Spatializing a Digital Text Archive about History

A. Bruggmann, S.I. Fabrikant

University of Zurich, Department of Geography, Zurich, Switzerland
{andre.bruggmann,sara.fabrikant}@geo.uzh.ch

1 Introduction

The amount of digital text data available in online libraries has risen dramatically in recent years. GoogleBooks or the Universal Digital Library (UDL) initiatives illustrate this impressively. The rapid evolution of vast digital text data archives has spurred the growth of an interdisciplinary Digital Humanities (DH) community, as [1] puts it, the once inaccessible has suddenly become accessible. Researchers in the humanities and social sciences have recognized the big potential digital text archives might offer to gain new insights on long-standing research questions. Especially interesting are unstructured or semi-structured digital libraries in this context, as text documents have been central to the humanities and social sciences long before digitization. Along these developments, the need for automatically extracting new knowledge from text corpora using advanced data mining methods and information reduction techniques has risen as well. Text corpora also bear exciting research avenues for spatially aware disciplines and research fields, including geography, GIScience, and the interdisciplinary geographic information retrieval (GIR) community. This is because text documents often contain explicit and implicit spatio-temporal and thematic information, which can be automatically extracted, reorganized, visualized and analyzed for knowledge generation.

We would like to introduce our interdisciplinary research project for discussion at the workshop which offers one avenue for bridging the gap between the digital humanities and GIScience. We combine methods from geographic information retrieval (GIR) and geovisual analytics (geoVA) in order to gain new insights from a digital dictionary about the history of Switzerland [2]. In own prior work, we illustrated how spatio-temporal analysis methods can be applied to a history text archive and how long-standing geographic principles (e.g., Tobler's law) might be verified with data not typically employed in GIScience [3,4]. For the workshop, we present preliminary findings of this spatio-temporal approach, and raise future work ideas for discussion. We are particularly interested in incorporating sentiment analyses in our work, in order to assess *how* (historical) places are referred to in texts over time. In our transdisciplinary approach we aim to create a geographic information observatory. We illustrate this by combining well established methods in GIScience to mine a typical data source for social science and humanities researchers, and by incorporating methods typically employed in social science (e.g., sentiment analysis) to expand the methodological toolkit in interdisciplinary GIScience research dealing with multivari-

ate (text) data. Below we sketch current ongoing work as a starting point for discussion.

2 Space and Time in the HDS

2.1 Methods

We chose the Historical Dictionary of Switzerland (HDS) as prototypical text data source for our approach, as it represents a common example of an online digital text archive in the humanities. It contains explicit and implicit spatial, temporal, and thematic information in semi-structured text documents, in this case, about the history of Switzerland. The dictionary is multi-lingual (i.e., German, French, and Italian), and consists of 36,188 articles, categorized in *thematic contributions* (e.g., events), *geographical entities* (e.g., municipalities), *biographies* and articles about historically important *families*. In this version, there are no possibilities to browse or query the articles by space, time or theme.

In a first step we automatically retrieved spatial and temporal information from the German version of the HDS. Other language versions will be considered in future work. We employed the approach presented in [5] to automatically extract spatial information from the articles. This resulted in 169,094 toponyms (e.g., cities, single objects, forests, rivers and lakes, mountains) stored in a database. Next, we automatically retrieved temporal information (e.g., dates such as 07/06/1856 or time periods such as 19th century) from the HDS by employing Heideltime [6]. As a result, 510,357 temporal annotations were additionally stored in the database.

We generated graph models to conceptualize co-occurrence relationships of the 40 most often mentioned toponyms in the HDS articles over time, following the approach proposed by Hecht and Raubal [7]. We chose *centuries* as the temporal unit of analysis, and thus aggregated extracted temporal annotations to centuries (e.g., the date 07/06/1856 is a member of the category 19th century). Each article was assigned century weights, according to the frequency of respective temporal annotations occurring in the articles (e.g., article A: 19th century: 0.5, 20th century: 0.3, 21st century: 0.2). We used this temporal article weighting scheme to compute the strength of toponym relationships by summing up the weighted relationships in articles where the two toponyms co-occur. In other words, the more often two toponyms co-occur in articles with a high percentage of temporal annotations categorized as 19th century, the higher the weight for their relationship in the 19th century. Finally, we visualized the relationship graphs in a series of network visualizations, following the spatialization framework by Fabrikant and Skupin [8]. A more detailed description of the approach is presented in [4].

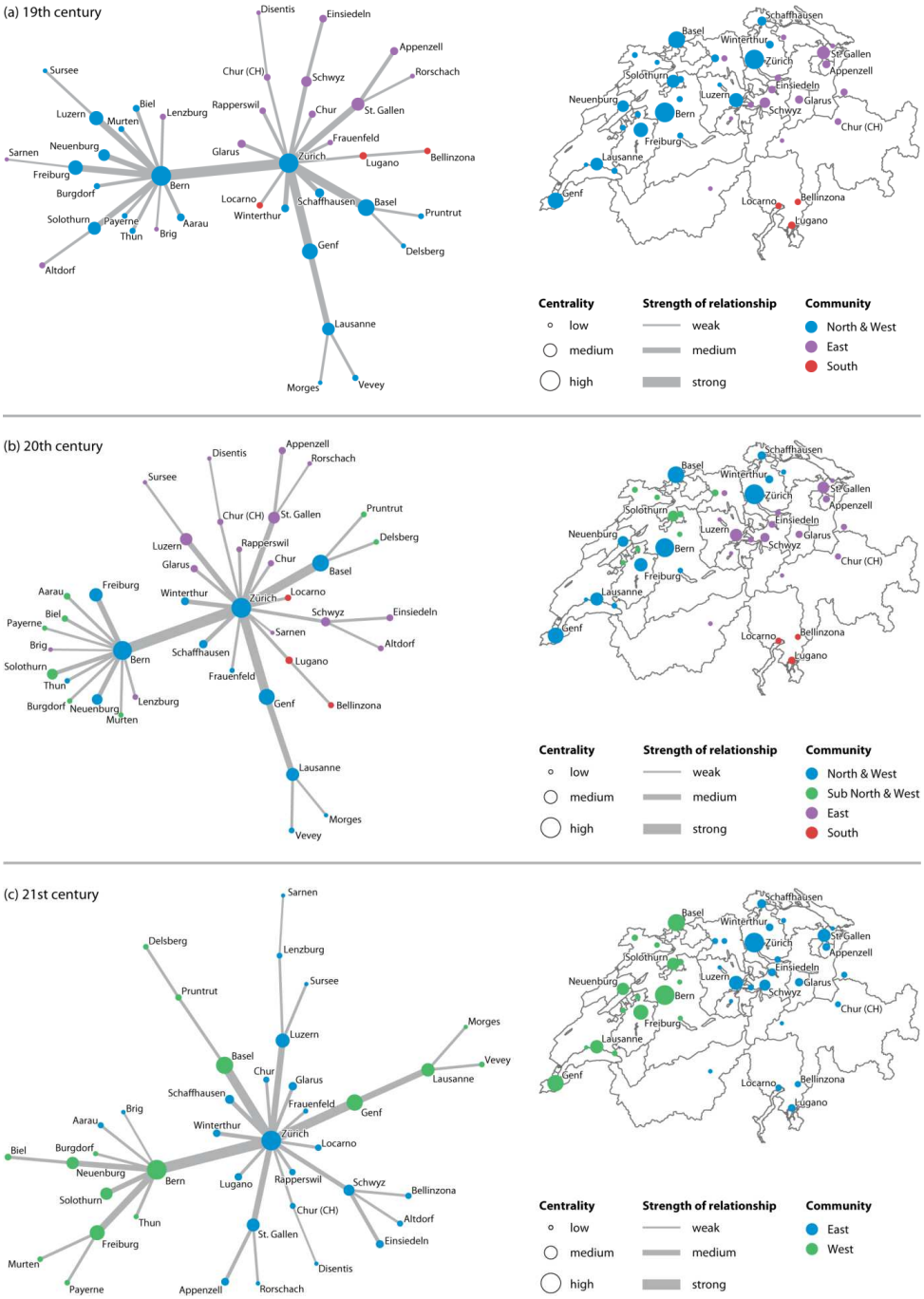

Fig. 1. Toponym relationship from the 19th to the 21st century [4] (map data source: swisstopo, `http://www.swisstopo.admin.ch/`)

2.2 Results and Discussion

In Figure 1 three network visualizations are shown, each depicting the structural most salient relationships (i.e., using a minimum spanning tree) between toponyms in Switzerland over the last three centuries. The graphs were visualized with the Network Workbench [9] using the GEM layout algorithm to avoid edge crossings. Toponyms that have strong relationships in a specific century are visualized closer together on the network, and are connected with a larger edge than those that have weaker relationships. The size of the nodes represents the importance of a toponym (i.e., its strength) in the network, calculated by summing all weighted relationships with all other toponyms in the network; the larger the node, the higher its importance. We also ran the Blondel community detection algorithm on the complete network in order to delineate toponym clusters which separate densely connected toponyms within a community from weakly connected toponyms outside a community [10]. The same information is depicted geographically on a map of Switzerland in Figure 1, with the 20 most frequently occurring toponyms labeled for reference.

By analyzing the network structures in Figure 1, the increasing degree (i.e., directly connected nodes) of *Zürich* (the financial capital) compared to *Bern* (the political capital) becomes salient over time. In the 19th century, the degree for *Zürich* and *Bern* is almost the same (i.e., 14 and 13), but in the 21st century *Zürich*'s degree increases to 15 and *Bern*'s degree decreases to eight. Furthermore, Tobler's [11] first law of geography ("Everything is related to everything else, but near things are more related than distant things") is evident in the community structure over time. The communities form contiguous spatial clusters in the maps in each time slice, except for the green cluster in the 19th century which divides the blue cluster into two parts. Further, we observe a merge of the red cluster in the Italian speaking part south of the Alps (i.e., Lugano, Locarno, and Bellinzona) with the German speaking blue cluster in the north of the Alps in the 21st century which could be due to the opening of the Gotthard road tunnel in 1980 that connects Southern Switzerland with Northern Switzerland. Further results and a detailed discussion is presented in [4].

3 Next Steps

At this point, we only considered the spatial and the temporal information contained in the HDS. In ongoing and near future work we wish to focus on extracting and analyzing thematic information buried in the HDS corpus, and combine this with the reorganized spatio-temporal data. Below we sketch ideas on how to achieve this, and for which we would like to get feedback at the workshop.

3.1 Thematic Analysis

In order to explain found relationships between toponyms, we will study the topicality of articles that connect toponyms in the HDS in more detail. One possible approach for this is Latent Semantic Analysis (LSA) [12], including Topic Modeling [13]. We will employ the Text Visualization Toolbox (TVT) in MATLAB [14] for LSA. We

aim to again automatically group articles with similar thematic content by using the Blondel community detection algorithm, as mentioned earlier. Automatic labeling of uncovered clusters could be achieved by the tf-idf method, for example. The joint visualization of how thematic clusters in toponym relationships might change over time could further help to explore reasons why toponym relationships might have changed over time. A first conceptual approach how one might visualize this in a (static) network is illustrated in Figure 2 below.

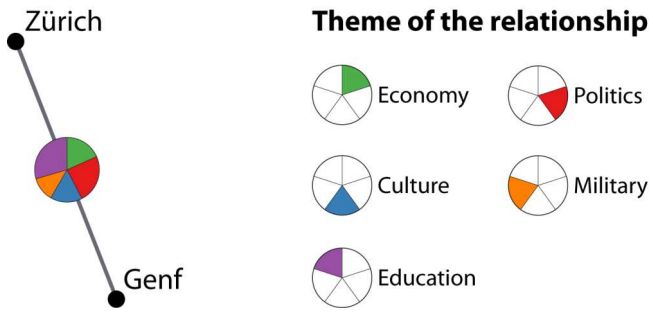

Fig. 2. Themes associated with the toponym relationship *Zürich* and *Genf*

A pie chart is placed on the edge linking the toponym *Zürich* with *Genf*. Distinct color hues represent the thematic groups and the size of the pie chart sectors reflects the importance of a specific theme for this toponym relationship.

We will have to face various challenges to implement the illustrated approach above and we would like to raise them at the workshop for discussion. Firstly, with Topic Modeling, for instance, one needs to determine the number of extracted topics a priori. One possible way to do this is to match topic numbers with currently available themes in the HDS. Another possibility might be to test the performance of different Topic Modeling solutions (i.e., with varying number of topics) using the perplexity measure [15]. Second, the (automatic) labeling of the extracted topics will be a major challenge. Even though we have already worked on methods how to label and distinguish article groups thematically [16], most methods that we have tested so far still require time consuming manual cross-validation and human interpretation. Third, the more toponyms one aims to visualize in a network, as shown in Figure 2, the more difficult the network might be to read and interpret. One solution to mitigate this, is to adopt a cyclic user-centric design approach, that is, to create different visualization solutions, discuss them with historians, and adapt them accordingly.

3.2 Sentiment Analysis (SA)

In a further step, we wish to employ sentiment analysis to the HDS. One the one hand, we aim to get a better understanding about *how* (historical) places in Switzerland are described by Swiss historians. While one would not expect to detect an authors' or editors' positive or negative sentiments towards a specific topic in a scholarly lexicon, one could expect that the content of an article, for example, about an infectious dis-

ease (e.g., Cholera) should be identified and classified as negative with sentiment analysis, as it might contain many words and expressions which are commonly used in such life-critical contexts (e.g., death). Further, we envision to study the changes of sentiment towards a particular place over time, by incorporating the re-organized temporal information (see Section 2).

General Inquirer (GI) is a well-known and widely used sentiment analysis tool [17]. The GI tags and counts words in English texts by classifying them into *negative*, *positive*, *strong*, *active*, and many other categories. Tetlock et al. [18] used GI to parse reports of firms to quantify their financial performance. In GIScience, sentiments expressed in Tweets has been analyzed, and visualized on maps (e.g., [19]), using similar approaches. However, as GI only works with English texts, we have developed a prototype SA tool in Python which works similarly to the GI, but incorporates the *SentiWS* dictionary [20]. The *SentiWS* categorizes German words into *negative* (e.g., Gefahr = danger) and *positive* (e.g., Freude = joy) sentiments. It contains about 3,500 German sentiment words, mainly based on an automatic translation of *GI*'s word categorization. Other German sentiment dictionaries such as the Berlin Affective Word List Reloaded (BAWL-R) [21] will be considered in future work.

In a first attempt, we calculated the sentiment scores for all HDS articles. Table 1, lists HDS articles which were ranked most negatively and positively (English translations in brackets, where necessary) according to the sentiment scores (i.e., *negative* / *positive*, etc.) which were assigned to the HDS articles by applying our own prototype SA tool. We calculated the corresponding sentiment scores by summing up the weights of the *negative* and *positive* words in the HDS articles, and normalized the score by the document length. Finally, we multiplied all sentiment scores by 1000. The category *thematic contributions* seems to score highest compared to the other categories. We thus include these articles in our study for now, the remaining categories will be considered in future work.

Table 1. Top five negative and positive themes

Negative HDS articles	Score	Positive HDS articles	Score
Articulans	-19.5	Ehrschatz (legitimation fee)	5.2
Cannstatt, Gerichtstag von	-17.2	Pfarrhäuser (parochial house)	4.2
Pro Libertate	-15.8	Fête des Vignerons	4.0
Helvetische Annalen (Helv. Annals)	-12.7	Comics	3.9
Cholera	-11.5	Hülsenfrüchte (legume)	3.6

This first analysis looks promising, especially when considering the negatively scored articles. For example, the article *Articulans* is mainly about banishments of people, and about people who were sentenced to death and executed. The article *Cholera* describes how Switzerland was affected by this disease. These articles contain a great deal of negatively weighted words (e.g., Tod = death, Streit = fight, Panik = panic). However, interpreting the positively ranked articles seems not to be as straightforward. According to the literature, reasons could be that negative information is more salient and thus has more impact on readers, than positive information [18], and also because of frequently appearing negations of positive words in certain contexts [22].

We will have to face various further challenges regarding sentiment analyses in planned future work. The choice of an appropriate method in this context is of major importance. For example, there have not been many studies about how to implement and evaluate sentiment analyses in languages other than English (i.e., German). A further challenge will be the effective and efficient visualization of the results. We plan to depict historically most important places (i.e., toponyms) on a map of Switzerland, and visualize the potentially changing sentiments with which they are associated in different time periods, for example, by using pie charts (i.e., different periods of time = different sectors in the pie chart). We can then assess the valence of a place by analyzing its association with negative or positive texts, and how this might have changed over time. We will discuss results and visualization ideas with the historian target group, in order to evaluate our solutions.

4 Summary

Our contribution to create a geographic information observatory incorporates a combination of GIScience methods applied to data typically employed in the humanities. We have sketched ideas and illustrated our transdisciplinary approach in combining geographic information retrieval with geovisual analytics to retrieve, reorganize, and visualize text information about space, time, and theme in a semi-structured digital humanities text corpus. We also suggest how sentiment analysis applied to articles in a history dictionary could help to reveal geographically relevant context of historical places, and how these places are described by historians.

We further illustrate how long-standing geographic principles (i.e., Tobler's law) can be verified in non-typical geographic data contexts (i.e., text corpora for the humanities), and how the methodological toolkit in GIScience can be expanded by incorporating methods, borrowed from social science and the humanities (e.g., sentiment analysis). We further contribute to the digital humanities by applying sound cartographic techniques to make uncovered thematically relevant information visually salient to a user. In doing so we hope to help historians to generate new research hypotheses, and to get a better understanding of complex spatial, temporal, and thematic relationships buried in text archives.

References

1. Jockers, M.L. (2013): Macroanalysis – Digital Methods & Literary History. University of Illinois Press.
2. Historical Dictionary of Switzerland (HDS), http://www.hls-dhs-dss.ch/
3. Bruggmann, A., Fabrikant, S.I. (2014): How to visualize the geography of Swiss history. In: Huerta, Schade, Granell (eds.) Connecting a Digital Europe through Location and Place. International Conference on Geographic Information Science, AGILE 2014, Jun. 3-6, 2014, Castellón, Spain, ISBN: 978-90-816960-4-3.
4. Bruggmann, A., Fabrikant, S.I. (in press): Spatializing time in a history text corpus. In: Proceedings of the 8th International Conference on Geographic Information Science, GIScience (extended abstracts), Sept. 23-26, 2014, Vienna, Austria.
5. Derungs, C., Purves, R.S. (2014): From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. International Journal of Geographical Information Science 28(6), 1272-1293, DOI: 10.1080/13658816.2013.772184.
6. Strötgen, J., Gertz, M. (2013): Multilingual and Cross-domain Temporal Tagging. Language Resources and Evaluation 47(2), 269-298.
7. Hecht, B., Raubal, M. (2008): GeoSR: Geographically Explore Semantic Relations in World Knowledge. In: Bernard, L., Friis-Christensen, A., and Pundt, H. (eds.) 11th AGILE International Conference on Geographic Information Science.
8. Fabrikant, S.I., Skupin, A. (2005): Cognitively Plausible Information Visualization. In: Dykes, J., MacEachren, A.M., Kraak M-J. (eds.) Exploring Geovisualization, 667-690.
9. NWB Team (2006): Network Workbench Tool 1.0.0., http://nwb.slis.indiana.edu
10. Blondel, V.D., Guillaume, J-L., Lambiotte, R., Lefebvre, E. (2008): Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, DOI: 10.1088/1742-5468/2008/10/P10008.
11. Tobler, W. (1970): A Computer movie simulating urban growth in the Detroit region. Economic Geography 46(2), 234-240.
12. Landauer, T.K., Foltz, P.W., Laham, D. (1998): An Introduction to Latent Semantic Analysis. Discourse Processes 25(2&3), 259-284.
13. Steyvers, M., Griffiths, T. (2007): Probabilistic Topic Models. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis.
14. Hespanha, S.R., Hespanha, J.P. (2011): Text Visualization Toolbox – a MATLAB toolbox to visualize large corpus of documents, http://www.ece.ucsb.edu/~hespanha
15. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003): Latent dirichlet allocation. Journal of Machine Learning Research 3, 993-1022.
16. Bruggmann, A. (2012): Netzwerkvisualisierung der Ostschweiz. Master Thesis, University of Zurich, Zurich.
17. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M. (1966): The General Inquirer: A Computer Approach to Content Analysis, MIT Press.
18. Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2008): More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance 63(3), 1437-1467.
19. Mitchell, L., Frank, M.R., Harris, K.D., Dodds, P.S., Danforth, C.M. (2013): The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PLoS One 8(5).
20. Remus, R., Quasthoff, U., Heyer, G. (2010): SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Resources and Evaluation (LREC'10).

21. Võ, M.L-H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M.J., Jacobs, A.M. (2009): The Berlin Affective Word List Reloaded (BAWL-R). Behavior Research Methods 41(2), 534-538.
22. Loughran, T., McDonald, B. (2011): When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66(1), 35-65.