Stylometric Analysis of Church Slavonic Texts of Czech Origin

Miroslav Vepřek
Palacký University Olomouc

Radek Čech
Masaryk University in Brno

Faculty of Arts

MUNI ARTS

Research goals

- applicability of stylometric methods to (Old) Church Slavonic texts
- heuristic approach (without prior assumptions)
  - may these methods lead to the identification of new perspectives?
- stylometric analysis as a means of testing existing hypotheses
  - based on philologcal research

Material (Corpus) of Czech CS Texts

- Church Slavonic language
  - solely written language
  - used in different Slavic areas from the end of the 9th century in various variants (local „redactions“)

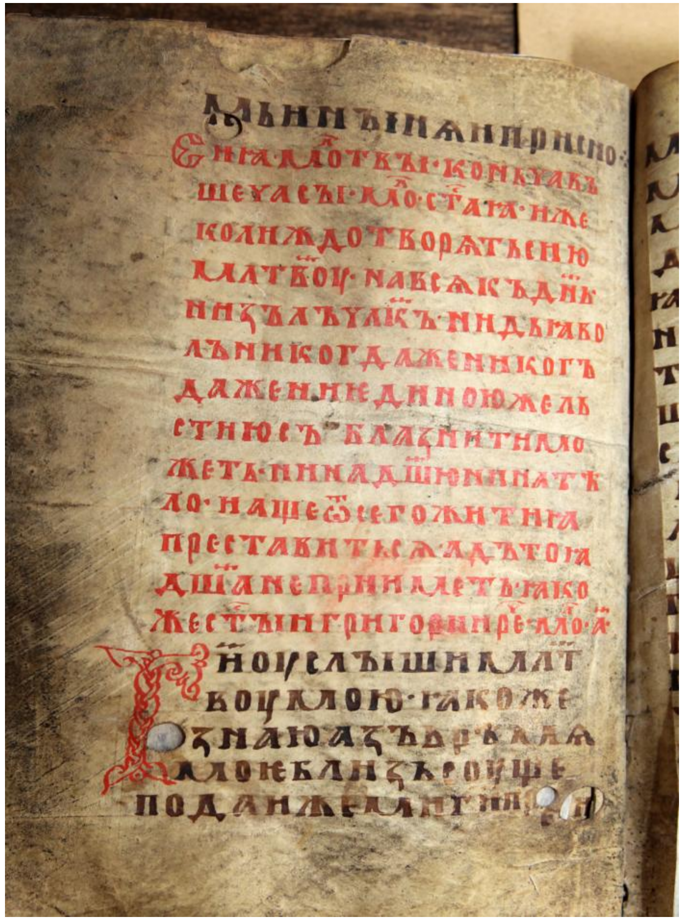
- 13 literary texts of various lengths
  - the *Forty Gospel Homilies of Gregory the Great* (93.358 tokens)
  - the *Prayer of Confession of Sins* (533 tokens)

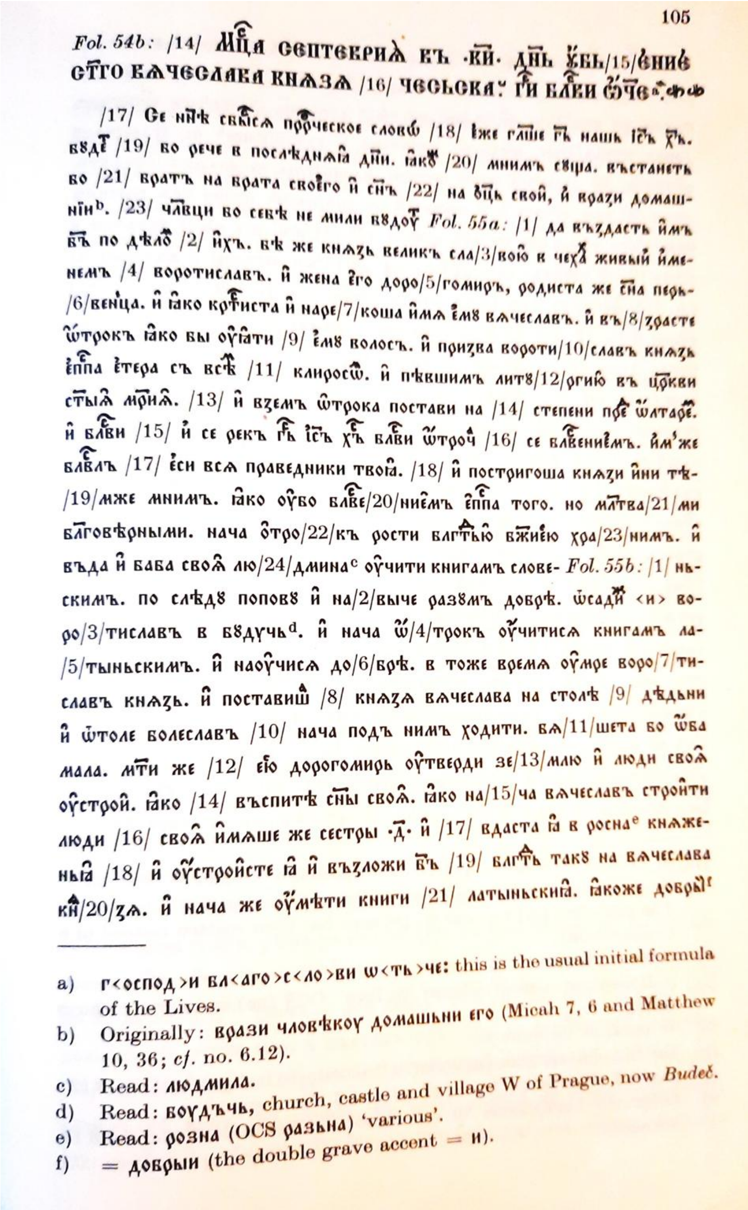
- proven or at least hypothetical Czech origin (10th-11th centuries)
  - preserved in later copies in different redactions of CS (mostly East Slavic)

Material (Corpus) of Czech CS Texts

- original writings and translations from Latin
- various genres
  - legends, prayers, homilies, a pseudo-Gospel
- the *Codex Suprasliensis* and *Vita Methodii* added for a comparison

Digitalization





Fol. 54b: /14/ **Мца септевриа въ ·ки· днь оувь/15/ение стго вачеслава кназа /16/ чесьска: ги блви ѿче[a].**

/17/ Се нн҇ѣ свысѧ прорческое словѡ /18/ еже глше гь нашь ісъ хъ. вꙋдеть /19/ во рече в послѣднѧя днн. ꙗко /20/ миимъ свѣра. въстанеть во /21/ вратъ на врата своего и снъ /22/ на ѡца свой, и врази домашнін[b]. /23/ члвци во севѣ не мили вꙋдоуть *Fol. 55a:* /1/ да въздасть имъ бъ по дѣлѡ /2/ ихъ. вѣ же кнѧзь великъ сла/3/вою в чеха живый именемъ /4/ воротиславъ. и жена его доро/5/гомирь, родиста же сна перь/6/веньца. и ꙗко крстиста и наре/7/коша имѧ емꙋ вачеславъ. и въ/8/зрасте ѿтрокъ ꙗко бы оустати /9/ емꙋ волосъ. и призва вороти/10/славъ кнѧзь еппа етера съ всѣ /11/ клиросѡ. и пѣвшимъ литꙋ/12/ргию въ цркви стыа мрнѧ. /13/ и вземъ ѿтрока постави на /14/ степени пре ѡлтаре. и блви /15/ и се рекъ гь ісъ хъ блви ѿтрочѧ /16/ се блвениемъ. имже блвлъ /17/ еси всѧ праведники твоꙗ. /18/ и постригоша кнѧзи ини тѣ/19/мже мнимъ. ꙗко оубо блве/20/ниемъ еппа того. но млтва/21/ми блговѣрными. нача отро/22/къ рости блгтью бжиею хра/23/нимъ. и въда и баба своꙗ лю/24/дмина[c] оучити книгамъ слове- *Fol. 55b:* /1/ нь- скимъ. по слѣдꙋ поповꙋ и на/2/выче разꙋмъ добрѣ. ѡсади <и> во/ро/3/тиславъ в бꙋдꙋчь[d]. и нача ѿ/4/трокъ оучитисѧ книгамъ ла/5/тыньскимъ. и наоучисѧ до/6/брѣ. в тоже времѧ оумре воро/7/тиславъ кнѧзь. и поставиша /8/ кнѧзѧ вачеслава на столѣ /9/ дѣдьни и ѿтоле болеславъ /10/ нача подъ нимъ ходити. вѧ/11/шета бо ѡба мала. мти же /12/ ею дорогомирь оутверди зе/13/млю и люди своꙗ оустрой. ꙗко /14/ въспитѣ сны своꙗ. ꙗко на/15/ча вачеславъ строити люди /16/ своꙗ имѧше же сестры ·д· и /17/ вдаста ꙗ в росна[e] кнѧже/ныꙗ /18/ и оустройсте ꙗ и възложи бъ /19/ блгть такꙋ на вачеслава кн/20/зѧ. и нача же оумѣти книги /21/ латыньскиꙗ. ꙗкоже добры[f]

---

a) г<оспод>и бл<аго>с<ло>ви ѡ<ть>че: this is the usual initial formula of the Lives.

b) Originally: врази чловѣкоу домашьни его (Micah 7, 6 and Matthew 10, 36; *cf.* no. 6.12).

c) Read: людмила.

d) Read: боудъчь, church, castle and village W of Prague, now *Budeč*.

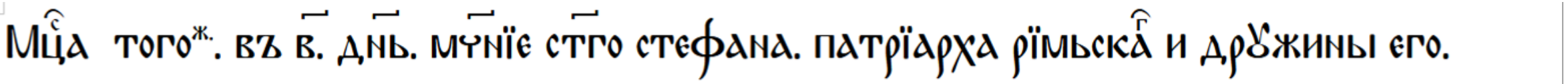
e) Read: розна (OCS разьна) 'various'.

f) = добрыи (the double grave accent = и).

Digitalization

- font Bukyvede
- unification of punctuation
- expansion of abbreviations
- special character for the numerical value of letters
- unification of variant graphemes: и (и, й, і, ї, ı), оу (оу, у, ѹ), etc.

Digitalization



м‹ѣ›с‹а›ца тогож‹е› · въ в* · д‹ь›нь · м‹ж›ч‹е›нїе с‹ва›т‹аѥ›го стефана · патрїарха рїмьска‹ѥ›г‹о› и дружины его ·

мѣсаца тогоже · въ в* · дьнь · мжчени сватаѥго стефана · патриарха римьскаѥго и дроужинъı его ·

Methods

- average token length
- moving average type-token ratio
- distances between texts
  - the most frequent words
  - Cosine delta distance

Average token length (ATL)

- length of the word measured in graphemes

$$ATL_{text} = \frac{\sum_{i=1}^{N} L_i}{N}$$

L … length of the word

N…number of words in text

Moving Average Type-Token Ratio (MATTR)

- vocabulary diversity (lexical richness)
- based on the type token ratio
- measures lexical diversity using moving windows

Peter loves Mary. John loves Mary too.

Moving Average Type-Token Ratio (MATTR)

- vocabulary diversity lLexical richness)
- based on the type token ratio
- measures lexical diversity using moving windows

Peter loves Mary. ***John*** *loves* *Mary too.*

$$TTR_1 = \frac{V}{N} = \frac{4}{5} = 0.8$$

Moving Average Type-Token Ratio (MATTR)

- vocabulary diversity (Lexical richness)
- based on the type token ratio
- measures lexical diversity using moving windows

*Peter **loves Mary**. **John** loves Mary too.*

$$TTR_2 = \frac{V}{N} = \frac{3}{5} = 0.6$$

Moving Average Type-Token Ratio (MATTR)

- vocabulary diversity (Lexical richness)
- based on the type token ratio
- measures lexical diversity using moving windows

Peter loves ***Mary***. ***John loves*** *Mary* ***too***.

$$TTR_3 = \frac{V}{N} = \frac{4}{5} = 0.8$$

Moving Average Type-Token Ratio (MATTR)

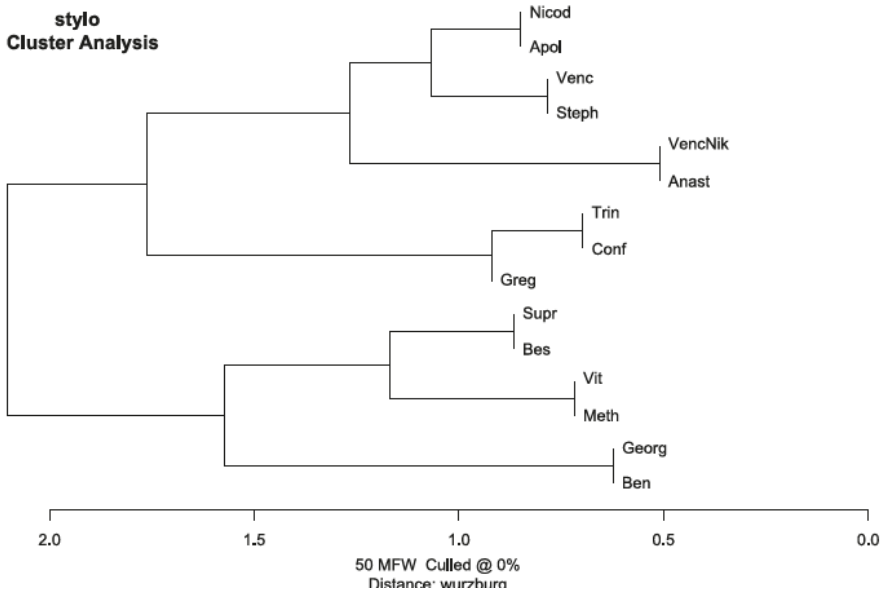
- vocabulary diversity (Lexical richness)
- based on the type token ratio
- measures lexical diversity using moving windows

Peter loves Mary. John loves Mary too.

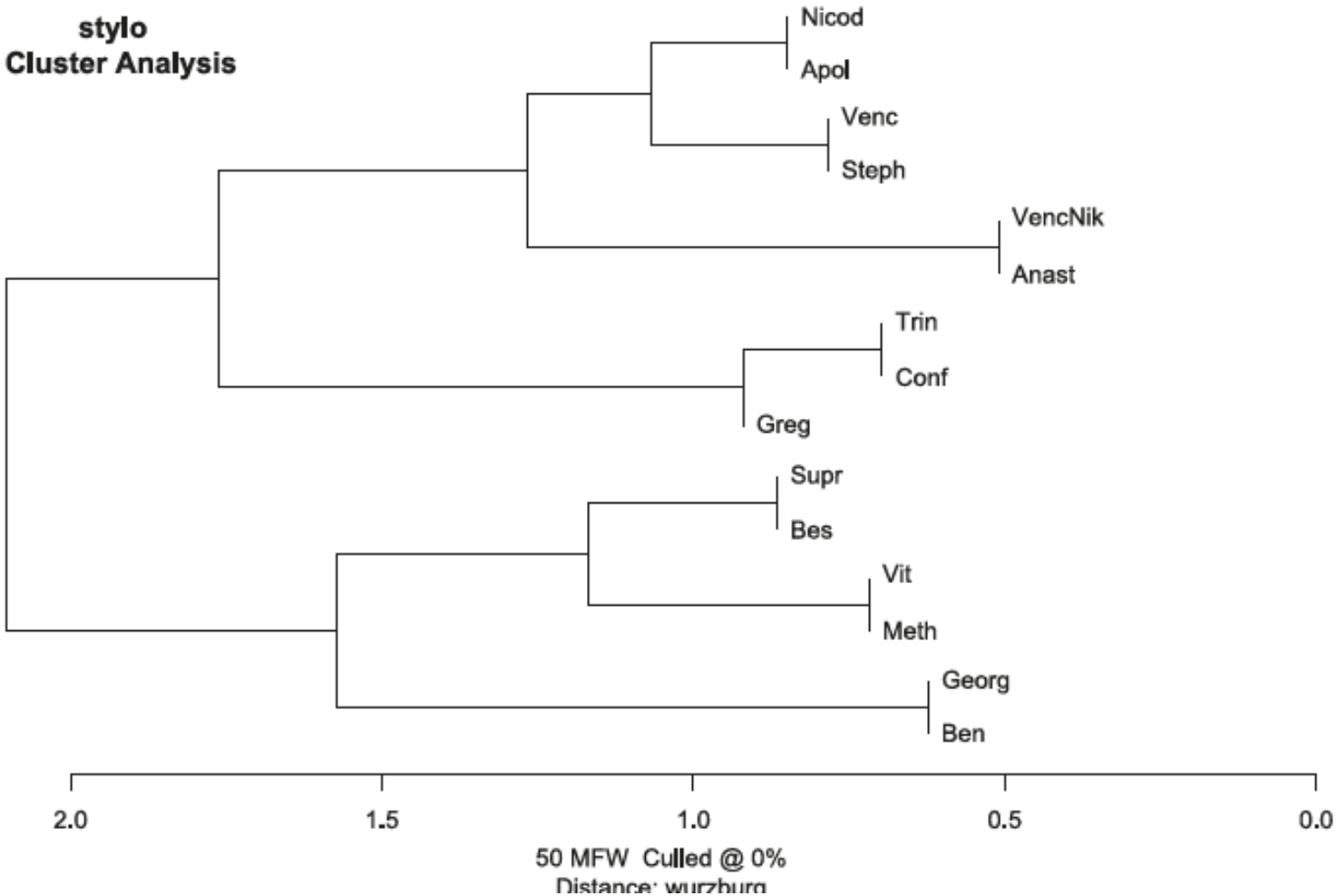
$$MATTR = \frac{0.8 + 0.6 + 0.8}{3} = 0.733$$

Distances between texts

- relative frequency of the most frequent words
  - in this analysis: 50 and 200
- Cosine Delta distance
- hierarchical cluster analysis
- Stylo software



Distances between texts



Results

- two sets of data:
  1. the *Forty Gospel Homilies of Gregory the Great* (*Bes*) as 1 file
  2. *Bes* divided into 46 parts (mainly according to chapters)

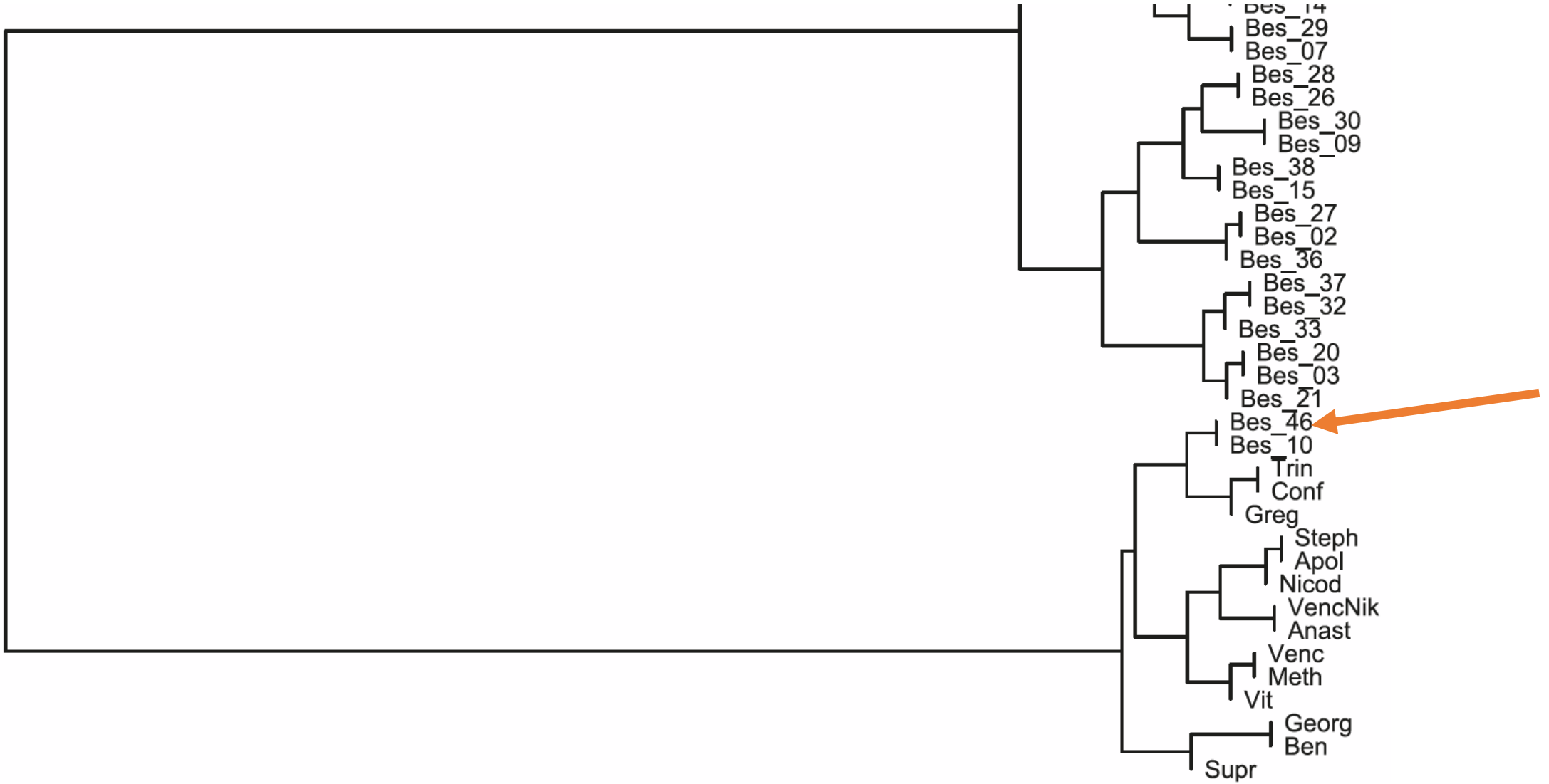
- reasons
  - to reduce the large differences in text length
  - to account for possible internal heterogeneity of *Bes*

- in general, **genre** characteristic is the main factor

A) Prayers

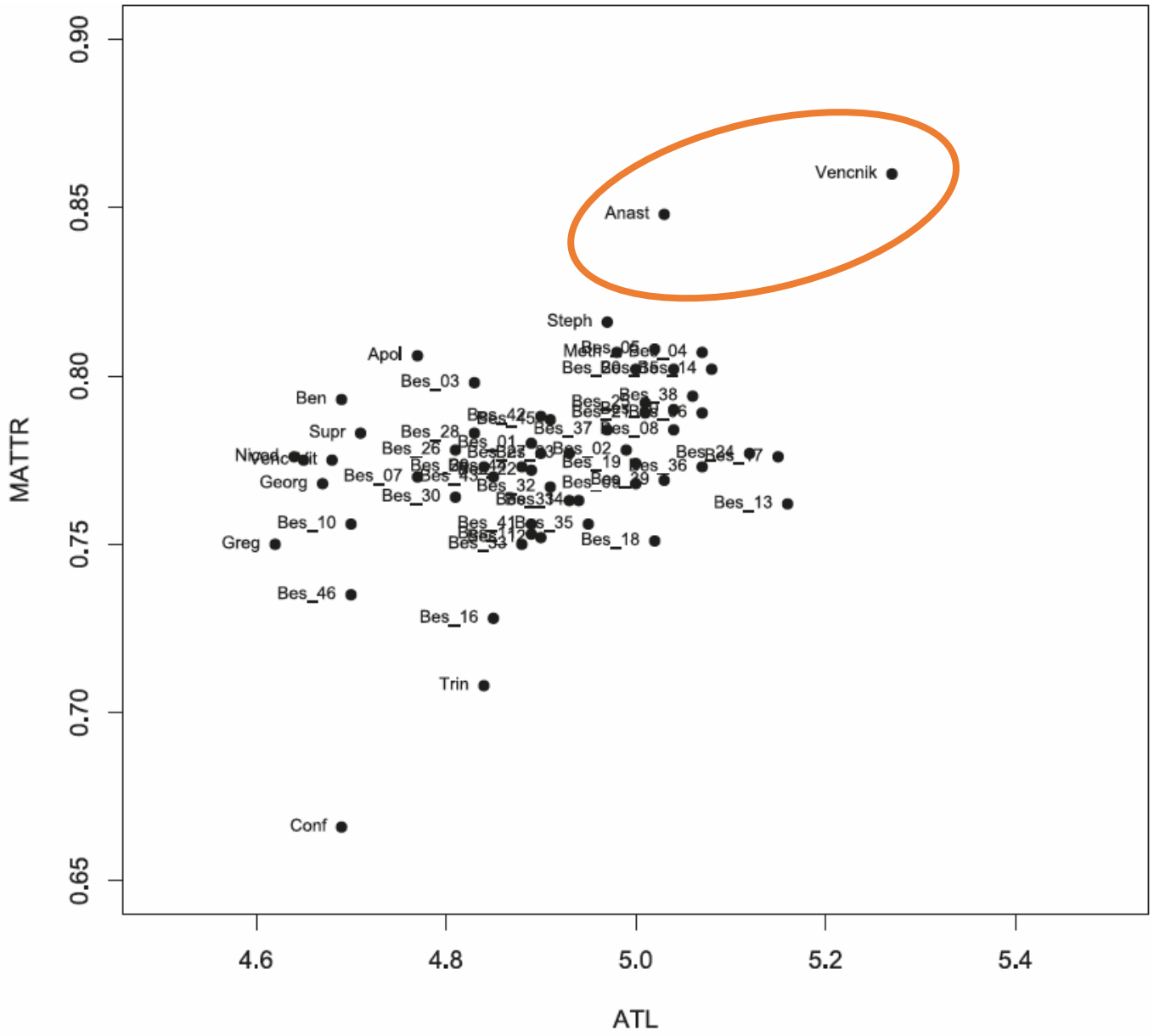
- *Trin*, *Conf*, *Greg*
  - stable cluster at both MFW 50 and MFW 200
  - MATTR – relatively low lexical diversity
  - correlation between MATTR and ATL

- *Bes*, segment 46 (prayer added to the base text)
  - joining *Trin, Conf, Greg* cluster


Bes_29
Bes_07
Bes_28
Bes_26
Bes_30
Bes_09
Bes_38
Bes_15
Bes_27
Bes_02
Bes_36
Bes_37
Bes_32
Bes_33
Bes_20
Bes_03
Bes_21
Bes_46
Bes_10
Trin
Conf
Greg
Steph
Apol
Nicod
VencNik
Anast
Venc
Meth
Vit
Georg
Ben
Supr

B) Legends

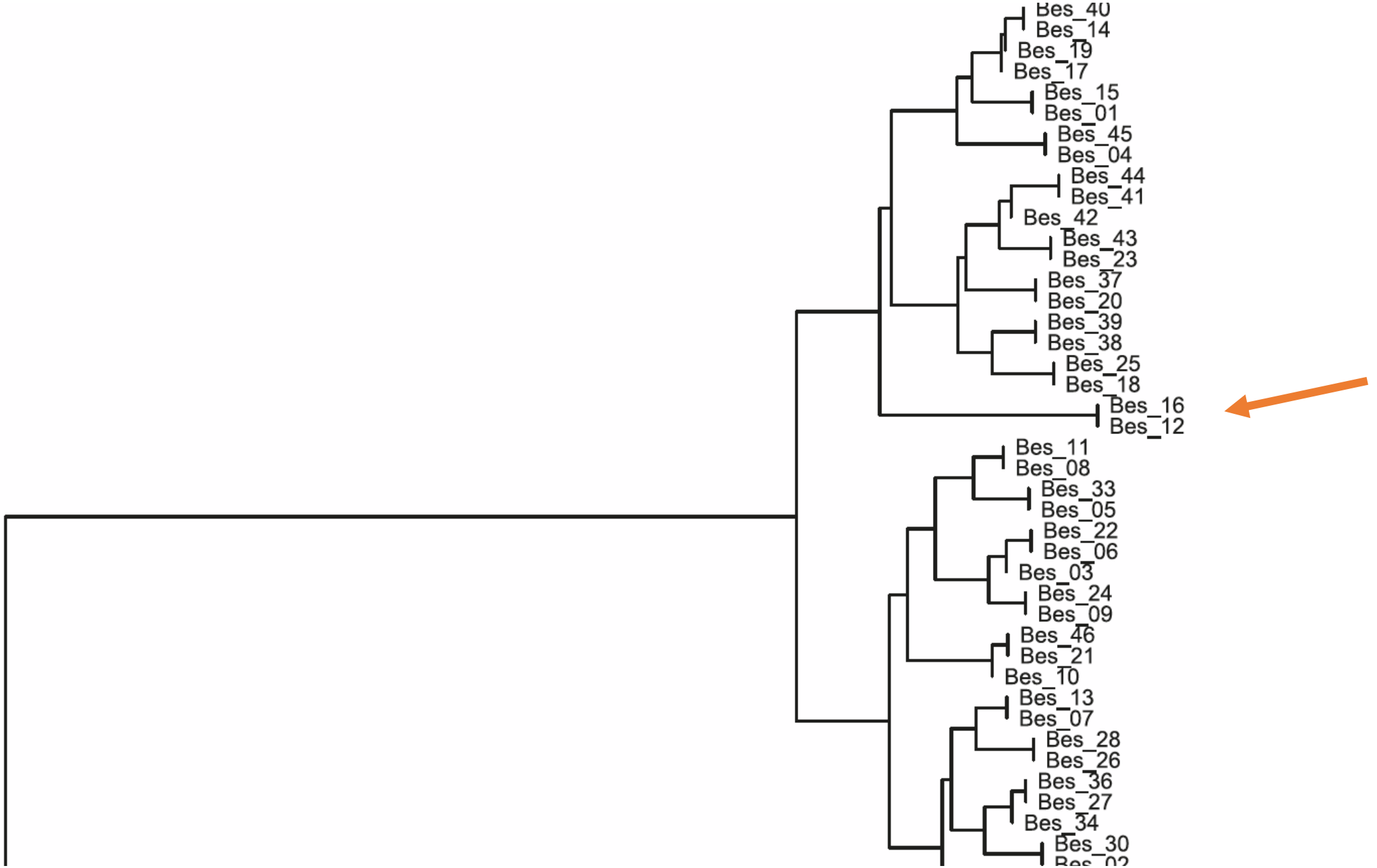
- two basic groups according to MFW
- *VencNik* and *Anast*
  - joint cluster in every analysis (MFW, MATTR, ATL)
  - supports the philological hypothesis of the same origin, perhaps even the same author/school
- *Ben* and *Georg*; *Meth* and *Vit*
  - possible influence of manuscript preservation?


MATTR
0.90
0.85
0.80
0.75
0.70
0.65
ATL
4.6
4.8
5.0
5.2
5.4
Vencnik
Anast
Steph
Apol
Ben
Supr
Georg
Greg
Conf
Trin
Bes_03
Bes_10
Bes_46
Bes_16
Bes_18
Bes_13
Bes_30
Bes_07
Bes_04
Bes_38
Bes_37
Bes_08
Bes_02
Bes_19
Bes_36
Bes_32
Bes_35

C) „Segmented“ *Bes*

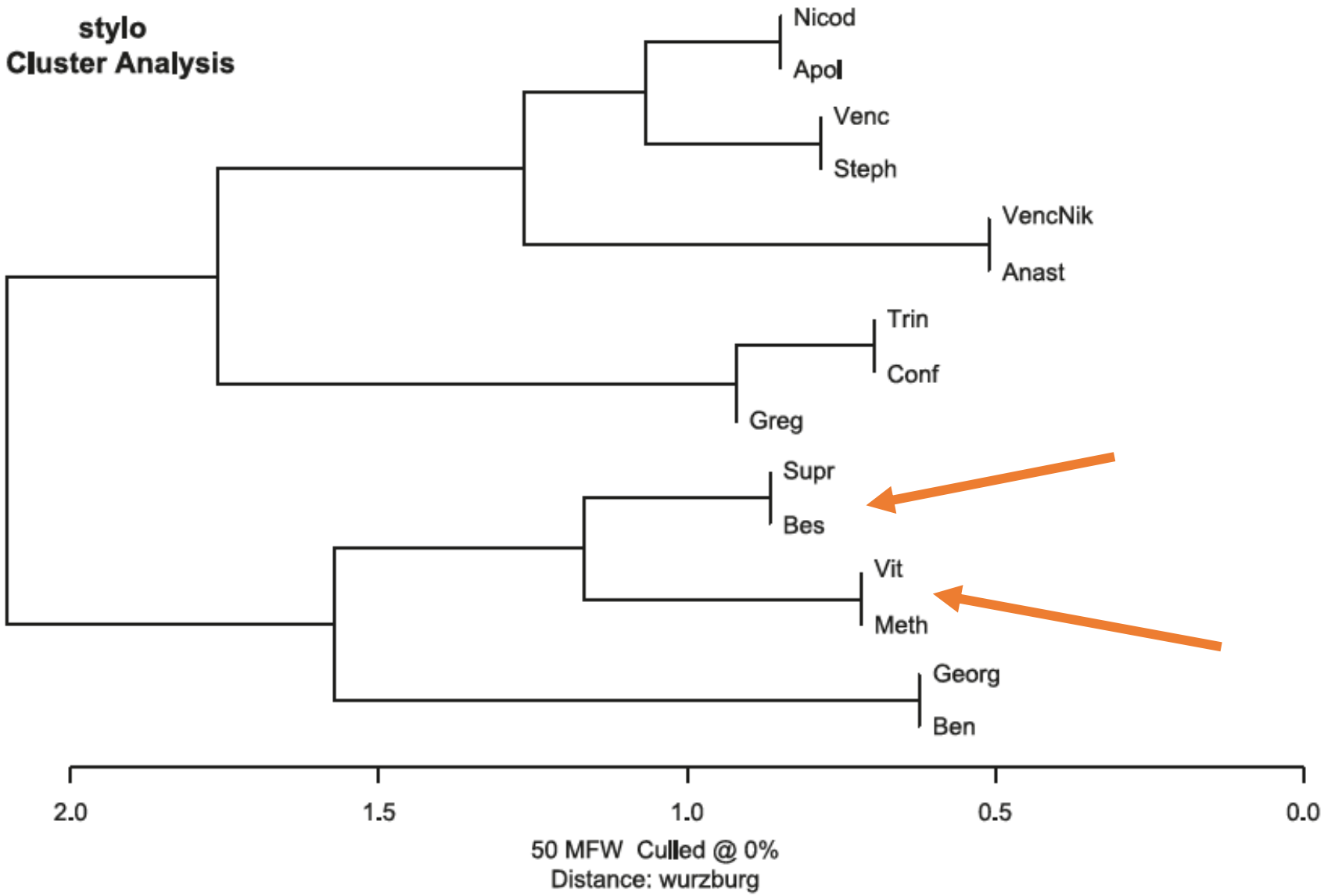
- some passages from a variant manuscript in the edition
  - they are not grouped in any parameter
  - no significant influence of manuscript characteristics!

- one homily (no. 9) twice in the CS translation (not an identical text!) – the same cluster


Bes_40
Bes_14
Bes_19
Bes_17
Bes_15
Bes_01
Bes_45
Bes_04
Bes_44
Bes_41
Bes_42
Bes_43
Bes_23
Bes_37
Bes_20
Bes_39
Bes_38
Bes_25
Bes_18
Bes_16
Bes_12
Bes_11
Bes_08
Bes_33
Bes_05
Bes_22
Bes_06
Bes_03
Bes_24
Bes_09
Bes_46
Bes_21
Bes_10
Bes_13
Bes_07
Bes_28
Bes_26
Bes_36
Bes_27
Bes_34
Bes_30

D) Texts of different origin

- *Supr* and *Meth*
  - methods cluster them mainly according to a genre
- *Supr* and *Bes*
  - both contain homilies
  - however, translations from Greek x Latin
- *Meth* and *Vit*
  - the same manuscript
  - the origin of *Vit* in the early 10th century
  - some scholars even suggested the Great Moravian origin


stylo
Cluster Analysis
Nicod
Apol
Venc
Steph
VencNik
Anast
Trin
Conf
Greg
Supr
Bes
Vit
Meth
Georg
Ben
2.0
1.5
1.0
0.5
0.0
50 MFW Culled @ 0%
Distance: wurzburg

Conclusions

- stylometric methods appear to be plausible and yield relevant results
- genre is the essential categorization factor
- similarities between texts may also reflect a possible common origin
- whether a text is translated or original is not the main criterion for relatedness
- this type of research can support existing hypotheses and highlight new connections between texts

Thank you!