Khepri – a Modular View-Based Tool for Exploring (Historical Sociolinguistic) Data

Eetu Mäkelä Tanja Säily Terttu Nevalainen

July 5, 2016

1 Motivation

Digital humanities needs tools that better support the core processes of humanistic inquiry. This includes support for handling uncertainty and incompleteness in the data, for interactive exploration, and for fluidly moving between close and distant reading (Gibbs and Owens 2012; Drucker 2011; Jänicke et al. 2015; Caviglia, Ciuccarelli, and Coleman 2012; Uboldi and Caviglia 2015).

The Khepri tool presented here is part of a project to develop a modular set of components that take these requirements into account, and can be connected and configured to respond to the needs of a particular humanities task and data. Khepri targets data stored as Linked Data (Heath and Bizer 2011), a set of scalable standards that has gained widespread adoption particularly in the sphere of cultural heritage.

2 Development process

To ensure the tools developed meet the needs of humanities users, they are being developed iteratively, utilizing participatory design in relation to case studies, as advocated by the field of design science (Hevner et al. 2004; Peffers et al. 2007; Wieringa 2009). The task of the computer scientist is to see beyond these individual studies; to identify common components allowing the tools to generalize beyond the projects under scrutiny.

To date, a variety of collaborations have been embarked upon, from the prosopographical study of the Republic of Letters[1], through supporting engagement with WW1 primary sources (Mäkelä, Törnroos, et al. 2015), to developing a contextual network for Finnish fiction (Mäkelä, Hypén, and Hyvönen 2013). Together, these span a range of research questions, as well as types of data.

Through these collaborations, a prevalent process of inquiry was identified – the need to explore and contrast differently constrained subsets of a dataset. For instance, this might be looking at the correspondence networks of different individuals and comparing them, or looking at how possible values of a linguistic variable behave with respect to each other as well as associated metadata.

To support this process, Khepri utilizes the view-based paradigm (Mäkelä 2010), where data is presented simultaneously from different perspectives, with each

[1] `http://www.republicofletters.net/`

perspective acting both as a visualization as well as a means to constrain what is shown. A proper implementation of the paradigm also allows for speedy informed variation of parameters, and thus interactive exploration.

Because the views interact in a defined way, they can be developed as separate components targeting major visualization classes such as geographical, temporal or statistical. Each individual Khepri instance can then select from these the views suitable for that particular use.

Thus far, most of the work has been preparatory, with the functionalities simulated through ad-hoc disconnected components, tied together and supplemented by manual work of the computer scientist. However, now a first complete tool for a particular task has been developed. This instance has been configured for historical sociolinguistics.

3 Khepri for historical sociolinguistics

Historical sociolinguistics is the study of language in relation to social factors through time. An example research question would be to chart the role of gender, age and socioeconomic status in the diffusion of the English progressive (as in *I am writing*). From the viewpoint of the Khepri tool, this is interesting because it requires combining access to unstructured text with access to the structured (meta)data describing their authors.

This is also the area where current tools fall short, for while corpus tools (e.g. CQPweb (Hardie 2012), Korp (Borin, Forsberg, and Roxendal 2012) and WordSmith[2]) enable querying texts by linguistic features, they poorly support walking from the texts to the attributes of the authors. On the other hand, tools for visually exploring structured data (e.g. Palladio[3], Europeana4D[4] and RAW[5]) do not support interacting with text corpora.

This makes research currently very labor-intensive. For instance, if one wishes to study the aforementioned progressive, one first searches for instances of *-ing* in the corpus using a corpus tool. The instances are then exported into Excel to analyze them and eliminate false hits such as gerunds (*My favourite hobby is writing*). Next, the number of hits produced by each person is calculated using another sheet that lists the authors by gender, age, socioeconomic status and time period. These numbers are then exported for statistical analysis and visualization. Because the corpus texts, spreadsheets, visualizations and statistical analyses are not connected to each other, the exploration and interpretation of the observations is cumbersome and time-consuming at every stage.

3.1 The User Interface Configuration of Khepri for Historical Sociolinguistics

The Khepri for historical sociolinguistics interface is depicted in Figure 1. The interface is divided into three columns, with the views contained in each having different primary purposes.

[2] `http://www.lexically.net/wordsmith/`
[3] `http://palladio.designhumanities.org/`
[4] `http://www.tinyurl.com/e4d-project`
[5] `http://raw.densitydesign.org/`

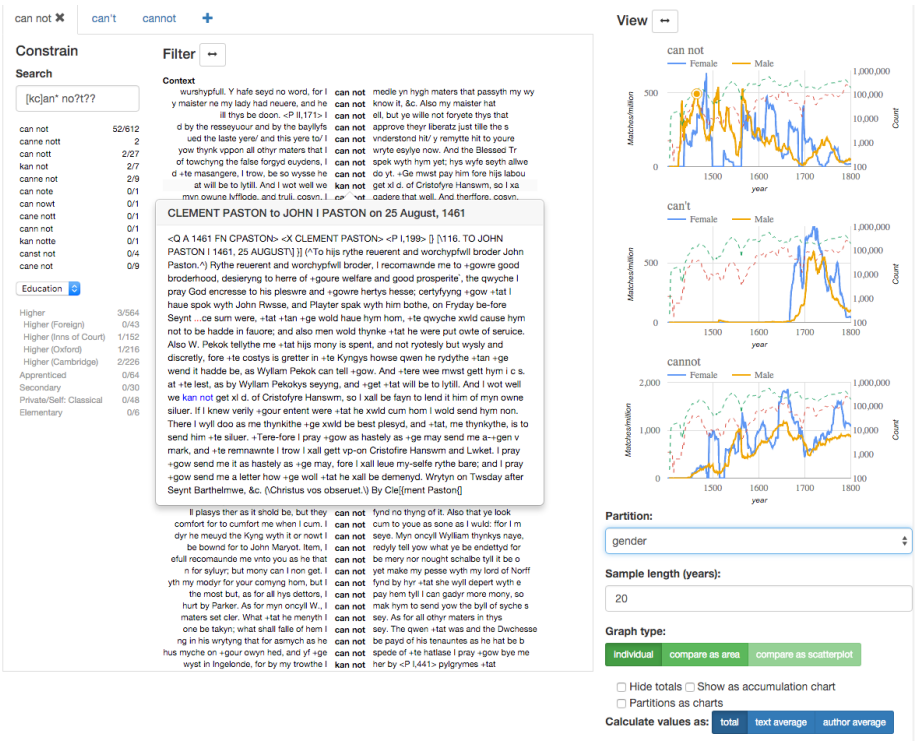

Figure 1: the Khepri for Historical Sociolinguistics interface

On the left are views aimed primarily at producing a subset of interest. The first view is for text search. Below the query, matching keywords from the data are presented for evaluation. Notice that two sets of counts are given. One shows the overall amount of hits for a keyword in the corpus, while the other takes into account constraints set in other windows. This way, the view acts not only as a selector, but also as a statistical breakdown of the current subset.

Below the keyword search view, the user can add metadata views. Here for example, a view visualizes and allows one to constrain the data through the lens of the author's education.

The second column shows the items in the current subset. Matches are shown in their textual context, with metadata and additional context available on mouse-over. While tuned for close reading, this view also acts as a filter. Clicking on an item removes it from the current subset. For linguistic research, this is important as the inclusion or exclusion of a particular example of a phenomenon may depend on contextual hints and background knowledge that cannot be defined as search parameters, but require manual evaluation.

When focusing on close reading, the column can be expanded to occupy the whole right-hand side of the interface. Expanded, the view shows additional metadata, such as the author and year of the texts. The view can also be sorted according to these properties, as well as grouped by them, so that for example only a listing of the authors, or the linguistic types (e.g. different words ending with *-ing*) is shown, with the individual matches revealed by expanding.

To further help in keeping a close reading task organized, the interaction

between this view and the constraining views has been designed so that it is easy to temporarily restrict the matches shown to only those from e.g. a particular spelling, or a particular social class.

Finally, the column on the right is intended primarily for visualization. In fact, it can visualize and contrast multiple subsets of the data. To facilitate this, the first two columns are subsumed in a tabbing container, with each tab containing the query state of a single subset. In the example of Figure 1, these are spelling variants of the negated auxiliary verb *cannot* (open compound, contraction, closed compound).

By default, the frequency of each subset is visualized as its own line chart. However, numerous options affecting this are provided, drawn from best practices in the field (Hinneburg et al. 2007). For example, separate lines can be graphed for each of the values of a particular metadata property. In Figure 1 for example, each chart contains lines for male and female writers, showing that the usage of the form "can not" seems to follow an approximately linear decline for men, but not for women.

To prevent misinterpretations arising from small samples, each graph can be accompanied by a dotted logarithm representing the size of the corpus as a whole for that metadata value. The interface also supports bootstrapping to visualize confidence intervals. As this takes considerable time to calculate, it should only be enabled when a seemingly significant discovery needs verification.

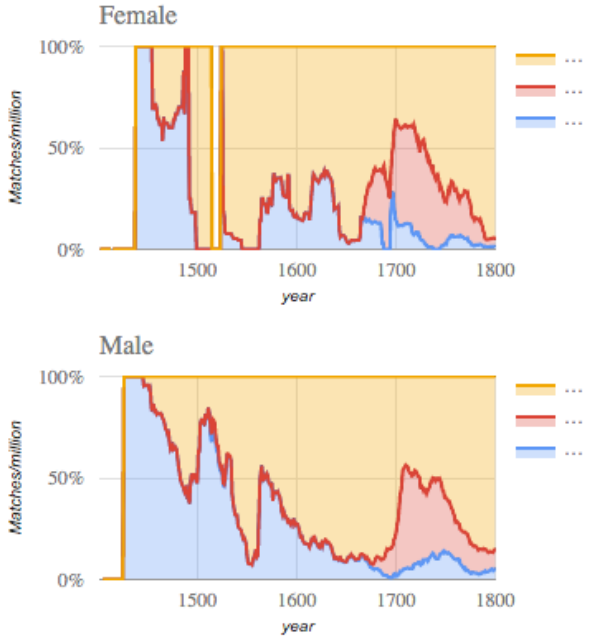

Figure 2: Area charts showing the relative proportions of "can not" (blue), "cannot" (yellow) and "can't" (red) by time and gender.

The interface also offers alternative charts. For example, when comparing possible values of a single linguistic variable, the area chart visualization shown in Figure 2 is appropriate. In addition, a motion chart visualization (Figure

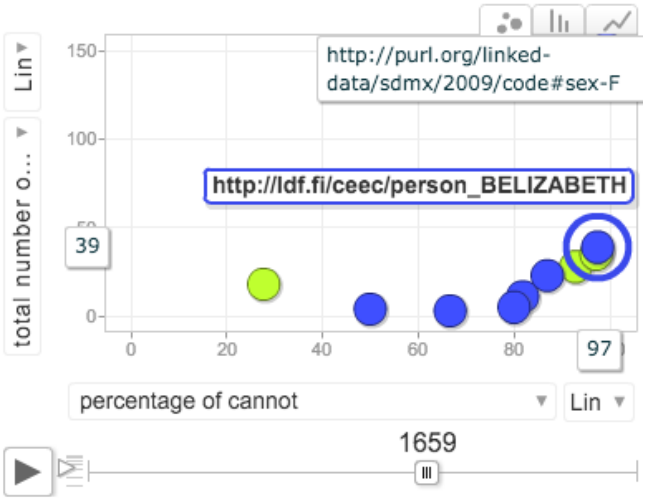

Figure 3: Motion chart showing how many percent of individual actors' use is of the compound form "cannot".

3, inspired by the static scatterplots in Nevalainen, Raumolin-Brunberg, and Mannila (2011)) is provided, used to see how different individuals relate to the variable under study, and even how they change their use through time.

In line with the view-based querying paradigm, all visualizations also act as selectors, enabling delving deeper into interesting phenomena. Through them, one can for example constrain the instance list to show only usage by women in a particular timespan, or in the case of the motion chart, even the use of a single individual.

4 Discussion and future work

Khepri for historical sociolinguistics is the first complete version of the tool. It is also only in its second iteration, so will continue to improve based on feedback. However, already it has been received with excitement, making possible research that was previously too time-consuming to attempt.

With the architecture of the tool now in place, other instances will soon follow, targeting next the Republic of Letters and Finnish fiction use cases. This can be said because all the views created are actually generic, and can be pointed to different data by reconfiguring. For example, text search is also useful for locating individuals or books, while the metadata facets directly target structured data already. The views requiring most modification are the statistical charts, but even here work will be fine-tuning to match differing metrics. Correspondingly, any visualizations developed for other scenarios can be ported here, to for example visualize the language phenomena on maps.

References

Borin, Lars, Markus Forsberg, and Johan Roxendal (2012). "Korp – the corpus infrastructure of Språkbanken". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.

Caviglia, Giorgio, Paolo Ciuccarelli, and Nicole Coleman (2012). "Communication Design and the Digital Humanities". In: *Proceedings of the 4th International Forum of Design as a Process*.

Drucker, Johanna (2011). "Humanities approaches to graphical display". In: *Digital Humanities Quarterly* 5.1, pp. 1–21.

Gibbs, Fred and Trevor Owens (2012). "Building better digital humanities tools: Toward broader audiences and user-centered designs". In: *Digital Humanities Quarterly* 6.2.

Hardie, Andrew (2012). "CQPweb - combining power, flexibility and usability in a corpus analysis tool". In: *International Journal of Corpus Linguistics* 17.3, pp. 380–409. ISSN: 1384-6655. DOI: 10.1075/ijcl.17.3.04har.

Heath, Tom and Christian Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers. DOI: 10.2200/S00334ED1V01Y201102WBE001.

Hevner, Alan R. et al. (2004). "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1, pp. 75–105. ISSN: 02767783.

Hinneburg, Alexander et al. (2007). "How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change". In: *Literary and Linguistic Computing* 22.2, pp. 137–150. DOI: 10.1093/llc/fqm006.

Jänicke, Stefan et al. (2015). "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges". In: *Eurographics Conference on Visualization (EuroVis) - STARs*. Ed. by R. Borgo, F. Ganovelli, and I. Viola. The Eurographics Association. DOI: 10.2312/eurovisstar.20151113.

Mäkelä, Eetu (2010). "View-Based User Interfaces for the Semantic Web". D.Sc. dissertation. PhD thesis. Aalto University, School of Science and Technology, Espoo.

Mäkelä, Eetu, Kaisa Hypén, and Eero Hyvönen (2013). *Fiction Literature as Linked Open Data - the BookSampo Dataset*.

Mäkelä, Eetu, Juha Törnroos, et al. (2015). *World War 1 as Linked Open Data*. Submitted for review.

Nevalainen, Terttu, Helena Raumolin-Brunberg, and Heikki Mannila (2011). "The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change". In: *Language Variation and Change* 23 (01), pp. 1–43. ISSN: 1469-8021. DOI: 10.1017/S0954394510000207.

Peffers, Ken et al. (2007). "A Design Science Research Methodology for Information Systems Research". In: *Journal of Management Information Systems* 24.3, pp. 45–77. ISSN: 07421222.

Uboldi, Giorgio and Giorgio Caviglia (2015). "Information Visualizations and Interfaces in the Humanities". English. In: *New Challenges for Data Design*. Ed. by David Bihanic. Springer London, pp. 207–218. ISBN: 978-1-4471-6595-8. DOI: 10.1007/978-1-4471-6596-5_11.

Wieringa, Roel (2009). "Design science as nested problem solving". In: *Proceedings of the 4th international conference on design science research in information systems and technology*. ACM, p. 8.