ParHistVis: Visualization of Parallel Multilingual Historical Data

Aikaterini-Lida Kalouli* and **Rebecca Kehlbeck*** and **Rita Sevastjanova***
and **Katharina Kaiser** and **Georg A. Kaiser** and **Miriam Butt**
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

The study of language change through parallel corpora can be advantageous for the analysis of complex interactions between time, text domain and language. Often, those advantages cannot be fully exploited due to the sparse but high-dimensional nature of such historical data. To tackle this challenge, we introduce ParHistVis: a novel, free, easy-to-use, interactive visualization tool for parallel, multilingual, diachronic and synchronic linguistic data. We illustrate the suitability of the components of the tool based on a use case of word order change in Romance *wh*-interrogatives.

1 Introduction

Historical linguistics has begun to work with parallel corpora, exploiting the advances in corpus linguistics that facilitate the creation, linkage and analysis of large data sets. For some discussions as to the advantages of using parallel corpora see, e.g., Wälchli (2009); Enrique-Arias (2013). Aspects that stand out are: a) the direct comparability of concrete examples across time periods; b) the ease of analysis due to the known structure of the base text which makes it possible to look selectively at a small number of passages in which relevant structures are likely to occur; c) the facilitation of analysis of languages for which the researcher has no deep knowledge, based on the better known languages. Despite these advantages, it is challenging to use parallel corpora with state-of-the-art statistical/learning methods, because such data is often a) too sparse; b) but too large and too high-dimensional for manual inspection; c) a learning approach necessarily reduces the dimensionality so that important aspects that could in principle be gained from the use of parallel texts are lost. For our study on word order and language change in Romance *wh*-interrogatives, our goal is to investigate the strict word order observed in Old Romance (Kaiser, 1980; Schulze, 1888; Lapesa, 1992) and the more flexible word order of Modern Romance (Ordóñez, 1997; Rizzi, 2006), based on a parallel corpus of French and Spanish Bible translations of the 12th, 16th and 20th centuries. In particular, it is of interest to determine what factors might interact to determine word order, e.g., particles or the type of interrogative pronoun (Bayer and Obenauer, 2011). However, the relatively small size of our corpus for statistical methods but large size for manual investigation of the interacting factors and the unsuitability of existing visualizations for the inspection of parallel, multilingual, diachronic texts pose a challenge.

To tackle this challenge, but also to assist researchers with similar issues, we designed ParHistVis (Parallel Multilingual Historical Visualization). ParHistVis is a novel, freely-available,[1] easy-to-use, interactive visualization tool for parallel, multilingual, diachronic and synchronic data of a) the same time period across languages; b) of different periods of the same language; c) across languages. The tool employs methods of the field of Visual Analytics (VA) (Keim et al., 2008) and Computational Linguistics. It is suitable for researchers with little or no experience with computational approaches: after defining an input data file, they can directly interact with the visualization. Thus, our contributions are twofold: first, we present an easy-to-use, freely available, interactive tool suitable for the visualization of parallel multilingual data. Concretely, we show what aspects of parallel data can be efficiently explored using streamgraphs and Sankey diagrams. Second, we describe how the tool can be used via a concrete use case: the investigation of word order change in Romance *wh*-interrogatives.

*The first three authors had an equal contribution.

[1]The tool is available under https://typo.uni-konstanz.de/parhistvis/

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 109–114
Florence, Italy, August 2, 2019.

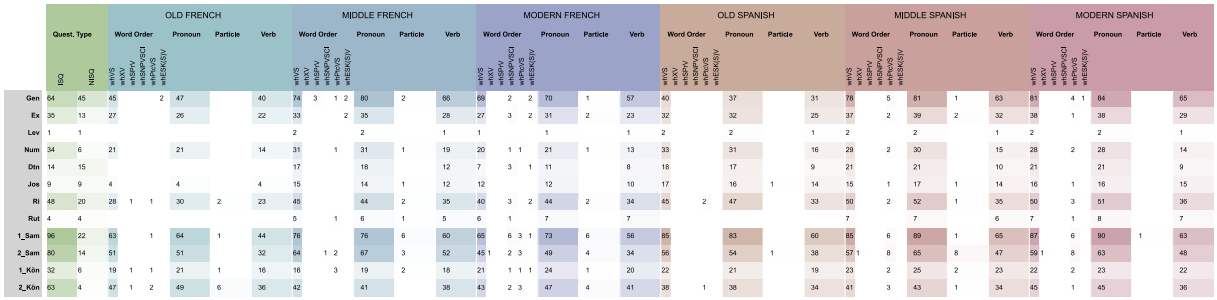


Figure 1: The aggregated matrix view of the books of the Old Testament across time periods and languages.

2 Relevant Work

Visualization as a means of illustration has a long tradition in linguistics, e.g., through spectograms for sound waves, tables for paradigms or graphs and attribute-value matrices for syntactic information. Besides such traditionally established visualizations, recent years have seen the emergence of new visualization ideas coming out of the field of VA (Keim et al., 2008) for the analysis and representation of linguistic data (Sun et al., 2013; Liu et al., 2014; Gan et al., 2014). A considerable amount of research has specifically focused on the visualization of historical linguistic change. One strand of research has focused on the visualization of word meaning across time (Sagi et al., 2009; Rohrdantz et al., 2011; Hilpert, 2011; Tahmasebi and Risse, 2017; Jatowt et al., 2018), while others have approached the same area with state-of-the-art embeddings (see Kutuzov et al. (2018) for a review). Another strand of research has concentrated on visualizing diachronic information in historical dictionaries, e.g., Theron and Fontanillo (2015) and linguistic evolution within the discourse (Lyding et al., 2012). Other work has visualized syntactic historical change (Butt et al., 2014; Schätzle et al., 2017; Schätzle, 2018). This work situates itself in the middle of those approaches, attempting to present a general, easy-to-use tool that can be employed for historical change of any kind (syntactic, semantic, etc.), but particularly targeting parallel, multilingual data.

3 The ParHistVis Tool

The tool works through a web-browser interface and is fully implemented in JavaScript. The only requirement is a tabulated file with the data to be visualized. The file can contain parallel, multilingual text, synchronic or diachronic, with each aligned piece of text (across languages or across time) associated with a row and identified by a unique ID. The rows can contain different columns, each of them encoding linguistic annotations that the researcher has assigned to that specific piece of text/row. In what follows, we call these linguistic annotations *dimensions* and their possible subcategories *features*. These dimensions can be specific to a particular language or time period or be associated with the whole row, i.e. the whole aligned text across languages and periods. The loading of the file in the tool is easy and fast: the researcher creates an online document of her file, e.g., a Google Sheets Document, and feeds its link in the provided field of the interface. This connects the document with the tool and the visualization instantly appears. This method is user-friendly and avoids complex handling of the documents usually found in a server-client environment. An additional merit is that the user can update the input document anytime and the changes will be automatically reflected in the visualization.

3.1 Parallel Analysis of Linguistic Change

One simple but essential requirement for the efficient study of parallel data is that the researcher can indeed observe each data point in a parallel way for each time period and, if multilingual data is available, for each language. Although this is possible with common tools like Excel, such a large document can quickly become overwhelming. To facilitate the direct comparability that parallel corpora enable, our tool builds upon this existing metaphor of a *matrix* visualization, as such a method preserves all dimensions of the data, in contrast to others which use dimensionality reduction techniques and crucial information gets lost. In this initial matrix view, the data follows the format of the input file but is structured in a colorful visualization: the languages and time periods are on the horizontal axis, allowing for interlingual and diachronic analysis of the data, and the course of the corpus is on the vertical axis, allowing for

intralingual and synchronic comparisons (see Figure 1). Each time period of each language is assigned a different color. The user can choose to filter a subset of the data, i.e. the features she is interested in, by selecting the corresponding columns. These columns will be automatically highlighted and the rest of the dimensions will be blended out to enable a more focused view on the data. In this detailed view, the user can observe the data in a qualitative way. For example, she can hover over the ID of a row and get the specific text associated with that ID, for each separate time period and language. General trends concerning the whole corpus can also be observed by zooming out on the matrix, e.g., we could observe that the filtered features appear only in the second half of the corpus (on the vertical axis) or only in the later time periods (on the horizontal axis).

3.2 Aggregated View of Linguistic Change

Although the detailed matrix view is suitable for inspecting individual data points, it does not facilitate general quantitative observations for the whole corpus or another natural grouping of the data. But these observations are of interest when comparing the same text across time and languages. We therefore offer an aggregated matrix view. Here, the user can select which data points should be aggregated; the tool offers a standard aggregation option but also makes educated guesses for other natural groupings of the data. The standard option is the aggregation of all data points of the whole corpus. Other aggregation options are offered based on the unique IDs: the tool searches for any reasonable pattern occurring in the IDs and suggests this as a natural aggregation, e.g., IDs with the same prefix will be aggregated to a group. The aggregation function merges all values of each feature of the subgroups contained in the aggregation and calculates their sum (Figure 1). Additionally, a colormap encodes the frequency of the features: the lighter the color of a given feature, the lower its frequency across the aggregated dimension; the darker the color, the higher the frequency.

3.3 Streamgraphs for Pattern Recognition

The aggregated view is suitable for general quantitative observations. Nevertheless, through the aggregation the features of the categorical dimensions are collapsed and thus interesting patterns may disappear. Moreover, the summed aggregated numerical dimensions can be overwhelming

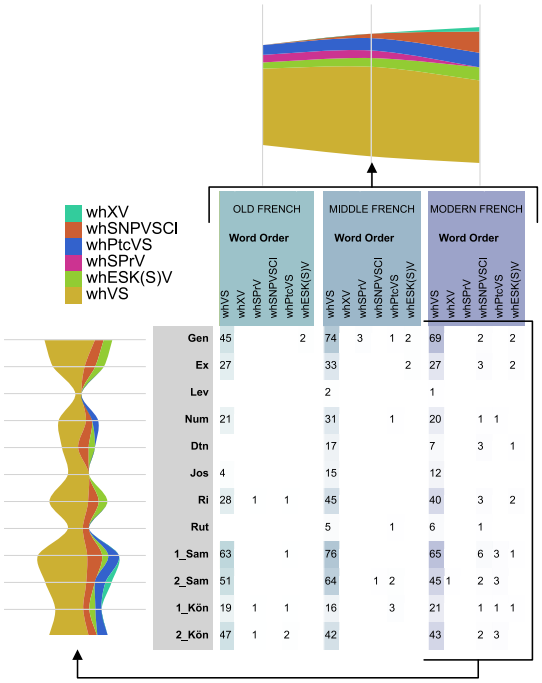
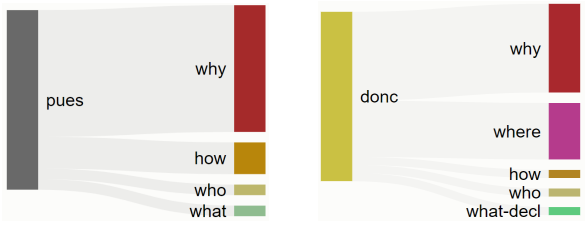


Figure 2: The streamgraphs of word order in French across time (top right) and of word order in Modern French across the aggregated Bible books (bottom left).

when trying to identify patterns. To tackle these drawbacks, the aggregated data is further visualized through *streamgraphs*. A streamgraph (also known as ThemeRiver) is a type of stacked area graph displaced around a central axis, resulting in a flowing, organic shape. Streamgraphs were popularized by Byron and Wattenberg (2008) for movie box office revenues but were already used for topic modelling by Havre et al. (2002) and have been applied to prosody visualization (Martin et al., 2010). Streamgraphs are commonly used to show changes of different categories across a single dimension, e.g., time, where categories might appear or disappear at different times. The height of each individual stream shows how its value has changed over time and the length shows its duration. This allows a comparison of the width of individual features visually, highlighting trends and outliers. Colours are used to differentiate between categories. Such high-dimensional data could also be represented by Parallel Coordinates (Inselberg, 1985), if time and space were considered "simple" quantifiable dimensions. However, as highlighted by Kehrer and Hauser (2013), the independent dimensions of time and space tend to play a central role in spatio-temporal data and should thus be considered independently. The properties of streamgraphs are thus suitable for parallel his-



(a) Middle Spanish (b) Modern French

Figure 3: The Sankey Diagram of the interactions between particles and interrogative pronouns.

torical data. To the best of our knowledge, this is the first work to apply streamgraphs to parallel, multilingual historical data. In ParHistVis, for a selected dimension, two streamgraphs are displayed: a) one for the frequency of the features of this dimension over the aggregated dimension for the specified time period (Figure 2, bottom left) and b) one for the frequency of the features of the dimension over the time periods, if this feature exists in a diachronic scale (Figure 2, top right). By hovering over a stream, the exact frequency of the feature visualized is displayed.

3.4 Sankey Diagrams for Pattern Interaction

Although streamgraphs offer a useful at-a-glance overview of the frequency of the dimensions, they cannot provide any insight into potential interactions between them. However, in the study of language change it is crucial to be able to discover such interactions, as most changes are the outcome of a series of interacting factors. Specifically, in parallel data there is a need for comparing how a concrete interaction has behaved across time or language. We make this kind of visualization available by incorporating *Sankey Diagrams*. These diagrams are traditionally used for visualizing (energy) flows; the entities under investigation are represented as nodes. The links among them are represented with edges with a width proportional to the importance of the flow. The diagrams have already gained attention in the digital humanities, e.g., in the visualization of migration flows and evolution (Abel, 2018), but also in literature (e.g., Campbell et al. (2018)). Here, the user selects the dimensions for which she wants to observe potential interactions. The features of these dimensions are depicted as nodes and the interactions between them as flows connecting them; the thickness of the flow shows the extent of the correlation. Again, a colormap helps the user distinguish between the interacting dimensions. An example of a Sankey Diagram is shown in Figure 3. The tool does not make predictions about potential interactions to display but lets the user define themselves the dimensions which might show an interesting interaction. This is especially useful for historical data where the data might be too sparse for the tool to be able to find any statistically interesting interactions but the user still wants to observe preliminary patterns and tendencies.

4 Use case

The visualizations above were obtained as part of our study on Romance *wh*-interrogatives. We used a subset of the parallel, multilingual corpus made available by Kalouli et al. (2018). This subcorpus contains three French and three Spanish Bible translations of the 12th, 16th and 20th centuries. We semi-automatically annotated this corpus for a) word order in interrogatives, b) interrogative pronouns and verbs of speaking introducing questions and c) particles used with interrogatives. The ultimate goal was to investigate the differences between the strict word order in Old Romance vs. the greater word order variation in Modern Romance, in correlation with interrogative pronouns, the introducing verbs of the interrogatives and particles. (Old and Modern) Romance languages are characterized by a relatively high stability with respect to word order in *wh*-interrogatives. They generally exhibit the fronting of the *wh*-phrase (*wh*-ex-situ) in combination with subject-verb inversion (*whVS*), so there is a strict adjacency between the *wh*-phrase and the verb. In Modern Romance, however, there is some variation with respect to these word order constraints. Many Modern Romance languages exhibit, mostly under very specific conditions, *wh*-ex-situ interrogatives without subject-verb inversion and allow for non-adjacency of the *wh*-element and the verb with certain elements. With this high-dimensional research question we are interested in a linguistic development within one language, as well as across different languages and time periods, with various interacting factors. ParHistVis can ideally assist us: although a detailed linguistic analysis is beyond the scope of this paper, we can show how the different views facilitate the study of this kind of data. With the color encoding in the matrix view in Figure 1, we can already make at-a-glance observations, e.g., there is a relatively

high number of interrogative pronouns in the beginning and the end of the corpus.[2] By using the streamgraph visualization (Figure 2) we can exactly inspect two phenomena attested in the literature: the emergence of complex inversion in Modern French (Roberts, 1993), i.e. the orange stream (*whSNPVSCl*) first appears in Middle French and increases its frequency in Modern French, and the diachronic non-adjacency of the *wh*-element and the verb when a particle is present, i.e. the blue (*whPtcVS*) stream stays stable over time. The visualization with the Sankey Diagram also offers interesting insights: arguably, some interrogative pronouns allow for more variation in the sentence structure, e.g. allow for particles (cf. e.g. Ordóñez (1997)). If we select to view the interaction of the interrogative pronouns and the particles in Middle Spanish and Modern French, Figure 3 shows us that the interrogative pronoun *why* allows for more frequent use of the particle *pues* and *donc* in Spanish and French, respectively, than other pronouns. Through the streamgraphs and Sankey Diagrams, similar observations can be made for other dimensions of the dataset. More importantly, the different available views allow the user to switch between them and inspect patterns that arise from these higher-level graphs. With this, the researcher can recognize and evaluate patterns in an otherwise too multifactorial dataset.

5 Conclusion

We presented ParHistVis, a visualization tool for parallel, multilingual, synchronic and diachronic linguistic data. We showed how the different views of the tool facilitate the inspection of the data, based on our study on word order change in Romance *wh*-interrogatives.

Acknowledgements

We thank the German Research Foundation (DFG) for the financial support within the projects P8 “Questions Visualized” and P2 “Word Order Variation in Wh-questions: Evidence from Romance” of the Research Group FOR 2111 Questions at the Interfaces at the University of Konstanz.

[2]For our particular use case this observation is not of great importance. However, for other studies of historical change, e.g. a corpus study on different genres, it is interesting to observe how specific patterns develop across those genres.

References

Guy J. Abel. 2018. Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015. *International Migration Review*, 52(3):809–852.

Josef Bayer and Hans-Georg Obenauer. 2011. Discourse particles, clause structure, and question types. *The Linguistic Review*, 28(4).

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehé, and Daniel A. Keim. 2014. V1 in Icelandic: A Multifactorical Visualization of Historical Data. In *Proceedings of the LREC 2014 "Visualization as Added Value in the Development, Use and Evaluation of Language Resources (VisLR)"*, pages 33–40.

Lee Byron and Martin Wattenberg. 2008. Stacked Graphs Geometry Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252.

Sarah Campbell, Zheng-Yan Yu, Sarah Connell, and Cody Dunne. 2018. Close and Distant Reading via Named Entity Network Visualization: A Case Study of Women Writers Online. In *Proceedings of the 3rd Workshop on Visualization for the Digital Humanities. VIS4DH*.

Andrés Enrique-Arias. 2013. On the usefulness of using parallel texts in diachronic investigations. Insights from a parallel corpus of Spanish medieval Bible translations. *Corpus Linguistics and Interdisciplinary Perspectives on Language*, 3:105–115.

Qihong Gan, Min Zhu, Mingzhao Li, Ting Liang, Yu Cao, and Baoyao Zhou. 2014. Document visualization: an overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(1):19–36.

S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20.

Martin Hilpert. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16:435–461.

Alfred Inselberg. 1985. The Plane with Parallel Coordinates. *The Visual Computer*, 1:69–91.

Adam Jatowt, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi, and Antoine Doucet. 2018. Every word has its History: Interactive Exploration and Visualization of Word Sense Evolution. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1899–1902.

Egbert Kaiser. 1980. *Strukturen der Frage im Französischen. Synchronische und diachronische Untersuchungen zur direkten Frage im Französischen des 15. Jahrhunderts (1450-1500).* Tübinger Textbeiträge zur Linguistik, 142. Narr, Tübingen.

Aikaterini-Lida Kalouli, Katharina Kaiser, Annette Hautli-Janisz, Georg A. Kaiser, and Miriam Butt. 2018. A multingual approach to question classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2715–2720. ELRA. ISBN: 979-10-95546-00-9.

Johannes Kehrer and Helwig Hauser. 2013. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513.

Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual Analytics : Definition, Process, and Challenges. In A. Kerren, editor, *Information Visualization*, pages 154–175. Springer, Berlin.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rafael Lapesa. 1992. La interpolación del sujeto en las oraciones interrogativas. In M. Ariza, editor, *Actas del II Congreso Internacional de Historia de la lengua espaola, Vol. I.*, pages 545–553. Pabellón de Espaa, Madrid.

Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. 2014. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393.

Verena Lyding, Ekaterina Lapshinova-Koltunski, Henrik Dittmann, and Christopher Culy. 2012. Visualising Linguistic Evolution in Academic Discourse. *Proceedings of the European Chapter of the Association of Computational Linguistics (EACL) 2012*, pages 44–48.

JR. Martin, M. Zappavigna, and P. Dwyer. 2010. *Visualising appraisal prosody*, pages 44–75. Continuum, London.

Francisco Ordóñez. 1997. *Word order and clause structure in Spanish and other Romance languages*. Ph.D. thesis, The City University of New York.

Luigi Rizzi. 2006. Selective residual v-2 in Italian interrogatives. In P. Brandt and Eric Fu, editors, *Form, Structure and Grammar. A Festschrift Presented to Gúnther Grewendorf on Occasion of His 60th Birthday.*, Studia Grammatica, 63, pages 229–241. Akademie Verlag, Berlin.

Ian Roberts. 1993. *Verbs and Diachronic Syntax. A Comparative History of English and French.* Kluwer, Dordrecht.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 305–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, pages 104–111.

Christin Schätzle. 2018. *Dative Subjects: Historical Change Visualized*. Ph.D. thesis, Universität Konstanz, Konstanz.

Christin Schätzle, Michael Hund, Frederik Dennig, Miriam Butt, and Daniel Keim A. 2017. HistoBankVis: Detecting Language Change via Data Visualization. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 32 in NEALT Proceedings Series, pages 32–39, Linköping. Linköping University Electronic Press.

Alfred Schulze. 1888. *Der altfranzsische direkte Fragesatz. Ein Beitrag zur Syntax des Französischen.* Hirzel, Leipzig.

Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749. INCOMA Ltd.

Roberto Theron and Laura Fontanillo. 2015. Diachronic-information visualization in historical dictionaries. *Information Visualization*, 14(2):111–136.

Bernhard Wälchli. 2009. Advantages and disadvantages of using parallel texts in typological investigations. *STUF - Sprachtypologie und Universalienforschung*, 60(2):118–134.