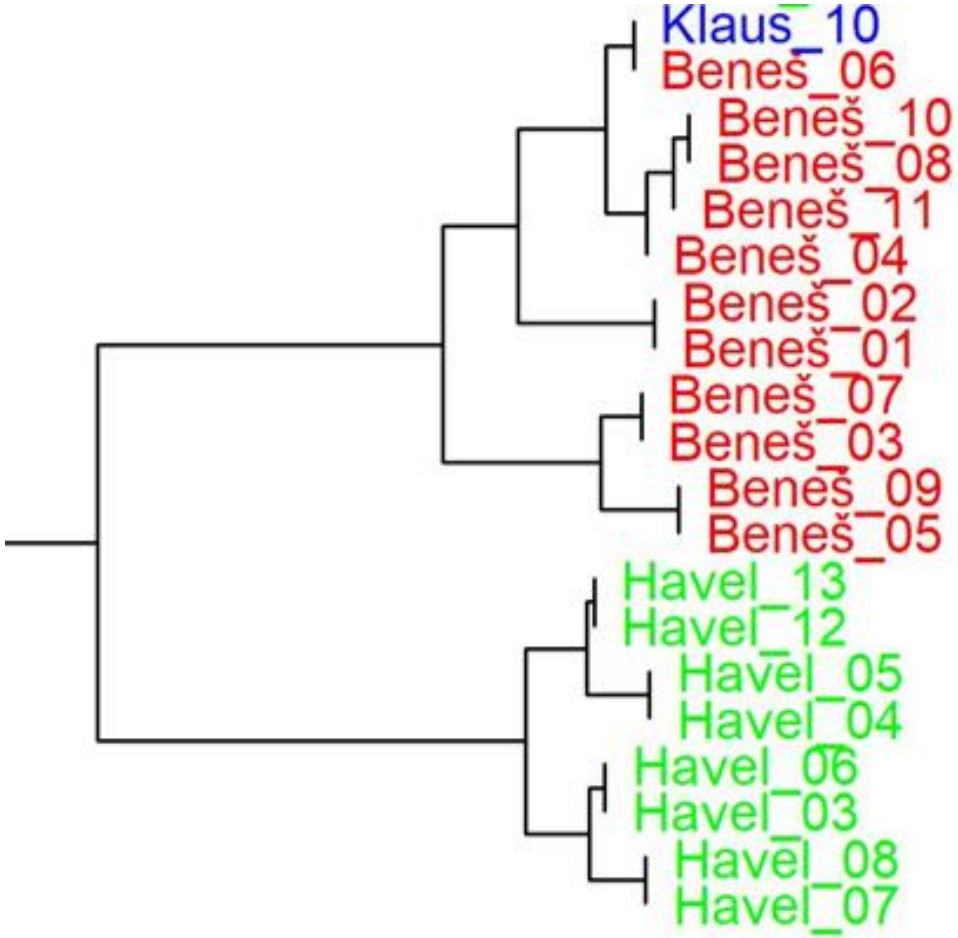
Syntactic functions

- democratic addresses
  - 22 most frequent syntactic relations
- communist addresses
  - 18 most frequent syntactic relations
- configuration
  - culling = 0
  - n-gram size = 1

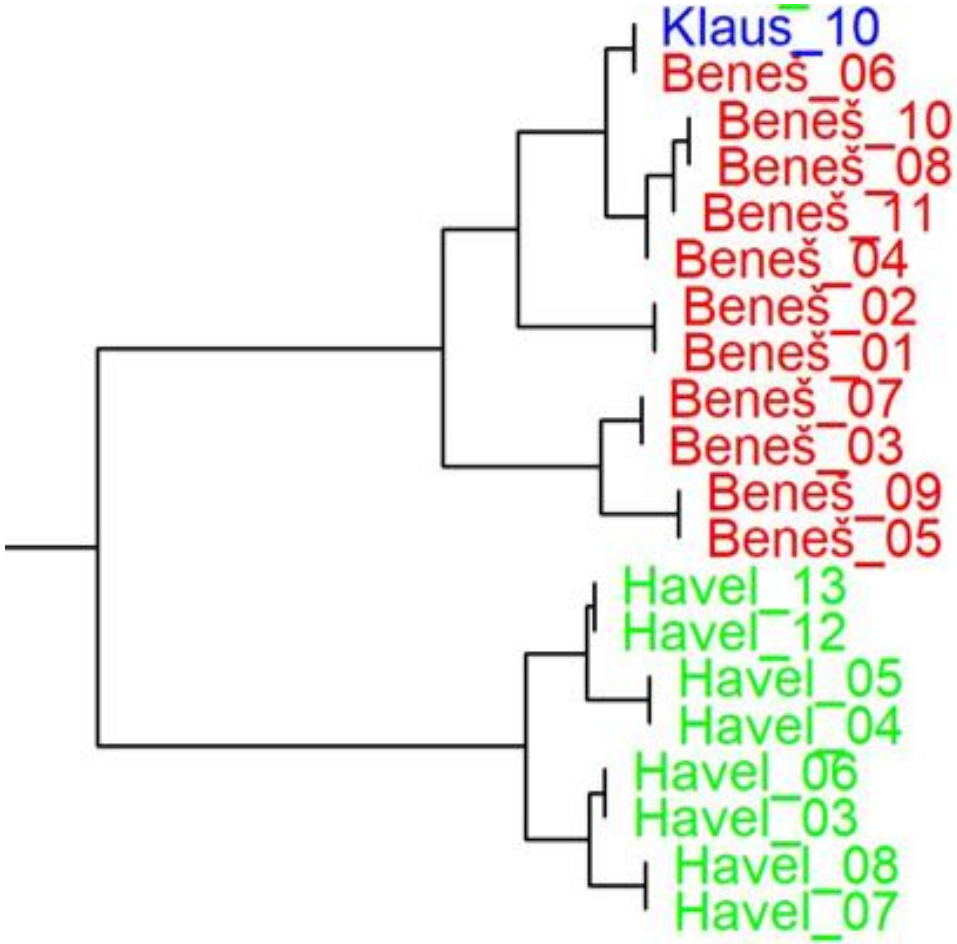
Results – democratic presidents

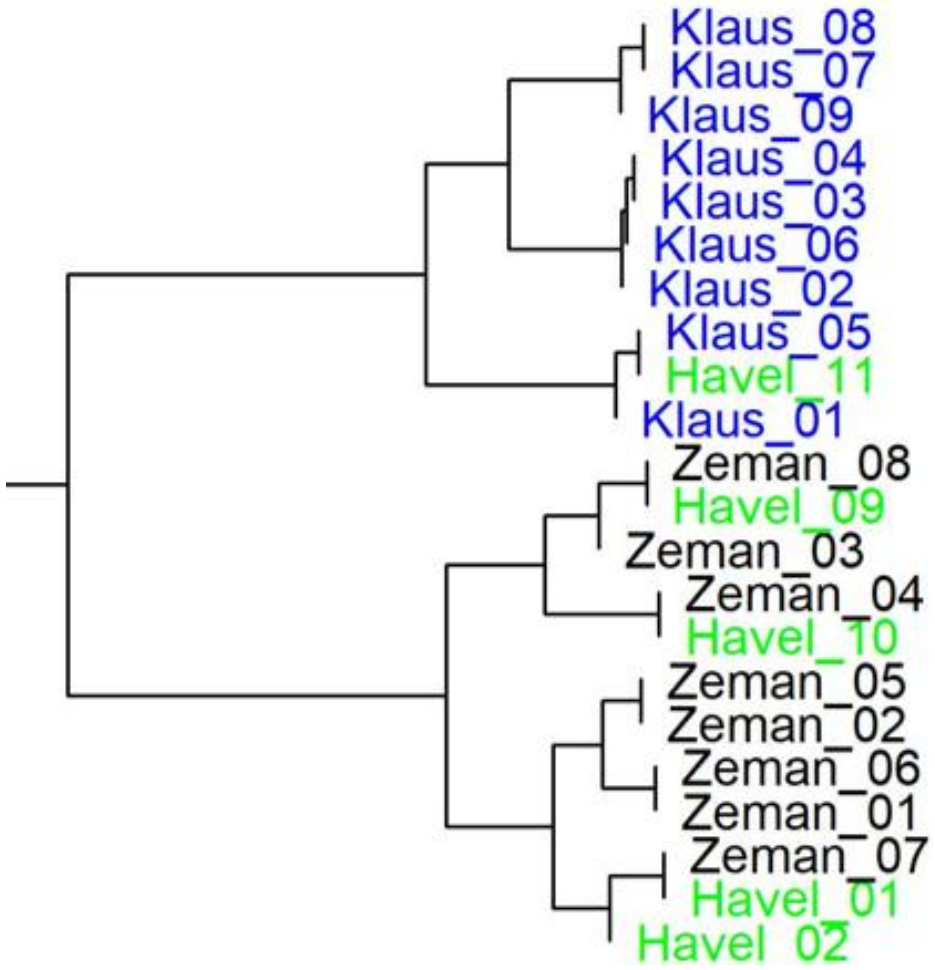


Results – democratic presidents

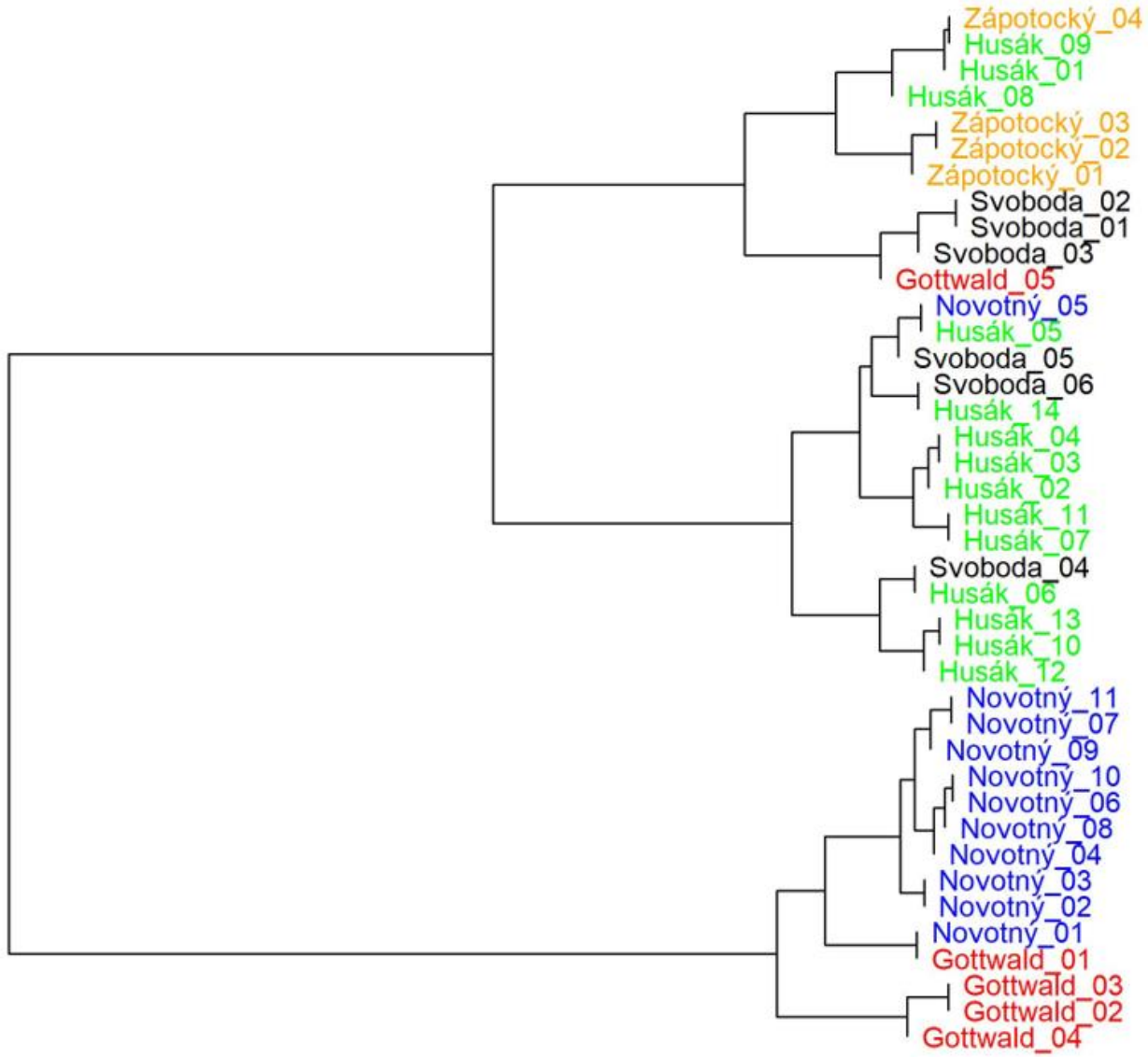


Results – democratic presidents

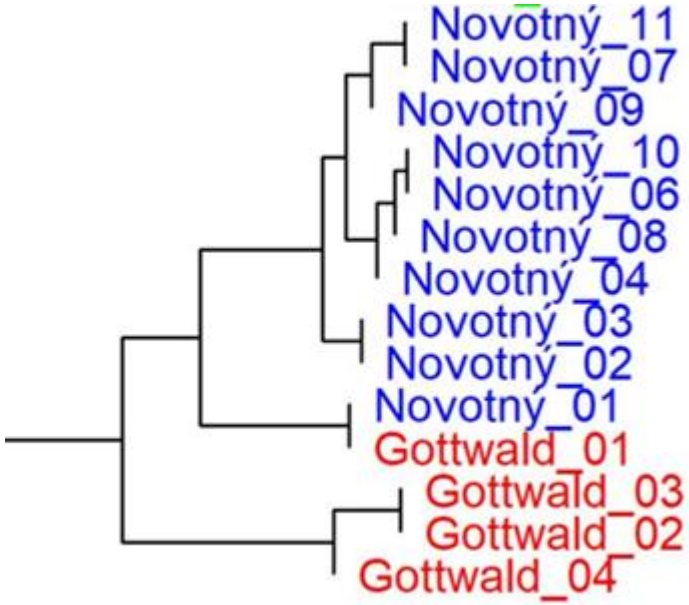




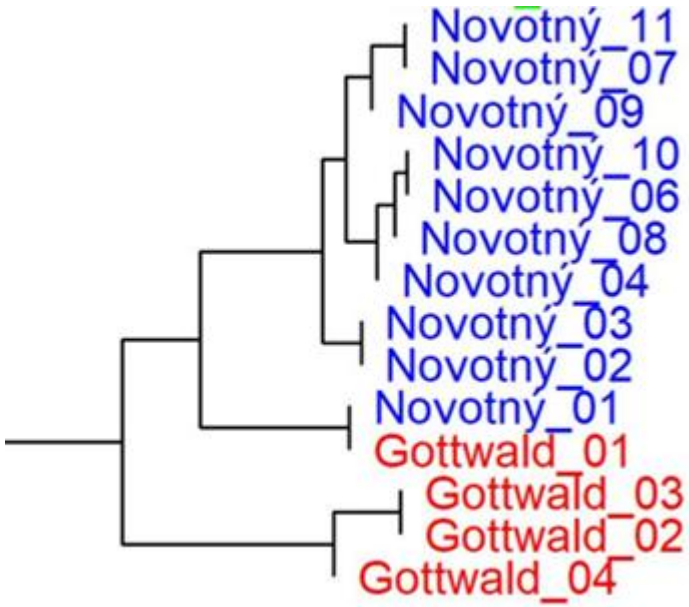
Results – communist presidents

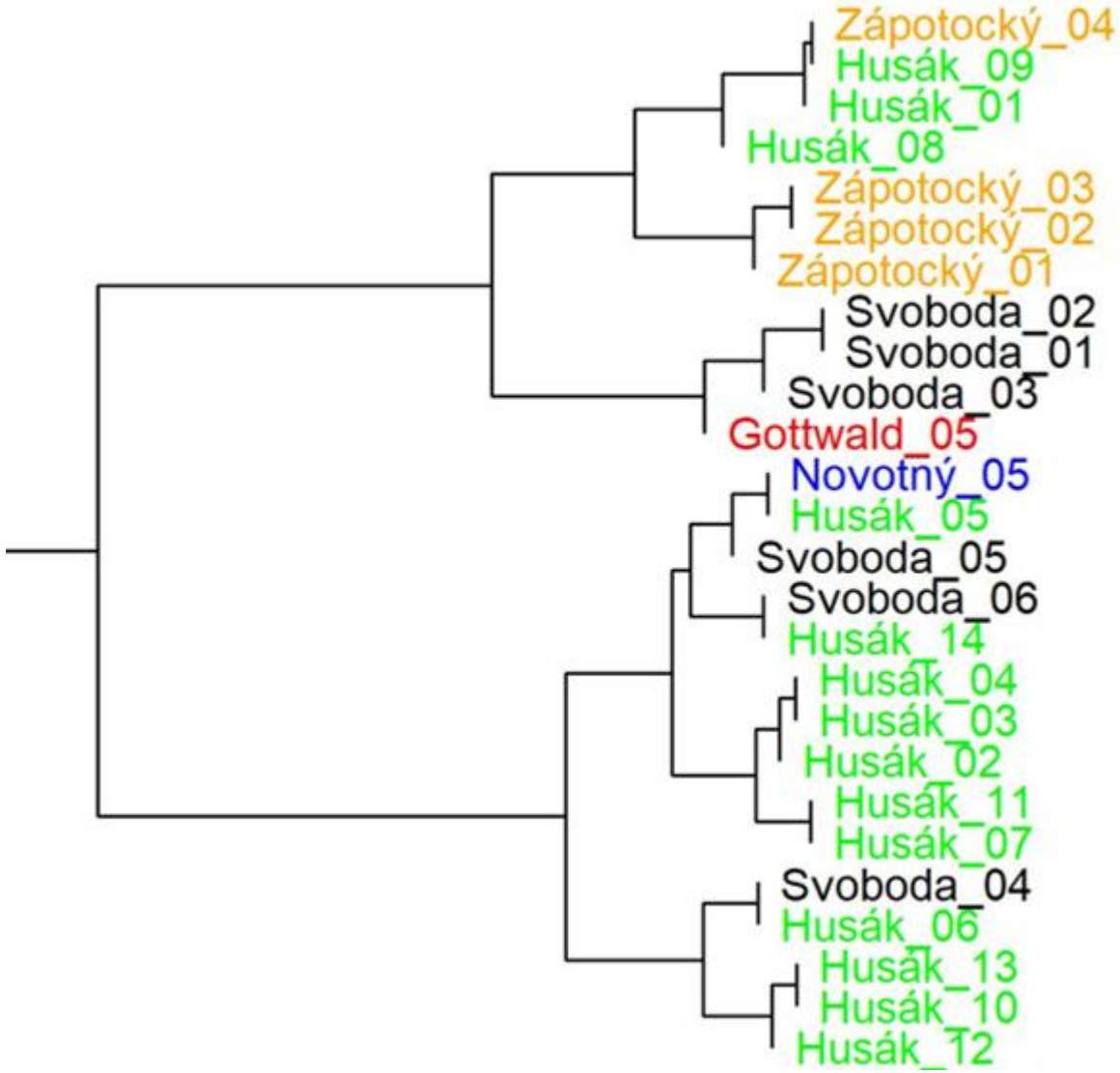


Results – communist presidents



Results – communist presidents

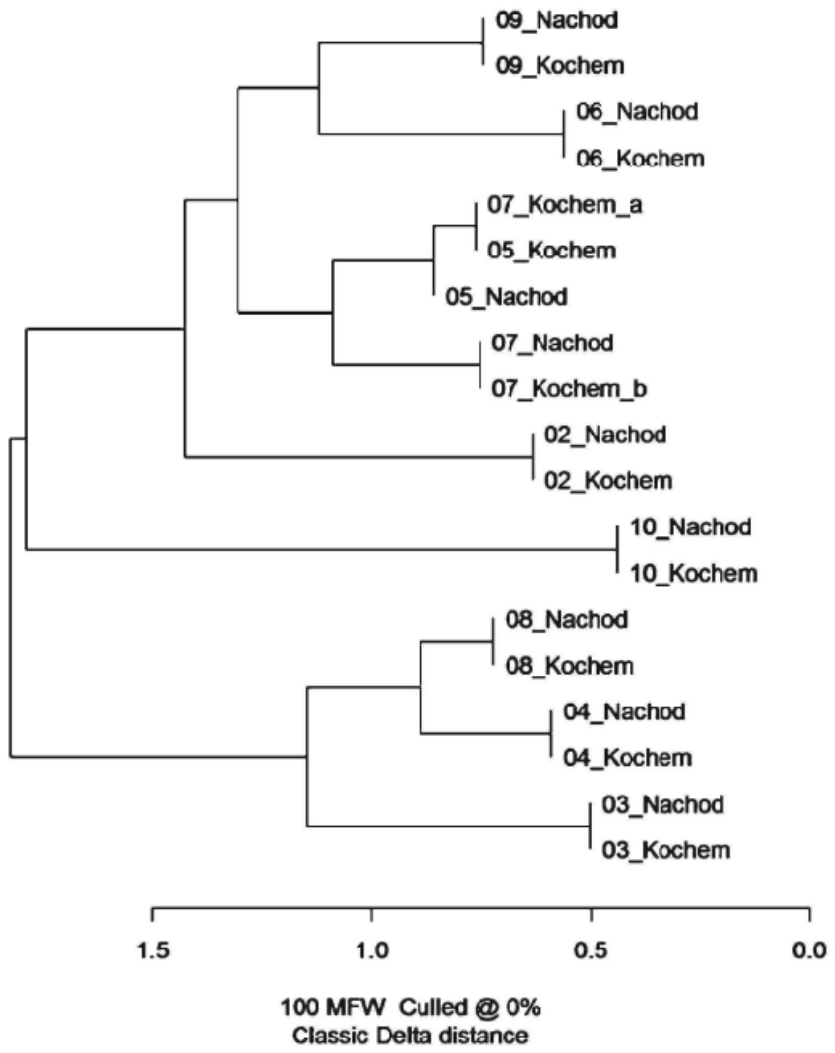




Baroque Prayers

- 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



Thank you
for your attention!