International Symposium on Visual Analytics Science and Technology (2010), pp. 1–6
J. Kohlhammer and D. Keim (Editors)

Comparative Visual Analysis of Cross-Linguistic Features

C. Rohrdantz[1] and T. Mayer[2] and M. Butt[2] and F. Plank[2] and D.A. Keim[1]

[1]Computer Science Department, University of Konstanz, Germany
[2]Linguistics Department, University of Konstanz, Germany

Abstract

So far Visual Analytics approaches have been developed for a wide array of research areas, mainly with focus on industrial or business applications. The field of linguistics, however, has only marginally incorporated visualizations in their research, e.g. using simple tree representations for the syntax structure of sentences. This paper suggests a new interesting field of application demonstrating how Visual Analytics is able to support linguists in their research. Therefore, one concrete linguistic phenomenon, named Vowel Harmony, is visually analyzed so that it can be compared across a variety of different languages. The developed approach covers the whole pipeline of Visual Analytics methods necessary to solve this tasks: data processing, feature extraction and the creation of an interactive visual representation. The outcomes are very convincing and allow to see at-a-glance whether vowel harmony is present in a language and indicate the vowel interdependencies that are involved in the harmonic process.

1. Introduction

Early approaches for the visualization of text data have been mainly focused on providing topical overview of document collections ([HPK95], [WTP*99], [Wis99]) as well as detailed topical insight into document collections ([Hea95], [FD00], [HHWN02], [DZG*07]), and are mostly related to the field of Information Retrieval. Since then, a lot of related visualizations have evolved. Some of the recent approaches are based on much more sophisticated text processing methods also involving some linguistic knowledge but still focus on topical content of text documents ([CCP09], [SOR*09]). In addition, approaches that are concerned with the visual analysis of affective content in large document collections (e.g. news articles, weblog entries or customer reviews) appeared recently in the field of opinion and sentiment analysis ([GPM*07], [GBB*08], [WRM*09], [OHR*09]).

While all of the mentioned work primarily deals with extracting and visualizing the topical content of text data, little work has been dedicated to the visual exploration of other text features and natural language phenomena. [HKLK97] have obtained visual syntactic category clusters by generating self-organizing maps based on word context vectors. [KO07] have extracted and visualized diverse statistical text properties on different hierarchy levels for literature analysis and authorship attribution, and [AC07] have extracted and visualized detailed text features to enable a visual classification of documents that is not only based on the topic content but also on style and sentiment. [WV08] created the "Word Tree" visualization that was primarily aimed at visualizing the content structure of texts but can also be used to visualize language features as shown by the example of a tree containing Greek nominal suffixes. Recently [AHM09] have introduced an interactive tool for the correction of erroneous machine translation output with visual components.

However, to the best of our knowledge there are no published approaches in the field of the visualization and visual analysis that try to visually compare a large set of languages with respect to linguistic properties. This paper is devoted to fill this gap by visually analyzing one exemplary cross-linguistic feature called Vowel Harmony for a large set of languages.

Vowel Harmony (VH) is an assimilatory phonological process by which vowels are pronounced in accordance (or harmony) with their environment (see [vdHvdW95] for an overview). Most often, preceding vowels trigger the shape of the vowel that follows them (see (6) for an overview), leading to a kind of domino effect within a certain linguistic domain Ů usually a phonological word. Languages differ as to whether they have harmonic processes or not and which features are involved, with closely related languages mostly sharing the same (or similar) features. A famous instance of VH is found in Turkish, where grammatical mark-

ers are pronounced differently in harmony with the preceding vowel. For example, the Turkish plural suffix is pronounced -ler or -lar depending on the last vowel of the stem. If the vowel has the feature [FRONT], i.e., if it is articulated in the front of the mouth, the plural marker is realized as ler (e.g., evler ŚhousesŠ, cÿ Íoler ŚdesertsŠ, ÍortĺulerŚcoveringsŠ, kediler ŚcatsŠ); if it is [BACK], i.e., if it is articulated in the back of the mouth the plural marker has the form lar (e.g., adamlar ŚmenŠ, toplar ŚballsŠ, komsÿular ŚneighborsŠ, kapýlar ŚdoorsŠ). However, most languages do not conform to VH and even VH languages always also show signs of disharmony. Besides, they differ with respect to how many and which features are active in the harmony process. In this paper it will be demonstrated how Visual Analytics can support linguists in detecting the degree and kind of VH involved in a language and readily compare different languages with respect to vowel harmonic processes. One important point is the automatic data analysis involving data preprocessing, statistical feature extraction and vowel ordering which are described in Section 2. In a consecutive step two matrix-visualizations are designed that help to track the probability and association strength of vowel successions within words and provide an insightful visual fingerprint for the vowel distributions in a language (see Section 3). Next, in Section 4 a case study is provided that shows that accurate hypotheses about VH can be derived from the matrix visualizations without any prior knowledge about a language. Finally, in Section 5 a conclusion as well as a research outlook is given.

2. Automatic Processing

2.1. Data gathering and processing

As data foundation for each investigated language a type list was compiled containing all different word forms appearing in the Bible. It is preferable to work on the Bible types instead of using a dictionary, because VH often occurs in inflected word forms. Moreover, it is preferable to work on type rather than on token level, because otherwise highly frequent tokens will bias the results. Taking the gathered list all vowel successions within types are counted and summed up. For this purpose, we define a vowel succession as a binary sequence of vowels within a word. Consecutive vowels have to be separated by at least one consonant, otherwise they will be ignored. To give an example, the word "harmonic" would contribute to the count of the vowel succession "o follows a" which we will refer to as (a-▷o) and to the count of the vowel succession (o-▷i). The resulting sums are saved in a matrix, an example is provided in Table 1.

2.2. Statistics

The simple matrix with the counts of vowel successions gives a rather general overview. Some high or low values are salient and usually it can be seen that some vowels appear with a much higher overall frequency than others. For most languages the strong variance between the overall frequencies of distinct vowels is the dominating effect visible in the matrix.

Method	Precision (Pos.)	Recall (Pos.)	Precision (Neg.)	Recall (Neg.)
Standard	69.83%	85.67%	88.59%	54.57%
Improved	71.28%	84.48%	86.90%	66.74%

Table 1: Example of a matrix with vowel succession counts for the Finnish Bible. The successions go from the row letter to the column letter. The succession (a-▷e) for instance occurred 1940 times.

In order to give some more detailed insight the succession probabilities are calculated. That means that if a certain vowel is observed, then it is calculated with which probability (in percent) certain other vowels are expected to be observed next. The values for succession probabilities are then saved in a probability matrix, analog to the matrix of absolute succession counts. Of course, still highly-frequent vowels in most cases have a higher probability of succeeding any other vowel than low-frequent vowels.

This leads us to apply a test for the statistical significance of deviations in the distribution of vowel successions. The aim is to find out if the deviation of an observed vowel succession from an expected vowel succession is statistically significant.

	e	not(e)
a	A = 1940	B = 7260
not(a)	C = 6354	D = 19861

Table 2: Example of the fourfold matrix for the succession (a-▷e) in Finnish. The expression "not(a)" stands for the set of all vowels except "a" and the same with "not(e)". Note that the four cells of the matrix have names (A, B, C and D) that are important for the formulas 1 and 2.

To get a significance value the fourfold χ^2 formula (see Formula 1, [Rum70]) is applied. The higher the values, the more significant in a statistical sense is the deviation of observed frequencies from expected frequencies. The test quantifies the influence of the independent variable (e.g. "a" in Table 2) on the dependent variable (e.g. "e" in Table 2).

$$\chi^2 = \frac{(A+B+C+D)\cdot(A\cdot D - C\cdot B)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \quad (1)$$

The χ^2 value depends on the sample size and therefore is not easily interpretable and comparable among sets of different size. To overcome this problem the correlation coefficient ϕ was applied (see Formula 2, [Rum70]).

$$\phi = \sqrt{\frac{\chi^2}{(A+B+C+D)}} \quad (2)$$

The ϕ coefficient represents the association strength and, when calculated directly from the fourfold matrix the ϕ values lie between -1 and +1, where a negative sign indicates a negative association among the two binary variables. Consequently, another matrix is created containing these association strength values, which we denote as ϕ matrix.

2.3. Matrix Arrangement

To make relations between similarly behaving vowels visible it is essential to sort the rows and columns of the matrices in a meaningful way. Only if a certain pixel coherence can be guaranteed will interesting vowel succession patterns become evident as motivated in [Ber83] and [HF06]. To enable a sorting of vowels first of all a numerical dissimilarity between vowels needs to be calculated. To do so, for each vowel a feature vector is created that corresponds to the ϕ-values of its matrix row. Next, a distance function between feature vectors has to be defined that quantifies the dissimilarity of the ϕ-vectors of two vowels at a time. Different distance functions were created and tested and the one that yielded the best results can be found in Formula 3. The rationale behind the formula is that pairs of vectors containing different signs at the same index are considered rather dissimilar.

$$dist(x,z) = \sum_{i=1}^{n} d(x_i, z_i), \tag{3}$$

$$\text{where } d(x_i, z_i) = \begin{cases} 1 & \text{if } sign(x_i) \neq sign(z_i), \\ (x_i - z_i)^2 & \text{else.} \end{cases}$$

The distance measure in Formula 3 is then used in the sorting process. The first row in the matrix of any language is fixed as the row belonging to the vowel with the smallest Unicode value (usually the vowel [a]). Then a nearest neighbor sorting is done: The most similar vowel row in the ϕ-matrix (vector in high-dimensional feature space) to the /a/-vector is searched and the corresponding vowel is placed in the second position. Next, to this second vowel again the most similar vector is searched among the remaining ones. This procedure is iteratively repeated until there is no vowel left.
After sorting the vowel rows, the vowel columns are sorted in exactly the same order. We also tried to sort columns and rows independently but came to the conclusion that this was not desirable as the diagonal of the matrix lost its general meaning (self-successions). Our tests showed that having the same row and column order is an important visual clue that helps in understanding the matrix and is more beneficial for the analysis process than an independent sorting of rows and columns.

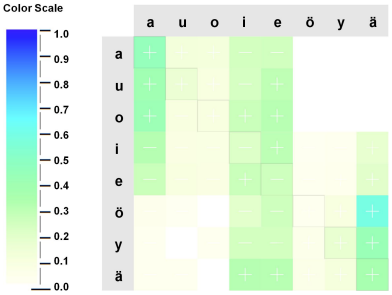

Figure 1: The visualization represents the probability matrix for the Finnish Bible types that has been sorted automatically. The "+" and "−" signs indicate whether a vowel succession occurred more or less frequently than expected when assuming vowel independence. One interesting finding that can be deduced from the visualization is that there are two blocks of vowels that almost never combine, viz. the block {a,u,o} and the block{ö,y,ä}.

3. Visualization and Visual Analysis

The numerical matrices generated with the analysis methods described in Section 2 were then transformed into visualizations for further analysis. Therefore, a straight forward visual representation was designed, maintaining the basic matrix metaphor and mapping the numerical entries to colors (see Section 3.1). Most importantly, the matrix rows and columns were sorted according to vowel similarity as described in Section 2.3 in order to make patterns become visible.

3.1. Data mapping and design

In the matrix with the succession probabilities all values inherently lie in the interval [0,1] and thus can be directly mapped to a color scale. In order to get many distinguishable color shades a bipolar color scale was chosen, ranging from bright yellow to dark blue (see Figure 1 for an example). For the matrix with the statistical association strength (ϕ) values of vowel successions two unipolar color scales were used. Vowel successions occurring more frequently than expected (positive ϕ) were colored in blue and vowel successions that were less frequently observed than expected (negative ϕ) got a red color. The higher the ϕ value was, the more saturated became the color. Because of the skewed data distribution with many values close to 0 a square root transfer function was applied. Thus, a larger color range was reserved for the densely populated area of low absolute ϕ values. See Figure 2 for the Finnish example.

Again, it has to be pointed out that a meaningful sorting of the matrix rows and columns is crucial for the visual analysis process. Figure 3 reveals that many interesting features are no longer clearly visible without sorting.

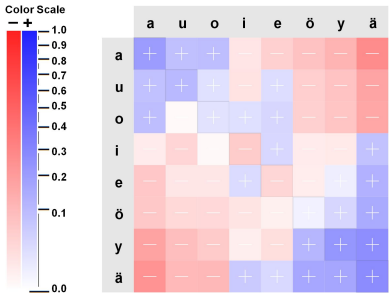

Figure 2: The visualization represents the ϕ matrix for the Finnish Bible. In this case the "+" and "−" symbols provide a redundant mapping. Now, blocks of vowels that belong together can clearly be seen. As before, {a,u,o} build one block, {ö,y,ä} another independent block, and {i,e} cannot unambiguously be assigned to one of them. In fact, this conforms nicely to the categorization linguists have for Finnish vowels: {a,u,o} are back vowels, {ö,y,ä} are front vowels, and {i,e} are neutral vowels, which explains that they do not adhere to one of the blocks.

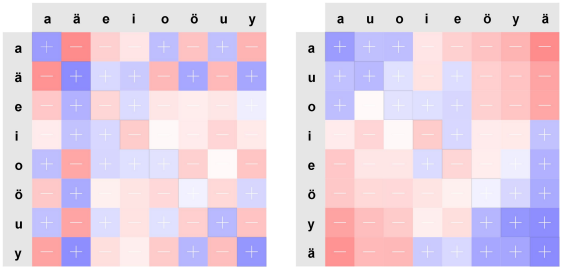

Figure 3: The left visualization has a default vowel sorting (alphabetical order) and shows no easily perceivable pattern at all. The right matrix which was automatically sorted, in contrast, reveals that there exists an interesting pattern.

3.2. Comparative Visual Analysis

When performing the described analysis for a large amount of different languages vowel harmonic patterns become easily visible (see Figure 4). Apart from Maori, all of the top 7 languages actually contain different kinds of VH. The strongly colored diagonal in Maori sticks out and can be explained by syllable reduplication which leads to a statistically salient amount of self successions. The strongest effect can be perceived in Turkish which is known to have rather strict and complex harmony patterns that become nicely visible here.

4. Case Study: Udihe

In this case study we want to examine to what extent our approach is able to help researchers in detecting VH and predicting the involved features. We therefore chose to investigate Udihe, a language that might be suspected contain

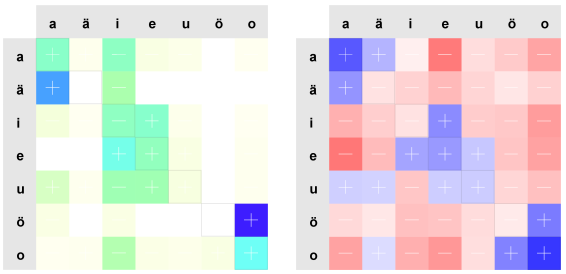

Figure 5: The probability matrix (left) and the ϕ-matrix (right) for the Udihe text fragment containing about 2450 words.

VH because it is related to other VH languages. Yet we did not know beforehand what kind of harmonic patterns are active in the language. We were able to get hold of a text with a length of 2450 words which according to previous experiments should be enough for detecting reliable patterns.
In order to generate a hypothesis about possible vowel harmonic patterns, first of all we must find out whether there is harmony present. We find three indicators for harmony:

- The average ϕ-value of Udihe (0.097) is the second highest after Turkish. This indicates that a strong effect like VH is present in the language.
- A look at the probability matrix (Figure 5, left) reveals that some successions are very probable and others very improbable which is a characteristic of vowel harmonic languages.
- There are blue blocks along the diagonal of the ϕ matrix as can be seen in Figure 5, right.

If a vowel succession is very probable and at the same time has a highly positive association (ϕ value) this is an indication for a harmonic pattern. Clearly, in Figure 5 this is the case for the transitions (a-▷a) and (ä-▷a) as well as (o-▷o) and (ö-▷o) as can be seen in Figure 5. As the vowel /i/ is very probable after any other vowel (except /ö/) it is very unlikely to be a successor within a harmonic pattern. In both matrices the same block in the /e/ column is salient and indicates the harmonies (i-▷e), (e-▷e) and (u-▷e). The hypotheses were derived from the visualizations without any prior knowledge about Uhide and the results are very satisfying as they correspond nicely to what grammarians find in their analyses [NT01, p. 74]. This example shows that it is possible to readily generate accurate hypotheses about VH in languages from such a visualization and without reading a single word.

5. Conclusions

Instead of using linguistics resources and methods for the visual data analysis of documents, we use visual data analysis methods on documents to analyze language. By analyzing human language, a new promising field for Visual Analytics

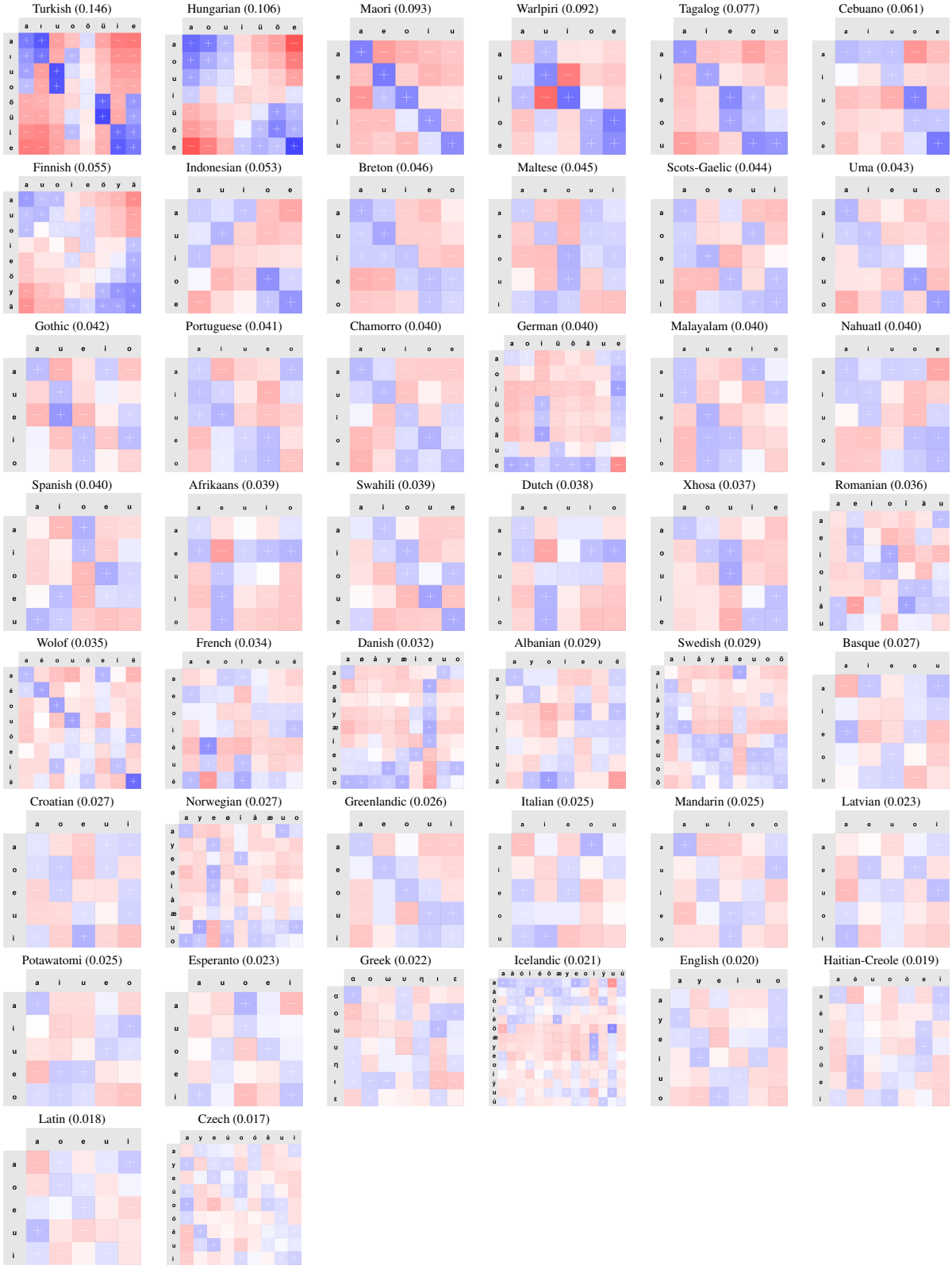

Figure 4: The ϕ matrices for 44 languages ordered according to decreasing average ϕ values (rounded average in parentheses) from left to right and top to bottom.

research is suggested. The aim is to support linguists in their sometimes pain-staking corpus work and provide them new perspectives on cross-linguistic data, allowing a deeper and more accurate insight, and possibly revealing yet unknown findings. This paper shows one beneficial example for such a collaboration by visually analyzing Vowel Harmony. With the applied statistics a first approximation for the degree of WH a language contains can be quantified. Second, the visualization enables linguists to compare a large set of languages with respect to VH patterns. Third, based on the visual impression from the ϕ-matrix and the probability matrix accurate hypotheses can be derived about which exact vowel harmonic dependences are involved. This is shown in a case study on the language Udihe, where VH patterns could be predicted accurately just based on the visualization.

References

[AC07] ABBASI A., CHEN H.: Categorization and analysis of text in computer mediated communication archives using visualization. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2007), ACM, pp. 11–18.

[AHM09] ALBRECHT J., HWA R., MARAI G.: The chinese room: Visualization and interaction to understand and correct ambiguous machine translation. *Computer Graphics Forum (also in 2009 Eurographics/IEEE Symposium on Visualization, Proceedings of) 28* (Jun 2009), 1047–1054.

[Ber83] BERTIN J.: *Semiology of graphics*. University of Wisconsin Press, 1983.

[CCP09] COLLINS C., CARPENDALE S., PENN G.: Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum (Proceedings of Eurographics, IEEE-VGTC Symposium on Visualization (EuroVis '09))* (2009), vol. 28, pp. 1039–1046.

[DZG*07] DON A., ZHELEVA E., GREGORY M., TARKAN S., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (New York, NY, USA, 2007), ACM, pp. 213–222.

[FD00] FEKETE J.-D., DUFOURNAUD N.: Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), ACM, pp. 47–55.

[GBB*08] GAMON M., BASU S., BELENKO D., FISHER D., HURST M., KÖNIG A. C.: Blews: Using blogs to provide context for news articles. In *2nd AAAI Conference on Weblogs and Social Media* (2008), American Association for Artificial Intelligence.

[GPM*07] GREGORY M., PAYNE D., MCCOLGIN D., CRAMER N., LOVE D.: Visual analysis of weblog content. In *International Conference on Weblogs and Social Media* (2007).

[Hea95] HEARST M. A.: Tilebars: visualization of term distribution information in full text information access. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1995), ACM Press/Addison-Wesley Publishing Co., pp. 59–66.

[HF06] HENRY N., FEKETE J.-D.: Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 677–684.

[HHWN02] HAVRE S., HETZLER E., WHITNEY P., NOWELL L.: Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 9–20.

[HKLK97] HONKELA T., KASKI S., LAGUS K., KOHONEN T.: Websom - self-organizing maps of document collections. In *Neurocomputing* (1997), pp. 101–117.

[HPK95] HONKELA T., PULKKI V., KOHONEN T.: Contextual relations of words in grimm tales, analyzed by self-organizing map. In *in Symbol ProcessingŠ, in Stefan Wermter and Ron Sun (eds) Hybrid Neural Systems* (1995), Springer.

[KO07] KEIM D. A., OELKE D.: Literature fingerprinting: A new method for visual literary analysis. In *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 115–122.

[NT01] NIKOLAEVA I., TOLSKAYA M. V.: *A Grammar of Udihe*. Mouton de Gruyter, 2001.

[OHR*09] OELKE D., HAO M., ROHRDANTZ C., KEIM D. A., DAYAL U., HAUG L.-E., JANETZKO H.: Visual opinion analysis of customer feedback data. In *VAST '09: Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology* (2009).

[Rum70] RUMMEL R. J.: *Applied Factor Analysis*. Nortwestern Univ. Pr., 1970, pp. 298–299.

[SOR*09] STROBELT H., OELKE D., ROHRDANTZ C., STOFFEL A., KEIM D. A., DEUSSEN O.: Document cards: A top trumps visualization for documents. In *InfoVis '09: Proceedings of the 2009 IEEE Symposium on Information Visualization* (2009).

[vdHvdW95] VAN DER HULST H., VAN DE WEIJER J.: Vowel harmony. In *The Handbook of Phonological Theory*, Goldsmith J., (Ed.). Basil Blackwell Ltd, 1995, ch. 14, pp. 495–534.

[Wis99] WISE J. A.: The ecological approach to text visualization. *J. Am. Soc. Inf. Sci. 50*, 13 (1999), 1224–1233.

[WRM*09] WANNER F., ROHRDANTZ C., MANSMANN F., OELKE D., KEIM D. A.: Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)* (2009).

[WTP*99] WISE J. A., THOMAS J. J., PENNOCK K., LANTRIP D., POTTIER M., SCHUR A., CROW V.: Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 442–450.

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1221–1228.