The fun of exploration: How to access a non-standard language corpus visually

Roberto Theron, Eveline Wandl-Vogt
Department of Computer Sciences and Automation (University of Salamanca)
Institute for Corpus Linguistics and Text Technology (ICLTT, Austrian Academy of Sciences)
theron@usal.es, eveline.wandl-vogt@oeaw.ac.at

Abstract

Historical dictionaries and non-standard language corpora can greatly benefit from a visual access in order to grasp the inherent tangled and complex nature of the knowledge encapsulated in them. Although visual analytics has been used to tackle a number of language and document related problems, most dictionaries are still reproducing the book metaphor in which Web pages substitute the paper and the user experience is only enhanced by means of hyperlinks. Although fields such as dialectology and dialectal lexicography have incorporated Geographic Information Systems and advanced computational linguistics features, spatio-temporal dynamics can be discovered or understood if appropriate visual analytics techniques are used to surpass the idea of these linguistic resources as alphabetically ordered lists. In this paper we present the work carried out in this direction for the Dictionary of Bavarian Dialects in Austria. By means of multiple-linked views an access that fosters the exploratory analysis of the data is enabled.

Keywords: visual navigation, variational linguistics, historical corpora

1. Introduction

Variation in language use is one of the main concerns in linguistics and its study has been greatly benefited by the use of computers. The early adoption of electronic corpora have paved the way to deeper understanding of the variations that can be found among speakers, associated with age, gender, geography, etc. Thus, electronic corpora have enabled approaching the *life cycle* of linguistic variability in order to shed light on the dynamics of a given language across time and space.

By combining the strengths of machine and human data processing, visual analytics science has established itself as a very successful approach to a great variety of problems across domains, in which the complex nature of the data and its volume have proven to be too great a challenge for previous efforts. The visual analytics community has emphasized the fact that spatio-temporal analysis is not exclusively needed and done by highly qualified specialists (Andrienko et al., 2010); on the contrary, any person concerned about the development of their communities, regions or countries would want to take advantage of the data being collected. This is the case for any language for which vast amounts of information have been recorded and made available through dictionaries and corpora.

Although different methodologies and techniques have been developed from corpus and computational linguistics, and in some cases those advances have been coupled with different forms of visual interactive aids (Lee and Kretzschmar, 1993; Masron et al., 2005), the majority of electronic dictionaries are still regarded as flat lists of entries. This paper presents our preliminary results in an effort aimed at providing a different access to historical dictionaries and corpora by means of visual analytics techniques, in which the exploratory analysis is fostered so the idea of a flat dictionary is surpassed.

In the past, there have been several visualization efforts mainly aimed at escaping the idea of dictionaries as flat lists of entries. During the past decade the navigation of the web of words by means of hyperlinks was enabled and these word relationships are often visualised using different types of graph drawing techniques (Visual Thesaurus[1]). While this is an advantage for dictionary users, there are still several aspects that can dramatically change the users' experience: an enhanced access to the wealth of sources related to dictionaries tailored to the particular needs and tasks of such users can indeed surpass this very notion of a flat dictionary.

The previous statement is especially relevant when non-standard data are considered. Space and time become the steering wheel of any inquiry. In this regard, there are few works that take full advantage of visual interactive access to non-standard dictionaries. In (Theron et al., 2011; Theron and Fontanillo, 2013) one of the first efforts in this direction can be found. Furthermore, in (de Vriend et al., 2011) it is suggested that a working environment offering full support for using visualization as a research tool could take dialect geography into the era of eScience.

The rest of the paper is organised as follows: in the next section, we briefly explain the case of the Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich [WBÖ], from its origin to its present form. In the third section we discuss how the WBÖ data sources have been enhanced thanks to GIS technology. The fourth section is devoted to show how the same data can be exploited from a visual analytics approach, providing a more comprehensive user experience. Finally, we expose the main conclusions and future works that we envision.

[1] Visual Thesaurus [www.visualthesaurus.com/] [accessed: 15.1.2014]

2. 100 years of variational linguistics in a nutshell

Dealing with non-standard language corpora concerning time (historical language) and space (variants, regional language, dialects) is a challenge, especially when focusing usability, navigation and access. To exemplify the challenges and suggested solutions, we picked the example of the Dictionary of Bavarian Dialects in Austria.

The main reasons for doing so are obviously: the dictionary and its material are digitally available, yet the project stands for the type of a traditional, sophisticated, so called territorial dictionary. The materials in the dictionary and those ones used for the dictionary are highly heterogeneous. The dictionary and its corpus offers historic (8th century -) as well as recent materials (- 1998). Furthermore, the corpus offers materials from written sources as well as from questionnaires of spoken language. Professionals and laypersons collected the data in a period for about 100 years in written and digital format.

The project started in 1911 to introduce novel standards in lexicography and give a complete and detailed overview of the Bavarian dialectal variants in the Austrian-Hungarian Monarchy and the Kingdom of Bavaria, by presenting the complete lexicons of these areas with detailed definitions and contextualized examples, by recording the authentic pronunciation, by defining the grammatical coding for each word entry, by tracing the etymology of each lexical item and by registering expert knowledge in the fields of rural techniques, traditional folk medicine and customs. From 1913-1932 109 questionnaires and 9 auxiliary questionnaires with approx. 24,000 detailed questions were sent to selected municipalities. In 1961 the Austrian and Bavarian Academy decided to publish two separate parts of the lexicon: Part I dealing with Austria (with the exception of the province of Vorarlberg) and the (former) German-speaking parts across the current borders of Italy, Slovenia, Slovakia, Hungary and the Czech Republic, and part II dealing with the Bavarian dialects in Germany. Part I, the WBÖ, has been published since 1963 (4 volumes, 8 parts: A-E).

In 1993 the project *Datenbank der bairischen Mundarten in Österreich (DBÖ)* / Database of Bavarian dialects in Austria was initiated aiming at the digitalisation of the material for the dictionary and of source material (e.g. paper slips, cited texts) as well as background information (e.g. cv of collectors and co-workers).

In 1998 a rationalisation concept was issued targeting the dictionary as a (virtual) unit consisting of the printed dictionary and the complementary database. In 2008 the first step in the development of the system *Datenbank der bairischen Mundarten in Österreich* / Database of Bavarian dialects in Austria electronically mapped (dbo@ema) was taken. Since 2010 (Wandl-Vogt, 2010), some thousand example entries are digitally available; furthermore, background data and registers are online and interactively accessible.

In 2013, first steps towards a machine-readable dictionary and connection with Linked (Open) Data were taken (Wandl-Vogt and Declerck, 2013).

3. Access to the dbo@ema

In its current form, dbo@ema can be explored in a traditional way: the user can access the data according to different categories (lemma, bibliography, person, location). For each of these categories, the user would need to access the desired information following a list of pages ordered alphabetically.

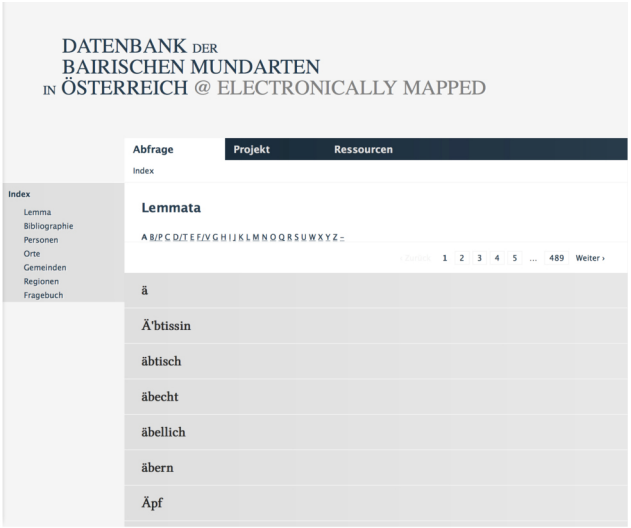

Figure 1: Access to the dbo@ema by means of alphabetical lists.

Figure 1 shows a screenshot of dbo@ema in which a list of lemmas is used to provide direct access to the desired lemma. Once the user has selected a particular item of information, a page containing relevant information is shown, e.g., type of lemma, multimedia sources, etc. (see Figure 2).

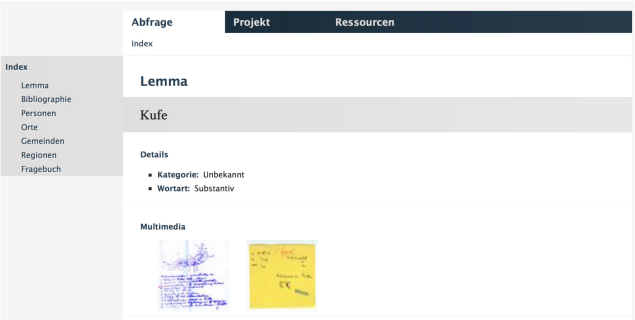

Figure 2: Details of the lemma 'Kufe' including multimedia sources.

The same procedure can be followed when searching for persons, regions, etc. In these cases, since most of the information is georeferenced, an interactive map is included, so the user can grasp the spatial influence of the searched item.

Figure 3 shows the details page for Werner Bauer, including personal information and a list of sources related to this author.

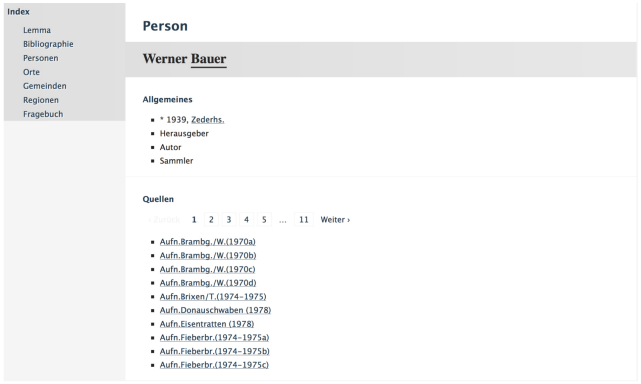

Figure 3: Details related to author Werner Bauer.

Accordingly, Figure 4 shows an interactive map of the spatial distribution of data related to the same author.

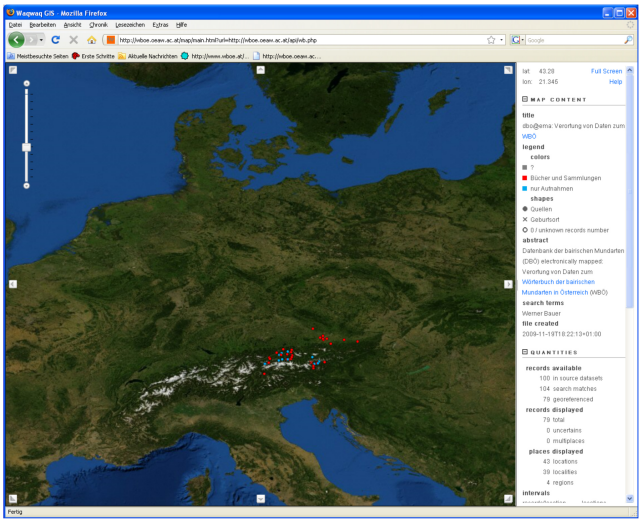

Figure 4: Spatial distribution of data related to the author Werner Bauer..

Finally, the user can go inspecting the details of each spatial point in the map, available by means of tooltips (Figure 5).

Although this provides a method of accessing the spatial distribution of the data stored in the corpus, it is only useful for directed searches, where the user actually knows what she is looking for. Exploratory tasks are not supported, since the context of information is lost once the user access the information of a given lemma, person, location, etc.

Furthermore, although details can be navigated by means of hyperlinks, maps are independent, so a click on a hyperlink directs the users to a new Web page; also the interactions in the map are limited to its current information and the user cannot access further information from the tooltips. Thus, we approached the same data sources from a different/complementary perspective; we introduced some visual analytics techniques in order to provide a more comprehensive access to the corpus data.

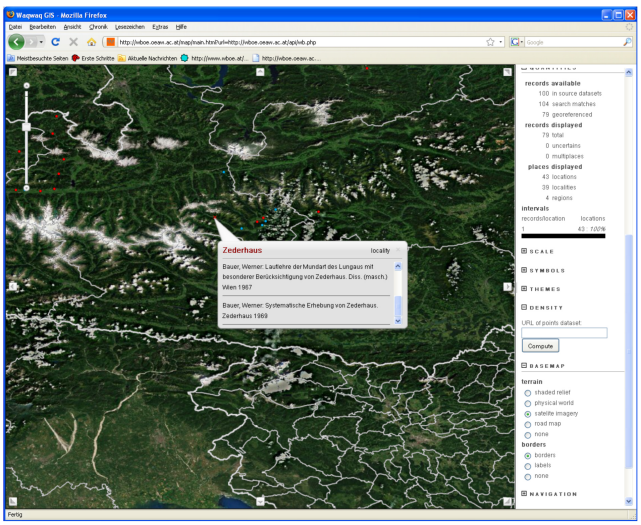

Figure 5: The user may explore the spatial distribution of the data and access individual information of a particular point..

4. Visual Analysis in the dbo@ema

One of the first issues that we wanted to address was providing an overview of the spatial distribution of the data that in turn could be further explored by means of user interaction. We wanted also to maintain the access to all the details originally available in the Web version.

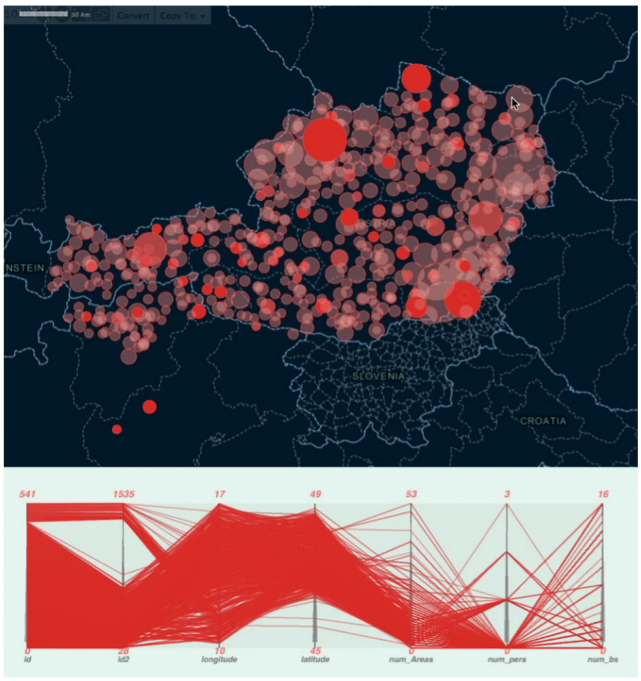

Figure 6: Graduated point symbol maps and parallel coordinates views that enable visual analysis in dbo@ema.

Figure 6 shows two views conveying quantitative information related to regions (it can be done for communities and localities similarly). For each region the number of subregions, persons, lemmas and documents is computed. On

top, a graduated point symbol (Flannery, 1971) or bubble map is shown, in which the size opacity of each bubble. In the case of figure 6, the bubble size depends on the number of sub regions, while the opacity depends on the number of documents related to that region. These visual encodings can be used to convey results of dialectometry studies such as maximum similarity indices.

At the bottom of figure 6, a parallel coordinates (Inselberg, 1985) plot (PCP) can be seen, which conveys the same information by means of polylines (each polyline is linked to a bubble). This way the user can explore the whole dataset and discover patterns related to distribution of sources, etc. The PCP is interactive, so the user can filter data according to different criteria (i.e., filter out the regions that have less than 3 persons related to them and within a particular geographic area (longitude and latitude values), and so on). Finally, once a particular region is on focus, the detailed information is shown, and further information can be retrieved (e.g., details of documents or persons) and shown upon user's request). Both views feature several ways of interaction that enable a typical sense-making loop, going from the overview, filtering data, to details on-demand (Keim et al., 2008).

Figure 7 shows an example in which the user has focused the analysis on the central area of the map, using filters for the latitude and longitude of the data points. The user is inspecting the region Kärnten with a high (16) number of related documents.

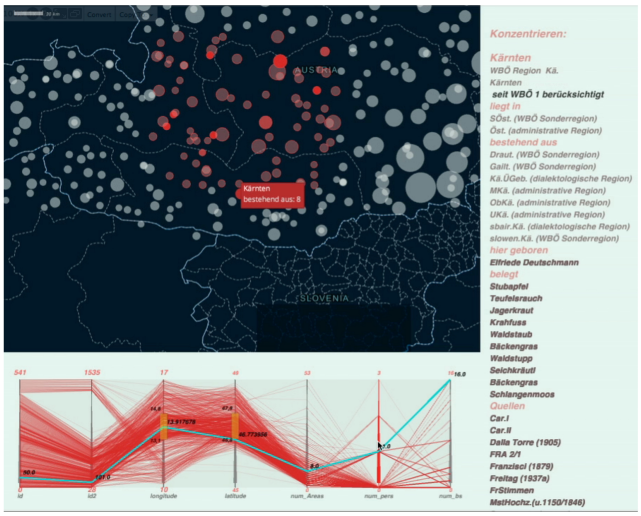

Figure 7: By means of interactions, the user can easily explore the spatial dynamics of dbo@ema.

5. Conclusion

The use of visual analytics can greatly enhance the experience of users, enabling exploratory analysis that contribute to the discovery of spatio-temporal dynamics that would remain hidden with traditional flat access to dictionaries and corpora. We have assessed the potential of exploration of georeferenced data for regular dictionary visitors.

In this first version of our prototype we have used mainly information available in the dbo@ema, but it has been designed from a visual analytics approach, to facilitate the addition of dialectometry studies such as maximum similarity indices.

However, the temporal dimension has not been fully exploited, and the visual analytics research has developed efficient techniques that can be integrated in our work.

6. Acknowledgements

This research was supported by the Spanish MINECO grant FI2010-1623.

7. References

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M. J., Schumann, H., and Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600.

de Vriend, F., Boves, L., Van Hout, R., and Swanenberg, J. (2011). Visualization as a research tool for dialect geography using a geo-browser. *Literary and Linguistic Computing*, 26(1):17–34.

Flannery, J. J. (1971). The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 8(2):96–109.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). *Visual analytics: Scope and challenges*. Springer.

Lee, J. and Kretzschmar, W. A. (1993). Spatial analysis of linguistic data with gis functions. *International Journal of Geographical Information Science*, 7(6):541–560.

Masron, T., Rainis, R., Ghazali, S., Lah, S. C., Sazali, A., Ghani, A. A., and Ghafor, S. A. (2005). Using gis to grasp dialectal variation. In *Proceedings of Map Asia 2005 Conference, Jakarta, Indonesia*, pages 1–10.

Theron, R. and Fontanillo, L. (2013). Diachronic-information visualization in historical dictionaries. *Information Visualization*. Published online, doi: 10.1177/1473871613495844.

Theron, R., Fontanillo, L., Esteban, A., and Seguín, C. (2011). Visual analytics: A novel approach in corpus linguistics and the nuevo diccionario histórico del español. In Carrió, M. L. and Candel, M. A., editors, *Actas del III Congreso Internacional de Lingüistica de Corpus. Las Tecnologías de la Información y las Comunicaciones: Presente y futuro en el Análisis de Corpus.*, pages 335–342.

Wandl-Vogt, E. and Declerck, T. (2013). Mapping a traditional dialectal dictionary with linked open data. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., and Tuulik, M., editors, *Proceedings of the eLex 2013 conference, Electronic lexicography in the 21st century: thinking outside the paper.*, pages 460–471.

Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis*, 2:35–53.