Exploring Lemma Interconnections in Historical Dictionaries

Alejandro Benito[*]
Departamento de
Informática y Automática
Universidad de Salamanca

Roberto Therón [†]
Departamento de
Informática y Automática
Universidad de Salamanca

Antonio Losada [‡]
Departamento de
Informática y Automática
Universidad de Salamanca

Eveline Wandl-Vogt[§]
Austrian Centre for Digital
Humanities
Austrian Academy of
Sciences

Amelie Dorn[¶]
Austrian Centre for Digital
Humanities
Austrian Academy of
Sciences

Abstract

In this paper we discuss the goals, motivations and other particularities of a visual exploratory analysis tool for historical dictionaries of the Bavarian dialects in Austria. As an input data set we employ the digitized version of the Historical Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich or WBÖ), an initiative started in 1963 which compiles more than five million paper slips collected during the years 1911-1998 in different areas of current Austria, the Czech Republic, Hungary and northern Italy. In the 1990s these paper slips started to be progressively digitized, becoming part of the Database of Bavarian Dialects (DBÖ) and in 2010 nearly 32.000 records related to plant names were made available online on the DBÖ electronically-mapped (dbo@ema) platform[1]. In more recent efforts, the project *exploreAT!: Exploring Austria's Culture Through the Language Glass*, which this work is part of, started in 2015. Its aim is "to reveal unique insights into the rich texture of the German Language, especially in Austria, by providing state of the art tools for exploring the unique collection (1911-1998) of the Bavarian Dialects in the region of the Austro-Hungarian Empire", and it has originated other publications in conference proceedings that we encourage the reader to review [8, 19].

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—User-centered design

1 Introduction

Among the many definitions for the concept of a dictionary provided by several authors over the years, one was taken under special consideration in our research: "A lexicographical product which shows inter-relationships among the data" [18]. Consequently, one of the goals of our tool is to serve the purposes of a variety of scholars and also the general public curious to explore such interrelationships found in the historical dictionaries under study in an easy and fun way. In order to fulfill this initial requirement, we designed our tool on top of a set of well-defined computational tasks: **Spatio-temporal analysis**, **fast full-text search** and **network analysis (NA)**, which are exposed to the final user by means of data visualization. All these computational tasks serve the ultimate purpose of browsing, exposing, projecting and exploring these interrelationships by applying well-known data visualization techniques. Thus, the role of data visualization is specially important in our approach, given the profile of such final users, which is expected not to be necessarily technical or academic. This means final users might not know -or even want to know- the inner workings of any of the aforementioned computational techniques, but still want to benefit from the advantages they can bring to their activity. Data visualization helped to reduce the cognitive load involved in the exploration tasks and also in lowering the user's level of digital mastery needed to operate the proposed tool resulting from our research.

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 research of non text-based approaches to the field of humanities [5].

Directly linked to the topic of this paper, we evaluated one of the first attempts to create a digital edition of a dictionary that could be explored visually. At the core of this work resides an ad-hoc search engine with potent natural language processing and string search capabilities that allow final users to launch fuzzy searches in order to detect mispellings and alternative spellings of headwords [16]. In our study we place searching capabilities at the core of the architecture that supports the proposed pilot tool.

For text-based materials, John et al. [13] 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. [12] 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 [20] 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.

In recent work, Bernard et al. [3] 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. [17] 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. The main aim is simple yet innovative: to identify tendencies in the usage of certain works referring to the same concepts in different languages across time. As a

[*] e-mail: abenito@usal.es
[†] e-mail:theron@usal.es
[‡] e-mail:alosada@usal.es
[§] e-mail:eveline.wandl-vogt@oeaw.ac.at
[¶] e-mail:amelie.dorn@oeaw.ac.at

[1] https://wboe.oeaw.ac.at

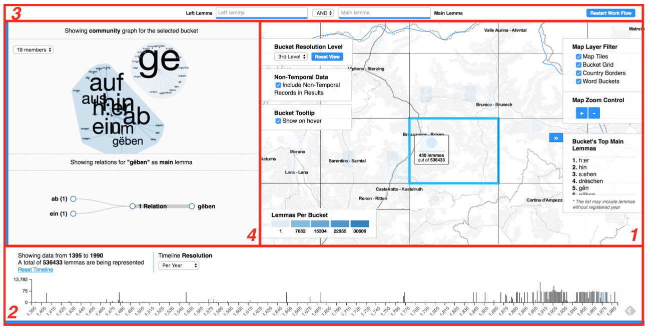


Figure 1: Proposed tool interface: 1) Spatial projection/map. 2) Temporal projection or timeline. 3) Textual search bar. 4) Network analysis view.

solution, they present a web application with three, two-way linked-views that enables geographical, textual and network analyses to be performed at the same time. Network analysis is achieved by means of a force-directed graph in which nodes represent the different concepts and connecting edges their coincidences in meaning. In a similar manner, we employ a network visualization in our pilot tool to present words as interrelated lemmas.

Regarding the proposed workflow of our tool, it has been deeply influenced by work of data visualization experts, and particularly the extension of the *visualization mantra*: "Analyze first, show the important, zoom, filter and analyze further, then details on demand" [14]. This workflow has proven to behave specially well with massive data sets such as the ones employed in our research, resolving the problematic announced years ago by the cited authors: "[...] current and especially future data sets are complex on the one hand and too large to be visualized straightforward on the other hand," [14].

3 Pilot Visualization Tool

Our 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 Fig. 1 we present a screen shot of the tool, showing all of its views.

The interface depicted in Figure Fig. 1 shows four views, of which only three are made available from the beginning to the user (the network analysis area is shown or, conversely, hidden dynamically depending of the current stage of the analysis the user is at). In a similar approach to work by other authors [1], we also propose a dynamic multiple linked views exploration system. We expand and adapt work on the reactivity of web applications and visualization systems by some authors [9, 15] and introduce the component of the search engine as a data-management entity that plays the role of predictable state container often seen in these approaches. Below we provide a short explanation of each of these views:

1. **Spatial view:** Here the data with available coordinate information is projected. It supports the panning, zooming and filtering by selecting specific visual elements on the map.

2. **Timeline:** This histogram is linked to the spatial view and makes a representation of the records with temporal information (projected over the Y axis).

3. **Textual search bar:** This element collects the text queries performed by the user. It employs instant search or "Search as you type": In this technique filtering is performed on each user keystroke, leaving out of the current selection the documents that does not match the search criteria.

4. **Network analysis area:** This area shows the graph representations of the networks present in the current selection. It helps the user to identify patterns and hidden connections between lemmas in a certain geographical area. Closing the linked-view cycle, this view relates to the timeline and textual search bar and permits a fast tuning of the temporal and textual filters. For this purpose two views are employed, as it can be seen in the image. The top view is a modified force directed graph that represents lemma co-occurrences, visually grouped in communities computed by an algorithm which is run on demand [4]. Communities with less than the average members are filtered out from the graph by default as they are considered less relevant, but they can be brought back into the visualization by employing the control on the top left corner. The bottom view is a tree graph that, employing the same data input as the force-directed graph, sorts such occurrences in numerical order for an specific lemma, depending on its role withing the words (prefix or lexeme). This lemma acts as the root of the tree and its position can be switched according to the user's needs using the corresponding control.

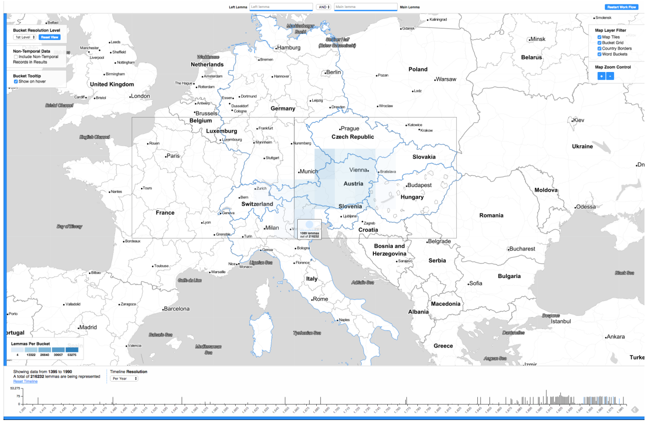


Figure 2: Initial state. Hovering on the different aggregation buckets displays stats about the relevant section of the data.

In Figure Fig. 2 the initial state of the interface is shown. Bucket aggregations at the lowest supported resolution level are displayed on screen, providing a general view of the dictionary data. The user can in turn slice the data by filtering using one of following dimensions: text, time or space. When the number of results has been reduced to a reasonable amount, NA can be launched for a specific portion of terrain, revealing more insights on the underlying nature of the data (See Sect. 4 for examples of this work flow).

3.1 A search engine for the dictionaries

As we anticipated in previous sections, a key feature of the pilot tool to implement was the full-text search. Also, another important software requirement was to build a web-ready piece of software able to run in browsers. For these reasons, the chosen data storage and search engine was the open-source, Apache-licensed ElasticSearch documental search engine. Despite its short lifetime, ElasticSearch has become a strong software alternative in real-time analysis of human-readable, machine-generated computer logs. Given the similarity of these formats to those employed in our context and taking previous work of other DH practitioners as example [11], ElasticSearch presented itself as a suitable solution for indexing and querying our data.

The results of this addition were promising. With more than two million different indexed documents, the search engine adapted very well to work in a web environment, acting as a predictable state container in the context of the reactive paradigm that we employ in our tool. Along with its strong text search capabilities, the search engine offers the possibility to obtain summaries of the response data sets by using *aggregations*[2]. These summaries are condensed in special data structures called *buckets* that, upon receipt, are processed by data controllers which in turn trigger the necessary changes in the linked-view system. The calculation of these metrics is embedded at the core of the search engine and it can be pre-programmed by supplying appropriate data-modeling mappings. Therefore, the common tasks of querying data and obtaining summaries performed substantially faster in comparison to other simpler data-management alternatives.

4 Use Cases

This section presents the results of a short test session performed by the team of lexicographers that collaborated in this research. The two use cases that were found during the session demonstrate the advantages of the proposed workflow in the free exploration of the dictionary and its ability to facilitate the extraction of different kinds of knowledge.

4.1 Colour term usage

At the beginning of the session, the participants chose to base the exploration on 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 cultural knowledge [7]. In this case, they looked for common referents and associations of red (*rot*) with other terms, as this practice can provide insight on the cultural ramifications of such colour [7]. In order to do this, they looked for the presence of red (*rot*) in compound words such as *weinrot* (red wine) or *blutrot* (red blood) and compared the results obtained for different regions. As a first step, the colour term *rôt* (note the pronunciation mark on the "o") was manually entered in the lexeme input box, while the prefix input box was intentionally left blank. Non-temporal data was selected to be included in the results since temporal analysis was not considered relevant in this case. Then, the spatial view presented the results returned by the search engine, showing an even spatial distribution of the query results. Finally, they selected a bordering area between modern Italy and Austria containing 272 sources to perform the NA task. Results are presented in the interface depicted in Figure Fig. 3. Notice how the geographical area for which NA is run is highlighted in the map.

4.2 Using fuzzy searches to find similar headword pronunciations

This second use case expands the work flow by centering the attention on one of the top combinations of red found in the previous example: the lemma *prinn*. The participants launched a new textual query by using the contextual menu of the particle, which returns results of headwords containing this particle at the beginning of the compound word. In addition, they opted to add fuzzy parameters to the search query in order to analyze a different kind of phenomenon.

Variations in the pronunciation and written representation vary greatly among dialect areas. On the segmental level, vowels are a particularly prominent example of showing high variability. Fuzzy searches are therefore particularly well suited to tackling such queries and have also been employed by other authors [16]. When the search term was modified to "prinn∼", obtained results included lemmas that varied in maximum one character from the input term.

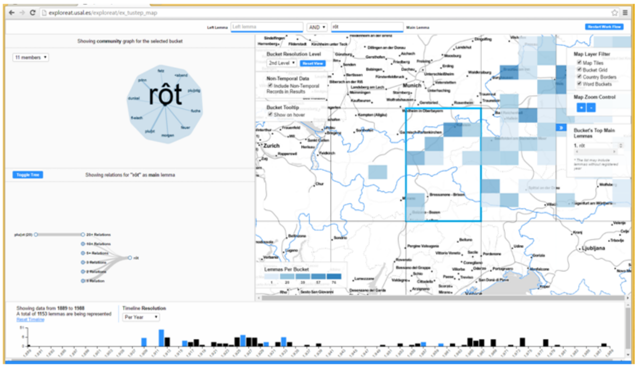
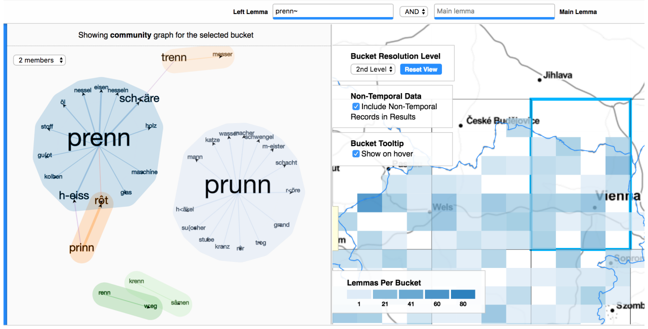


Figure 3: Visual output of the colour term red (*rôt*) in compound words and its possible referents.



Left Lemma: prenn~ | AND | Main lemma: Main Lemma

Bucket Resolution Level: 2nd Level | Reset View

Non-Temporal Data: ☑ Include Non-Temporal Records in Results

Bucket Tooltip: ☑ Show on hover

Lemmas Per Bucket: 1, 21, 41, 60, 80

Figure 4: Network visualization showing two pronunciations of the same underlying form *prenn/prinn*

In turn, NA was launched for the region of Vienna, producing intriguing results (See Figure Fig. 4).

The network visualization showed the network with the different communities detected by the algorithm which are dominated, as expected, by terms matching the fuzzy query. At first sight, the users' attention is drawn to the more populated communities of *prenn* and *prunn*, which are not connected and therefore appear distant in the force-directed graph, *repelling* each other. This is not the case, however, for the *prinn* community, which is connected to the *prenn* node by two different paths: *h-eiss* () and *rôt*. At the top of the visualization appears yet another particle, *trenn* (to separate), which is less tightly connected to *prenn* than the previous one, having only one lemma in common, *sch<äre* (scissors).

The original sources of the records displayed confirm there is a reason why *prenn* and *prinn* appear closer in the graph: *prenn/prinn h-eiss* (very hot) and *prenn/prinn rôt* (burning red) are manifestations of the same underlying word form, both meaning the same. On the other hand, the team of experts stated that they could not find any semantic coincidences between *prenn* and *trenn*, because the introduction of the fuzzy character modified in this case the lexical root and thus the origin of these two words was completely different. This comment also confirmed the validity of the chosen visualization, which placed the node farther apart than it did with the lemma *prinn*.

5 Discussion and Future Work

DH prove a highly interesting field for the application of classic computational techniques in novel ways. The necessity to create open standards, methods and frameworks able to adequately direct this experimentation has become an academic imperative. Envisioning new interactive and plastic techniques capable of managing the increasing volume of data related to humanistic studies and foster

[2] https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations.html

knowledge extraction has become a core component of data visualization in recent years and will be the origin of many of the future applications of this discipline.

This work draws attention to the key role of data visualization in enhancing the accessibility to computational methods usually employed in linguistics. Furthermore, we put into practice a novel architecture specially oriented towards the analysis of big data sets and the creation of responsive visualizations in the web using open standards. The software methodologies and paradigms adopted during the conception of the pilot tool were adequate for the problem domain of this research, specially those concerning the alternation of micro-prototypes and testing sessions with the team of lexicographers involved in this investigation, which contributed greatly to the suitability and usefulness of the resulting software. Additionally, the reactive paradigm proved to adapt well to the inclusion of the search engine as a state manager although and we felt it was good enough for a first attempt. However, in future research they will have to be more tightly connected in order to create more complex interfaces, such as adding more logic to the controllers that allow a complete exploitation of the search engine capabilities.

We identified other possible lines of work that arose over the course of this research. Visual treatment of uncertainty would have a tremendous impact on the study of cultural evolution and the dating and classification of dictionary entries with missing information. The study of the past is, by definition, uncertain and in consequence some of the sources in the study could not be dated accurately by automatic means, while others presented incomplete or missing information. Incorporating elements that are able to transmit this uncertainty and lack of information in visual terms to the user, as showcased in previous work [2, 6], will suppose one of the most important challenges in future investigations.

Regarding the network analysis capabilities, we noticed certain degree of detachment (in interface terms) between the elements in the NA area (and more specifically, the graph nodes) and their counterparts on the map. This fact made the establishment of a direct visual correspondence difficult for some of the participants and as a consequence, the multidimensional nature of the sources was at times hard to grasp. We also noticed this problematic arose more often in medium/large sized computer screens. A common solution for this issue implies merging the two representations, creating a new visualization that allows spatial and network analysis to be performed using the same visual elements and areas of the screen, as seen in other famous visualizations. Moreover, time could also be brought into this same view as other authors have attempted in their work [10]. This could enable more self-contained and effective forms of dynamic network analysis and novel ways to represent the temporal uncertainty that is present in some of the sources.

The analysis of massively populated graphs still needs to be addressed in future versions of the tool. When the number of nodes analyzed reaches the order of thousands, the amount of edges displayed on screen grows exponentially, producing the unwanted, so-called *hairball graphs* and impeding the extraction of knowledge. A common approach to solve this issue involves clustering the data in order to create a more intelligible representation. Although we looked into possible alternatives, more work is needed to be done in conjunction with the team of lexicographers collaborating in this research in order to find a combination of algorithms able to produce meaningful results. Treatment of overlapping nodes in the graph has not been addressed yet and calls for more elaborate solutions like collision detection or similar approaches. If these questions are to be solved, NA will be promoted to a first-class element in our analysis, placed at the same level of time and space. This addition will open new doors to more complex and exciting ways of multivariate analysis.

References

[1] L. Anselin, I. Syabri, and O. Smirnov. Visualizing multivariate spatial correlation with dynamically linked windows. *Urbana*, 51:61801, 2002.

[2] S. Barthelme. Visual uncertainty (a bayesian approach). 2010.

[3] 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.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

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

[6] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. In *Computer Graphics Forum*, vol. 31, pp. 1015–1024. Wiley Online Library, 2012.

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

[8] B. . D. Dorn, Wandl-Vogt. The colour of language! exploiting language colour terms semantically for interdisciplinary research. In *Poster presented at Conference on Progress in Colour Studies (PICS), London, UK.*, pp. 1–8. 2016.

[9] F. M. Facca, S. Ceri, J. Armani, and V. Demaldé. Building reactive web applications. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pp. 1058–1059. ACM, 2005.

[10] K. Grossner and E. Meeks. Topotime: Representing historical uncertainty. *DH2014, Proceedings from Digital Humanities*, 2014.

[11] T. Hauswedell and M. Wevers. Reporting the empire: The branding of metropolises and empire in the pall mall gazette 1870-1900. In *DH Benelux2 Book of Abstrcts for the Second Digital Humanities Benelux Conference*, p. 78–80, 2015.

[12] 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.

[13] 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.

[14] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual Analytics: Scope and Challenges*, pp. 76–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-71080-6_6

[15] C. Kelleher and H. Levkowitz. Reactive data visualizations. In *SPIE/IS&T Electronic Imaging*, pp. 93970N–93970N. International Society for Optics and Photonics, 2015.

[16] C. D. Manning, K. Jansz, and N. Indurkhya. Kirrkirr: Software for browsing and visual exploration of a structured warlpiri dictionary. *Literary and Linguistic Computing*, 16(2):135–151, 2001.

[17] T. Mayer, J.-M. List, A. Terhalle, and M. Urban. An interactive visualization of crosslinguistic colexification patterns. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation, At Reykjavik, Iceland, Volume: VisLR : Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pp. 1–8. 2014.

[18] S. Nielsen. The effect of lexicographical information costs on dictionary making. *Lexikos*, 18(1), 2008.

[19] R. Therón and E. Wandl-Vogt. New trends in digital humanities. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM '16, pp. 945–947. ACM, New York, NY, USA, 2016. doi: 10.1145/3012430.3012630

[20] 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.