Network Visualization for Historical Corpus Linguistics

externally-defined variables as node attributes

Timo Korkiakangas — University of Oslo

Korkiakangas & Lassila [in review]:

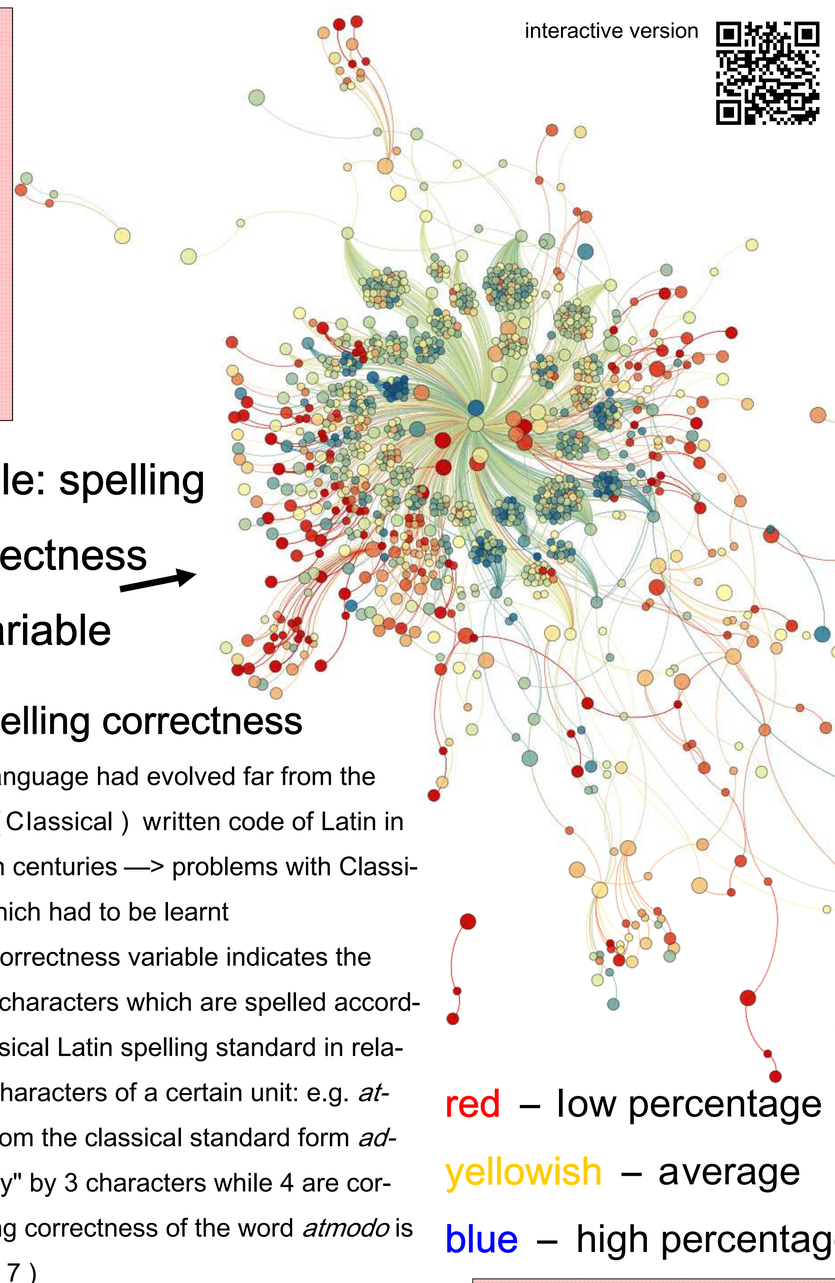
Objectives

1) to find out whether network visualization can support philological and historical-linguistic argumentation in a corpus-based study (does the method have demonstrable advantages compared to ordinary cross-tabulations?)

2) to clarify the scientific premises for the use of network environment to display externally-defined values of linguistic variables

3) to trace early medieval scribes' effective language skills, language attitudes, and stylistic as well as morphosyntactic preferences by setting their writing performances in the geographical, chronological, and social context

interactive version



Implementation

- visualizing connections between linguistic phenomena and extra-linguistic socio-historical entities (document, scribe, writing place, year)
- each node represents an entity and is provided with the values of the linguistic variables, which are derived corpus-linguistically
  - the distributions of these (continuous) linguistic variables are visualized on the network graph as colour transition in Gephi

Data: early medieval Latin documentary texts

- Late Latin Charter Treebank, version 2 (LLCT2): c. 480,000 words, 1,040 Italian (Tuscan) documents from between AD 714-897
- documents (i.e. charters) as a privileged material for examining the spoken/written interface (precise writer, date, and location metadata, no transmission history)
- written on parchment, preserved as originals
- private documents related to buying and selling of landed property
- lemmatic, morphological, syntactic (dependency grammar), and light semantic annotation layers (annotation standard by the Perseus Guidelines)
- TEI P4 XML, Prague Dependency Treebank style

LLCT2 network

- trimodal network, created out of the documents (1,040), scribes (220), and locations (84) underlying the LLCT2 treebank
- unweighted edges
- approximate map background; Gephi's Geo Layout and Force Atlas 2 algorithms

Example: spelling correctness variable

Spelling correctness

- the spoken language had evolved far from the conservative (Classical) written code of Latin in the 8th and 9th centuries —> problems with Classical spelling, which had to be learnt
- the spelling correctness variable indicates the percentage of characters which are spelled according to the Classical Latin spelling standard in relation to all the characters of a certain unit: e.g. *atmodo* differs from the classical standard form *admodum* "greatly" by 3 characters while 4 are correct —> spelling correctness of the word *atmodo* is 57% (i.e. 4 in 7)
- technically, the number of misspelled characters is obtained by calculating the Levenshtein edit distance between each word attested in LLCT2 and the normalized, standard version of that word

Note!

This is no network analysis proper: no network metrics are calculated!

red – low percentage

yellowish – average

blue – high percentage

Theoretical considerations and choices

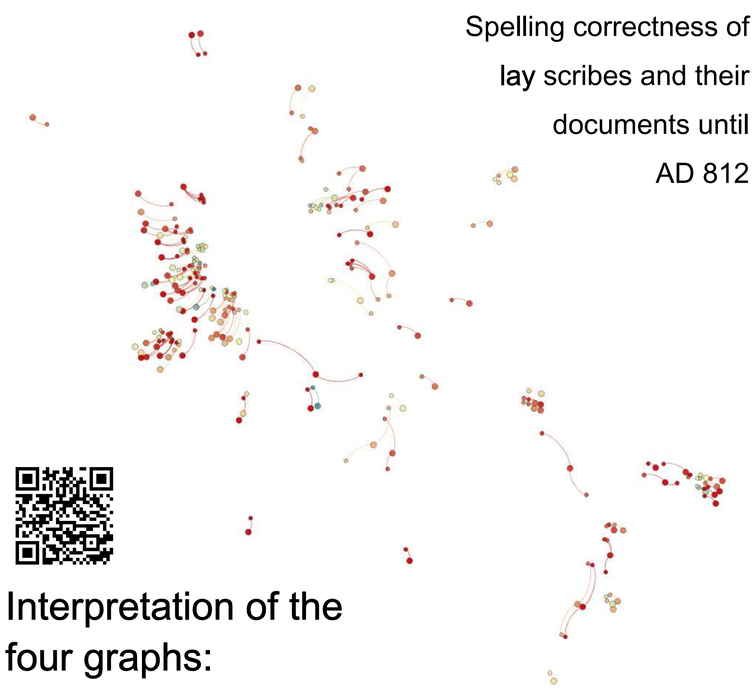
- for research purposes 1) the network visualization must be objective and (as) replicable (as possible), 2) the network graph must be maximally distinctive visually

—> the gradient colour must be rooted in the statistical distribution of the (continuous) linguistic variables

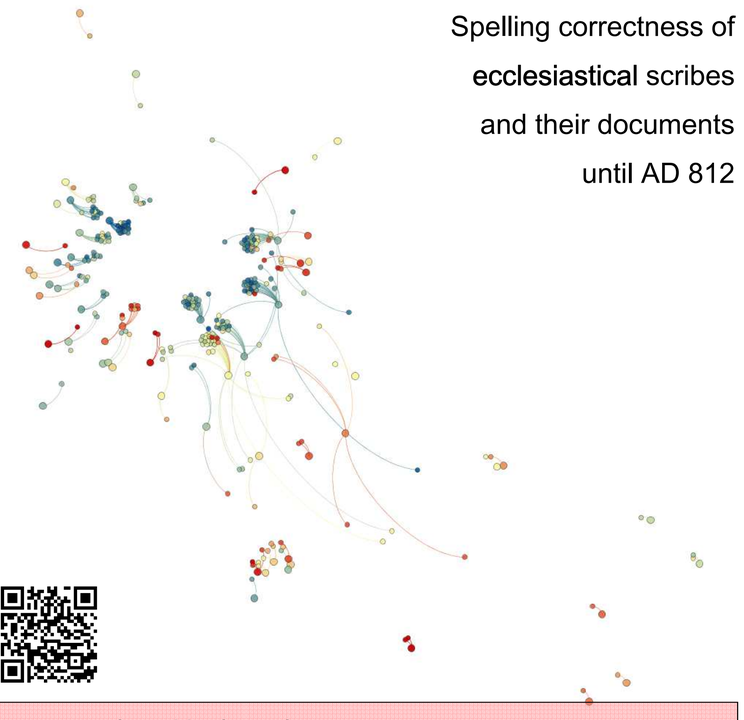
—> the thresholds of the maximal red and maximal blue are set two standard deviations away from the mean, which is marked with maximal yellow

CASE STUDY: Administrative reform and its consequences for spelling correctness?

Spelling correctness of lay scribes and their documents until AD 812

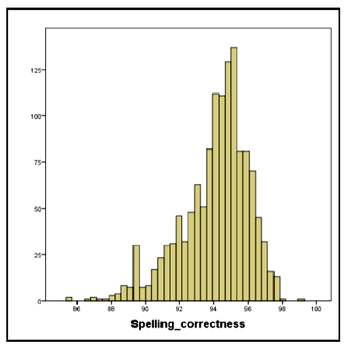


Spelling correctness of ecclesiastical scribes and their documents until AD 812



Interpretation of the four graphs:

- ecclesiastical scribes (clerics) worked in churches, lay scribes were employed by lay rulers
- 812 is a historical watershed: count Bonifatius I practically excluded the ecclesiastical scribes from official document production > this must have had consequences for the quality of documents if the best spellers were ousted suddenly
- the best prolific spellers seem to have been active in Lucca
- after 812 the spelling correctness level of both the lay and of the few remaining ecclesiastical scribes is close to or above the mean <> radical replacement of the officials
- centralization and consolidation of document writing in Tuscia > the role of Lucca as the administrative centre is emphasized

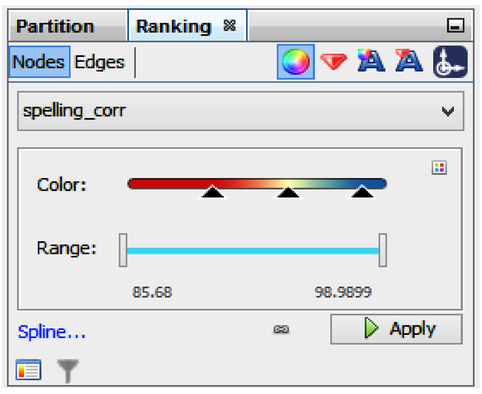


Distribution of the spelling correctness variable

Some statistics

Minimum	85.68
Maximum	98.99
Mean	94.09
Standard Deviation	1.98
Skewness	-0.90
Kurtosis	0.88

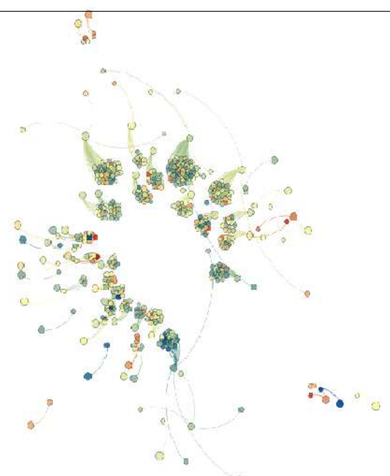
- the Gephi gradient colour palette (right) is adjusted so that the first handle marks the mean minus two standard deviations, (90.1%) the middle handle marks the mean (94.1%), and the third handle marks the mean plus two standard deviations (98.1%)



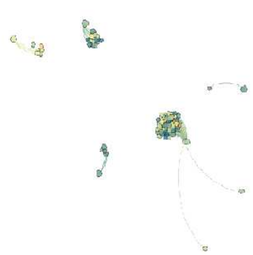
Linguistic variables

- this far, spelling correctness, Classical Latin prepositions, genitive plural form, and <*ae*> diphthong have been analyzed – more to come!
- the visualized linguistic features reflect the language change that had taken/was taking place in the Latin of the 8th and 9th centuries
- the features are operationalized as variables which quantify the variation of those features in the LLCT2 treebank

Spelling correctness of lay scribes and their documents after AD 812



Spelling correctness of ecclesiastical scribes and their documents after AD 812



red – low percentage yellowish – average blue – high percentage

Some biliography

Adams J.N. Social variation and the Latin language. CUP, 2013.

Araújo T. & Banisch S. Multidimensional Analysis of Linguistic Networks. Mehler A., Lücking A., Banisch S., Blanchard P. & Job, B. (eds) *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Springer (Berlin, Heidelberg), 2016, 107-131.

Bergs A. *Social Networks and Historical Sociolinguistics: Studies in Morphosyntactic Variation in the Paston Letters*. Walter de Gruyter (Berlin), 2005.

Korkiakangas T. & Lassila M. Visualizing linguistic variation in a network of Latin documents and scribes. Manuscript submitted to *Journal of Data Mining and Digital Humanities*.