Interactive Text Visualization with Text Variation Explorer

Harri Siirtola, Poika Isokoski
TAUCHI / VIRG
School of Information Sciences
University of Tampere, Finland
harri.siirtola@uta.fi,
poika.isokoski@uta.fi

Tanja Säily, Terttu Nevalainen
Department of English
University of Helsinki, Finland
tanja.saily@helsinki.fi,
terttu.nevalainen@helsinki.fi

***Abstract*—Digitalization is changing how research is carried out in all areas of science. Humanities is no exception – materials that used to be hand-written or printed on paper are increasingly available in digital form. This development is changing how scholars are interacting with their material.**

We are addressing the problem of interactive text visualization in the context of sociolinguistic language study. When a scholar is reading and analyzing text from a computer screen instead of a paper, we can support this by providing a dashboard for reading, and by creating visualizations of the text structure, variation, and change.

We have designed and developed a software tool called Text Variation Explorer (TVE) for sociolinguistic language study. It is based on interactive visualization with a direct manipulation user interface, and aimed for exploratory corpus linguistics.

The TVE software tool has proven to be useful in supporting the study of language variation and change in its social contexts, or sociolinguistics. It is, to a certain degree, language-independent, and generic enough to be useful in other linguistic contexts as well.

We are now in the process of designing and implementing the next iteration of TVE. We present the lessons learned from the first version, discuss the old and the new design, and welcome feedback from the communities involved.

***Keywords*-Information visualization; text visualization; interaction**

I. Introduction

Text is a challenging data type to visualize. Firstly, text itself is a visual encoding, and it does not provide much to vary. In the spirit of Bertin [1], we can vary the size, color, orientation, style, and typeface of text, but this usually hinders the readability and spoils the reading experience. Secondly, text is not just a sequence of characters that always has the same meaning – a fragment of text detached from context may carry a completely different meaning, or might even be open to several interpretations. Finally, visualizations and visual summaries of text are often crafted to avoid seeing detailed data, which is unacceptable in many tasks involving text.

 doi:10.1109/IV.2016.57

A. Sociolinguistics

Sociolinguistics is the study of language in social and cultural context. Typical background variables in sociolinguistic studies include, e.g., the author's age, gender, ethnicity, domicile, social class, and the date of speaking/writing. Examples of quintessential research questions involve which social and linguistic factors influence language variation and change, how language change begins and proceeds, and the effect of the change upon linguistic structure and communication [19]. Also of interest is how personal and communal styles of speech and writing evolve in interaction. Research methods cover the full spectrum, including the information visualization approach [16], and often the research material is compiled into a corpus of representative texts.

B. Text visualization

Text visualization is a thriving subfield of information visualization. Kucher and Kerren [8] have recently made a survey of text visualization techniques and maintain an interactive, online browser of currently published techniques [17]. At the writing of this paper (the browser is constantly updated, initially there were 141 techniques) the browser lists and categorizes 272 text visualization techniques. Searching for linguistically-motivated text visualizations returns six techniques, including the first version of our tool.

Often we visualize data to avoid seeing it in detail, because the sheer volume of the data might render the close reading impossible. In sociolinguistic research, the close reading is an essential activity, and visualization provides a dashboard for the reader – not unlike the dashboard in a car.

This paper describes experiences gained from the development of *Text Variation Explorer (TVE)*, and presents our plans for the version we are currently developing.

II. Interactive text visualization

Text visualizations can be broadly divided into three classes based on how they combine the text and visualization. We can characterize them as direct, indirect, or hybrid visualizations. Please see Kucher and Kerren [8] for an extensive classification.

A. Direct text visualizations

The direct visualizations rely on the visual properties of text, in a 'bertinian' sense: size, color, location, orientation, style, value, and shape (pattern deliberately left out). A popular example in this class is the tag cloud (see [18] for the history), although it does not present the whole text, just the *n* most frequent words. Tag clouds have been criticized for using just the size to carry information — other visual variables have only a decorative function (color, location, orientation), and may mislead the reader. A classic direct text visualization is *SeeSoft* which used color to encode programmers in software systems and characters in novels [4].

B. Indirect text visualizations

The indirect text visualizations quantify some aspect of text, and visualize the numbers with a suitable technique. A popular approach is to compute term vectors or weights of each word in a document and visualize the vectors. This approach allows the use of, e.g., *Self-Organizing Maps* (SOMs) [7] or *ThemeScapes* [10] to create maps of the document or document collection. From the sociolinguistic perspective, these techniques serve as overviews, but lose connection to the detail.

C. Hybrid visualizations

Finally, the hybrid visualizations combine the text and the visualization in a meaningful way. This requires easy movement between the text, the corresponding spot in the visualization, and vice versa.

A significant portion of the text visualization tools have a 'zero-interaction' user interface, i.e., they are static. However, in an exploratory tool it is essential to be able to interact, to adjust visualization parameters and test ideas.

III. Text Variation Explorer

Text Variation Explorer (TVE) is a linguistically-oriented visualization tool for gaining insight into text [14], [15]. It was not designed to be a tool to run statistical tests (there are unsurpassed tools for that [11]), but rather a tool for quick exploration of text structure, complexity, and variation. It is a tool to raise questions rather than give answers. In the following, we describe the design goals, the features, and the use of TVE.

A. Language-independence

TVE was designed to be as simple and general as the intended task reasonably allows. One of the issues was how to keep this kind of a tool as language-independent as possible. We chose to limit the input into plain text (Unicode, utf8) and leave the definition of a 'word' to the user. What constitutes a word is defined by negation, by giving the set of characters that can't appear in a word (Fig. 1). The input is then parsed into words according to this set, and the resulting word count is displayed. These choices essentially limit the use of TVE to western, left-to-right written languages.

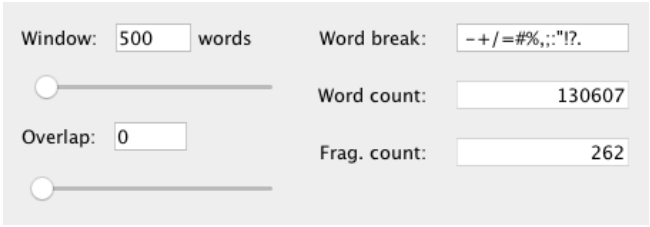

Figure 1. Definition of 'word' and setting of the sample size and overlap.

B. Size of text window

In corpus-linguistic analysis one of the important parameters is the window size of the text sample, which affects how some of the linguistic measures behave. Often the length of the sample window is set to a value that has been used in previous research (400 words is common), although it would pay off to explore a range of values. TVE allows a quick experimentation by providing a direct-manipulation slider both for the text window size and overlap (Fig. 1).

C. Measures

Besides language-independence, another important design goal in TVE is to keep it responsive, even with novel-length inputs and beyond. The linguistic measures were limited to the three most important ones that are highly useful when analyzing the structure, complexity, and variation of text. These measures are type-token ratio, hapax legomena, and average word length. The *type-token ratio* is the proportion of unique words in a sample, the *hapax legomena* is the proportion of words appearing exactly once in a sample, and the *average word length* as characters is self-explanatory. These three measures are displayed as a line graph (Fig. 2), and they describe the vocabulary richness and style of text. The first two of these measures are affected by the change of text window size, and they are known to stabilize around 1,300 words [6].

The interplay of these measures reveals to an expert interesting things about the text. As a trivial example, when all three measures have a low or high value, it may signal that the text type is *dialog* (conversation of two or more people) or *narrative* (representation of an event or a series of events), respectively.

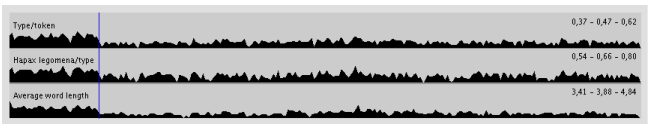

Figure 2. The linguistic measurements as a line graph: type-token ratio, hapax legomena, and average word length.

D. Text clustering

There is also a more generic method to explore the structure of text. The user can define a list of words that is used to compute frequency vectors for those words, per each text fragment (Fig. 3). When a text fragment is selected, the word list view will show the corresponding word frequencies.

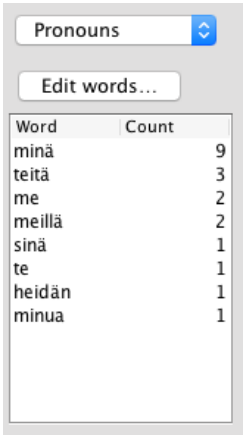

Figure 3. Frequency vectors.

Then the principal component analysis (PCA, the WEKA library [5]) is used to compute a user-defined number of text fragment clusters, based on the first two principal components. The result of the computation is shown as an xy-plot of the two first components, and the points representing text fragments are colored according to the assigned clusters (Fig. 4). The points of each cluster can also be represented with a minimal convex hull having a transparent color-coding. The word frequency vectors can also be exported from TVE to continue the analysis with other tools.

In Fig. 6 “Seven Brothers” by Aleksis Kivi, the national author of Finland, is read into TVE. If we define a list of Finnish Pronouns (Fig. 5), then we can cluster the text fragments into three according to them. The areas indicated are dominated either by ‘she/he’ (‘hän’), ‘they’ (‘he’), or ‘I’ (‘minä’). Based on close reading of the text fragments falling within each area, it seems that the ‘she/he’ area corresponds to narratives within the narrative of the novel, such as folk tales told by the brothers, while the ‘they’ area indicates narrative sections of the novel itself, i.e., what the brothers

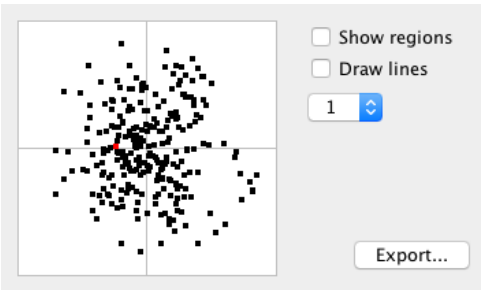

Figure 4. Principal Component View (PCA). Each point is a text fragment.

were doing, and the ‘I’ area indicates dialog. In a sociolinguistic corpus, differences in the use of personal pronouns might be interpreted, e.g., as different communicative styles employed by people representing different social groups.

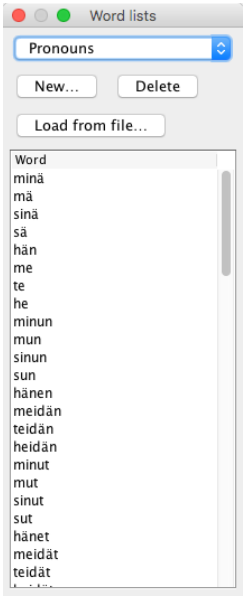

Figure 5. The list of Finnish pronouns, opened from the ‘Edit words...’ button in Fig. 3.

E. Interaction

TVE provides two methods to interact with the text. Firstly, there are sliders to adjust the text window size and overlap (Fig. 1). As the sliders are manipulated, the line graph view (Fig. 2) will be continuously updated, making it possible to review a large number of parameter settings in a short time. Secondly, there is a three-way brushing interaction between the data views (Fig. 7). When a text fragment is selected either in the text view, line graph, or the principal component view, all the other views are updated to show the same fragment. The text view will also scroll to the corresponding point to make the text fragment visible.

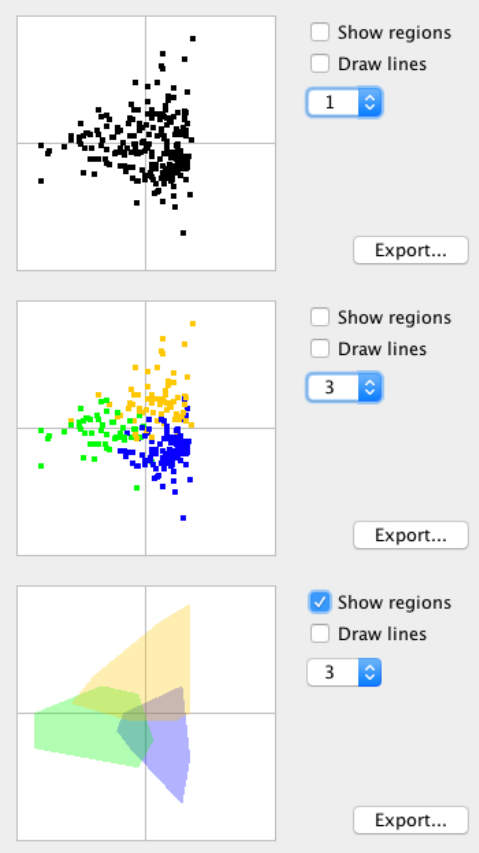

Figure 6. The "Seven Brothers" by Aleksis Kivi, clustered according to pronouns. The view of first two principal components (top), data clustered into three (middle), and clusters shown as regions (bottom). The text window size was set to 370 words.

F. Use cases

It is known that there is a gender difference in the use of pronouns in the Corpus of Early English Correspondence [3]. Women use more pronouns than men [12]. This sociolinguistic difference can be seen in TVE by clustering the text samples according to pronouns and dividing the text fragments into two. The first principal component signifies gender fairly accurately. Knowing this fact one might be interested to ask how homogenous is the language use of women writers. If we look at two historical figures, Dorothy Osborne (1671) and Lady Arabella Stuart (1605), the clustering according to pronouns does not suggest difference in use. However, if we cluster according to function words [2], we see that their use of language is quite different (Fig. 8).

TVE has also been used in comparing the different versions of International Corpus of English (ICE) although the current user interface is not really designed for this. TVE was able to point out similarities and differences at the level of corpus, genre, and subgenre [9].

IV. Discussion

We have seen that TVE can provide a quick overview of similarities and differences across corpora, highlighting sections that require more careful analysis. It can also be used to explore variation both across and within social categories. However, both of these functions could be enhanced by further development.

In the following, we have gathered the development ideas and issues we have collected from the users of TVE. We refer to the current version as 'TVE' and the new version as 'TVE2'. TVE has been demonstrated in seminars and conferences, and it has been used in teaching as well.

A. From texts to corpora

In the current TVE, the input is simply pasted into text view, and the design assumption was that we analyze a single text. Obviously, it is possible to paste several texts into text view, in conjunction, so we added a special non-word marker, 'dammocmark', that can be placed between texts, and which is shown as a blue line in the line graph (Fig. 7). However, this afterthought does not really solve the problem, because the line graph runs out of pixels to represent the text fragments. TVE2 will have a setup screen to define the files that a corpus consists of, and separate screens for each text and the corpus they form. Texts can be analyzed separately, and we can use a more suitable method to visualize the corpus.

B. Metadata and scatterplot

Sociolinguistic text materials are invariably described by metadata, such as author, year, gender, ethnicity, domicile, social class, data, etc. TVE2 will read metadata from a csv text file with a header. The header defines the names for metadata items and avoids fixing the set of metadata in advance. We add a general-purpose scatterplot to TVE2 where the user can set the x, y, size, and color variables. The available set of variables includes the metadata items and linguistic measures, so it will possible to create plots like 'hapax legomena of men over time'.

C. PCA and line graph views

Users find it confusing that changing the sample size just by one word may produce a completely different PCA view. This is due to the nature of principal component analysis, and is the correct behavior. The interpretation of PCA view can be simplified if we show the highest frequency word on top of the corresponding area (Fig. 9).

In addition, TVE2 will implement the vector space model [13] for text fragments and use PCA to show the similarities between corpus texts.

As noted earlier (subsection III-C), it is usually the interplay of linguistic measures that gives insight about the text. TVE2 will have an option to stack the line graphs, which will amplify the changes and make it easier to locate interesting spots (Fig. 10).

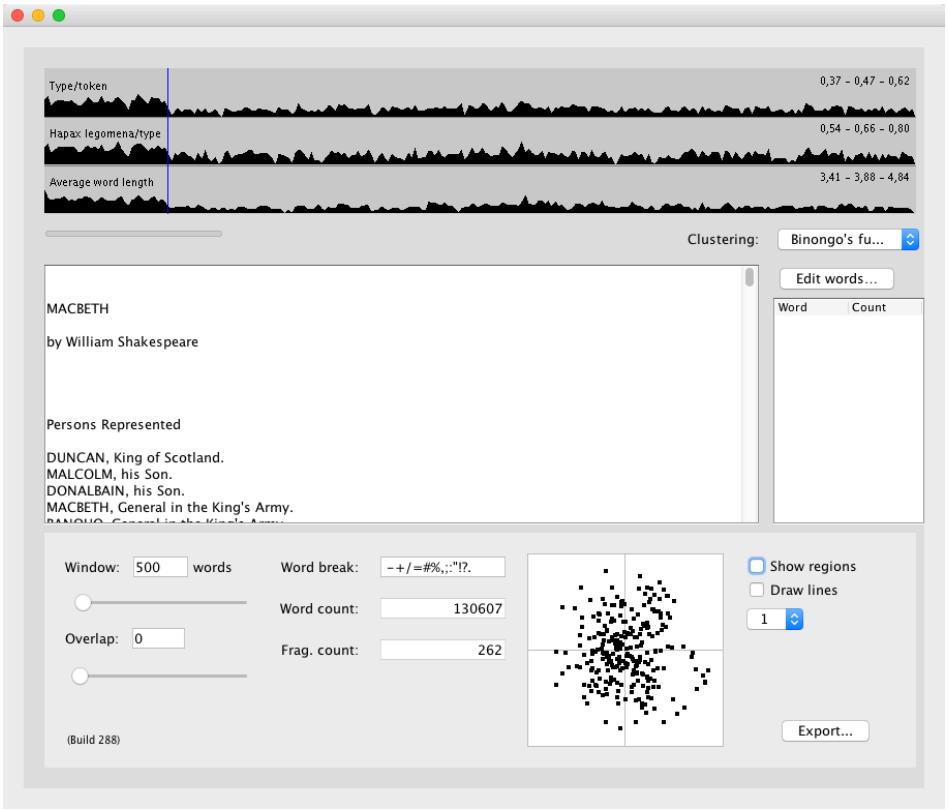

Figure 7. The user interface of TextVariationExplorer application.

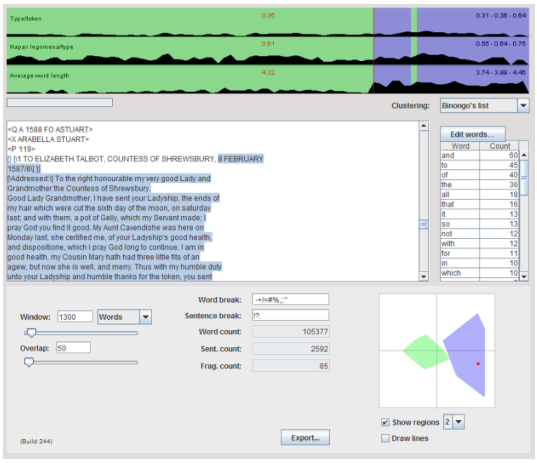

Figure 8. Clustering according to function words.

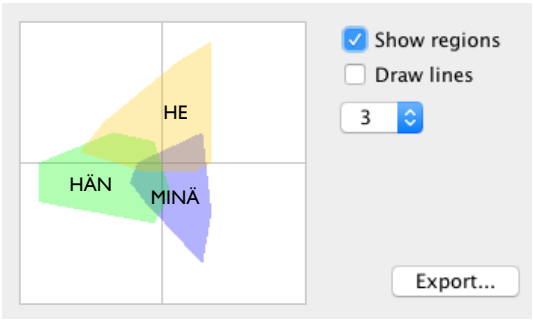

Figure 9. PCA clusters labeled with the highest-frequency word (continues the example of Fig. 4).

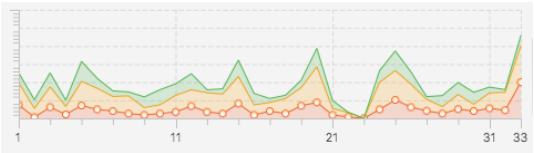

Figure 10. Linguistic measures as a stacked line graph.

D. Data export for analysis

TVE can export the frequency data of user-defined word vectors as csv files. TVE2 is able to export the complete text fragment data in R [11] format for further analysis.

E. Wildcards in word vectors

The current text clustering is based on the frequencies of words that are given literally, i.e., only the exact match is considered (subsection III-D). This is a problem with languages like Finnish where we have about 15 cases for nouns, which are expressed by suffixes. Adding a single noun to the wordlist would then mean inserting about fifteen

entries. In TVE2, the wordlist entries will allow wildcards.

V. Conclusion

We have presented and discussed the design of the Text Variation Explorer tool we are developing for exploratory corpus linguistics and for sociolinguistics. Both the current and the upcoming versions are freely available [14].

We would welcome any ideas or experiences about visualizing sociolinguistic data you may have, especially if you have tried our Text Variation Explorer.

Acknowledgments

This research was funded by the Academy of Finland, Digital Humanities Programme, project 'Interfacing structured and unstructured data in sociolinguistic research on language change (STRATAS)', sub-project #293441.

References

[1] J. Bertin, *Graphics and Graphic Information-Processing*. Berlin: Walter de Gruyter, 1981, 273 p. Translated by William J. Berg and Paul Scott.

[2] J. N. G. Binongo, "Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9–17, 2003.

[3] CEEC, *Corpus of Early English Correspondence*, Compiled by T. Nevalainen, H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi, and M. Palander-Collin, Department of English, University of Helsinki, 1998.

[4] S. G. Eick, J. L. Steffen, and E. E. Sumner Jr., "Seesoft – a tool for visualizing line oriented software statistics," *IEEE Trans. Softw. Eng.*, vol. 18, no. 11, pp. 957–968, 1992. [Online]. Available: http://dx.doi.org/10.1109/32.177365.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278.

[6] D. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *IEEE Symposium on Visual Analytics Science and Technology (VAST 2007)*, 2007, pp. 115–122. DOI: 10.1109/VAST.2007.4389004.

[7] T. Kohonen, *Self-Organizing Maps*, Third Extended, ser. Springer Series in Information Sciences. Springer-Verlag, 2001.

[8] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, Apr. 2015, pp. 117–121. DOI: 10.1109/PACIFICVIS.2015.7156366.

[9] T. Nevalainen and T. Säily, *Comparing like with like? tools for exploring families of corpora*, Talk at ChangE 2013, Helsinki, 2013. [Online]. Available: https://www.cs.helsinki.fi/u/tsaily/presentations/change2013_tn_ts.pdf.

[10] K. A. Pennock and D. B. Lantrip, "Themescapes: A landscape representation of themes in text," in *Symposium on Advanced Intelligence Processing and Analysis*, Tysons Corner, VA, USA, 1995, p. 47.

[11] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: http://www.R-project.org/.

[12] T. Säily, T. Nevalainen, and H. Siirtola, "Variation in noun and pronoun frequencies in a sociohistorical corpus of English," *Literary and Linguistic Computing*, vol. 26, no. 2, pp. 167–188, 2011. [Online]. Available: http://dx.doi.org/10.1093/llc/fqr004.

[13] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975. [Online]. Available: http://doi.acm.org/10.1145/361219.361220.

[14] H. Siirtola, *Text Variation Explorer (TVE)*, Computer program, 2011. [Online]. Available: http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve.html.

[15] H. Siirtola, T. Säily, T. Nevalainen, and K.-J. Räihä, "Text Variation Explorer: Towards interactive visualization tools for corpus linguistics," *International Journal of Corpus Linguistics*, vol. 19, no. 3, pp. 417–429, 2014. [Online]. Available: http://dx.doi.org/10.1075/ijcl.19.3.05sii.

[16] M. Stone, "Information visualization: Challenge for the humanities," in *Report of a Workshop Cosponsored by the Council on Library and Information Resources and the National Endowment for the Humanities*, Council on Library and Information Resources, 2009, pp. 43–56. [Online]. Available: http://www.clir.org/pubs/abstract/reports/pub145.

[17] *Text Visualization Browser*, 2016. [Online]. Available: http://textvis.lnu.se.

[18] F. B. Viégas and M. Wattenberg, "Timelines: Tag clouds and the case for vernacular visualization," *Interactions*, vol. 15, pp. 49–52, 4 2008. DOI: http://doi.acm.org/10.1145/1374489.1374501.

[19] U. Weinreich, W. Labov, and M. Herzog, "Directions for historical linguistics: A symposium," in, W. P. Lehmann and Y. Malkiel, Eds. University of Texas Press, 1968, ch. Empirical Foundations for a Theory of Language Change, pp. 95–188.