Timo Korkiakangas, University of Oslo, 3 Jan 2018

Network visualization for historical corpus linguistics: externally-defined variables as node attributes

In my poster presentation, I will explore whether and how network visualization can benefit philological and historical-linguistic research. This will be implemented by examining the usability of network visualization for the study of early medieval Latin scribes' language competences. Thus, the scope is mainly methodological, but the proposed methodological choices will be illustrated by applying them to a real data set. Four linguistic variables extracted corpus-linguistically from a treebank will be examined: spelling correctness, classical Latin prepositions, genitive plural form, and <ae> diphthong. All the four are continuous, which is typical of linguistic variables. The variables represent different domains of language competence of the scribes who learnt written Latin practically as a second-language by that time (Korkiakangas 2017, Korkiakangas & Lassila [submitted]). Even more linguistic features will be included in the analysis if my ongoing project proceeds as planned.

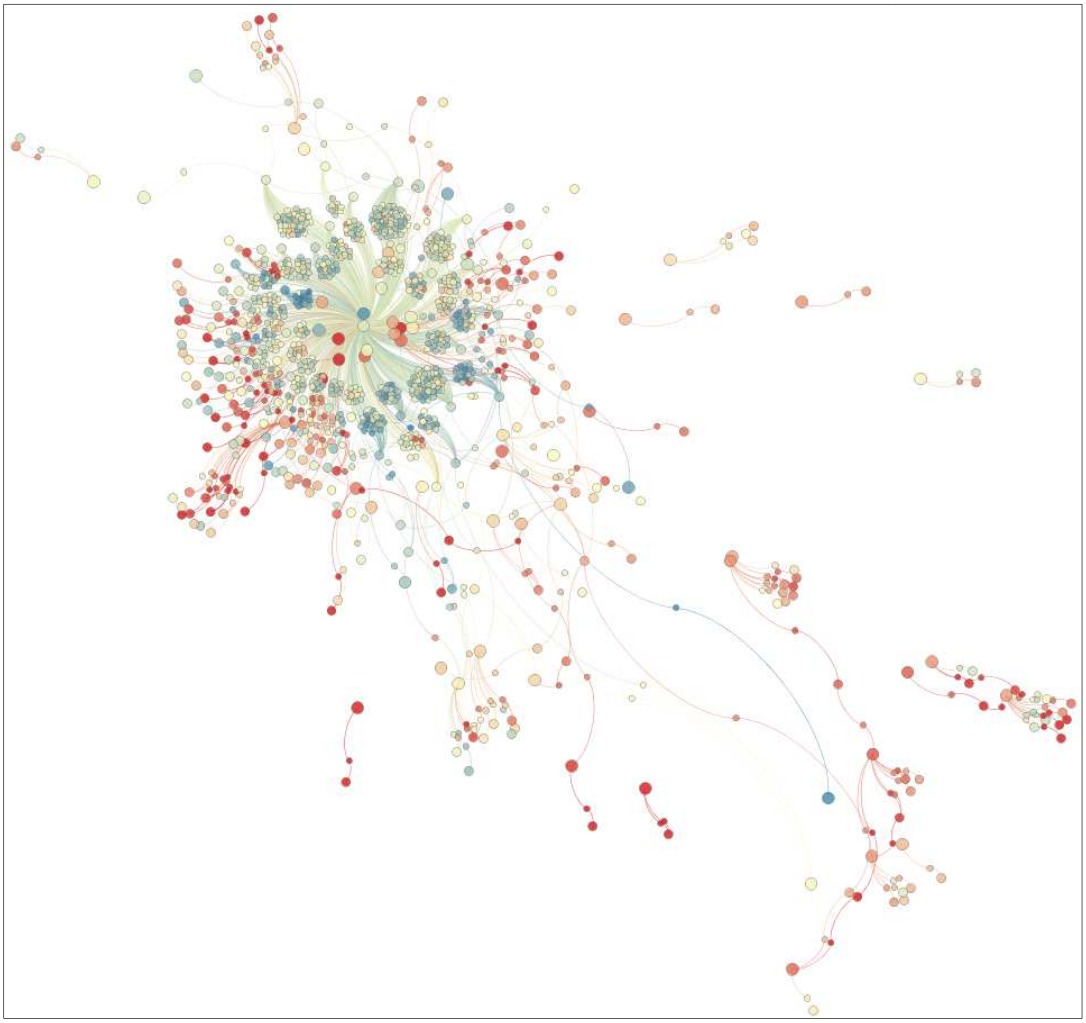
Thus, the primary objective of the study is to find out whether the network visualization approach has demonstrable advantages compared to ordinary cross-tabulations as far as support to philological and historical-linguistic argumentation is concerned. The main means of visualization will be the gradient colour palette in Gephi, a widely used open-source network analysis and visualization software package. As an inevitable part of the described enterprise, it is necessary to clarify the scientific premises for the use of network environment to display externally-defined values of linguistic variables. It is obvious that in order to be utilized for research purposes, network visualization must be as objective and replicable as possible.

By way of definition, I emphasize that the proposed study will not deal with linguistic networks proper, i.e. networks which are directly induced or synthesized from a linguistic data set and represent abstract relations between linguistic units (Araújo & Banisch 2016). Consequently, no network metric will be calculated, even though that might be interesting as such. What will be visualized are the distributions of linguistic variables that do not arise from the network itself, but are derived externally from a medium-sized treebank by exploiting its lemmatic, morphological, and, hopefully, also syntactic annotation layers. These linguistic variables will be visualized as attributes of the nodes in the trimodal "social" network which consists of the documents, persons, and places that underlie the treebank (cf. Bergs 2005). These documents, persons, and places are encoded as metadata in the treebank. The nodes are connected to each other by unweighted edges. The number of document nodes is 1,040, scribe nodes 220, and writing place nodes 84. In most cases, the definition of the 220 writer nodes is straightforward, given that the scribes scrupulously signed what they wrote, with the exception of eight documents. The place nodes are more challenging. Although 78% of the documents has been written in the city of Lucca, the disambiguation and re-grouping of small localities of which little is known was time-consuming and the results not always fully satisfying. The nodes will be set on the map background by utilizing Gephi's Geo Layout and Force Atlas 2 algorithms.

The linguistic features that will be visualized reflect the language change that took place in late Latin and early medieval Latin, roughly the 3rd to 9th centuries AD (Adams 2013). The features are operationalized as variables which quantify the variation of those features in the treebank. This quantification is based on the numerical output of a plethora of corpus-linguistic queries which extract from the treebank all constructions or forms that meet the relevant criteria. The variables indicate the relative frequency of the examined features in each document, scribe, and writing place. For the scribes and writing places, the percentages are calculated by counting the occurrences within all the documents written by that scribe or in that place, respectively.

The resulting linguistic variables are continuous, hence the practicality of the gradient colouring. In order to ground colouring in the statistical dispersion of the variable values and to conserve maximal visual effect, I customize the Gephi default red-yellow-blue palette so that the maximal yellow, which stands for the middle of the colour scale, marks the mean of the distribution of each variable. Likewise, the thresholds of the maximal red and maximal blue are set equally far from the mean. I chose that distance to be two standard deviations away from the mean. In this way, only around 2.5% of the nodes with the lowest and highest values at both ends of the distribution are maximally saturated with red and blue while the rest, around 95%, of the nodes feature a gradient colour, including the maximal yellow in between. Following this rule, I will illustrate the variables both separately and as a sum variable. The images will be available in the poster. The sum variable will be calculated by aggregating the standardized simple variables (cf. Korkiakangas & Lassila [submitted]).

The below image illustrates, by way of an example, the distribution of the spelling correctness variable within the LLCT2 network. The spelling correctness indicates the percentage of characters which are spelled according to the Classical Latin spelling in relation to all the characters of a document. For example, the word form *atmodo* differs from the classical standard form *admodum* "greatly" by three characters. The correct characters are four and, thus, the spelling correctness percentage of the form *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 (Korkiakangas 2017). The red colour stands for a low spelling correctness percentage and blue for a high percentage. The interactive version of the graph with node labels can be consulted at http://bit.ly/2Abk3Bv. That version is realized by SigmaJS tools (http://sigmajs.org/).



The graph shows that the spelling correctness values above the mean are mostly concentrated around one spot, which represents the city of Lucca. Even more importantly, all the substantial blue clusters of high-value documents are written in Lucca, whereas the blue location nodes outside Lucca are due to sporadic high-value documents (and scribes). Conversely, most of the red and reddish low-value nodes are situated outside Lucca, e.g. in Pisa and in peripheral southern and south-western localities. All this elicits the conclusion that classical spelling was cherished primarily in Lucca, the administrative and cultural centre of Tuscia. In sum, the applied distributionally-based principle of gradient colouring seems to be suitable at least for variables which are not too badly divergent from normal distribution. The result is a graph with easily observable colour patterns that are, at the same time, grounded in statistical reality (Korkiakangas & Lassila [submitted]).

The preliminary conclusions also include the observation that network visualization, as such, is not a sufficient basis for philological or historical-linguistic argumentation, but if used along with statistical approach, it can support argumentation by drawing attention to unexpected patterns and – on the other hand – to irregularities. However, it is the geographical layout of the graphs that gives the most of the surplus in regard to traditional approaches: it helps in perceiving patterns that would have otherwise failed to be noticed.

The treebank on which the analyses are based is the Late Latin Charter Treebank (version 2, LLCT2), which consists of 1,040 early medieval Latin documentary texts (c. 480,000 words). The documents have been written in historical Tuscia (Tuscany), Italy, between AD 714 and 897, and are mainly sale or purchase contracts or donations, accompanied by a few judgements as well as lists and memoranda. LLCT2 is still under construction and only the first half of it is already provided with the syntactically annotated layer, thus making it a treebank proper (i.e. LLCT, version 1). The lemmatization and morphological annotation style are based on the Ancient Greek and Latin Dependency Treebank (AGLDT) style which can be deduced from the *Guidelines for the Syntactic Annotation of Latin Treebanks* (Bamman & al. 2007). Korkiakangas & Passarotti (2011) define a number of additions and modifications to these general guidelines which are designed for Classical Latin. For a more detailed description of the LLCT2 and the underlying text editions, see Korkiakangas (2017). Documents are privileged material for examining the spoken/written interface of early medieval Latin, in which the distance between the spoken and written codes had grown considerable by the Late Antiquity. The LLCT2 documents have precise dating and location metadata and they survive as originals.

Bibliography

Adams J.N. Social variation and the Latin language. Cambridge University Press (Cambridge), 2013.

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

Bamman D., Passarotti M., Crane G. and Raynaud S. Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3), 2007 http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf.

Barzel B. and Barabási A.-L. Universality in network dynamics. *Nature Physics*. 2013;9:673-681.

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

Ferrer i Cancho R. Network theory. Hogan P.C. (ed.) *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press (Cambridge), 2010, 555–557.

Korkiakangas T. Spelling Variation in Historical Text Corpora: The Case of Early Medieval Documentary Latin. *Digital Scholarship in the Humanities*, 2017. https://doi.org/10.1093/llc/fqx061

Korkiakangas T. and Lassila M. Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material. Mambrini F., Sporleder C. and Passarotti M. (eds) *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3), Sofia, December 13, 2013. Bulgarian Academy of Sciences (Sofia), 2013, 61-72.

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

Korkiakangas T. and Passarotti M. Challenges in Annotating Medieval Latin Charters. Journal of Language Technology and Computational Linguistics. 2011;26,2:103-114.