Quotations, Relevance and Time Depth:

Medieval Arabic Literature in Grids & Networks

Petr Zemánek and Jiří Milička

Intertextualita versus idiomy

- Přiznané a nepřiznané citace
- Co s citacemi, které nejsou naznačeny metadaty?

- Kolak and Schilit (2008)
  - GoogleBooks

Tolerantní algoritmus

- *Erbsensuppe ist ein nahrhaftes essen für die Familie.*
- *Erbsensuppe ist für die Familie ein nahrhaftes essen.*
- *Erbsensuppe ist ein nahrhaftes Familie-essen.*
- *Ein nahrhaftes essen für die familie ist Erbsensuppe.*
- *Ein nahrhaftes essen ist Erbsensuppe für die familie .*
- *Für die familie ist Erbsensuppe ein nahrhaftes essen.*

(James Krüss, Mein Urgroßvater und ich)

Winding its way among countless islands, and imbedded in mountains, the "holy lake" extended a dozen leagues still further to the south. With the high plain that there interposed itself to the further passage of the water, commenced a portage of as many miles, which conducted the adventurer to the banks of the Hudson, at a point where, with the usual obstructions of the rapids, or rifts, as they were then termed in the language of the country, the river became navigable to the tide.

Winding its way among countless islands, and imbedded in mountains, the "holy lake" extended a dozen leagues still further to the south. With the high plain that there interposed itself to the further passage of the water, commenced a portage of as many miles, which conducted the adventurer to the banks of the Hudson, at a point where, with the usual obstructions of the rapids, or rifts, as they were then termed in the language of the country, the river became navigable to the tide.

Winding its way among countless islands, and imbedded in mountains

Word	Frequency in the corpus
winding	1
its	342
way	84
among	148
countless	2
Islands	10
and	4545
imbedded	2
in	2623
mountains	28

Word	Frequency in the corpus
winding	1
countless	2
imbedded	2
Islands	10
mountains	28
way	84
among	148
its	342
in	2623
and	4545

Word	Frequency in the corpus
winding	1
countless	2
imbedded	2
Islands	10
mountains	28
way	84
among	148
its	342
in	2623
and	4545

To the north stretched the limpid, and, as it appeared from that dizzy height, the narrow sheet of the "holy lake," indented with numberless bays, embellished by fantastic headlands, and dotted with countless islands. At the distance of a few leagues, the bed of the water became lost among mountains, or was wrapped in the masses of vapor that came slowly rolling along their bosom

✓

To the north stretched the limpid, and, as it appeared from that dizzy height, the narrow sheet of the "holy lake," indented with numberless bays, embellished by fantastic headlands, and dotted with countless islands. At the distance of a few leagues, the bed of the water became lost among mountains, or was wrapped in the masses of vapor that came slowly rolling along their bosom

✓✓

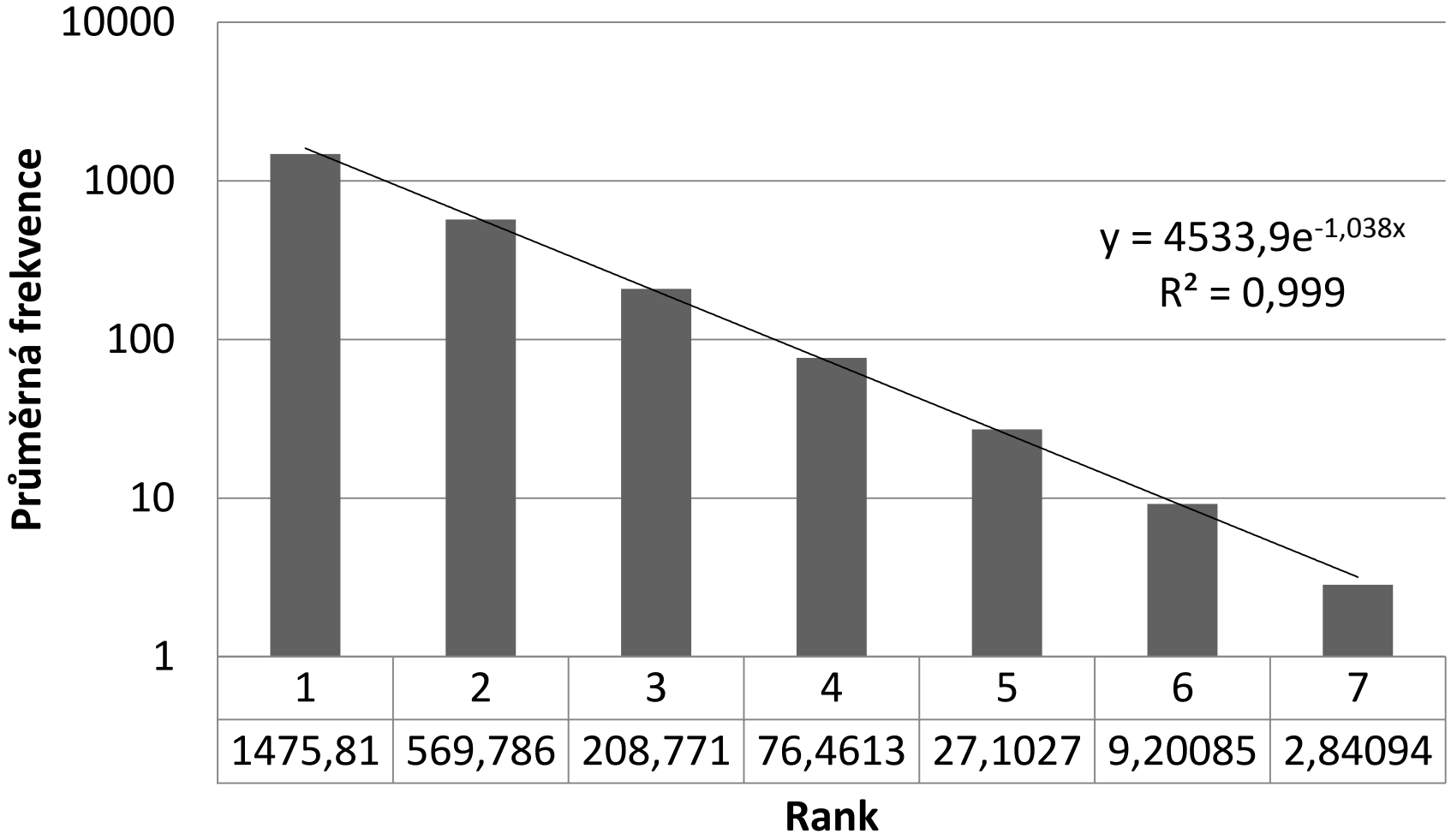
To the north stretched the limpid, and, as it appeared from that dizzy height, the narrow sheet of the "holy lake," indented with numberless bays, embellished by fantastic headlands, and dotted with countless islands. At the distance of a few leagues, the bed of the water became lost among mountains, or was wrapped in the masses of vapor that came slowly rolling along their bosom

✓✓✕✕

To the north stretched the limpid, and, as it appeared from that dizzy height, the narrow sheet of the "holy lake," indented with numberless bays, embellished by fantastic headlands, and dotted with countless islands. At the distance of a few leagues, the bed of the water became lost among mountains, or was wrapped in the masses of vapor that came slowly rolling along their bosom

✓✓✕✕✕✕

Babička Boženy Němcové



Hypertextualizace korpusu

CLAUDia:

Corpus Linguae Arabicae Universalis Diachronicus

420 milionů slov (tokenů)

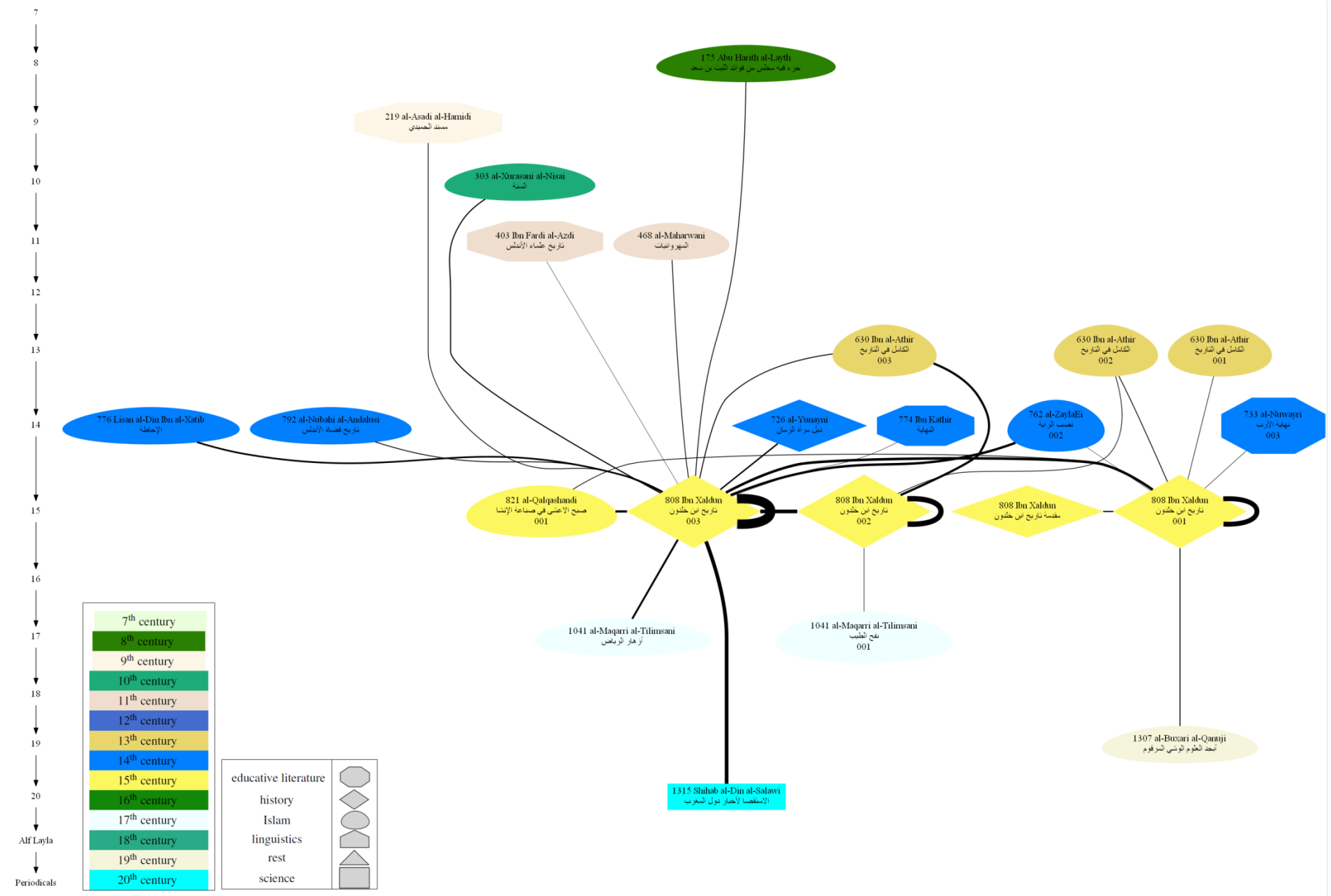
Od 7. do poloviny 20. století

Cca. 2000 děl

Spousta žánrů

$$\Xi_{i,j} = log_2 \frac{h \frac{M_{i,j}}{N_i N_j}}{\sum_{(k,l) \in K} \frac{M_{k,l}}{N_k N_l}}$$

- *i*, *j* ; *k*, *l* … páry textů
- *N* … počet tokenů
- $M_{i,j}$ … počty tokenů, kteé jsou částí citací textu *j* v textu *i*
- *h* … parametr, který si volí uživatel
- *K* … množina všech párů textů v korpusu


7
8
9
10
11
12
13
14
15
16
17
18
19
20
Alf Layla
Periodicals
175 Abu Harith al-Layth
219 al-Asadi al-Hamidi
303 al-Xurasani al-Nisai
403 Ibn Fardi al-Azdi
468 al-Maharwani
630 Ibn al-Athir
630 Ibn al-Athir
630 Ibn al-Athir
776 Lisan al-Din Ibn al-Xatib
792 al-Nubahi al-Andalusi
726 al-Yunayni
774 Ibn Kathir
762 al-ZaylaEi
733 al-Nuwayri
821 al-Qalqashandi
808 Ibn Xaldun
808 Ibn Xaldun
808 Ibn Xaldun
808 Ibn Xaldun
1041 al-Maqarri al-Tilimsani
1041 al-Maqarri al-Tilimsani
1307 al-Buxari al-Qanuji
1315 Shihab al-Din al-Salawi
7th century
8th century
9th century
10th century
11th century
12th century
13th century
14th century
15th century
16th century
17th century
18th century
19th century
20th century
educative literature
history
Islam
linguistics
rest
science

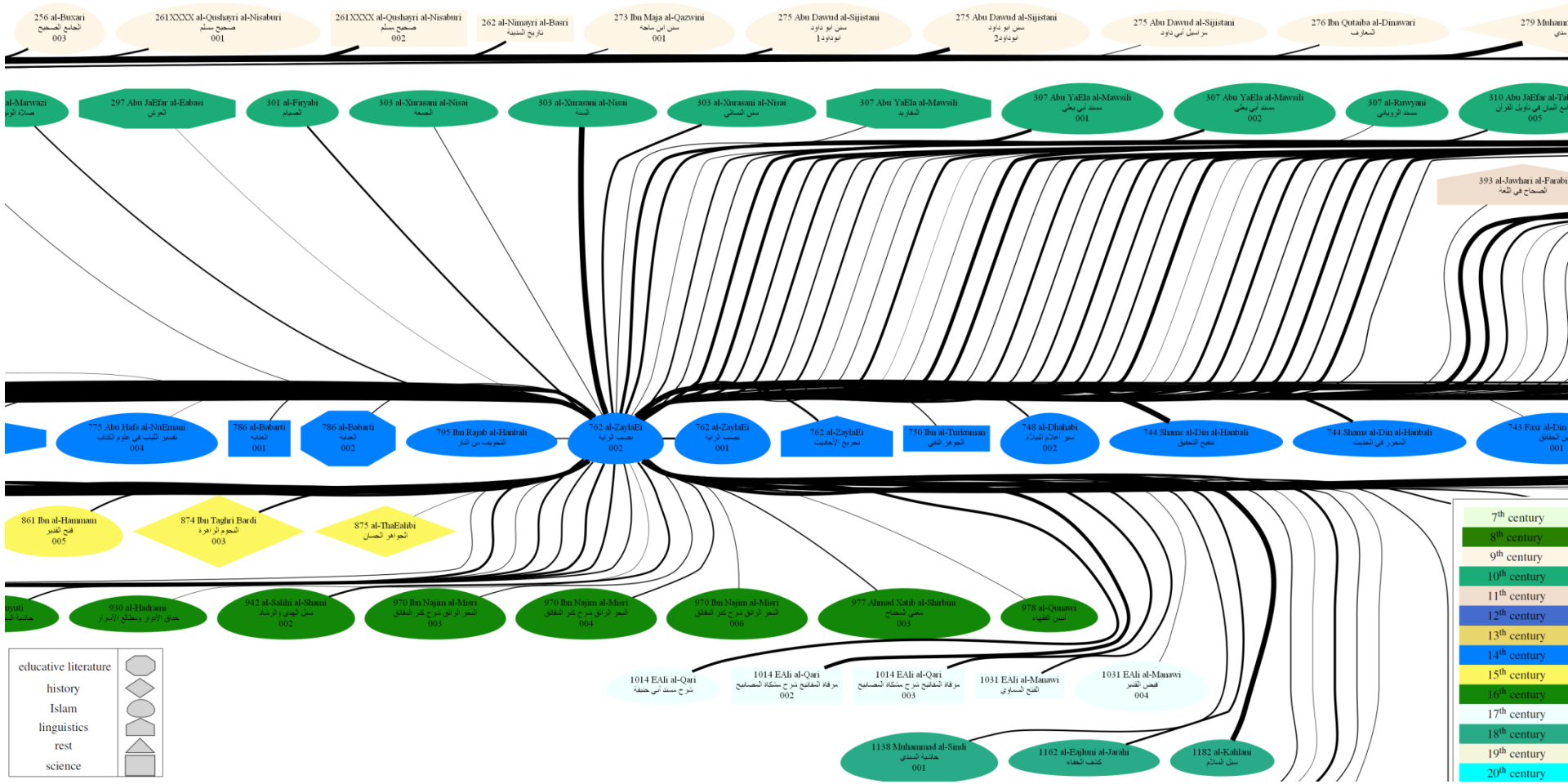
Centralita

$$C_D(i) = \sum_{j \in J} \frac{M_{i,j}}{N_i N_j}$$

- *J* … množina všech textů v korpusu
- *i, j* … text
- *N* … počet tokenů
- $M_{i,j}$ … počet tokenů, které jsou částí citátů textu *i* v textu *j*

Author	Title	Degree Centr.	Cited C_D	Citing C_D	Cited (edges)	Citing (edges)
Az-Zaylai	Nasab ar-Raya II.	0.0958	0.0278	0.0681	70	12
Al-Isbahani	Axbar Isbahan	0.08257	0.0789	0.0036	23	5
Al-Isbahani	Axbar Isbahan	0.07763	0.0001	0.0775	0	2
An-Nasa‘i	Sunna	0.07277	0.0597	0.0130	155	0
An-Nasa‘i	Mir‘at al-Jinan	0.04562	0.0038	0.0418	35	13

Zajla´í


256 al-Buxari
261XXXX al-Qushayri al-Nisaburi
261XXXX al-Qushayri al-Nisaburi
262 al-Nimayri al-Basri
273 Ibn Maja al-Qazwini
275 Abu Dawud al-Sijistani
275 Abu Dawud al-Sijistani
275 Abu Dawud al-Sijistani
276 Ibn Qutaiba al-Dinawari
279 Muhamm
297 Abu JaEfar al-Eabasi
301 al-Firyabi
303 al-Xurasani al-Nisai
303 al-Xurasani al-Nisai
303 al-Xurasani al-Nisai
307 Abu YaEla al-Mawsili
307 Abu YaEla al-Mawsili
307 Abu YaEla al-Mawsili
307 al-Ruwyani
310 Abu JaEfar al-Ta
393 al-Jawhari al-Farabi
775 Abu Hafs al-NuEmani
786 al-Babarti
786 al-Babarti
795 Ibn Rajab al-Hanbali
762 al-ZaylaEi
762 al-ZaylaEi
762 al-ZaylaEi
750 Ibn al-Turkuman
748 al-Dhahabi
744 Shams al-Din al-Hanbali
744 Shams al-Din al-Hanbali
743 Fakr al-Din
861 Ibn al-Hammam
874 Ibn Taghri Bardi
875 al-ThaEalibi
930 al-Hadrami
942 al-Salihi al-Shami
970 Ibn Najim al-Misri
970 Ibn Najim al-Misri
970 Ibn Najim al-Misri
977 Ahmad Xatib al-Shirbini
978 al-Qunawi
1014 EAli al-Qari
1014 EAli al-Qari
1014 EAli al-Qari
1031 EAli al-Manawi
1031 EAli al-Manawi
1138 Muhammad al-Sindi
1162 al-Eajluni al-Jarahi
1182 al-Kahlani
educative literature
history
Islam
linguistics
rest
science
7th century
8th century
9th century
10th century
11th century
12th century
13th century
14th century
15th century
16th century
17th century
18th century
19th century
20th century

Další práce

- experimentování s různými délkami nejkratšího n-gramu a různou mírou povolené variability;
- porovnání komplexity grafů různých subkorpusů selektovaných podle různých kritérií;
- porovnání různých měřítek pro centralitu
- detailní interpretace jednotlivých hran;
- porovnání s jinými korpusy;
- síť autocitací v rámci jednotlivých textů.

Kenneth R. Beesley. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. *ACL Workshop on Arabic Language Processing: Status and Perspective.* Toulouse, France: 1–8.

Carl Brockelmann. 1996. *Geschichte der Arabischen Literatur,* (4 Volume Set). Brill, Leiden (1st edition: 1923).

Tim Buckwalter. 2004. Issues in Arabic Orthography and Morphology Analysis. *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING.* Geneva: 31–34.

William Paulo Ducca Fernandes, Eduardo Motta and Ruy Luiz Milidiú. 2011. Quotation Extraction for Portuguese. *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology.* Cuiabá: 204–208.

Okan Kolak and Bill N. Schilit. 2008. Generating Links by Mining Quotations. *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia.* New York: 117–126.

Kathleen Kuiper. 2010. *Islamic Art, Literature and Culture.* Rosen Publishing Group.

Andrew J. Lane. 2005. *A Traditional Mu'tazilite Qur'an Commentary: The Kashshaf of Jar Allah al-Zamakhsari (d.538/1144).* Brill, Leiden.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Seattle: 989–999.

Bruno Pouliquen, Ralf Steinberger and Clive Best. 2007. Automatic Detection Of Quotations in Multilingual News. *Proceedings of Recent Advances in Natural Language Processing 2007.* Borovets.

Xiaolin Shi, Jure Leskovec and Daniel A. McFarland. 2010. Citing for High Impact. *Proceedings of the 10th annual joint conference on Digital libraries.* New York: 49–58.

Otakar Smrž. 2007. *Functional Arabic Morphology. Formal System and Implementation.* Doctoral Thesis, Charles University, Prague.

Díky za pozornost!

jiri@milicka.cz

petr.zemanek@ff.cuni.cz