Bruggmann, A., Fabrikant, S.I., (2014). Spatializing a digital text archive about history. Proceedings (extended abstracts), 8th International Conference on Geographic Information Science, Sep. 23, 2014, Vienna, Austria.

Spatializing time in a history text corpus

A. Bruggmann and S. I. Fabrikant

University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zurich, Switzerland
Email: {andre.bruggmann; sara.fabrikant}@geo.uzh.ch

1. Introduction

Due to recent mass digitization initiatives of large text archives (e.g., GoogleBooks), the online access to massive amounts of text documents has risen dramatically. These efforts offer exciting new ways to explore human knowledge encapsulated in text. While text documents have been central to the humanities and the social sciences long before digitization, text sources are still largely untapped for spatio-temporal analyses in GIScience.

We aim to fill this gap and present a theory-driven framework that applies geographic information retrieval (GIR) and geovisual analytics (GeoVA) to an online dictionary about Swiss history. We chose the Historical Dictionary of Switzerland (HDS 2014) available in German, French and Italian as a prototypical online text archive, as it specifically includes spatial, temporal, and thematic information. Even though the 36,188 HDS documents implicitly contain spatio-temporal information, there are no such browsing or query possibilities in its current version. In own prior work (Bruggmann and Fabrikant 2014) we illustrate how spatial relationships between toponyms mentioned in text documents can be automatically extracted, re-organized semantically, and presented to an information seeker in static cartographic maps, and spatialized displays. In this paper, we specifically focus on how to automatically extract and visualize temporal information from the HDS, as to allow an information seeker to explore whether and how spatial relationships between historically relevant Swiss toponyms might have changed over time.

2. Methods

Following the methodology presented in Bruggmann and Fabrikant (2014), we first retrieved 169,094 toponyms from the HDS articles in German, by first identifying candidate toponyms with the Swissnames gazetteer (swisstopo 2014), and resolving disambiguation issues (Derungs and Purves 2014). We then re-organized the retrieved spatial data by assuming a (semantic) relationship between two toponyms, if they both co-occurred in the same article (Hecht and Raubal 2008). By example for this case study, we focus on the forty most often mentioned Swiss toponyms in the HDS. To analyze the potentially changing nature of toponym relationships over time, we employed *HeidelTime* (Strötgen and Gertz 2013) to automatically extract 510,357 temporal annotations from HDS text corpus, including *dates* (e.g., 07/09/1984), *periods of time* (e.g., 18th century) and *other temporal information*. In this paper, we exemplify our approach using centuries as the temporal unit of analysis, even though other temporal resolutions are possible. We used this temporal unit to weigh toponym relationships in each article. In other words, if two toponyms co-occur in articles that contain a high percentage of temporal annotations categorized as 20th century, their relationship is assigned a higher weight for the 20th century, compared to two toponyms that only co-occur in articles that have few annotations categorized as 20th century. Finally, we are able to visualize the extracted spatio-temporal toponym relationships, based on Fabrikant and Skupin's (2005) empirically validated spatialization framework.

3. Results and Discussion

We depict the extracted toponym relationships covering the last three centuries as a series of spatialized networks in Figure 1, where toponyms with stronger relationships are placed closer to one another on the network than those with weaker relationships. We constructed the network displays for each century separately, using the GEM layout algorithm to avoid edge crossings, and by applying the minimum spanning tree (MST) pathfinder algorithm available in the Network Workbench (NWB Team 2006) to visualize only the structurally most important relationships. Line width represents the strength of toponym relationships in the network. Toponym importance was calculated by summing all weighted relationships with all other toponyms in the network. Varying node sizes shows this: the larger the node, the higher the toponym importance in the network. We also ran the Blondel et al. (2008) community detection algorithm to investigate whether extracted toponym relationships might form node clusters that are more densely connected within the group, than with the rest of the network, and to identify whether these clusters might change over time. We visualized toponym clusters with differently colored nodes in Figure 1. Similarly, we depicted this information on a map of Switzerland, with the twenty most frequently occurring toponyms labeled for reference.

Focusing on the two most central nodes in the networks shown in Figure 1, i.e., *Zürich* (the financial capital) and *Bern* (the political capital), one can detect the steady increase of *Zürich*'s degree (i.e., the number of connected nodes) over time compared to *Bern*'s. While the degree for *Zürich* (14) and *Bern* (13) is about the same in the 19th century, *Zürich*'s degree rises to 15 nodes in the 21st century, compared to *Bern*'s, which drops to only eight. Hence *Zürich*'s well established importance as Switzerland's major economic hub today can be traced back with our semantic analysis of the HDS articles. Figure 1 shows that *Zürich*'s degree accelerated in the 20th and at the beginning of the 21st century.

Strikingly, Tobler's (1970) first law of geography ("Everything is related to everything else, but near things are more related than distant things") is also evident. The colored toponym nodes form contiguous spatial clusters in the maps in each time slice. The relationship dynamics of the blue and green clusters is interesting. The green toponym cluster *Sub North & West* appears in the 20th century as a sub-cluster of the blue colored *North & West* toponym cluster. One possible reason for this could be due to the separatist movements in the western parts of this region after WW II, resulting in the creation of the new Canton of *Jura* (located northwest of the city node labeled *Solothurn*) in 1979.

The central nodes *Zürich* and *Bern* are no longer located in the same cluster in the 21st century. *Zürich* now emerges as the center hub city for the eastern half of Switzerland, and *Bern* for the western half, respectively. The prior distinct toponym cluster in the Italian speaking region south of the Alps (i.e., *Lugano*, *Locarno*, *Bellinzona*) merges with the German speaking blue toponym cluster in the 21st century. One important reason for this, also connected to the rise of *Zürich*'s economic importance, may be the opening of the Gotthard road tunnel in 1980 which connects Southern Switzerland with its northern parts. The network visualizations provide another lens to view the hierarchical toponym relationship structure of over time, for example, by showing *Zürich*'s rising connectivity in the course of time, and also by detailing toponym hierarchies in hub nodes and peripheral nodes.

These encouraging results already illustrate how semantic analyses of space and time concepts extracted automatically from text documents in combination with geovisual analytics approaches can prove useful to assess the dynamics of spatial structure in a history text corpus over time.

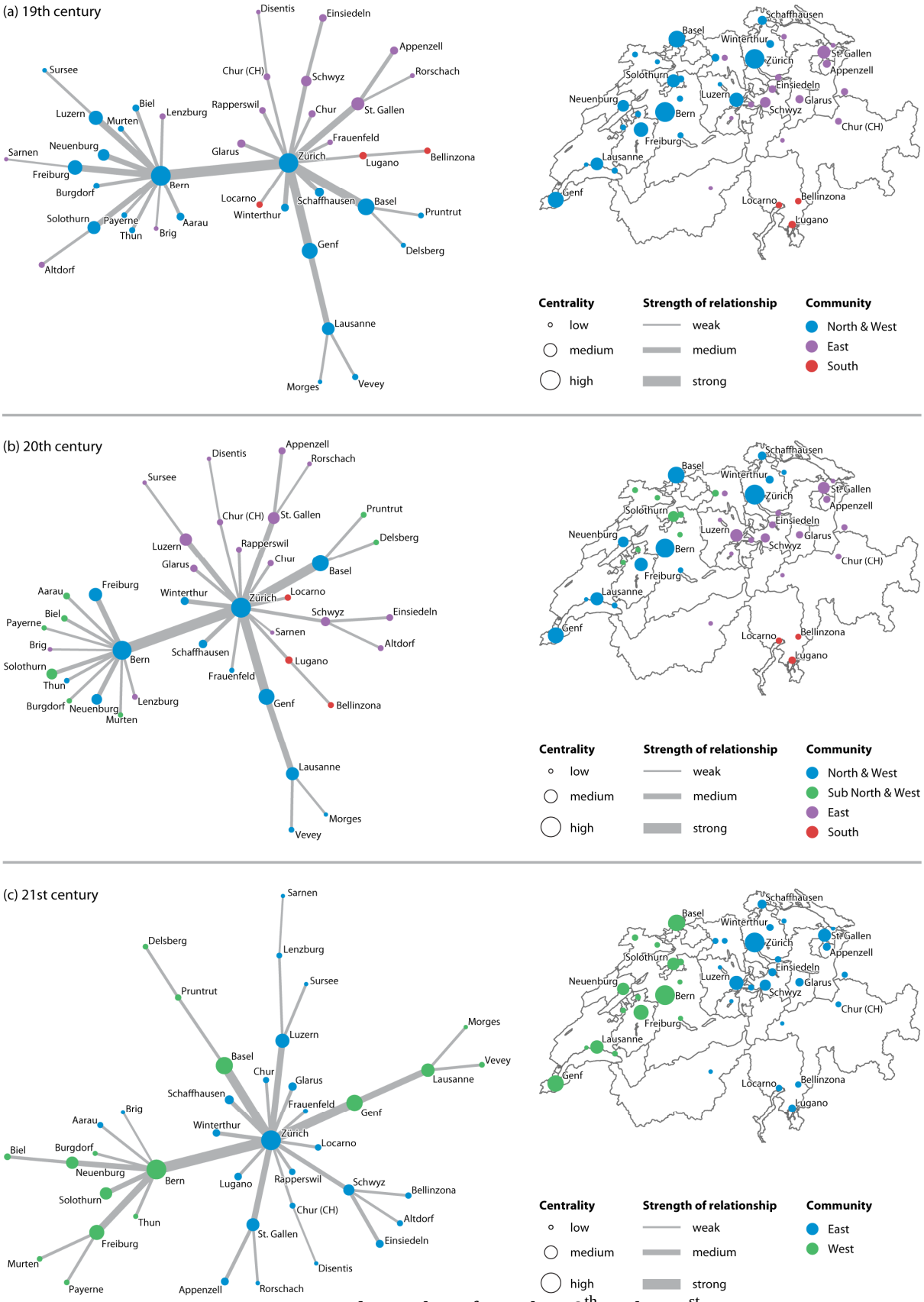

Figure 1: Toponym relationships from the 19th to the 21st century (map data source: swisstopo 2014).

4. Summary and Outlook

This paper introduces a novel text analysis framework based on GIR and GeoVA to automatically uncover and visualize latent spatio-temporal relationships buried in a history text corpus. In doing so, we hope to contribute our GIScience perspective to future interdisciplinary research projects in the digital humanities where space and time matter.

In future work, we aim to integrate thematic information analyses into our framework, as to identify the topicality of toponym relationships (e.g., economy, politics, culture, etc.) and how these might change over time. Finally, we will develop an interactive (online) user interface (e.g., using D3 technology) to extend the current HDS with spatio-temporal browsing and search capabilities.

Acknowledgements

We would like to thank Curdin Derungs, Jannik Strötgen and Julian Zell who specifically helped us to implement the GIR part of our research. We are also grateful to Ross S. Purves and Damien Palacio for their invaluable feedback on this research project.

References

Blondel V D, Guillaume J-L and Lambiotte R, 2008, Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. DOI: 10.1088/1742-5468/2008/10/P10008

Bruggmann A and 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. *Proceedings,* International Conference on Geographic Information Science, AGILE 2014, Jun. 3-6, 2014, Castellón, Spain. ISBN: 978-90-816960-4-3.

Derungs C and 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

Fabrikant S I and Skupin A, 2005, Cognitively Plausible Information Visualization, In: Dykes J, MacEachren A M and Kraak M-J (eds), *Exploring Geovisualization*, 667-690.

Hecht B and Raubal M, 2008, GeoSR: Geographically Explore Semantic Relations in World Knowledge. In: Bernard L, Friis-Christensen and Pundt H (eds), *11th AGILE International Conference on Geographic Information Science*.

Historical Dictionary of Switzerland (HDS), 2014, http://www.hls-dhs-dss.ch/ (April 2014).

NWB Team, 2006, Network Workbench Tool 1.0.0. http://nwb.slis.indiana.edu (April 2014).

Strötgen J and Gertz M, 2013, Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2): 269-298.

swisstopo, 2014, SwissNames. http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html (April 2014).

Tobler W, 1970, A Computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240.