Alejandro Benito, Antonio G. Losada, Roberto Therón, Amelie Dorn, Melanie Seltmann, and Eveline Wandl-Vogt. 2016. A spatio-temporal visual analysis tool for historical dictionaries. In Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '16), Francisco José García-Peñalvo (Ed.). ACM, New York, NY, USA, 985-990. DOI: https://doi.org/10.1145/3012430.3012636

A spatio-temporal visual analysis tool for historical dictionaries

Alejandro Benito[1], Antonio Losada[1], Roberto Therón[1], Amelie Dorn[2], Melanie Seltmann[2], and Eveline Wandl-Vogt[2]

[1] *Department of Computer Science and Automation, University of Salamanca, Salamanca, Spain , email: abenito@usal.es, alosada@usal.es,theron@usal.es*

[2] *Austrian Academy of Sciences, Austrian Centre for Digital Humanities, Vienna, Austria, email: amelie.dorn@oeaw.ac.at,melanie.seltmann@oeaw.ac.at, eveline.wandl-vogt@oeaw.ac.at*

Abstract

The *exploreAT!* project aims to give insights into the richness of the German language in the Austrian area through a rich and unique collection of dialect words of the Bavarian dialects recorded during the former Austrian-Hungariann Monarchy period and beyond. Originally collected by means of questionnaires, words were noted in handwriting on individual paper slips, covering topics from nature and food to religious festivities, etc. Once digitized, the full database contains around 3.5 million single data entries with an estimated 200,000 headwords, which requires substantial effort if the analysts want to access specific information from the data set. It should also be noted that the data presents a high heterogeneity in terms of its nature and origin (from questionnaires, collectors, scientists, spoken language, hand written notes, etc.), which calls for the creation of a homogeneous database containing all of the available information.

In this paper we present a tool aimed to improve the comprehension of that massive amount of data through visualization means, thus trying to help in the reach of meaningful conclusions and the acquisition of valuable insights in easy and fast ways. With it, analysts can discover cultural issues and access them through means of language and visualization. This is possible thanks to a multidimensional approach to data analysis based on the use of maps, projections and other visualization artifacts. To reach our goal, a team of experts with different backgrounds worked together trying to close the gap between the Humanities and Computer Sciences fields through the creation of our prototype and its multiple iterations.

1 Introduction

Interdisciplinary approaches to data analysis are widely practiced in Digital Humanities (DH). During recent years, much emphasis has been placed on the devising of sustainable digital infrastructures and new and innovative ways of big data exploration in order to exploit and shed light on a multitude of aspects, thus increasing local and global accessibility.

Living in an age that is increasingly dominated by visual impacts and stimuli, the use of novel methods and tools that allow for visual exploration, grouping, analysis and relational demonstrations of different kinds of data have received heightened attention across different fields, including Digital Humanities [10][3][7][6][12]. Various visualization techniques employed across studies at the DH2014 conference[1] were evaluated by Verbert [10] and showed that graphs, geographic maps and 3D models in that order were among the most popular models. Trees, histograms, word clouds and scatter plots, however, were used less often, as depicted in Figure 1. Studies evaluating the usefulness and usability of visualization techniques in the Digital Humanities field, however, are still scarce.

2 Related Work

It is believed that visualization can be extremely helpful while working in humanities, as it can make arguments relevant to its researchers in easier and faster ways. This calls for a more prominent

[1] DH2014 - Annual Conference of the Alliance of Digital Humanities Organizations, Lausanne, Switzerland.

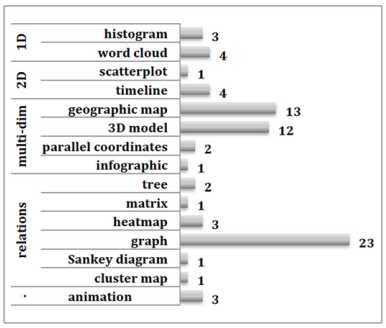

Figure 1: Visualization techniques used by works presented at DH14, as studied by Verbert [10].

research of non text-based approaches to the field of humanities [3].

For text-based materials, however, John et al. [7] developed a method for visualizing the combination of distant and close reading, arguing that a good visualization strategy is vital to understand large amounts of quantitative data. Similarly, Jähnichen et al. [6] devised an interactive visualization of topic models, where the main topics of a text are automatically modeled and visually implemented so users can browse through relations between documents, topics and words, and navigate through the data by concatenating single exploration tasks. Visual analysis techniques have also been used in combination with social network analysis as in the "Early Modern Network Of Networks" (EMNON) presented by Wilson [12] to access, explore and participate in the reconstruction of the social network of scholars working in Europe and America between 1500 and 1750 and thus show how social relationships drove an intellectual change in this period on a global scale.

Bernard et al. [2] introduced a digital library system for time series research data. Based on an overview visualization, the user can delve into large collections of data in an exploratory way while utilizing different views (geographical, calendar-based,...) to reach multiple and different conclusions. Also, Mayer et al. [8] created a visual solution to the CLICS database, which includes information about 200 different languages. It presents usage tendencies for different words with similar definitions in different languages, allowing to analyze temporal evolutions and enabling comparisons between those languages' tendencies.

3 The ??xploreAT! Project

In this paper we introduce and exemplify a novel visual exploration tool, developed and applied in the context of a recent collaborative DH project, *exploreAT!* - exploring Austria's culture through the language glass [11]. In what follows we provide information on the *exploreAT!* project as a general background to the visual analysis tool and take two thematic use cases in Section 5 to validate our tool.

The visual analysis tool presented in this paper is being developed in the context of *exploreAT!*, a DH project which aims to give insights into the richness of the German language in the area of the former Austro-Hungarian Empire [11][5]. *exploreAT!* draws on a rich collection of dialect words of the Bavarian dialects in Austria (Bairische Mundarten in Österreich). Originally collected by means of questionnaires in the first half of the 20th century for the purpose of dictionary making, words were noted in handwriting on individual paper slips. Topics covered range from nature, food, professions or customs to religious beliefs and festivities. Having gone through different stages of digitization, the large database contains around 3.5 million single data entries with an estimated 200,000 headwords. Parts of the collected material have been published in print as a 5 volume dictionary (WBÖ), and as an online database (DBÖ). In the current project, it is envisaged to explore these data in terms of relevant concepts in relation to four specific aspects (infrastructure, elexicography, citizen science and visualization). For this purpose certain thematic groups originating from the questionnaires (colours, plant names and foods) have been chosen and are being developed in greater detail in the project.

exploreAT!, carried out at the Austrian Centre for Digital Humanities (ACDH-ÖAW), is implemented in a collaborative effort together with its partner institutions Universidad de Salamanca (USAL), Dublin City University (DCU) and the Zentrum für Soziale Innovation (ZSI). In this paper

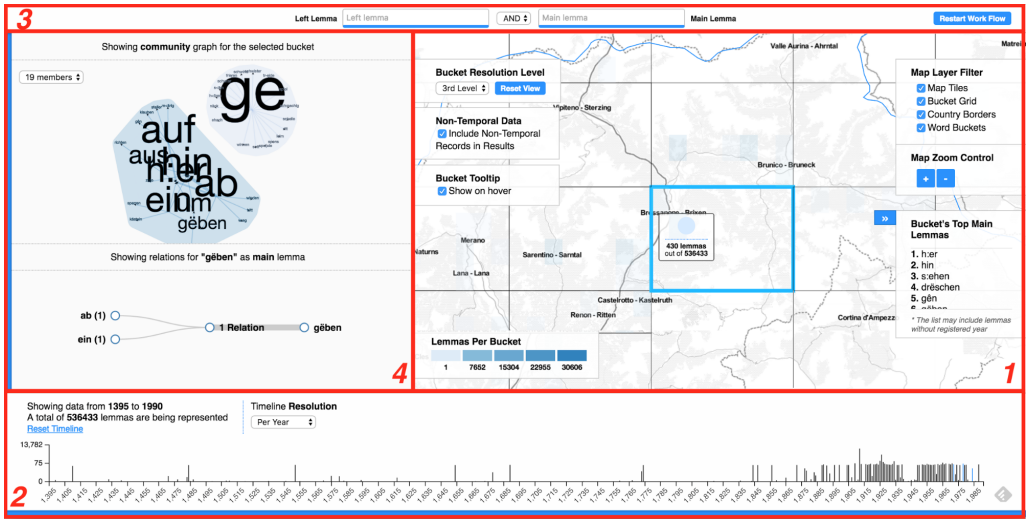

Figure 2: Proposed tool interface: 1) Spatial projection/map. 2) Temporal projection or timeline. 3) Textual search bar. 4) Network analysis view.

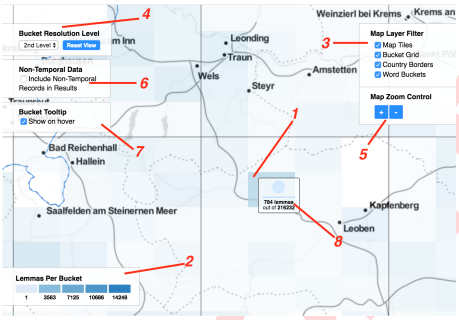

Figure 3: Map view. 1) Geohash bucket. 2) Scale. 3) Layers. 4) Resolution. 5) Zoom. 6) Show/hide data without temporal dimension. 7) Show/Hide bucket information on hover. 8) Bucket information.

we focus on the aspect of visualization and the development of the pilot tool for data exploration.

4 Pilot Visualization Tool

Our functional prototype is a multidimensional, visual analysis tool that allows the exploration of the data mentioned in earlier sections. Despite supporting several dimensions, the analysis workflow is guided by the spatial dimension and the user interface is thus greatly based on the use of maps, projections and other visualization artifacts built on top of those two. In Figure 2 we present a screen shot of the tool, showing all of its views.

The interface depicted in Figure 2 shows four views, of which only three are available from the beginning to the user (the fourth view is shown or, conversely, hidden dynamically depending of the current stage of the analysis the user is at). The proposed visual workflow is based on well-known Shneiderman's visualization mantra: "Overview first, zoom and filter, then details-on-demand" [9].

The application first loads displaying a general view of the data that serves as a starting point for the analysis task (overview first). In our approach we calculate, on-the-fly, general metrics for each of the different geohashes [1] comprising the different geographical areas present in the data. The map, which is the main view, leads the multidimensional analysis of the data and sports the visual artifacts shown in Figure 3.

Each geohash encodes a rectangular portion of terrain: the longer a geohash string is, the smaller is the rectangle it defines in the map. This approach is very convenient for working with search engines that allow fast textual searches like the ones used in this prototype. The tool presents a different colour for each rectangular portion associated with a geohash based on the number of

Figure 4: Usage of different resolutions for the retrieved map data.

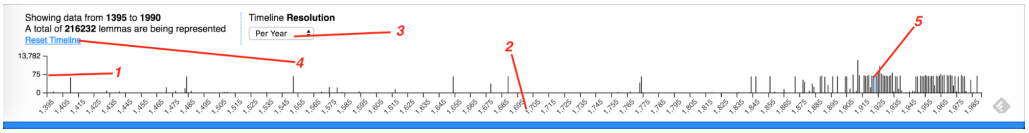

Figure 5: Detail of the timeline view. 1) Scale. 2) Representation of the spatial dimension. 3) Resolution control. 4) Explanatory text and reset control. 5) Bars and highlighting.

results found, according to the scale presented in Figure 3.2, which changes dynamically.

Several layers present different information on the map. These are the following:

- Map Tiles: It holds the images shown in the map, which contain information about the terrain, municipalities, etc. This layer is basic to give a proper contextualization of the information presented in other layers.
- Bucket Grid: Displays the contours of the geohashes of the immediately superior resolution level to the one currently selected. It allows quick identification of the parent geohash and enhances the navigation through different resolution levels.
- Country Borders: An extra layer to represent political borders between countries is used as well. Given that the analyzed data comes from historical dictionaries, this layer helps researchers to identify the current country a region belongs to.
- Word Buckets: Lastly, the buckets/geohash layer presents information in the way previously explained, employing a varying colour scale.

All these layers can be hidden and shown upon user's request. This kind of interaction allows the removal of unnecessary information which could not be useful for the current analysis stage in which the researcher is at through the controls shown in Figure 3.3.

In Figure 4 the same area of the map is presented, showing data at different resolution levels, which are controlled by the interface element depicted in Figure 3.4. In a similar way, the zoom level can also be changed, producing a projection change in the map for this effect.

Given that the analyzed data set is divided in different subsets, each one containing a combination of dimensions, we decided to initially present in the map the subset that contains both spatial and temporal information and to project the subset that contains temporal information (even if some part of it does not have spatial dimension) in the timeline. However we could not ignore the fact that there are subsets that do not contain any of those dimensions but that could hold important information for the analysis task. This is the reason why we enable the inclusion of this data by means of the control shown in Figure 3.7. Lastly, we show the bucket information view (Figure 3.8) when the user hovers the mouse pointer over one of the buckets shown in the correspondent layer, following a details-on-demand approach. This little view exposes the exact number of matches for a given search that fall into the selected bucket and the percentage of the total this number accounts for, providing contextual information.

In an analogous way to the spatial projection of the data shown in the map view, we implemented the already mentioned timeline projection (Figure 5) which represents the data using the temporal information associated to it (if available).

The functionality of this view, which is basically a histogram, is simple yet powerful. In the x-axis the temporal dimension is presented, whereas in the y-axis we present the number of occurrences for a certain year, according to a certain scale (this last one highlighted in Figure 5.1). The histogram dynamically changes its scale depending of the analyzed data by calculating its maximum, mean and minimum values and redrawing the chart. The act of selecting a portion of

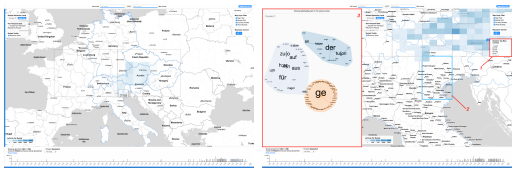

Figure 6: Exploration process. The user starts working with the base view and once he selects a bucket information changes on screen. 1) Words contained in the selected bucket. 2) The selected bucket is highlighted so context is not lost after the map is reprojected. 3) The SNA view appears from the left once a bucket is selected, plotting the words related to it and their relations.

the timeline (known as brushing) is also implemented, allowing to modify the information depicted in the map's buckets or rectangular regions. On top of that, when the user hovers over a bucket in the map view the related bars are highlighted in the histogram. More functionality was added, like the one allowing resolution changes in the same fashion the map view does. Also a summary view, which presents the minimum and maximum years of the results and the total amount of hits fetched from the database was included. Lastly, a reset UI control allows analysts to clear the current selection of a portion of the timeline, resetting it and the map view.

In a second stage of the workflow, user interaction is expected. It is then when the user would fix his mental state in the application by means of the UI controls already discussed, whereas it is by zooming or further filtering the data, producing more smaller slices that allow a deeper level of visual detail and analysis. The prototype supports three types of filtering:

- Spatial: By means of the interactive elements presented in the map view.
- Temporal: By using the controls in the map.
- Textual: Which allows complex searches over the original text data.

In our proposed workflow, we expect the user to combine these 3 kinds of filtering until he is able to obtain a sufficiently small portion of the data that allows adequate knowledge extraction.

Spatial filtering is done mainly by means of interacting with the spatial buckets presented in the correspondent layer of the map. When the user selects and clicks one bucket (Figure 6.2), a series of animations and actions are triggered and its area is highlighted: First, a zoom projection is applied over the geographical zone delimited by the selected bucket. Then the map resolution changes to adapt to this new level of zoom, as more level of detail is acceptable now. Following this, the search engine returns a new result set according to this new selection and resolution level. When information is retrieved, the SNA view (Figure 6.3) comes out the left of the map view and at last the user's current mental state is represented in the interface in the same fashion we explained in previous sections. This interaction workflow is to be repeated indefinitely, most probably combined with the other two types of filtering in order to get to different results.

5 Use Cases

5.1 Bread Types and Naming

In the first use case we will have a look at different pastries for All Saints' Day in Austria and their distribution. It fits into the subproject *explore.bread.AT!* which researches different bread types and their naming. As a staple food, bread plays an essential role in everyday life and shows a great variety in occurrence. Particularly in Austria, a staunchly Catholic nation, there is a wide variety in bread for feast days as All Saints' Day. For our research we query on "*aller-h-eiligen*" (All Saints) as a left lemma and include non-temporal data. As there is not only bread data in the database, we have to manually exclude the entries which do not refer to any bread or pastry type. It is notable that over the whole data set the distribution of the right lemmas for the left lemma "*aller-h-eiligen*" reflect the distribution of the lexicon: there is a small amount of lemmas which occur very often but a wide amount of lemmas which only occur once or twice. There are 12 different lemmas for bread which occur with *aller-h-eiligen* but three of them are very similar to

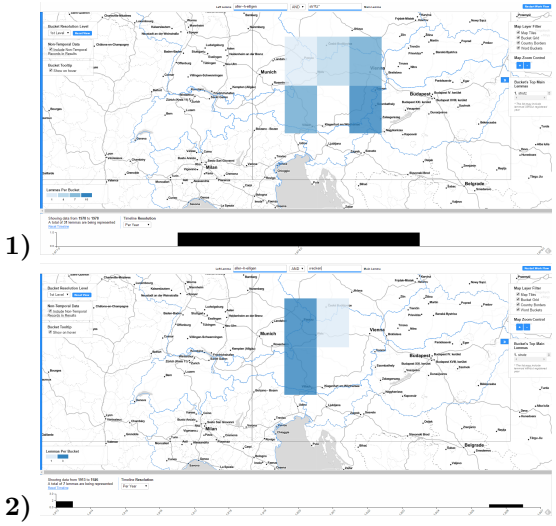

Figure 7: 1) Geographical distribution of *(aller-h-eiligen)strutz*, *-strützel* and *–strutzen*. 2) Geographical distribution of *(aller-h-eiligen)wecken*.

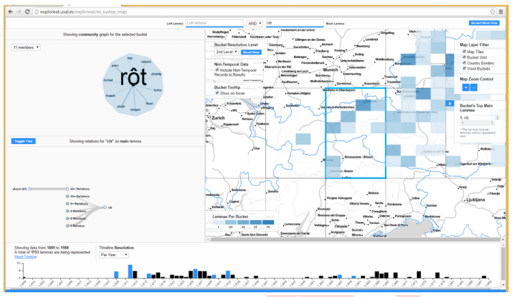

Figure 8: Visual output of the colour term red (*rôt*) in compound words and its possible referents.

each other: *strutz*, *strützel* with umlaut *ü* and diminutive *-el* and *strutzen* with suffix *-en*. *Strutz* (a long plaited bun) is the lemma which occurs the most with 21 of 50 hits. If we combine *strutz*, *strützel* and *strutzen* there are up to 31 hits. Only the lemma *wecken* occurs more than twice, with 7 hits. Within the several buckets the different distributions distinguish from each other. The different forms of *strutz* can be found in 4 of 7 buckets as the bucket's top main lemma (Figure 7). In two buckets wecken is the top main lemma (Figure 8). Also, only in two buckets neither *strutz* nor *wecken* do occur.

5.2 colour Term Usage

Another use case that exemplifies the application of our visualization tool within *exploreAT!* is the one related to the usage of colours through the language. colours form an integral part of our lives and are also a prominent topic spanning across several academic disciplines. Moreover, colour concepts play an important role in the representation of our cultural knowledge [4].

Here we apply our pilot visualization tool to colour terms in the database. More specifically, we look at red (*rot*) in compound words, e.g., *weinrot* (wine red), *blutrot* (blood red), in order to determine their most typical referent. As a first step, the colour term *rôt* was manually entered in the "right lemma" box, while the "left lemma" box was intentionally left blank. Non-temporal data was selected to be included in the results. Then, a community graph and relations plot was generated.

Figure 8 shows the visual output for "red" (*rôt*) and its referents as found in our data. The community graph shows that the colour red is linked to a variety of different concepts. The relations plot revealed that most compounds start with "*blut / plu{ot*" (blood), "*plu{otrôt*" (blood red), or "*brenn / prinn*" (burning), *prinnrôt* (burning red), and this finding holds across different areas (buckets). Other results include "*glos / glut / feuer / blutig / fuchs / ...*" (glow/smolder/fire/bloody/fox/...)

to varying degrees across different regions.

Hence by means of our pilot visualization tool we could demonstrate that the term "red" is most typically associated with concepts of blood (*blut*) and fire-related terms (*prinn / glos / glut*), with a variety of other connected terms that vary in numbers and distribution across areas.

6 Conclusions

Both Humanities and Computer Science disciplines been developing new ideas during their existence and have merged multiple times to solve different problems. Still, there is a much needed effort to improve the relation between these two fields in order to cover some issues not properly exploited yet.

This is not an easy task due to the distance that exists between the two scopes. Although we got to build a fully functional tool based on collaborative work, there is a noticeable gap between humanists and computer scientists' ways of thinking that needs to narrow to allow the achievement of better results. This comes from the approaches each of those groups have taken to complete their tasks during the evolution of their working fields over the time, which should turn into a more blend situation in the future.

In this paper we have presented a prototype aimed to visually explore linguistic data through several dimensions by using multiple methods already employed for text documents analysis such as network analysis, natural language processing and geographical representations (this being the main pillar of the system due to the data structure). By following an iterative development we were able to evolve from simple prototypes to the final version of the tool, thus getting better results.

With Humanities being a field accustomed to classical data analysis approaches, tools as the one proposed here provide solutions to well known problems in a much more intuitive way thanks to the employment of visual methods. By combining the efforts of researchers from the fields involved in the *exploreAt!* project, and adding knowledge from the society, it is to expect that prototypes as ours ease the comprehension of huge data sets in much faster and easier ways than those used nowadays.

7 Discussion and Future Outlook

As already mentioned, there is a high contrast between Humanities and Computer Sciences as of now. As the gap between these disciplines becomes smaller, better outcomes are to be expected. Our prototype aims to start building this bridge but still has multiple angles to cover that were not explored during this first collaborative approach.

One of the ideas for future developments goes through the implementation of a system that takes care of the uncertainty of the data in terms of, for example, the date words were recorded for the first time or the place they appeared at. This, combined with the inclusion of fuzzy searches, would allow the analysts to have more information at hand which presented in the correct way would in turn make the visualizations and ultimately the whole system more powerful while looking for certain answers.

Another feature of the prototype that could be further improved is the force directed graph that highlights relations between words. In zones where there are a lot of words recorded, the reading of the visualization outcome becomes an impossible task, so tuning the creation of the communities represented in it to make them more compact would ease the comprehension while analyzing the data. This, though, would require an effort on the side of the linguists and humanities researchers as they are the ones who truly know where the real value of the relations between words are, or how they could be better grouped and represented in the graph.

Finally, we're looking at the possibility of combining the digital approach of our prototype with a more classical approach in which copies of the original word manuscripts (the paper slips where words were originally described and registered) would be accessible from our tool. This, combined with the inclusion of citizen science-driven features (such as the possibility of completing missing fields of the records stored in the database) would provide a system useful not only for the researchers involved in deep linguistic and historical studies, but also for any user with access to the tool.

NOTICE

The images showcased in this paper can be found in high resolution at https://goo.gl/6OCfcB for a better and clearer viewing.

References

[1] Gustavo Niemeyer. Geohash. http://geohash.org. Accessed: 2016-09-10.

[2] J. Bernard, D. Daberkow, D. Fellner, K. Fischer, O. Koepler, J. Kohlhammer, M. Runnwerth, T. Ruppert, T. Schreck, and I. Sens. Visinfo: a digital library system for time series research data based on exploratory search—a user-centered design approach. *International Journal on Digital Libraries*, 16(1):37–59, 2015.

[3] E. Champion. Seeing is revealing: A critical discussion on visualisation and the digital humanities. In *Digital Humanities 2015*, June-July, Sydney, Australia.

[4] G. Deutscher. *Through the Language Glass: Why the World Looks Different in Other Languages*. Henry Holt and Company, 2010.

[5] A. Dorn, E. Wandl-Vogt, J. Bowers, B. Piringer, and M. Seltmann. exploreat! – perspectives of exploring a dialect language resource in a framework of european digital infrastructures. In *1st International Congress on Sociolinguistics*, September, Budapest, Hungary 2016.

[6] P. Jähnichen, P. Oesterling, G. Heyer, T. Liebmann, G. Scheuermann, and C. Kuras. Exploratory search through interactive visualization of topic models. In *Digital Humanities 2015*, June-July, Sydney, Australia.

[7] M. John, S. Koch, F. Heimerl, A. Müller, T. Ertl, and J. Kuhn. Interactive visual analysis of german poetics. In *Digital Humanities 2015*, June-July, Sydney, Australia.

[8] T. Mayer, J.-M. List, A. Terhalle, and M. Urban. An interactive visualization of crosslinguistic colexification patterns. *09: 00–10: 30–Morning Session, Part I 09: 00–09: 10–Introduction 09: 10–09: 40 Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, An Interactive Visualization of Crosslinguistic Colexification Patterns*, 11(15):1.

[9] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[10] K. Verbert. On the use of visualization for the digital humanities. In *Digital Humanities 2015*, June-July, Sydney, Australia.

[11] E. Wandl-Vogt, B. Kieslinger, A. O'Connor, and R. Therón. exploreat! - perspektiven einer transformation am beispiel eines lexikographischen jahrhundertprojekts. In *DHd (Digital Humanities Im Deutschsprachigen Raum) 2015*, February, Graz, Austria.

[12] E. A. Wilson. Building the early modern digital university: Using social network analysis and digital visualization tools to bring the early modern network. In *Digital Humanities 2015*, June-July, Sydney, Australia.