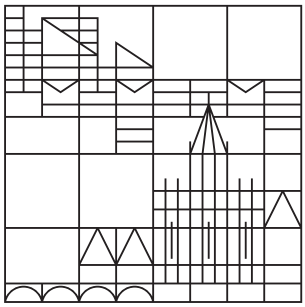
Visual Analytics of Change in Natural Language

Dissertation zur Erlangung des akademischen Grades
eines
Doktors der Naturwissenschaften

vorgelegt von
Christian Thomas Rohrdantz

an der

Universität Konstanz



Mathematisch-Naturwissenschaftliche Sektion
Informatik und Informationswissenschaft

Tag der mündlichen Prüfung: 03. Dezember 2013
1. Referent: Prof. Dr. Daniel A. Keim
2. Referentin: Prof. Dr. Miriam Butt
3. Referent: Prof. Dr. Marc H. Scholl

Abstract

This thesis describes novel computer science research on visual analytics methods for the detection and understanding of diverse phenomena of change that can be observed either within natural language text or based on it. The term *change* refers to the observable variation of features and patterns over time. In particular, two different kinds of phenomena are under research. The first part of the thesis deals with the diachronic change of linguistic features, namely *language change*. It includes pioneering work in the intersection of the disciplines of historical linguistics typological comparison of languages and visual analytics and contributes to the broader field of digital humanities or enhanced humanities (eHumanities). The second part of the thesis deals with visual analytics methods for the interactive detection and exploration of sudden unexpected changes in the information content conveyed by a large-scale text data stream. The research fills gaps in the previous work on time-related visual text analytics, demonstrates the commercial potential of such methods, and systematically outlines future research challenges for the live analysis and visualization of large-scale text data streams.

Zusammenfassung

Die vorliegende Dissertation beschreibt neuartige informationstechnische Forschungsansätze der Visuellen Datenanalyse, welche sich der automatischen Entdeckung und interaktiven Exploration verschiedener Arten von Veränderungen widmen, die sich in bzw. mit Hilfe von Textdaten beobachten lassen. Mit Veränderungen sind beobachtbare Variationen von Datenmerkmalen und -mustern gemeint, die sich über die Zeit hinweg ergeben. Dabei sind zwei verschiedene Arten von Veränderung Gegenstand der Forschung. Im ersten Teil der Dissertation werden historische Veränderungen sprachwissenschaftlicher Merkmale untersucht, sogenannter Sprachwandel. Dieser Teil leistet Pionierarbeit im Bezug auf die Forschung in der Schnittstelle zwischen den Gebieten des historischen und typologischen Sprachvergleichs und der Visuellen Datenanalyse und trägt damit auch zu einer Weiterentwicklung des weitergefassten Forschungsfeldes der Digitalen Geisteswissenschaften (eHumanities) bei. Der zweite Teil der Dissertation behandelt die interaktive Entdeckung und Ergründung von plötzlich und unerwartet auftretenden Veränderungen des Informationsinhalts eines großen Textdatenstroms. Diese Forschung füllt Lücken im vorherigen Stand der Technik zur zeitorientierten Visuellen Analyse von Textdaten, zeigt das Potential für eine wirtschaftliche Verwertung solcher Methoden auf und gibt einen systematischen Ausblick auf die zukünftig zu meisternden Herausforderungen für die Forschung im Bereich der Echtzeitanalyse und -visualisierung von großen Textdatenströmen.

Acknowledgments

First and foremost, I would like to thank my supervisors Daniel Keim and Miriam Butt for their advise, support, and encouragement. On the one hand they gave me the opportunity to pursue my own ideas and develop my own research agenda, while on the other hand they were always available to support me with their valuable advise. I was able to profit a lot from their expertise and enormous experience in research. I would also like to thank Marc Scholl for taking the time and effort of being part of the thesis committee.

Moreover, I would like to thank all the numerous students, PhD candidates, and senior researchers with whom I had the pleasure and honor to work with. Their names can be found throughout the different chapters of this thesis. Especially, my colleagues from the Data Analysis and Visualization Group and the Research Initiative "Computational Analysis of Linguistic Development" at the University of Konstanz contributed to my research through many inspiring research discussions.

I am also very grateful that I was given the opportunity to do research within an industrial context during my four stays at the Hewlett-Packard Research Labs in Palo Alto, California. I would like to thank Ming Hao, Umeshwar Dayal, Lars-Erik Haug, and Meichun Hsu for the fruitful and enduring collaboration.

Last but not least, I would like to thank the persons that contributed to this thesis in a very different manner: My parents Hildegard and Rüdiger, my brother Florian, and my partner Gitte who would always support me in my plans.

Contents

“Mir fällt nichts ein. Mir fällt etwas auf.” (Alfred Hrdlicka)[1]

[1]Seen in: Werkschau in der Kunsthalle Würth, Schwäbisch Hall, 2008

Chapter 1

Introduction

The topic of this thesis is:

*Computer Science research on **visual analytics** methods for the detection and understanding of diverse phenomena of **change** that can be observed either within **natural language text** or based on it.*

The following paragraphs will shed light on what that means in particular. Each paragraph refers to one of the key terms printed in bold font.

Natural Language Text / Written Language Today, there are about 7,000 known living human languages[1] also termed as natural languages, in contrast to artificially created languages such as programming languages. Natural language as such consists of different elements and accordingly linguistics, the research field that is dedicated to natural language, "is traditionally subdivided into phonetics, phonology, morphology, syntax, semantics, and pragmatics" [116, p.112]. From the natural language processing point of view, the different elements are briefly defined in literature [101, p.15] as follows:

i "*Phonetics and phonology - knowledge of linguistic sound.*

ii *Morphology - knowledge of meaningful components of the words.*

iii *Syntax - knowledge about the structural relationship about the words.*

[1] http://www.ethnologue.com last revised on March 6th, 2013

iv *Semantics - knowledge of meaning of words.*

v *Pragmatics - knowledge of how language is used to accomplish goals.*"

In principle, natural language is an auditory medium for communication, however, written language nowadays has also become an essential means for information exchange. The invention of writing systems marks a decisive step in the history of success of humanity. According to Nissen et al. [132, p.4] "the first appearance of writing was somewhere between 3500 and 2800 B.C.; the likeliest dating would place this emergence at ca. 3100 B.C.". First archaic writing systems were mostly aids to memory, non-linear, and consisted of pictographic and ideographic symbols, cf. [162, p.34]. They were the origin of more sophisticated systems and "in consequence of a series of fortuitous developments, the Latin alphabet has become the world's most important writing system" [54, p.7]. Such a mostly phonographic linear writing system consisting of letters that encode sounds (phonograms) has caused language to extend from being a mere auditory medium for communication to being a visual means for full-fledged natural language communication. Unlike other media for visual communication, e.g. symbols[2] or drawings, written language is suitable for conveying extremely precise and abstract descriptions, resulting in a huge number of consequences. To give just one example, how could a modern society work without written laws?

Through the invention of writing systems, abstract knowledge that previously had to be spread in a word-of-mouth manner face-to-face from generation to generation could be separated from the speaker. Today, advances in both education and technology have made written language a central part of our everyday lives, "writing is a skill practiced by about 85 per cent of the world's population" [54, p.7], and there remain only very few societies that do not make use of written language, cf. [162, p.11]. The number of people being able to communicate in different languages is also constantly rising and English has become a medium of communication used all over the world. Estimates say that 1.5 billion people spread over the whole planet speak English either as a first, second, or foreign language [59].

[2]Examples are piled stones or knotted strings (quipus), cf. [162, p.30]

Initially, the main advantage of written text was that it enabled asynchronous communication. A communicator could send a message to a recipient far-off in space and time and did not have to meet her/him in person. With the invention and spread of printing in the Late Middle Ages a large number of recipients could be reached with a minimized effort. The last technological breakthrough in the history of written communication, however, is a very recent one: On-line electronic communication enables real-time conversation and the massive spread of text messages at a scale that goes far beyond the reading and digestion abilities of individuals.

Change The term *change* in the context of this thesis refers to observable variation of features and patterns over time. The phenomena under research include diachronic change of linguistic features, so-called *linguistic change* or *language change* (see [76]), as well as sudden content changes in text streams.

One subject of research investigated within this thesis is natural language as a medium for communication. The main aspect of interest is that this medium is known to be prone to change, which is an important subject of study for *historical linguistics* (cf. [77]). Exploring language change is a challenging task, because the mechanisms involved can be quite diverse and have complex interactions. In recent literature, "these changes are not treated as phenomena amenable to explanation from a single source: they constitute a dynamic domain of complex, complementary, and correlated processes" [26, p.1]. Hock and Joseph [78] distinguish different types of "major linguistic changes that affect languages under all circumstances" [78, p.13]: *Sound change*, *analogy*, which "may have profound effects on word structure (also known as morphology)" [78, p.9], *semantic change*, and *syntactic change*. In addition, there is *change resulting from language contact*: "A number of other changes take place only when different languages (or dialects) are in contact with each other" [78, p.13]

In addition to this, there is another interesting aspect of change observable in text data. Not only does language, as a medium for communication, change over time, but also the information or content communicated changes. Changes in text content are a further subject of research investigated in this thesis. Over the last years the amount of information being communicated as

text, e.g. through the Web, has been increasing at a fast pace. On-line text streams constitute a rich body of information which is of interest for different real-world application tasks. For example, companies may get feedback on their products and services that may help them to monitor and improve the customer satisfaction. Especially, sudden changes in feedback content may point to previously unknown issues.
In summary, in this thesis both language and content change will be explored. In particular, four areas of research are identified, where visual analytics methodology is crucial to support domain experts to investigate complex subjects in large data sets. For all four areas I develop new visual analytics approaches that help to solve existing analysis problems.

1. Historical change of language that has occurred way back in history before language was recorded. These phenomena are explored by comparing today's languages (based on text) in the context of their genealogical and geo-spatial proximities (cf. [130]), which is the task of the linguistic subfield *typology*. Comrie [32] characterizes typology by two features: "1. it draws on data from a wide range of languages; 2. it is data-driven rather than theory-driven". Typology is thus a research field facing challenging data analysis issues and can potentially profit from advanced data analysis methodology as provided by the field of visual analytics. Details are given in Chapter 3 *Traces of Change: Cross-Linguistic Visual Analytics for Language Comparison*.

2. Historical change in lexical semantics. The research of *distributional semantics* is based on the assumption that the context of a word contains information about its meaning, and that investigating different contexts of a certain word over time can point to shifts in word meaning. For such investigations massive amounts of data are available and so far data-driven historical comparisons have rather rarely been performed, in contrast to investigations in the area of morphology and phonology. Investigating historical change in lexical semantics is challenging. It is difficult to pin-point the lexical semantic-content of a word, several different senses of one word may co-exist, the usage frequencies of different senses may change over time, new senses may come up, and established

senses may lose slightly in importance. In order to quantify, trace, and understand such changes in lexical semantics, massive amounts of data have to be analyzed automatically. Yet, automated processing is subject to inaccuracies, the quality of results is highly parameter dependent, and in some cases results may even contain systematic errors. This leads to a need to visually explore and interpret the analysis results, which are not only interesting for historical linguists, but also for lexicographers. Visualization offers possibilities for an explorative investigation of time-related multivariate data. Apart from the core distributional semantics analysis, visualization helps also to support the investigation of the distribution and spread of cross-linguistic phenomena over different languages, countries, and sources. Details are provided in Chapter 4 *Visual Analytics of Diachronic Change in Lexical Semantics.*

3. In the domain of online-communication sudden short-term changes in language use within a closed domain are likely to indicate a change in information content. The detection and analysis of such changes in content are challenging and relevant for many real-world application scenarios from business intelligence or public security. Often, changes in text content, e.g. changes in word frequency, word context or sentiment, indicate real-world events or issues that may be critical for analysts to detect. In this thesis the main focus is on the detection of sudden changes in large sets of time-stamped customer comments. Those changes may indicate emerging problems reflected by complaints about the product or service quality. Other related tasks for different analysis domains are treated as well. Novel methods for the retrospective analysis of past data archives are discussed in Chapter 5 *Visual Analytics of Diachronic Change in Text Content.*

4. In some application domains the detection of sudden changes in text content is also critical in real-time analytics scenarios. The consideration of real-time analytics brings additional challenges both for automated analytics as well as visualization. After a thorough discussion of these mostly unresolved challenges, novel methods for real-time text stream analysis will be discussed in Chapter 6 *Real-time Analytics and Visualization of*

	No time component	Historical analysis	Live analysis
Language change	Chapter 3	Chapter 4	-
Change in content	-	Chapter 5	Chapter 6

Table 1.1: Structured overview on the content of the different chapters.

Change in Text Content.

Table 1.1 gives an overview over the contents of the different chapters and their relation with respect to the subjects of investigation and the role of time.

Visual Analytics *Visual analytics* (see [167]) has been defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [168]. The visual analysis of text data is one subfield of visual analytics that has increasingly attracted attention in the recent years. In Ben Shneiderman's heavily cited *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations* [158], one of the rather early and most influential papers within the research field of *information visualization*, textual documents are seen as just one example for a one-dimensional data type, because text is "organized in a sequential manner" [158]. Only in 2002, text (including hypertext) became acknowledged as a data type on its own in another very influential data type taxonomy from Daniel Keim [88], because it "can not be easily described by numbers and therefore, most of the standard visualization techniques can not be applied. In most cases, first a transformation of the data into description vectors is necessary before visualization techniques can be used" [88].

Consequently, tracking the change of a text feature, content feature, or language feature requires computationally modeling, extracting, and measuring that feature. The better such a measure approximates the real phenomenon, the more accurately changes can be tracked. Examples for such measures can be *quasi-semantic properties* [90] of text as described in detail by Oelke [134] or language features as described by Wälchli [174]. In this work, I will build on existing measures, when available, and extend and refine them or develop new ones, where necessary.

There are several reasons why automatically extracted features and measures have to be conveyed visually for further interactive exploration:

1. Often, the analysis has an exploratory nature, that is, it is not quite clear beforehand what kinds of patterns can be expected in the data.

2. Often, features have to be put into the context of other features for exploration, because it is the interplay of the features that is of interest.

3. In many cases, for example, when investigating content change, interesting patterns of change may be automatically detected and preselected for further exploration. In the end, however, a human analyst has to verify the finding, gain understanding, and draw conclusions or generate hypotheses. To enable her/him to do so, an interactive access to the underlying text data has to be granted.

Content and Contributions of this Thesis In the recent past abundant computer science research has addressed the analysis of content features in text *without* considering time components. Yet, for the areas dealt with in this thesis (see Table 1.1) only a limited amount of previous related research exists. This is especially true for the visual analysis of linguistic features and the live analysis. In all areas the aim of this thesis is to fill gaps in the current research. In particular, this thesis summarizes research that aims to give answers to different research questions centered around phenomena of change in language as they can be observed in digitalized written texts:

- How can we support researchers from typology and historical linguistics in arriving at a better understanding about language change that happened before the invention of writing, based on the textual material available today? How can we support them in hypothesizing about causes and impact factors for such language change?

- What kinds of potentially ongoing language change can be tracked based on large amounts of written records from the more recent past?

- How can interesting content changes in on-line communication be detected and revealed to analysts? What kinds of real-world application problems can be solved with innovative visual analytics systems and what are the major challenges for future research?

To this end, research in the field of visual analytics for linguistic and time-oriented text analysis is conducted and described in this thesis. The main focus is on designing novel visual data analysis methods that support uncovering, understanding, and tracking *change* in natural language and language use as it can be detected in digital text collections.

The main high-level contributions of this work can be summarized as follows:

- The first part of this thesis (Chapter 3 and 4) contributes to the upcoming field of *digital humanities* in that it opens up a new area of research: The visualization of natural language data for linguistic research on language change. Novel methods are suggested for visually analyzing phonology, morphology, and lexical semantics and for the cross-linguistic comparison of language features that have been either extracted automatically or manually.

- In the second part (Chapter 5 and 6) novel techniques are suggested that enable the detection of interesting temporal bursts of text patterns independent from pre-defined aggregation intervals. This enables, for example, the detection of relatively high temporal accumulations of both generally frequent and infrequent terms with the same set of methods. It is demonstrated, that the underlying concepts are applicable to a wide range of time-stamped text resources and live text streams.

- The research summarized in this thesis has led to contributions for overview articles and surveys [89, 147, 150]. Chapter 6 discusses open challenges in real-time visual analytics of text data.

Parts of this thesis were published in different publications listed in the order of their appearance in this thesis:

- Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. Computer Graphics Forum, 31(3):935-944, 2012.

- Christian Rohrdantz, Thomas Mayer, Miriam Butt, Frans Plank and Daniel A. Keim. Comparative visual analysis of cross-linguistic features.

Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010), pages 27-32, 2010.

- Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank and Daniel A. Keim. Visualizing Vowel Harmony. Linguistic Issues in Language Technology, 4(Issue 2):1-33, 2010.

- Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt and Daniel A. Keim. Consonant Co-occurrence in Stems across Languages: Automatic Analysis and Visualization of a phonotactic Constraint. Proceedings of the ACL 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING 2010), pages 67-75, 2010.

- Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, Frans Plank. Towards Tracking Semantic Change by Visual Analytics. ACL (Short Papers) 2011: 305-310.

- Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. Lexical Semantics and Distribution of Suffixes - A Visual Analysis. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, 2012.

- Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, Daniel A. Keim: Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009), 2009.

- Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, Daniel A. Keim: Feature-Based Visual Sentiment Analysis of Text Document Streams. ACM TIST 3(2): 26, 2012.

- Ming C. Hao, Christian Rohrdantz, Halldór Janetzko, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Meichun Hsu. Integrating Sentiment Analysis and Term Associations with Geo-Temporal Visualizations on Customer Feedback Streams. SPIE 2012 Conference on Visualization and Data Analysis (VDA 2012), 2012.

- Ming C. Hao, Christian Rohrdantz, Halldór Janetzko, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, Meichun Hsu, and Florian Stoffel. Visual Sentiment Analytics of Customer Feedback Streams Using Geo-Temporal Term Associations. Information Visualization 12(3-4): 273-290, 2013.

- Christian Rohrdantz, Daniela Oelke, Milos Krstajic and Fabian Fischer. Real-Time Visualization of Streaming Text Data: Tasks and Challenges (Best Paper Award). Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek 2011, 2011.

- Daniel A. Keim, Milos Krstajic, Christian Rohrdantz and Tobias Schreck. Real-Time Visual Analytics for Text Streams. IEEE Computer 46(7): 47-55, 2013.

- Milos Krstajic, Christian Rohrdantz, Michael Hund and Andreas Weiler. Getting There First: Real-Time Detection of Real-World Incidents on Twitter. Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA, 2012.

In addition, there are a number of related publications that I was involved in, but that only indirectly contributed to the content of this thesis:

- Hendrik Strobelt, Daniela Oelke, Christian Rohrdantz, Andreas Stoffel, Daniel A. Keim and Oliver Deussen. Document Cards: A Top Trumps Visualization for Documents. IEEE Transactions on Visualization and Computer Graphics, 15(6):1145-1152, 2009.

- Daniela Oelke, Ming C. Hao, C. Rohrdantz, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug and Halldór Janetzko. Visual Opinion Analysis of Customer Feedback Data. Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST '09), pages 187-194, 2009.

- Slava Kisilevich, Christian Rohrdantz and Daniel A. Keim. "Beautiful picture of an ugly place". Exploring photo collections using opinion and

sentiment analysis of user comments. Computational Linguistics & Applications (CLA 10), pages 419-428, 2010.

- Daniel A. Keim, Daniela Oelke and Christian Rohrdantz. Analyzing Document Collections via Context-Aware Term Extraction. Proceedings of Natural Language Processing and Information Systems (NLDB '09), Springer Berlin / Heidelberg, pages 154-168, 2010.

- Christian Rohrdantz, Steffen Koch, Charles Jochim, Gerhard Heyer, Gerik Scheuermann, Thomas Ertl, Hinrich Schütze and Daniel A. Keim. Visuelle Textanalyse. Informatik-Spektrum, 33(6):601-611, 2010.

- Thomas Mayer, Christian Rohrdantz, Frans Plank, Miriam Butt and Daniel A. Keim. A Quantitative Approach to the Contrast and Stability of Sounds. QITL-4 4th Conference on Quantitative Investigations in Theoretical Linguistics, pages 59-64, 2011.

- Ming C. Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug and Meichun Hsu. Visual Sentiment Analysis on Twitter Data Streams (Poster Paper). Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST 2011), 2011.

- Christian Rohrdantz, Milos Krstajic, Mennatallah El Assady and Daniel A. Keim. What's Going On? How Twitter and Online News Can Work in Synergy to Increase Situational Awareness. Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA, 2012.

- Slava Kisilevich, Christian Rohrdantz, Veronica Maidel, and Daniel A. Keim. What do you think about this photo? A novel approach to opinion and sentiment analysis of photo comments. Int. J. Data Mining, Modelling and Management, 5(2):138-157, 2013.

- Christian Rohrdantz, Florian Mansmann, Chris North and Daniel A. Keim. Augmenting the educational curriculum with the Visual Analytics Science and Technology Challenge: Opportunities and pitfalls. In-

formation Visualization, SAGE, Published online before print April 11, 2013.[3]

- Andreas Weiler, Marc H. Scholl, Franz Wanner, and Christian Rohrdantz. Event Identification for Local Areas Using Social Media Streaming Data. In Kristen LeFevre, Ashwin Machanavajjhala, and Adam Silberstein, editors, Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks, DBSocial 2013: 1-6.

- Thomas Mayer and Christian Rohrdantz. PhonMatrix: Visualizing co-occurrence constraints of sounds. ACL (Conference System Demonstrations) 2013: 73-78.

- Andreas Lamprecht, Annette Hautli, Christian Rohrdantz and Tina Bögel. A Visual Analytics System for Cluster Exploration. ACL (Conference System Demonstrations) 2013: 109-114.

Finally, parts of the research presented within this thesis that I conducted during my four stays at Hewlett Packard Labs, Palo Alto, CA, have contributed to the filing and publishing of five US and two WO Patent Applications. To date, one of the patents has been issued.

The remainder of this thesis is structured as follows. Chapter 2 puts this thesis in context with the state of the art and related work and outlines which research gaps it fills. The Chapters 3, 4, 5, and 6, as already mentioned, detail on the novel research approaches and scientific contributions provided. Finally, Chapter 7 completes the thesis with concluding remarks and perspectives.

[3] http://intl-ivi.sagepub.com/content/early/2013/04/09/1473871613481693.abstract last revised on May 6th, 2013

Chapter 2

Visual Text Analysis

Contents

This chapter describes the current state of research with respect to the use of visual analytics in linguistic research (Section 2.1) and visual analytics in time-oriented text mining (Section 2.2). For both fields I will briefly describe the state of the art, identify open issues for research, highlight how the work presented in this thesis is embedded in these research areas, and point out which previously existing research gaps it fills.

2.1 Visual Analytics in Linguistic Research

The analysis of large text collections has emerged as a subfield of visual analytics only within the last few years. The goal of almost all approaches in this subfield is to enable analysts to gain insight into the topics and content

contained in large document or text collections. Over the years more and more sophisticated computational linguistic methods have been integrated into such visual analytics approaches. However, visual analytics approaches that have the aim of directly supporting theory-driven linguistic research are very rare. At the same time, until recently linguistic research has only marginally incorporated visualizations into its investigations. The work presented in this thesis is thus cutting-edge in terms of integrating visual analytics and linguistic research.

2.1.1 State of the Art

There are two centrally relevant works that have so far set the research agenda with respect to the integration of computational linguistics and visual analytics. One is a Ph.D. dissertation by Christopher Collins with the title *Interactive Visualizations of Natural Language* at the University of Toronto [30], and the other is a Ph.D. dissertation by Daniela Oelke with the title *Visual Document Analysis: Towards a Semantic Analysis of Large Document Collections* at the University of Konstanz [134].

Collins' Thesis: Collins coins the term *linguistic visualization divide* in his thesis, which refers to the "gulf separating sophisticated natural language processing algorithms and data structures from state-of-the-art interactive visualization design" [30]. Through five design studies he gives examples on how this linguistic visualization divide can be overcome combining "sophisticated natural language processing algorithms with information visualization techniques grounded in evidence of human visuospatial capabilities." [30]. Two of the design studies are meant to support computational linguistic research in the area of natural language processing suggesting an innovative use of visualization methods. The further design studies deal with content analysis and also with augmenting real-time computermediated communication.

Oelke's Thesis: Oelke describes a concept related to Collins' linguistic visualization divide. She states that for text analysis "Some semantic aspects are too complex to find good computational approximations. And even if we do have a good approximation, often, there still exists a gap between the computa-

tional efforts and the analysis goals" and concludes that "the analysis process therefore has to be designed in a way that the user can be incorporated to bridge the semantic gap" [134]. In order to enable this Oelke suggests a framework for analyzing document collections. The idea is to identify one or more semantic aspects of the text, called *quasi-semantic properties*, that are relevant for solving an analysis task. "This permits to targetly search for combinations of (measurable) text features that are able to approximate the specific semantic aspect" [134], these combinations are named *quasi-semantic measures*. Concrete implementations of the abstract framework and quasi-semantic measures are discussed and evaluated for the application areas of literature analysis, readability analysis, term extraction, and sentiment and opinion analysis. All examples include visual interfaces and visualizations that support the different steps of the analytic process.

Further approaches: Further related approaches using visualization to support linguistic tasks are rather sparsely scattered and have appeared at different venues of different research communities. On the technical level the approaches are quite diverse and can hardly be compared. For example, Honkela et al. [81] have obtained visual syntactic category clusters by generating self-organizing maps based on word context vectors. Later, Wattenberg and Viégas [178] 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. Further subfields of computational linguistics that have used visualization are machine translation [4, 40] and discourse parsing [188], where the output can be interactively explored and corrected. One of the very few examples where visualization is applied to investigate a phenomenon of language change was published in 2012 by Lyding et al. [112]. They use a parallel coordinates display to visually explore the distribution of modal verbs in academic discourse. Several academic disciplines can be compared for two points in time in order to detect changes.

2.1.2 Open Issues

Both theses mentioned in the previous subsection share the conclusion that combining computational methods with interactive visualizations enables text analyses that go far beyond what can be achieved with standard methods. Typically, analysis tasks come from the context of business, marketing, and security applications where large text collections have to be explored. Systems designed to support such analyses usually incorporate linguistic and natural language processing methods to achieve a higher analytic quality and grant deeper insight. In other words, visual text analytics profits from advances in (computational) linguistic research. In the case of Collins's thesis, also the contrary case is given: Visualization methods were used to support linguistic research and improve linguistic methods. However, the improved methods (machine translation and automatic speech recognition) belong to the subfield of computational linguistics, which naturally interfaces with other computational methods such as visualization. The computational linguistic researchers profiting from his novel visualizations, already brought a high affinity for computational methods: "Preliminary discussions revealed that they spent most of their time sitting at a computer, programming." [30].

Apart from Collins' work, the related work with respect to visual analytics systems that support linguistic researchers in their tasks is not quite advanced, often rather vague with respect to the tasks that shall be supported, and also rather superficial on a technical level. The most obvious gap in previous research is that there is a lack of visual analytics approaches that support subfields of linguistics that do not have a long tradition of performing computer-aided research. Still, manual data analysis is widely spread for example when it comes to the research of historical language developments and cross-linguistic comparisons. Manual data analysis is very accurate on a detail level, but is not scalable for the exploration of large data repositories. On the other hand, mere computational analyses, as they are usually performed within the research field of *corpus linguistics*, do not provide enough flexibility, because a concrete analysis proceeding has to be determined beforehand and the analytic process cannot be interacted with and guided on-the-fly. Thus, the insight is limited and the trustworthiness of results cannot be confirmed easily. At the same time, more and more linguistic data is becoming available

in digital format and waiting to be explored in-depth.

2.1.3 Goals of this Thesis

This thesis opens up a novel area of research that is about to become a new subdiscipline in linguistics and computational linguistics. In that sense some of the presented research is groundbreaking. The Chapters 3 and 4 push the previous state of research, introducing novel visual analytics approaches which support subfields of linguistic research that traditionally rely on manual analyses: Linguistic Typology and Historical Linguistics.

First, in Chapter 3 we integrate an extensive amount of available information about languages into one visual data analysis environment to support the field of areal typology in its research. We show how past language change in phonology and morphology can be traced and explored by performing cross-linguistic comparisons of multi-variate language features. To this end, in the first part of the chapter language features are presented both in areal and genealogical contexts. In the second part of the chapter, we present a novel matrix-based visual analytics method, that enables linguists to compare different languages with respect to complex features, i.e. vowel and consonant sequence patterns within words. We show that with our method, languages containing special sound patterns can easily be depicted visually based on processing limited fragments of text. For both parts of the chapter we show that it is important to arrange visualizations in a way that interesting visual patterns are likely to emerge. In the provided case studies we demonstrate that a meaningful spatial sorting or ordering of visual objects, based on their feature values, makes unexpected interesting patterns in the data visible.

In Chapter 4 we introduce novel computational methods for tracking and understanding change in lexical semantics coupled with interactive visual result representations. In the first part of the chapter we show that methods from topic modeling are well-suited to induce word senses from word contexts. The visualization is generated fully automatically from a large diachronic corpus and reveals the appearance of new word senses. This includes a description of the new word sense, the point in time when it appeared first and the frequency development over time in relation to other senses of the same word. In the second part of the chapter we suggest visualizations to investigate the dynamics of

the cross-linguistic spread of new coinages like words ending in the suffix *-gate*.

For feature extraction, in most cases we bear Oelke's framework in mind using concepts similar to her quasi-semantic properties in order to computationally model the linguistic research tasks.

The part of the thesis dealing with visual analytics for linguistic inquiry is cutting-edge in that it brings visual analytics research to a new application domain. Only after first common work had been published in 2010 and 2011, first workshops and conferences in this field have shown up and led to repeated citations of our work: The EACL 2012 Joint Workshop of LINGVIS & UNCLH Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources, April 23-24, 2012, Avignon, France[1], the AVML 2012 Conference on Advances in Visual Methods for Linguistics, York, United Kingdom, September 5-7, 2012[2], and the Workshop on the Visualization of Linguistic Patterns at the Annual Conference of the German Linguistic Society 2013[3], on March 12-15.

Interdisciplinary collaborations between computer science and humanities are an upcoming line of research termed *enhanced* or *digital humanities*. This thesis promotes such interdisciplinary research efforts bringing together visual analytics and linguistics. As part of the conclusions (see Section 7.2.1) also best practices, pitfalls, and lessons learned are discussed, which in turn may potentially be beneficial for other branches of the digital humanities.

2.2 Visual Analytics in Time-Oriented Text Mining

2.2.1 State of the Art

The time-oriented analysis of large text collections has become a productive area of research in the last years. In the following a survey of approaches for

[1] https://sites.google.com/site/lingvisunclh/ last revised on January 29th, 2013

[2] http://www.avml2012.wordpress.com last revised on January 29th, 2013

[3] https://dgfs.de/en/news/arbeitsgruppen-der-jahrestagung-2013.html last revised on January 29th, 2013

the visualization of document time series will be given, where a number of fundamental distinctions can be made:

- Approaches differ in that they either display meta-information or text content or both.
- Most but not all approaches have a visual representation for the timeline.
- There are different visual objects that can be displayed, mostly keywords, documents or topics.
- Most approaches also enable some sort of interactive exploration.
- Tasks and goals of the visualization of document time series can be quite diverse.
- Some approaches shall enable real-time analytics, which is especially challenging.

All of the mentioned distinction criteria will now be discussed in detail. In addition Table 2.1 and 2.2 give an overview which of the approaches, published within the last years, fall into which of the categories.

Meta-information

The vast majority of the visualization approaches for document time series focus on displaying the text content of documents over time. However, some approaches deal with document time series, but are not interested in the text content at all. Instead, they display different kinds of meta-information in relation to time, e.g. author information [171], citation information [28], tags and associated images from Flickr [50], the document structure [172] or the geo-spatial distribution of documents. Some more recent approaches integrate both meta-information and text content into small visual text analysis systems like iBlogVis [83], VisGets [45], Parallel Tag Clouds [31], TIARA [109,156], the Visual Backchannel [46], SensePlace2 [113], Discursis [9], and two systems for off-line event detection and exploration in social media [25, 48]. The focus of the following paragraphs will be on approaches working with the text content.

Timeline

When visualizing developments over time a frequent choice is to grant one of the valuable positional variables of the screen to the time dimension. Consequently, the majority of the approaches work with a timeline. Yet, some approaches use the two positional variables in order to distribute data items on a 2D plane and express similarities in the data space through spatial proximity on this 2D plane. While some approaches couple both a 2D display and a timeline display through linking and brushing others rely only on the 2D display, as can be seen in the columns *Timeline* and *2D Display* of Table 2.1 and 2.2. The latter is a common choice for live displays. When working without a timeline, often the time information is mapped to color hue [37, 74] or brightness [5, 7]. The only example of an approach integrating both a 2D distribution and timelines into one single view are the SparkClouds [104]. Yet, the approach does not have just have one single timeline, but one separate timeline for each data item displayed on the 2D plane. In Discursis [9] only the temporal order of items is preserved. The CloudLines approach [94] allows a logarithmic scaling of the timeline granting more space to the current data than to the past data.

Visual objects

It is common to all approaches that they rely on text documents as input data. However, there are different possible choices of what kind of objects to display. It can be distinguished between approaches that visualize keywords [5, 21, 37, 104], documents [94, 181] or topics. Of course, also any combination of the mentioned visual objects is possible and most approaches actually combine at least two different kinds of objects in their views. The most common combination is to display topics and in addition descriptive keywords for the topics, see Table 2.1 and 2.2. There are even approaches that show all three, i.e. documents within topic clusters for which descriptive keywords have been extracted [7, 74]. Other approaches track individual keywords, but interpret the keywords as topics [55, 71]. Finally, some approaches show documents and keywords without topic modeling [31, 36, 83, 170].

Interactive exploration

By far not all approaches address the issue of interactive exploration. However, some approaches have a special focus on interaction, e.g. most of the approaches relying on coordinated views, see column *Coordinated Views* in Table 2.1 and 2.2. Especially challenging is the interactive exploration in a live environment, as will be detailed later in Section 6.1. For further exploration, apart from text, sometimes additional data types are considered. Besides meta-information, like authorships, this includes images related to the text [46,50] or geo-spatial information about the text sources [25,31,45,48,113].

Tasks and goals

So far we have omitted the motivation that researchers have for visualizing document collections over time. While there is quite a number of approaches having a focus on tracking topics or events in news there are also diverse other purposes. Recently, the development of opinions or sentiments contained in text data [36, 41, 95, 156] and topic developments in scientific publications [35, 47] have received further attention. In addition, the development of discourse within conversations [9, 21] has been explored.

Most approaches have been designed with clear tasks in mind, and typically use cases and case studies provide evidence for the usefulness. One problem, however, from which especially those approaches dealing with news data suffer, is that real target users often were not available as test users and instead the authors had to simulate supposed expert exploration behavior.

Approach	*Meta-info*	*Text Content*	*Live*[1]	*Timeline*	*2D Display*	*Topics*[2]	*Keywords*[3]	*Documents*	*Coordinated Views*	*Year*
TimeMines [164]		+		+		+	(+)			2000
ThemeRiver [71]		+		+		(+)	+			2002
Stream MDS [181]		+	+		+			+		2003
Dynamic Discourse [21]		+			+		+			2003
AuthorLines [171]	+			+		+[4]				2004
HistoryFlow [172]	+[5]			+				+[5]		2004
TextPool [5]		+	+		+		+			2005
LiveIN-SPIRE [74]		+	+	(+)	+	+	(+)	+	+	2005
Themail [170]		+		+			+	+		2006
CiteSpace II [28]	+				+			+		2006
Flickr Tags [50]	+	(+)	+		+		+			2007
T-Scroll [85]		+	+	+		+	(+)			2007
NewsLab [58]		+		+		+	(+)			2007
Narratives [55]		+		+		(+)	+			2008
iBlogVis [83]	+	+		+			+	+		2008
VisGets [45]	+	(+)		+	+		+	(+)	+	2008
Meme-tracking [106]		+		+	(+)	+[6]	(+)[6]			2009
Story flow [151]		+	+	+		+	(+)			2009
ParallelTagClouds [31]	+	+			+		+	+	+	2009
TIARA [109]	+	+		+	+	+	+		+	2009
TIARA II [156]	+	+		+		+	+			2010
SparkClouds [104]		+		+	+		+			2010

[1] The column *live* contains only those approaches that have been designed and tested for live analysis. Potentially, further approaches could be extended to be applicable to live streams.

[2] In parentheses if single keywords are tracked as topic representatives.

[3] In parentheses if keywords are provided only additionally as topic/cluster labels or on demand.

[4] Newsgroup Threads

[5] Structure of a document

[6] Quotations

Table 2.1: Overview of the different approaches and their features.

Approach	*Meta-info*	*Text Content*	*Live*[1]	*Timeline*	*2D Display*	*Topics*[2]	*Keywords*[3]	*Documents*	*Coordinated Views*	*Year*
Vox Civitas [41]		+		+		+[4]	+	+	+	2010
ArticleThreads [95]		+	+	+		+	(+)			2010
Visual Backchannel [46]	+	+	+	+	+	+	(+)		+	2010
Context Preserving Word Clouds [37]		+		+	+		+		+	2010
Semantic Preserving Word Clouds [183]		+			+	+	+			2011
StreamIT [7]		+	+		+	+	(+)	+		2011
Discursis [9]	+	+		+	(+)	(+)	+	+		2011
TextFlow [35]		+		+	+	+	+		+	2011
CloudLines [94]		+		+			(+)	+		2011
ParallelTopics [47]		+		+	+	+	(+)	+	+	2011
SensePlace2 [113]	+	+		+	+			+	+	2011
EventRiver [111]		+	+	+		+	(+)			2012
TextWheel [36]		+	(+)	+	+		+	+		2012
IncrementalNews [97]		+		+		+	(+)			2012
Social Media Events [25]	+	+		+	+	+	(+)	+	+	2012
LeadLine [48]	+	+		+	+	+	(+)	+	+	2012

[1] The column *live* contains only those approaches that have been designed and tested for live analysis. Potentially, further approaches could be extended to be applicable to live streams.

[2] In parentheses if single keywords are tracked as topic representatives.

[3] In parentheses if keywords are provided only additionally as topic/cluster labels or on demand.

[4] Topic changes are marked along the timeline

Table 2.2: Overview of the different approaches and their features.

2.2.2 Open Issues

The column labels used in Table 2.1 and 2.2 were determined after a careful review of past research. They summarize the main dimensions in which previously published approaches can be distinguished. Some trends are revealed

when examining the tables: Whereas the first approaches either opted for having a timeline or a 2D display, most of the recent approaches integrate both through coordinated views. A less clear, but somewhat similar tendency is observable with respect to the displayed data objects (topics, keywords, documents). Early publications focus on one or two of them, while more recent works integrate all three options into one system. In conclusion, more and more information is displayed in the analysis systems to allow for increasing data volumes and increasingly complex analysis tasks. Yet, a clear gap is that while visualizations and interactions have become more sophisticated over the years, novel methods for automated text mining have only rarely been integrated into visual text analysis systems for both live and off-line analyses. Mostly, the automated analysis part is limited to topic modeling, which indeed is a quite advanced processing step, but it is mostly applied in a straightforward off-the-shelf manner, and the fact that it brings a lot of problems is usually omitted. Among others, the main issues are that especially smaller or only temporally present topics are likely to remain undiscovered and topic modeling is not applicable for live analysis. In general and especially recently, only very few live analysis systems have been presented. One explication could be that it is quite challenging to process and visualize today's large text streams in real-time so that actionable knowledge can be derived. Convincing solutions are still missing. More details on the live analysis are provided in Chapter 6.

Only recently, automatic event detection components have been integrated into time-oriented visual text analysis systems [25, 48], however, the systems are not real-time capable either.

Another issue of the current approaches is that even though time is often displayed as a continuous variable represented by a timeline, the data values are in most cases subject to aggregation along the time domain. The time variable is subdivided into so-called *time slices*, which in turn are not visualized explicitly. Almost all of the approaches sticking to the ThemeRiver metaphor share this issue, including the original ThemeRiver work [71]. Depending on the dataset, the documents are aggregated by time slices of one day, one month or even one year. Statistics are derived for each of the time slices, and the values are displayed in a way that adjacent time slices are connected "with smooth and continuous curves" [71]. While for some long-term developments

and sparsely scattered data such an aggregation and interpolation might be useful, this kind of visualization is certainly problematic in many real-world data analysis scenarios. The aggregation may lead to the disappearance of temporally local patterns and the interpolation at the same time mislead the analyst by not making the aggregation explicit.

2.2.3 Goals of this Thesis

In Chapter 4 we bring time-oriented text mining to the linguistic domain, introducing novel methods for the visual analysis of lexical-semantic change. In Chapter 5 a new visual analytics approach is suggested for the analysis of sudden changes in text content, more specifically sudden unexpected accumulations of negative user comments relating to the same issue. As part of this approach a new automatic event detection approach is coupled with a novel visualization that distorts the timeline and re-integrates exact temporal relations using a so-called *time density track*, in order to provide insight for unevenly-spaced text time series. Both the automatic and visual processing have the advantage that they do not require pre-defined exploration intervals, but discover temporal clusters independent from that. This also makes the analysis very interesting for live applications. Chapter 6 discusses which special challenges arise when analyzing text streams in real-time and shows how the methods from Chapter 5 can be extended to be applicable in a streaming environment integrating basic ideas and concepts from anomaly detection. It is demonstrated that the analysis can be performed in real-time consuming only limited storage resources, trigger updates only when new interesting issues come up, and be a good foundation for steered, importance-driven topic modeling.

In conclusion, the contributions of this thesis fill gaps both with respect to the integration of automated text mining into time-oriented visual text analytics systems and real-time text analytics.

Chapter 3

Traces of Change: Cross-Linguistic Visual Analytics for Language Comparison

Contents

Languages are complex systems that are prone to change and *language change* has always been happening. Albeit, only since language has been documentable through writing systems, and later voice recordings, language change can actually be observed and researched as it has happened. The further we go back in history, however, the less research material is available and it is often not very representative for language use in daily life.

Yet, information about language change before language was documented is not completely lost. Each of today's languages is the result of language change. By observing and comparing the different results, i.e. different languages, it is possible to infer past language change or at least to speculate about it (cf. [130]). For example, if many closely related languages share a certain feature it is an indication that this feature has been inherited from their common ancestor, a so called *protolanguage*, which "refers to the earliest form of a language family presupposed by all of its descendants. Reconstruction of a protolanguage is never secure." [124, p.18]. If one of the languages differs in a certain feature from all of its' closely related languages, there is a quite high probability that this feature may have undergone language change. Traditionally, linguistic researchers in the fields of *typology* and *historical linguistics* would analyze their data manually and base their research on observations. New corpus based approaches and automatization have then provided them with large amounts of multivariate data. Yet, they face the lack of suitable tools for solving their complex and specialized tasks.

Thus, these researchers have a need for visual analytics approaches in order to gain a better understanding of linguistic variation and language change by doing cross-linguistic comparison: "In typology, the aim is to discover constraints on variation across languages and principles that account for the observed variations, in order to establish general (and possibly universal) prop-

erties of human languages as well as the range of potential differences among languages" [56, p.10/11]. One methodology for reconstructing language history is the "comparative reconstruction based on corresponding forms in related languages/dialects" [124, p.5]. Section 3.1 introduces a novel visual analytics approach that enables researchers to compare multiple univariate language features in the context of language genealogy and geography. Section 3.2 describes a visual analytics approach that enables the comparison of complex language features.

3.1 Cross-Linguistic Comparison of Language Features in Genealogical and Areal Contexts

This section builds on the following publication:

Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. Computer Graphics Forum, 31(3):935-944, 2012 [146][1]

Some additional parts contained in this section have been submitted for publication and are currently under review.

This section presents a novel visual analytics approach that helps linguistic researchers to explore the world's languages with respect to several important tasks: (1) The comparison of manually and automatically extracted language features across languages and within the context of language genealogy, (2)

[1]Most of the publication was written by myself and I took the lead on the computer science research part of the paper, while Michael Hund did most of the programming work. Only some incremental parts were programmed by me, in particular the automatic sorting of languages. Bernhard Wälchli and Thomas Mayer contributed the linguistic knowledge, tasks, and findings. Daniel Keim gave advice on the project. Further people that we also acknowledge in the publication are Östen Dahl for help with bringing some of the N.T. data into an easily processable form, Ljuba Veselinova for help with the language data, and Michael Cysouw and Miriam Butt for valuable suggestions and comments. For all parts of the publication that were not written by myself I reference the original work. The inline images are also reprinted from our publication [146],

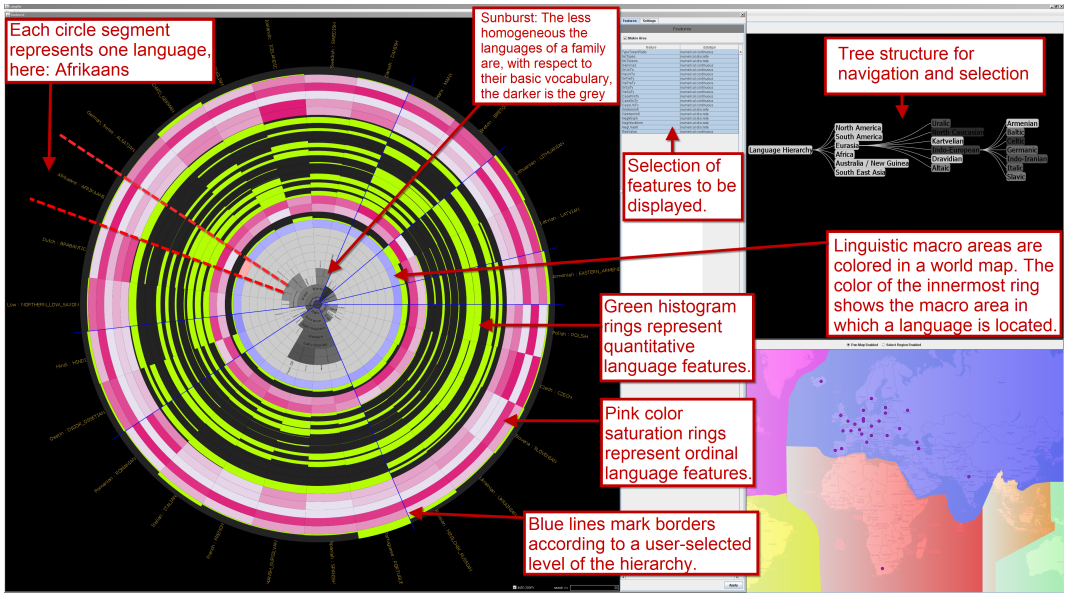


Figure 3.1: Overview on the core components of the system. Reprinted from [146], © 2012 The Eurographics Association and Blackwell Publishing Ltd.

the exploration of interrelations among several of such features as well as their homogeneity and heterogeneity within subtrees of the language genealogy, and (3) the exploration of genealogical and areal influences on the features. We introduce the World's Languages Explorer, which provides the required functionalities in one single visual analytics environment, see Figure 3.1. Contributions are made for different parts of the system: We introduce an extended Sunburst visualization whose so-called *feature-rings* allow a cross-comparison of a large number of features at once, within the hierarchical context of the language genealogy. We suggest a mapping of homogeneity measures to all levels of the hierarchy. In addition, we suggest an integration of information from the areal data space into the hierarchical data space. With our approach we bring visual analytics research to the application field, *historical comparative linguistics*, and *linguistic* and *areal typology*. Finally, we provide evidence of our system's performance in this area through a real-world application case study conducted by domain experts.

3.1.1 Background

> "*There are app. 6,900 modern natural human languages [...], many of them endangered or moribund. The comparative analysis of the*

world's languages is a considerable challenge, which is traditionally addressed from three different sides. Historical-comparative linguistics deals with language families (genealogically related languages) which derive from largely homogeneous reconstructed proto-languages, such as Indo-European, through structural divergence in language change. Areal linguistics investigates how intensive language contacts seduce languages to converge structurally in linguistic areas such as South East Asia or Mesoamerica. Linguistic typology explores the full range of linguistic variability in terms of structural features, such as word order and number of grammatical cases. While typology traditionally tries to explain the distribution of structural features with other structural features, modern research has shown that linguistic diversity is not randomly distributed over the world, but that there are macro-areal patterns of continental or even hemispheric size [...] which must be due to very old language contacts and/or genealogical relations that are not demonstrable with standard historical methods. This reunited the three disciplines in areal typology, which investigates typological, genealogical and areal properties in their interplay. [...] divides the world into six regions (macro areas) where massive language contacts are most likely to have occurred. Wherever features in genera - genealogical units with a time depth of app. 3,500 years, such as Germanic or Romance - are not distributed the same way across all six regions this is taken as evidence for a non-random distribution. Areal typology investigates among other things genealogical stability of features and their propensity to areal diffusion." [146][2]

In recent years an increasing number of manually edited language data has been created, digitalized, and made available to the public, some of which will be described in Section 3.1.3. An alternative to this time consuming procedure is to extract typological features automatically from parallel texts, i.e. translations of the same source text into different target languages [38]. Like that, languages can be directly analyzed on the level of language use without presupposing expert knowledge for the researched languages.

[2]Part of our joint publication written by Bernhard Wälchli. Original version contains further references.

Despite of the increasing availability of automatically and manually generated language features, until now, linguistic researchers have only marginally availed themselves of visualizations or advanced interactive visual interfaces for doing cross-linguistic comparisons and exploration. The World Atlas of Language Structures[3] offers a variety of language properties that are mapped to the geo-positions where the respective languages are spoken, more detailed information will be provided in Section 3.1.3. Another approach combines a world map with other visual representations for the analysis of meaning evolution [166]. Finally, the *Multitree* tool [186] enables the user to visually access information about language relationships displayed as a node-link tree diagram. Yet, no work exists that combines both a geo-spatial and a hierarchical representation and would allow a visual comparison of multiple features at once.
In this section, the goal is to describe a visual analytics system, the World's Languages Explorer, that enables the analysis of languages with respect to several research questions that domain experts have, such as:

- Are certain language features homogeneous within certain branches of the genealogy and diverse across different branches? This might be a trace of language change before written record.
- Are there any outliers, that is, languages where a certain feature value surprisingly deviates from that of other closely related languages?
- If so, is this outlier value similar to that of other unrelated, but geographically close languages? This might point to a language change that was triggered by language contact, which is of special interest to linguists.

More details are provided in Section 3.1.4.

This design study (see Figure 3.1) contains several contributions to the field of visual analytics: We display the language genealogy as a Sunburst visualization and complement it with our *feature rings* which allow a cross-comparison of several features at once, within the hierarchical context of the language genealogy. Feature rings have different representations depending on whether they display quantitative, ordinal, or nominal features. Moreover, we suggest

[3] http://wals.info last revised on March 6th, 2013

a mapping of homogeneity measures to all levels of the hierarchy. We also propose different means of integrating areal information into the hierarchical data space. A further contribution is that we bring visual analytics research to a new application field, namely *historical comparative linguistics*, and *linguistic* and *areal typology*.

The description is structured as follows: In Section 3.1.2, we give an overview over how this approach relates to other methods and techniques for visual data exploration. In Section 3.1.3 we briefly outline our automatic feature extraction and further data sources containing manually edited language features. Section 3.1.4 gives insight into the concrete tasks and requirements linguistic researcher have. In Section 3.1.5 we introduce our new system and give a detailed explanation of design decisions and our contributions. Section 3.1.6 next provides two application case studies showing real findings relevant to linguistic researchers. In Section 3.1.7 we discuss advantages and limitations of our approach and finally provide a conclusion.

3.1.2 Related Work

The core part of our approach is a visual display of the language genealogy, which is a hierarchical data structure. According to Shneiderman's terminology [158] our data has a *high fanout*, that is, it potentially contains thousands of leaf elements, but the leaf-root distance is usually very short. Different approaches for plotting hierarchical data will be discussed in this section. For our purposes we need to compare multiple language features across languages within and across different hierarchical categories. To the best of our knowledge so far no other approaches have been published that pursue this as a main goal. However, there are different approaches that plot relations among different nodes in a hierarchy, and approaches that combine geo-spatial information and hierarchical data.

Plotting hierarchical data

In the literature, several basic approaches for displaying hierarchical data can be found (1) Node-link tree diagrams, (2) Icicle Plots, (3) Treemaps, and (4)

Radial space-filling layouts.
(1) Node-link tree diagrams are the most intuitive and natural way of plotting hierarchies. In contrast to most other methods, node-link tree diagrams are not space-filling and many different layouts exist. Interesting extensions include the Hyperbolic Tree [102, 103], which lays out the hierarchy on a hyperbolic plane, and the three-dimensional animated Cone Tree [143]. The integrated change of focus interaction makes it a useful display for browsing large hierarchies. Node-link tree diagrams can also be plotted in a 3D space like in the case of the animated Cone Tree [143].
(2) Icicle plots [99] are space-filling rectangular versions of trees. The levels of the hierarchy are displayed as horizontal stripes from top to bottom and the elements of the hierarchy divide the stripes vertically into parts. Each hierarchy level requires the same amount of space as the leaf level.
(3) Both the concepts of nested and non-nested Treemaps have first been published by Johnson and Shneiderman [87, 157] and since then have become very popular and were extended in different ways for different purposes. A brief overview of the history of Treemaps by Shneiderman and Plaisant can be found online[4]. Interesting extensions include Ordered and Quantum Treemaps [18], Voronoi Treemaps [15], and Generalized Treemaps [173]. Treemaps grant almost the whole display space to the leaf nodes making internal nodes in non-nested Treemaps only visible as space separators, i.e.,in Treemaps the hierarchy is conveyed through containment. They show their strengths when the focus of analysis lies on the leaf nodes and especially when these leaf nodes have different sizes that are important to explore. The shapes of leaf nodes may differ considerably, even if they are all granted the same space.
(4) Radial space-filling layouts like the Information Slices [8] and later the Sunburst display [161] have the advantage that they do not grant as much space to the inner nodes as Icicle Plots, but still have a space-filling representation for them. Typically, the amount of nodes in a hierarchy increases with increasing distance to the root. In the Sunburst visualization the amount of display space available at each level of the hierarchy increases analogously.

Other visualizations for hierarchical data can be found but have not become as popular as the aforementioned ones. Examples include the Cheops system

[4]http://www.cs.umd.edu/hcil/treemap-history/ last revised on March 6th, 2013

[17] that emphasizes on browsing and exploration of complex hierarchies but not on the analysis.

Hierarchical and relational data

A visualization that integrates both hierarchical and relational data are Arc-Trees [128], a combination of a Treemap with an arc diagram. The Treemap grows only in horizontal direction and linking arcs connect two nodes of the hierarchy if they are related. Another technique with a similar purpose is Holten's Hierarchical Edge Bundles [79], which can be combined with different hierarchical visualizations. Elements in the hierarchy are connected with colored bundled links if they are related. Both ArcTrees and Hierarchical Edge Bundles, however, require the link space to be rather sparsely populated and do not scale for fully connected graphs. In addition, while it can be conveyed that two items are related, different types of relations are hard to express and the links are not suitable for performing feature comparison across elements in the hierarchy.

A further possibility of combining hierarchical data with relational clues is to provide a matrix display that shows relations in the cells. Either the axes elements of the matrix are leaf nodes of a hierarchical node-link tree structure as in the Matrix Browser [189] or the hierarchy is conveyed through a Treemap-like recursive subdivision of the matrix as in [169]. Fully connected graphs are not a problem for these latter approaches, but overall only a limited number of nodes can be displayed; otherwise the matrix will grow too big. For all of the mentioned techniques, there is no intuitive way to use the visualization for feature comparison.

Hierarchical and geo-spatial data

An approach that combines hierarchical and geo-spatial data are Flow Maps [24, 137]. Flow Maps lay a tree structure over a map in order to indicate geo-spatial movements (flows). The tree structure is the result of a hierarchical clustering of locations. Thus, the hierarchy directly depends on the geo-locations and is of a binary nature. In our scenario, in contrast, the hierarchy is predefined and not directly related to geo-spatial distributions. In addition, the authors state that "good flow maps contain a moderate number of nodes (less than

100)" [137], which is not the case for our data. Recently, an improved layout based on spiral trees has been introduced [24]. A different approach to combine geographic and hierarchical information into one visual display is to consider spatial ordering when creating space-filling rectangular layouts like Treemaps. One option is to take longitude and latitude values into account when splitting the Treemap rectangles as in Mansmann's Geographic HistoMap Layout [117]. Wood and Dykes [182] follow the same fundamental idea with their Spatially Ordered Treemaps. Later Slingsby et al. [159] suggest a further version of geographically ordered space-filling rectangular layouts. All of the mentioned approaches share the property that either the upper levels of the hierarchy are geospatial, e.g., areas and subareas, or that in each leaf node geo-spatial distributions have to be displayed. The latter option causes difficulties when having many geo-spatial locations and limits the possibilities for conveying feature information in leaves, because the most important visual variables are already used to convey the geo-spatial dimensions.

3.1.3 Data and Resources

In this section we briefly describe our approach to automatic feature extraction as well as the external data sources used for our approach.

Automated Linguistic Feature Extraction

> "*Many simple linguistic features can be extracted from parallel texts, such as the New Testament, by indirect measurement [...]. This holds especially for morphological typology. Morphological typology is a traditional field within linguistic typology concerned with assessing the degree of cross-linguistic variation in morphology, the internal structure of words [...]. Five families of values — (i) degree of synthesis, (ii) amount of prefixing and suffixing, (iii) case, (iv) amount of internal inflection, and (v) synthetic vs. analytic negation marking — are extracted automatically from electronic parallel texts (here the Gospel according to Mark) in a diverse world-wide convenience sample of 161 languages and in 125 languages of Papua New Guinea*[5].

[5] http://www.pngscriptures.org last revised on March 6th, 2013

Languages differ in how much information is packed in a word (degree of synthesis). In parallel texts languages with more complex morphology have more types of word-forms with lower token frequency than languages with less complex morphology. Types are the set of unique word-forms in a text and tokens are all instances of word-forms. One way to measure degree of synthesis is hence type-token ratio, another one is using trigonometry in token-type diagrams [...]. The simplest kind of morphology is concatenative morphology distinguishing parts of words (morphemes) of three kinds: stems (lexical element), and prefixes and suffixes (grammatical elements preceding or following stems). Given the distribution of lexical elements in parallel texts is known in one language, the forms of a lemma can be extracted for all languages with considerable accuracy. In the set of extracted forms invariant strings will be stems and variable strings prefixes and suffixes (depending on position). This allows us to measure the degree of prefixing and suffixing in different languages [...] . The amount of case marking can be estimated effectively from extracting just the forms of proper names by the same method since proper names do not usually vary in any other grammatical category except case [174]. Finally, the amount of analytic vs. synthetic marking of negation (whether negation is expressed in a word, as in English (not) or as an affix, as in Czech (ne-) is measured with two different algorithms [175]. All extracted features are continuous." [146][6]

External data sources

Ethnologue Ethnologue[7] is a Web source containing information about 6,909 languages. According to Ethnologue, these are all of the world's living languages. Among other metadata genealogical relationships between languages are provided, which we use for generating a complete genealogy, which can be used in the tool.

[6]Part of our joint publication written by Bernhard Wälchli. Original version contains further references.

[7]http://www.ethnologue.com/ last revised on September 11th, 2012

The Automated Similarity Judgment Program (ASJP) The ASJP[8] provides another widely used language resource on the Web. Namely, phonetically transcribed basic vocabulary lists containing 40 lexical items that are known to be relatively stable over time. The 40-items are a subset of the so-called *Swadesh list*, a basic vocabulary list often used by lexicostatisticians. As the name of the program already suggests, the ASJP data can be used to automatically compute similarities among languages. In addition, for each language a single-point geo-location is provided, i.e., one latitude and one longitude value per language. We use Version 13 of the ASJP database [180] which contains information on 4,816 different languages.

The World Atlas of Language Structure (WALS) WALS [49] "is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors (many of them the leading authorities on the subject)."[9] The version of the WALS database we used, contains 76,492 datapoints for 192 features and 2,678 languages. That means, that only about 15% of the data table are populated with entries. In addition, a language genealogy is provided as well as single-point geo-locations for languages. All features are either nominal (e.g., word order types such as Subject-Verb-Object or Subject-Object-Verb) or ordinal.

3.1.4 Analysis Tasks and Goals

The analysis goals and tasks are described by the domain experts:

> "*In principle, there are four reasons why languages can share a certain feature [...]. Beyond the trivial case (a) where all languages share the feature and (b) features are shared by chance, linguists are interested especially in whether (c) features are shared due to genealogical inheritance or (d) due to areal contact (borrowing).*
> *In order to be able to distinguish between (c) and (d) both genealogical*

[8] http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm last revised on September 11th, 2012

[9] Cited from the WALS Web page: http://www.wals.info last revised on September 11th, 2012.

(hierarchical) and areal (geo-spatial) information has to be combined in one visualization. The combination of both types of information can be achieved from two different angles. On the one hand, it can be checked for a given areal pattern whether the languages that are included all belong to the same family and thus lead to a clustering of the same feature at a certain region of the world or whether there is a real contact situation where unrelated or distantly related languages share a feature. On the other hand, it is of interest to check for a given family whether the feature values are the same or similar for individual languages or whether an unusual feature value occurs, which can be attributed to the fact that the language is spoken in some other area and therefore might have borrowed that divergent feature from the languages in that area.

A further advantage of the Sunburst visualization is that a considerable number of features (both nominal and numeric) can be visualized together without much data reduction which allows for a direct introspection of the degree of homogeneity or heterogeneity of a large dataset or parts of it (some families or features being more heterogeneous than others). There are no visualizations of typological data up to now that can achieve this goal.

Hence, the visualization has multiple aims. It helps linguists to formulate hypotheses about feature inheritance or borrowing in individual cases and to assess the degree of homogeneity of the complex datasets or parts of them in direct comparison to each other. The visualization further allows to see which features are more stable genealogically than others." [146][10]

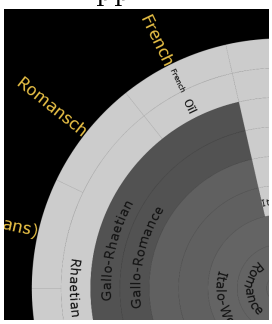
3.1.5 Integrating the Hierarchical and Geographic Data Space for Visual Feature Comparison

This section contains detailed explanations about our approach which are also illustrated and exemplified in Figure 3.1 for a better understanding. Our goal is to give a complete overview of all available language resources integrating

[10]Part of our joint publication written by Thomas Mayer. Original version contains further references.

automatically extracted and manually edited language features with genealogical and areal information into one visual analytics system. As a core part, we suggest a novel Sunburst display that was implemented building on *prefuse* [73] and Christopher Collins' RSF-Tree implementation[11]. It enables the visual exploration of different types of language features, even combined at the same time:

1. The homogeneity of *distance-based features* is plotted to the inner nodes of the Sunburst. Distance-based features may be any abstract data features. The only requirement is that their pairwise distance can be calculated according to a metric distance function. Examples are the edit-distance of Swadesh lists or geographic distances among languages. Of course, for any single or multivariate quantitative feature, homogeneity can also be calculated and mapped to the inner nodes.



The saturation of the grey tone of an inner node, indicates whether the languages of the corresponding family on average have small distances (light grey) or large distances (dark grey). Apart from providing additional information, this coloring also helps to perceive the hierarchical relations easily.

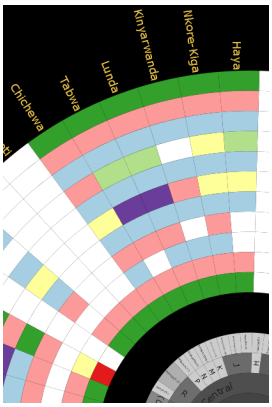
2. The *quantitative, ordinal, or nominal features* are plotted to the outer rings of the Sunburst display. For each feature dimension one ring is reserved, the value for a certain language is mapped to the color, brightness or degree of fill of the ring segment belonging to that language. Examples for such values are the quantitative value showing the degree of prefixing of a language's words, or the nominal value of a language's word order type, as described in Section 3.1.3. The segments belonging to one feature dimension are aligned in one ring, readily enabling the comparison across languages in accord with the Gestalt law of continuity.

Plotting language features

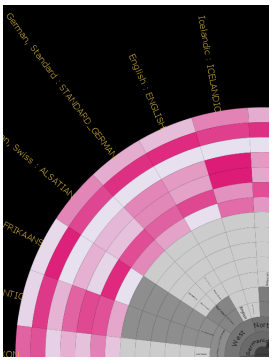
Mackinlay's fundamental research [114] has shown that the choice of suitable visual variables to convey information depends on the data types. In our case,

[11] http://faculty.uoit.ca/collins/research/docuburst/index.html last revised on September 11th, 2012

the two generally most valuable visual variables, namely the x and y Position, are already used to display the hierarchy. Consequently, we pick the next best choice according to Mackinlay's research to plot the language features. This next best choice is different for quantitative, ordinal, and nominal data.

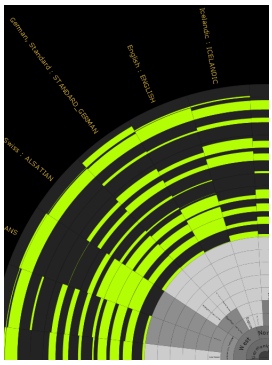


Nominal Data: For the nominal data we use different color hues to encode different categories. We follow the suggestions for color maps provided by the *color brewer* (see http://colorbrewer2.org). It has to be remarked that we aim to create two notably different color maps keeping the hues as disjoint as possible. The reason is that if two adjacent nominal feature segments have the same color, they appear as a visual pattern calling the attention of the user. This is beneficial if the two segments are located within the same feature ring, i.e., two closely related languages share the same feature value. However, if the two segments are located in different feature rings the coincidence in color is meaningless. To avoid the second case, two adjacent nominal feature rings get different color maps that are as disjoint as possible. To do so, we use a color map from color brewer that contains 11 different colors for nominal data, which are about as many colors as can be readily distinguished. As typically a nominal feature dimension in our data has only 5 or 6 categories, usually we can split this color map into two disjoint color maps. In this case, the first ring gets the first colors of the color map and the second ring gets the further colors. Of course, in cases with more nominal categories, it cannot be guaranteed that the color maps for the two adjacent rings do not overlap, but at least the number of overlapping colors is minimized. Missing values can be colored in white or grey.



Ordinal Data: For conveying ordinal features Mackinlay identifies different density or color saturation values to be suitable. We decided to take different color saturation values. Thus, we divide the spectrum of all color saturation values by the number of different ordinal values for a feature. We thus get a set of ordered color tones of the same color hue that can be distinguished easily. We decided to select a pink hue as this stands out and is sufficiently

dissimilar to the hues used for the nominal data.



Quantitative Data: For quantitative data it makes sense to use the variable *size* in order to reveal relative differences among the feature values. The quantitative feature rings in our approach show values in a histogram, where the height of the bars corresponds to the normalized feature value. Again we chose a hue that is dissimilar to the hues used for the nominal data.

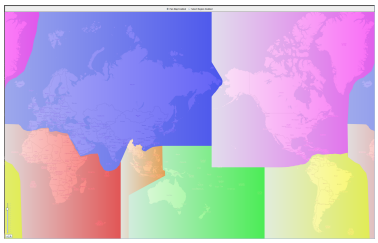
Visually conveying the data type

The meaningful mapping, however, is not the only reason why we decided to use different visual variables for the different data types. A further advantage is that the user is able to recognize the data type of a random feature ring shown to her/him immediately. This is especially valuable if one user creates a visualization selecting a set of features with mixed data types and shows it to another user. Yet, we still enable users to change the pre-configured mapping, for example s/he can also choose color saturation values to represent quantitative features.

Integrating areal information

Geographic information is also integrated into the Sunburst display. We have already described the option to map the homogeneity (average) of geographic distances among languages in the same branch to the grey tone of the corresponding inner node in the hierarchy. Further options are explained in the following.

Macro areas as nominal dimension



As mentioned before, the world can be divided into macro areas, within which intense language contact is known to have happened. At the same time, language contact between different macro areas used to be rare. We allow the user to choose

between two ways of integrating the information about macro areas into our extended Sunburst visualization:

1. The macro areas can be incorporated as a new first level into the language genealogy hierarchy. This means that the root of the tree has no particular meaning. Next, the languages are split up according to contact regions and only below according to the language genealogy.

2. The macro areas can be incorporated as the innermost ring into the display. In this case, the macro areas can be seen as another nominal data dimension and the ring segments will be colored according to the coloring of the macro areas on the world map. The user has the option to choose increasing color saturation values within macro areas either from east to west or north to south to receive more detailed information about the location of a language.

Interactive linking of the world map

To explore the exact geo-spatial distribution of languages the Sunburst display is interactively coupled to the world map. Through linking and brushing, the geo-spatial distribution of all languages belonging to a selected branch is displayed on this separate world map. Each language has exactly one point on the world map, because this is what the data gives us. A small circle is displayed at the language position colored according to a user-selected language feature. At the same time, the user can select arbitrary areas on the world map and create a Sunburst containing only those languages that are located in the selected area. In addition, the user has the option to ignore the coloring of the macro areas and create a bipolar color map for the selected area as shown in Figure 3.4.

User Interaction

The interactive linking between the Sunburst and the world map is only one way of interacting with the display. The user is interactively involved in the data analysis process right from the start, see Figure 3.2. For example, s/he is asked to specify the data types of the feature dimensions and able to change them anytime, in case of errors. Both the world map and the Sunburst enable

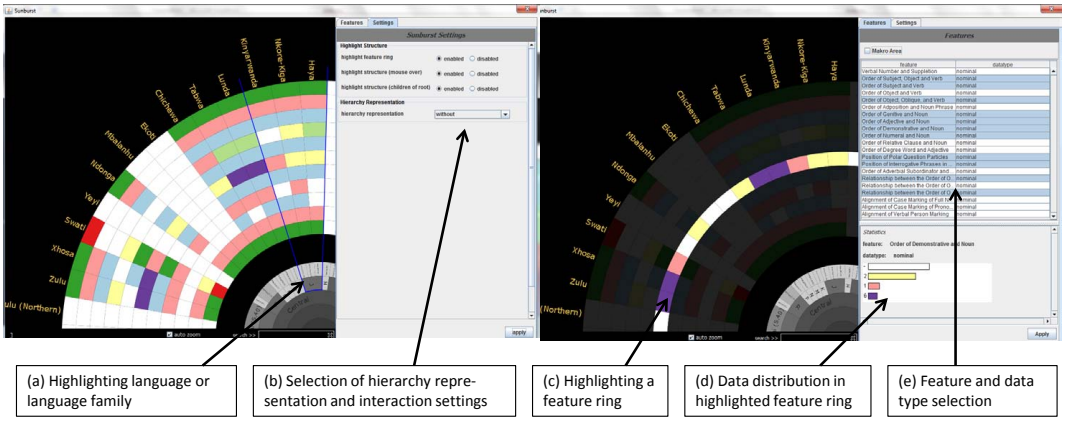


Figure 3.2: User Interaction with the Sunburst and feature rings. Reprinted from [146], © 2012 The Eurographics Association and Blackwell Publishing Ltd.

panning and zooming interaction. In the Sunburst, the user can select to focus on different aspects, e.g., single language families or languages or features that s/he is currently interested in. While this information is highlighted all other information will be covered with a semi-transparent dark grey color tone in order not to distract the user. In addition, the global data distribution for the highlighted feature ring is displayed in a further panel.

Automated ordering of the tree layout

The order of languages (and features) in the display layout may have a strong impact on the appearance of visual patterns and thus be crucial for an analyst when trying to spot unexpected peculiarities. For that reason it makes sense to order the leaves (languages) in the tree layout according to their similarity. Even though the hierarchical structure restricts the number of possible language permutations, it is still a computationally complex problem. In our investigations we used an optimized algorithm [16] from the field of bioinformatics, which solves the computation in $O\left(4^k n^3\right)$, where k is the upper bound on the number of children of each internal node, and n is the number of leaf nodes. Starting from the leaf-nodes, optimal sortings for subtrees are computed bottom-up. In particular, for each subtree the optimal ordering is computed for each combination of leftmost and rightmost leaf elements (languages). Only half of the combinations have to be actually computed, because

of the symmetry of metric distance functions, which we apply. That is the optimal ordering between node a and node b is the inverse of that between node b and node a. In the upper levels of the hierarchy the optimal orderings from below have to be combined and again computed for each combination of all possible leftmost and rightmost leaf elements. The number of leaf-node permutations are restricted by the hierarchy constraints of the subtrees. Of course, for large hierarchies with a high fan-out a run-time of $O\left(4^k n^3\right)$ may still be prohibitive. We re-implemented and integrated the algorithm into our system and for the limited datasets we have, the computation of an optimal ordering may still take several days. Therefore, it is recommended to run it as a batch process before analysis and save the ordering to be reloaded without further computational effort anytime. Our experiments have shown that it makes sense not only to order the tree to maximize the sum of the pairwise similarities of the leaves, but that it can also be valuable to minimize this sum. In that case, subtrees are sorted to appear the least homogeneous possible. Consequently, highly homogeneous subtrees will stand out visually and can easily be identified at a glance. For the sorting of the leaves a metric similarity or distance function between pairs of languages is defined. As a first step, each language is represented as a feature vector, where each dimension corresponds to one language feature. Next, any common distance or similarity function can be applied. If we have only numerical values, for example, we apply the Euclidean distance for sorting.

3.1.6 Case Studies

The development of the tool was an interdisciplinary effort involving linguistic researchers. We met regularly to assure a correct understanding of the linguistic data and tasks and discuss further steps. With the support of our novel visual analytics approach domain experts were able to generate new hypotheses relevant to their field and confirm old ones. The following case studies report on the experimental work and findings.

> "*In order to be able to discriminate between cases of language contact and inheritance from a proto-language [...] it is necessary to combine both the genealogical (hierarchical) and the areal (geo-spatial) infor-*

mation about languages. As mentioned before, the impact of a contact scenario can be inspected from two perspectives: (i) looking at geographical distributions (areal patterns) and checking whether all languages in the given area are from the same family; (ii) looking at a particular family (or genus) and checking whether all languages exhibit the same feature values and are spoken in the same region. We will concentrate on the latter aspect with two application case studies of our Sunburst visualization which enables the user to check for a larger amount of features whether there are outliers within the family that result from the fact that a language is spoken in a different area. As to the language properties, we experimented mainly with the automatically extracted features which have been inferred from the parallel Bible texts [...]. These features give a good approximation of what linguists have analyzed manually and are also interesting for contact situations for which the visualizations are designed. In order to test the visualization for its usability, a number of language families have been inspected by the domain experts among us. Several interesting findings could be inferred from the visual representation of the features." [146][12]

Case Study 1

The first case study was conducted by the domain experts.

"*First, we will concentrate on a particular case which can be most easily explained for non-experts. For this purpose, we look at the more familiar Indo-European language family, which also includes the prominent European languages English, French or German. Figure 3.1 shows the Sunburst representation of the Indo-European languages in our sample and their hierarchical structure of subfamilies (genera). In addition, the innermost ring of the visualization shows the color-coded macro-area in which the respective language is spoken. It can be seen at-a-glance that the languages are spoken in the same macro-area (Eurasia), with the sole exception of Afrikaans, which is located on the African continent. Furthermore, Afrikaans can easily be detected*

[12]Part of our joint publication written by Thomas Mayer.

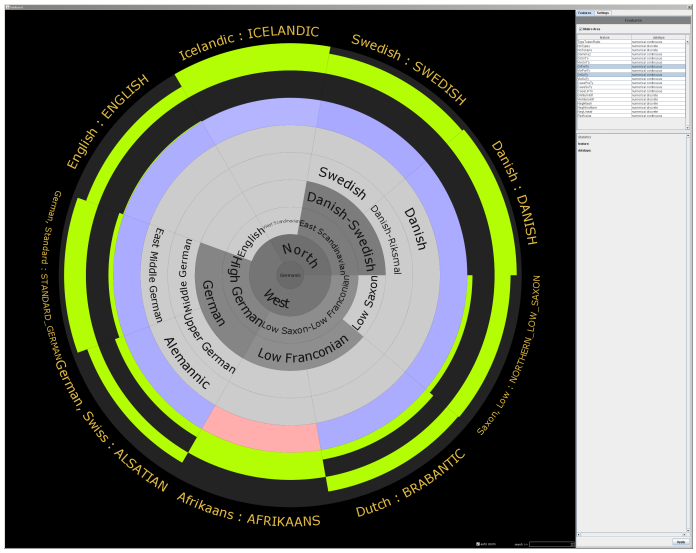


Figure 3.3: Detailed look into two quantitative features for the Germanic languages. The macro area (color of the innermost ring) and the two numerical features (outer rings) are clearly different for Afrikaans compared to the other Germanic languages in the sample. Reprinted from [146], © 2012 The Eurographics Association and Blackwell Publishing Ltd.

as an outlier with respect to its neighboring languages, i.e. many feature values deviate strongly. Part of this effect is due to the fact that some features are correlated, which can be seen when looking at their distribution over all languages, however, Afrikaans is visually salient independent from that.

For the sake of simplicity, we select only the family of Germanic languages and look at just two features in the Sunburst visualization, namely the synthesis parameters of prefixation (morphological material occurring before the stem) or suffixation (morphological material occurring after the stem). In the visualization in Figure 3.3, both features are depicted in the outer rings of the Sunburst, with the prefixation feature as the inner ring of the two and suffixation as the outermost ring. When looking at both feature rings, it can immediately be seen that Afrikaans is not only peculiar because of its areal status but also regarding the feature values that it has. In comparison to the adjacent (West) Germanic languages Afrikaans has a higher prefixation and a lower suffixation value. This is particularly interesting because it is in a contact situation with surrounding African languages (our sample contains the Bantu languages Zulu and Xhosa, which are also spoken in South Africa). Bantu languages are notorious for their extensive use of prefixes to convey grammatical meaning on the verb.

The comparatively higher prefixation value for Afrikaans thus might be caused by the influence of the Bantu morphological patterns. On closer inspection, however, it turns out that Afrikaans makes extensive use of the perfect construction involving the past participle with "ge-" (similar to Dutch or German). The synthetic past tense forms (the so-called imperfect tense) where a further distinction for different persons (first person singular, third person singular, etc.) is made in suffixes have disappeared except for a few vestigial cases [44]. The fact that a further distinction in suffixes does not exist with the past participles, which are now dominant in the language to convey reference to a past event, results in a lower suffixation value for the language with respect to other (West) Germanic languages. Whether the use of the perfect instead of the past tense is a direct influence of the contact languages or merely due to the geographic separation of Afrikaans with respect to other Germanic languages (especially its sister language Dutch), however, remains to be investigated. Yet the visualization easily enables the linguist to check for such suspicious patterns which can later be inspected in more detail." [146][13]

Case Study 2

While the Germanic Languages in general are well-studied, for other language families the available knowledge can be very limited. For only a few of the numerous languages spoken in Papua New Guinea grammar books are available. Translations of Bible texts, however, can be gathered for quite a lot of them[14]. The features automatically extracted from those can be seen in Figure 3.4. In the following I summarize the observations and findings that my collaborators from linguistics were able to make (without having closer knowledge about the individual languages): AUSTRONESIAN languages are rather more homogeneous in their feature values than PAPUAN languages which according to my collaborators is in line with their well established genealogic relationship. Another quite homogeneous group is HUON-FINISTERRE, the domain experts identified a high degree of synthesis, no morphological negation, very little

[13]Part of our joint publication written by Thomas Mayer.
[14]http://www.pngscriptures.org last revised on January 11th, 2013

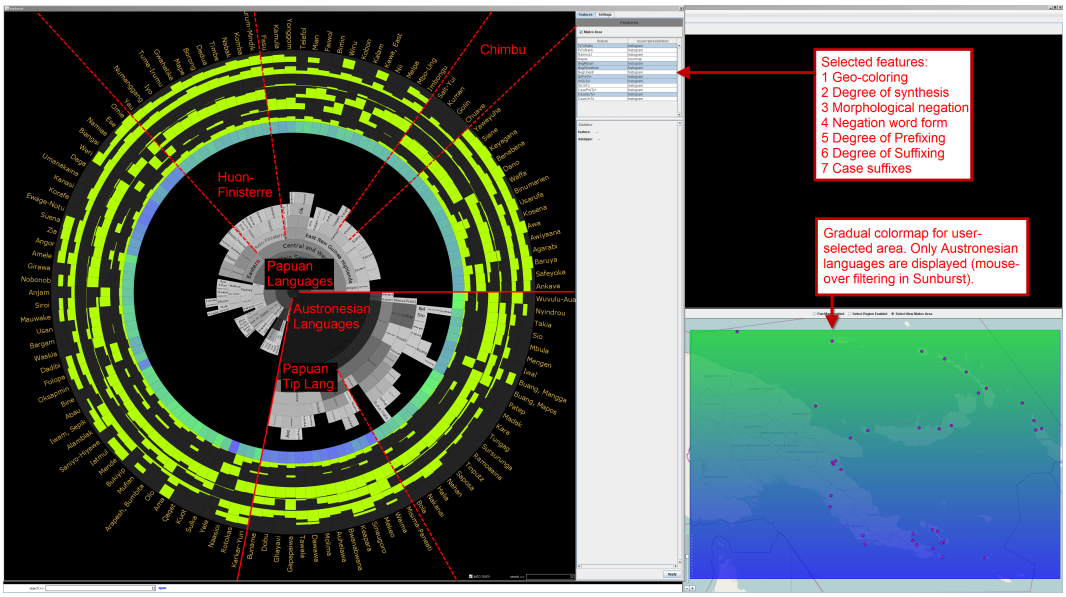


Figure 3.4: High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea. Reprinted from [146], © 2012 The Eurographics Association and Blackwell Publishing Ltd.

prefixing, and much suffixes among which case suffixes. A further observation is that within East New Guinea, genealogical subgroupings clearly emerge when several features are considered. The Eastern subgroup, for instance, is characterized by high synthesis, analytic negation, no prefixes, moderately high suffixing, and case suffixes. The Chimbu subgroup called the linguists attention, because it is distinguished by morphological negation, lack of case and rather low degree of synthesis. Much heterogeneity is observable in East Papuan, which according to the domain experts is well in line with the fact that this is not quite an established family such as Austronesian. Within the Austronesian family the subfamily of the Papuan Tip languages can be distinguished both with respect to certain features and the geo-locations. The case study demonstrates that domain experts are able to quickly make insightful observations when using the interactive visualization. The observations concern both details on certain languages and certain features as well as groups and clusters that become visible at different levels of the genealogical hierarchy.

Case Study 3: The effect of leaf ordering

This section discusses the impacts of re-ordering the tree nodes in order to maximize or minimize the pairwise leaf similarities within the subtrees. The same data as in the previous case study is used and the emergence of visual patterns is discussed.

Ordering to *maximize* pairwise leaf similarities Figure 3.5 again shows the Papua New-Guinea languages provided by the domain experts, but this time sorted according to similarity. One of several interesting sectors is shown in Figure 3.6, where the languages AMA and KARKAR-YURI nicely fit to the SEPIK language family, with respect to their automatically extracted feature values. At the same time, especially ABAU deviates from the other SEPIK languages. Only looking at these features, it does not become clear why ABAU should be considered a SEPIK language and AMA and KARKAR-YURI not, because the latter ones *look* much more like it. The next logical step is to explore the geo-spatial distribution of these languages. ABAU and KARKAR-YURI are centered in the SANDAUN Province, while AMA and all of the other SEPIK languages of the sample are centered in the EAST SEPIK Province. Again, from a naive data analysis perspective one would come to the conclusion that if ABAU really is a SEPIK language, the same should be true for KARKAR-YURI and especially for AMA. This is an interesting starting point for linguists to explore, where the similarities among languages stem from. Is the current classification insufficient or even wrong? Or is it correct, but the features reflect effects of language contact? Or may the visualization just help to discover an error in the data or a disadvantage of the automatic feature extraction method?

This is just one example out of many, where the visualization can point the analyst to potentially interesting findings that have to be further explored by domain experts. In other cases it enables the analyst to spot outlier languages (see Figure 3.7 and 3.8), or languages that are similar in features, but belong to different subtrees of the genealogy (see Figure 3.9, 3.10, and 3.11). It is also of interest to compare patterns in the genealogy with patterns in the geographic data space, as in Figure 3.12.

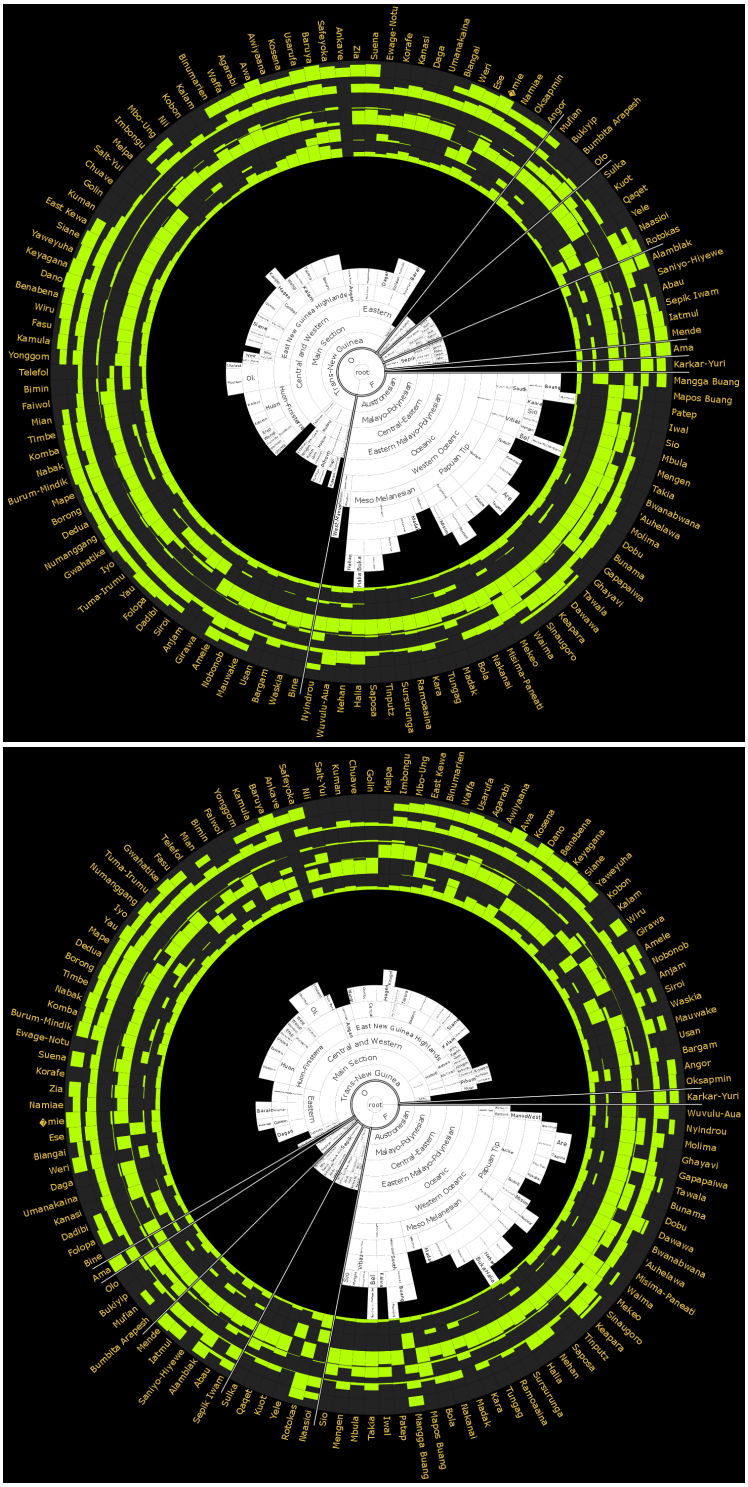


Figure 3.5: High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea with leaves ordered to *maximize* (top) and *minimize* (bottom) the pairwise leaf similarity for neighbors. Details will be highlighted in the Figures 3.6 to 3.13.

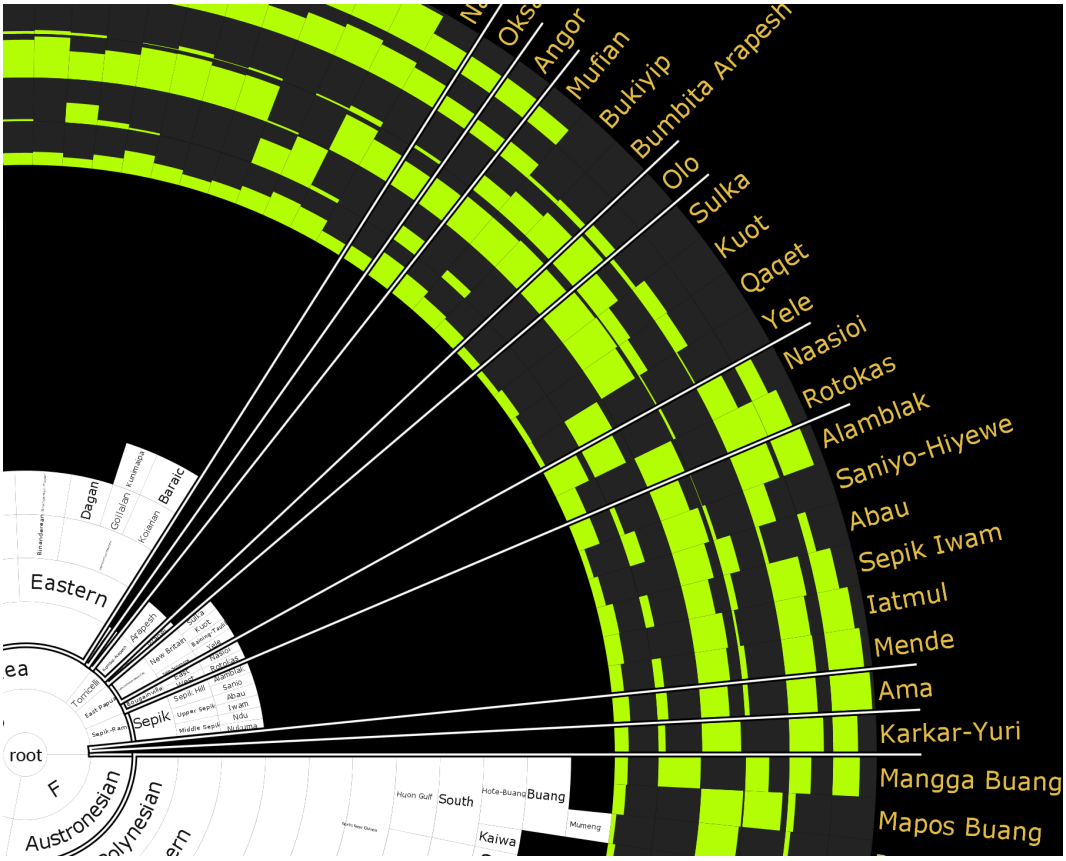


Figure 3.6: High-resolution screenshot showing automatically extracted features for a subset of languages from Papua New Guinea with leaves ordered to maximize the pairwise leaf similarity of neighbors.

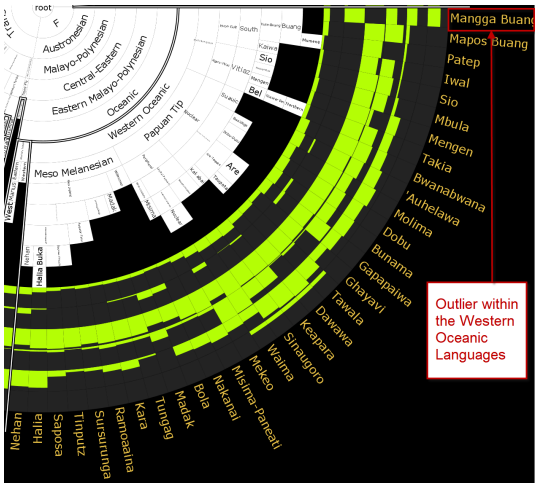


Figure 3.7: There are outlier languages like Mangga Buang within the Western Oceanic Languages

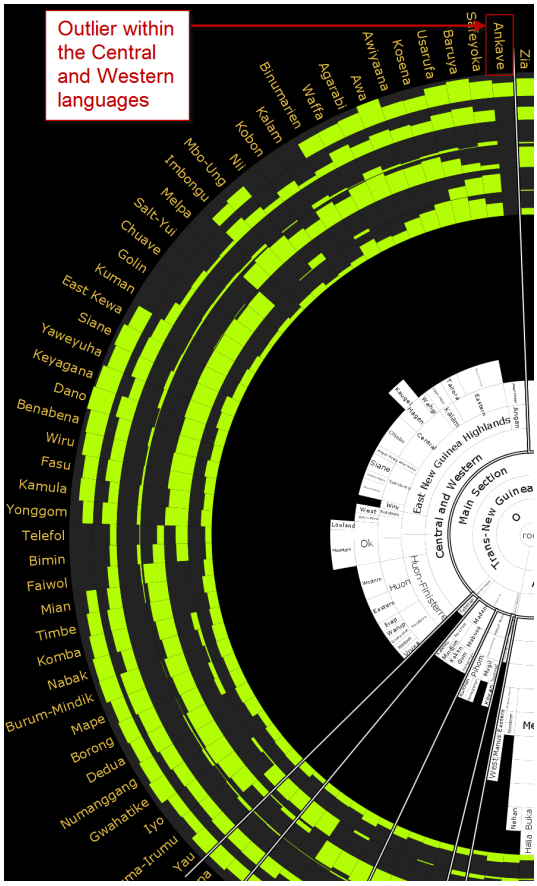


Figure 3.8: There are outlier languages like ANKAVE within the CENTRAL AND WESTERN LANGUAGES

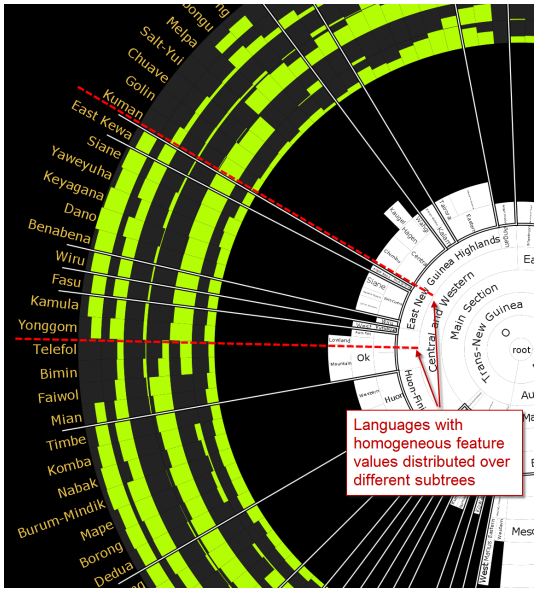


Figure 3.9: When looking just at the feature values a homogeneous circle segment ranging from EAST KEWA down to YONGGOM stands out. While all of the languages belong to the CENTRAL AND WESTERN language family, they diversely distribute over different subfamilies. Whether this could be due to language contact or a controversial categorization of subfamilies is an open question that domain experts could investigate.

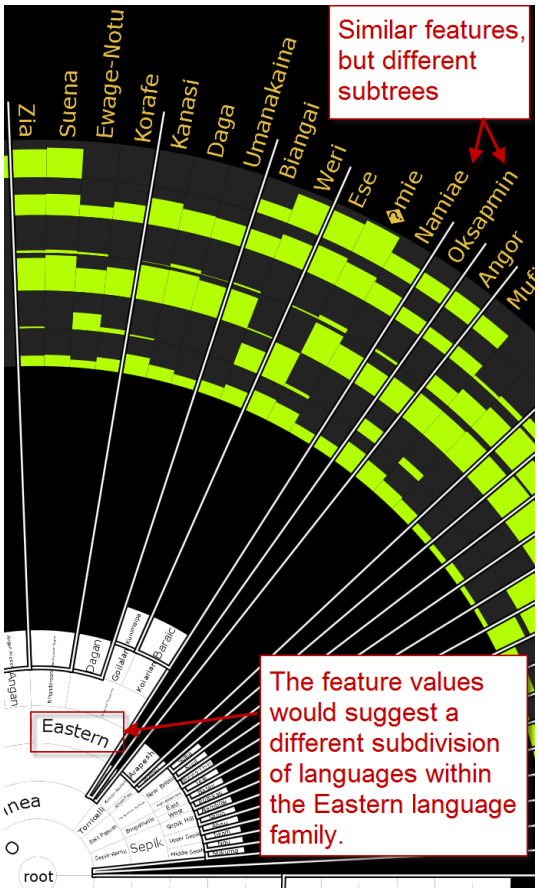


Figure 3.10: The languages NAMIAE is more similar in features to OKSAPMIN than to any other language of its EASTERN language family. However, OKSAPMIN does not belong to this family. The features suggest that either both or none of them should belong to the EASTERN language family. When looking just at the feature values of the languages contained by the EASTERN language family, one would expect a different division of subfamilies. For example, KORAFE and KANASI are quite similar and divided, and the same is true for WERI and ESE, whereas within the established families clear differences in features are perceivable.

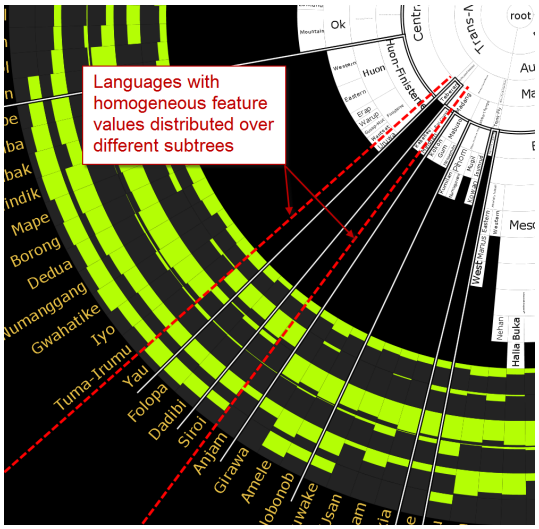


Figure 3.11: When looking just at the feature values for languages ranging from YAU down to SIROI are quite homogeneous, but the hierarchy reveals that they belong to three apparently quite different language families.

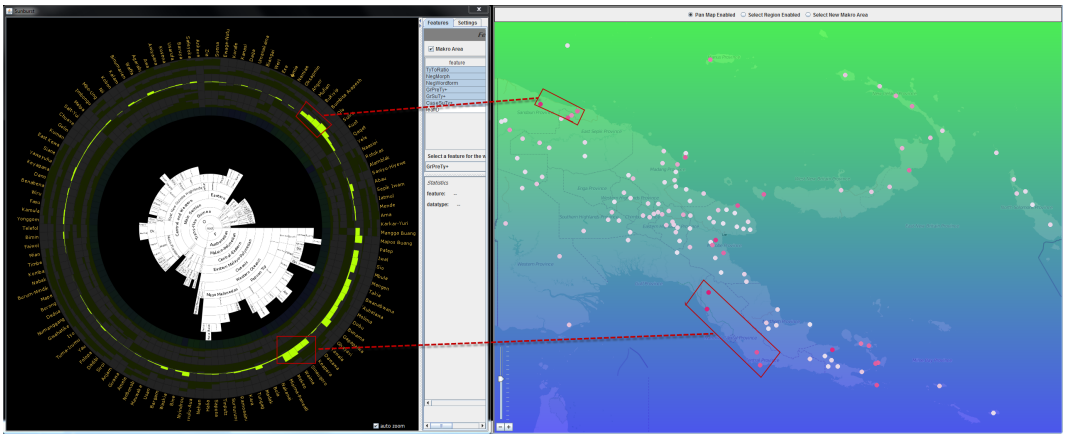


Figure 3.12: Example for an exploration using both displays. The user highlights a feature in the genealogy and the same feature is plotted on the map (white = low value, red = high value). The salient languages cluster in both views.

Ordering to *minimize* pairwise leaf similarities Ordering data objects (in this case languages) in a way that similar ones will be grouped is a strategy often pursued in information visualization. In contrast, it is not quite clear at first sight why we should also try to do exactly the opposite. Yet, in our particular analysis setting there is one analysis task where an ordering according to dissimilarity might be quite useful. If a subtree is highly homogeneous in the data space, the algorithm will not be able to arrange the corresponding languages in a way that the subtree visually appears inhomogeneous. Thus, an extraordinarily homogeneous subtree will stand out much more when the tree is ordered according to dissimilarity, so that in this case the homogeneous subtree can be spotted easily. Two examples are provided in Figure 3.13. The fact that the HUON-FINISTERRE languages in this figure still appear homogeneous backs up the finding made in the previous case study on the same data, but without leaf ordering.

3.1.7 Discussion and Conclusion

Discussion In this paragraph we discuss different aspects, problems, and open issues of this research project as well as lessons learned from our interdisciplinary collaboration.

For the application development the involvement of domain experts from the

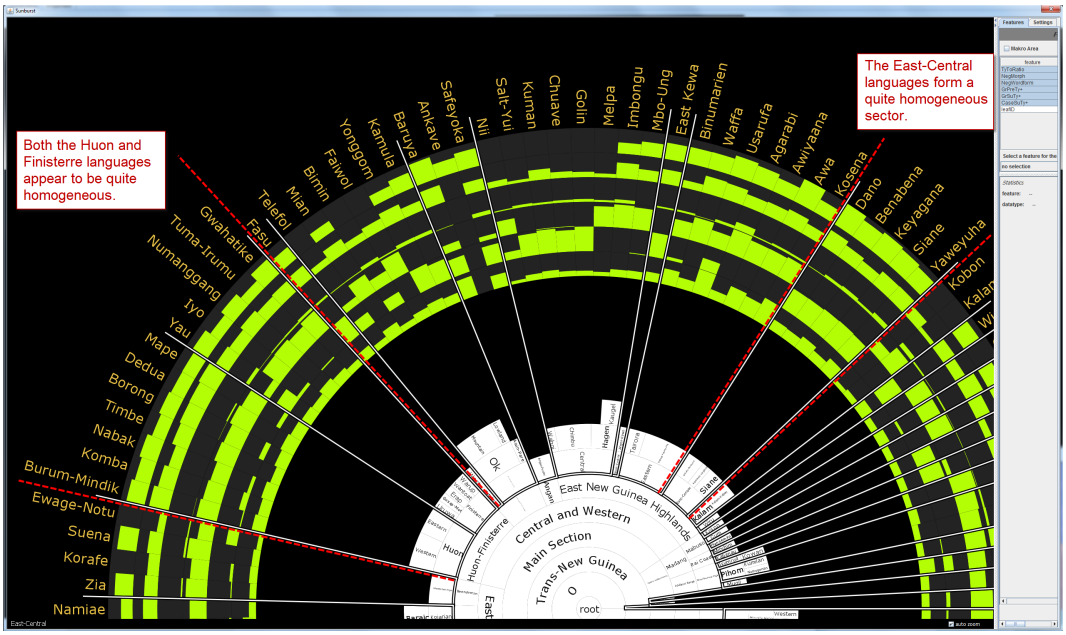


Figure 3.13: High-resolution screenshot showing automatically extracted features for a subset of languages from Papua New Guinea with leaves ordered to *minimize* the pairwise leaf similarity of neighbors. Two homogeneous sectors stand out.

very beginning on was extremely useful, much more than expected. The incorporation of a deep understanding of the domain and its data prevented us going into a wrong direction when designing the application. The concrete analysis tasks of the domain experts provided a good guidance to design the system.

A problem from the analysis perspective is the sparseness of the data. While the genealogical information for about 6,900 languages is available, most features are only available for a few hundred languages. However, linguists are about to collect more and more such data and for the future it is to be expected that the data issue will become less critical.

The number of languages is much larger than the number of different features. Currently, we have less than 250 features. The number of feature rings, consequently, is limited and the whole data set can still be visualized on a large high-resolution display. However, given that the number of features to be displayed grows heavily in the future, an Icicle plot may be more suitable than a Sunburst when displaying the whole dataset at once. The reason is that each

additional feature ring for the Sunburst needs more space than the previous one and the display grows in x and y direction. An Icicle plot would still allow to map information to the inner nodes of the hierarchy and the display would only grow in y direction and become more quadratic as the number of features approaches the numbers of languages. However, we could observe that in a typical analysis case only a limited set of features is available or of interest. For this setting the Sunburst makes a better use of the screen space.
While the hierarchically structured genealogical information is the core of our display, the integration of geo-spatial views is yet on a fairly basic level. Several open issues can be identified that could help to improve the integration in the future. First, working with distorted maps might be useful to grant more space to data-wise densely populated regions. Ideally, the distortion would be constantly re-calculated according to the current selection of languages, which is a challenging task. Secondly, the division of the world map into macro areas is linguistically motivated and easy to understand, but coarse grained. For arbitrary smaller regions the user can create customized color maps, similar to a two dimensional color mapping of location proposed by Wood and Dykes [182]. Further, it is known that populations, and with them languages, are more likely to spread within the same climate zone than across climate zones [60]. The actual likelihood of language spread in each direction also depends on natural borders like seas, mountains, and deserts. It would certainly be an interesting topic for future interdisciplinary work to generate a color map that encodes a "spread"-distance between languages.

Conclusion In this section we introduced a new field of application for visual analytics: Historical comparative linguistics, linguistic typology, and areal typology. We provided background information about the research in this field including concrete tasks and requirements and available data sources. In our approach we demonstrated how linguistics research can profit from visual analytics. In particular, we suggested an extended Sunburst visualization with *feature rings* in order to enable the comparison of several features at once in the context of a language genealogy. We discussed different ways to design the feature rings that are optimized for either of the data types nominal, ordinal, and quantitative. We showed that ordering languages and language features

according to similarity supports the visual analytics process, because it makes visual patterns emerge that would not be visible in alternative orderings. In a second step, we linked the hierarchical display with a geo-spatial visualization and suggested ways of integrating the geo-spatial information into our Sunburst.

Domain experts were involved in the development from the beginning on to assure that their tasks and data were correctly understood. Their suggestions were considered during the development. In the end, they used the final version of our tool and were able to generate new hypotheses relevant to their field and confirm old ones. Visualization also showed to be a good means to discuss hypotheses and theories.

While the domain for which this application was designed at first sight might appear to be narrow, typological comparison and investigations of historical change engage a large research community. One of the data resources investigated, the WALS data, has even become a standard resource for teaching in linguistics. In addition, further research communities with related tasks, like the variation genetics field in biology, could potentially profit from the presented application.

In future work the geo-spatial component of the approach could be extended and experiments with further interaction techniques could be conducted. In addition, the reasoning behind some of the design choices made could be strengthened by controlled user studies.

3.2 Cross-Linguistic Comparison of Complex Language Features

This section builds on the following publications:[15]

C. Rohrdantz, T. Mayer, M. Butt, F. Plank and D. A. Keim. Comparative visual analysis of cross-linguistic features. Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010), pages 27-32, 2010.

T. Mayer, C. Rohrdantz, M. Butt, F. Plank and D. A. Keim. Visualizing Vowel Harmony. Linguistic Issues in Language Technology, 4(Issue 2):1-33, 2010.

In this section we research complex language features, i.e. features that do not consist of one or multiple univariate language features, but form a graph-like structure of interdependent features. In particular, the novel method introduced here is a visual analysis of statistics on sound successions, which furthers the investigation of phonological and morphological characteristics of languages. The sound successions are visually organized into a matrix display, revealing binary sound dependencies. First, we introduce the novel technique along with a case study on a phenomenon called *vowel harmony* (VH). Then, we show how this approach can also be applied to support hypothesis generation when investigating related data such as consonant successions and sound replacements.

[15]The two publications have a high degree of overlap, one was a workshop and the other a journal publication. Thomas Mayer and I shared the work equally, he contributed the linguistics research and I contributed the computer science research and did the programming, except for some linguistic pre-processing computations. Miriam Butt proof-read the text and gave advice. Frans Plank and Daniel Keim also gave advice. Further people that we also acknowledge in the journal publication are Maria V. Tolskaya and Irina Nikolaeva for providing the Udihe texts, the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) for the Warlpiri Bible sections, and Bernhard Wälchli and John Goldsmith for valuable comments on an earlier version of the work. For all parts of the publications that were not written by me I reference the corresponding original work.

3.2.1 Background

In this section a novel visual analytics approach is introduced to research vowel harmony.

> "*Vowel harmony is an assimilatory phonological process by which vowels are pronounced in accordance (or harmony) with their environment [...]. Most often, preceding vowels trigger the shape of the vowel that follows them, leading to a kind of domino effect within a certain linguistic domain (usually the 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 markers 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 'houses', çöler 'deserts', örtüler 'coverings', kediler 'cats'); 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 'men', toplar 'balls', komşular 'neighbors', kapılar 'doors'). However, most languages do not contain VH and even VH languages always also show cases of disharmony. Besides, they differ with respect to how many and which features are active in the harmony process.*" [148][16]

More details about VH have been provided by my research partner from linguistics, Thomas Mayer, as part of our joint journal publication [119].

VH is a complex language feature because it cannot simply be calculated and reduced to a single number or set of numbers, but rather consists in patterns of interdependence among statistics for different vowels. We demonstrate how visual analytics can support linguists in detecting the degree and kind of VH (or related phenomena) involved in a language and readily compare different languages with respect to vowel harmonic processes. The data and resources

[16]Part of our joint publication mostly written by Thomas Mayer. Further references are contained in the original work.

used are described in Section 3.2.2, and the tasks and analysis goals are outlined in Section 3.2.3. One important point is the automatic data analysis involving data preprocessing, statistical feature extraction, and vowel ordering which are described in Section 3.2.4. As a next step, matrix visualizations are designed that help track the probability and association strength of vowel successions within words and provide an insightful visual fingerprint for the vowel distributions in a language. Section 3.2.5 evaluates to what degree the results depend on the amount of data available. Next, in Section 3.2.6 several case studies are provided that show that the results nicely conform with existing linguistic knowledge and also that accurate hypotheses about VH can be derived from the matrix visualizations without any prior knowledge about a language. In Section 3.2.7, we demonstrate that the visual analytics approach can support hypothesis generation in other related tasks. Next, in Section 3.2.8 we discuss how existing visualizations, like the *droplet maps* technique, can be used to extend the analysis to sequences of more than two items. Finally, in Section 3.2.9 a discussion, conclusion, and a research outlook are provided.

3.2.2 Data and Resources

The data used as a basis for our work was extracted from Bible texts — using Bible texts means that we have data available for languages for which texts are not otherwise readily available. For each investigated language a type list was compiled containing all the different word forms appearing in the Bible. It is better to work on the Bible types instead of using a dictionary, because VH is best detected in inflected word forms, which are usually not contained in dictionaries. Moreover, it is better to work on a type rather than on token level, because otherwise highly frequent tokens might bias the results. For each list of types of a language, all vowel successions within the types are counted and analyzed. Vowels have been automatically determined with Sukhotin's algorithm [163] and manually edited for each language. 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. For 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 3.1.

	a	ä	e	i	o	ö	u	y
a	3548	20	1940	1893	831	0	944	24
ä	35	944	806	820	10	138	33	266
e	1623	1144	1495	1608	419	56	497	187
i	1580	854	1514	1044	376	46	355	135
o	1384	7	1032	902	284	0	294	8
ö	7	125	54	39	0	3	1	18
u	1464	6	1085	850	315	1	547	8
y	39	656	368	368	35	75	4	251

Table 3.1: 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.

3.2.3 Analysis Tasks and Goals

The goal is to provide linguists with a visual analytics methodology that supports them to explore the interdependence of sounds, like vowels, in an arbitrary language only requiring a limited fragment of digitalized text. For single languages the aim is to get a notion about the correlation of sounds and, most importantly, about groups of sounds that heavily succeed or avoid each other. It is anticipated that, apart from classical cases of VH, other kinds of interesting patterns might also appear. A special interest is the cross-linguistic comparison of sound succession phenomena and their variation across different languages.

3.2.4 A Statistics-based Matrix Visualization

Statistics

The simple matrix with the counts of vowel successions (as in Table 3.1) 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.
In order to provide more detailed insight into the relevant patterns, we calculated the succession probabilities. There are two kinds of probabilities that we can consider, the probability of observing the next vowel ($nextProb$, see equation 3.1) and the probability of observing the previous vowel ($prevProb$,

see equation 3.2).

$$nextProb(v_x, v_y) = \frac{count(v_x - \triangleright v_y)}{\sum_{i=1}^{n} count(v_x - \triangleright v_i)} \tag{3.1}$$

where $x, y \in \{1 \cdots n\}$
and $count(v_x - \triangleright v_y)$ is the frequency of successions from $vowel_x$ to $vowel_y$
and n is the overall amount of different vowels in the language under investigation.

$$prevProb(v_x, v_y) = \frac{count(v_x - \triangleright v_y)}{\sum_{i=1}^{n} count(v_i - \triangleright v_y)} \tag{3.2}$$

where $x, y \in \{1 \cdots n\}$
where $count(v_x - \triangleright v_y)$ is the frequency of successions from $vowel_x$ to $vowel_y$
and n is the overall amount of different vowels in the language under investigation.

In the case of *nextProb* that means that if a certain vowel is observed (as a first vowel in a binary succession), then it is calculated with which probability certain other vowels are expected to be observed next. In the case of *prevProb* that means that if a certain vowel is observed (as a second vowel in a binary succession), then it is calculated with which probability certain other vowels are expected to have been observed previously. The values for the two kinds of succession probabilities are then saved in a *nextProb* and *prevProb* matrix, analog to the matrix of absolute succession counts. Of course, still highly frequent vowels in most cases have a higher probability of succeeding and preceding any other vowel than infrequent 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. To get a significance value the fourfold χ^2 formula (see Formula 3.3, [152]) is applied. The higher the χ^2 values, the more significant in a statistical

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

Table 3.2: 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 equations 3.3 and 3.4.

sense is the deviation of observed frequencies from expected frequencies. The test quantifies the influence of the independent variable (e.g. *a* in Table 3.2) on the dependent variable (e.g. *e* in Table 3.2).

$$\chi^2 = \frac{(A+B+C+D)\cdot(A\cdot D - C\cdot B)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \tag{3.3}$$

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 3.4, [152]).

$$\phi = \sqrt{\frac{\chi^2}{(A+B+C+D)}} \tag{3.4}$$

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

Apart from that we tested further statistical measures, namely the *t test*, *likelihood ratio test*, and *pointwise mutual information*. In the end, the ϕ statistics turned out to be the most useful choice. There are basically two issues about using tests for statistical significance, like the χ^2 *test*, *t test*, or *likelihood ratio test*: (1) distributions of phenomena in natural language are far from being random and even minor correlations of many features, like sounds, are highly significant in a statistical sense. (2) The statistics are dependent on the sample size. The more data we base the statistics on, the more significant the effects will be. This makes a cross-linguistic comparison of significance values mostly meaningless, as for different languages we command different amounts

of data. It indeed does make sense, from an analytical point of view, to rank correlations of phenomena within the same language by significance. However, the next open issue is how to normalize the resulting values for visualization. Even a simple *divide-by-max* normalization might be misleading, because it cannot be deduced that if a value is twice as high, the effect would be twice as significant.

As the ϕ value measures the strength and not the significance of an association it is better suited for our purposes. As long as the sample is large enough and representative the values will be accurate. In addition, they lie in the interval]-1,+1[and thus do not have to be normalized.

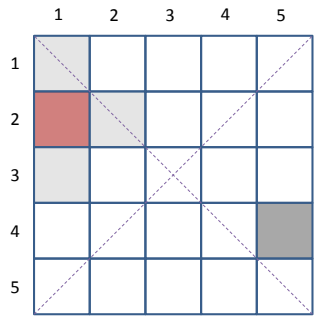
Matrix Arrangement

To make the relations between vowels with similar behavior maximally visible, it is essential to sort the rows and columns of the matrices in a meaningful way. Vowels showing similar behavior should be grouped into blocks, a task that is closely related to *blockmodeling* [133] in graph analysis. As in our case the matrices are rather small, typically languages contain between 5 and 8 vowels, all possible sortings can be tested, even though in principle it is a NP-hard problem. As all different sortings can be tested, what we need is an optimality criterion in order to automatically determine which of all possible sortings is the optimal one to be presented to the user. To enable a sorting of matrix cells, first of all, a numerical dissimilarity between matrix cells needs to be calculated. Hence, a distance function is required that quantifies the dissimilarity of two matrix cells, i.e. the dissimilarity of their ϕ-values. Different distance functions were created and empirically tested. Among the tested versions the most satisfying results were achieved with the distance function provided in Formula 3.5. The rationale behind the formula is that pairs of cells x and y with different algebraic signs are considered rather dissimilar.

$$d(x,y) = \begin{cases} 1 & \textit{if } sign(x) \neq sign(y), \\ (x-y)^2 & \textit{else.} \end{cases} \tag{3.5}$$

One constraint we enforce during the sorting is that the row and column orders of vowels have to be identical. 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.



Next, we defined characteristics for an optimal sorting and mapped it to a function in order to calculate the quality of a certain matrix sorting. For each matrix cell, first, the similarities to its directly adjacent cells are taken into account in order to get blocks of similar cells. In the displayed example we focus on the red cell, for which the adjacent cells are colored in light gray. In addition, we also consider the similarity of a cell to its *centroid reflection*. The latter is identified performing a point reflection through the matrix centroid. The centroid reflection of the red cell is colored in dark grey. Among different tested strategies that try to guarantee a certain symmetry within the whole matrix, using point reflections turned out to be the most favorable option. Of course, each relationship between two cells is only taken into account once during the process of evaluating the whole matrix.

Visualization and Visual Analysis

The numerical matrices generated with the described analysis methods are 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. Most importantly, the matrix rows and columns were sorted according to vowel similarity in order to make patterns become visible.

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 achieve many distinguishable color shades a bipolar color scale was chosen, ranging from bright yellow to dark blue (see Figure 3.14 for an example showing the *nextProb* values represented by color).

Alternatively, the *nextProb* values can be displayed as bar charts within the

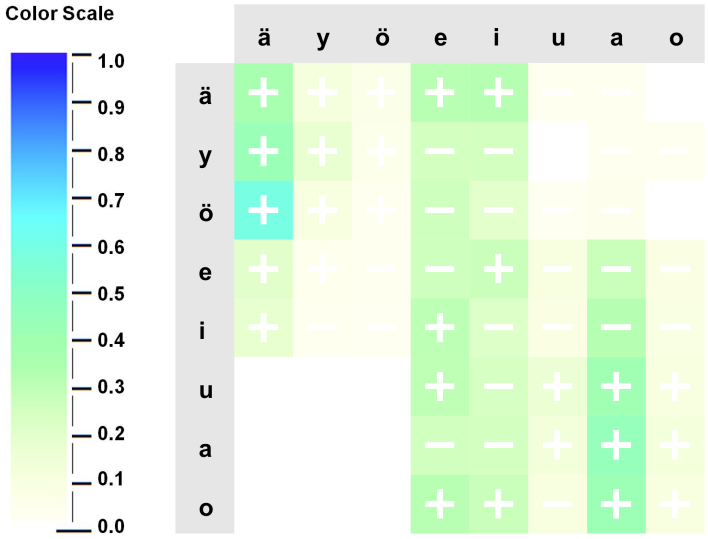


Figure 3.14: The visualization represents the probability matrix with $nextProb$ values for the Finnish Bible types that has been sorted automatically. The + and − signs indicate whether a vowel succession occurred more or less frequently than expected when assuming vowel independence. One interesting finding that can be deduced from the visualization is that there are two blocks of vowels that almost never combine, viz. the block {ä,y,ö} and the block {u,a,o}.

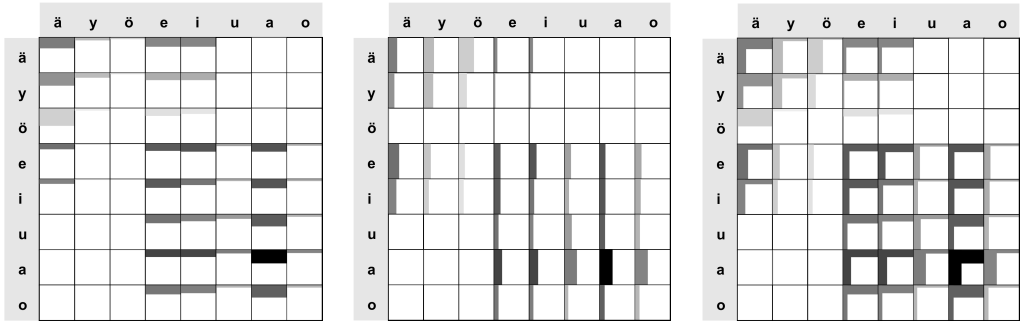


Figure 3.15: The matrices showing the *nextProb* values (left), the *rightProb* values (middle), which have been introduced in Section 3.2.4, and both values at once (right). Again, it becomes visible that the block {ä,y,ö} and the block {u,a,o} avoid each other. In addition, some other interesting effects caused by the varying overall frequencies of different vowels become evident. For example, if we observe an ö it is likely that the previous vowel was an ä (transition ä -▷ ö in middle matrix) and the next vowel is an ä (transition ö -▷ ä in left matrix). However, if we observe an ä, the likelihood of observing an ö before or after is quite low (transitions ö -▷ ä in middle and transition ä -▷ ö in left matrix).

matrix. Each row contains bars that grow from top to bottom and the height of the bar indicates the probability, ranging from 0% (no bar) to 100% (whole matrix cell filled by the bar). The probabilities of all horizontal bars in one row will sum up to 100%. The color saturation of a bar indicates the amount of data on which the calculation of the probability is based. The more data the more saturated the bar. If a bar has a low saturation the observed effect may be less reliable. The leftmost matrix in Figure 3.15 shows the example for the Finnish matrix.

In analogy to the *nextProb* values, the *prevProb* values can also be displayed as bar charts within the matrix. In this case, each column (not row) contains bars that grow from left to right and the height of the bar indicates the probability, ranging from 0% (no bar) to 100% (whole matrix cell filled by the bar). The probabilities of all vertical bars in one column will sum up to 100%. The matrix in the middle of Figure 3.15 shows the example for the Finnish matrix.

Both kinds of bars, of course, can also be integrated into one visualization. The rightmost matrix in Figure 3.15 shows the example for the Finnish matrix. We name such a matrix a probability-bar matrix.

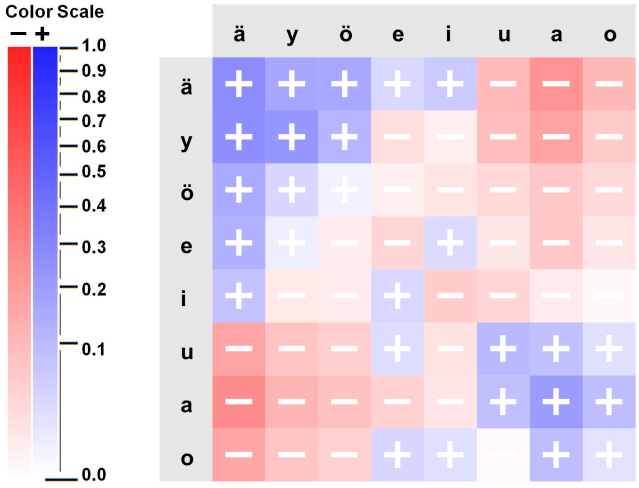


Figure 3.16: The visualization represents the ϕ matrix for the Finnish Bible. In this case the "+" and "−" symbols provide a redundant mapping. Now, blocks of vowels that belong together can clearly be seen. As before, {ä,y,ö} build one block, {u,a,o} another independent block, and {e,i} cannot unambiguously be assigned to any of them. In fact, this conforms nicely to the categorization linguists have for Finnish vowels: {u,a,o} are back vowels, {ä,y,ö} are front vowels, and {e,i} are neutral vowels, which explains why they do not adhere to one of the blocks.

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 absolute ϕ value was, the more saturated 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 3.16 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 3.17 reveals that many interesting features are no longer clearly visible without sorting.

Comparative Analysis of Vowel Patterns

When performing the described analysis for a large number of different languages vowel harmonic patterns become easily visible (see Figure 3.18). Apart from Maori and Tagalog, all of the top 7 languages actually contain different

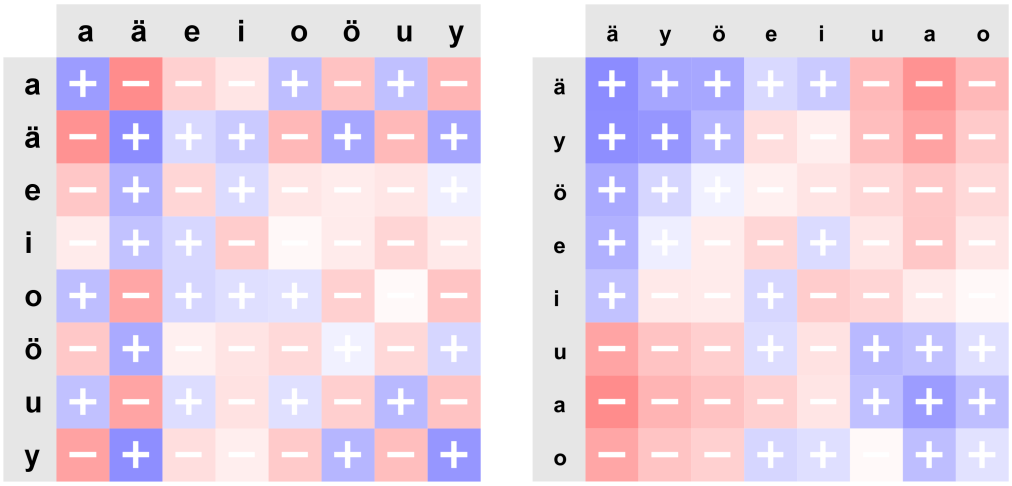


Figure 3.17: The left visualization has a default vowel sorting (alphabetical order) and shows no easily perceivable pattern at all (reprinted from [119], ©2010, CSLI Publications). The right matrix which was automatically sorted, in contrast, reveals that there exists an interesting pattern.

kinds of VH. The strongly colored diagonal in Maori stands out and is actually not due to VH per se, but to a process of syllable reduplication, which leads to a statistically salient amount of vowel self successions. The strongest effect can be perceived in Turkish which is known to have rather strict and complex harmony patterns that are rendered clearly visible with our approach. More details on succession probabilities and absolute vowel frequencies can be gathered from the probability-bar visualization (see Figure 3.19). In this visualization, patterns are harder to discover than in the ϕ matrix visualization, but it contains more detailed information. While the ϕ matrix visualization is more perceptually effective, the probability matrix visualization is more expressive. Both complement each other well.

We could observe that languages containing VH usually have a heavily skewed data distribution in the matrix of absolute succession counts (as in Table 3.1). This observation is a potentially good starting point for the operationalization of VH. Hence, the degree of deviation from equal distribution within the matrix of a language can indicate the degree of VH tendencies within that language. We measure the statistical significance of the deviation for each language applying the likelihood ratio test to the whole matrix of absolute succession counts. Three points are important to consider:

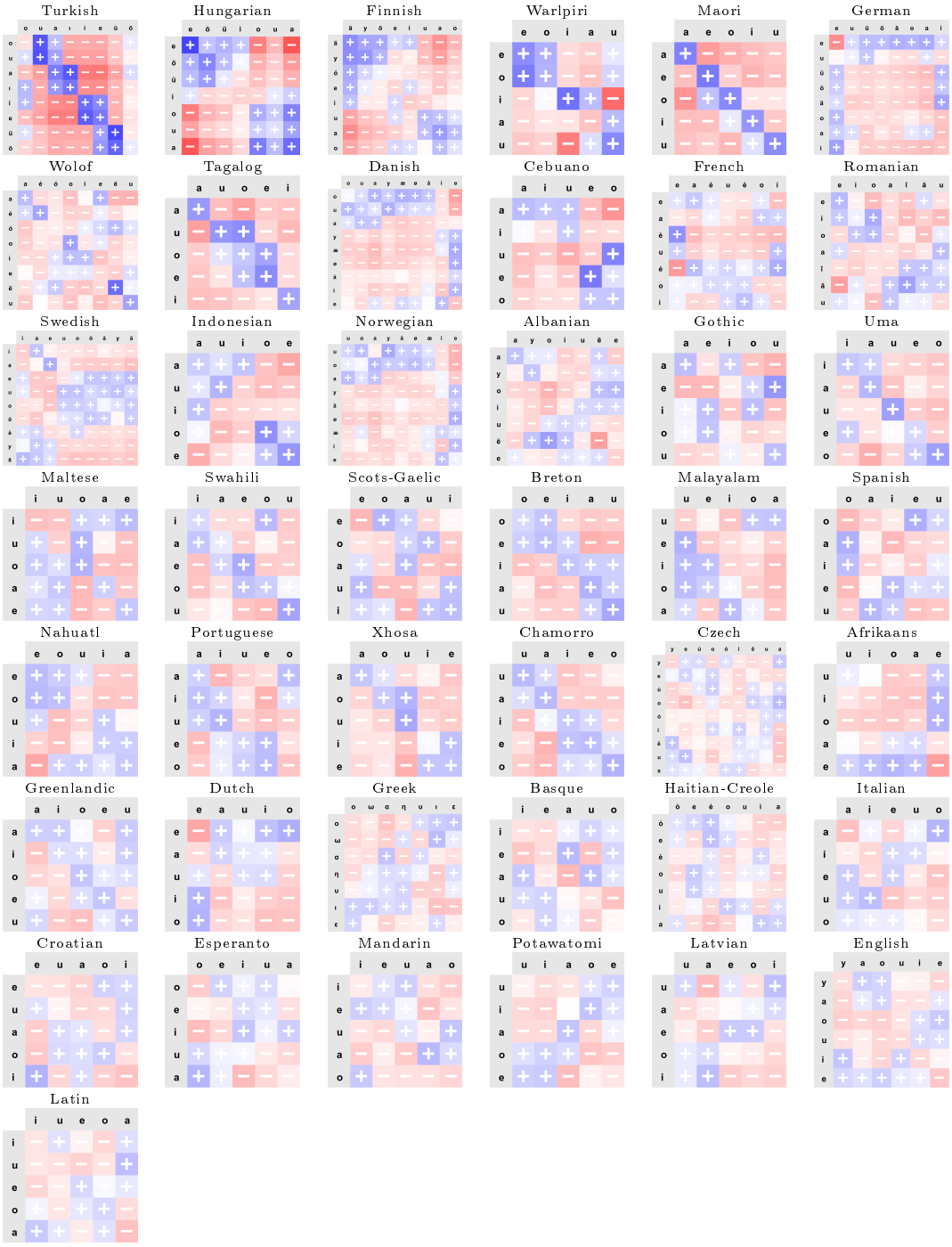


Figure 3.18: The ϕ matrices for 43 languages ordered according to decreasing log-likelihood ratio values (as displayed in Figure 3.20) from left to right and top to bottom.

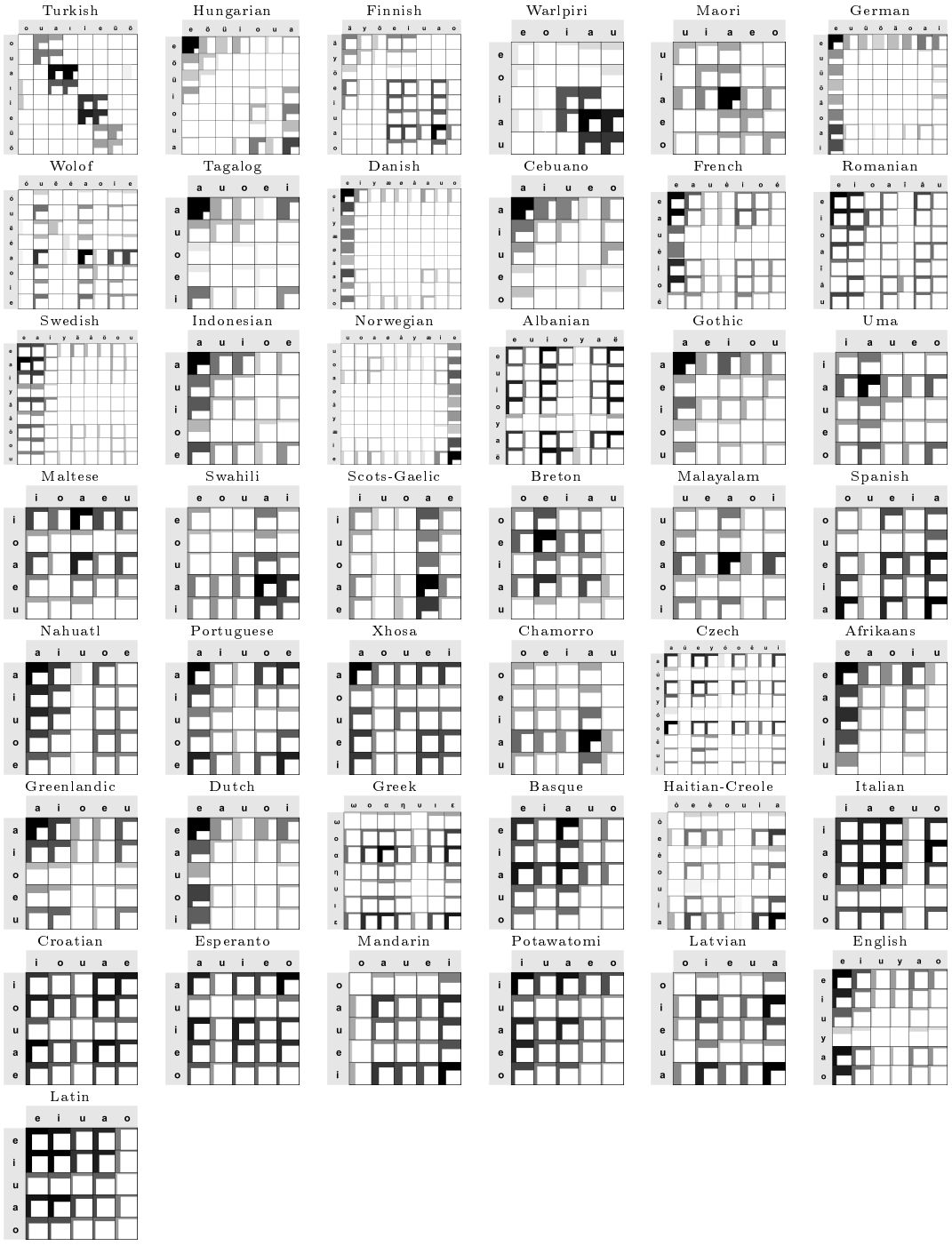


Figure 3.19: The probability matrices for 43 languages ordered according to decreasing log-likelihood ratio values (as displayed in Figure 3.20) from left to right and top to bottom.

1. Everything is highly significant, because for all languages the vowel distributions are far from being random. Therefore, it does not make sense to look up the resulting values of the test in a significance table — p will almost always be close to 1. Despite of that it makes sense to compare the likelihood ratio values relatively to one another. The higher the value is, the stronger is the VH tendency.

2. The more vowels a language has, the more degrees of freedom we have in the statistical test. The same absolute likelihood ratio value becomes less and less significant the more degrees of freedom we have. As vowel harmonic languages tend to contain more vowels, we still decided to compare the absolute values.

3. The amount of data influences the significance of an effect. Consequently, we normalize all matrices as if we would have observed exactly the same number of vowel occurrences for each, in our case 1000.

Figure 3.20 shows the distribution of the likelihood ratio values. Some languages stick out: First, the three well-known VH languages Turkish, Hungarian, and Finnish. Less strong effects are contained in Warlpiri and Maori, which also is consonant with what is known about their linguistic structure as will be detailed in Section 3.2.6. The results suggest that this operationalization of VH could be useful and constitute a good *quasi-semantic feature* indicating the degree of VH contained in a language. As the feature consists in a single numerical value its distribution in the context of genealogy and geography, and its correlation with other features could be explored with the methodology suggested in Section 3.1.

3.2.5 Evaluation: Minimum Amount of Data Required

Unfortunately, for many less well-documented languages there are no large-scale textual resources — like the Bible — available. With the purpose of finding out whether the proposed Visual Analytics approach can be applied to a wider range of languages, the scalability of the introduced statistics and visualizations was systematically tested for smaller text fragments. We examined empirically how many different words (types) are required so that we can derive reliable knowledge about vowel harmony in a language.

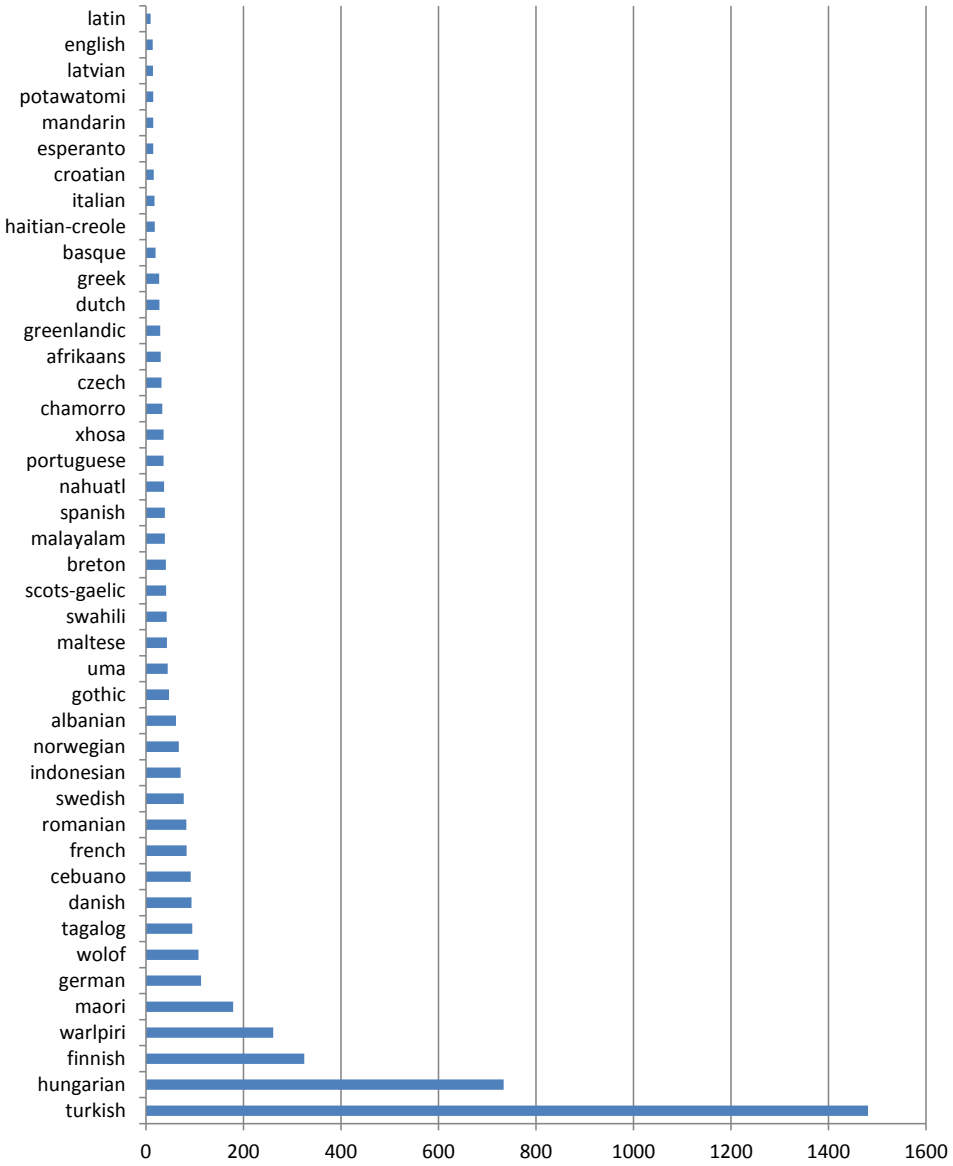


Figure 3.20: Likelihood ratio values for the matrices of absolute language succession counts.

To do so, the resulting ϕ-matrices obtained by using a whole Bible[17] were defined to be a gold standard. Then, it was tested how close the results obtained using smaller textual resources approached this gold standard. For each language we drew random type lists from the Bible, ranging from 10 up to 1500 types — in intervals of 10 types.

In a next step, for each obtained type list a ϕ matrix was created in order to compare it to the previously created gold standard ϕ matrix. Then, for each amount of types — from 10 to 1500 — the mean deviation of matrix entries from the gold standard was calculated. In Figure 3.21 this mean deviation is plotted on the y-axis against the number of types on the x-axis.

As can be seen, about 500 to 1000 types are already sufficient to achieve a good convergence. The mean deviation of matrix entries (ϕ values) for almost all languages lies below 0.03 for an amount of 500 random types. The convergence depends on a set of different influence factors like the word lengths, the number of vowels a language has and of course also the presence of vowel harmony. The well-known vowel harmonic languages Turkish, Hungarian and Finnish are among the languages that have a rather fast convergence (see green lines in Figure 3.21). This confirms the assumption that if there is a clear vowel harmony present in a language, the effect will already be visible in short texts.

3.2.6 Case Studies: In-depth Cross-linguistic Investigations

In this section some of the results and findings for a larger number of languages are discussed. Figure 3.18 shows the ϕ-matrices of all languages for which we had a suitable Bible corpus at our disposal. The languages are ordered according to the strengths of the effects they contain from left to right and top to bottom according to the automatically determined likelihood ratio values, which indicate the degree of VH. Vowel harmonic languages tend to have rather saturated blue blocks along the first diagonal of their matrix (left top to right bottom) and rather saturated red blocks on the inverse diagonal. Most languages in the first two rows feature this at least to a certain extent. Yet, from the second line on diagonal orientation begins to vanish slightly and

[17] Depending on the language under consideration the number of types in the Bible ranges from 2,000 to 70,000.

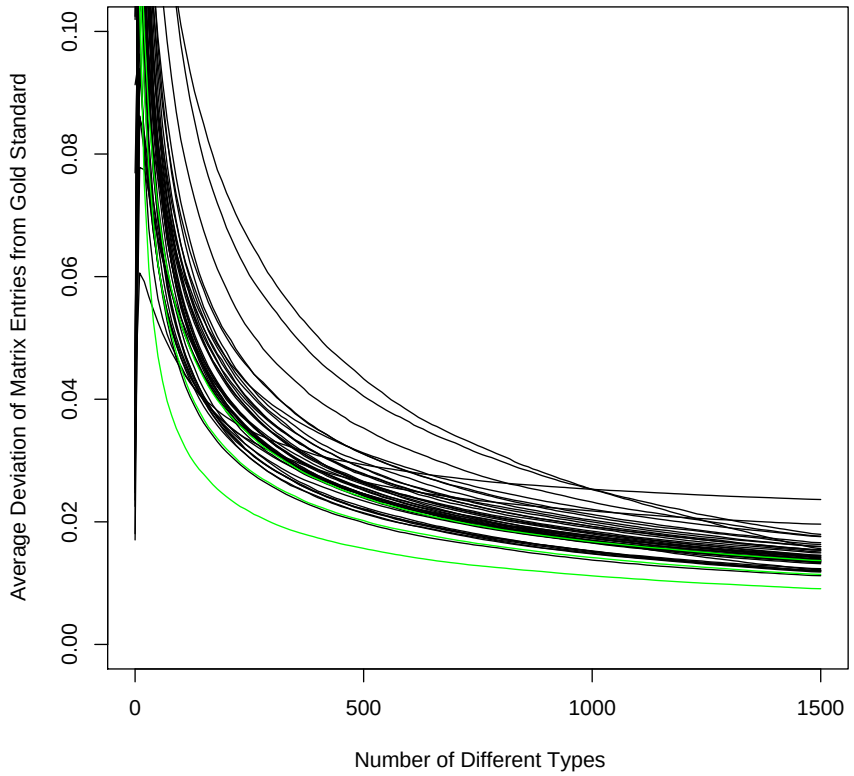


Figure 3.21: This plot shows the mean deviation of the ϕ matrix entries for smaller type lists from the gold standard entries (whole Bible type list). In order to smooth the curves and reduce the clutter we took the average curve of 1000 trials. In total it shows the results for 44 languages, whereby the vowel-harmonic languages Turkish, Hungarian, and Finnish are colored in green. Turkish has the fastest convergence among all languages. Reprinted from [119], ©2010, CSLI Publications.

the color saturation becomes less intense. Although there may be some effects visible in the further matrices surely the first couple of matrices are the most interesting ones.

A look at the first couple of probability matrices (see Figure 3.22) can also reveal interesting information. It shows that languages known to have strict vowel harmony (like Turkish, Finnish and Hungarian) are very prohibitive with respect to non-harmonic vowel successions. They simply do not occur and therefore large very bright areas appear in the upper right and lower left parts of the matrices. In Turkish the effect again is so strong that vowel harmony could be detected from the probability matrices only.

Case Study: Turkish and Finnish

The information that domain experts can read from the matrices is as follows:

> "*The Turkish matrix shows the palatal harmony as two complementary blue blocks in the /a/- and /e/-columns whereas the labial harmony clusters are represented as adjacent 2-cell blocks [...], indicating that there are no neutral vowels in the harmony processes present. The fact that the rows of the matrix can be filled twice with blue blocks shows that two harmony processes (labial and palatal) are active in the language.*
>
> *The Finnish matrix shows a less clear-cut picture. Nevertheless two main blocks (in the upper left and bottom right corner) are visible and illustrate the harmony clusters [...]. Unlike in Turkish, the harmony blocks are separated by two rows and columns in the middle of the matrix (representing the vowels /e/ and /i/), which indicates that the harmony contains neutral vowels. Remember that the matrix rows and columns have been sorted automatically and have not been arranged with the knowledge of which blocks should stand out.*" [119][18]

Case Study: Warlpiri and Maori

While Turkish and Finnish were known to the domain experts for containing vowel harmony, other languages revealed different interesting patterns in their

[18] Part of our joint publication written by Thomas Mayer.

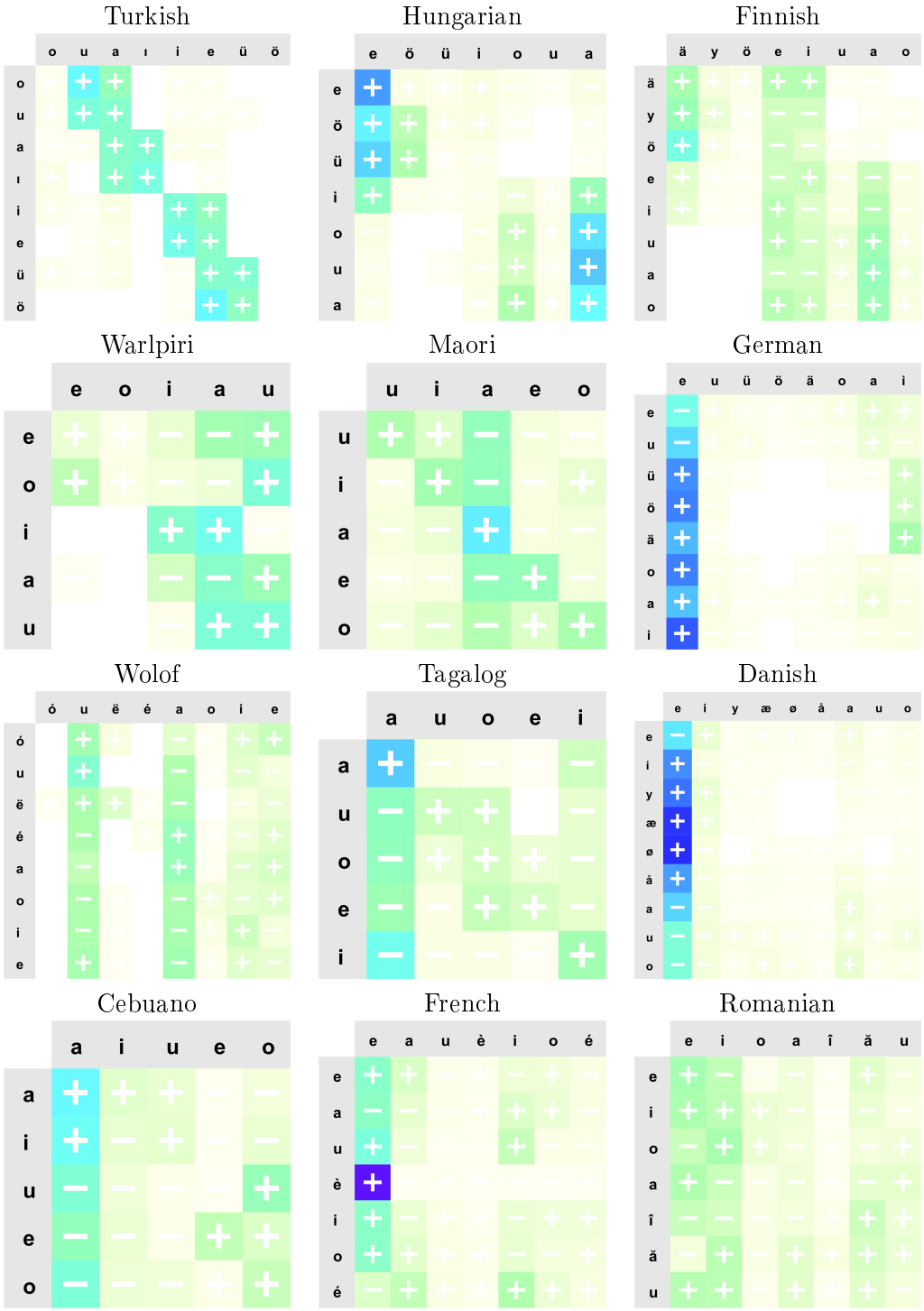


Figure 3.22: The probability matrices of the 12 top ranked languages. For languages known to have a strong harmony like Turkish, Finnish and Hungarian the matrices contain empty blocks in the upper right and lower left. This is another visual characteristic that indicates vowel harmony because certain vowel successions are prohibitive in vowel harmonic languages.

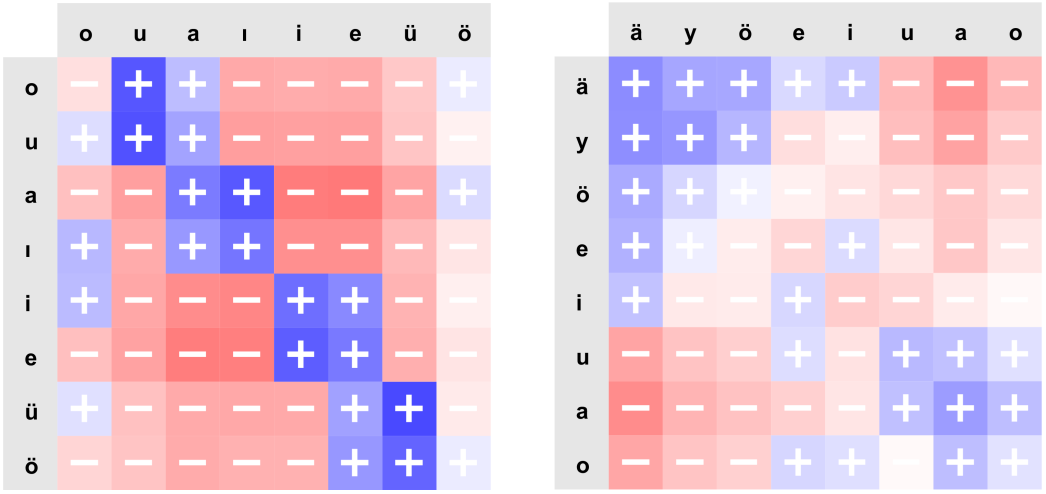


Figure 3.23: The ϕ-matrix for Turkish (left) and Finnish (right)

vowel distributions. Two examples are Warlpiri and Maori, where the domain experts made the following observations:

> "*In the ϕ-matrix of Warlpiri (see Figure 3.24) there is a conspicious block that involves the letters /u/ and /i/. Both vowels are not likely to occur together in words but rather have themselves as successor vowels. [126, p. 84] describes Warlpiri as having vowel harmony (both regressive and progressive) only involving the vowels /i/ and /u/.*[19] *Verbs with root-final /i/ change it to /u/ if the past tense suffix -rnu is added (regressive assimilation). Progressive assmiliation changing /u/ to /i/ shows up with a large proportion of the nominal suffixes and enclitics. Consider the following words with the corresponding suffixes (see [126, p. 86]):*
>
> 1. *kurdu-kurlu-rlu-lku-ju-lu*
>
>   *child*-PROP-ERG-*then-me-they*
>
> 2. *minija-kurlu-rlu-lku-ju-lu*
>
>   *cat*-PROP-ERG-*then-me-they*
>
> 3. *maliki-kirli-rli-lki-ji-li*
>
>   *dog*-PROP-ERG-*then-me-they*

[19] Notice that Warlpiri has a very small vowel inventory of only three vowels /a, i, u/. The occurrences of /e, o/ is due to loanwords or proper names from English.

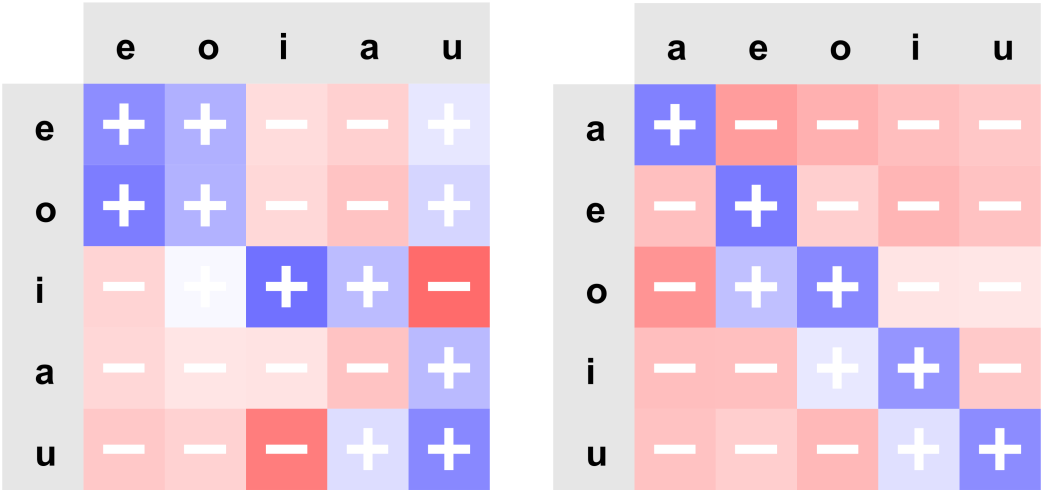


Figure 3.24: The ϕ-matrix for Warlpiri (left) and Maori (right)

> *As can be seen in 3, all subsequent suffixes change their vowels to /i/ if the last vowel of the stem is /i/.*
>
> *Warlpiri therefore shows harmony in both directions. However, since our calculations do not take into account the direction of harmonic spreading, both harmony processes strengthen the results in Figure 3.24. The positive and negative cells in the matrix clearly show the non-cooccurrence of /i/ and /u/.* " [119][20]
>
> "*Maori does not have VH, but an effect not altogether dissimilar is produced by its morphology of reduplication. Partial reduplication affixes an abstract CV syllable, and these segments are then specified through features of the corresponding CV segments of the base. In contradistinction to VH, the vowels of the reduplicand and the base are in fact identical, rather than only sharing the harmonic features; also, there are no patterns of dispreferred vowel sequences in the case of reduplication alongside the preferences.*" [119][21]

Case Study: Udihe

In this case study we wanted to test whether the methodology was not only suitable for confirming or illustrating existing knowledge, but could potentially

[20] Part of our joint publication written by Thomas Mayer.
[21] Part of our joint publication written by Thomas Mayer.

also lead to the generation of new knowledge. For this purpose we gathered a short text fragment with a length of only 2450 words from an almost extinct language called Udihe, which is estimated to have only 230[22] speakers. Since it belongs to the Altaic family of languages, and other Altaic languages like Turkish contain vowel harmony, it was likely that Udihe might also contain such an effect. However, we were not aware if this was really the case and what a potential vowel harmony in Udihe could look like. To explore the language we applied our approach to the small text fragment, which according to the stability tests performed in Section 3.2.5 should already be sufficient to detect 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 among all tested languages after Turkish. This indicates that a strong effect like vowel harmony is present in the language.
- A look at the probability matrix (Figure 3.25) 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 as can be seen in Figure 3.25 (left). Here, the effect is not as clear as for the other vowel harmonic languages.

In Figure 3.25 we left out the vowel /ü/ because it appeared only in 3 successions within the corpus, so that no reliable statistics about it could be derived. It has been learned from previous observations that both probability- and ϕ-matrix are important in order to track vowel harmony. 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, this is the case for the transitions (o-▷o) and (ö-▷o) as well as (ä-▷a) and (a-▷a) as can be seen in Figure 3.25. 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.

[22] http://www.ethnologue.com/show_language.asp?code=ude revised on Feburary 10th, 2012

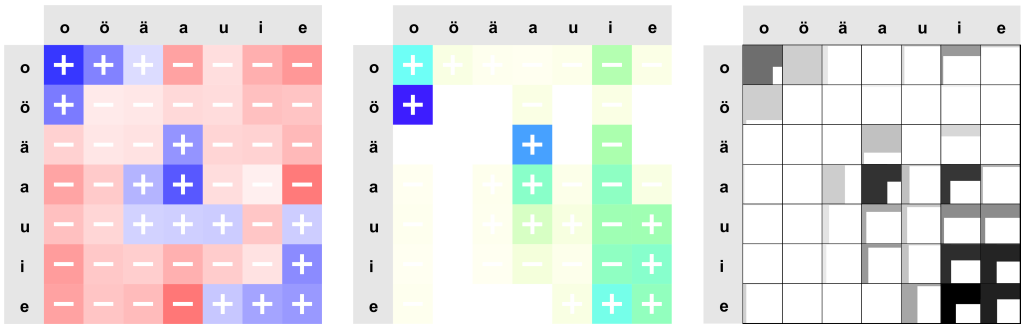


Figure 3.25: The ϕ matrix (left), colored *nextProb* matrix (middle), and probability-bar matrix (right) for the Udihe text fragment containing about 2450 words.

In both matrices the same block in the /e/ column is salient and indicates the harmonies (u-▷e), (i-▷e) and (e-▷e). The only further feature that is slightly conspicuous is the succession (u-▷a), but the effect is weaker than for (u-▷e). Table 3.3 summarizes these findings.

trigger		successor vowels
a, ä, (u)	→	a
o, ö	→	o
e, i, u	→	e

Table 3.3: Hypotheses about probable harmonies in Udihe

It has to be remarked that the hypotheses shown in Table 3.3 were deduced without any prior knowledge about Udihe except that it could be expected to find vowel harmonic patterns there. The Udihe text was just fed into our program and the original text was not even looked at during hypothesis generation.

The accuracy of the results is very satisfying. The predicted harmony patterns found in the visualizations (Table 3.3) correspond to what grammarians find in their analyses [131, p. 74].
This example shows that it is possible to quickly generate accurate hypotheses about vowel harmony in languages without reading a single word.

Case Study: Tracking Language Development

It is also possible to treat word beginnings and word endings in the same way as vowels. While this incorporation of word limits weakens the statistical effect of word internal vowel successions, it can reveal other interesting connections. Figure 3.26 provides an example where some interesting similarities and differences are pointed out and numbered in the visualization:

1. For Norwegian /e/ is the only letter that is less frequent after word beginnings than expected. For Swedish this is the case for two letters /e/ and /a/.

2. For Norwegian, /e/ is much more probable to be the last vowel in a word than expected and for Swedish /e/ and /a/ are much more probable to be the last vowel in a word than expected.

3. In Norwegian, after any vowel, /e/ is more probably observable than expected — except after /e/ itself. In Swedish, after any vowel, /e/ and /a/ are more probably observable than expected — except after /e/ (and /a/).

As both languages are closely related this fact indicates that at some point in time they might have developed differently. Thus, from the visualization one could derive the hypothesis that one of both languages had an innovation that the other one did not have.

This example suggests that even though the matrices are calculated from contemporary resources they can possibly also reveal information about language development. As the current state of a language is the result of a development process, detailed insights into today's language can help to understand more about language change.

This can also be observed when looking at the German matrix, where domain experts could find another trace of historic change:

> "*Even though the harmonic process of umlaut is no longer active in the language and the former triggers for umlaut (/i/ and /j/ in the following syllable) have mostly disappeared in the relevant environments due to weakening processes in unstressed vowels, the general pattern is still visible. As can be seen in Figure 3.27, the umlauted vowels*

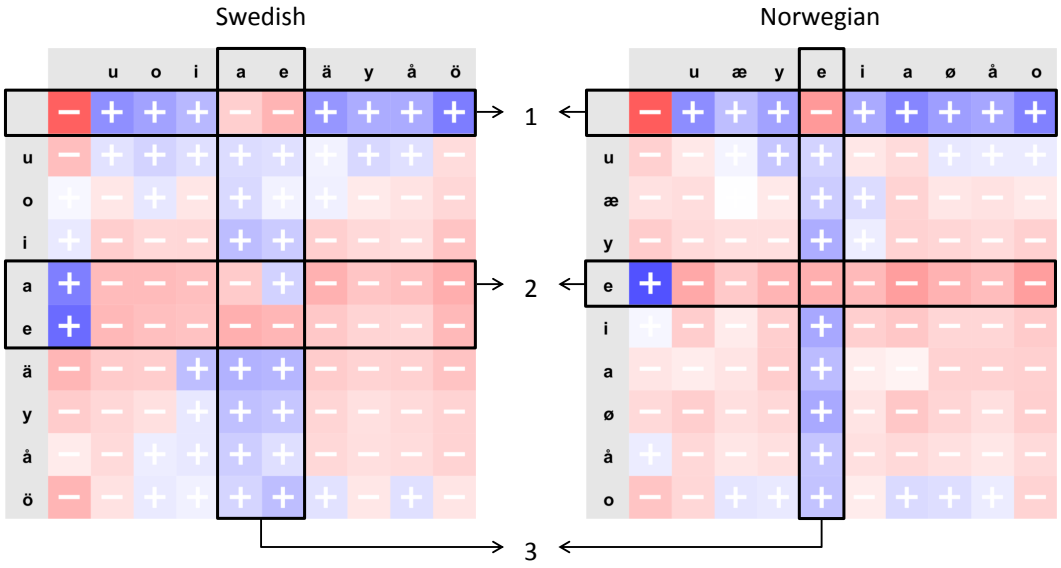


Figure 3.26: In this example word beginnings (first row) and word endings (first column) are incorporated. The left matrix was generated from the Swedish Bible and the right one from the Norwegian Bible. Three visual findings are numbered.

> */ü, ö, ä/ occur more frequently before the vowel /i/ than before other vowels (except /e/, which is the most frequent successive vowel for all vowels). There are only a few suffixes left that still have an /i/ and trigger umlaut at the same time (e.g., -in as in Französin as compared to Franzose, -ig as in völlig as compared to voll or -lich as in köstlich as compared to Kost). However, with respect to the whole distribution of vowels the pattern can still be detected although German orthography does not reflect the pronunciation of words properly.*" [119][23]

Conclusion of the Case Studies

The case studies suggest that the novel methodology, combining statistical analysis methods with visualization, can support the discovery of language patterns that otherwise might not become evident. It could be shown that existing knowledge about languages could be confirmed and additional insight was gained from getting an appropriate visual representation of vowel succession patterns. While the technical sophistication of the matrix display is

[23]Part of our joint publication written by Thomas Mayer.

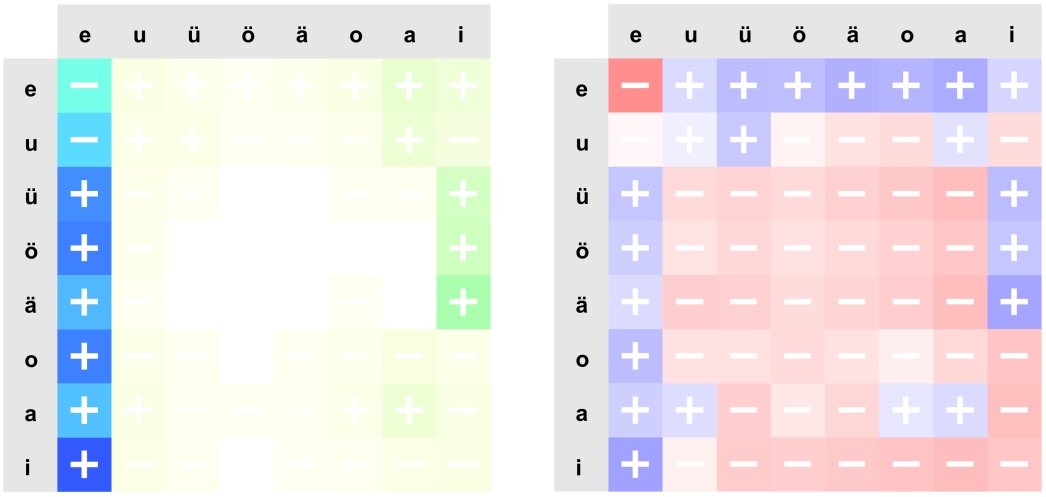


Figure 3.27: The probability matrix (left) and the ϕ-matrix (right) for German.

limited, the technique is easy to apply and easy to interpret. This enables quick test runs on other data with immediate visual feedback and makes it a good means for exploratory investigations and hypothesis generation. Some examples will be provided in Section 3.2.7.

3.2.7 Extended Use for Hypothesis Generation

This subsection builds on my contribution to the following publication:

T. Mayer, C. Rohrdantz, F. Plank, P. Bak, M. Butt and D. A. Keim. Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. Proceedings of the ACL 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING 2010), pages 67-75, 2010.[24]

This subsection contains a case study on how the use of the previously introduced visual analytics technique can be extended to related analysis tasks and further data sources and support the detection of unexpected patterns,

[24]Thomas Mayer wrote most of the publication, I contributed the visualization and statistical analyses. Peter Bak contributed the geo-spatial visualization. Miriam Butt proof-read the text and gave advice. Frans Plank and Daniel Keim also gave advice. Further people that we also acknowledge in the publication are Aditi Lahiri and two anonymous reviewers for valuable comments and suggestions. For all parts of the publication that were not written by me and that I will use in this thesis, I reference the original work.

that may in turn lead to the generation of new hypotheses. It shows how the use of visual analytics can enrich research in a domain like linguistics.

Analyzing Consonant Patterns

After analyzing vowel patterns a next intuitive step was trying to apply the same methodology to consonant patterns. The idea of the domain experts was to investigate languages for a phenomenon called *Similar Place Avoidance* (SPA), which had been claimed to be a universal tendency. With the goal to support or reject the claim, we applied our technique to more than 4500 languages.

In simple words, the SPA theory says that successive consonants within words, i.e. consonants only separated by vowels, tend *not* to have a similar place of articulation. Usually, three or four places of articulation are distinguished:

> "*One approach (PTCK) is based on the arrangement in Pozdniakov and Segerer [140] and distinguishes four places of articulation for labial (P), dental (and alveolar) (T), (alveo-)palatal (C) and velar (K) consonants. A second grouping (LCD) only distinguishes three places of articulation: labial (L), coronal (C) and dorsal (D)*" [119][25]

Table 3.4 [120][26] shows how the different places of articulation can be mapped to the Automated Similarity Judgment Program (ASJP) [180] orthography and the *International Phonetic Alphabet* IPA.

For more details about SPA and linguistic background I would like to refer the interested reader to the original publication [120].

The data used came from the ASJP and has already been described in Section 3.1.3. An advantage of the data set is that it contains the basic vocabulary for more than 4500 languages and thus enables a broad investigation. Furthermore, the ASJP orthography is a phonographic writing system, i.e. letters can be directly mapped to sounds independent from the language. A disadvantage is that per language only a limited number of words and with it a limited number of consonant-consonant-successions is given. Considering the lack of

[25]Part of our joint publication written by Thomas Mayer.
[26]The table was provided by Thomas Mayer as part of our joint publication.

LCD	PTCK	ASJP	IPA
L	*P*	p, b, m, f, v, w	p, ɸ, b, β, m, f, v, w
C	*T*	8, 4, t, d, s, z, c, n, S, Z	θ, ð, n̪, t, d, s, z, ts, dz, n, ʃ, ʒ
	C	C, j, T, l, L, r, y	ʧ, ʤ, c, ɟ, l, ʎ, ɭ, ʎ, r, ɾ, j
D	*K*	5, k, g, x, N, q, G, X, 7, h	ɲ, k, g, x, ɣ, ŋ, q, ɣ, χ, ʁ, ħ, ʕ, ʔ, h, ɦ,

Table 3.4: Assignment of consonants to symbols reprinted from [120]. All varieties of "click"-sounds have been ignored.

other data sources we still came to the conclusion that it would be useful to run our experiments on this data, keeping the disadvantage in mind.

The first thing we did is to join the consonant-consonant-successions aggregated over *all* languages and visualize them. The assumption was that if SPA was really universal, then it should be observable in a joint set of all languages. At the same time, we tested whether it was useful to distinguish the two different subcategories dental (and alveolar) (T), and (alveo-)palatal (C).

Figure 3.28 shows the resulting association values ϕ of place successions. It can be seen very clearly that T and C behave very similarly. A further interesting observation is that in the joint data for all languages places of articulation actually tend to alternate (negative diagonal values for self-successions). As revealed in the succession graph of Figure 3.28, the places of articulation do not remain the same, but change to the closest alternative(s). In the case of P and K the closest distinct places of articulation (T and C) are preferred. In the case of T and C, however, this is somewhat different. Apparently, direct alternations between both are less probable. One plausible explanation could be that they are not distinct enough and thus either K or P are preferred as a following place of articulation, both having roughly the same distance. These observations led us to merge the places T and C in our further analyses and distinguish labial, coronal, and dorsal consonants only, as in Figure 3.29.

Note that the cross pattern on the left in Figure 3.29, which now emerges

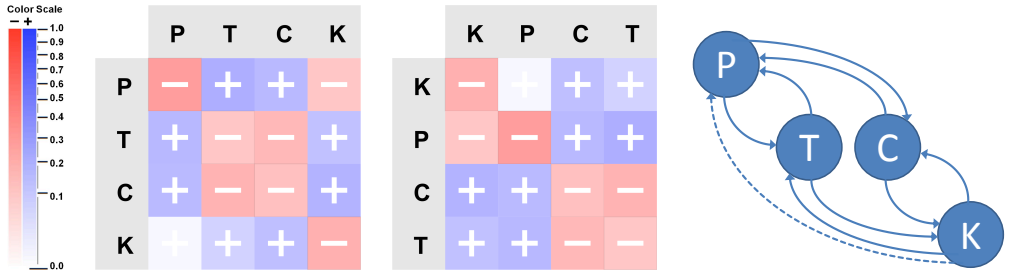


Figure 3.28: Successions of P, T, C, and K in all languages. The + and − signs indicate the polarity of a succession (going from row to column category). The color saturation of the background indicates the strength of association. In the left figure, places of articulation are sorted according to their position in the oral cavity, in the middle figure an automatic similarity sorting of matrix rows and columns was applied. The right part of the figure shows an alternative view, where only on those successions are displayed that have a positive association. Reprinted from [120], © 2010 Association for Computational Linguistics.

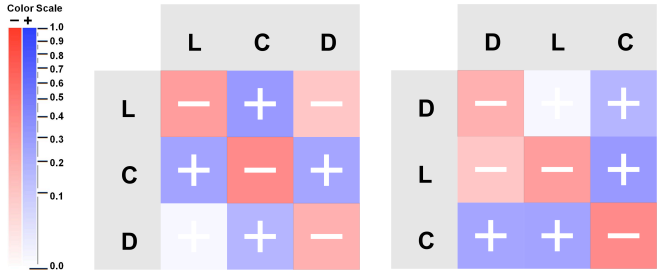


Figure 3.29: The ϕ matrix considering only the three main categories for all the data across languages. In the left figure, the categories are sorted according to their position in the oral cavity. In the right figure, the categories are sorted automatically, which shows that D and L are more similar to each other than D and C. Reprinted from [120], © 2010 Association for Computational Linguistics.

very clearly, reinforces the hypothesis that the closest distinct place of articulation is preferred as successor.

Still, the open question remained whether SPA applies to every individual language. The answer we can give is "most probably yes". The uncertainty lies in that for some languages we have only very few consonant-consonant successions and can thus not derive reliable statistics. However, the more data we have for languages, the clearer the tendency shows up, which is illustrated in a series of simple plots.

We examined the distribution of ϕ values for self-successions of places of articulation in about 3,200 languages. Self-successions correspond to the diagonal values of the ϕ matrices from the upper left to the lower right. As can be seen in the histogram in Figure 3.30, the peak of the distribution is clearly located in the area of negative association values. In the box-plots of Figure 3.31, which show the distributions for all three places of articulation separately, it is clearly visible that for each of the three places of articulation at least 75% of the languages included show negative associations. Furthermore, it can be seen that most outliers disappear when taking only the languages for which most data is available and thus statistics are more reliable. The same can be seen in the scatter plot in Figure 3.32, where the average ϕ value is always negative if the number of successions exceeds a certain threshold. For all three categories, the figures demonstrate that the same place of articulation is generally less frequently maintained than expected if there were no interdependencies between consonant occurrences.

3.2.8 Beyond Binary Sequences: Using Droplet Maps for Visualizing Vowel Patterns

The *Droplet Map* is a visualization technique for sequences that was originally designed in order to show movement sequences in geo-spatial data. It stems from David Spretke, Patrick Jungk, and Peter Bak, who developed it jointly in the Data Analysis and Visualization Group at the University of Konstanz.[27] Together with the inventors of the technique I worked on applying and adapting the technique to analyze vowel sequences.

We used the same data extracted from Bible texts as in the previous investigations on Vowel Harmony, see Section 3.2.2. In contrast to the matrices, Droplet Maps can display sequences of more than two items. Consequently, more detailed distributional patterns and also information about typical word lengths and word endings might be revealed.

The visual mapping is as follows:

- Parallel vertical lines mark the positions of vowels within a word. In

[27] The technique is also described in Patrick Jungk's Master Thesis that seems not to have been published.

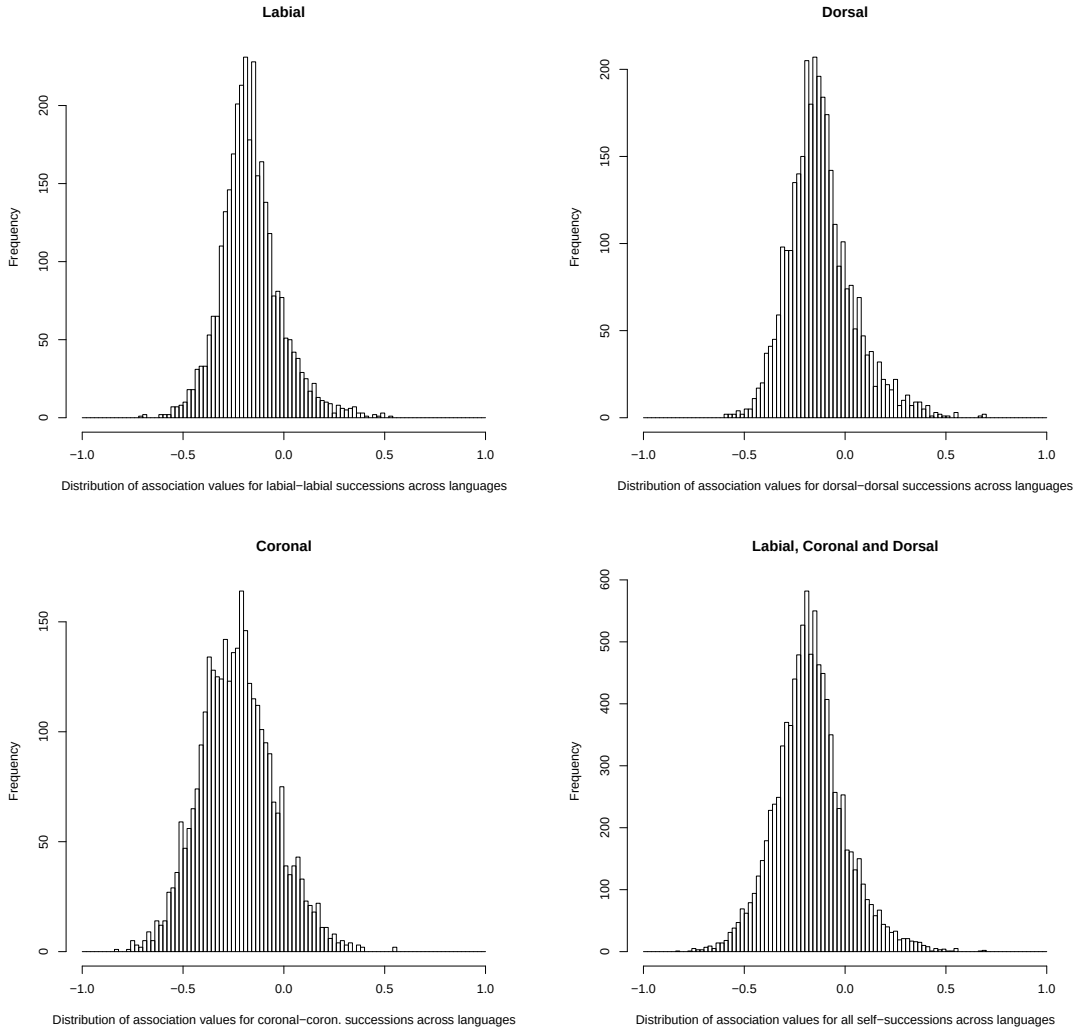


Figure 3.30: Histograms showing the distribution of association strength values (ϕ) for self-successions of places of articulation in more than 3200 languages. Lower right subfigure reprinted from [120], © 2010 Association for Computational Linguistics.

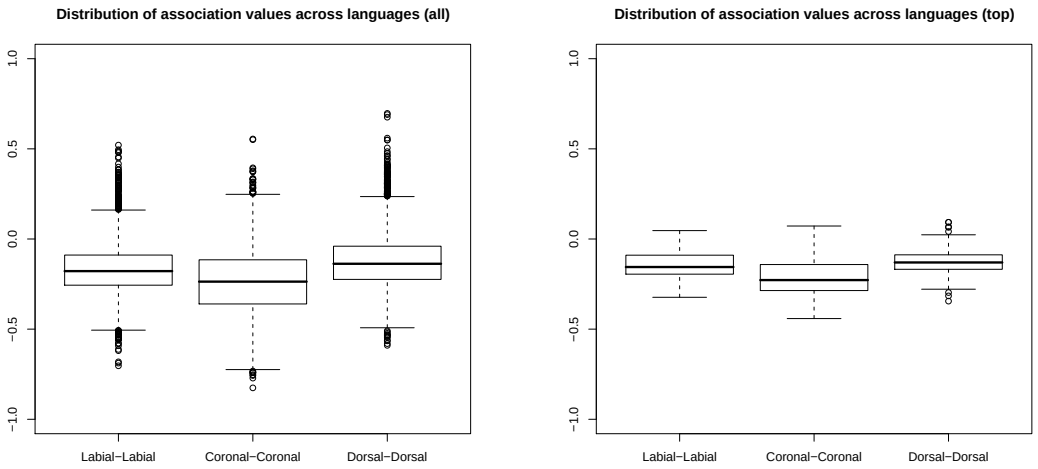


Figure 3.31: Boxplots showing the distribution of association strength values (ϕ) for self-successions of places of articulation. For the left boxplots about 3,200 languages were considered for which the Swadesh lists contained more than 20 successions. For the right boxplots only the top 99 languages were considered for which the Swadesh lists contained at least 100 successions, thereby removing most outliers and reducing the variance. The visualizations support the hypothesis that positive ϕ values may only be due to random effects when not having enough data for a language. Reprinted from [120], © 2010 Association for Computational Linguistics.

many languages there are not more than 5 or 6 vowels in a word, i.e. mostly 5 or 6 lines are displayed.

- Each vowel is represented by a different color. Most languages do not contain more than 8 vowels, which is still a number of colors that should be easily distinguishable.

- Whenever a certain vowel appears at a certain position, within a word of a language, the corresponding color will appear as a rectangle at the corresponding axis. The more words that have the vowel at the position, the bigger the corresponding rectangle will be.

- In fact, there are typically two adjacent rectangles of one color plotted on a vertical line. One left to the line having a size proportional to the number of incoming vowel transitions and one right to the line having a size proportional to the number of outgoing vowel transitions. The right rectangle is necessarily smaller or equal to the left one. A large difference

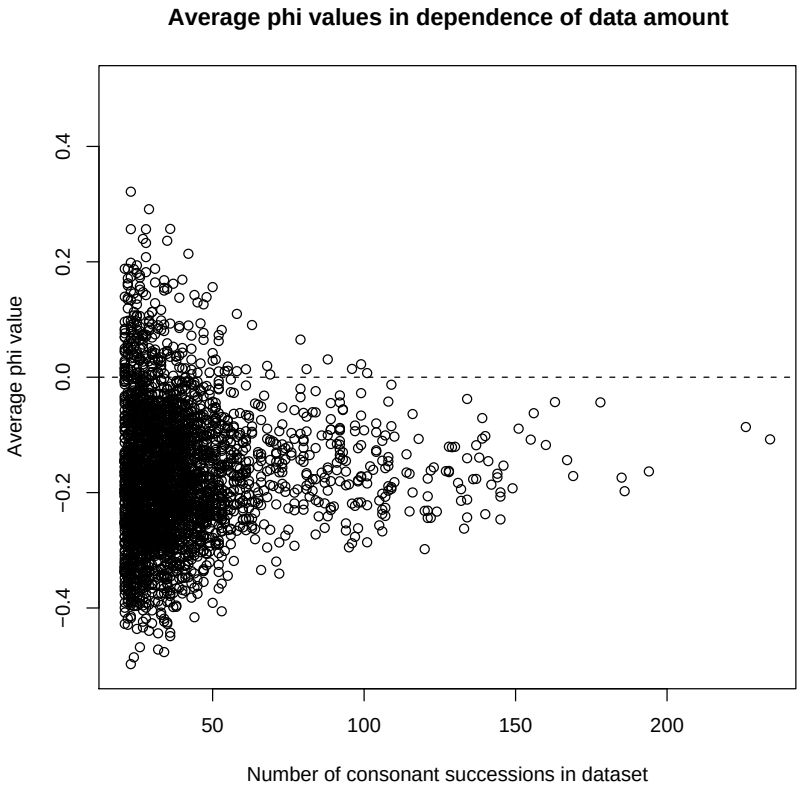


Figure 3.32: The scatter plot displays the average ϕ values for self-successions of all places of articulation depending on the number of consonant successions (CVC) for each language in the sample. About 3,200 languages were considered for which the Swadesh lists contained more than 20 successions. Reprinted from [120], © 2010 Association for Computational Linguistics.

in size indicates that a lot of words end with the corresponding vowel at the corresponding position.

- The number of incoming and outgoing transitions is additionally plotted as a line, connecting the rectangles of adjacent vertical lines. The thickness of the line depends on the number of corresponding transitions. For example, if many of the words of a language have an i at position 2 and next an e at position 3, then there will be a thick line between the outer right rectangle of line 2 with the color of i and the outer left rectangle of line 3 with the color of e.

- The vertical positions of the rectangles are heuristically determined in order to minimize the number of line crossings.

One example for a language with a complex vowel harmony system is Turkish shown in Figure 3.33. Two similarly behaving subgroups of vowels can be detected that barely interact with each other: The back vowels *ı*, *a*, *u*, *o* and the front vowels *i*, *e*, *ü*, *ö*. Each vowel in a subgroup patterns similarly to a vowel in the other subgroup, where *ı* corresponds to *i*, *a* to *e*, *u* to *ü*, and *o* to *ö*. The transitional restrictions within the subgroups can be explained by harmony constraints.

Similar languages can also be compared easily: Figure 3.34 shows that Swedish words usually end with the vowels *e* and *a* to similar extents, while Norwegian words end almost exclusively with *e*. As already mentioned in Section 3.2.6 this difference between the two closely related languages is a result of language change.

Linguistic Expert Study As in this case domain experts had not been involved in developing the visualization a rather informal expert study should reveal whether the method was useful to them. I asked three linguistic researchers to use *Droplet Maps* to explore vowel patterns in different languages. The experts were asked to individually select 5 out of 44 languages and formulate their hypotheses about the vowel patterns of each language a priori, given that they had any expectancies. The experts picked Afrikaans, Finnish, French, German, Gothic, Hungarian, Indonesian, Maori (2), Nahuatl (2), Spanish, Swahili, Turkish, and Wolof.

In general, it could be observed that the experts needed some time to become familiar with this kind of visual representation. From the second and third language on their exploration became a lot quicker and more confident as they could compare the current language to already seen ones. It could be observed that the experts were able to solve different kinds of tasks with the help of *Droplet Maps*:

- Researching influences one language might have had on another one. One expert explored whether Nahuatl might have undergone influences from Spanish, but no relations between both languages became visible. In contrast, when Afrikaans was examined for its relationship to Dutch, it became obvious that both languages shared most characteristics of their vowel distributions.

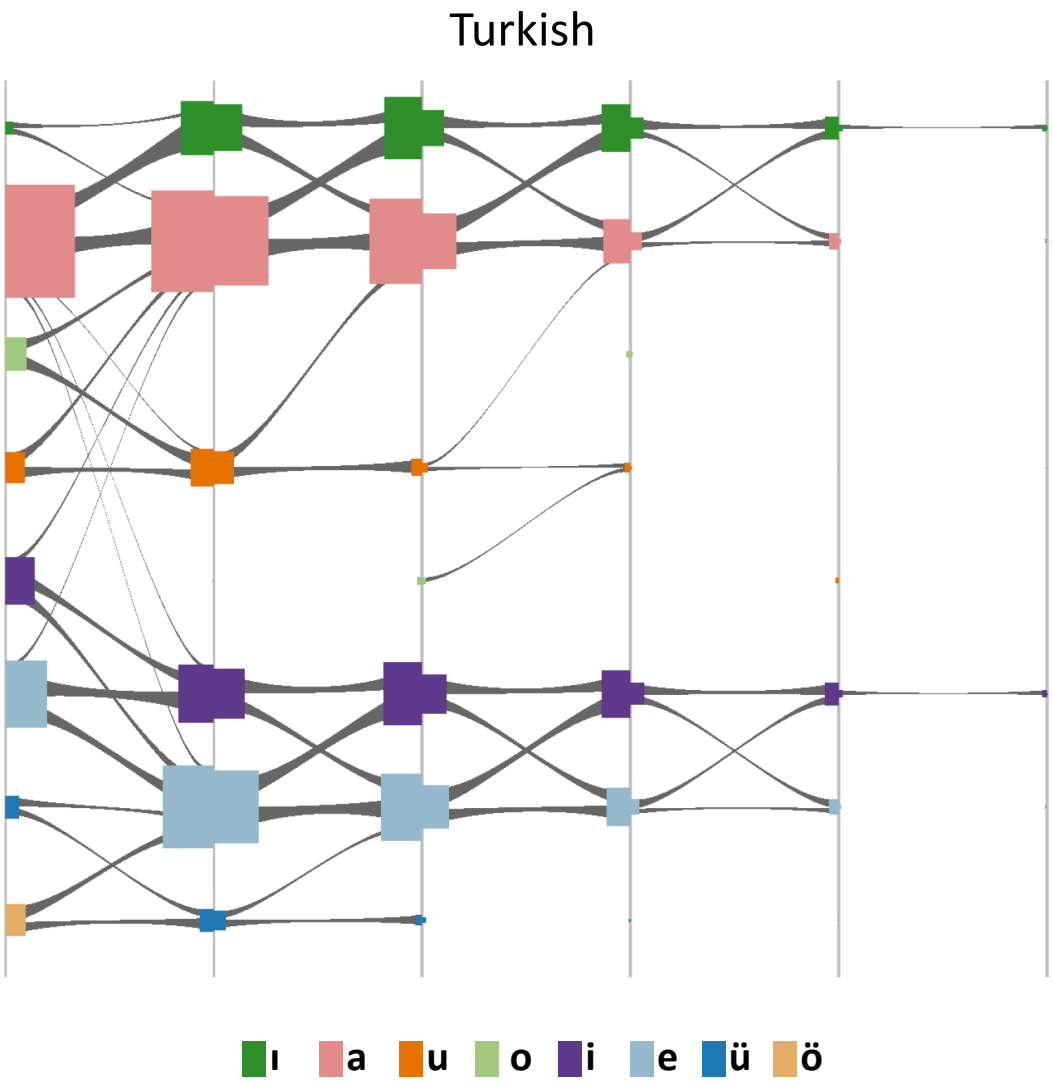


Figure 3.33: Turkish vowel transitions pattern clearly as Turkish contains a highly structured vowel harmony. Only transitions were plotted that are based on at least 200 Bible types, thus minimizing the noise introduced by proper names, borrowings, and foreign words, etc.

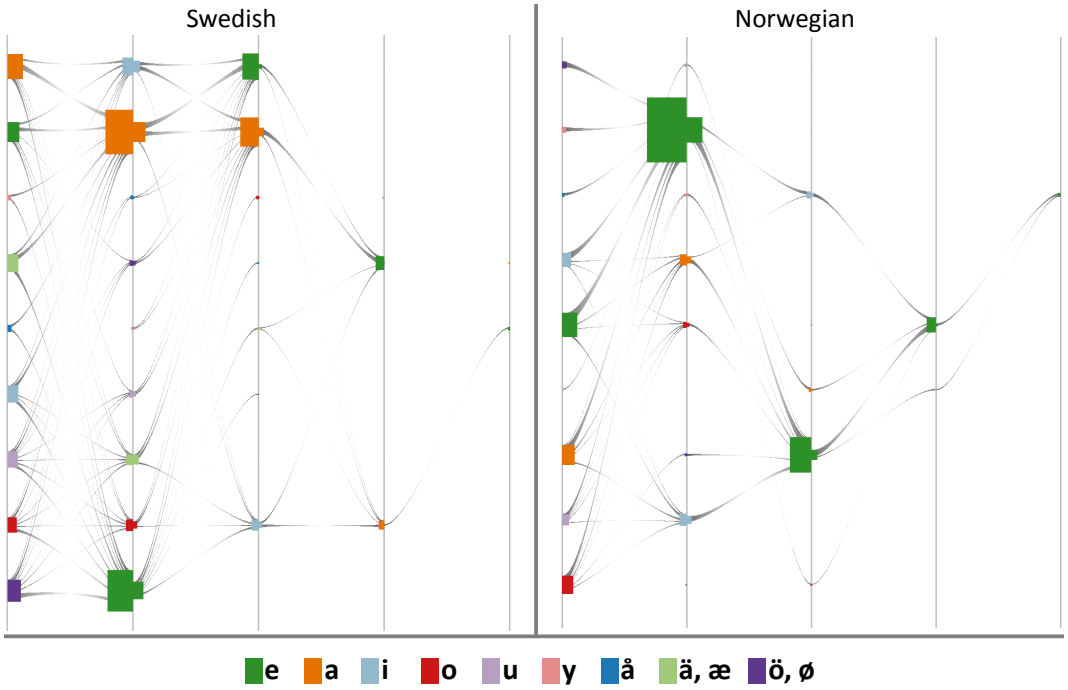


Figure 3.34: Comparison of Swedish and Norwegian vowel sequences where only the most prominent links, based on more than 50 Bible types, are displayed. Swedish words usually end with *a* and *e*, Norwegian words with *e*.

- Recognizing vowel harmony in Finnish, Hungarian, and Turkish including detailed observations about vowel clusters and neutral vowels.
- Recognizing suffix and prefix patterns. In German, French, and Afrikaans the vowel *e* heavily dominated word endings. In Swahili, vowels occurring in person and time prefixes (*a*, *i*, *u*) were remarkably frequent at the word beginnings.
- Recognizing syllable reduplication in Maori, where the same vowel tended to follow again rather than other vowels.
- Recognizing heavy reliance on single vowels, e.g the prominent role of the vowel *a* in Indonesian and Maori, or the vowel *e* in German and Afrikaans.
- Recognizing the tendency of a language to be agglutinating by the length of typical vowel sequences.

The experts also discovered two limitations and improvement potentials of the method. First, digraphs have been partly ignored in the preprocessing, by

considering only the first of the two vowels. That means, whenever two vowels were not separated by at least one consonant, the second vowel was ignored. This is not a problem of the visualization as such, but of the preprocessing. However, it lead to biases in the visualization of languages with frequent digraph occurrences like German and especially Gothic and Wolof. A possible solution is to work with phonetic transcriptions of words instead of orthography. Another drawback of the current method is that positive correlations among succeeding vowels are much more evident than negative correlations. This is due to the fact that the eye is drawn to links between vowels and not intrigued so much by the absence of particular links or the absence of clutter in general. One way to partly overcome this potential disadvantage is to provide the option to draw links between vowels based on statistical measures expressing, for example, a strong negative association. Using such measures instead of taking only absolute occurrence counts, yet, comes at the cost of introducing further clutter.

3.2.9 Discussion and Conclusion

One of the lessons learned from the research described in this chapter is that when designing visualizations for domain experts it is very important to understand the practitioner's perspectives and needs. For example, it is quite valuable if a visualization is simple to understand, that is, the visual detection and interpretation of patterns should be easy for persons not used to working with visualizations.

With respect to this criterion the matrix visualization was quite successful as domain experts were able to use it and to transfer it to other related tasks. For them, in many cases, it is much more important that they can readily process their data and get immediate visual feedback, than to have a large number of options to configure and manipulate the visualization.

In addition to the case studies discussed, there are further application examples where the matrix display has successfully pointed to novel findings. The interested reader is referred to the original publication:

T. Mayer, C. Rohrdantz, F. Plank, M. Butt and D. A. Keim. A Quantitative Approach to the Contrast and Stability of Sounds. QITL-4 4th Con-

ference on Quantitative Investigations in Theoretical Linguistics, pages 59-64, 2011. [121]

Another important finding is that a visualization can only be as good as the data extraction, preprocessing, and automatic analysis. In earlier versions of the matrix display the data was visualized directly instead of first deriving statistics from it, like e.g. the ϕ values. The insight that could be achieved was very limited. Sorting also proved to be very important to make visual patterns emerge. Of course, there must be some basic assumption about the kinds of patterns that might be contained in the data, before the matrix can be sorted in an appropriate way.

Finally, one rather unexpected insight we gained is that sometimes the *absence* of data or data values can be heavily interesting, however, usually only the *presence* of data or data values results in visual representations. Displaying also negative associations in the matrix turned out to be a good means to highlight the absence of a data.

Chapter 4

Visual Analytics of Diachronic Change in Lexical Semantics

Contents

Trying to track phenomena of language change by automatically analyzing historical documents brings several problems:

1. Many ancient documents have not been digitalized yet and training Optical Character Recognition (OCR) systems for ancient hand writings is challenging.

2. Existing digitalized historical corpora like the *Penn Corpora of Historical English*[1] are heavily biased both with respect to time and sources. The further back in history, the fewer sources and evidence are available. Most old documents were authored by clerics and noblemen and the content and vocabulary use hardly reflects the everyday speech of the common people in that times.

3. The lack of standardized orthography even within documents of the same author makes it hard to discern the same lexical units across documents and time.

Considering the current data situation there is little hope that automated analytics methods can be of great support for linguistic researchers investigating phonological, morphological or syntactic change in diachronic corpora. Such methods will have to remain limited to search and retrieval tasks and the big share of the investigative work will have to continue to be done manually.

Nevertheless, there are phenomena of language change that are observable even in contemporary corpora of rather short time spans, namely changes in lexical semantics. This refers to changes in word meaning which can be identified and tracked by analyzing changes in the context of a word, following Firth's famous quote "you shall know a word by the company it keeps" [53].

Section 4.1 will show how methods originating from the field of topic modeling can help in characterizing the temporal development of word meaning and especially the appearance of new word senses.

In continuation, Section 4.2 deals with the emergence and spread of new suffixes and the lexical semantics of new coinages made based on these suffixes.

[1] http://www.ling.upenn.edu/histcorpora/ last revised on Nov. 27th, 2012

4.1 Tracking Change in Word Meaning through Topic Modeling

This section builds on the following publication:

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, Frans Plank. Towards Tracking Semantic Change by Visual Analytics. ACL (Short Papers) 2011: 305-310. [145][2]

4.1.1 Background

There are different basic meaning changes that may occur with words. For example, words can acquire more positive connotations over time (amelioration) or negative connotations (pejoration). Words may as well acquire new meanings (semantic broadening) or lose certain aspects of meaning over time (semantic narrowing), see [34, p. 200 ff.] for more details.

The computational modeling of word senses is based on the assumption that the meaning of a *target word* can be inferred from the words in its immediate context, to which we will refer as *context words*. Research in this area mainly focuses on two related tasks: Word Sense Disambiguation (WSD) and Word Sense Induction (WSI).

The goal of WSD is to classify occurrences of polysemous words according to manually predefined senses. One popular method for performing such a classification is Latent Semantic Analysis (LSA) [39] as in the heavily cited paper *Automatic Word Sense Discrimination* of Hinrich Schütze [155]. However, other methods are also suitable for the task, see Navigli [127] for an extensive survey.

The aim of WSI is to induce different word senses of a target word from text corpora without presupposing certain senses. This goal is much more

[2]The publication was written in equal parts by Annette Hautli, Thomas Mayer and myself. While both of them did the linguistic research part, I did the computer science research part. Miriam Butt helped with the writing, proof-read the text, and gave advice. Frans Plank and Daniel Keim also gave advice. The programming was done by Zdravko Monov and myself. For all parts of the publication that were not written by myself I reference the original work.

difficult to achieve, as it is not clear beforehand how many senses should be extracted and how an abstract description of a sense could be automatically computed. Recently, however, Brody and Lapata [23] have shown that Latent Dirichlet Allocation (LDA) [20] can be successfully applied to perform word sense induction from small word contexts. Unlike in this existing work, we do not only label "each instance with the single, most probable sense" [23], but also take the probability distributions of contexts into account for analysis and visualization.

The original idea of LSA and LDA is to learn *topics* from documents, whereas in our scenario word contexts rather than documents are used, i.e., a small number of words before and after the word under investigation.

In the way we use LDA it does not typically assign one word context unambiguously to a certain sense (Figure 4.1), but assigns different probabilities to a word context as belonging to different senses. By having a large number of word contexts, it is possible to determine degrees of overlap among different senses, which can and do differ over time. In addition to each context having a probability for belonging to a certain sense, each word within that context is assigned to one sense. This means that a certain word context could be assigned to sense X with a high probability, while some of its individual words could be assigned to a different sense Y.

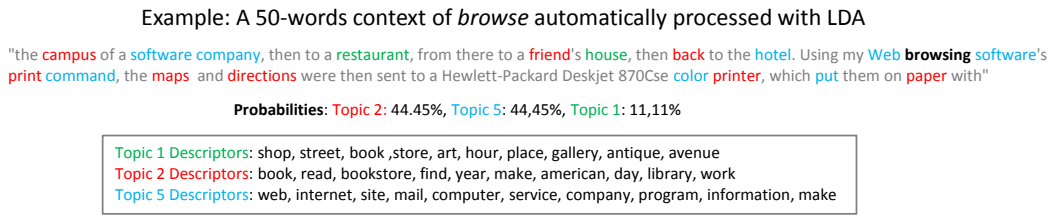


Figure 4.1: Example for automatically generated topics/senses for a word context. Each word in this context of *browse* was automatically assigned to different color-coded topics/senses. Consequently, the whole context can be assigned to different topics/senses with different probabilities. Characteristic terms describing one topic/sense are listed in the box.

More recently, researchers have been adding a diachronic component to the investigations on word senses, trying to detect and track changes in word meaning over time. Sagi et al. [153] have demonstrated that broadening and narrowing of word senses can be tracked over time by applying LSA to small

word contexts in diachronic corpora. Cook and Stevenson have investigated the semantic change types amelioration and pejoration, i.e. "a word sense changes to become more positive or negative, respectively" [33]. Based on different diachronic corpora the associations of a target word with positive and negative words was measured using *point-wise mutual information* and then compared for the two categories. A further approach from Heyer, Holz, and Teresniak [75, 80] uses a volatility measure adapted from the field of econometrics to assess changes in word context. However, they do not try to identify sense dimensions, but rather aim to find changes in news topics relating to a target word.

The outlined previous approaches for detecting diachronic changes in word senses are limited to general analyses, looking for example for narrowing or pejoration. In contrast, our aim is to go beyond the existing approaches by bringing together the two tasks of word sense induction and the tracking of semantic changes, resulting in a much more detailed analysis of change in word senses. The goal is to be able to find the point in time when a new word sense appears and to get a hint about its origin, i.e. to see whether it is related to one of the other already existing senses. In addition, we aim to enable a quantitative comparison of different prevailing word senses over time. In order to investigate the data and derived models in detail, we design suitable visualizations for exploration.

4.1.2 Data and Resources

For our investigations we use the New York Times (NYT) Annotated Corpus,[3] which contains 1.8 million articles from the daily newspaper editions between 1987 and 2007. While in this case the investigation is done in English, in principle it is language independent.

In order to be able to do meaningful analyses, we focus on investigating target words that are likely to have had changes in meaning within the 20 years under investigation. Most of the selected target words have acquired a new dimension of meaning due to the spread new technologies, like computers and the internet, e.g. *to browse* or *to surf*.

[3] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19, last revised on November 28th, 2012.

The following external resources where used for the implementation of the vector space reduction: the *singular value decomposition* SVD method for LSA was taken from the JAVA MATRIX PACKAGE JAMA,[4] *multi-dimensional scaling* [100] (MDS) was done with the package MDSJ[5] [138] and the toolkit MALLET[6] [122] was used for LDA.

4.1.3 An Interactive Visualization for Semantic Change

In order to investigate semantic change we designed a processing pipeline consisting of automated and visual methods. The automated data processing involved context extraction, vector space creation, and sense modeling. As Schütze [155] showed, looking at a context window of 25 words before and after a target word provides enough information in order to disambiguate word senses. Each extracted context is complemented with the time stamp from the corpus. To reduce the dimensionality, all context words were lemmatized and stop words were filtered out.

For the set of all contexts of a target word, a bag-of-words vector space was created and a global LDA model was trained. In the course of the experiments it became clear that the LDA topics/senses were much better interpretable than the dimension of the vector space after applying LSA (or MDS), see Section 4.1.4 for an example. Due to this, the temporal analysis was limited to the LDA topics, where each context is assigned to its most probable topic/sense Contexts for which the highest probability was less than 40% were omitted because they could not be assigned to a certain sense unambiguously. The distribution of contexts across different senses and the distribution of senses over time was then visualized.

Visualization

In order to visually analyze the development of word contexts over time, an interactive visualization tool was created, which displays the results of projecting words contexts in 2D using MDS, LSA or LDA. The tool contains a component with a separate visual representation for each target word occur-

[4] http://math.nist.gov/javanumerics/jama/ last revised on March 11th, 2013
[5] http://www.inf.uni-konstanz.de/algo/software/mdsj/ last revised on March 11th, 2013
[6] http://mallet.cs.umass.edu/ last revised on March 11th, 2013

rence (see Figure 4.3) and a component which provides aggregated views on the data (see Figure 4.4).

Plotting individual contexts

Figure 4.2 shows the initial view of our tool and gives an overview of the possibilities for exploring individual contexts, using LDA. The word under investigation is the verb *to browse*.

In the scatterplot, each context is represented by one dot. The axes correspond to LDA, LSA or MDS dimensions. In this case, 7 senses (dimensions) have been automatically learned and two at a time can be selected to be mapped to the axes for visual inspection. Here, the x-axis sense is characterized by the terms on top (shop, street, book, store, art, etc.). The further to the right a dot is situated, the more the corresponding word occurrence relates to the described x-axis sense. Accordingly, the y-axis is characterized by the terms on the left (software, microsoft, internet, netscape, window, etc.). The further to the bottom a dot is situated, the more the corresponding word occurrence relates to the y-axis sense. Contexts that belong to both senses are displayed along the diagonal of both axes. Yet, in this example screenshot there are no such cases.

As a further visual variable, the color of a dot indicates the time when the context appeared. The bipolar color map ranges from light green (year 1987) to dark purple (year 2007), optimized to contain a large number of distinguishable color tones. The color map on the right allows for arbitrary time intervals to be chosen with the sliders situated on the left (start-slider) and right (end-slider) of the color map, labeled with (a). Figure 4.2 shows contexts from 1987 to 1994. As can be seen, many word occurrences in this time relate to the x-axis sense, but only two strongly relate to the y-axis sense, labeled with (b). This can also be seen in an additional view where for each of the two selected sense axes the word occurrences (red dots) are plotted against time.

The context of a word occurrence can be displayed by mouse-over-interaction. Apart from these main features, the tool offers more options to optimize the visual display, including zooming, changing dot size and dot opacity, as well as strategies to reduce clutter.

Figure 4.3 shows further examples, where in each subfigure different com-

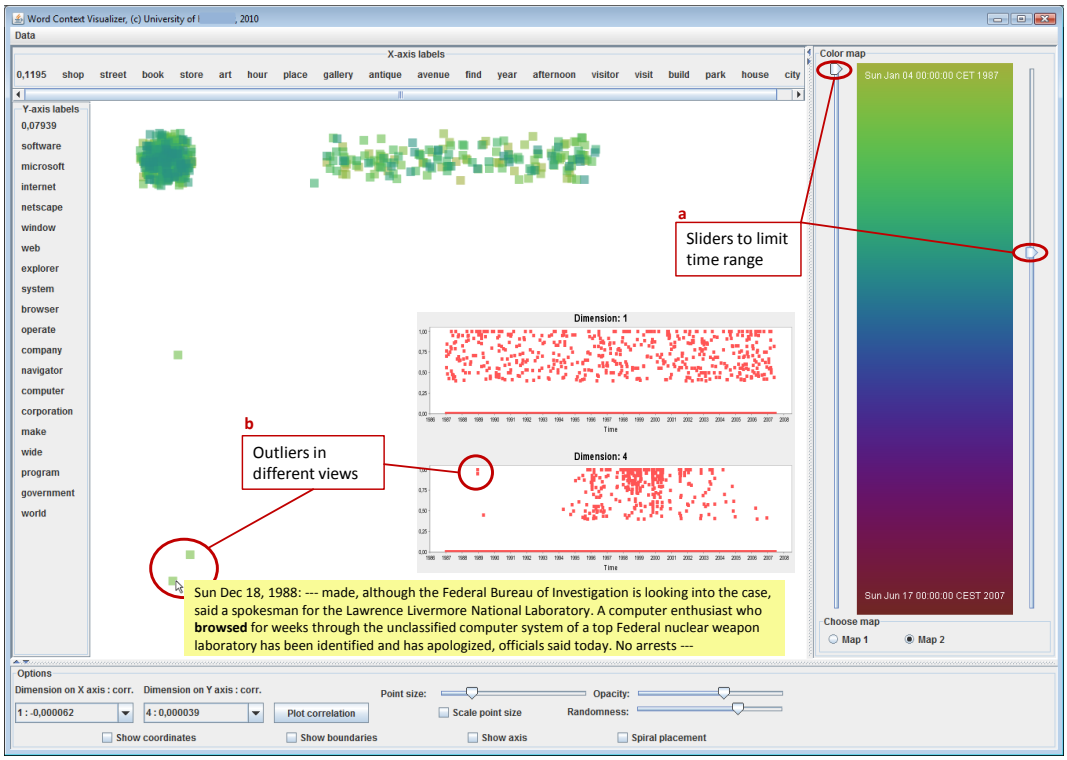


Figure 4.2: The word context visualization tool. Dimension 1 has been selected to be mapped to the x-axis and dimension 4 has been selected to be mapped to the y-axis.

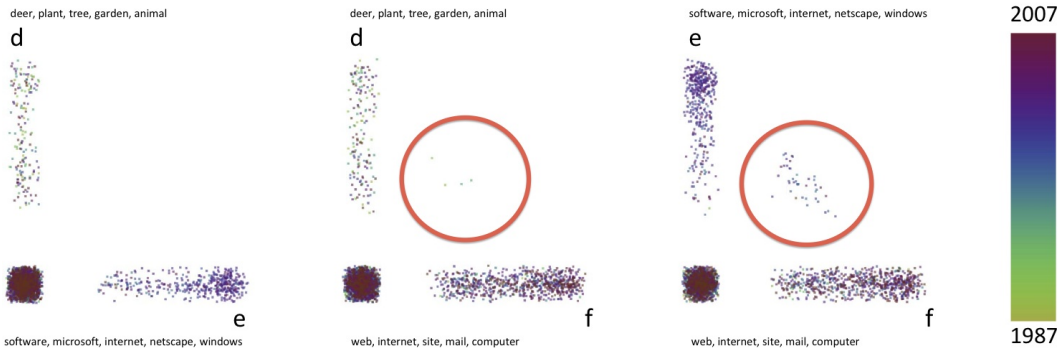


Figure 4.3: Pairwise comparisons of different senses for the verb *to browse*. In each subfigure different combinations of LDA dimensions are mapped on the axes. Reprinted from [145], © 2011 Association for Computational Linguistics.

binations of senses of *to browse* are plotted. In this case the y-axis goes from bottom to top. A random jitter has been introduced to avoid overlaps. Contexts in the middle (not the lower left corner, but the middle of the graph, e.g., see *e* vs. *f*) belong to both senses with at least 40% probability. In cases where the middle of the plot is populated with many data points, the axis senses share many ambiguous contexts can usually be considered to be similar. By mousing over a colored dot, its context is shown, allowing for an in-depth analysis. With the help of the time sliders analysts can filter the data for arbitrary time ranges to gain a better feeling for the diachronic development of the different senses.

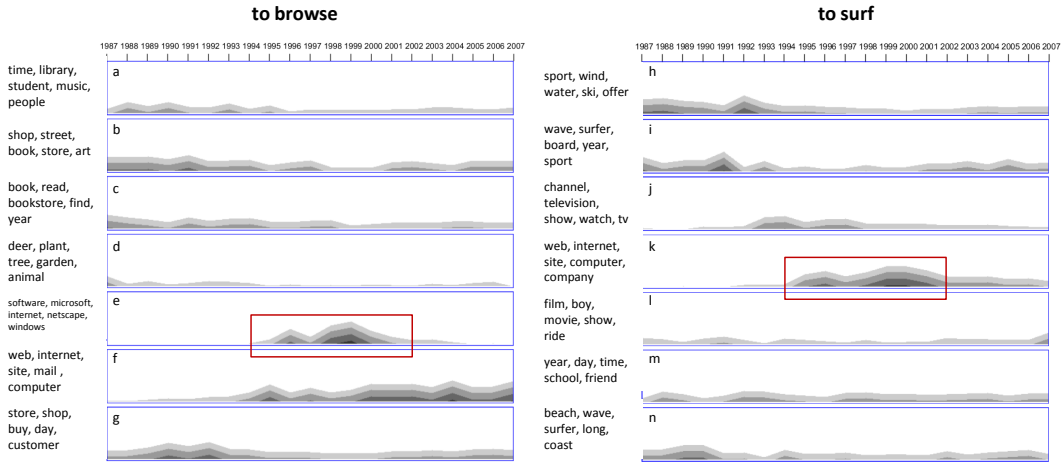


Figure 4.4: Temporal development of different senses concerning the verbs *to browse* (left) and *to surf* (right). Reprinted from [145], © 2011 Association for Computational Linguistics.

Aggregated data plotting

While plotting every word occurrence individually offers the opportunity to detect and inspect outliers and investigate the relatedness of different senses, aggregated views on the data are able to provide further knowledge on overall temporal developments.

Figure 4.4 provides another view of the tool, where the percentage of word occurrences belonging to the different senses is plotted over time. For the verbs *to browse* and *to surf* seven senses have been learned with LDA. Each sense corresponds to one line and is described by the top five terms identified by LDA. The higher the grey area at a certain x-axis point, the more of the contexts of the corresponding year belong to the specific sense. Each shade of grey represents 10% of the overall data, i.e., three shades of grey mean that between 20% and 30% of the contexts can be attributed to that sense.

This method of presenting the data focuses less on the detection of outliers and more on general trends. It can easily be seen that certain senses appear at particular points in time, e.g., the senses belonging to the lines labeled *e*, *f*, *j*, and *k* in Figure 4.4. This provides a strong indication that the outlined senses might correspond to new ways of word usage.

4.1.4 Case Studies

In order to be able to judge the effectiveness of our new approach, we chose key words that are likely candidates for a change in use in the time from 1987 to 2007. That is, we explored the contexts of target words relating to the relatively recent introduction of the internet. The advantage of these terms is that the cause of change can typically be located precisely in time.

Browsing and Surfing Figure 4.4 shows the temporal sense development of the verbs *to browse* and *to surf*, together with the topmost descriptive terms for each sense. Sense *e* for *to browse* and sense *k* for *to surf* pattern quite similarly. Inspecting their contexts reveals that both senses appear with the invention of web browsers, peaking shortly after the introduction of the groundbreaking Netscape Navigator (1994). For *to browse*, another broader sense (sense *f*) concerning browsing in both the internet and digital media collections shows

a continuous increase over time, dominating in 2007.

The first occurrences assigned to sense f in 1987 are "browse data bases", "word-by-word browsing" in databases and "browsing files in the center's library", referring to physical files, namely photographs. We speculate that the sense of browsing physical media might haven given rise to the sense which refers to browsing electronic media, which in turn becomes the dominating sense with the advent of the web.

Figure 4.3 shows pairwise comparisons of word senses with respect to the contexts they share, i.e., contexts that cannot unambiguously be assigned to one or the other. Each context is represented by one dot colored according to its time stamp. It can be seen that senses d (animals that browse) and e (browsing the web) share no contexts at all. Senses d (animals that browse) and f (browsing files) share only few outlier contexts. In turn, senses e (browsing the web) and f (browsing files) share a fair number of contexts, which is to be expected, as they are closely related. Single contexts, each represented by a colored dot, can be inspected via a mouse roll over. This allows for an in-depth look at specific data points and a better understanding how the data points relate to a sense.

Figure 4.5 shows the first and second MDS dimension of the *browse*-contexts. While Subfigure 4.5(a) shows the contexts of the whole time range, Subfigures 4.5(b), (c), and (d) show smaller selected time spans. One bias in (a) is that the contexts are plotted in temporal order and newer contexts potentially cover older ones. For this reason the interactive selection of time spans is important. In (b) it becomes obvious that older contexts are all located on the same spot, whereas in (c) they are scattered across the whole plane and in (d) they are more limited to certain parts of the plane again.

This implies that the two main sense dimensions uncovered with MDS are not present in the beginning (1987-1993), then increase between 1993 and 2003 and again lose some of their importance between 2003 and 2007. This observation is backed up by the plot of Figure 4.6, where the same phenomenon becomes visible when each of the two sense dimensions is plotted against the time dimension.

Further experiments revealed that there is no noteworthy difference between the top-dimensions when performing LSA and MDS. Additionally, the difference

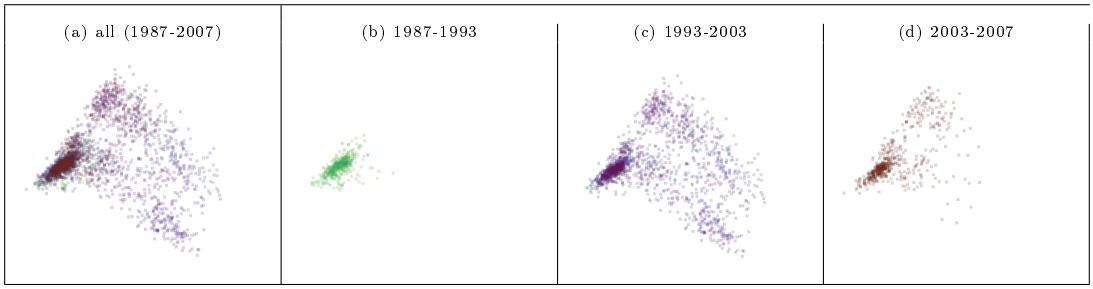


Figure 4.5: First and second MDS dimension of the "to browse"-context, different time ranges have been selected.

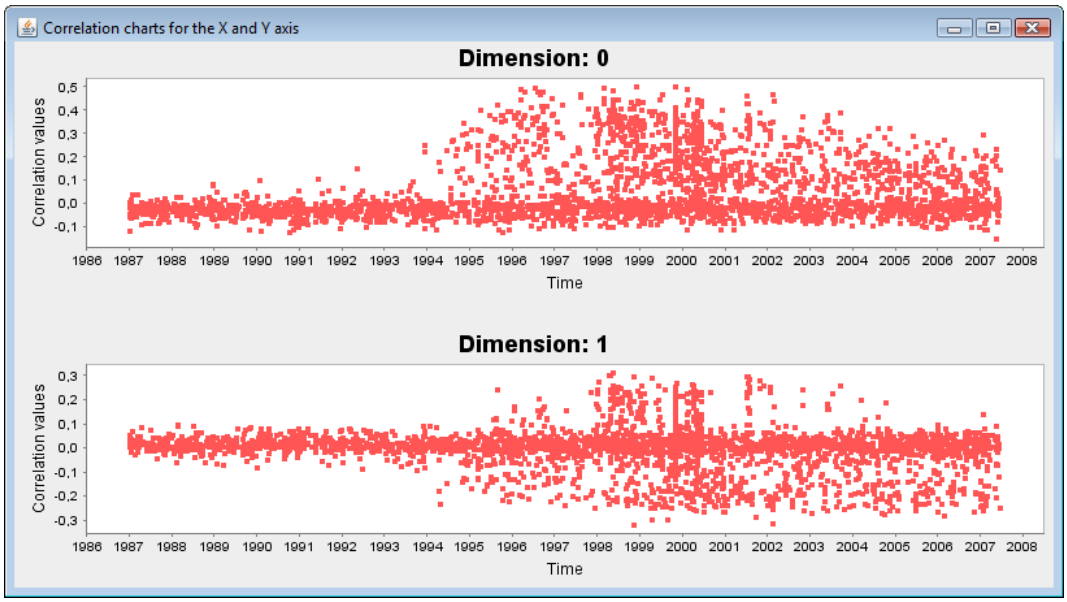


Figure 4.6: The coordinates of the first and second MDS dimension of the *browse*-contexts plotted against time.

between MDS applied on a distance matrix of pairwise Euclidean distances and applied on pairwise cosine distances was only marginal.

While the first MDS dimensions are generally able to show developments over time, it is necessary for the understanding of the developments to read a number of contexts and compare contexts at different locations in the plane. The axes as such are not interpretable. This leads to the main advantage when applying LDA: The sense labels give specific hints as to the kind of development observed in the data.

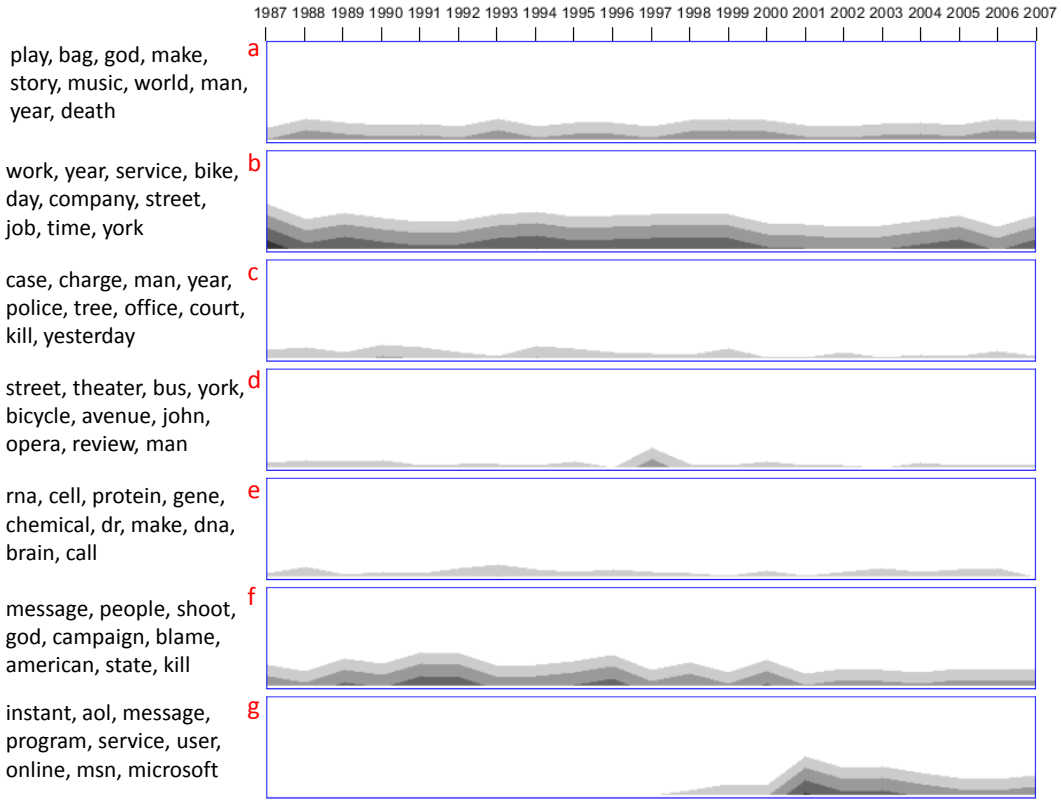


Figure 4.7: Development of different word senses for *messenger*.

Messenger Figure 4.7 shows seven senses induced with LDA from the contexts of the target word *messenger* in the NYT corpus as well as the respective sense developments over time. Not all of the senses can be interpreted easily, yet, the analysis points to some interesting findings. Most of the less clear senses refer to human messengers in general, and bike messengers in particular. This includes senses referring to *messenger bags* or *shoot-the-messenger* quotes. One clear-cut sense is that of messengers in biochemistry (sense *e*).

Another clear-cut sense has only come up in 1997 and become much stronger in 2001 (sense g): It is about online instant messaging. The appearance of this new sense coincides with the first release of the very popular AOL Instant Messenger in May 1997[7].

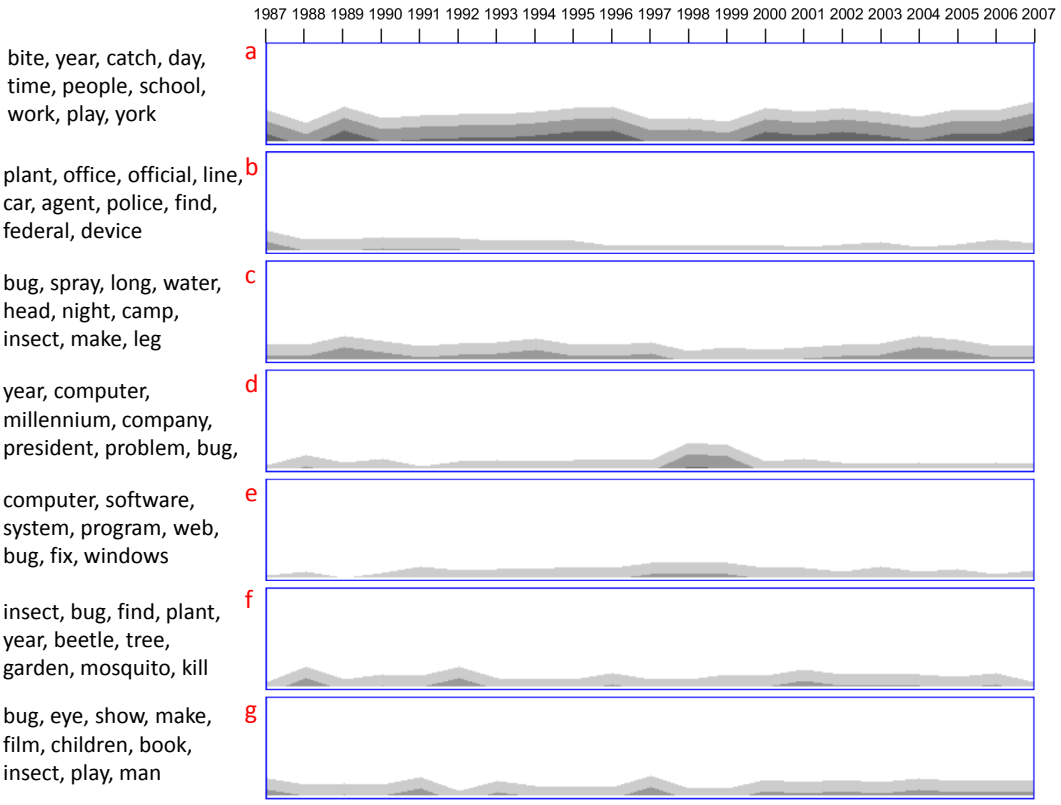


Figure 4.8: Development of different word senses for *bug*.

Bug Figure 4.8 shows seven senses induced with LDA from the contexts of the target word *bug* in the NYT corpus as well as the respective sense developments over time. Several senses can be discerned. Apart from the insect (senses c and f), a bug can be a wiretap (sense b), and refer to errors in computer software (senses d and e). While sense e is more general, sense d refers mainly to the famous millennium bug, known as *Y2K*, and consequently peaks shortly before the year 2000.

[7] http://en.wikipedia.org/wiki/AOL_Instant_Messenger last revised on February 5th, 2013

	LSA dimensions
1	web 0.40, internet 0.38, software 0.36, microsoft 0.28, windows 0.18
2	microsoft 0.24, software 0.23, windows 0.13, internet 0.13, netscape 0.12
3	microsoft 0.27, store 0.22, shop 0.20, windows 0.19, software 0.16
4	shop 0.32, netscape 0.23, web 0.23, store 0.19, software 0.19
5	book 0.48, netscape 0.26, software 0.17, world 0.13, communication 0.12
6	internet 0.58, shop 0.25, service 0.16, computer 0.13, people 0.11
7	make 0.39, shop 0.34, site 0.16, windows 0.13, art 0.08
...	...
15	find 0.30, people 0.22, year 0.19, **deer** 0.16, day 0.15

Table 4.1: Descriptive terms for the top LSA dimensions for the contexts of *to browse*. For each dimension the top 5 positively associated terms were extracted, together with their value in the corresponding dimension.

4.1.5 Evaluation: LDA vs. LSA

The case studies give empirical evidence for the good performance of the suggested approach in the investigation of diachronic word sense developments. Apart from that a further evaluation has been provided by my collaborators as part of the original publication, where the information that can be gained from the visualization is compared to knowledge from different lexical resources. The interested reader is referred to the original work [145].

In addition, we compared the quality of the computed LSA dimensions with the LDA senses, and show it for one example case. Table 4.1 shows the LSA dimensions learned from the contexts of the verb *to browse*. The top five associated terms for each dimension have been extracted as descriptors. The dimensions are heavily dominated by senses strongly represented in the corpus (e.g., browsing the web). Infrequent senses (e.g., animals that browse) only occur in very low-ranked dimensions and are mixed with other senses (see the bold term *deer* in dimension 15). A meaningful clear-cut distinction between different senses, as with LDA, is not achieved with LSA. This observation was confirmed in further test cases.

4.1.6 Discussion and Conclusions

The discussion of the suggested approach from a domain experts perspective:

> "*When dealing with a complex phenomenon such as semantic change,*

one has to be aware of the limitations of an automatic approach in order to be able to draw the right conclusions from its results. The first results of the case studies presented in this paper [the current section; note from the author] show that LDA *is useful for distinguishing different word senses on the basis of word contexts and performs better than* LSA *for this task. Further, it has been demonstrated by exemplary cases that the emergence of a new word sense can be detected by our new methodology.*

One of the main reasons for an interactive visualization approach is the possibility of being able to detect conspicuous patterns at-a-glance, yet at the same time being able to delve into the details of the data by zooming in on the occurrences of particular words in their contexts. This makes it possible to compensate for one of the major disadvantages of generative and vector space models, namely their functioning as "black boxes" whose results cannot be tracked easily.

The biggest problem in dealing with a corpus-based method of detecting meaning change is the availability of suitable corpora. First, computing semantic information on the basis of contexts requires a large amount of data in order to be able to infer reliable results. Second, the words in the context from which the meanings will be distinguished should be both semantically and orthographically stable over time so that comparisons between different stages in the development of the language can be made. Unfortunately, both requirements are not always met. On the one hand words do change their meaning, after all this is what the present study is all about. However, we assume that the meanings in a certain context window are stable enough to infer reliable results provided it is possible that the forms of the same words in different periods can be linked. This of course limits the applicability of the approach to smaller time ranges due to changes in the phonetic form of words. Moreover, in particular for older periods of the language, different variants for the same word, either due to sound changes or different (or rather no) spelling conventions, abound. For now, we circumvent this problem by testing our tool on corpora where the drawbacks of historical texts are less severe but at the same time interesting

developments can be detected to prove our approach correct." [145][8]

The intelligent integration of existing methods for the analysis of word senses over time is innovative. Whereas the previous approaches were only able to compute a set of numbers indicating general changes in word meaning, like broadening or pejoration, this is a first attempt to provide a clearer notion of what word senses can be found at different points in time. The case studies demonstrate that especially newly emerging word senses can readily be detected. While in some cases the clear point in time can be identified, when a new word sense has started appearing and spreading, it is not to be expected that there are clear points in time for the disappearance of a word sense. However, the visualization shows tendencies of decreases in usage, which might indicate a long-term extinction of a word sense. In addition, further effects become visible that would probably not have been detected without the use of visualization. For example, the temporary diffusion of the word sense "software error" of the target word *bug* shortly before the year 2000.

In addition to the exploration over time, the item-based views allow to visually depict whether two senses are closely related or not. This novel feature of our approach is especially interesting to use, when a new word sense appears. It may help to hypothesize whether the new sense originates from one of the old senses.

Instead of using LDA, it is also possible to use Hierarchical Dirichlet Processes (HDP) [165]. The advantage is that HDP infers the number of topics to be generated from the data and thus does not depend on this parameter. Yet, it depends on other user-given parameters. A recent study finds "that the two models achieve similar levels of induction quality" [185]. Thus, in future work experiments with HDP should be conducted.

4.2 Analysis of the Appearance of new Suffixes

This section builds on the following publication:

Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. Lexical Semantics and Distribution of Suffixes - A Visual

[8] Written by my collaborators from linguistics as part of our joint publication.

Analysis. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, 2012. [149][9]

This Section deals with relatively new derivational morphemes that have come up in the last century. The investigation is done based on three suffixes originating in English: *-gate*, *-geddon* and *-athon*. In particular, we develop visual analytics processes that support domain experts in their research targeting two main tasks. First, to help specify the semantics of these suffixes and discover potential developments of their semantics over time. Second, to characterize the dynamics of the spread of new coinages, containing the suffixes, within and across languages.

First, we give some background information on this research topic (Section 4.2.1), describe the data and resources used (Section 4.2.2), and outline the analysis goals (Section 4.2.3). Next, Section 4.2.4 provides methods to investigate the semantics of the suffix *-gate* and potential developments over time. Section 4.2.5 introduces visual mappings that help to reveal cross-linguistic spread and productivity of the suffixes under investigation. Finally, Section 4.2.6 provides a brief discussion of the presented research and a conclusion.

4.2.1 Background

The research background from linguistics was described by the domain experts I collaborated with as follows:

> "*It is well-known that parts of a compound can begin to lead an additional life as derivational suffixes, or even as stand-alone items. A famous example is* burger*, which is now used to denote a food-item (e.g.,* burger, cheese burger, veggie burger*) and is originally from the word* Hamburger*, which designates a person from the German city of*

[9]The computer science research part of the paper was conducted by Andreas Niekler and myself. Andreas Niekler did the optimized topic modeling and also wrote the corresponding section. All the other parts with computer science focus were written by myself. Annette Hautli and Miriam Butt wrote the linguistic research part of the paper. Daniel Keim gave advice. Volker Rehberg and myself did the programming, that also partly built on implementations from Zdravko Monov. As also acknowledged in the publication Thomas Mayer helped with comments on previous versions of the work. For all parts of the publication that were not written by myself I reference the original work.

Hamburg. These morphemes are generally known as cranberry morphemes *(because of the prolific use of* cran*). Some other examples are* -(o)nomics, -(o)mat *or* (o)rama. *While it is well-known that this morphological process exists, it is less clear what conditions trigger it and how the coinage "catches" on to become a regular part of a language." [149]*[10]

To the best of our knowledge there is no other work that aims to computationally model and assess the development of new derivational suffixes both with respect to diachronic and cross-linguistic implications. The only other work that comes close to our approach is a study about the spread of the German suffix *-itis* into non-medical contexts [110]. However, the focus there is mainly on productivity.

4.2.2 Data and Resources

Data & Statistics

Our investigations are based on two different data sets, one is a diachronic news corpus, the New York Times Annotated Corpus[11] containing 1.8 million newspaper articles from 1987 to 2007. To generate the second data set, we performed an online scan of the EMM news service,[12] which links to multilingual news articles from all over the world and enriches them with metadata [11,96]. Between May 2009 and January 2012, we scanned about eleven million news articles in English, German, and French.

For both data sources, we extract a context of 25 words before and after the words under investigation, together with the timestamps. In the case of the EMM data, we also save information on the news source, the source country and the language of the article. In a manual postprocessing step, we clean the dataset from words ending in the suffixes by coincidence, many of which are proper names of persons and locations.

From the EMM metadata, we can attribute the employment of the suffixes to the countries they were used in. Table 4.2 shows the figures for the *-gate*

[10] Written by my collaborators from linguistics research as part of our joint publication
[11] http://www.ldc.upenn.edu/ last revised on March 11th, 2013
[12] http://emm.newsexplorer.eu/ last revised on March 11th, 2013

suffix, what language it was used in, and its country of origin. We can see that the suffix was used in many countries and different world regions between May 2009 and January 2012.

Lang.	**Country**
English	GB (1142), USA (840), Ireland (364), Pakistan (275), South Africa (190), India (131), Australia (129), Canada (117), Zimbabwe (73)
French	France (2089), Switzerland (429), Belgium (108), Senegal (30)
German	Germany (493), Switzerland (151), Austria (151)

Table 4.2: Usage of the suffix *-gate* in different languages/countries. For each language only the countries with the most occurrences are listed.

Among the total 7,500 *-gate* appearances, *Rubygate* – the affair of Italian's ex prime minister Silvio Berlusconi with an under-aged girl from Morocco – was the most frequent word with 1,558 matches, followed by *Angolagate* with 1,025 matches and *Climategate* with 752 matches. All in all, app. 700 supposedly new coinages could be found in the EMM data, most of which have appeared only once. A list of all these coinages is provided in the appendix of this thesis. The NYT corpus has 1,000 matches of *-gate* words, the top ones were *Iraqgate* with 148, *Travelgate* with 122, and *Irangate* with 105 matches. The frequency of *-geddon* and *-athon* was much lower.

4.2.3 Analysis Tasks and Goals

Our subject of investigation are the three productive suffixes *-gate*, *geddon*, and *-athon*.

> "*What these suffixes have in common is that they trigger neologisms in various languages and all of them seem to carry some lexical semantic information. Whereas -gate, which was coined by the Watergate affair, is used for scandalous events or affairs, -geddon seems to denote a similar concept but more of a disastrous event, building on its original use in the Bible. Usually, -athon, coming from marathon, denotes a long-lasting event.*" [149][13]

[13] Jointly written part of our joint publication.

We assume that the lexical semantic content of these suffixes can be modeled with standard topic models.

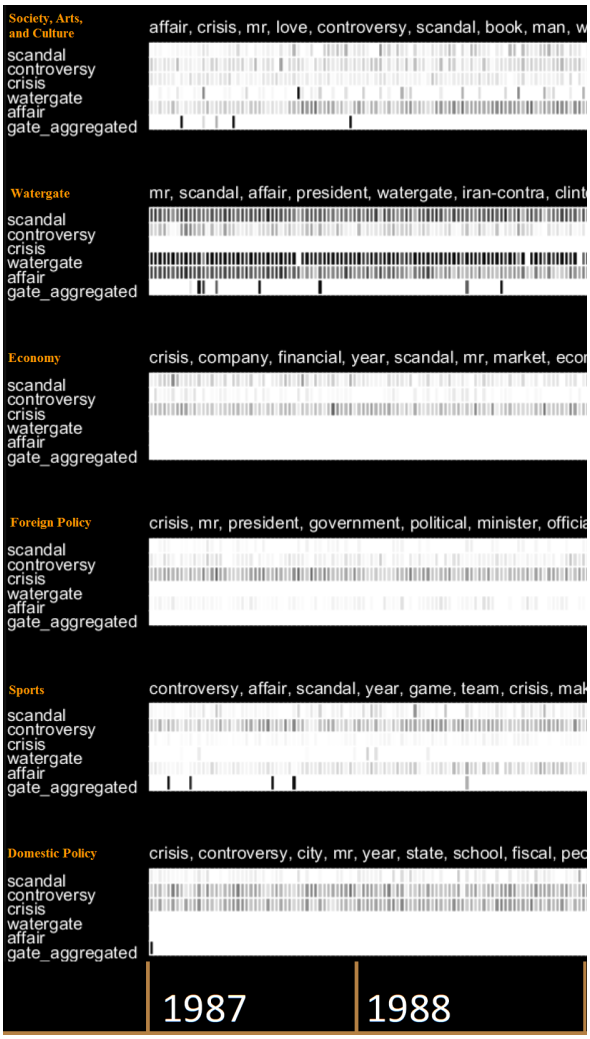
Topic Modeling

The purpose of using topic modeling in this approach is to discover meaning relationships between the suffixes and semantically related words, i.e. we want to determine from the word contexts whether *-gate* words share context features with words such as *scandal* or *affair*. As previously (see Section 4.1), again we apply LDA topic modeling to learn a certain number of topics from the contexts of the *-gate* coinages. Topics in this case can be interpreted as word meanings or usage contexts. For more details about the concrete implementation of the optimized topic model that we applied, please see our original publication [149]. As input to the topic modeling we extracted from the NYT corpus all appearances of new coinages ending on the suffix *-gate* together with their contexts, i.e. 25 words before and after the appearance. The same was done for the words *Watergate*, *scandal*, *affair*, *crisis*, and *controversy*. The intuition was that these target words were considered to be likely to share some meaning components with the new coinages. The open questions were, whether the *-gate* suffix had managed to gather some exclusive meaning component over the years and whether it was otherwise more closely attached to the meaning of one of this alternative target words. We decided to request 6 topics from the topic modeling. In an extreme case each of the topics would describe the meaning of only one of the target words. However, it could also be expected that different target words could have a considerable overlap in meaning. The most interesting to us, would be the overlaps in meaning with the *-gate* suffix.

4.2.4 Diachronic Analysis of Word Sense Developments

The topics extracted from the target word contexts based on the NYT corpus were further investigated with respect to the correlation between the lexical semantic content of the suffixed words and a development over time. For this purpose we designed a pixel visualization (see Figure 4.9), mapping the data facets to the visual variables as follows:

Figure 4.9: The diachronic distribution of the words under investigation over the 6 topics learned from the New York Times Corpus. The high resolution screenshot is zoomable in the pdf version. Adapted from [149], © 2012 Association for Computational Linguistics.



The data is divided according to the topics, mapping each topic to one horizontal band. The descriptive words of a topic as found by LDA are listed above its band. In addition, each topic is manually assigned an interpretive label. In Figure 4.9 these labels are at the far left of a topic band and here they are printed in orange.

Each topic band is further subdivided according to the words under investigation. Under the label "gate-aggregated", all words with *-gate* suffixes (except *Watergate*) are summarized. The bands are aligned with a time axis and vertically divided into cells, each cell representing one week of data. The cell color indicates whether the corresponding word under investigation occurred within the corresponding topic in the corresponding week. The white color means that there was no such occurrence, whereas the darkest black is assigned to the cell of the week where most occurrences (*max*) of a word under investigation are found, independent from the topic. Other occurrence counts are colored in grey tones according to a linear mapping into the normalized color range from white=0 to black=*max*. Note that the normalization depends on the word under investigation, i.e. is relative to its maximal occurrence.

In Figure 4.9, the data had to be split into two chunks to fit the page. The upper part shows the years from 1987 to 1997 and the lower part from 1997 to 2007. There are several possibilities for user interaction: A semantic zoom allows the data to be displayed in different levels of time granularity, e.g. day,

week, month, year. By mousing over a cell, the underlying text passages are displayed in a tooltip.

Findings Figure 4.9 shows that the topics are dominated by different words under investigation, i.e. the words under investigation cannot be clearly separated into self-contained meanings. This mixture indicates that the words under investigation have similar meanings, but that in different contexts they are used in different combinations. The descriptive terms of a topic are plotted right above its band in Figure 4.9. Following this descriptive terms we manually assigned a label to each topic:

1. Society, Arts, and Culture: This seems to be the most general topic with the broadest usage of the words under investigation. The descriptive terms show that it is a lot about interpersonal relations and dominated by *affair*. In 1989/1990 the play *Mastergate* becomes visible in the *gate-aggregated* band.

2. Economy: This topic is strongly related to *crisis* and apart from the moderate frequency of *scandal*, other words are rarely used in this context. Apparently, financial scandals were usually not described attaching the suffix *-gate* in the years between 1987 and 2007.

3. Foreign Policy: This is another topic dominated by *crisis*, with moderate occurrences of *controversy*. Some *-gate* coinages also appear.

4. Sports: Here, *controversy* is the dominating element, with a raised frequency of *affair* and small frequency of *scandal*. Again, *-gate* coinages appear from time to time, with a slightly increased frequency towards the end.

5. Domestic Policy: The dominant words are *controversy* and *crisis*. It's noteworthy that *controversy* is a lot more frequent here than for Foreign Policy. Especially in the last years *-gate* coinages appeared from time to time.

In sum, we find that there are preferred contexts in which *-gate* is used. A lot of new *-gate* coinages seem to either reference Watergate explicitly or appear in very similar contexts. Apart from that they mainly appear in topics to do with society, arts, and culture. On the other side, in topics that have to do with the economy, *-gate* is hardly used. In the researched corpus, the lexical semantic content of *-gate* seems to be most closely linked to the word

affair, while there are no clear trends over time indicating a meaning change of the *-gate* suffix. All of these conclusions, however, have to be taken with a pinch of salt, because the overall frequency of the *-gate* coinages is relatively low and observations may thus be biased by random effects.

Analyzing the Semantic Field

In order to find out more about the meaning of the suffix *-gate* we conducted a closer investigation of the distribution of other words across the contexts of different very recent *-gate* coinages. To this end, we extracted all coinages in English that appeared at least 10 times in the EMM corpus and investigated the words in the context. For each of those context words we extracted two feature values with the aim to address the following questions:

- How discriminating is it for a single coinage? For each of the context words we calculated the relative occurrence frequency with each of the -gate coinages. Next, we subtracted the average relative frequency over all coinages from the maximal relative frequency.

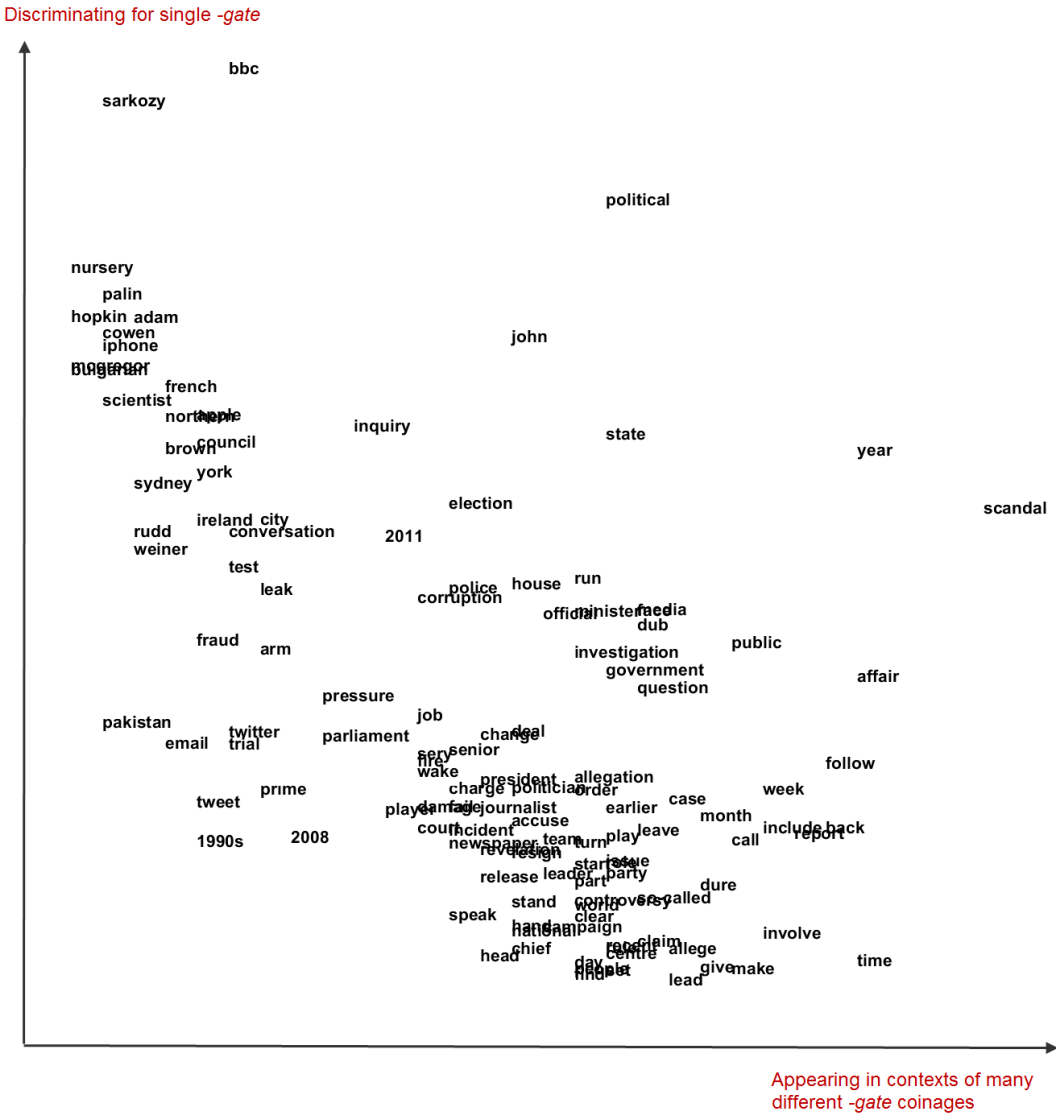
- How descriptive is it for the suffix *-gate* in general? For each of the context words we counted with how many different coinages it co-occurred at least once.

We mapped the context words on a 2D plane where the first feature was mapped to the y-axis and the second feature to the x-axis. Figure 4.10 shows the original positions of all words and Figure 4.11 shows a version where some words were slightly shifted to remove overlap.

On the right side of both figures words appear that occur in the contexts of many different coinages. These words either specify the meaning of the suffix *-gate* more closely, e.g. *scandal*, *affair*, *investigation*, *to dub*, *to call* or are time specifications (*year*, *week*, *month*, *time*).

On the upper left of both figures, context words can be found that are very specific for single coinages, mostly proper nouns like *sarkozy*, *bbc*, *palin*, etc. These words relate to the topics and subjects involved in scandals.

In between the two extremes, in the middle of the display, we can find terms that describe more general topics relating to the coinages. These context words



Appearing in contexts of many different *-gate* coinages

Figure 4.10: Words occurring in the contexts of *-gate* coinages *without* overlap removal.

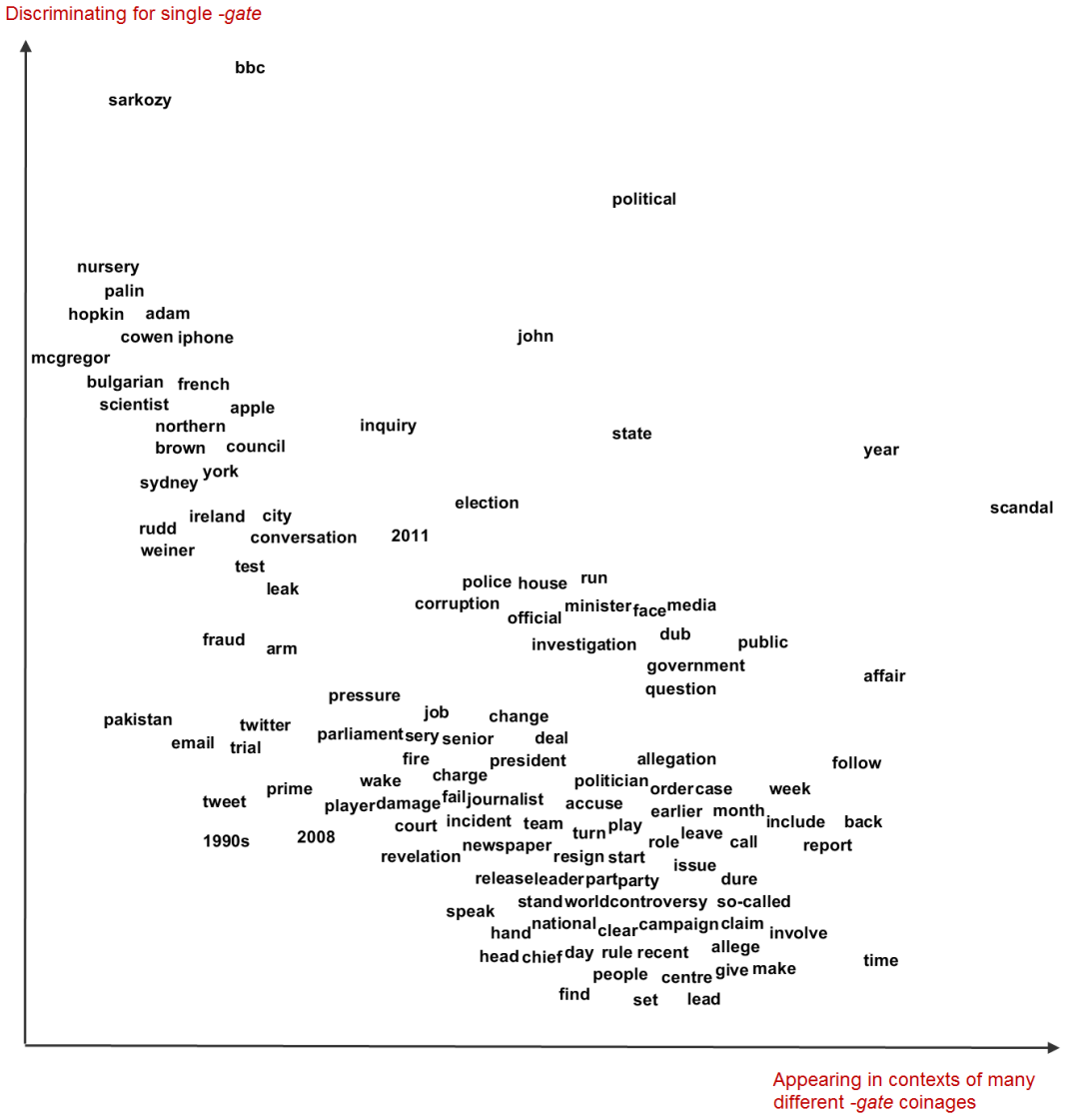


Figure 4.11: Words occurring in the contexts of -*gate* coinages *with* overlap removal.

also partly define the sense of the suffix. Many of them relate to politics, e.g. *political*, *state*, *election*, *minister*, *government*. This indicates that the most widely spread *-gate* coinages often refer to political scandals and affairs.

4.2.5 Diachronic Analysis of Cross-Linguistic Spread and Productivity

The morphological productivity of a suffix is interesting to explore for linguists. We investigate this productivity using the EMM data in English, German, and French:

> "*The cases of suffixation presented above should also be considered from the standpoint of morphological productivity. For Baayen [12], morphological productivity is a complex phenomenon in which factors like the structure of the language, its processing complexities and social conventions mingle. Whereas he focuses on the the correlation between productivity and frequency, we can take into account another variable for productivity. In particular, we can consider the number of newspapers that use a certain term. This will normalize the measures usually taken in that a term like Watergate, which is highly frequent and mentioned in a variety of sources is more productive than a term that occurs frequently, but only in one source. Using this methodology we can at least partly circumvent the problem of productivity effects that are merely based on the specific style of one particular newspaper.*" [149][14]

First, we visually evaluate the productivity of the different suffixes plotting the sum of different coinages against time, see Figure 4.12. As can be expected, in all three cases there is a steeper slope in the beginning of the monitored period. This is an artifact because all older coinages, that had been around before the monitoring started, will be observed for the first time. As more time passes all plots show a linear overall trend, indicating that the rate with which new coinages appear remains somewhat constant. Yet, there are some local oscillations in the rate that become more visible in the plots of *-geddon* and *-athon* coinages, which are in general much more infrequent than *-gate* coinages.

[14] Written by my co-authors as part of our joint publication.

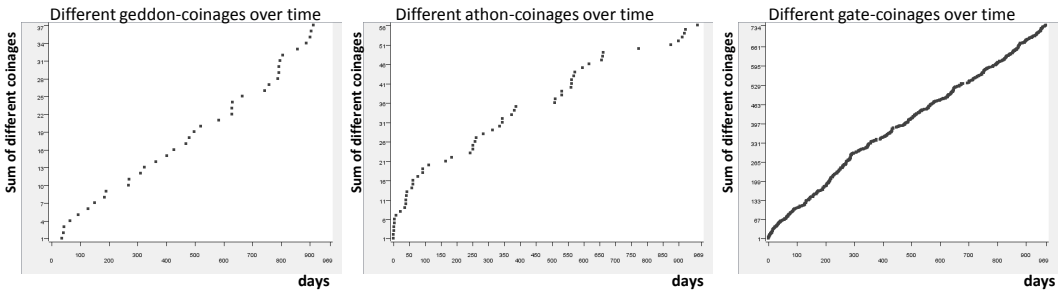


Figure 4.12: The number of different coinages containing the suffixes under investigation (on the y-axis) plotted against the number of days passed during the monitoring process (on the x-axis). Reprinted from [149], © 2012 Association for Computational Linguistics.

It can be concluded that over the last two-and-a-half years the suffixes kept their rate of productivity in English, German, and French newswire texts fairly constant.

To get a first insight into the cross-linguistic productivity of the new coinages we customized a visualization with the Tableau software.[15] Figure 4.13 shows the appearances of the 15 most frequent *-gate* coinages across the three languages over time. Along the y-axis the data is divided according to *-gate* coinages and languages, whereas the x-axis encodes the time. Whenever a certain coinage appears in a certain language at a certain point in time, a colored triangle is plotted to the corresponding position. The color redundantly encodes the language for easier interpretation.

The simple overview visualization shown in Figure 4.13 already shows many interesting patterns. The most salient patterns can be summarized as:

1. No language barrier: The top *-gate* coinages belong to scandals that are of international interest and once they are coined in English they immediately spread to the other languages, see *Rubygate*, *Climategate*, *Cablegate*, *Antennagate*, and *Crashgate*. Only in the case of *Angolagate* and *Karachigate* there is a certain delay in the spread, possibly due to the fact that it was coined in French first and initially did not achieve the same attention as coinages in English.

2. Pertinacity partly depends on language: Some *-gate* coinages reappear over and over again only in individual languages. This especially holds for words that were coined before the monitoring started, e.g. *Sachsgate*, *Oil-*

[15] http://www.tableausoftware.com/ last revised March 12th, 2013

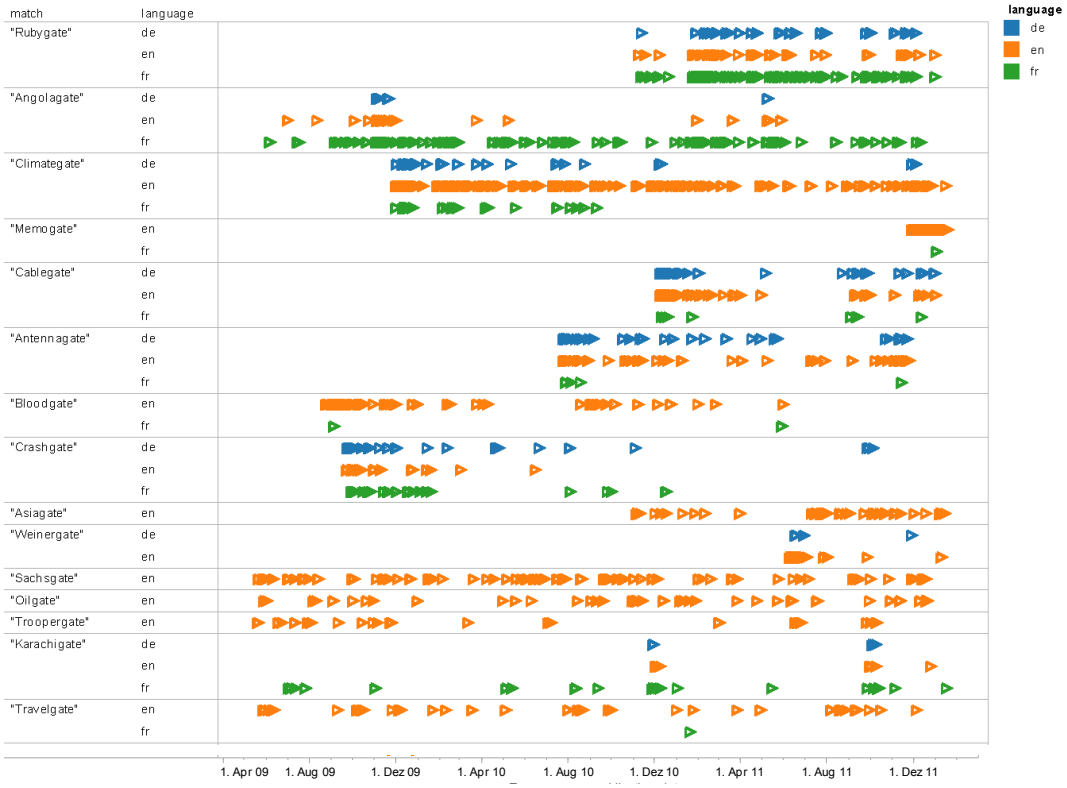


Figure 4.13: The appearances of the 15 most frequent *-gate* coinages over time and across the different languages. Adapted from [149], © 2012 Association for Computational Linguistics.

English	Great Britain ●, USA ●, Ireland ●, Pakistan ●, India ●, Australia ●, Canada ●
French	France ■, Switzerland ■, Belgium ■
German	Germany ▲, Switzerland ▲, Austria ▲

Figure 4.14: This is the legend for all of the following Figures in this Section. It shows the icons for the most frequent language-country combinations. The shape indicates the language and the color the country. Further shapes and colors represent languages and countries that appear only with a very low frequency.

gate, *Troopergate*, and *Travelgate* which all persist in English. Examples can be found for other languages, e.g. *Angolagate* for French. Interestingly, in German *Nipplegate* persists over the whole monitored period, but only in German, and even outperforms its German spelling equivalent *Nippelgate*.

3. Some coinages are special: Some of the recent coinages such as *Memogate*, *Asiagate*, and *Weinergate* reach an extremely high frequency within very short time ranges, but can be found almost exclusively in English. The exact spread over different countries and news sources will be subject of further investigation later in this section. It has to be noted that many of the infrequent coinages appear only once and are never adopted.

As mentioned before, a complete list of the app. 700 supposedly new coinages with *-gate* suffix as extracted from app. 11 million online news articles in English, French, and German can be found in the appendix.

Spread across News Sources and Countries

Figure 4.13 clearly shows that *Memogate*[16] is heavily mentioned in English speaking news sources within a short time range. We developed a further visualization that shows how these mentions diachronically distribute over different news sources and countries. In Figure 4.15 each article mentioning *Memogate* is represented by a colored icon. The y-axis position encodes the news source, the x-axis position encodes the temporal dimension. The shape of an icon indicates the language of the article and the icon color indicates the country, see Figure 4.14 for more details about the visual mappings. In

[16] http://en.wikipedia.org/wiki/Memogate_%28Pakistan%29, last revised on December 6th, 2012

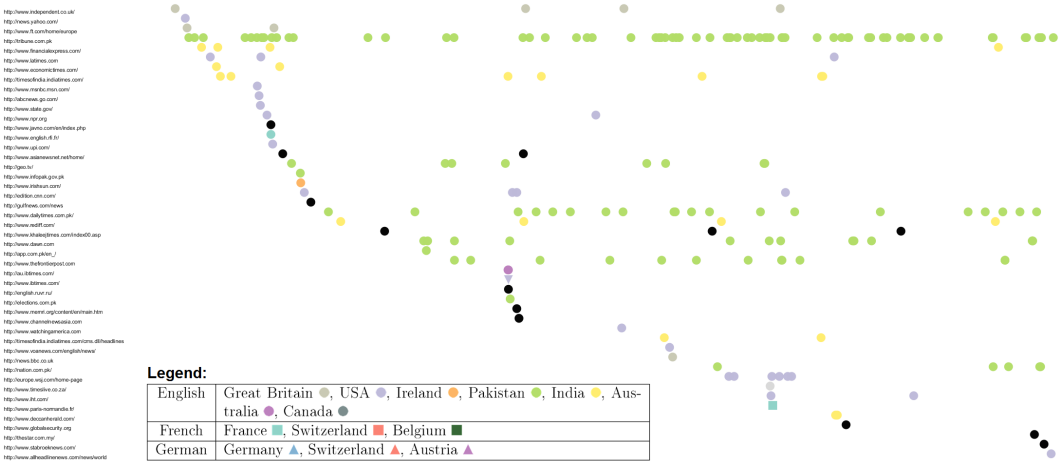


Figure 4.15: Detailed analysis of the *Memogate* based on same data as in Figure 4.13 using alternative visual mappings: Diachronic spread over different countries and news sources. Sources from Pakistan (green) and India (yellow) dominate.

the *Memogate* example circles (English) heavily dominate. The dominating countries of origin of the news sources, in this case are Pakistan (green), India (yellow), and the USA (purple).

Findings: While the first three mentions of *Memogate* could be found in British and American newspapers, early on it was adopted by one Pakistani newspaper[17] (fourth line from the top) and used so heavily that it kept being adopted and became constantly used by further sources from Pakistan and also India. Apparently, individual sources may have a huge influence on the spread of a new coinage.

When visualizing further of the top frequent *-gate* coinages interesting patterns emerge and according to their visual appearances, certain coinages can be grouped as behaving similar:

- **Rubygate**[18] is a scandal about an alleged relationship between Italy's former prime minister Silvio Berlusconi and an under-aged girl nicknamed "Ruby", see Figure 4.16. There is a gap between the first appearances and the next wave of mentions which makes the coinage spread more widely. While news sources from Germany catch up later all of

[17] http://tribune.com.pk/ last revised on December 17th, 2012

[18] http://en.wikipedia.org/wiki/Silvio_Berlusconi_underage_prostitution_charges, last revised on December 6th, 2012

a sudden, but do not repeat the coinage a lot, French-speaking news mention it over and over again during a longer period of time.

- **Angolagate**[19] is a scandal about corruption and illegal arms-sales to Angola involving several prominent French politicians, see Figure 4.17. The coinage seems to have appeared first in French news and it took a while until it suddenly started spreading internationally. A follow-up event concerning the trials once more triggers news mentioning it, this time also further news sources pick it up. In between, several mostly French news sources apparently keep the coinage alive repeating it continuously.

- **Climategate**[20] is a revelation suggesting that scientists manipulate data and surpress critics, see Figure 4.18. This is a good example for the quick international spread of a coinage, given that the subject concerns many countries. The first three mentions are made in three different languages, which confirms the hypothesis that there is no language barrier.

- **Cablegate**[21] is a scandal concerning the leak of United States diplomatic cables, see Figure 4.19: The visualization for this coinage reveals two interesting aspects. First, the heavy involvement of German-speaking news sources in the initial dispersion of the coinage is noteworthy. Secondly, it becomes evident that there are two "waves" of dispersion, whereas in the second wave mostly other, new sources make use of the coinage.

- **Crashgate**[22] and **Antennagate**, an issue with the Iphone 4 mobile reception, see Figure 4.20, are two representatives for scandals that appear in a burst, immediately spread internationally, and vanish quickly. For *Antennagate* there is a very small second burst at the end, which is not as big though as for *Cablegate*.

[19] http://fr.wikipedia.org/wiki/Affaire_des_ventes_d%27armes_%C3%A0_l%27Angola, last revised on December 6th, 2012

[20] http://en.wikipedia.org/wiki/Climatic_Research_Unit_email_controversy, last revised on December 6th, 2012

[21] http://en.wikipedia.org/wiki/United_States_diplomatic_cables_leak, last revised on December 6th, 2012

[22] http://en.wikipedia.org/wiki/Renault_Formula_One_crash_controversy, last revised on December 6th, 2012

- **Bloodgate**[23] and **Sachsgate**,[24] see Figure 4.21, are two representatives for coinages that have quite a number of occurrences, but are limited to few sources and countries. In both cases almost exclusively news sources from Great Britain and Ireland use the coinage. Here, *Bloodgate* once more is an example having a first and second wave of bursts. *Bloodgate* was a scandal relating to the usage of fake blood capsules in a Rugby game between an English and an Irish team, which explains the local interest.

- **Troopergate**[25] and **Karachigate**,[26] see Figure 4.22, are two representatives for coinages that keep on being around having small bursts, involving new news sources, over and over again. *Troopergate* actually dates back to the US Presidential election campaign 2008, i.e. several months before the monitoring had started. The scandal will appear again in the next chapter in the context of a completely independent study. Interestingly, the online encyclopedia Wikipedia lists two further different scandals with the same name.[27]

4.2.6 Discussion and Conclusion

We have presented initial experiments with respect to the application of topic modeling and visualization to gain a better understanding of language change involving the appearance of new coinages and their lexical semantics. We investigated three relatively new productive suffixes, namely *-gate*, *-geddon*, and *-athon* based on their occurrences in newswire data. Even though our initial data set was huge, the occurrences of the suffixes are comparatively rare and so we only had enough data for the suffix *-gate* to investigate the contexts it occurs in with an optimized topic modeling. The results indicate that it is used in broader contexts than *affair* or *scandal*, with which it is

[23] http://en.wikipedia.org/wiki/Bloodgate, last revised on December 6th, 2012

[24] http://en.wikipedia.org/wiki/Russell_Brand_Show_prank_telephone_calls_row, last revised on December 6th, 2012

[25] http://en.wikipedia.org/wiki/Alaska_Public_Safety_Commissioner_dismissal, last revised on December 6th, 2012.

[26] http://en.wikipedia.org/wiki/2002_Karachi_bus_bombing, last revised on December 6th, 2012

[27] http://en.wikipedia.org/wiki/Troopergate, last revised on December 6th, 2012

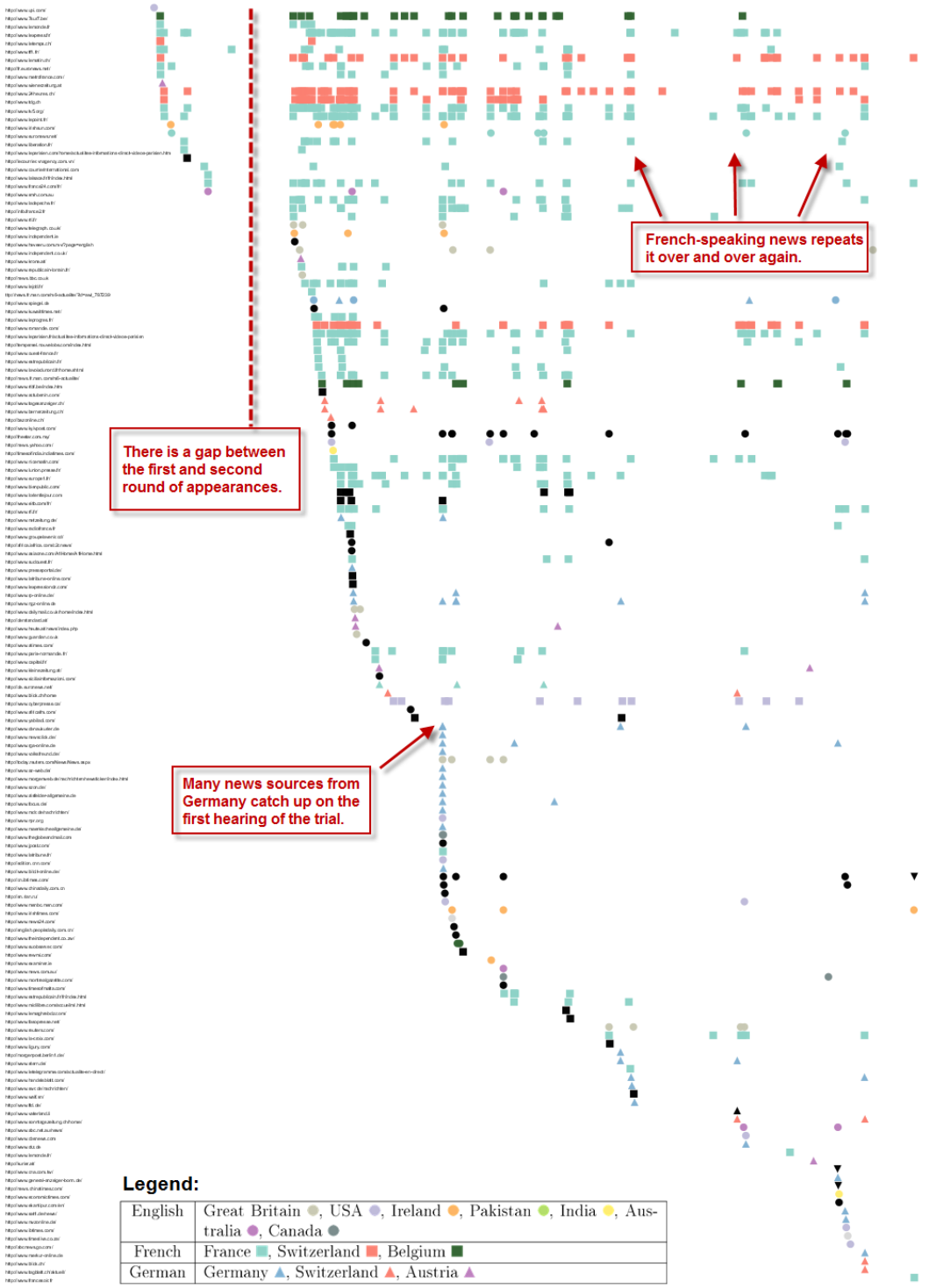


Figure 4.16: Detailed analysis of the diachronic spread of *Rubygate* across sources, countries, and languages.

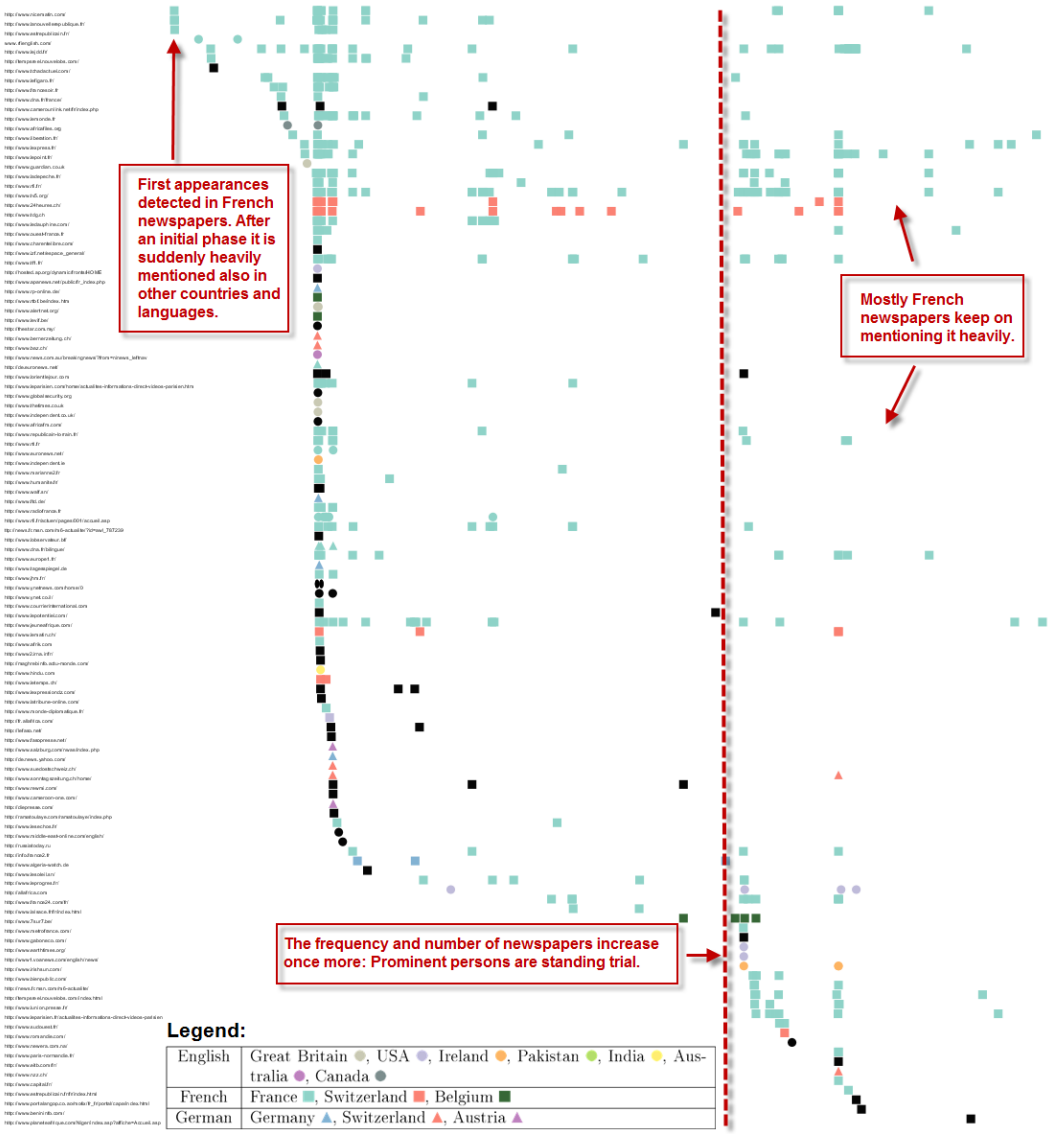


Figure 4.17: Detailed analysis of the diachronic spread of *Angolagate* across sources, countries, and languages.

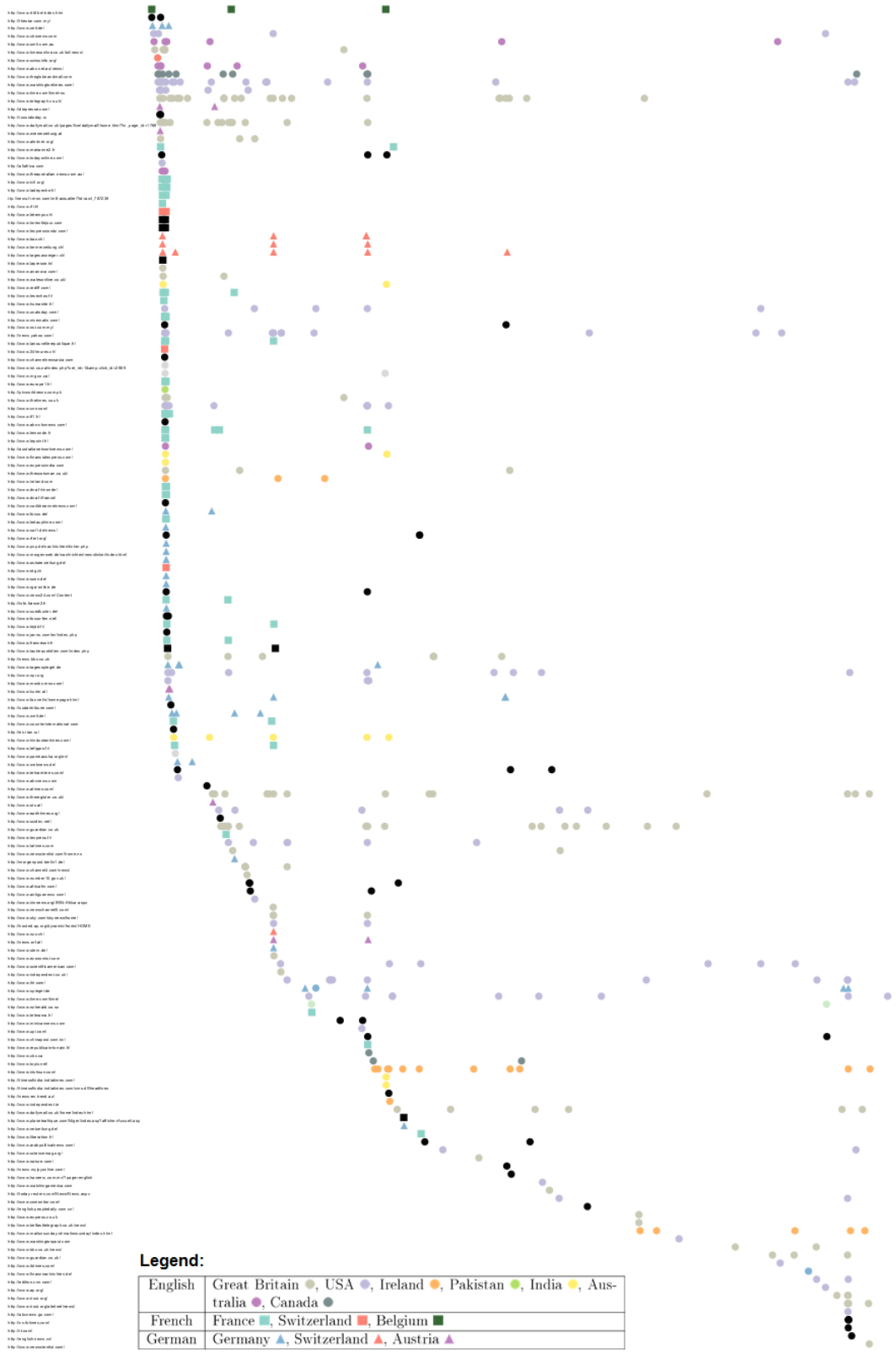


Figure 4.18: Detailed analysis of the diachronic spread of *Climategate* across sources, countries, and languages.

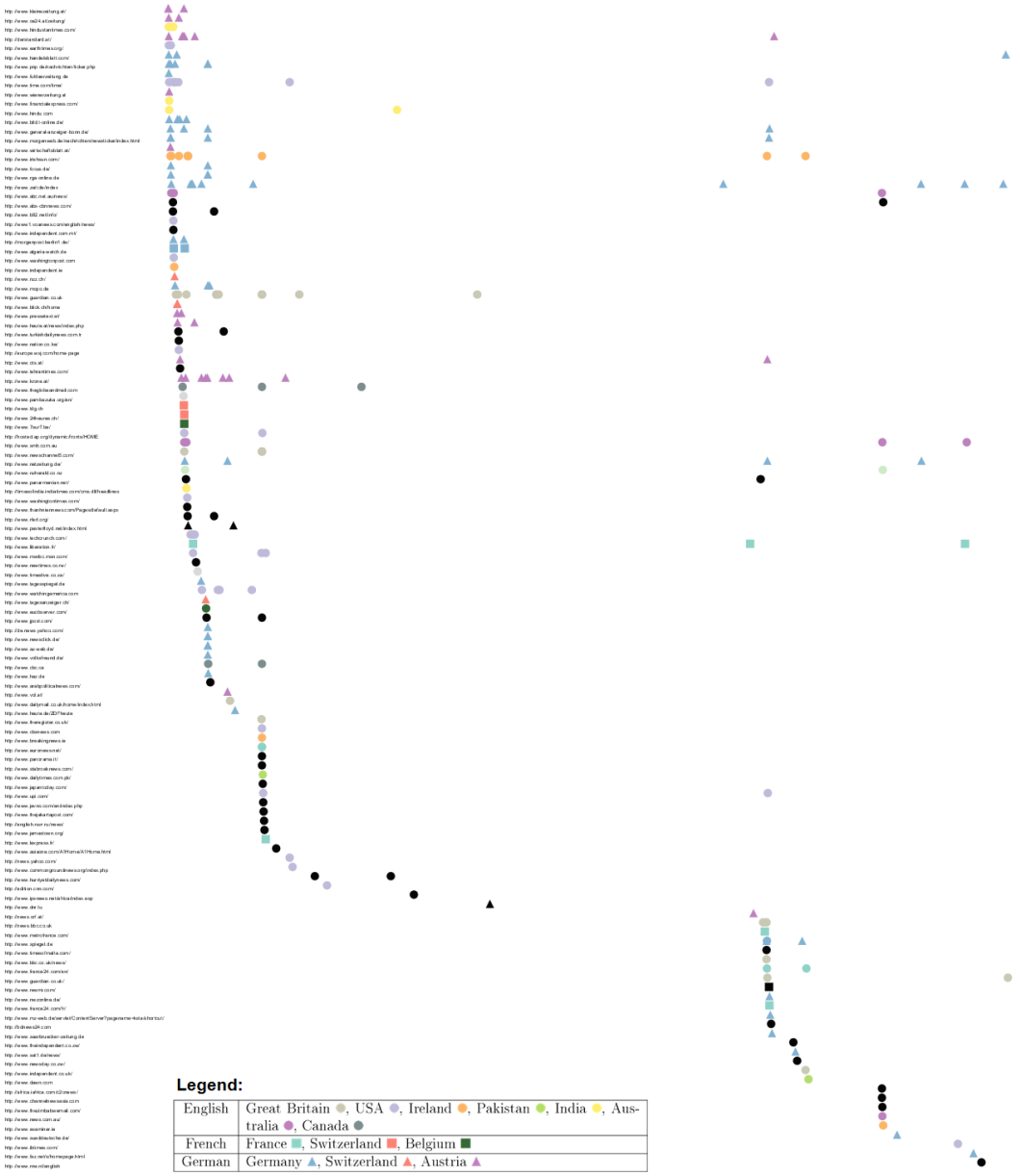


Figure 4.19: Detailed analysis of the diachronic spread of *Cablegate* across sources, countries, and languages.

Figure 4.20: Detailed analysis of the diachronic spread of *Crashgate* (top) and *Antennagate* (bottom) across sources, countries, and languages.

Figure 4.21: Detailed analysis of the diachronic spread of *Bloodgate* (top) and *Sachsgate* (bottom) across sources, countries, and languages.

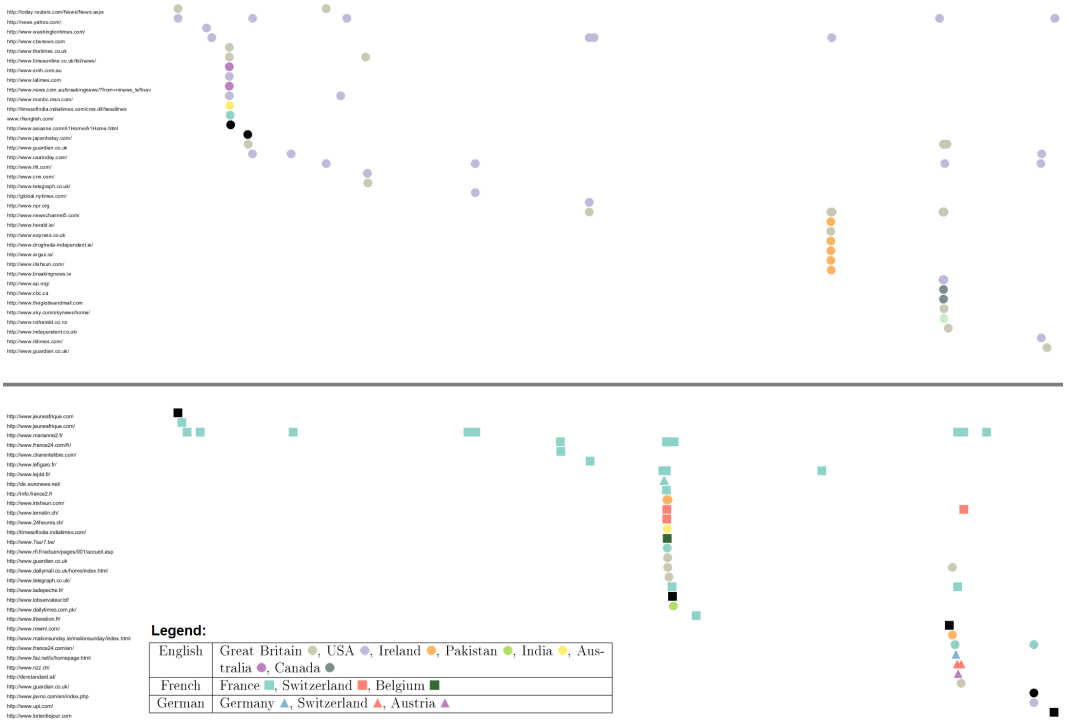


Figure 4.22: Detailed analysis of the diachronic spread of *Troopergate* (top) and *Karachigate* (bottom) across sources, countries, and languages.

most related, and that the most widely spread coinages still often refer to the domain of politics. Different domains of usage could be distinguished, even though a clear development over time could not be detected based on the NYT corpus. Investigating the multilingual EMM newswire data it became evident that all three suffixes under investigation have a relatively stable rate of appearance. Many more different *-gate* coinages could be found, though. We could observe that *-gate* was usually attached to one specific single event, and especially in many of the less frequent coinages the suffix was combined with proper names of persons, institutions, or locations. In contrast, *-athon* and *-mageddon* coinages seem to be easier to generalize. For example, the two most widely spread coinages *Snowmageddon* and *Carmageddon*, while initially referring to a certain snow storm and a certain traffic jam, have been applied to further such events and can be found listed in resources such as the Urban Dictionary.[28]

Whereas *-geddon* and *-athon* are mostly used to convey a semantic component, many uses of the *-gate* suffix seem to have a rather rhetorical or pragmatic component. First, because the meaning is usually redundantly codified using paraphrases like "the scandal dubbed...", "the so-called... affair", and secondly, coining a name for an issue, affair, scandal or whatsoever, helps to make it persistent and implicitly places it on the same level as other former *-gate* issues and ultimately as the *Watergate* affair. In a sense, the *-gate* suffix is an example for how language can be subtly exploited to form opinions. The chief executive of the Rugby team involved in the *Bloodgate* scandal indirectly supports this hypothesis stating: "You would be incredibly naive to think (the Bloodgate stigma) will ever disappear completely. Things like that don't."[29]

Our results suggest that the spread of coinages with *-gate* suffix seems to depend much more on external non-linguistic factors, such as journalistic or public interests in the corresponding scandal and its news dynamics, than on linguistic factors. Still, it might be interesting to see whether morphological or phonological features play a role when new coinages are termed. For example, we could observe a tendency of having two syllables in front of the suffix *-gate*:

[28]http://www.urbandictionary.com/define.php?term=Carmageddon, last revised on March 14th, 2013

[29]BBC Sport: http://news.bbc.co.uk/sport2/hi/rugby_union/8588557.stm, last revised on December 6th, 2012

There was **no** *Berlusconigate* or *Sarkozygate* in our data, but a *Sarkogate* and a *Merkelgate*. A big scandal centered around Berlusconi, in contrast, was named *Rubygate*.

The research described within this chapter shows how the combination of large-scale data processing with carefully designed visual displays is able to support research in lexical semantics. Automated algorithms first find relevant word occurrences in extremely large amounts of text data and either provide statistical models of their contexts or information about sources, source languages, and source countries. In a second step, the derived information is displayed in a way that enables the detection of expected and unexpected patterns. Some previously existing hypotheses could be investigated, but more importantly, in an interdisciplinary effort we were able to come up with new interesting hypotheses on recent developments in lexical semantics. Visualization was key to that.

Chapter 5

Visual Analytics of Diachronic Change in Text Content

Contents

Whereas in the previous chapters the goal was to detect and explore changes of natural language as such, in this chapter the focus is on the visual analysis of changes in text content over time.

In the last decade the amount of textual information readily available in digital form has increased enormously. The amount of user-generated content has especially grown at a fast pace lately, as the Web 2.0 has enabled easy participation for all internet users. Large amounts of texts are provided through blogs, forums, wikis, twitter messages, companies' online surveys and feedback forms and also through more formal publications like RSS news feeds and online news websites.
These text sources constitute a rich body of information that is valuable to exploit for different stakeholders with different information needs. For example, political analysts want to see when and why political parties and persons are mentioned in negative contexts, and business analysts want to see when and why a certain product or product feature is mentioned in negative contexts. Methods from *text mining*, *natural language processing*, and *computational linguistics* can help to extract interesting features out of the raw text data. However, not only automatic algorithms for data analysis are important, but also the appropriate conveyance of detected peculiarities to the analyst and the offering of possibilities for interactive data exploration. In the case of such complex and ambiguous data as natural language text this requires possibilities to drill-down to the original text sources whenever needed in order to make sense of the automatic analysis, to enable an easy visual detection of interesting patterns, and to provide means to quickly generate or verify hypotheses. Methods from the fields of *visual analytics* and *information visualization* have

been demonstrably shown to support such tasks.
In many concrete text analysis scenarios one crucial requirement is to extract sentiments or opinions contained in the documents. For example, companies might be interested in what their customers like or dislike about their products and services or what sentiments are associated with the brand or its products in news. Similarly, organizations or individuals of public interest have to be aware of what is reported about them in news and how their decisions and statements are reflected. Of course, many more related examples can be found, where opinions or sentiments on certain topics have a high relevance. The vibrant field of *opinion* and *sentiment analysis* is dedicated to detect these kinds of statements from text.
As web communication and publishing is increasingly happening in real-time, a further particularly interesting issue from the data analysis perspective is the involvement of the time dimension. Temporal aspects like the distribution of text features over time can be important for different real-world applications.

This chapter is structured as follows. First, in Section 5.1 I introduce a pilot study that has the goal to provide a first impression of how visual analytics can help to detect different kinds of changes and anomalies in text content over time. The approach is both elementary and innovative in that it gives each document a visual representation and the ensemble of all these visual document objects forms an image where different kinds of temporal patterns emerge visually, whenever text content and sentiment change. From this first study I derive conclusions that help to build a more targeted and scalable approach solving the real-world analysis problem of detecting highly-relevant complaint patterns in customer feedback in Section 5.2. In Section 5.3 I show how further analyses exploring term associations can complement the time-oriented analysis. This second and third part contain a number of innovations both in the automatic processing and visualization. The corresponding research has contributed to the filing and publication of two World Intellectual Property Organization patent applications and five United States patent applications as part of an industry collaboration with the Hewlett Packard Research Labs in Palo Alto, California. To date, one of the applied patents has already been issued.

5.1 Pilot Study: Detection of Sentiment Anomalies in RSS Feeds

This section partly builds on the following publications:

Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, Daniel A. Keim: Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009), 2009. [177][1]

Franz Wanner, Christian Rohrdantz, Florian Mansmann, Andreas Stoffel, Daniela Oelke, Milos Krstajic, Daniel A. Keim, Dongning Luo, Jing Yang, and Martin Atkinson. Large-scale Comparative Sentiment Analysis of News Articles. Poster Paper presented at the IEEE Information Visualization Conference (InfoVis 2009). 2009.[2]

5.1.1 Background

In this initial pilot study, my aim is to explore what kinds of discoveries, regarding patterns of change, can be made by visually exploring time-stamped text data, without previously building models based on the data. The idea is to do just a basic data processing and directly plot the raw data (documents) as visual objects. This has the advantage that no a priori assumptions have to be made like choosing a suitable clustering algorithm, making assumptions on the number of clusters, binning or aggregating the data along the temporal dimension, etc. This avoidance of a priori assumptions supports an unbiased explorative analysis with the goal of bringing up interesting patterns and find-

[1] I joined the research of this paper, when fundamental decisions about the data and tasks and first visualization sketches had already been made. My contribution is limited to the details of the visual design including the usage of parallel lines, and semi-transparent distinguishable visual objects. I did all the programming and wrote the section describing details on the technique and design. For parts of the original publication not written by myself, I will cite the original publication.

[2] For this project again I did the programming of the visualization and contributed the idea of having an overlap reducing zoom instead of a geometrical zoom. The publication can be accessed online only: http://bib.dbvis.de/uploadedFiles/32.pdf

ings that might help achieve a better understanding on what kinds of patterns are contained in such text data. In this case we investigate *Rich Site Summary* (RSS) Feeds. RSS Channels are similar to news-tickers and usually contain a small piece of newsworthy text in a standardized XML format. Our approach was to monitor RSS Feeds dealing with the US Presidential election in the USA, one month before the election. The kind of patterns we were interested in, were sentiment patterns over time and especially changes in sentiment that reflect real-world events related to the electoral campaign. Section 5.1.2 reveals more details about the data processing, Section 5.1.3 gives details about the design decisions for the visual mapping and the possibilities for interactive exploration. Results are discussed as part of a case study in Section 5.1.4, and advantages and disadvantages of the approach are discussed as part of the conclusion in Section 5.1.5.

5.1.2 Data and Resources

The data we used was gathered from 50 different RSS News Feeds, that mainly dealt with the 2008 US presidential elections. The RSS Feeds were retrieved every 30 minutes during a time interval of one month (10/09/2008 - 11/10/2008). For every news item within a retrieved feed, we saved date, title, and description, as well as the id of the feed. Next, noise was eliminated out of the title and description. With noise we refer to strings that do not carry any content, such as URLs or strings consisting of special characters. The concatenation of title and description was then considered to be the content of the news item. Finally, we deleted those documents that did not contain any of the following signal words: *Obama*, *McCain*, *Biden*, *Palin*, *Democrat*, and *Republican*. More than 23,000 news items contained at least one of the six words.

For the automatic data processing we applied ad hoc solutions and heuristics. The applied methods work reasonably well, but have not been optimized as this was not the goal of this initial study. Pairwise similarities between news items were calculated by applying a similarity measure, which counts the number of non-stopwords that two items have in common (normalized by the length of the larger item). This relatively simple measure is quite effective for the short texts we have.

Another aspect of interest is the sentiment context of a news item. Therefore

every item is enriched with an automatically determined sentiment score. For this purpose we make use of a freely available list of words that evoke positive or negative associations. We count the number of positive and negative words and evaluate the whole news item as rather positive, if it contains in total more positive words and negative in the opposite case. The absolute relation of positive against negative words normalized by the item's length, provides our sentiment score. One important point to mention here is that the appearance of a candidate e.g. in a negative context, does not necessarily mean, that the item contains negative publicity for the candidate, but simply that s/he appears in a negatively connoted context. This becomes clear when we consider the example of news telling that racists planned to assassinate Obama (see Section 5.1.4). This was bad news for Obama, not about Obama, with a visibly negative connotation.

5.1.3 Item-based Plotting with Visual Aggregation

The visualization, on the one hand, aims to give a meaningful representation of the data and on the other hand is intended to be an appropriate starting point for the interactive exploration and discovery of interesting patterns. Figure 5.1 shows a screenshot of the visualization. Each line represents one day and each colored object depicts one news item. The news item's sentiment score is encoded by a vertical displacement of the news item. Colors encode whether the text mentions the Democratic party, the Republican party or both. Additionally, the shape of the news objects visualizes whether the first candidate, the second candidate or only the name of the party itself was mentioned. The following passages describe each of those aspects in detail.

Placement

Every news item is represented by an object in a 2D plane. The position of the object within the plane depends on the date the news was published. Thereby, the day it was published accounts for the line it will be placed in (as each line represents one day) and the time of day determines its horizontal position along this line. The exact vertical position depends on the sentiment score of the object. According to this value an object is slightly shifted upwards

(positive) or downwards (negative). Horizontal lines mark the position that a news item would have that is neither positive nor negative.

Coloring

Everything that is solely related to the conservatives (Republican party) is colored in red and everything purely related to the liberals (Democratic party) in blue. Gray news objects relate both to the liberals and the conservatives, which basically means that both camps are mentioned within the news' content.

Shape

The use of different shapes for the object allows us to make a distinction between (a) news items in which the first candidate of a party was mentioned, (b) news items where the second candidate, but not the first candidate, was mentioned, and (c) those containing none of the candidates, but only the name of the party. Figure 5.2 shows the visual appearance of the different shapes. We keep the horizontal interruptions, that are utilized to mark news items that talk about the second candidate, always at the same vertical position. If several neighboring objects refer to one of the second candidates only, this leads to a clear visual pattern of a continuous white horizontal interruption.

Semi-transparent plotting to support visual aggregation

We paint our news objects with a relatively low opacity. That means they are partly transparent, which comes with two advantages: First, the problem of overlapping news objects is reduced. In most cases every object is visible and can be differentiated clearly from its overlapping neighbors. Secondly, if multiple news items are put on top of each other, the overall opacity at this position increases, resulting in an object that is less transparent and can therefore be distinguished from objects that represent just one news item. The situation that several feeds bring the same news nearly at the same moment in time is often the case when the news is very important. That means that the less transparent news objects often represent news that is more important and surely more widely spread. The right part of Figure 5.2 visually illustrates the above mentioned design decisions.

Figure 5.1: Overview on all RSS feeds retrieved during 31 days. Several interesting patterns are visible, annotated with A, B, C, D, and E. Reprinted from [177].

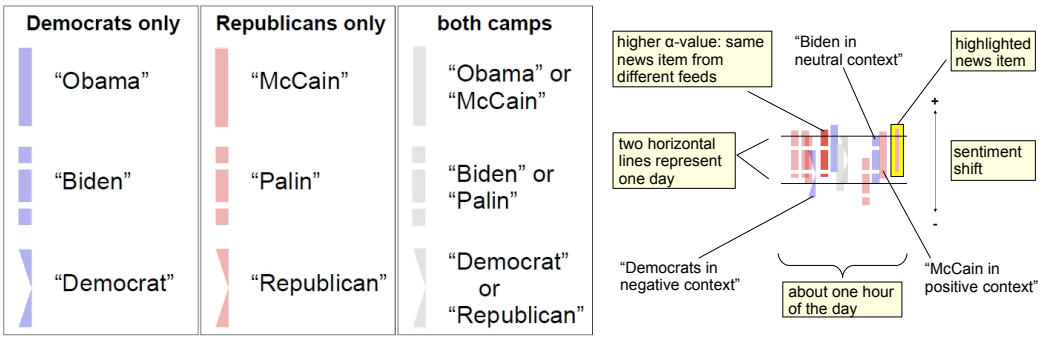


Figure 5.2: Symbols used to represent news items according to the appearance of certain keywords (left). The right side of the figure shows an example for the semantics of our visualization. Reprinted from [177].

Interaction

The visualization is designed for interactive data exploration. There are several possibilities of interaction with the tool:

- Zooming: Continuous zooming allows the analysis of certain parts at a greater level of detail.
- Details on demand: When the mouse is dragged over a news object, a tooltip appears containing date, time, feed id, and content of the item.
- Similarity search: With a mouse click on a news object, the search for similar news items is started. The news item itself and every other news object that is related to it is highlighted (please refer to section *Data Processing* for our definition of similarity). Figure 5.3 shows an example.
- Filtering: The user can select the different candidates/parties he is interested in. Another possibility to reduce the number of items that are displayed is to select one specific RSS feed. Both filtering mechanisms can be used to analyze in detail the behavior of one specific news provider or the development of news for a subset of candidates and/or parties.

5.1.4 Case Study: Discovery of Unexpected Patterns

For analysts it is interesting when the sentiment referring to certain entities shows unexpected behaviors over time. Different patterns of sentiment anomalies can be detected in Figure 5.1.

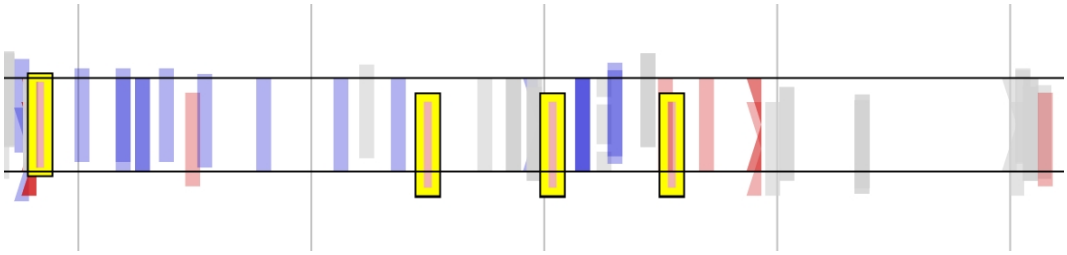


Figure 5.3: After selecting one news item, similar items are highlighted in yellow enabling the user to track specific topics (low threshold) or redundant postings (high threshold). Reprinted from [177].

The visually salient area marked in Figure 5.1 with (A) is enlarged in Figure 5.4. It shows an accumulation of interrupted red bars, which indicates the heavy mention of the key entity *Palin* only. The tendency shows a negative shift of the items, i.e. a negative sentiment trend. This visual pattern indicates bad news about the Republican candidate Sarah Palin dealing with the so-called *Troopergate* scandal. Palin is accused of abusing her power as Alaska Governor, when firing the state's public safety commissioner. It can be seen that after a while the negative news about Palin also become negative news for McCain who chose Palin. Finally, a positive outlier can be detected as the accused camp reacts on the accusations declining them with positive phrasing. The pattern is an example that might be relevant for analysts: Suddenly one entity is mentioned more frequently as usual and in a negative context. It also shows that single outliers within such temporal clusters can be interesting.

The visually salient area marked in Figure 5.1 with (B) is enlarged in Figure 5.5. It shows a large number of rather negatively connotated news items mentioning the presidential candidates from both camps. This was caused by a TV debate in which both candidates battled fiercely.
This example shows that sometimes the sudden co-occurrence of two entities, which usually do not co-occur at such a high frequency, is another kind of anomaly that analysts might want to explore further.

The visually salient area marked in Figure 5.1 with (C) is enlarged in Figure 5.6. It shows a large number of negative news mentioning Obama. When analyzing the corresponding texts it becomes evident that it is not actually bad news about Obama, but bad news for him: An assassination plot targeting Obama was uncovered.

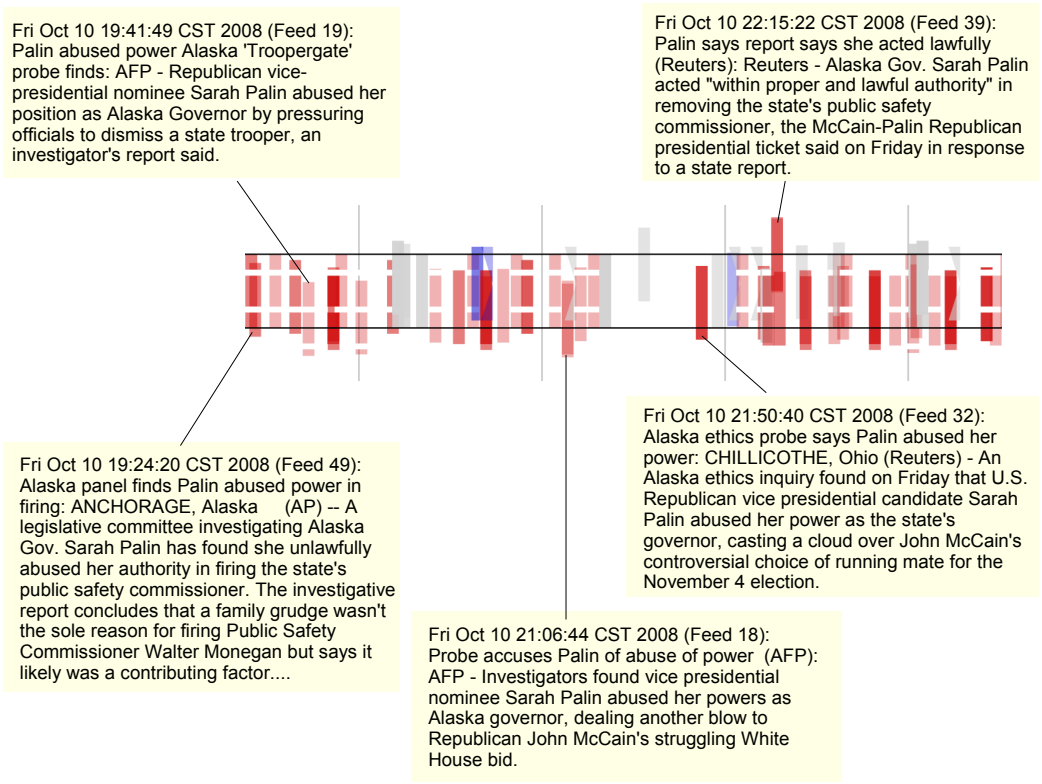


Figure 5.4: Enlarged version of area (A) from Figure 5.1. Reprinted from [177].

The example shows that it is important to carefully explore interesting patterns and have a look at the frequent words within the pattern or read a certain number of the text documents in order to gain a better understanding of sentiment anomalies.

The visually salient area marked in Figure 5.1 with (E) is enlarged in Figure 5.7. This example shows that sometimes different news signals of different strengths exist simultaneously. In this case, after the election by far most of the news is about the winner Barack Obama. However, another weaker news signal comes up that relates in a negative way to the Rebuplican candidate for the vice-presidency, Sarah Palin. In this case only the difference in sentiment makes the smaller burst visible.

5.1.5 Discussion and Conclusion

The presented visualization approach has several advantages that can be summarized as follows:

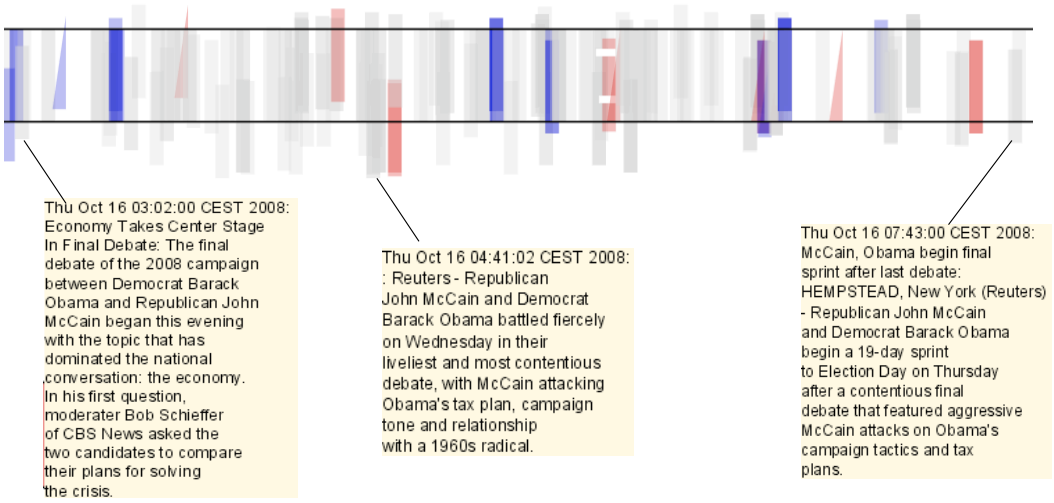


Figure 5.5: Enlarged version of area (B) from Figure 5.1. Reprinted from [177].

Advantages:

- Patterns do not have to be defined beforehand, no complex models have to be derived from the data for visualization. Thus, the method is rather generic, does not depend on a priori assumptions, and is readily applicable for exploration tasks.
- Compact display for overview providing a broader context.
- Partly transparent overplotting leads to visual aggregation and makes patterns emerge.

Yet, there are also some disadvantages that should be addressed with further complementary research:

- Scalability: Only 6 words (*Obama*, *Biden*, *Democrat*, *McCain*, *Palin*, *Republican*) monitored.
- The overplotting leads to the effect that not every item is clickable for details-on-demand.
- At each point in time the strongest signal might cover weaker interesting signals. Though, this is not the case if different signals have different sentiments.

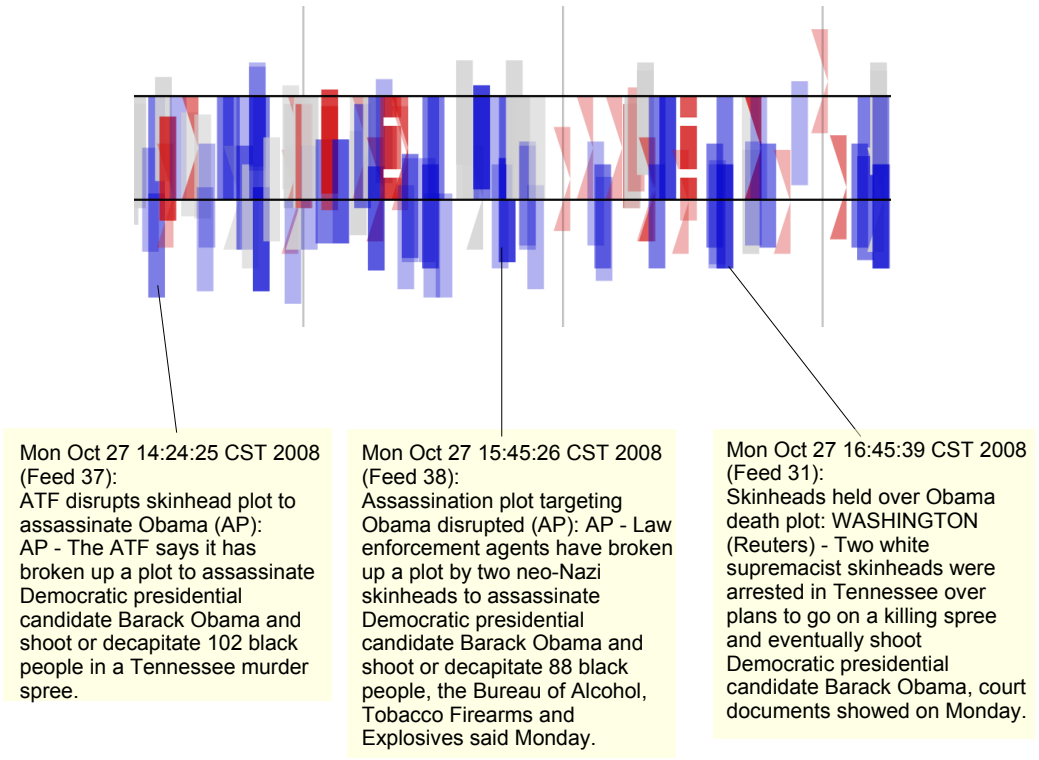


Figure 5.6: Enlarged version of area (C) form Figure 5.1. Reprinted from [177].

- The visualization does not show the sentiment that is conveyed about an entity, but the sentiment of the news context in which the entity is mentioned.

Some of the disadvantages can be partly overcome through interaction. For example, problems of data occlusion that might be caused by the potentially high amount of overplotting. In a small follow-up publication we introduced the idea of altering the geometrical zooming capability. Continuous zooming allows to analyze certain parts at a greater level of detail. From a certain zoom level on, the horizontal scale of the visual object representing news items (in this case triangles) is reduced while the background scale is still enlarged. This has the desired effect that the triangles are not simply becoming constantly larger but are separated when a further enlargement would not reveal additional insights. Thus, there always is a zoom level where each single news item will be displayed without overlap in order to allow a more in-depth analysis for a certain time interval, as illustrated in Figure 5.8.

Based on this pilot study a new approach was designed, which will be pre-

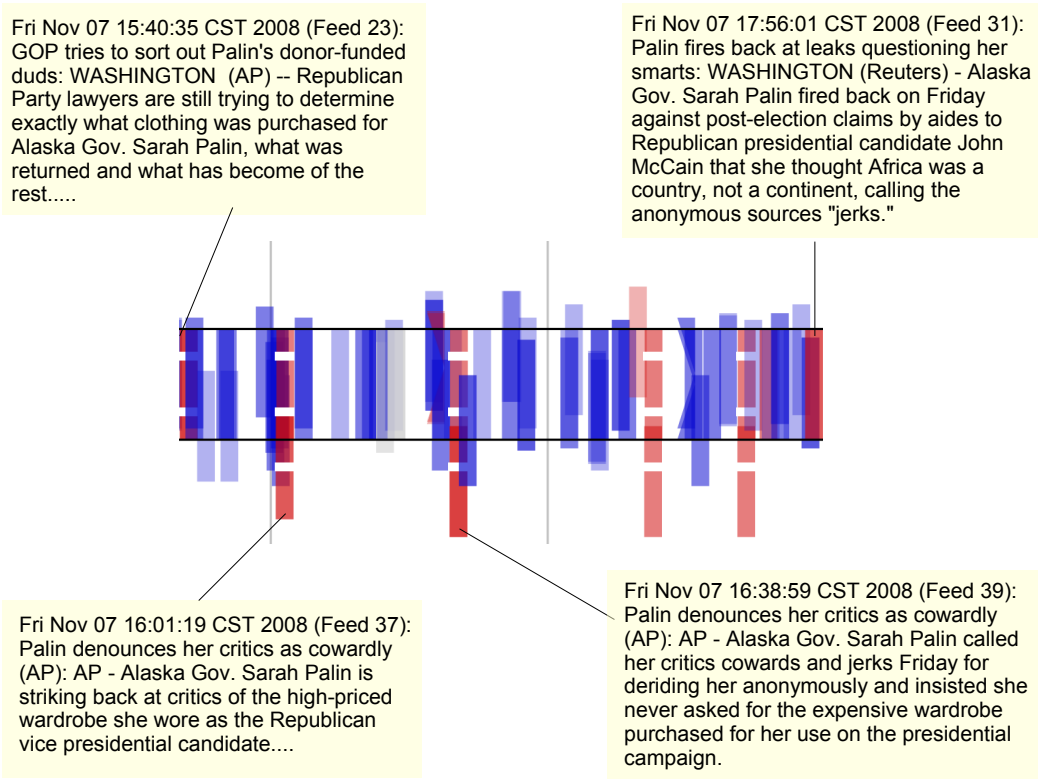


Figure 5.7: Enlarged version of area (E) from Figure 5.1. Reprinted from [177].

sented in the next section. The aim is to overcome the disadvantages while maintaining the advantages. In other words, the new visual analytics approach is scalable in that it enables the monitoring of a large number of different words. Overplotting is avoided. Whether the temporal accumulation of a word is meaningful is made dependent on the overall frequency of the word. For a generally infrequent word a small burst may already point to relevant findings. The sentiment analysis module shall be improved to do a more detailed analysis on which sentiment refers to which target word, instead of just considering the prevailing sentiment of the surrounding text snippet.

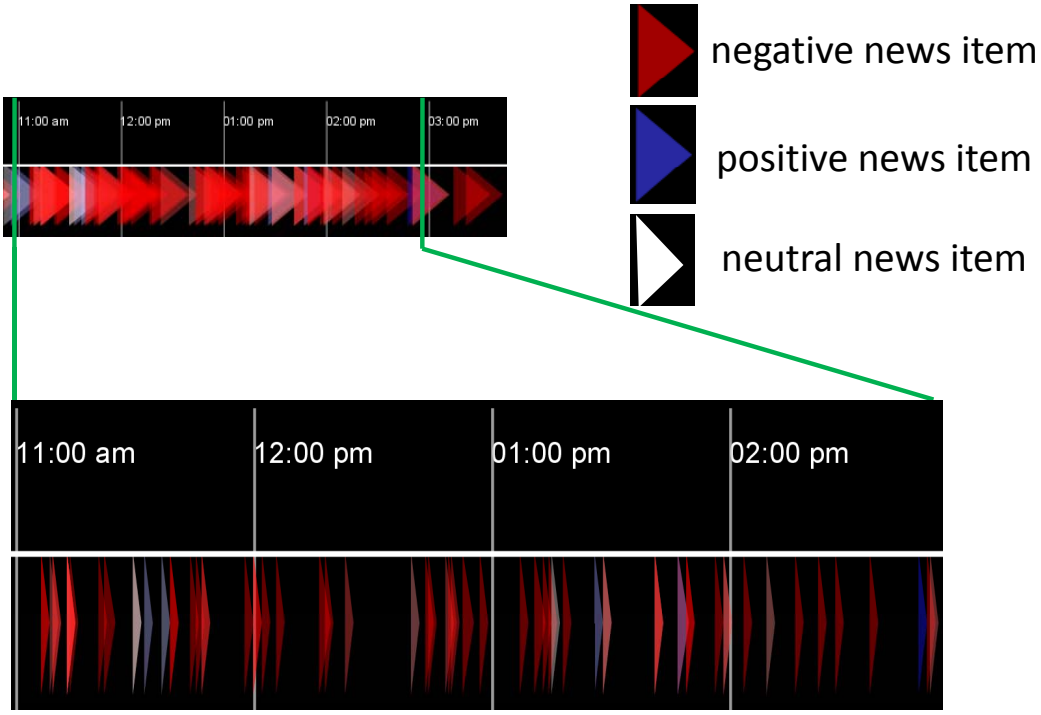


Figure 5.8: Non-overlapping zoom for an in-depth analysis of a certain time interval. Reprinted from our InfoVis 2009 Poster *Large-scale Comparative Sentiment Analysis of News Articles*.

5.2 Critical Time-Related Issues in Target-based Sentiment Analysis

This section builds on the following publication:

Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, Daniel A. Keim: Feature-Based Visual Sentiment Analysis of Text Document Streams. ACM TIST 3(2): 26 (2012).[3]

Furthermore, the ideas that I have developed as part of the research presented in this section have contributed to the filing and publication of a number

[3]For this publication I did all the research, programming, and almost all of the writing by myself. The only exception is the use and description of the pixel bar charts, which was done by Ming C. Hao. The other collaborators provided data, gave advice, feedback, and did proof-reading. As also acknowledged in the publication Meichun Hsu supported this research with her suggestions and encouragement. For all parts of the publication that were not written by myself I reference the original work.

of patent applications. One of the applications has already been issued.

United States Patent Grant US 8595151 B2: Selecting Sentiment Attributes for Visualization. Filing Date 08.06.2011. Publication Date 26.11.2013. Inventors: Ming C. Hao, Umeshwar Dayal, Christian Rohrdantz, Meichun Hsu, Mohamed Dekhil, and Riddhiman Ghosh [62].

World Intellectual Property Organization Patent Application WO/2012/044305: Identification of Events of Interest. Filing Date 30.09.2010. Publication Date 05.04.2012. Inventors: Ming C. Hao, Umeshwar Dayal, and Christian Rohrdantz [64].

United States Patent Application US 2012/0109843: Visual Analysis of a Time Sequence of Events Using a Time Density Track. Filing Date 27.10.2010. Publication Date 03.05.2012. Inventors: Ming C. Hao, Christian Rohrdantz, Umeshwar Dayal, Daniel Keim, and Lars-Erik Haug [69].

United States Patent Application US 2012/0060080: Visual Representation of a Cell-based Calendar Transparently Overlaid with Event Visual Indicators for Mining Data Records. Filing Date 03.09.2010. Publication Date 08.03.2012. Inventors: Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Christian Rohrdantz [63].

United States Patent Application US 2013/0046756: Visualizing Sentiment Results with Visual Indicators Representing User Sentiment and Level of Uncertainty. Filing Date 15.08.2011. Publication Date 21.02.2013. Inventors: Ming C. Hao, Christian Rohrdantz, Umeshwar Dayal [68].

This section describes automatic methods and interactive visualizations that are tightly coupled with the goal to enable users to detect parts of text document streams relevant for their tasks. In this scenario the interestingness is derived from the sentiment, temporal density, and context coherence that comments about features for different targets (e.g. persons, institutions, product attributes, topics, etc.) have. Contributions are made at different stages of the visual analytics pipeline, including novel ways to visualize salient temporal

accumulations for further exploration. Moreover, based on the visualization an automatic algorithm detects and preselects salient time interval patterns for different features in order to guide analysts. The main target group for the suggested methods are business analysts who want to explore time-stamped customer feedback to detect critical issues. Finally, application case studies on two different datasets and scenarios are conducted and an extensive evaluation is provided for the presented intelligent visual interface for feature-based sentiment exploration over time.

5.2.1 Background

More and more people use the Web and other online channels to convey their sentiments and opinions, for example on products, brands, and services. These customer comments are a valuable source of feedback and an external quality control for manufacturers and retailers. It is crucial for them to track such feedback and derive conclusions from it in order to arrive at improved decision-making processes and to eliminate sources of customer dissatisfaction. In this section we work with customer feedback sent to a company through online web surveys over the course of two years. We introduce a visual analytics pipeline in order to process, analyze, and visualize these data. Innovations are part of almost every step of the pipeline. The goal is to point business analysts to relevant time-related issues as described by customers and to offer novel visualization methods for interactive exploration.

5.2.2 Related Work

This section describes relevant related work on automatic and visual feature-based sentiment analysis and the visual analysis of time series.

Feature-based Sentiment Analysis

Feature-based sentiment analysis is a subtask of opinion and sentiment analysis. In literature the terms opinion and sentiment are often used interchangeably. For simplicity, in our approach we will use the term sentiment only.

Most approaches for feature-based sentiment analysis involve three or four consecutive steps:

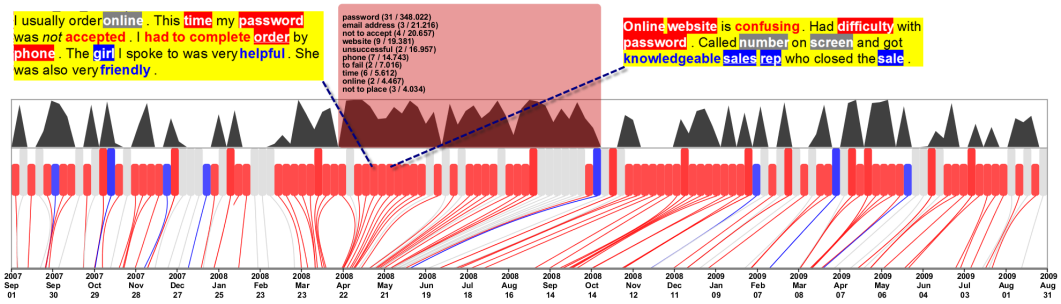


Figure 5.9: *Time density plot* of an issue on the feature *password* with associated terms (top with red background) and automatically annotated example comments (top with yellow background). Among 50,000 customer comments, received within two years, all those are sequentially displayed that contain the noun *password*. Each comment is represented by one vertical bar. The color indicates whether the noun *password* has been mentioned in a positive (blue), negative (red) or neutral (gray) context. The height of a bar can encode another data dimension. In this case we experimented with the uncertainty involved in the sentiment analysis, the lower the bar, the more uncertain. The curve plotted on top of the *sequential sentiment track* is a *time density track*: the curve is high if the comments below have been relatively close in time. Each document bar is connected with a link to its position along a linear time line on the bottom. Links are bundled according to their relative time density. An automatic algorithm detects and highlights interesting time intervals in the visualization that analysts should explore in detail. Mousing over single comments, the content is displayed and the coloring of words indicates what the sentiment analysis has found. All nouns get a background coloring according to their sentiment context, sentiment words get font colors and negation words are printed in italics. If the sentiment analysis of a noun was evaluated to be confident (little uncertainty) the corresponding word is underlined. Here, this is the case for *order* and *sales rep*.

1. Features for different targets (e.g. persons, organizations, products, services or topics) are detected either directly from the corpus or based on predefined word lists.

2. Sentiment words that describe the extracted features are searched for in the documents. Sentiment words are words that evoke positive or negative associations.

3. A mapping strategy aims at detecting which sentiment words refer to which feature, so that a sentiment score can be determined for each feature.

4. Some approaches visualize the results of the feature-based sentiment analysis and enable the user to interactively explore the results in detail.

For the first two steps abundant research has been published in the last years. For the sake of brevity I refer to comprehensive summaries given in [136] and [107] for details. Both features and sentiment words can be either learned from the processed text documents themselves, from external resources (like e.g. WordNet[4]) or they can be gathered from predefined lists. One special challenge is to identify sentiment words that have no general validity, but depend on the domain or even feature. For example, in a domain like *printer* an adjective like *fast* is feature-dependent, i.e. positive in the sentence "the printer prints fast" and negative in the sentence "the ink cartridge runs out fast".
Details about steps 3 and 4 are listed in the following paragraphs.

Sentiment-to-Feature Mapping Different approaches have been suggested in the past to determine which sentiment words refer to which feature. Some of them use distance-based heuristics, i.e. the closer a sentiment word is to a feature word, the higher is its sentiment influence on the feature. Such approaches operate on whole sentences [43], on sentence segments [42, 91] or predefined word windows [135].
Other approaches exploit advanced natural language processing methods, like

[4]http://wordnet.princeton.edu/ last revised on March 18th, 2013

typed-dependency parsers, to resolve linguistic references from sentiment words to features. There are several methods that resolve such references and thus can be used for feature-based sentiment analysis, although most of them were created for different purposes. Ng et al. [129] use subject-verb, verb-object, and adjective-noun relations for polarity classification. Qiu et al. [141] use dependency relations to extract both features (product attributes) and sentiment adjectives from reviews by a double propagation method. Popescu and Etzioni [139] extract pairs (sentiment word, feature) based on 10 extraction rules that work on dependency relations and Riloff and Wiebe [142] use lexico-syntactic patterns in a bootstrapping approach for subjectivity classification resolving relations between opinion holders and verbs.
Our method differs from the previous ones in that we use a predefined set of simple syntactic reference patterns that are based on part-of-speech sequences only, in order to resolve references - instead of using typed dependencies. In cases where this method is not able to resolve references, we rely on a distance-based heuristic. This approach also allows us to estimate a degree of uncertainty involved in the analysis.
Recently, another approach was published that takes uncertainty into account. The authors consider if customers do not express clear opinions, i.e. "customers' conflict and uncertainty about their opinions" [184] as well as the uncertainty involved in the automatic opinion analysis processing. As a result a feature mention can be both negative and positive at the same time. Their uncertainty score is based on two parameters: The smaller the difference between the negative and positive sentiment on a feature within a sentence and the longer the sentence, the higher the uncertainty. We also capture uncertainty in our analysis, however, we limit the analysis to the uncertainty the algorithm has when evaluating a sentence. In contrast to the existing approach, our method relies on basic linguistic knowledge and not only on distance-based heuristics. In addition, the sentence length is not relevant in our analysis as we consider only sentence segments.

Visual Exploration of Feature-based Sentiment Analyses Several approaches have been suggested for the visualization of the outcome of automatic feature-based sentiment analyses and for the enabling of further user explo-

rations. The Opinion Observer [108] visualization enables users to compare products with respect to the amount of positive and negative reviews on different product features. A more scalable approach for the same purpose that is able to display more products and features at once is that of the Summary Reports presented in [135]. The paper was co-authored by myself but is not within the scope of this thesis. It also provides further visualizations for the identification of groups of customers with similar opinions and correlations between individual feature scores and overall ratings. The AMAZING System [123] also visualizes the sentiment of product reviews on certain products over time. The number of positive and negative reviews are aggregated over months and displayed with line charts. In the pilot study of Section 5.1 and [177] a visualization is suggested for tracking sentiments expressed in RSS news feeds on political parties and their candidates during a presidential election.

Later, OpinionSeer [184] a novel visual analysis tool for hotel reviews was introduced, where uncertainty contained in reviews is visually represented and aggregated analyses can be performed e.g. on day, week, and month scale.

In contrast to the previous work, the approach presented in this section enables a much more detailed insight into the temporal development of sentiments on individual features. For each feature an interactive visualization is created that combines a sequence view with a linear time line and additionally conveys the uncertainty of the underlying sentiment analysis.

Visual Time Series Analysis

A comprehensive survey about the visualization and visual analysis of time series is given in [3]. Several further publications on the visual exploration of time series data are related to the TimeSearcher Project.[5] Methods especially designed for text time series are often based on a linearly scaled time line, aggregating events according to predefined time bins. Many of these approaches have been inspired by the ThemeRiver method [72]. A complete overview of related work is provided in Section 2.2 of this thesis. One particularity of our approach is that it deals with unevenly spaced data streams in which events

[5] http://www.cs.umd.edu/hcil/timesearcher/ last revised on February 11th, 2013

(here: feature occurrences) may appear with an arbitrarily skewed temporal distribution. That means that the data includes short time spans with high amounts of incoming data and large time spans that are only sparsely populated. In [10] several methods are presented to deal with unevenly-spaced auction data. The (interleaved) event index method is the most similar one to our time density plots. It distorts the time axis in order to grant the same amount of space to each event. While the temporal order is preserved the exact temporal relations are lost. Since the exact time between two consecutive events is not conveyed, the authors try to support the user by shading the time axis segments.
Our visualization complements the previous work, in that it displays data records in sequential order without overlap and empty space, while still conveying information about exact temporal relations.

5.2.3 Data and Resources

Our target users are industries, companies, and small businesses who want to explore their customer feedback. In addition to web surveys, we also can readily apply our visualization and pattern detection methods to time-stamped news, twitter data, hotel reviews, movie/recreations reviews, etc., as long as the quality of the sentiment analysis in that domain is reasonable. We use mostly standard methods for the automatic sentiment analysis that were suggested for mining customer reviews.

To apply our methods to datasets with different analysis scenarios and characteristics, in addition to customer web surveys, also RSS news feeds were explored.

- *Customer Web Surveys*: Web surveys give users the opportunity to directly comment, in a detailed way, on issues they liked or disliked about the product itself and its purchase, service, delivery, payments, etc. This kind of information can be especially valuable for companies as it might point them to problems that they had been unaware of beforehand, having negative effects on their business performance if the problems are not detected and eliminated in time. We gathered a dataset containing about 50,000 web survey responses sent to a company between 2007 and

2009. In this dataset the available time resolution is on daily basis and I assigned random hours and minutes to all day time stamps in order to get a more realistic data distribution.

- *RSS News Feeds*: RSS news feeds redistribute and spread current news. For example, they are interesting for political analysts who want to see when and why political parties and persons are mentioned in negative contexts. To explore the applicability of our methods for such a related task, we analyzed about 16,000 RSS news items collected from 50 feeds about the US-presidential election in 2008. The collection started about four weeks before the election and ended on the election day. The data is a subset of the data used in Section 5.1 and [177] with focus on the pre-election phase.

5.2.4 A Visual Analytics Pipeline for the Discovery of Time-Related Sentiment Patterns

Most of the previously mentioned feature-based sentiment analysis approaches deal with collections of customer reviews on a certain product, as can be found on retailer sites such as amazon.com. In contrast, this approach focuses on customer reviews that are directly sent to a company via a web survey. This direct feedback is not necessarily related to products but refers to any issue within the purchase and service process. Most importantly, not only the sentiment polarity but also the temporal and context coherence of customer comments are considered to detect critical issues that occur at certain points in time. This approach covers the whole pipeline of methods necessary to detect important sentiment pattern information in large document streams, and contributes to different stages of the analysis process by suggesting novel automatic and visual analysis approaches, see Figure 5.10. The required input for our analysis is rather generic in order to guarantee a wide applicability. It consists of a set of time-stamped texts. To give an overview of the analysis steps, they are listed in the following. Contributions are shortly explained:

- Linguistic Preprocessing

- Feature Sentiment Identification: In the sentiment-to-feature attribution

we aim to achieve a good coverage while being as accurate as possible. Therefore, we combine different methods to resolve sentiment-to-feature references and together with the analysis results we give an estimation for the uncertainty involved in the analysis. This is a minor contribution that is not central to the overall approach but I consider it interesting to explore.

- Context Identification: Nouns, adjectives, and verbs are considered to provide most of the relevant context. These parts of speech are extracted and saved separatly.

- Feature Time Density Calculation: Along the temporal dimension, we try to detect shifts in the occurrence frequency of a certain feature which may indicate time-related issues. The time density is calculated relative to the overall occurrence frequency of the feature. This allows us to detect interesting time patterns also for infrequent features.

- Visualization and Interactive Visual Analysis: To visualize sudden temporal accumulations of comments on one feature, we propose an innovative visualization method: *Sequential sentiment tracks* together with *time density tracks* are able to display unevenly-distributed feature occurrences without overlap and space-consuming gaps. In addition, the sequential sentiment tracks are linked to a linear time line in order to convey details about the time distribution. Here, we introduce *time density edge bundling* as an additional visual clue. Critical issues can readily be detected visually and explored in detail interactively accessing the relevant full text as a tooltip, as shown in Figure 5.9. To provide a global overview about the data distribution pixel map calenders are applied.

- Time Interval Pattern Detection: In order to guide an analyst and advise her/him of critical issues, we further propose a new time pattern detection algorithm that operates on past data. Interesting time spans for features will be filtered and ranked according to their importance scores. Patterns have to be comparatively dense in time, with a smooth time density curve, have to have a clearly negative sentiment connotation and the feature has to appear in similar and specific contexts within the

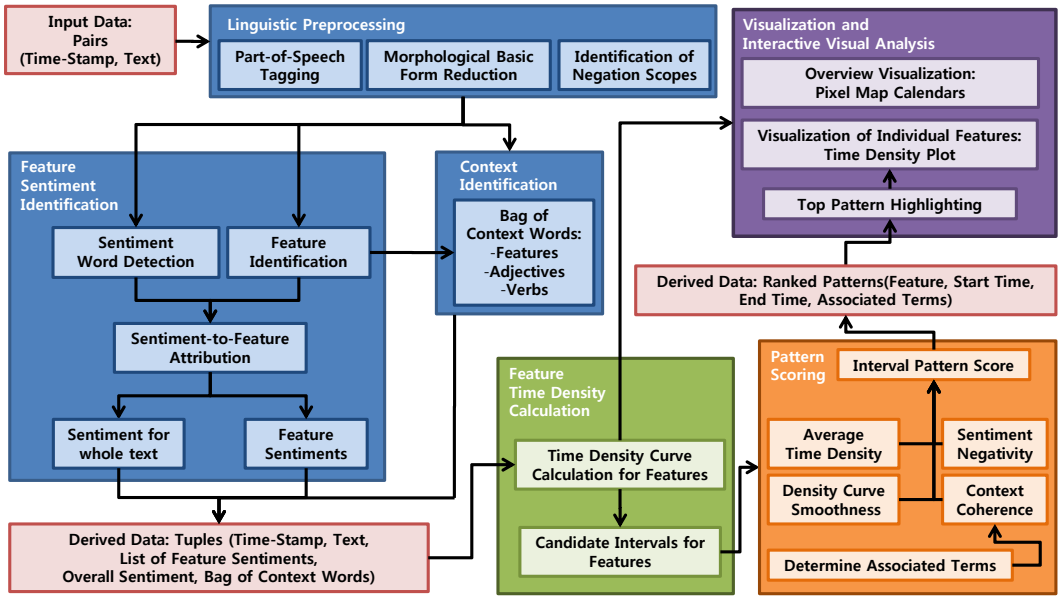


Figure 5.10: Overview of the steps involved in the visual analysis. Reprinted from [144], © 2012 Association for Computing Machinery (ACM).

documents of the pattern. With the purpose to determine the context coherence, terms are extracted that have a strong association with the pattern. Patterns are highlighted in the visualization and the associated terms are displayed in order to provide a quick insight.

This section gives details about the feature-based sentiment analysis, explains the visual analysis components, and details the detection of interesting time interval patterns. In Section 5.2.5 we provide application case studies discussing interesting results that were obtained on real data. An extensive evaluation of different parts of our approach including an expert user study is given in Section 5.2.6, where advantages and limitations are also discussed. Section 5.2.7, finally, provides a discussion and conclusion.

Feature-based Sentiment Analysis in Document Streams

The kind of data we deal with does not have a pre-defined limited topic coverage. There is no fixed set of features, i.e. we are interested in any kind of feature for any kind of target (persons, organizations, products, services, topics etc.). This also implies that we cannot define a domain- or attribute-dependent sentiment word list, but have to rely on sentiment words with general validity.

The feature-based sentiment analysis comprises several steps where we apply standard methods:

1. Linguistic Preprocessing: In a preprocessing step we apply part-of-speech tagging[6] and lemmatization. Next, predefined negation words and their scope are detected in sentences. For this purpose, a list of negation words (like *no*, *not*, *never*, *without*, etc.) and further negating words (like *lack*, *miss*, *stop*, etc.) is defined. Later, the polarity of sentiment words occurring after negations is inverted. The negation remains valid in the same sentence until one of the words or punctuation marks typically marking the end of a negation frame is encountered (e.g. ",", "-", *but*, *and*, *though*, *however*, etc.).

2. Feature Extraction: All nouns and compound nouns are extracted as candidate features. Whether a feature is interesting or not will only be determined in the later time-related analysis. Features and further content-bearing context words (verbs and adjectives) are saved together with the information whether they appeared in a negated context. The context words will be used when evaluating context coherences of interesting feature time interval patterns.

3. Sentiment Word Detection: The polarity categories (positive, negative) from the Internet General Inquirer[7] are applied in order to find sentiment words. The lists have been manually enhanced based on the analysis of common errors by removing some words and adding further colloquial words. The positive word list contained 1594 words after removing 40 and adding 90. The negative word list contained 2018 words after removing 14 and adding 138.

4. Sentiment-to-Feature Mapping: While the processing steps 1-3 are very similar to what has been done by other approaches before, for example [135], this step includes novelties in that it relies both on syntactic reference patterns and distance-based heuristics. It is described in detail in the following paragraph.

[6] http://opennlp.apache.org/ last revised on February 11th, 2013
[7] http://www.wjh.harvard.edu/ inquirer/ last revised on February 11th, 2013

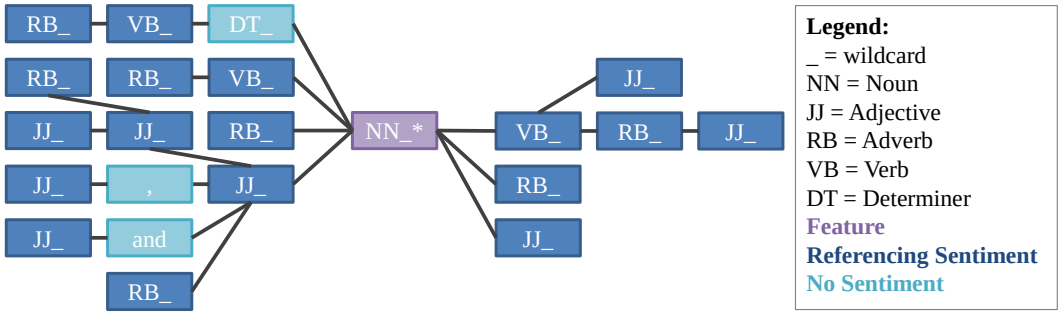


Figure 5.11: Syntactic Sentiment Reference Patterns. Word order patterns go from left to right, the level indicates the exact position. The first pattern at the top left e.g. would match a sentence like "I/PRP really/RB like/VBP this/DT printer/NN". The positive polarity of the verb *to like* would then be attributed to the noun *printer*. The graph summarizes the most frequent reliable patterns we could detect in our data and we therefore regarded in the analysis. In total, the graph covers 18 different patterns, one for each blue node. Reprinted from [144], © 2012 Association for Computing Machinery (ACM).

Sentiment-to-Feature Mapping in Document Streams As outlined in the related work there are distance-based methods for Sentiment-to-Feature mapping and methods based on typed-dependency parses. The first set of methods has the problem that it does not involve any linguistic knowledge and the latter type suffers from high computational complexity and error-proneness. A series of simple tests we conducted indicates that such a parsing is not feasible for large amounts of text documents due to the exponentially increasing processing time with increasing sentence length. To illustrate the effect we sent three requests to the Stanford Parser[8] and retrieved the quickest response time out of 20 trials: (1) *It rains.* (0.006s), (2) *It rains quite often.* (0.020s), (3) *It rains quite often here these days, but still not as much as in other places that I have visited during my last trip.* (0.824s). In addition, Oelke presents a series of experiments comparing feature mapping using typed-dependency parses and simple word distances and comes to the conclusion that "word distance mapping still is the better overall approach" [134, p.128]. This may be partly due to the fact that Internet users often use colloquial and sometimes ungrammatical language, which is hard to parse correctly.
On the other hand, it is not very accurate to rely only on distance-based heuris-

[8]http://nlp.stanford.edu:8080/parser/index.jsp last revised on February 11th, 2013

tics. Therefore, we chose to have a hybrid sentiment attribution approach and also account for the uncertainty involved.
In the first step, we make use of a set of manually defined syntactic reference patterns [92][9] (see Figure 5.11). The only preprocessing requirement is part-of-speech tagging, which has to be performed anyway in order to extract features. We determine three levels of certainty:

1. If a sentiment word stands in one of the syntactic pattern relations from Figure 5.11 to a feature, then this mapping is considered to be correct with a high certainty, i.e. we assign certainty level 1. In this case the certainty value is 1.

2. If no sentiment word could be found in such a syntactic relation, then we use the distance-based mapping from Ding [43]. We modify this mapping by not considering the whole sentence, but sentence segments [42] and word windows. First, we try to detect sentence segments by searching typical segment borders (*but*, *except*, ",", *though*, *however*, etc.). Next, we consider only the segment containing the feature and introduce a threshold for the maximal distance that is still to be considered, like in [135]. In an set of experiments with manually annotated data, we determined the best threshold for this reference window to be 10. If only sentiment words of one polarity, that means either only positive or only negative words, can be found within that sentence segment, then the certainty level is 2. In this case the certainty value is 2/3.

3. If both polarities are encountered within a reference window, the polarity with lower distance from the feature is assigned, but only with a certainty level of 3. If the feature itself is a sentiment word, e.g. *problem*, it is only regarded if no other sentiment words could be found in its reference window. Then, again we assign the feature-polarity with certainty level 3. In this case the certainty value is 1/3.

Finally, a sentiment value is saved for each feature occurrence. The sentiment value corresponds to the certainty value (1/3, 2/3 or 1) of an analysis comple-

[9]The patterns have previously been used successfully for resolving sentiment references in photo corpora as part of another publication that I co-authored, but that does not fit into the scope of this thesis.

mented with the algebraic sign of the assigned polarity (+ or −). The resulting sentiment value can then be conveyed to the user as part of the visualization of the analysis results. While the straightforward choice of certainty levels is not sufficient to exactly reflect the uncertainty involved in the analysis (see Section 5.2.6), it is a first meaningful step in that direction that brings two advantages: (1) These three levels can be deduced from the analysis and distinguished easily in a visualization. (2) We observed that it is important to sensitize analysts to that the accuracy of an automatic sentiment analysis is not nearly 100%. In addition, analysts are pointed to cases where they should manually assure the correctness of the analysis result if crucial to them. This can be done reading the annotated tooltips, as shown in Figure 5.9.

Overlap-free Plotting of Item-based Time Series

For the visual analysis of feature sentiment developments over time two complementary visualizations are used. In order to provide global overview of the overall data distribution, the existing technique of *pixel map calendars* is used [67]. To track concrete temporal developments of single features, with a focus on time spans with high data frequency, novel *time density plots* are applied. It has to be pointed out that in both visualizations each individual document gets a visual representation. Such a plotting on record-level allows details, like the full text and further data attributes, to be accessed and explored by mouse-over interaction, which is crucial to get a deeper understanding of the data.

Pixel Map Calendars

> "*Each data point is represented by one pixel and displayed in hierarchical bins along x and y dimension. For example, in Figure 5.12 x axis bins correspond to days and y axis bins to years with months, but also any other combination of time units (seconds, minutes, hours, days, weeks, months, years etc.) is possible. Within the bins of the pixel map calendar, pixels (documents) are plotted in temporal order based on their arrival sequence from bottom to top and left to right. There is always enough space to place the documents in the corresponding bins, because the size of each bin is calculated from the maximum number of*

documents in a day as illustrated in Figure 5.12. All bins have equal width in the pixel map calendar. Different bin heights are used for the different months according to the maximum number of documents in a day. As a result empty space is visible in the bins which do not have enough documents to occupy the bin, see Figure 5.12." [144][10]

While temporal distances within bins are no longer visible, this method is very scalable with respect to the amount of data that can be displayed: each document requires one pixel only. This makes pixel calendar maps a very suitable overview visualization and point of entry for further analyses. Feature occurrences can be explored in the context of selectable temporal granularities and in the context of the overall data distribution.

Time Density Plots

The basic idea of the *time density plots* is similar to the event index method [10], described in related work, as it does not use the x axis for conveying exact temporal relations but granting the same amount of space to each event (document containing a certain feature). In this approach, however, we suggest a *time density track* displaying both the temporal order of events on the x axis and the detailed temporal connections among events on the y axis. In the time density track, we omit the, for our purpose, less relevant information about the exact lengths of the time intervals during which no events occur, and focus on areas with a high density of events. These interesting time intervals get much more space than they would with linear time scaling and can easily be analyzed in detail with the overlap-free representation. For each feature one individual *time density plot* is created. The threshold that determines when detailed temporal relations are displayed depends on the average frequency of the respective feature. Thus, data streams and features of very different temporal resolutions and granularities - as they appear in our data - can be readily handled in the same manner. This novel approach can be generalized for application scenarios, where such temporal accumulations of feature occurrences are the main interest. Our basic visualization consists of two main parts that require the same space each, a *sequential sentiment track*

[10]Part of our joint publication partly written by my co-authors.

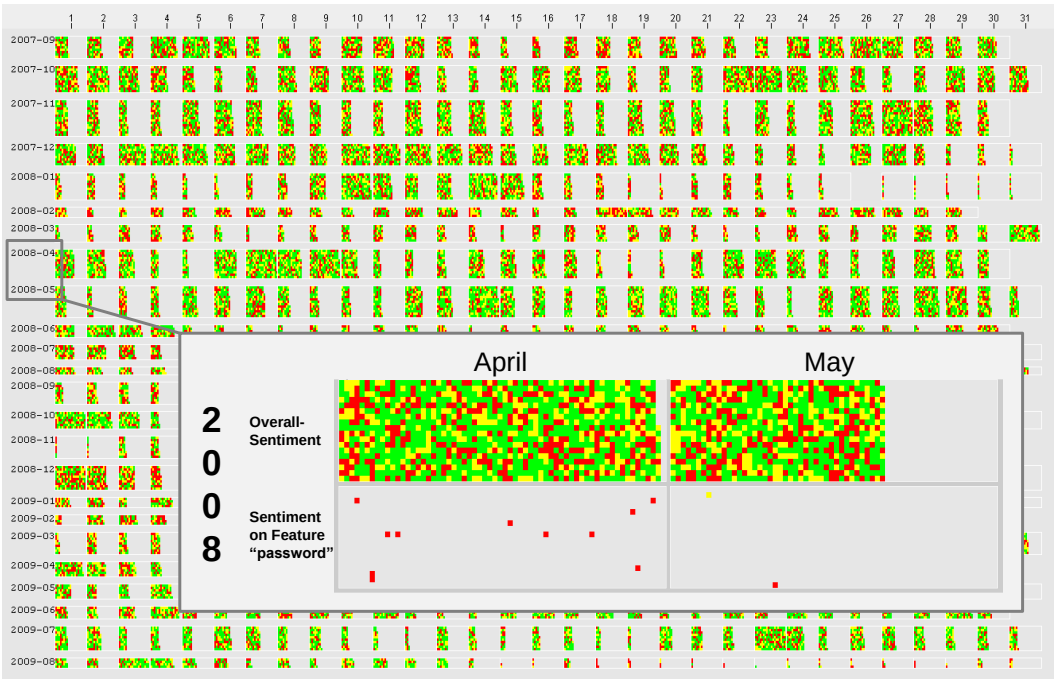
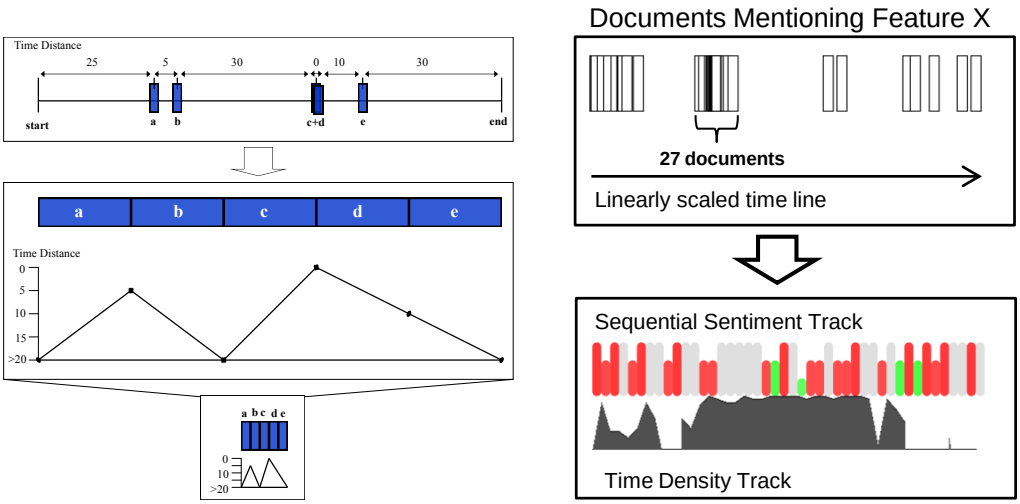


Figure 5.12: *Pixel map calendar*: Each document corresponds to one pixel and the color of the pixel indicates the overall-sentiment of the document, which corresponds to the average of all contained feature sentiments. If the overall sentiment is positive, the pixel is colored in green, if it is neutral, the pixel is colored in yellow and negative sentiments lead to a red pixel coloring. In the background the x axis bins correspond to days and y axis bins to years with months. Additionally, an enlarged view of April and May 2008 is provided, where the x axis bins correspond to months and the y axis bins to years. In this visualization the overall sentiment (top) can be compared to the sentiment on the feature *password* (bottom). It can easily be seen that *password* is a relatively infrequent term that mostly occurs in negative contexts. Reprinted from [144], © 2012 Association for Computing Machinery (ACM).

and a *time density track* either above or below; see Figure 5.9 for an example where it is above. The *sequential sentiment track* contains all occurrences of one feature in sequential order as they appear over time. The exact point in time is not relevant in this upper track, only the temporal order is maintained, so that both space-consuming gaps and over-plotting are inherently avoided. Each rectangular bar encodes one document that contains the feature and indicates by its color the polarity of the feature. The height of a bar depends on the certainty level of the analysis, i.e. the more certain the analysis, the higher the bar. The space that is needed in horizontal direction thus depends linearly on the number of documents in which the feature appears. Figure



(a) Details about the construction of time density curves.

(b) Example for a beneficial application of *time density plots*.

Figure 5.13: In the upper parts, all documents about one specific feature are plotted as they occur over time. The documents are shown in temporal order along a linearly scaled time line, each document being represented by one rectangle. In the upper part of (b) there are dense areas where many rectangles (documents) overlap which does not allow an appropriate analysis. In the lower part of (b) our new approach is shown that overcomes these problems and provides further insight for analysts. Reprinted from [144], © 2012 Association for Computing Machinery (ACM).

5.13(a) provides details and an example on how the *time density track* is created. In the upper box, documents a, b, c, d, and e are plotted as rectangles on a time line. Documents c and d have exactly the same time stamp and are thus plotted one on top of the other. The time distance between each pair of consecutive documents is given in the figure; for simplicity let us assume we have an overall time interval of 100 minutes. Then, a appears after 25 minutes, b after 30 minutes, and so on. As we observe 5 documents within 100 minutes we assume that if they were equally distributed over time, every 20 minutes (= 100/5) we should be able to observe one document - as this is the average time distance between successive documents. Therefore, we define that if the time gap between two documents is larger than the average (here: 20 minutes) there is no noteworthy temporal connection between both. If the time gap is smaller, then the temporal connection is interesting and reaches a maximum level of interestingness at a time gap of 0 (same point in time).

This temporal density is plotted as a curve below the document rectangles as indicated in the second box of Figure 5.13(a). When shrinking the document rectangles to their original size (lowest box of Figure 5.13(a)), we observe that we potentially save a lot of space without losing the relevant information, i.e. c and d occur at the same point in time, a and b are relatively close together, and e is still relatively close to c and d. The only information we lose is how long exactly the gap between b and c is and the gaps at the beginning and ending of the time interval. Yet, for our task this information is irrelevant; it is sufficient to know that in this case there is no interesting temporal proximity between two successive documents. In addition, we are now able to display the documents c and d without over plotting. This is a major advantage since the document contents can easily be explored by mouse-over interaction, and patterns become visible in detail. Figure 5.13(b) provides an example for a beneficial application on real-world data.

If exact global time points or time relations are relevant to users, so-called time density edge bundles can be added to the visualization. These link the document bars of the sequential sentiment track to a linear time line, as in Figure 5.9. Neighboring links attract each other if and only if the corresponding pair of documents is closer in time than the average time gap. The bundling brings an additional visual structure that hints at interesting temporal bursts.

Creation of the Visualization All documents containing a certain feature are extracted and ordered by time. They are plotted in sequential order represented by vertical bars (the *sequential sentiment track*) and colored according to the feature polarity. Positive contexts are encoded in blue (or green), neutral ones in gray, and negative ones in red. The height of a red or blue bar depends on the uncertainty involved in the analysis and corresponds to the certainty value: For certainty level 1 the bar has the default height, for certainty level 2 it has 2/3 of the default height and for certainty level 3 it has 1/3 of the default height. That means the higher the certainty, the higher the bar, the stronger the visual impact. Next, a *time density track* is plotted below or above the *sequential sentiment track*. The height of the time density curve below the border of two successive document bars is determined by the normalized temporal distance between the time stamps of both documents (see

Formula 5.1):

$$\text{time\_density\_height}(f, a, b) = max\left(0, \left(1 - \frac{\text{timedist}\,(a, b)}{\text{avgtimedist}\,(f)}\right)\right) \qquad (5.1)$$

where feature f, preceding document a, succeeding document b

The calculated time density values determine the height of the curve at the position of the border between the two corresponding documents. In the area between two borders, the curve is linearly interpolated. For obtaining the average gap time (avgtimedist) in live data streams, we will use moving or incremental averages, see Chapter 6 for more details.

The bars representing documents within the time density plots are connected by visual links to their positions at the time line. The links are usually not straight, but bent and bundled. Adjacent links will constitute a bundle, if and only if the corresponding documents have a time density value larger than 0. The bundling is computed through an iterative non-deterministic algorithm, detailed in the following pseudo code:

```
Algorithm to do the time density edge bundling for one feature:
Definition
  edge := cubic curve with 4 control points (array cp);
  cp[0] := x, y coordinates of the edge at the time density track;
  cp[3] := x, y coordinates of the edge at the linear time line;
  cp[1] := x of the edge at y = 1/3*(cp[3].y - cp[0].y);
  cp[2] := x of the edge at y = 2/3*(cp[3].y - cp[0].y);
Input
  L_e := List of ordered edges (for feature);
  L_td := Corresponding list of pairwise time density values;
  maxIter := Maximal number of iterations;
  iter := 1;
  while iter < maxIter
     L_i := Random order of all indexes between 0 and (L_e.length - 1)
     for i in L_i
        randomValue := randomly assign 0 or 1;
        if randomValue equals 0
```

```
            performLeftAttraction(i, L_e, L_td, iter, maxIter);
            performRightAttraction(i, L_e, L_td, iter, maxIter);
         endif
         else
            performRightAttraction(i, L_e, L_td, iter, maxIter);
            performLeftAttraction(i, L_e, L_td, iter, maxIter);
         endelse
      endfor
      iter++
   endwhile

where
   performLeftAttraction(i, L_e, L_td, iter, maxIter)
      if i > 0
         e := L_e at i;
         leftE := L_e at (i - 1);
         td := L_td at (i - 1);
         decay := (1 - iter/maxIter);
         e.cp[1].x := e.cp[1].x - decay * td * (e.cp[1].x - leftE.cp[1].x);
         e.cp[2].x := e.cp[2].x - decay * td * (e.cp[2].x - leftE.cp[2].x);
      endif

   performRightAttraction(i, L_e, L_td, iter, maxIter);
      if i < (L_e.length - 1)
         e := L_e at i;
         rightE := L_e at (i + 1);
         td := L_td at i;
         decay := (1 - iter/maxIter);
         e.cp[1].x := e.cp[1].x + decay * td * (rightE.cp[1].x - e.cp[1].x);
         e.cp[2].x := e.cp[2].x + decay * td * (rightE.cp[2].x - e.cp[2].x);
      endif
```

The built-in *decay* function leads to the effect that from iteration to iteration smaller shifts will be realized and the algorithm converges usually quite fast.

Temporal Sentiment Pattern Detection: Searching for Critical Issues

The visualization of temporal developments through the combination of a time density curve, time density edge bundling, and the sentiment polarity coloring draws the eye to interesting time spans. However, if all reasonable features have to be considered, an analyst will still have to skim through the time density plots of several hundred features. Visually scanning all time density plots one by one is exhaustive, time consuming, infeasible on big data or live streams, and error-prone in the sense that details might be overlooked. Consequently, we developed an automatic algorithm to preselect interesting patterns and guide the analyst. This algorithm is based on a scoring function that approximately encodes the criteria an expert would use while skimming through the time lines in search for critical issues. The algorithm automatically detects interesting parts of the visualization and shows them to the user, who can manually verify if they really form a pattern that brings useful new information. In order to discover patterns the algorithm tries to analyze the data according to the questions a data analyst has:

1. Does a set of documents mentioning a certain feature appear accumulated in a relatively short time range?

2. Is this subset dominated by negative sentiments about the feature?

3. Is the feature mentioned in similar and specific contexts, i.e. do people report about the same issue or about different ones?

The first question can be answered by separating documents according to the occurrence of features and by investigating their temporal distribution. In order to detect interesting time patterns within the documents mentioning a specific feature, first *candidate patterns* have to be identified. A candidate pattern is any pattern that corresponds to a relatively large block of documents with high time density. Time-dense means that all time distances between consecutive documents in this block are smaller than the average (for the current feature). Visually this corresponds to a portion of the time density curve that is constantly above zero - without interruption. A block is considered to be large if it is at least twice as long as the average time-density block for the same feature. In addition, as the main goal is to detect real-world issues, blocks that are dominated by negative feedback are of greater interest. Thus,

if a block according to our criteria is both time-dense and large, and if it contains more negative than positive comments, it is inserted into the candidate pattern list and regarded for further analysis. The following algorithm details the detection of such candidate patterns:

```
Algorithm to extract candidate patterns for individual features:
Definition
   pattern = list of tuples (time distance, sentiment value);
Input
   L_t := List of ordered time stamps (for one feature)
   L_s := Corresponding list of sentiment values
          multiplied with certainty values
Derived
   L_d := List of pairwise time distances of succeeding
          time stamps (calculated from L_t);
   d_avg := Average pairwise time distance of succeeding
            time stamps (calculated from L_d);
   L_p := new empty list of patterns;
   p_tmp := pattern, initialized with null;

   for d at k in L_d
      if d < d_avg
         if p_tmp is null
           p_tmp := new empty pattern;
         endif
         add (d, L_s[k]) to p_tmp;
      endif
      else
         if p_tmp is not null
            if isNegative(p_tmp)
               add p_tmp to L_p;
            endif
            p_tmp:= null;
         endif
      endelse
   endfor
   deleteShortPatterns(L_p)
```

```
Output
   L_p

where
- isNegative(pattern p) returns true if the sum of all sentiment
values in pattern p is negative
- deleteShortPatterns(List of patterns L_p) deletes patterns that have
less than 2 times the average pattern length for the same feature.
```

Next, all candidate patterns for a feature F are scored and ranked with respect to their importance for analysts. The score was empirically designed and consists of four factors:

1) DENSITY: The average height of the time density curve for a candidate pattern. The higher the curve is on average, the more densely the documents appear in time. In general, the smaller the relative time distance *D(x)* of a document x to the next document within the pattern P, the higher the density value of P. The time distance is normalized with the average time distance *avg(D(F))* among consecutive documents mentioning feature F. Within a pattern all time distances are necessarily smaller than *avg(D(F))* as this is a criterion for being a candidate pattern.

$$\text{density}(P) = \frac{1}{|\{x \in P\}|} \sum_{x \in P} \left(1 - \frac{D(x)}{avg(D(F))}\right)$$

2) SMOOTHNESS: The time density curves of many interesting patterns have a shape that clearly shows an increase, a plateau, and a subsequent decrease. On the other hand, there are larger patterns of events that are rather loosely connected in time, showing a zigzag pattern. The latter ones are usually less interesting, and this is why we give a higher score to patterns with smoother time density curves. The smaller the average normalized difference of succeeding time distances D among the documents x within the pattern P, the higher is the smoothness value of pattern P. Note that all time distances D necessarily are smaller than the average time distance for the same feature F as this is a criterion for being a candidate pattern.

$$\text{smoothness}(P) = 1 - \frac{1}{(|\{x \in P\}| - 1)} \sum_{i=0}^{i<(|\{x \in P\}|-1)} \left| \frac{D(x_i)}{avg(D(F))} - \frac{D(x_{i+1})}{avg(D(F))} \right|$$

3) SENTIMENT-NEGATIVITY: The higher the amount of documents that mention the feature in a negative context within a candidate pattern, the more likely the pattern might point to critical issues. The higher the certainty of the sentiments, the more confident. Therefore, the sentiments S on the feature in individual documents x are summed up. To get a positive score when having mostly negative sentiments, values are multiplied with -1.

$$\text{sentiment-negativity}(P) = \sum_{x \in P} -S(x)$$

4) CONTEXT-COHERENCE: There may be accumulations of negative comments on a feature that do not necessarily refer to the same issue. However, those candidate patterns are of most interest where all documents apparently report about the same issue, i.e. mention the feature in similar and specific contexts. A simple but effective heuristic was designed to take this context coherence into account. For every potentially content-bearing term (adjectives, nouns, verbs), appearing in at least two documents of a candidate pattern, it was evaluated how strongly this term T is associated with the documents $DOCS$ of a pattern P. To measure the significance of an association the *likelihood ratio test* [51] was used, which operates on a contingency table (see Table 5.1) and has been used before to measure the strength of word collocations [116].

	$DOC \in P$	$DOC \notin P$
$T \in DOC$	A	B
$T \notin DOC$	C	D

Table 5.1: Contingency table showing the number of documents DOC depending on a certain term T and a certain pattern P.

The document counts were used to calculate the likelihood ratio (see Equation 5.2), where A, B, C, and D correspond to the four cells in Table 5.1.

$$\begin{aligned}
&\text{likelihood ratio} = \\
&2 \cdot \left(A \log \left(\frac{A/(A+B)}{(A+C)/(N)} \right) + B \log \left(\frac{B/(A+B)}{(B+D)/N} \right) \right. \\
&\left. +C \log \left(\frac{C/(C+D)}{(A+C)/N} \right) + D \log \left(\frac{D/(C+D)}{(B+D)/N} \right) \right) \\
&\text{with } N = A+B+C+D
\end{aligned} \tag{5.2}$$

Next, the average likelihood value for the terms in a pattern is determined. This value will be higher for patterns that have a number of terms occurring significantly more likely within the documents of the pattern than in the other documents of the corpus. This is the case for patterns with coherent and specific contexts for a feature.

$$\begin{aligned}
&\text{context-coherence}(P) = \\
&\frac{1}{|\{T : |\{D : D \in P \wedge T \in D\}| > 2\}|} \cdot \sum_{T_i \in \{T : |\{D : D \in P \wedge T \in D\}| > 2\}} \text{likelihood ratio}(T_i, P)
\end{aligned}$$

In our empirical tests such a significance-based heuristic yielded much better results, than a standard bag-of-words vector space approach. The latter suffers from several problems. Such a vector space has a very high number of dimensions and each document typically only has entries for a few of such dimensions. The significance-based approach focuses on the relevant information from these vectors in that it evaluates the significance of the dimensions that several documents of the pattern share. As some of the content words are orders of magnitude more frequent than others, the coincidence of infrequent words is in general more meaningful. The context words that make up the significant dimensions are strongly associated to the documents of the pattern. These context words are also displayed when a pattern is highlighted. Often they give a good hint at the semantics of the issue. Examples will be provided in Section 5.2.5 and more details about significance-based association measures in general will be given in Section 5.3.

OVERALL SCORE: All four outlined factors get equal influence in the score of a pattern P, see Formula below:

$$\text{score}(P) = \text{density}(P) \cdot \text{smoothness}(P) \cdot \text{sentiment-neg.}(P) \cdot \text{context-coh.}(P)$$

In a series of experiments on different test data sets - customer web surveys, RSS news, and Twitter data - this score function yielded a very satisfying performance. However, it is possible for the analyst to adapt the weighting among the factors according to the current focus of search. The default score makes it possible to find interesting patterns quickly without requiring any pre-knowledge about the data or deeper insight into the algorithm. The effectiveness of the scoring function is demonstrated in Section 5.2.6 (Evaluation), where also the influences of the individual factors on the overall performance are shown.

5.2.5 Case Studies

One difficulty in the evaluation of an approach that helps to visually detect interesting temporal sentiment patterns in large documents streams, is the lack of appropriate ground truths. For two data sets, however, we were able to get at least some basic ground truth. In the following application case studies we provide empirical evidence and examples for the good performance of our method, by comparing its results with these basic ground truths.

Customer Web Surveys

With the help of the data manager from the company, who had provided the customer web surveys, we were able to construct a ground truth of issues that had occurred in the time span of the data set (about 2 years), see Section 5.2.6 for details. We ran our automatic pattern detection algorithm and extracted the top patterns. Out of the top 25 extracted patterns 15 were meaningful and pointed to relevant time-critical issues. The ground truth contains 17 time-critical issues, so that the recall is 88.24% (cf. Section 5.2.6). The further patterns also pointed to findings that were somehow relevant, but belonged to issues that were general and not so much time-related, so that we did not

count them in as true positives. The general issues popped up on busy sales days, when the overall data volume was higher. The increased overall volume lead to somewhat artificial temporal bursts that were (falsely) detected by our method as temporal issues. Sometimes it is not easy to distinguish between time-related and general problems. For example, the customer service in general is complained about quite frequently. In particular, customers dislike the poor English skills of some service representatives. On busy sales days more people contact the customer service and thus more people complain about this issue. We do not consider this a time-related problem though. However, on busy sales days customers also have to wait longer to get a service representative on the telephone. We consider this to be a time-related issue, because the issue of long waiting times was only present in a clearly defined time range. Further examples for temporal bursts that we do *not* consider as true positives are periodic patterns. For example, the compound noun *christmas present* was among the top-ranked patterns that have been automatically identified, see Figure 5.15. It shows a clearly shaped pattern, however, this is due to the expectable fact that only around Christmas customers complain about issues with *christmas present*s. In the evaluation *christmas present* for us is a false positive, i.e. it refers to a general rather than a time-related problem. The outlined examples already give an intuition that an accuracy of 100% in the detection of time-related issue is not realistic, because it requires human understanding.

In order to provide empirical evidence for the usefulness of our novel methodology and to provide a better understanding on the kind of time-related patterns that can be extracted, we will discuss some of the findings. The data manager stated that he learned several interesting things about his data and to his surprise, during the ground truth construction, discovered issues with the help of the tool that he had not been aware of. This fact shows that the approach supports the discovery of nuggets hidden in the data, that have a high practical relevance. Details are provided as part of the following descriptions and within the captions of the referenced figures. The first automatically detected issue, *packing list*, is about a truly time-related issue, when during a period of about four weeks packing lists *showed wrong total* charges, see Fig-

ure 5.14. The automatically associated words (printed in italics) already give an insightful hint at the concrete nature of the issue. The second issue was about *password* (see Figure 5.9), when the *website* suddenly would *not accept* some customers' *password*s anymore so that they could *not place* their order *online* and instead had to order by *phone*. The issue with rank 7 also becomes clear from the associated terms already: Packages were *held up* in *customs* for severals *day*s due to incorrect *paperwork* and were therefore delivered with *delay*, see Figure 5.16. Finally, the issue ranked at position 13 indicates that customers became *annoyed* when the *identical copy* of an *email* with a *survey* invitation was *sent* to them several *time*s.

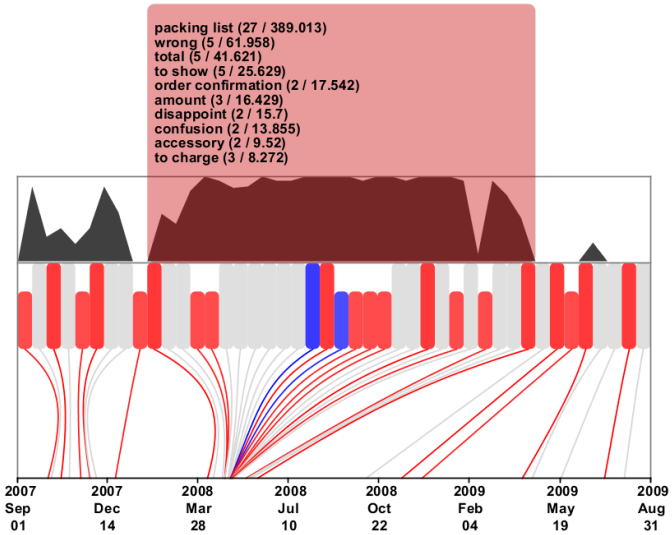


Figure 5.14: The top-ranked automatically detected pattern for the noun *packing list* is a true positive and was new to the domain expert.

Some of the real-world issues were detected with different features. For example, the issue discovered in the time density plot of *packing list* (see Figure 5.14 also appeared in the time density plots of *packing slip*, *charge*, and *slip*.

The issue detected in the time density plot of *survey* (see Figure 5.17) also appears in the time density plot of *email* (see Figure 5.19), but in the latter case the signal was too weak to be detected automatically. Taking a look at the time density plot of *email* it can be seen that there is a short plateau right in the middle of the time density track, but it is not really visually salient. In this case, both the automatic algorithm and the human analyst face similar problems.

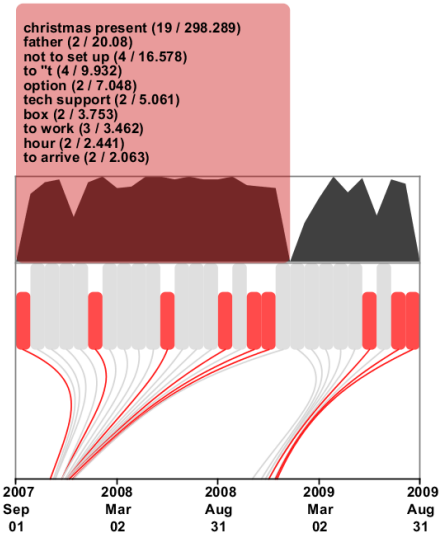


Figure 5.15: Pattern with rank 5 is an example for a false positive: Problems with *christmas presents* are of a general nature even though they only pop up around christmas.

It can be seen that patterns are detected in the feature time series of both frequent features and infrequent features like *packing list*, 44 comments, or *customs*, 26 comments. The two latter ones can otherwise easily remain undiscovered because of their low absolute frequency. Even the top issue *password*, 129 comments, is relatively infrequent considering the overall amount of documents, as can be confirmed for the pixel calendar map in Figure 5.12). The spike is only visible in the time density plot and coincides with a login issue that was not immediately corrected because it was not known at that time. In general, it can be observed that patterns may have varying time spans.

RSS News on Electoral Campaigns

RSS news feed items mentioning the (vice-)presidency candidates and their parties had been collected in the four weeks before the US presidential election 2008. In the pilot study of Section 5.1 three interesting events with negative sentiment connotation had been identified in this time range (see also [177]):

1. Sarah Palin was accused to have abused her power as Alaska's governor firing the state's public safety commissioner (*Troopergate*).

2. A plot of white supremacists to assassinate Barack Obama was uncovered.

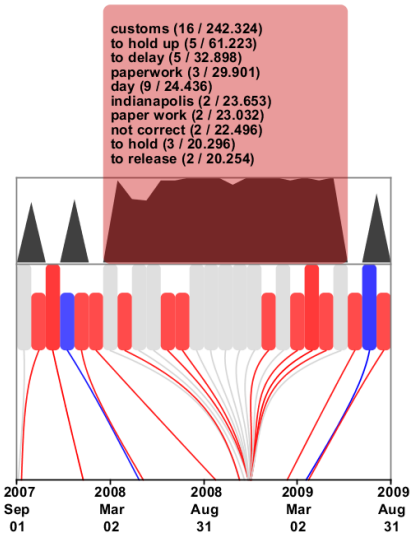


Figure 5.16: The pattern with rank 7 relating to *customs* pointed the domain expert to another relevant issue that he was not aware of before.

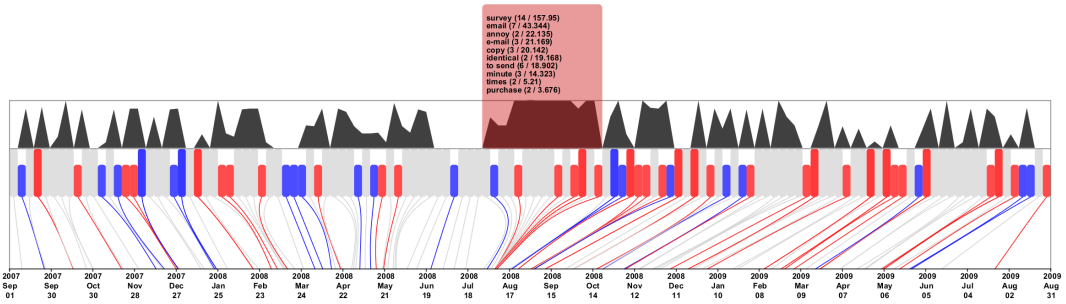


Figure 5.17: An example for an issue that the domain expert knew about before analyzing the data with the visual analytics system (noun *survey*). Still, it was interesting for him to explore it in detail within the global time context.

3. Obama and McCain attacked each other and battled fiercely in a TV debate.

When analyzing the same dataset with the automatic pattern detection, the goal was to detect additional sentiment patterns, that did not become evident in the overview visualization of Section 5.1. Indeed, in addition to the already known negative issues, further relevant findings could be revealed. A pattern pointing to the assassination plot was ranked as second important (see Figure 5.20) and Palin's alleged abuse of power showed up at rank 7 (see Figure 5.21). In both cases it becomes evident that different RSS channels send out the same

news at almost the same point in time. Often these are forwardings of news agency releases.

The previously unknown top patterns detected by our algorithm pointed to a number of further incidents with negative sentiment connotation. For

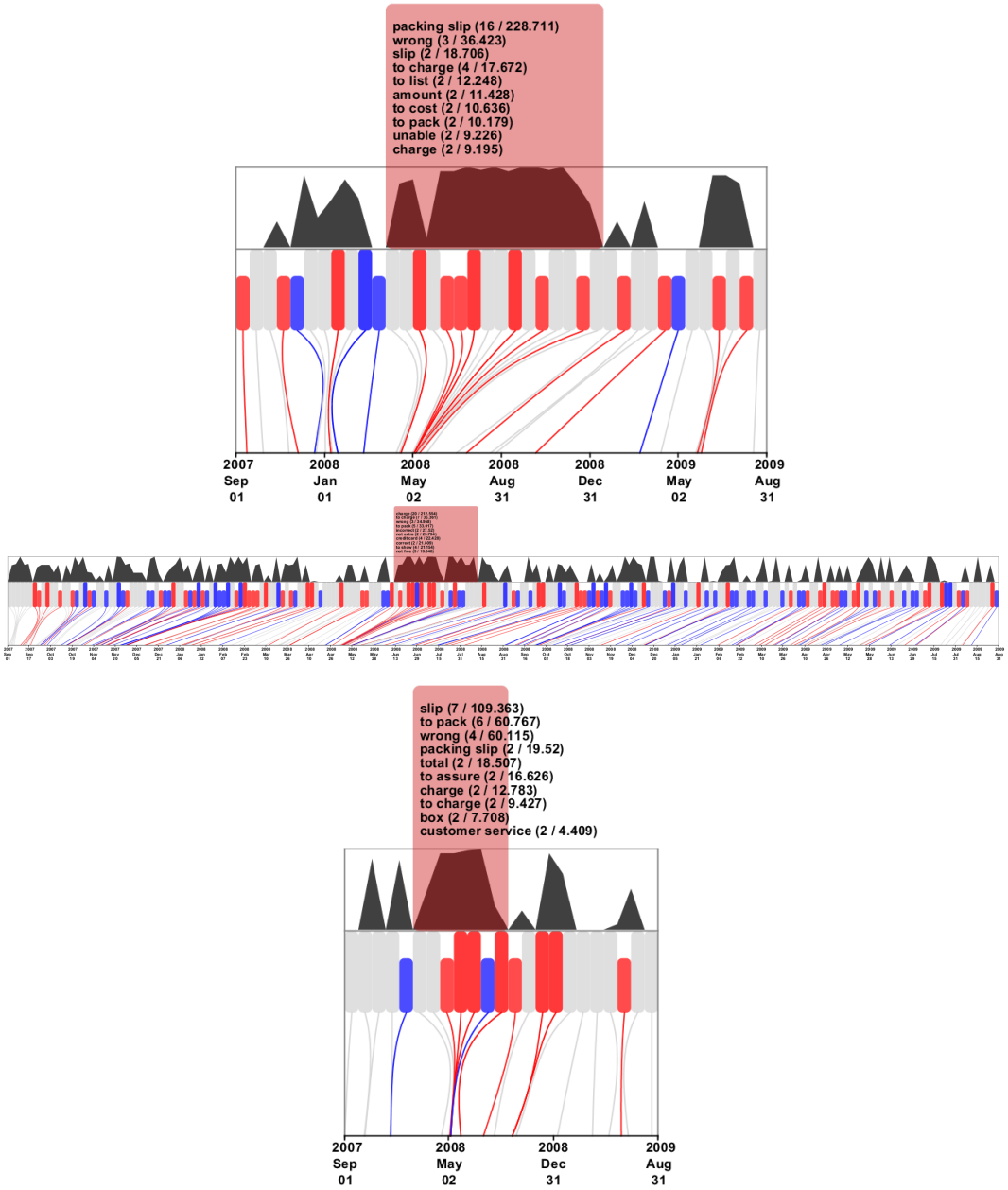


Figure 5.18: Some examples for further automatically detected patterns all referring to the same real-world issue as the top pattern *packing list*, namely *packing slip* (rank 14), *charge* (rank 31), and *slip* (rank 55), ordered from top to bottom.

example, Palin and McCain accused Obama of not being honest about his association with a former war protester (rank 6) and it was revealed that Barack Obama's aunt was living illegally in the US (rank 8). Furthermore, the brother of John McCain had to withdraw from campaign activities after calling 911 to complain about being stuck in traffic (rank 10). Also, a voter registration fraud is reported (rank 11).

However, the TV debate issue did not make it among the top patterns because it was not very negative. As mentioned before, many different sources post very similar news within short time spans, which are only slight variations of news agency messages. This facilitates the discovery of patterns. In such cases, however, the used visualization might not be the optimal choice as heavy bursts in data load may consume a lot of space along the horizontal axis. Here, Krstajic et al.'s CloudLines [94] could be a good alternative.

5.2.6 Evaluation

Apart from the application case studies, further components of our approach are evaluated individually: the automatic pattern detection and the modeling of uncertainty. In addition, an expert user study is provided to give further insight into the real-world applicability and usability of the system. Along the evaluation different features and limitations of our approach are discussed.

Automatic Pattern Detection

For the customer web survey dataset a ground truth of important issues was constructed. The data analyst who provided the dataset and had been working with it during data collection was able to name 9 important issues that he was aware of. In addition, with the help of first prototypical visualizations,

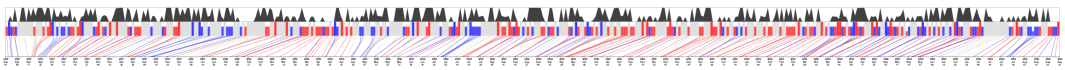


Figure 5.19: Time density plot of *email* containing information about the same issue as in the pattern with rank 13 (*survey*, see Figure 5.17). In this case the automatic algorithm could not detect the issue.

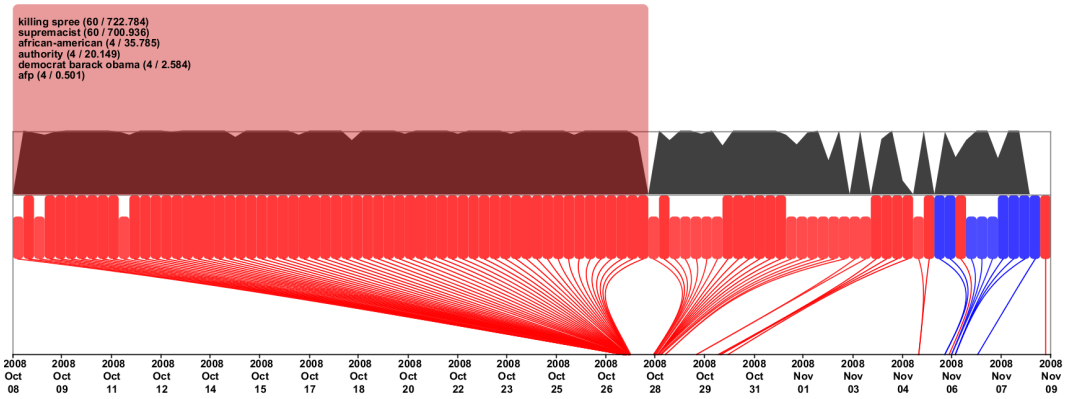


Figure 5.20: The time density plot of the noun *supremacist* (rank 2) points to an assassination plot agains Obama.

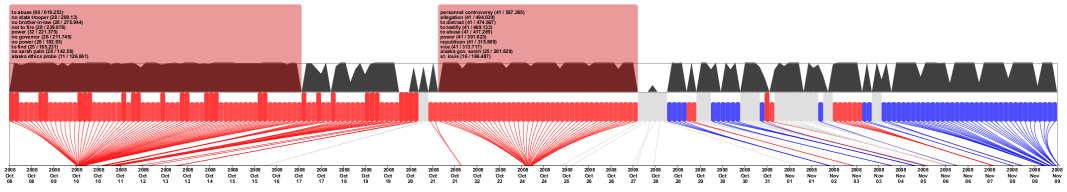


Figure 5.21: The time density plot of the noun *allegation* (rank 7) points to allegations of power abuse against Palin. Another pattern was discovered which shows that the news about the allegation came back as a second burst.

he was able to identify 8 further issues interesting to him. With this ground truth of 17 time-related issues to be found we evaluated the precision and recall of the automatic pattern detection algorithm. The top 25 patterns having the highest overall scores,[11] were extracted and it was verified whether these patterns actually pointed to one of the 17 known issues. The recall-precision diagram for the overall score is displayed in the upper left of Figure 5.22. Taking all 25 results, the precision is 60.00% and the recall 88.24%. In addition, the results for modified scores are provided, performing a sensitivity analysis. For each of the modified scores one factor of the overall score was left out and again a recall-precision diagram was created (Figure 5.22). The goal of this sensitivity analysis was to validate that each factor in the score is actually beneficial, which (except for the smoothness component) is true if all top 25 results are considered, see Table 5.2. However, some of the modified versions

[11] When a pattern pointed to an already detected issue the pattern was ignored in the evaluation. For example, when the pattern of *packing slip* was discovered at rank 14, the issue was already known (pattern *packing list* at rank 1), and therefore it was deleted an the pattern that originally was located at rank 15 got rank 14.

slightly outperform the original version, when retrieving less than 25 results. All in all, the content-based features (*context-coherence*, *sentiment*, and *uncertainty*) seem to be more relevant than the time distribution features (*density* and *smoothness*). It was interesting to see that considering the *uncertainty* has a beneficial influence on the overall score, although it just modifies the *sentiment-negativity*. We could observe that considering the *uncertainty* prevented features, that at the same time were sentiment words (e.g. *problem*, *waste*, *error*), from being weighted too high. The uncertainty modeling causes this kind of feature to get the certainty value 1/3 if no other sentiment words can be found in its surroundings.

Method	**Precision**	**Recall**
Original Score	60.00%	88.24%
Without Sentiment-Negativity	40.00%	58.82%
Without Context-Coherence	52.00%	76.47%
Without Uncertainty	52.00%	76.47%
Without Density	56.00%	82.24%
Without Smoothness	60.00%	88.24%

Table 5.2: Precision and recall values when extracting the top 25 patterns.

Uncertainty Assessment

It is well-known that automatic sentiment analyses cannot be 100% accurate, due to different reasons (ambiguity, implicitness, etc.). To enable analysts to judge how confident analysis results are, the uncertainty involved in the analysis is assessed and visually conveyed. To evaluate the uncertainty modeling, for both the customer web surveys and the RSS news feeds 201 feature mentions were annotated manually. For each dataset the first 3 mentions of the 67 most frequent features were considered. This was done in order not to bias the results, because sentiments appear to be easier to detect for some features than for others.

The resulting values are provided in Table 5.3. The numbers are quite different for both datasets and in general better for the customer web surveys, for which the analysis had been designed in the first place. For these surveys there are no considerable differences in accuracy among the three levels,

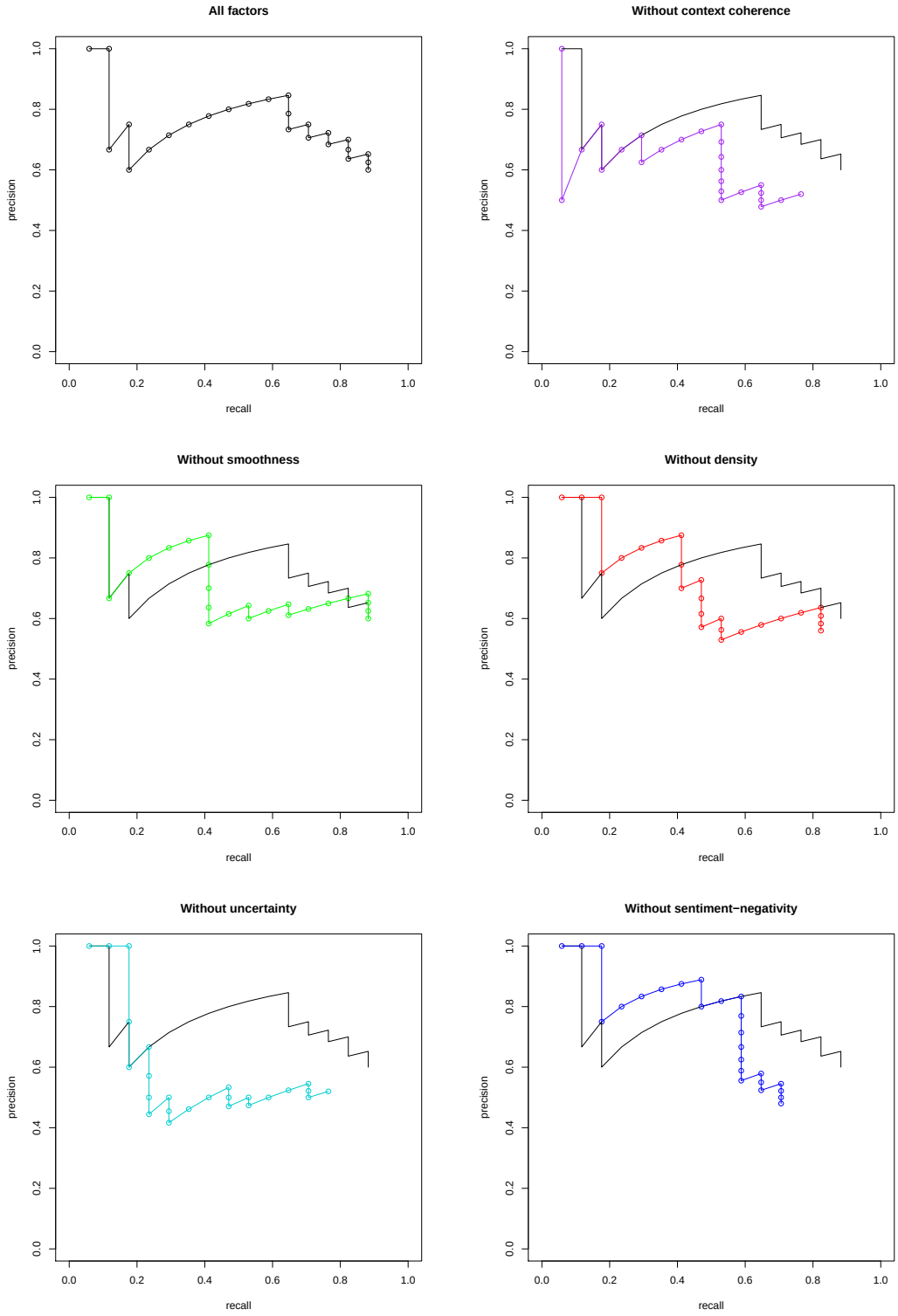


Figure 5.22: Recall-precision diagrams for the score containing all factors (upper left) and each of the modified scores where one factor is excluded. In these diagrams the recall-precision curve for the score with all factors is additionally displayed in black to enable comparison.

Dataset	Level 1	Level 2	Level 3	No Sentiment
Customer Feedback Accuracy	85.7%	84.8%	80.0%	19.7 %
Customer Feedback Proportions	27.9%	39.3%	2.5%	30.3%
RSS Feeds Accuracy	80.0%	52.9%	62.5%	74.6%
RSS Feeds Proportions	14.9%	51.7%	4.0%	29.4%

Table 5.3: Results of the sentiment analysis dependent on the certainty level. All values are rounded to one decimal. The proportions show the fraction of all feature mentions the algorithm assigned with the corresponding certainty level. The accuracy shows which fraction of the feature mentions of one certainty level have been assigned to the correct sentiment category (positive, negative, neutral). For example, 51.7% of all feature mentions within the rss feeds got a sentiment assigned with certainty level 2. For 52.9% of these the sentiment orientation was correct. This is still much better than chance, which would be 33.3% as there are three categories.

which would be a good argument to omit them. However, as shown in the evaluation of the pattern detection (Section 5.2.6), it is still beneficial to consider the uncertainty, because it works better for the special case where words are features and sentiments at the same time. The value in the category "No Sentiment" shows how many of the feature mentions, for which no sentiment could be identified, actually did not have any sentiment. In other words a low accuracy here indicates that quite a number of sentiments have not been detected, which is the case for the customer web surveys. This is quite different for the RSS feeds dataset. Apparently, many of the frequent nouns it contains are not mentioned in relation with a sentiment, which is very different for the customer surveys. In addition, a clear difference between Level 1 and the two other levels is visible. Considering the fact that the algorithm has three options for sentiment labels (positive, negative, neutral), the 52.9% accuracy for Level 2 is still much better than chance but at the same time not quite satisfying. One reason is that in news about electoral campaigns the different political camps are often named in the same sentence, either comparing them or citing representatives of one camp talking about the opponents. In these cases our automatic sentiment analysis has problems attributing sentiments to the correct entity.

Expert User Study

In order to gain a better understanding of the usefulness and usability of the system for target users an expert user study was conducted. We were able to get hold of 7 experts willing to participate. The study was performed based on a previous version of the system including the sequential sentiment track and the time density track, but not the additional features of having a linear time-line with connecting edges. It is to be expected that this additional layer would bring better results. The participants were asked to use the system to identify important time-related issues contained in the customer web survey dataset. To this end, they were first given explanations about the underlying concepts of the visualizations contained in the tool and were given the chance to become familiar with the possibilities for interaction. This introductory phase took about 5 to 10 minutes depending on the questions a participant had about the techniques and the tool.

Then, the participants were given 20 minutes to find as many time-related issues as possible. Moreover, they were asked to speak their thoughts aloud during the whole study, so it would be easier for the observing person to learn more about their strategies, assumptions, and problems when using the tool. For each time-related issue a participant believed to have found s/he should provide further information: The feature that had helped to discover the issue, a short description of the issue, the start and end time, the estimated number of comments read to identify and understand the issue, and finally the confidence in this analysis. The purpose of asking for a description of the issue as well as the start and end was to assure that the participant actually had gained an understanding of the issues s/he discovered and was able to identify time ranges. The further information was collected to evaluate the reading efforts required and the confidence participants had gained in their analysis. Participants were also given the option to state that they had discovered an issue that did not appear to be time-related but rather general.

Finally, the participants were asked whether they considered the tool as being useful and if so what they liked the most. Further questions were what they disliked about it and what suggestions for improvement they had. Also, they were asked whether they ran out of time and thought they could have discovered more issues easily or if they had a hard time discovering further

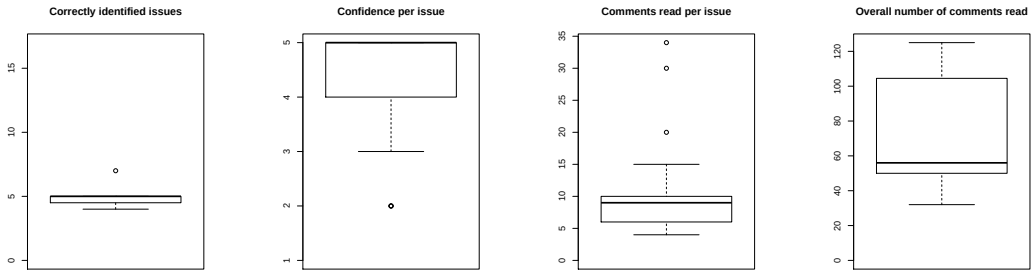


Figure 5.23: Numerical results of the study: (a) time-related issues detected correctly within 20 minutes (per person), (b) confidence in discovering time-related issues (per issue) on a scale ranging from 1 (not very confident) to 5 (absolutely confident), (c) estimated amount of comments read (per discovered time-related issue), (d) overall amount of comments read by each participant including comments that did not lead to the discovery of time-related issues. Reprinted from [144], © 2012 Association for Computing Machinery (ACM).

issues.

Results & Discussion

	Truly time-related	Truly general	Truly nothing
Time-related (participant)	35	0	3
General (participant)	2	10	0
Nothing (participant)	1	1	7

Table 5.4: The aggregated results for all participants. The choices of the participants (rows) in comparison with the ground truth (columns), only for those features participants actually had a chance to look at before running out of time.

First, we want to look at the numerical performance shown in Figure 5.23. On average participants were able to discover 5 time-related issues within 20 minutes and had to read 9.84 comments to verify them. They were pretty confident in the analysis with an average of 4.45 (with 5 being the maximum). Only 3 (7.9%) false time-related issues were identified, aggregating results for all participants. The 3 false hits correspond exactly to those that were identified with a confidence of 3 or lower. The overall amount of comments read (Figure 5.23 d) varies considerably. It could be observed that some participants read comments only superficially focusing on the sentences containing

the feature, while others read comments completely. The 74.57 comments participants read in average during the study only correspond to about 0.15% percent of the overall data (50,000 comments). Next, Table 5.4 shows the aggregated performances of all participants for all features which they managed to investigate within the 20 minutes. The overall precision is 88.14%, the precision for time-related issues 92.11%. The 5 time-related issues participants were able to detect in average correspond to a recall of 29.41%. However, the trade-off introduced by the fixed time constraint is decisive here. All 7 participants were confident that they would have discovered more issues when given more time. At the same time, one participant commented that if it was critical in a real-world analysis he would have read every single comment of an interesting pattern, which he did not during the study. All in all, the numerical evaluation of the performance proved that the system is actually usable and suitable for the outlined task of identifying time-related issues in large sets of documents minimizing the reading efforts for analysts.

More interesting than the numerical performance, however, were insights gained when observing the experts using the tool. A basic strategy all participants shared was starting to take a closer look at the patterns detected by the automated analysis. Yet, to our surprise participants applied quite different individual strategies to deal with these patterns. One participant first had a look at different interesting time density plots without exploring them in detail, in order to gain a better feeling for the visualization. In general, users got faster and more accurate the longer they worked with the tool. When running out of time one user commented that he probably had made an error at the beginning and had just realized this after researching further time density plots - which was actually true. When it came to exploring potentially interesting patterns, one user preferred reading negative comments with a high level of certainty while another user first read the comments at the beginning and end of the highlighted block to see whether they were similar. Several users tended to ignore comments with neutral rating and rather read the other comments. A further user preferred to first read the associated terms of a highlighted block and only then started reading the individual comments. He was the one that read the least number of comments to identify issues. While 4 participants

focused mainly on reading comments within highlighted patterns the other 3 also read a considerable number of comments not contained in patterns to see whether the reported issue might have persisted in time but just with a lower frequency. In general, it could be observed that the more homogeneous in content the comments within a pattern were, the quicker an issue was detected. This was more often the case when dealing with generally infrequent features than with frequent ones.

The answers to the survey questions at the end of the study revealed further details. All participants found the tool useful and when asked what they liked most made the following responses (several things could be named by one participant): 6 mentioned the automatic scoring and highlighting of interesting patterns, 3 mentioned the quick access to the documents via mouse-over and 2 named the combination of sentiment and time density track. Not liked about the tool was that when zooming out too far, objects could disappear from the screen. This was mentioned by 4 participants and could be fixed in the meantime. Another complaint was about the insufficient description of the interface and one user would have liked an autosize functionality to see a whole time density plot on the screen without zooming by himself. Further suggestions for improvement were given. One participant would have liked to be able to search and highlight keywords, respectively highlight the comments in which these keywords appeared. She wanted to see whether certain keywords appeared more frequently within the comments of a pattern than outside. Another participant would have liked to be able to combine two features to one time density plot, containing only documents where both selected features appeared.

As mentioned before, the target user group for our approach and also within the user study were experts. Of course, experts in data analysis are a very experienced user group, so the results cannot simply be generalized to common computer users. The similar level of experience of the participants may also have contributed to the relative homogeneity in performance, though their strategies were quite diverse. In conclusion, the basic concept was well-received, participants gave a mostly positive informal feedback and were able to identify time-related issues without major reading efforts. Apart from fixing minor interaction problems further methods for filtering and selection should

be integrated in future versions.

5.2.7 Discussion and Conclusion

In this section, we addressed a feature-based sentiment analysis process that tightly couples methods from automatic text processing, visual analytics, and information visualization. Our intelligent visual interface for several reasons is especially beneficial to explore sentiments over time. First, the uncertainty involved in the automatic analysis due to the complex and ambiguous nature of language requires further exploration: The textual content of customer comments can easily be accessed by mouse-over interaction for detailed insight and to derive meaning. Second, the fact that it is not quite clear beforehand what kind of interesting temporal patterns might be included in the data demands a broad feature coverage and visual exploration. To quickly analyze vast volumes of text documents over time, we suggest using *pixel map calendars* for general overview and to introduce *time density plots* for detailed insights. Time density plots combine both a visualization of a sequence on record level and a linear time-line in one view. Features of interest can be selected and visually explored for salient patterns, combining both visualizations. Furthermore, interesting interval patterns are automatically detected and scored. The score measures the importance of an interval based on the average height and smoothness of the time density curve, the sentiment negativity and uncertainty, and the context coherence. In many of the automatic processing steps, required to detect and rank issues, only a limited level of accuracy can be achieved. This is mostly due to the complex and ambiguous nature of language, which is very challenging for automated data analysis. In some cases, it makes sense to visualize the uncertainty involved in order to indicate potential inaccuracies, as done in the sentiment analysis case. In the overall score of patterns we partly overcome the inaccuracy issue by combining different score components. The broader and more diverse the components of the score are, the less prone it is to random noise. In our case this worked quite well. In the end, analysts still have to verify that the detected patterns are meaningful. Yet, the automatic detection guides their analysis, heavily reduces their workload, and points them to findings they might otherwise hardly make. Our new solutions were tailored for direct customer feedback through

web surveys and further tested on RSS news feeds. Previously unknown issues were discovered and features of interest could be explored in detail. In an extensive evaluation we were able to demonstrate the applicability, usability, and good performance of the approach, learned about the impacts of different parameters and discussed limitations.

5.3 Term Associations

This section builds on parts of the following publication:

Ming C. Hao, Christian Rohrdantz, Halldór Janetzko, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Meichun Hsu. Integrating Sentiment Analysis and Term Associations with Geo-Temporal Visualizations on Customer Feedback Streams. SPIE 2012 Conference on Visualization and Data Analysis (VDA 2012), 2012. [70][12]

In addition, the ideas that I have developed as part of the research presented in this section have contributed to the filing and publication of the following patent applications:

United States Patent Application US/2013/0054597: Constructing an Association Data Structure to Visualize Association among Co-occurring Terms. Filing Date 23.08.2011. Publication Date 28.02.2013. Inventors: Ming C. Hao, Umeshwar Dayal, Christian Rohrdantz, and Lars-Erik Haug [65].

World Intellectual Property Organization Patent Application WO/2012/167399: Sentiment Trend Visualization Relating to an Event Occurring in a Particular Geographic Region. Filing Date 08.06.2011. Publication Date 13.12.2012. Inventors: Ming C. Hao, Umeshwar Dayal, Baoyao Zhou, Cheng Chang, Meichun Hsu, Mohamed E. Dekhil, Riddhiman Ghosh, and Christian Rohrdantz [66].

[12] In this publication I was only responsible for the part on term associations. The corresponding research, programming, and writing was done by me. The other parts of the paper were done by the other authors and will not be used in this thesis.

So far sentiments about single target words were identified and unexpected temporal bursts were searched for in order to detect real-world issues and events. However, there may be further indicators like unexpected term co-occurrences, so-called term-term associations, or unexpected geo-spatial patterns relating to text documents, given that geo-coordinates can be attributed to text documents. In the previous section term-term associations indeed already did play a role as part of the *context-coherence* calculation. This score is calculated for all documents of a time pattern relating to a certain target word. If further terms are highly associated with these documents, the pattern gets a higher *context coherence* score. Not only unexpected bursts in the occurrence of individual terms are possible event indicators, but also unexpected bursts of the co-occurrence of two (or more) terms. In the latter case it could even be possible that the overall frequency of each of the co-occurring terms would remain the same and only the unexpectedness of the co-occurrence would point to a shift in content which might indicate an issue or event.

In this section I briefly want to show what kind of knowledge can be gained by exploring term-term and also term-geo-location associations. As part of an initial study I will explore the usefulness of different methods for measuring association strength and showcase potential applications. I will limit this first analysis to static datasets and omit the time dimension. Performing a time-dependent association analysis would require further research on how to identify useful time intervals on which to base the association exploration, which is not within the scope of this thesis. However, for static datasets the association exploration may also be quite useful as it can provide a broader view on the important terms and features bringing them into relation. To this end, information about which terms are associated has to be automatically extracted from the text resources and visually conveyed to the analysts to enable them to gain a better understanding of the data.

The most interesting term-term associations are those between two different target words and those between a target word and a potential sentiment word (adjective or verb). Hence, I want to determine which entity of interest is associated with which other entity of interest and which sentiment words are

associated with an entity of interest. Later, the aim is to go beyond binary associations (involving two words) and find n-ary associations (involving $n > 2$ words). For analysis I suggest to cluster these n-ary associations in the context of a Self-Organizing-Map [93] in order to provide a structured overview.

5.3.1 Background

The association strength of two terms can be measured based on their sentence-wise co-occurrence. From an analytic point of view this task is closely related to frequent item set mining. However, the typical support and confidence approach [160] is not very useful in the case of natural language, because term frequencies in text follow a long tail distribution as covered by Zipf's Law [190]. Some words are orders of magnitude more frequent than others and thus would be contained in many associations. Yet, highly frequent words usually carry less meaning than those with a moderate frequency and are therefore not very valuable to explore. Brin et al. [22] consequently suggest relying on statistical measures for cases such as text data. Manning & Schütze [116] apply different statistical association measures to assess term collocations: The hypothesis tests *t test* and *likelihood ratio* as well as *pointwise mutual information (pmi)*. For the sake of brevity I refer interested readers to the referenced book for details about these methods. The assumption behind the hypothesis tests is the null hypothesis that two items are independent. If this hypothesis can be rejected with a high level of confidence, the items can be considered to be associated (not independent). The more data that has been seen, showing evidence that supports the rejection of the null hypothesis, the higher the level of confidence.

5.3.2 Mining Term Associations: Novel Methods and Comparative Evaluation

In order to apply hypothesis testing for detecting term associations we first have to define the probabilities we work with. The Probability *P(a)* that a term a occurs in a sentence s of the corpus C is defined as:

$$P(a) = \frac{|\{s{:}s \in C \wedge a \in s\}|}{|\{s{:}s \in C\}|}$$

Assuming the independence of two terms a and b, the expected probability *P(a,b)* that both terms occur jointly in a sentence s of the corpus C is defined as:

$$P(a,b) = P(a) \cdot P(b)$$

The observed probability *P(a,b)* that both term a and term b occur jointly in a sentence s of the corpus C is defined as:

$$P(a,b) = \frac{|\{s{:}s \in C \wedge a \in s \wedge b \in s\}|}{|\{s{:}s \in C\}|}$$

Based on this, the mentioned methods are applied to find the top binary associations, i.e., pairs of terms that are highly associated on a sentence basis. The performance of the different methods will be compared and discussed. For our analyses we included only terms that we consider to be content-bearing, namely nouns, compound nouns, adjectives, and verbs. As mentioned before, one goal of extracting associations is to present them to a user with the intent of providing a more detailed insight into the results of the sentiment analysis. When extracting the top binary associations sometimes groups of associations show up that apparently belong together. For example, the top 100 associations for the web surveys included {website, easy}, {website, to navigate}, and {easy, to navigate}. Evidently, these associations belong to the same statement and should be merged. To this end we perform a merging of binary associations to triples and then iteratively to sets of more than 3 terms until no further merging is possible. The fundamental strategy was inspired by Agrawal and Srikant's apriori algorithm [1]. We found that the *pmi* is the only measure that can be extended in a straightforward manner to measure the association among more than two terms at a time. We calculate the *pmi* for n terms as:

$$I(a,b,...,n) = log_2 \left(\frac{p(a,b,...,n)}{p(a) \cdot p(b) \cdot ... \cdot p(n)} \right)$$

The prerequisite for getting an association containing a set of n terms is that all n distinct subsets containing *n-1* terms each, are also considered to be associations. To give an example, an association $\{a,b,c\}$ may exist if and only if $\{a,b\}$, $\{a,c\}$, and $\{b,c\}$ are considered to be associations. In addition,

the following two requirements have to be fulfilled:

1. $I(a,b,c) > max(I(a,b), I(a,c), I(b,c))$
2. $count(a,b,c) > lowerbound$

Where count(a,b,c) denotes the number of sentences in the corpus that have to contain the three items jointly. This number has to lie above a certain user-defined threshold we name *lowerbound*. This threshold is necessary to prevent getting associations that are underrepresented. We denote this merging step as *pmi merging*.

Sometimes, however, the use of synonyms prevents sets from getting merged. For example, in the web survey dataset we get the associations {website, easy, to navigate} and {website, easy, to use}. Basically, both associations address the same statement just with slightly alternating expressions; some people say it is easy to use the website and some say it is easy to navigate. To cope with such usage of synonyms, associations containing more than 3 terms and sharing at least 50% of their terms are merged as well. The threshold of 50% yielded good results in our tests, but the analyst could as well easily adapt this parameter. The two associations {website, easy, to navigate} and {website, easy, to use} share 2/3 of their terms and therefore result in the association {website, easy, to navigate, to use} which integrates the redundant information. We denote this step as *overlap merging*. To see whether both kinds of merging strategies for associations are beneficial to the analysis, we tested them for our data in a comparative evaluation provided later in this section.
After generating the associations, a sentiment value for each association is calculated. The process is slightly different for associations generated with *pmi merging* in comparison to associations generated with *overlap merging*. For an association generated with *pmi merging* necessarily a considerable number of sentences in the corpus exists ($>$*lowerbound*) that contain all terms of the association. For each of these sentences we sum up the sentiment values of all sentiment words contained in the sentence. A positive word contributes +1 and a negative word contributes -1 to the sum. The average sentiment value of all sentences is considered to be the sentiment of the association. For associations generated with *overlap merging* there might not exist a single sentence

t test	*likelihood ratio*	*phi*	*pmi*
free, shipping (1741)	free, shipping (1741)	mouth, taste (8)	mouth, taste (8)
great, service (1929)	day, next (995)	not friendly, not to user (21)	74xl, 75xl (7)
excellent, service (1225)	order, to place (761)	club, sam (21)	bang, buck (6)
day, next (995)	great, service (1929)	expectation, to exceed (51)	office home, student (6)
order, to place (761)	excellent, service (1225)	creative, kit (12)	god, to bless (8)
to keep, work (480)	to keep, work (480)	74xl, 75xl (7)	aol, yahoo (6)
good, work (494)	day delivery, next (313)	manner, timely (82)	creative, kit (12)
fast, service (599)	hour, phone (416)	free, shipping (1741)	not friendly, not to user (21)
free, next (523)	good, work (494)	bang, buck (6)	bait, switch (6)
hour, phone (416)	hour, to spend (268)	office home, student (6)	citizen, senior (10)

Table 5.5: The top 10 binary associations for the web surveys generated with each measure. Only pairs of terms co-occurring in at least 6 sentences were considered. The number of co-occurrences of each term pair is put into parentheses.

containing all terms. Such an association is the composition of m overlapping associations generated with *pmi merging*. All sentences that contain at least one of the m overlapping associations are taken into account. The average sentiment value of these sentences is considered to be the sentiment of the association. One advantage of taking average values is that these are more robust to noise. Even if single sentiment assignments are wrong, it can be expected that the average still correctly reflects the basic sentiment tendency.

Comparative Evaluation

The different methods applied to generate and merge term associations are compared and evaluated in the following.

Evaluation of Association Measures It was not quite clear which of the outlined term association methods would perform best on real world data. Consequently, we applied and evaluated them. In addition to the *t test*, *likelihood ratio* test, and *pmi*, we also applied a correlation coefficient (*phi*). In order to arrive at meaningful results we tested the methods on real data from web surveys. The same dataset as in Section 5.2 was used, consisting of app. 50,000 responses to a customer web survey containing 96,987 sentences; the results are shown in Table 5.5.

As can be seen, the two hypotheses tests tend to prefer rather frequent associations, whereas the two other measures tend to find more infrequent associations, which are often less general. In order to gain further insight we examined the frequency distribution among the top 100 associations. Figure 5.24 shows the distribution for the web surveys. The *pmi* and *phi* prefer

associations with a rather low overall frequency. Therefore, we regard both measures as not very suitable for our task. The *t test*, in contrast, tends to prefer associations with a very high frequency. The *likelihood ratio* test is the only measure that covers almost the whole frequency spectrum. In a more detailed manual analysis, we found that the *likelihood ratio* test is the best choice for our approach, as highly frequent associations are more interesting in the general case, although there are still many rather infrequent associations that could point us to interesting findings. After drawing the conclusion that the likelihood ratio test would be the most appropriate measure, we found a study that reinforces this finding: "For text analysis and similar problems, the use of likelihood ratios leads to very much improved statistical results. The practical effect of this improvement is that statistical textual analysis can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions, and it allows comparisons to be made between the significance of the occurrences of both rare and common phenomenon." [51].

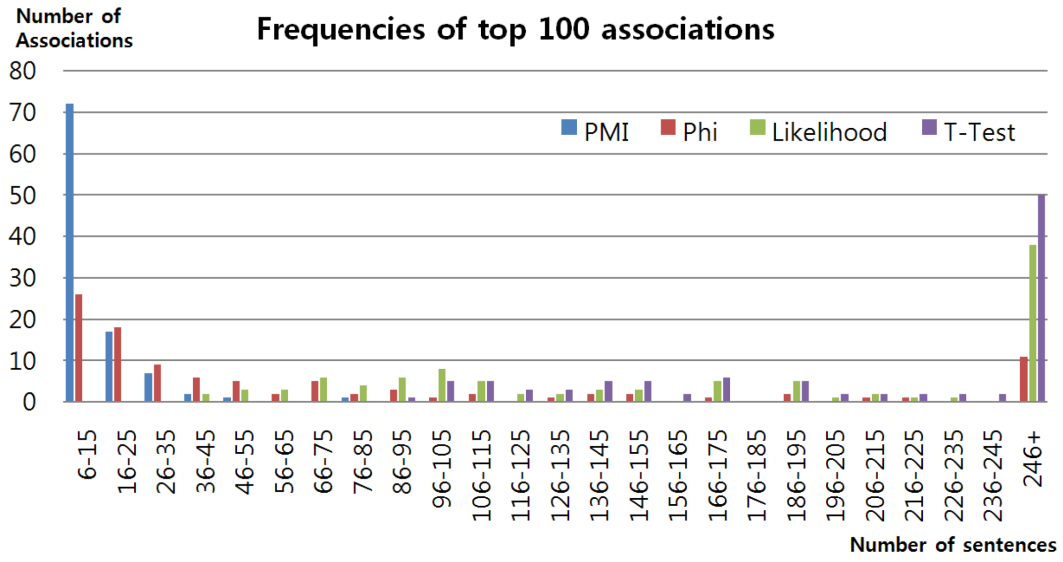


Figure 5.24: The frequency distribution of the top 100 associations extracted with each measure for the web surveys. The y-axis shows the frequency of associations in the corpus, i.e., in how many sentences associations occur. The x-axis reveals how many of the top 100 associations had a certain frequency. Reprinted from [70] © 2012 SPIE.

Web Surveys: Top 10 associations after *pmi merging*	Web Surveys: Top 10 associations after *pmi merging* and *overlap merging*
door, front, to leave (27)	free, overnight, shipping, price, delivery, fast, to love, appreciate
good, to keep up, work (55)	english, someone, to speak, people
hour, phone, to spend (154)	good, to keep up, work, to keep
english, someone, to speak (30)	easy, to navigate, website, to use
address, to deliver, wrong (63)	day shipping, free, next, day delivery
english, people, to speak (39)	address, to deliver, wrong, fedex
easy, to navigate, website (57)	day, next, to receive, to order
good, to keep, work (385)	hour, phone, to spend
courteous, helpful, knowledgeable (9)	door, front, to leave
day ship, free, next (153)	fedex, package, to leave

Table 5.6: The top 10 binary associations for the web surveys generated with each measure. Only pairs of terms co-occurring in at least 6 sentences were considered. The number of co-occurrences of each term pair is put into parentheses.

Evaluation of n-ary Associations To evaluate the performance of the suggested merging steps, we applied them to our data. First, *pmi merging* was performed and then additional merges were obtained through *overlap merging*. The results are shown in Table 5.6.

The n-ary associations are very useful. Often they can readily be interpreted as a statement, e.g. {easy, website, to use} indicates that the website is easy to use. Also, the *overlap merging* produces good results. For example, {good, to keep up, work} and {good, to keep, work} were merged into one association {good, to keep up, work, to keep}. In some cases our preprocessing algorithms were not able to find the particle *up* and relate it to *keep*. At least as part of the post processing this problem is now partly solved by merging terms together in the term association step.

5.3.3 A Self-Organizing Map for the Exploration of Term Associations

Often, one association represents one particular problem; for example, the association {address, to deliver, wrong, fedex} summarizes the complaints of customers that FedEx delivered their order to the wrong address. In many cases, such an interpretation of associations is quite obvious. However, in some cases it is still valuable for the analyst to have quick access to the sentences or whole documents that contain an association to understand or verify the meaning. Therefore, we provide information about the associations in an interactive visual interface. Instead of simply listing associations we want to enrich them with further information. We color each association with its

sentiment value, i.e. the average sentiment of sentences containing the association. Positive sentiments are mapped to green and negative sentiments to red; the color saturation indicates the sentiment value. Furthermore, we cluster associations according to the reviews to which they belong. While the associations can be interpreted as statements extracted from sentences, the association clusters can be interpreted as groups of statements often made within the same reviews. For the clustering a distance measure between two associations has to be defined. To do so, we create a high-dimensional vector for each association that has as many dimensions as there are documents in the data set. If an association is contained in a specific document, the entry in the respective dimension will be 1; otherwise, it will be 0. To calculate the distance between two associations we take the Euclidean distance between their vectors.

Instead of computing separate clusters of associations, we also want to reflect how the clusters relate. Therefore, we generate a self-organizing term association map (SOM) with a simple square topology. In order not to overwhelm the analyst, the map is limited to 16 clusters, but it can easily be extended to cover more clusters or show a different topology. More details about the visual representation are given in Section 5.3.4.

5.3.4 Case Studies

In order to explore whether our visual analysis interface reveals valuable additional information to analysts, we generated the corresponding SOM, see Figure 5.25. The SOM-nodes are displayed as rectangles and connected with lines. The thicker the line between two SOM-nodes, the closer the respective cluster centroids are in the underlying high-dimensional vector space. Based on this visualization we performed a use case analysis with the web surveys. At first sight, it becomes evident that there are many more positive comments than negative ones. With respect to negative associations two clusters are dominant. The cluster on the top right deals with problems regarding the language skills of the customer support teams, who some customers find difficult to understand. We also see that some customers complain that they have to spend hours on the phone in order to solve problems with the customer service. The other cluster (further down to the left) deals with the delivery of ordered

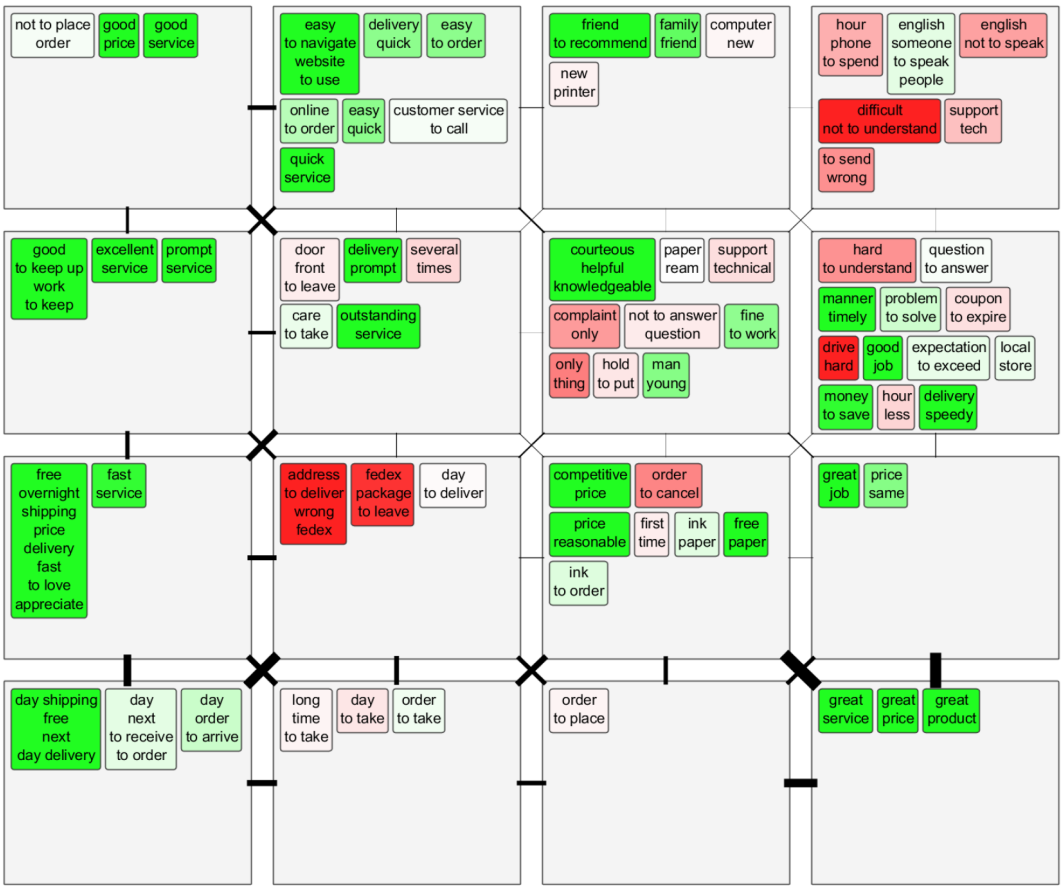


Figure 5.25: A clustering of all associations (small boxes) for the buyer's web surveys into a self-organizing map with 16 nodes (large boxes). Reprinted from [70] © 2012 SPIE.

items. Some people complain about FedEx delivering their order to the wrong address and some complain about FedEx leaving the package in front of the door. When the user moves the mouse over an association, the respective sentences from the underlying text data are displayed in a tooltip. This interaction reveals the problems that people have with FedEx leaving a package. Some are afraid that it could have been stolen and others complain about rain damage. The dominant positive feedback is easily analyzed. First of all, people love the good service and especially the free shipping and one-day delivery. Also, the competitive prices and free paper are appreciated, as well as the fact that the website is quick and easy to use for ordering online. Finally, people state that they would recommend the company to their family and friends. In comparison to standard "word cloud" visualizations, the additional structure provided by the term associations gives more insights by enriching words with semantic context information. However, the SOM visualization also reveals some limitations of the overall approach. When the real number of clusters in the data is larger than the number of SOM nodes, some SOM nodes necessarily show a mixture of several topics. In addition, preprocessing errors may also be revealed. For example, when hovering over the association {hard, drive} it can be seen that people do not have "hard times with their drives" as the strongly negative sentiment would suggest. They are simply making a comment about their *hard drive*, which is neither negative nor positive. The misleading representation is due to the fact that the preprocessing algorithm failed to detect *hard drive* as a compound noun and interpreted *hard* to be a sentiment referring to *drive*. A further interesting peculiarity is that the cluster on the upper left contains the associations {good, price} and {good, service}, and the cluster on the lower right contains {great, price} and {great, service}. These statements appear to be very similar, but are not clustered together. This is due to the fact that the persons who wrote *good price* never wrote *great price* in the same document and vice versa. Right now, two associations are clustered together if they appear together. To collapse the distinction between *good price* and *great price*, we have to cluster associations that appear in similar contexts but not necessarily within the same documents. In conclusion, this first visualization does not only provide us with a detailed insight into the document collection, but also points to potentials for improvement of the

underlying algorithmic processing. Both kinds of insights are key features of visual analytics.

5.3.5 Discussion and Conclusion

This initial study on the application potential of term associations in business-oriented text analysis leads to different conclusions. First, insights that can be gained from an analysis of term associations can be quite beneficial for the analysis. Second, there is a strong indication that the *likelihood ratio* is the most useful among all tested association measures. Finally, a self-organizing map can be a good point of entry for in-depth analysis. We also tested other projection methods like multi-dimensional scaling, but the high amount of clutter on the one hand, and a lack of clear groupings on the other hand, led to rather useless results. When applying the SOM approach to new datasets, the user may have to adapt the SOM topology and the *lowerbound* threshold, but other than that the method is generic. The SOM can be generated in a fully automatic way from the raw text data. The results on the static dataset are promising and it would certainly be interesting to explore changes in term associations over time. However, this is a challenging topic for future research that goes beyond the scope of this thesis.

The research presented in this chapter has pushed the state-of-the-art in time-oriented text mining and visualization. A new set of methods and algorithms was suggested to detect and explore temporal patterns in text content which are highly relevant for business analysts. In addition, it was shown that the exploration of term associations over time could provide a further analytical added value. Another relevant related approach that I contributed to is the extraction of geo-term-associations using a methodology similar to the presented one. Each customer is located in a certain geographic area and local complaint clusters did also lead to interesting findings. If only customers in certain areas complain about certain issues, this is also highly relevant to business analysts. The first research results in this direction are quite promising. For more details the interested reader is referred to the original publication:

Ming C. Hao, Christian Rohrdantz, Halldór Janetzko, Daniel A. Keim,

Umeshwar Dayal, Lars-Erik Haug, Meichun Hsu, and Florian Stoffel. Visual Sentiment Analytics of Customer Feedback Streams Using Geo-Temporal Term Associations. Information Visualization 12(3-4): 273-290, 2013.

Another interesting line of research that could potentially profit from the presented approach is the discovery of clusters coinciding in text content, time, and space. Here, I contributed to another publication:

Andreas Weiler, Marc H. Scholl, Franz Wanner, and Christian Rohrdantz. Event Identification for Local Areas Using Social Media Streaming Data. In Kristen LeFevre, Ashwin Machanavajjhala, and Adam Silberstein, editors, Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks, DBSocial 2013: 1-6. [179].

In this approach, term associations are computed incrementally in regular intervals, which makes the method applicable for streaming data. Yet, approaches for real-time analytics and visualization of streaming data face additional challenges, which will be treated and addressed in the next chapter.

Chapter 6

Real-time Analytics and Visualization of Change in Text Content

Contents

The previous chapter dealt with methods for time-oriented text mining for retrospective analyses on archived data. Some of the methods from Section 5.2 have been extended for real-time application and will be introduced in Section

6.2. Yet, before we come to the description of the real-time extensions, we provide a systematic overview of the different research challenges we face when performing real-time visual analytics of text streams. These challenges will be the topic of Section 6.1.

6.1 Real-time Visual Analytics of Text Streams: Overview and Challenges

This section partly builds on the following publications:

Christian Rohrdantz, Daniela Oelke, Milos Krstajic and Fabian Fischer. Real-Time Visualization of Streaming Text Data: Tasks and Challenges (Best Paper Award). Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek 2011, 2011. [150][1]

Daniel A. Keim, Milos Krstajic, Christian Rohrdantz and Tobias Schreck. Real-Time Visual Analytics for Text Streams. IEEE Computer 46(7): 47-55, 2013. [89][2]

There are different highly relevant real-word application scenarios where real-time text analytics may play an important role. For example, analysts in crisis response centers may gain situational awareness in emergency cases by monitoring user generated content that streams in through social media and similar web channels. Visual Analytics research in this area so far has been limited to retrospective analyses [25, 48, 113, 154]. However, providing a better understanding of the situation of people located in areas affected by natural disasters, accidents or terrorist attacks is highly desirable and might contribute to a more effective, efficient, and targeted allocation of aid.
Another relevant scenario is the analysis of financial news in real-time as it is

[1] The paper is a result of joint research discussions of all authors. The work was distributed according to different sections. Only the two sections that I was responsible for and wrote are used in this thesis. All authors cross-checked all sections and did proof-reading.

[2] The paper is a result of joint research discussions of all authors, which are therefore listed in alphabetic order. The work was distributed according to different sections. Only the two sections that I was responsible for and wrote are used in this thesis. All authors cross-checked all sections and did proof-reading.

known that the stock market may react to certain news. Traders who become aware of unexpected news before others do, have a potentially highly valuable trading advantage.
Furthermore, a topic like E-Democracy may become another important field of application, especially in Germany where participatory political decision making is an upcoming trend. This is reflected by the heated discussions about the construction of a new underground train station in Stuttgart ("Stuttgart 21"), where large parts of the public felt that the elected representatives ignored the interests of their voters. Moreover, in 2012 a relatively new political party in Germany ("Die Piraten")[3] was quite successful in demanding the inclusion of the public into political decision making through online web discussions open to everyone. Following this idea, the German Landkreis Friesland, a local administration area, decided to make its inhabitants participate accordingly through live Web discussions and polls ("liquid feedback")[4].
Finally, companies have a high interest in opinions that their customers convey through the Web. Besides direct customer feedback, as discussed in Section 5.2, this also includes reviews sent out to the public through web forums and social media. In some cases it may be crucial that a company reacts very timely to negative rumors in order to prevent that they "go viral" and strongly damage the reputation of the company. The same applies also to the reputation of politicians, hence, the British government has recently introduced a fast "Twitter strike force"[5] [6] to react quickly to so called shitstorms on Twitter.

As outlined in our survey on visual analytics of live streams [150], there are different kinds of challenges in this upcoming research field. Technical and algorithmic challenges, challenges of visually providing temporal context, dynamic visualization challenges, and challenges of supporting sense-making.

An overview of how the different kinds of challenges relate, is given in Figure 6.1. The numbers in the following enumeration relate to those in the figure.

1. **Incoming Data Objects:** The basic choice an application developer

[3]http://www.piratenpartei.de/ last revised on January 10th, 2013
[4]http://www.zeit.de/digital/internet/2012-11/liquid-feedback-friesland last revised on January 10th, 2013
[5]http://news-round.com/news/shitstorm-fight-camerons-fast-twitter-strike-force/ last revised on May 14th, 2013
[6]http://www.spiegel.de/politik/ausland/britische-regierung-startet-twitter-offensive-a-897226.html

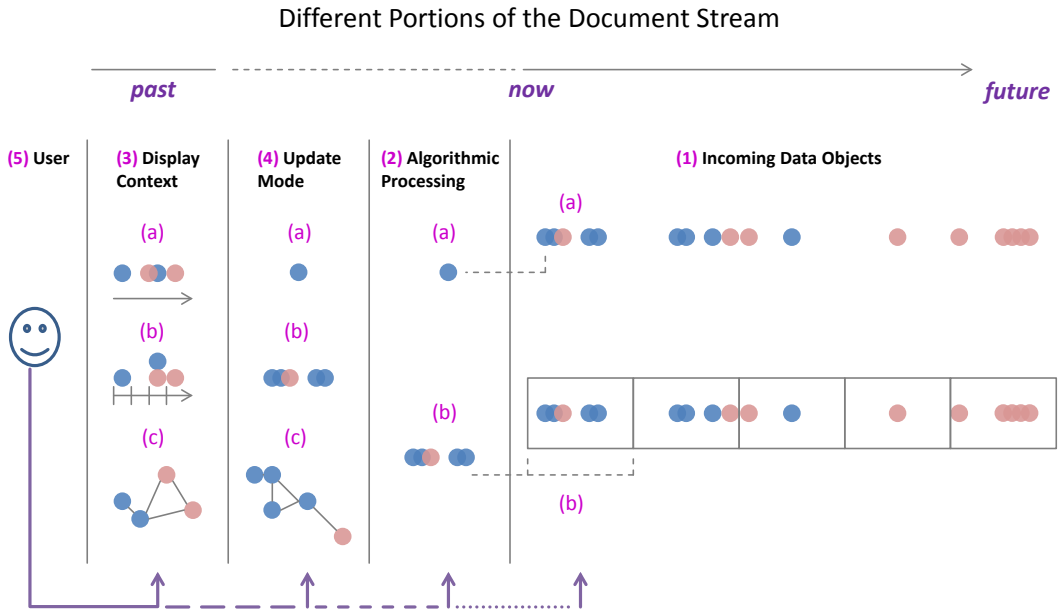


Figure 6.1: Different parts of the document stream and the corresponding processing options. Documents stream in from the future, the currently incoming documents are processed, and the visualization reflects the recent past.

has to make is whether s/he wants *(a)* to process each incoming data object individually or *(b)* to process data in chunks. The chunks can be determined either by fixing a time interval or by fixing the number of data objects to be contained in a chunk.

2. **Algorithmic Processing:** Text streams often contain strong bursts in data load and require scalable incremental algorithms, which are not available for some of the more sophisticated analysis steps. Figure 6.2 intends to give a feeling about the basic accuracy and efficiency of different text processing models [19, 116]. Of course, for each model different methods exist that again are subject to a trade-off between analysis quality and speed. Another aspect is that there is a certain interdependence between some of the models. To give one example, full syntactic parsing can help to improve anaphora resolution at the cost of introducing an additional delay. While the exact positions of the models within the diagram are subject to discussion, the rough tendency is clear. Most of the word-based methods scale quite well to real-time, sentence-based methods have quite different characteristics regarding their basic accuracy and efficiency, and most of the document-based methods are only

suitable for off-line processing. In some cases the reason is the long processing time and in other cases the lack of incrementality. For example, Latent Dirichlet Allocation (LDA) [20] needs to process the dataset as a whole to achieve good results. Hierarchical Dirichlet Processes (HDP) [2] partly overcome this problem, but cannot fully prevent a degeneration of the topic accuracy over time. In conclusion, a real-time text analytics system is mostly limited to a somewhat superficial analysis of the text. In order to reach the goal of providing a deeper insight, the extracted low-level text features must be combined thoroughly in an interactive visualization. An application developer has to keep in mind that some of the processing methods *(a)* can be performed on individual documents or *(b)* require a certain minimal amount of documents in order to produce meaningful results.

3. **Display Context:** In a visualization new textual input has to be put into context with previously seen text content, to enable the tracking of developments in the text message stream. Application developers have different options for providing context. Popular choices are *(a)* to visually represent each document individually along a timeline, *(b)* to show aggregations of documents along a timeline, or *(c)* instead of using a timeline exploiting the whole 2D layout space in order to put documents into a sophisticated visual structure, reflecting e.g. similarities. Only a limited context from the past can be displayed due to the constantly growing amount of data and the methods should be able to tolerate fluctuations in the data load.

4. **Update Mode - Dynamic visualization challenges:** As new data arrives the display constantly has to be updated in order to keep the analyst "up-to-date". It is hard to determine a suitable update rate and mode and there is a critical trade-off between making changes visible and at the same time keeping the display as constant as possible in order not to overload the analyst. Application developers have the option *(a)* to trigger an update for each single document, or to trigger updates whenever a certain amount of data has arrived or whenever a certain amount of time has passed. In the latter case the data can be added to

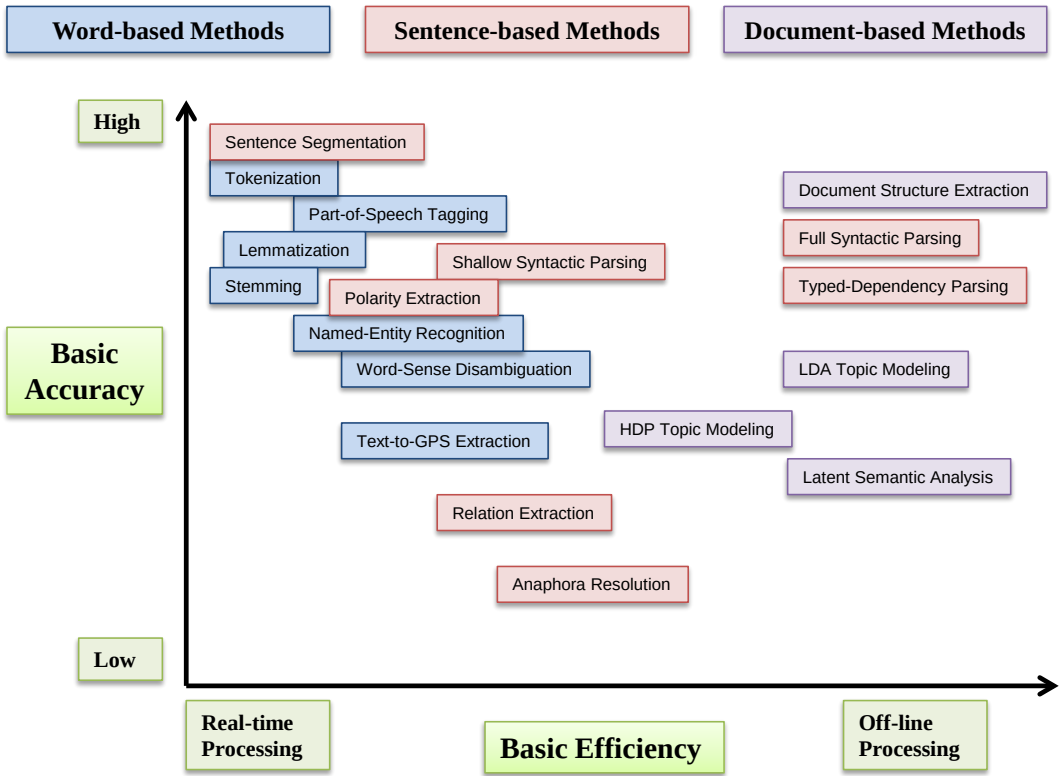


Figure 6.2: Overview of the most widely used text processing models for data analysis and a rough categorization of their basic average accuracy and efficiency.

the visualization (option *(b)*) or the current view can be entirely replaced by a new view created from the new data only (option *(c)*). Finally, good solutions are required to deal with two conflicting triggers for display updates: Stream-driven updates occur when new data arrives and is to be displayed and user-driven updates occur when the analyst tries to interactively manipulate the display for in-depth exploration.

5. **User - Supporting sense-making in complex analysis tasks:** To support sense-making a high-level semantic structure has to be derived from the text stream. The units of such a structure can be *thematic clusters*, *stories*, *topics*, *events*, etc. Over time, such units typically grow or shrink, overlap, split, merge, and (dis-)appear and are therefore hard to display incrementally.

The next two subsections will give more details about the technical and algorithmic challenges (relating to points 1 and 2) as well as the dynamic

visualization challenges (relating to point 4). Solutions to these challenges are necessary prerequisites for real-time visualization systems.

Technical and algorithmic challenges The technical and algorithmic challenges include all kinds of database related issues [13]. Most importantly, fast algorithms and data structures are needed that enable real-time processing and that can deal with incremental updates. Additionally, methods should not depend on a priori assumptions regarding data load, arrival rate or thresholds, because streaming data may have unpredictable contents. Moreover, we assume that streams cannot be stored completely because of their enormous size and that consequently on-the-fly processing and visualization is required. As mentioned, approaches usually either process each incoming data item individually or store data items in a buffer according to a predefined time frame, and then process the buffer content in regular intervals. However, it is not quite clear how to come up with suitable time frame sizes and consequently some approaches enable the user to modify this parameter dynamically [5, 85].

There are different ways to address some of the outlined challenges. Wong et al. [181] suggest making the algorithmic analysis dependent on the volume of incoming data. If in the short term there is a high data load, the trade-off is between being less accurate, as in [181], and temporarily buffering data as done by Alsakran et al. [7] who buffer "document items during peak times and handle them in idling periods."

Another important issue we mentioned previously is the derivation of a higher-level semantic structure, such as for example *topics*, *stories*, *events*, or *theme clusters*, in real-time. Ishikawa and Hasegawa [85] cluster documents in real-time incorporating a "document forgetting model", i.e. older documents have less influence on the clusters. The clustering is an incremental k-means algorithm, which has the drawback that the parameter k has to be to predefined. Zhang et al. [187] introduce an evolutionary hierarchical Dirichlet process that allows the number of clusters to vary over time, but they do not comment on a potential real-time capability of their approach. Rose et al. [151] cluster keywords for each time interval using a hierarchical agglomerative clustering algorithm in order to learn themes without having to predefine the number of clusters. Themes of adjacent time intervals are grouped into "stories" according

to their similarity. These stories can split and merge over time.

Another issue is querying in real-time. Hetzler et al. [74] allow the user to define a network combining different dynamic queries and visualize the corresponding result sets in the same structure.

For many approaches, there is no explicit description of how they behave in the initialization phase, i.e. when the stream starts and the first data items come in. While some algorithms might not need data from the past others heavily depend on it and will require some time to become stable and meaningful.

Dynamic visualization challenges Visualizations must necessarily change over time to reflect changes of the stream. As mentioned, there are two different approaches to visually express changes in the data. The first and more straight-forward solution is to create individual views for different time intervals and sequentially provide this series of views. Here, the challenge is to relate one view to the next view and moreover ensure that the user is able to perceive the differences between both. Usually, some sort of transition is used like the animation of appearing and disappearing words in the Parallel Tag Clouds [31]. Many standard text visualizations can potentially be adapted to this scenario.
The second and more demanding possibility is to provide a continuously evolving visualization reflecting a continuous data flow. In such a scenario the current view is constantly modified in accord with the newly arriving data. The modification is triggered by an update and usually expressed through motion. A nice example for the use of motion is [5]. The authors review the perception of motion and derive design guidelines for visualizing changes in live text streams. More recent approaches go in a similar direction: Dörk et al. [46] chose to "use primarily motion to represent data change" and Alsakran et al. [7] point out that "the dynamic procedure of this change is critical for reducing change blindness". However, even if changes are nicely traceable using motion, too many changes at once will overwhelm the user. The system of Alsakran et al. [7] addresses this issue by having at most one label change when a new document arrives and buffering documents during peak times at the cost of introducing delays.

The problem of having a lot of change in the display on updates is also given in the case of visualizations that display the arrival order of items by position. Hao et al. [67] discuss different ways of shifting displayed items on the arrival of new items minimizing the movement in the display. Chin et al. [29] discuss how standard visualizations can be extended to support the analysis of dynamic data and experiment with different ways of updating their dynamic spiral timeline keeping the display more ore less constant.
In the case of frequently or even continuously updating displays, support for interactive exploration is a further challenge. While a lot of approaches enable some user interactions like filtering and highlighting search queries on-the-fly, in-depth exploration is difficult. Interaction with a visualization while the stream keeps coming in, brings the problem of "adapting its display to stream content and user interaction" [5]. Those are two potentially conflicting triggers for changing the display. What if the user might like to explore a peculiarity further? Does he have to worry that it might disappear in the next moment when new data comes in (and could be lost forever)? One solution to prevent this is to freeze a certain point in time for exploration, the user can pause the stream for exploration [7, 74] or take interactive snapshots to save the current situation for a later off-line exploration as in [74]. Many other approaches, however, do not address this problem explicitly or assume that the whole stream can be saved and retrieved.

6.2 Real-time Analytics of Critical Event Episodes in Document Streams

Parts of the scientific contributions that I will provide in this section have been made during a research visit at Hewlett-Packard Labs in Palo Alto, California. In his Bachelor Thesis, Michael Hund, a student that I co-supervised, has extended some of the concepts presented here to be applicable to a 10% Twitter live stream in real-time [82]. Parts of this have also been published as:

Milos Krstajic, Christian Rohrdantz, Michael Hund and Andreas Weiler. Getting There First: Real-Time Detection of Real-World Incidents on Twitter. Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics

"Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA, 2012. [98]

6.2.1 Background

The amount of text data becoming available through real-time document streams is steadily increasing. Widely used examples for document streams are professionally edited news feeds or user-generated social media services like Twitter. In addition, there are further user-generated streams that may not be accessible for the public as for example customer feedback that a company receives through web surveys (see Section 5.2). The detection of interesting developments, hot topics, or critical issues in real-time is crucial for analysts from various domains. For example, authorities can learn about catastrophes or terrorist attacks that eye witnesses report using hand-held devices, or companies can learn about issues regarding their products, service, or reputation monitoring customer feedback, social media, or news feeds. In these scenarios the real-time component plays an important role. The sooner an issue will be discovered, the sooner actions can be taken on it and potentially prevent it from becoming worse.

As described in Section 6.1, from the analysis perspective the real-time monitoring of text or document streams is quite challenging. Documents are a rich source of diverse information that may quickly overload the analyst. Certainly, a reduction of the information to a relevant subset of manageable size will be required when monitoring document streams. Yet, it cannot be expected that the analyst will spend hours continuously focusing her/his attention on a visual interface displaying information filtered from the stream. In many scenarios a constant human monitoring is not feasible, but also not really necessary given that the system is able to trigger an alert once something unexpected occurs. In the case of alerts the analyst should be provided with information about the cause of the alert as well as context information both in content and time. In this section we show that the detection and visual analysis of event episodes in real-time is a practicable and beneficial way of monitoring document streams. In a sense, it is an extension of the off-line methods described in Section 5.2. The terminology, however, is adapted to match the conventions of the real-time

data management community. To be more specific, we consider the occurrence of specific words at given points in time as *events* and show that *event episodes* point to interesting findings and sometimes even critical issues in different domains such as customer feedback, or microblogs. In particular, we extend the previously described methods for the off-line detection and visualization of critical issues (here: event episodes) in document streams to operate in real-time. In addition, we provide empirical evidence for the usefulness of real-time visual analytics of event episodes in document streams in general, and the usefulness of the presented techniques in particular evaluating their performance and discussing findings in different analysis contexts and settings. For this purpose we test our methods on two datasets for which some basic ground truth is available. The first dataset consists of approximately 87.000 web surveys with customer feedback sent to company and is an extension of the dataset used in Section 5.2. The second data set contains more than 1.000.000 microblog messages, similar to Tweets, and was provided as part of the VAST Challenge 2011 Mini Challenge 1[7]. A basic ground truth was provided as part of the contest.

Motivation

As outlined in Section 2.2, for the visual analysis of temporal developments in text collections lately approaches based on topic modeling have gained a lot of popularity, for example TIARA [109], TextFlow [35], ParallelTopics [47], and approaches for the off-line exploration of Twitter data [25, 48]. However, topic modeling in real-time has several problems: (1) The underlying algorithms are computationally expensive and large datasets lead to processing times that make an instant real-time analysis infeasible. (2) Typically, chunks of newly arrived data are processed in regular intervals, which may already introduce a considerable delay. One possibility is to use an incremental topic modeling method with the disadvantage that newly emerging topics will only show up after a certain while, when the model has finally adapted and one of the old topics has been iteratively transformed into this new topic. The other possibility is that for each chunk different models are created and the topics from one time interval are heuristically mapped to topics from the previous and next

[7] http://hcil.cs.umd.edu/localphp/hcil/vast11/ last revised on February 13th, 2013

time interval. (3) The number of topics that should be requested as an output typically cannot be predefined in a straightforward manner and using a fixed number of topics in a streaming environment is problematic. This parameter choice may have a crucial impact on the output. (4) Additionally, in order to guarantee a good quality of the results they may have to be manually fine-tuned in iterative post-processing steps [105]. A recent approach has made topic modeling more scalable with respect to the number of documents that can be included in the model using a special distributed architecture [176]. However, even this approach learns one topic model for the whole time span of interest and is therefore limited to retrospective non-incremental analyses. In conclusion, topic modeling brings a lot of useful information for exploration tasks and is the state-of-the-art method for gaining insight into document collections. Yet, tracking topics in real-time is an approach that brings many unresolved problems.

Consequently, the fundamental strategy we pursue for monitoring document streams is different in order to achieve real-time performance for instant analysis. Instead of focusing on topics, we suggest to track the behavior of individual words and search for sudden and unexpected temporal accumulations of certain words. As in Section 5.2, if the frequency of a word (*event*) becomes much higher than it could be expected based on observations from the past, then we speak of a *candidate event episode*. The relevance of a candidate event episode will be determined considering additional information from the contained documents, such as context coherence and sentiment (cf. Section 5.2). The advantage of tracking individual words is that the event episode detection and visualization can be readily performed incrementally in real-time storing only limited amounts of data. In addition, we are able to capture issues that are reflected only by a relatively small number of rather infrequent words and thus constitute weak signals that may otherwise easily submerge in the multitude of established long-term topics. In order to provide topic context at a certain point in time we use relevance-based topic modeling. Based on the event episode detection methods, we are able to filter relevant documents at any point in time and limit the topic modeling to these. This brings the advantage that the datasets become so small that a real-time topic analysis is feasible.

Goals and Contributions

The contributions of this section are twofold: On the one hand we demonstrate the usefulness of visually analyzing event episodes in document streams for the purpose of becoming aware of sudden unexpected developments, emerging hot topics, and critical issues in real-time. On the other hand we provide three technical contributions:

- We extend the off-line algorithm for the detection and scoring of event episodes in document streams to perform real-time alerting. We suggest an intelligent data management so that only limited memory resources are required.

- We introduce a new way of incrementally displaying document streams with a linear time line and integrate it with an extended version of the *time density plots* technique in order to provide a visualization that enables the analyst to monitor and access the document stream for real-time analysis and understanding.

- We introduce a new way of performing topic modeling in real-time for analytical reasoning by limiting the computation at each point in time to temporarily relevant documents only.

The rest of this section is structured as follows: In Section 6.2.2 we explain how our technique relates to other research approaches from different research fields. In Section 6.2.3 we describe the real-time detection and scoring of event episodes in document streams. In Section, 6.2.4 we introduce several new ways of providing relevant semantic and topical context for analytical reasoning in real-time. Next, in Section 6.2.5 we describe the real-time capable visualization and its integration with the *time density plots* extension. In Section 6.2.6, we show the suitability of the approach in analysis use cases on real-world data. An extensive performance evaluation of the presented methods is provided in Section 6.2.7 focusing on the temporal development of the storage requirements for the automatic detection and scoring. Finally, in Section 6.2.8 we briefly summarize and discuss the content of this section.

6.2.2 Related Work

The research provided in this section relates to work from different areas. Some of the fundamental motivations and concepts are shared with the field of *Anomaly Detection*, which "refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains." [27]. In particular, we are interested in *Collective Anomalies*, where according to Chandola et al. [27] "the individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous". Only little work has been done for the detection of anomalies in text data, one of the few examples is [52]. Often *Anomaly Detection* in text is referred to as *Novelty Detection* [118] and focused on *Topic Detection and Tracking (TDT)* [6]. Baker et al. [14] suggest an approach to TDT based on probabilistic generative models. Topics are learned and documents belonging to a topic are clustered into what they call events. A related approach using Gaussian mixtures was presented by Hansen et al. [61]. In contrast to these top-down approaches monitoring general topics, clusters or classes, we choose a bottom-up approach monitoring specific words. This has the advantage that we do not need to perform expensive operations for topic modeling or costly operations in high-dimensional vector spaces which may work only with a considerable delay depending on the data load. Moreover, we do not need fixed update intervals, each document triggers an update, and we are able to detect issues that are only sparsely represented with respect to the overall data. The closest to our approach are strategies that have not yet been applied to text, i.e. approaches for the detection of *event episodes*. "An episode is a collection of events that occur relatively close to each other in a given partial order." [115]. In our approach the occurrence of a word in a document at a given point in time is an *event*. Unlike the original work on event episodes we do not mind the order, because we require all events in an episode to belong to the same type. That is, in our event episodes we are interested in temporal sequences of documents containing the same word and put a focus on the time density within the sequence, and on further features derived from the containing documents.

Within the visual analytics field recently approaches have been introduced where event episodes in text data can be visually detected and explored. While all of them have their particular strengths, none of them enables a real-time monitoring and comparison of event episodes relating to different event types. An item-based plotting as in the pilot study of Section 5.1 will inevitably lead to clutter when the data load is high. The CloudLines approach [94] allows a comparison of event episode patterns for different event types and in-depth exploration, but yet there is no straightforward way to compute them incrementally without introducing a certain delay. Therefore, its applicability in time-critical real-time scenarios is limited. The time density plots approach, introduced in Section 5.2, scales the time lines differently for different features, which makes a comparison of features hard. It is the only approach that comes with an integrated automatic event episode detection, which, however, was not designed for real-time detection. Our aim is to extend the event episode detection to real-time analysis tasks. The memory usage shall be minimized to allow a high scalability with respect to data load. Together with the automatic detection, we improve the visual display to provide global time contexts for the real-time monitoring.

6.2.3 Automatic Event Episode Detection and Scoring in Real-time

In the data mining and machine learning literature [115] an *event* is a pair of an *event type* and a time stamp. For our specific scenario this definition has to be extended as follows:

- An *event* is a triple (W, D, t), where W is word occurring in a document D at the time stamp t.
- The word W defines the *event type* and the set of all event types is V, the vocabulary that is monitored. We can either monitor a predefined set of event types (words) or have an unclosed set of event types, e.g. all nouns and compound nouns as discovered by a part-of-speech tagger or all named entities as discovered by an algorithm for named entity

recognition. There is also the possibility to join clusters of related words or sets of synonyms into one event type.

- An *event episode* E is a sequence of *events* having the same *event type* and unexpectedly low time distances. That means, that the gap between two successive events within the event episode has to be smaller than a certain threshold, for example the incrementally computed average time gap avg_gap for the respective event type in analogy to the off-line processing presented in Section 5.2. As interesting issues in text streams usually have a bursty nature, and in the real-time analysis there is a limited amount of data that we can store in-memory, we restrict this threshold further. In the real-time case the gap between two successive events within the event episode has to be smaller than a fraction λ of the average time gap avg_gap for the respective event type, where $\lambda = 0.25$ is a reasonable value. In a general case, an event episode for an event type W_k is defined as follows: $E(W_k) = ((W_k, D_1, t_1), (W_k, D_2, t_2), ..., (W_k, D_n, t_n),)$ with $dist(t_i, t_{i+1}) \leq \lambda \cdot avg\_gap(W_k, t_i)$. Note that one document D can be part of different event episodes for different event types.

Apart from the time stamp and event type information, we have to save the whole document, because it contains interesting information for scoring candidate episodes, such as context and sentiment. Our approach consists of two separate processing steps: The detection of candidate event episodes and the scoring of candidate event episodes, which will be described in the following.

Detection of Candidate Event Episodes

The algorithm for detecting candidate event episodes is an extended version of the time density method presented in Section 5.2 and in contrast to the previous approach is able to cope with incremental updates for real-time event episode detection. The algorithm will be explained step by step:

1. Split up the document stream: As illustrated in Figure 6.3 each document can contain different event types, i.e. words of interest. The document stream will be split up according to event types into one separate stream per

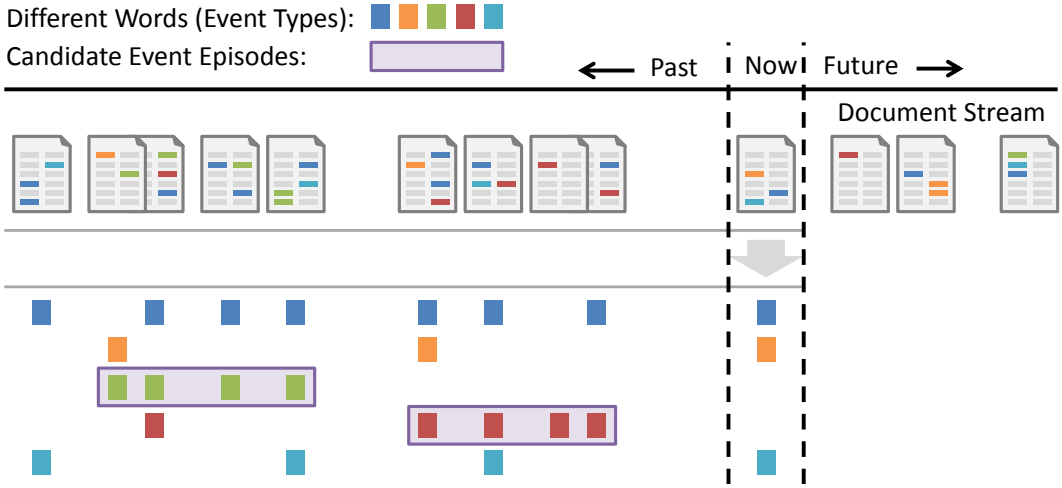


Figure 6.3: The document stream (top) is split up into different parallel event type streams (bottom), for which candidate event episodes are detected.

event type. Each event type stream can be scanned independently in parallel processes to automatically detect candidate episodes.

2. Scan each event type stream for candidate episodes: For each event type some data has to be saved into a database DB during the detection of candidate event episodes:

- The incrementally computed average time gap avg_gap of the event type.
- The time stamp when the event type was seen last $last\_t$.
- The current candidate event episode E, if existent.
- The expiration date exp_date of the current episode, that is the point in time when the episode will be finished if no further event of the current event type can be observed.
- The point in time when the monitoring process was initiated $start\_t$, which is independent from event types and just has to be stored once globally.

The following flow chart in Figure 6.4 shows the detection of candidate event episodes E for a specific event type W: It has to be noted that the update of the average gap between events of the same type is parameter dependent. The

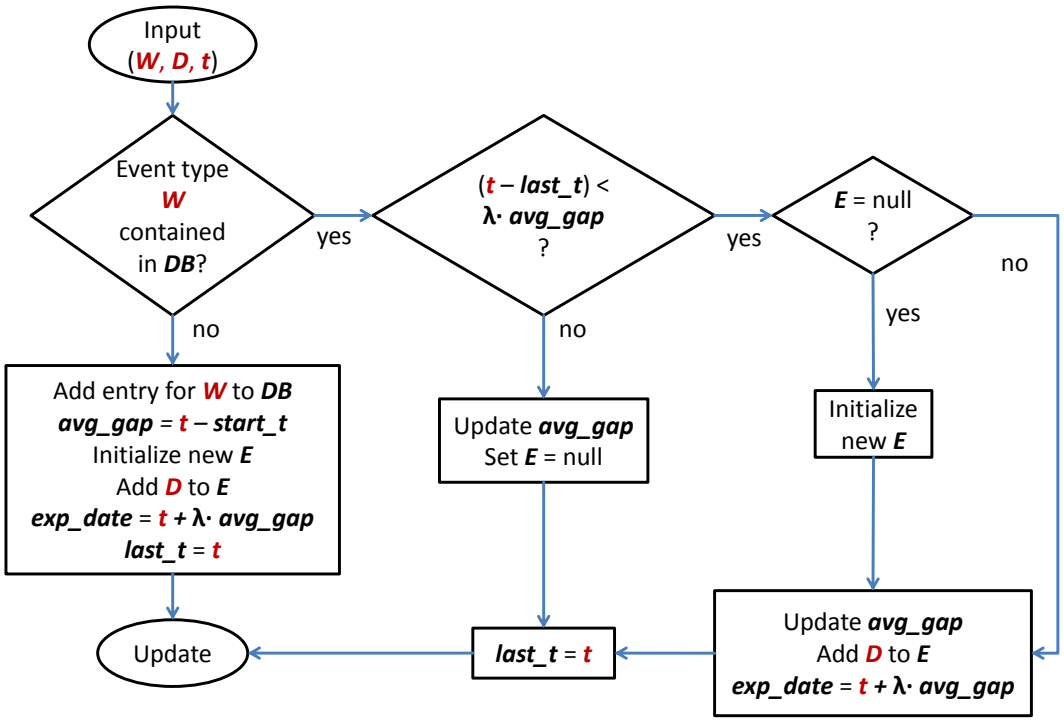


Figure 6.4: Flow chart shows the detection of candidate event episodes E for a specific event type W.

average gap calculated at the previous event is updated based on information about the time gap between the current and the previous event. There are two different options. It is possible to compute the average gap in a way that only the recent past is considered, similar to a moving window approach (mov_avg_gap), or alternatively in an incremental manner ($incr\_avg\_gap$):

$$mov\_avg\_gap(t_i) = (1 - a) \cdot mov\_avg\_gap(t_{i-1}) + a \cdot (t_i - t_{i-1})$$

or

$$incr\_avg\_gap(t_i) = \frac{1}{i} \sum_{k=1}^{i} (t_k - t_{k-1})$$

where t_0 is the starting time of the stream monitoring.

In the first case, the parameter a determines the influence the current gap has with respect to the past and can be modified by the user. Thus, more

recent gaps have a higher influence on the moving average than older gaps. The influence of older gaps decreases exponentially, and the moving average can therefore quickly adapt to global trends like a global decrease or increase in frequency. By default a is set to 0.1 which according to our tests is a good choice for different data sets.

In the second case all time gaps get the same influence, resulting in an incremental average that is more stable over time. According to our empirical observations, the results are better in this second case. Yet, we did not perform a formal evaluation on that, because it is not quite clear where a useful ground truth could come from.

3. Save metadata for candidate episodes: As long as a document belongs to at least one of the current candidate episodes it will be stored in memory. In addition, further metadata will be extracted from the document and saved, for example a bag of content words. For us content words are nouns, verbs, and adjectives detected by a part-of-speech tagger. By storing these words separately in a bag of words we retain the most relevant information about the document content in a structured way. For each content word that ever appears in the course of the monitoring we also increase a global counter. This counter allows us to estimate how specific this content word is; in general, we consider the less frequent words to be more specific. Moreover, depending on the data set and task we also save a sentiment score for each event. The metadata attached to a document will be useful when scoring the candidate episode, which will be detailed later in this section.

4. Delete expired candidate episodes: As mentioned before each candidate episode has an expiration date exp_date. When the expiration date has come, we know that the episode is obsolete, because too much time has passed since observing the last event of the corresponding type. At any point in time all existing candidate episodes are stored in the order of their expiration date within a priority queue. As time passes it will be checked in regular intervals if there are any expired episodes and these will be deleted. If and only if a document contained by an expired episode is not contained by any other existing episode, it will be deleted from main memory and stored only

on hard disk together with the attached metadata about content words and sentiments etc. This allows us to keep the amount of data to be stored in the main memory for instant processing low.

Scoring of Candidate Event Episodes

A candidate event episode consists of a set of documents that is attached to a certain event type. That means the documents share a certain word of interest and are closer in time than it could be expected. However, we have to consider the possibility that this might just be due to a random accumulation. In order to assure that a candidate event episode is meaningful it will be scored. Actually, the purpose of scoring a candidate event episode is twofold: Apart from excluding the possibility that it was just a random effect, in addition, we have to determine whether it appears to be interesting with regard to the current analysis task. Only those candidates with a high score, for example above a certain threshold, will be considered real event episodes in the further analyses and visualizations. On the one hand, the score has to be based on the limited amount of information that we are able to store and access in real-time. On the other hand, the more data aspects we consider in the score, the less likely it will be biased towards overrating single aspects.
We suggest a modification of the off-line scoring function suggested in Section 5.2, and explain how it can be computed in real-time. The scoring components and the rationale behind them are briefly outlined in the following:

- Time density: The higher the relative time density within a candidate event episode, the less likely it is to have happened by chance.
- Episode length: The more events a candidate event episode contains, again the less likely it is due to a random effect.
- Event type sentiment: In many cases it is critical to become aware of events that have a negative connotation, like catastrophes, accidents, or customer complaints. In these cases it makes sense to limit the analysis to candidate event episodes that are referenced with negative sentiments. The stronger the negative sentiment, the more interesting.

- Context coherence: The fact that several documents share one event type, does not necessarily mean that they mention it in the same context or sense. It has to be assured that the content apart from the event type is also similar. Otherwise, the episode is probably not meaningful.

The shorter the relative time gap between two events is, the higher is their time density value. The average of the time density values within the episode E gives us the *time_density* score of the episode:

$$\text{time\_density}(E) = \frac{1}{|\{D \in E\}| - 1} \sum_{i=1}^{i<|\{D \in E\}|} \left(1 - \frac{t_{i+1} - t_i}{avg\_gap}\right)$$

The *episode_length* is simply the amount of documents contained in an episode E:

$$\text{episode\_length}(E) = |\{D \in E\}|$$

We use the feature-based sentiment analysis algorithm suggested in Section 5.2 in order to determine a sentiment value for the event type W in each document D. Then, we sum up the values for all of the episode's documents into the score component named *sentiment_negativity*. We multiply the sum with *-1* in order to get high scores for negative sentiments. Of course, this component can easily be modified if the interest would lie for example in positive sentiments or in ambiguous sentiments.

$$\text{sentiment\_negativity}(E) = -\sum_{D \in E} Sentiment(W, D)$$

In the evaluation of the off-line scoring *context coherence* was a valuable score component. In order to determine the context coherence the *likelihood ratio test* was used, which operates on a contingency table (see Table 6.1). This method for hypothesis testing will be applied to any content word CW contained in the documents of the episode. The test yields high values if the investigated content word occurs statistically significantly more frequent within the documents of the event episode than in the whole document stream observed so far.

All necessary values can be calculated from the dynamically stored data:

- The first column of Table 6.1 only refers to documents contained in the

	$D \in E$	$D \notin E$
$CW \in D$	a	b
$CW \notin D$	c	d

Table 6.1: Contingency table showing the number of documents D depending on a certain content word CW and a certain event episode E.

event episode. These documents are stored together with their content words as long as the event episode has not expired. For each content word CW it can easily be counted in how many documents within the event episode it is present (cell a) and in how many it is not present (cell c).

- The second column of Table 6.1 refers to the whole document collection apart from the current event episode. As mentioned before, for each content word observed during the monitoring process there is one global counter, which counts in how many different documents this word has appeared ($count(CW)$). In addition, it is sufficient to know how many documents $sum(D)$ have been observed during the monitoring process in order to calculate the cells b and d. Specifically, $b = (count(CW) - a)$ and $d = (sum(D) - a - b - c)$.

Then, the four cells a, b, c and d are used to calculate the likelihood ratio (see Equation 6.1):

$$\begin{aligned} &\text{likelihood ratio} = \\ 2 \cdot \Bigg(& A \log \left(\frac{A/(A+B)}{(A+C)/(N)} \right) + B \log \left(\frac{B/(A+B)}{(B+D)/N} \right) \\ & + C \log \left(\frac{C/(C+D)}{(A+C)/N} \right) + D \log \left(\frac{D/(C+D)}{(B+D)/N} \right) \Bigg) \\ &\text{with } N = A + B + C + D \end{aligned} \tag{6.1}$$

In the different datasets we used, we could observe that due to language variation, usually, the number of content words strongly associated with an episode is below 10, even if the documents of the episode have similar contents. Consequently, it is sufficient to consider the top 10 (or similar number) of

associated words for an episode. In the off-line processing their likelihood values are summed up to get a score for the context coherence. This sum will be high for episodes that have a number of CWs occurring significantly more likely within the documents of the event episode than in the other documents. The presence of several strongly associated content words indicates event episodes that have a very specific and coherent context. In the live processing, however, this method has a disadvantage: The absolute likelihood ratio values depend on the overall amount of data observed. As in the streaming environment this amount is constantly increasing, it might be problematic to compare values for different points in time. To overcome this issue, we use the likelihood ratio only for ranking the words, but sum up scores that do not depend on the overall amount of data observed. We use the correlation coefficient ϕ, which tells us how strongly a content word is correlated with the event episode. This coefficient can be derived from the very same contingency table:

$$\phi = \frac{(a \cdot d - c \cdot b)}{\sqrt{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)}} \tag{6.2}$$

As also infrequent words can have high correlations, we omit words that do not appear more than once within the episode and multiply the correlation value of the remaining content words with their frequency within the episode, which is the same value as in cell a of the contingency table: $freq(CW_i, E)$. Both correlation and frequency are important, however, the correlation is of a higher relevance. Therefore, we square the correlation in analogy to the χ^2 significance test which is the product of a squared correlation coefficient and the overall amount of data seen. The multiplications are then summed up.

$$\text{context\_coherence}(E) = \sum_{i=1}^{10} (\phi(CW_i, E)^2 \cdot freq(CW_i, E))$$
$$\textit{where } \text{likelihood ratio}(CW_i, E) \geq \text{likelihood ratio}(CW_{i+1}, E)$$
$$\textit{and} freq(CW_i, E) \geq 2$$

Finally, we integrate all of the described components as factors into one formula. It is optional to weight the factors differently, by default all weights are equal ($\alpha = \beta = \gamma = \delta = 1$):

$$\text{score}(E) = \text{time\_density}(E)^{\alpha} \cdot \text{episode\_length}(E)^{\beta}$$
$$\cdot \text{sentiment\_negativity}(E)^{\delta} \cdot \text{context\_coherence}(E)^{\gamma}$$

In different empirical tests we determined the following parameter setup to be useful for our datasets: $\alpha = \beta = 2, \delta = \gamma = 1$. The first two factors depend on temporal document distributions and are less prone to noise than the last two, which extract scores directly from the text.

6.2.4 Relevance-based Context and Topic Analysis

For the most relevant event types and documents identified in real-time, we suggest further analytics steps. First, we calculate similarities among concurrently top-scored event episodes in order to put them into context. Secondly, at any given point in time, we use the event episode information in order to filter a reduced set of highly relevant documents. On this reduced set we are able to run topic modeling in real-time or near-to-real-time.

Event Episodes Similarity

One particularity of our bottom-up approach monitoring individual words/event types separately, is that two (or more) different event types may point to the same real-world issue. We could observe that this may have two different causes. First, both of the two words describe the same issue and they co-occur in the same documents. Second, the two words are synonyms and either one of them is used to describe the issue. Independent from the cause the analyst should be notified that two concurrent event episodes are closely related. On the one hand this supports the assumption that it is a non-random effect and on the other hand the analyst might want to complement the detailed analysis of one of the episodes with information from the other episode.
In order to detect closely related event episodes we evaluate their pairwise similarities. As it is not feasible to do this for any pair of candidate episodes in real-time, we limit this similarity analysis to a fixed number k of top-scored event episodes at each point in time. Whenever a new episode enters the top k episodes, its similarity to all other event episodes among the top k will be calculated. As mentioned there can be two different kinds of similarity, for both

of which we have designed different similarity measures: A *co-occurrence-based* and a *content-based* approach.

Co-occurrence-based Similarity

In order to determine the co-coccurrence-based similarity between two event episodes E_1 and E_2 we apply the Jaccard Index [86] on documents. Taking the two sets of documents belonging to E_1 and E_2 we divide the size of the intersection of both sets by the size of the union:

$$\text{document\_sim}(E_1, E_2) = \frac{|\{D : (D \in E_1) \wedge (D \in E_2)\}|}{|\{D : (D \in E_1) \vee (D \in E_2)\}|}$$

The case of maximal similarity is that both sets are identical, in this case the *document_sim* will be 1. In the case of maximal dissimilarity, in which both sets are completely disjoint, the *document_sim* will be 0. In all other cases the resulting value will be the closer to 1, the higher the proportion of joint documents is. In the customer feedback data the two simultaneous episodes with the highest similarity belong to the event types *survey* and *email*. Both episodes are related to an error with respect to an email invitation for survey participation. In the microblog data there are quite a number of event episode pairs that have a similarity of 1, i.e. two different event types lead to exactly the same event episode.

Content-based Similarity

In order to determine the content-based similarity between two event episodes E_1 and E_2 we investigate whether both episodes share context words, that are strongly associated to both. If at least two documents of an episode E contain a content word CW, we say that $CW \in E$ and then apply the following formula:

$$\text{content\_sim}(E_1, E_2) = \sum_{(CW \in E_1) \wedge (CW \in E_2)} (\phi(CW, E_1) \cdot \phi(CW, E_2))$$

If there exist content words that are strongly correlated with both episodes, these will contribute a large summand. As there are typically only few of

such content words, we could observe that the sum is quite expressive with respect to the content-based similarity. In the customer feedback data the two simultaneous episodes with the highest similarities belong to the event type pair *{packing list, packing slip}*. In this case, the two corresponding episodes do not share a single document, because customers either use one or the other word, but not both at a time. As the respective event types refer to the same issue though, they share a lot of common vocabulary, which could be detected by our method. For the Microblog data the most similar simultaneous event episodes belong to *{shortness, breath}* and actually relate to one symptom: shortness of breath.

Relevance-based Topic Modeling for Real-time Analytical Reasoning

Earlier in this chapter, we have identified two steps as highly important when processing text streams: 1. to filter the relevant information, 2. to derive a higher-level semantic structure from it. The detected event episodes fulfill both criteria, however, even though the relations among related event episodes can be detected and revealed, this approach does not provide as much context information as topic modeling. Consequently, we suggest using the event episode detection as a filter step for real-time topic modeling. As mentioned before, at any given point in time, a document will be kept in memory if and only if it contains at least one event type that is currently more frequent than expected. This ensures that the documents kept in memory contain all relevant information about the automatically detected issues. Of course, the set could be further restricted to documents that contain at least ($x > 1$) event types, that are currently more frequent than expected and that at the same time are contained in candidate event episodes having a length of at least ($y > 1$). Considering only the filtered documents we are able perform a relevance-based topic modeling at any given point in time.

Fortunately, the number of documents kept in memory (with $x > 0$ and $y > 0$) for the given datasets is already constantly low at any point in time, as our empirical experiments in Section 6.2.7 will show. Such small sets of filtered documents can be processed by state-of-the-art topic modeling in a split second on a common workspace computer. We use the MALLET[8] implementation for

[8]http://mallet.cs.umass.edu/topics.php last revised on May 3rd, 2013

topic modeling and restrict the documents to those with $x > 0$ and $y > 2$, i.e. process only those documents containing at least one event type contained in an episode of at least length 3. Moreover, we reduce the input documents to content words (nouns, verbs, adjectives) only to get a further speed-up. As there is no optimal way to fix the number of topics $\#topics$ beforehand, we currently use a simple heuristic that makes this number dependent on the amount of documents $\#docs$ kept in memory: $\#topics = \sqrt{\frac{\#docs}{2}}$. The analyst can request to get the topics at any point in time and can make use of them to get a better understanding of currently ongoing issues. Of course, it would be also possible to automatically generate and present topics periodically, according to fixed time intervals or as soon as a fixed amount of data has streamed in. The topics of one period could be matched with topics from adjacent periods, however, this goes beyond the scope of this thesis.

6.2.5 Visual Analytics of Event Episodes in Real-time

The detection and scoring of event episodes is a helpful automatic analysis step. However, in the end a human analyst has to explore, interpret, and understand the detected event episodes. This involves accessing and reading documents, and putting the event episodes into a temporal context. Consequently, it is necessary to provide the analyst with a visual display where s/he is able to perform her/his exploration tasks. In contrast to the off-line exploration, the exact time points of events are of a higher relevance in the real-time analytics scenario. To convey this information we provide a novel approach to item-based time series visualization with the following characteristics:

Visual Representation of Documents: For each incoming document one visual object with triangular shape is added to a timeline representation. The size of the shape depends on local data distribution characteristics, inspired by the CloudLines approach [94], where the size of a visual object (circle) depends on the data density at the respective point in time (in relation to the maximal global data density). In our approach the width of the triangle, i.e. the space it consumes along the time axis, is the average time gap avg_gap[9] at that point in time (see Figure 6.5). Like that, two visual objects

[9]For all figures displayed in this and the following sections we use the $incr\_avg\_gap$ as defined in Section 6.2.3.

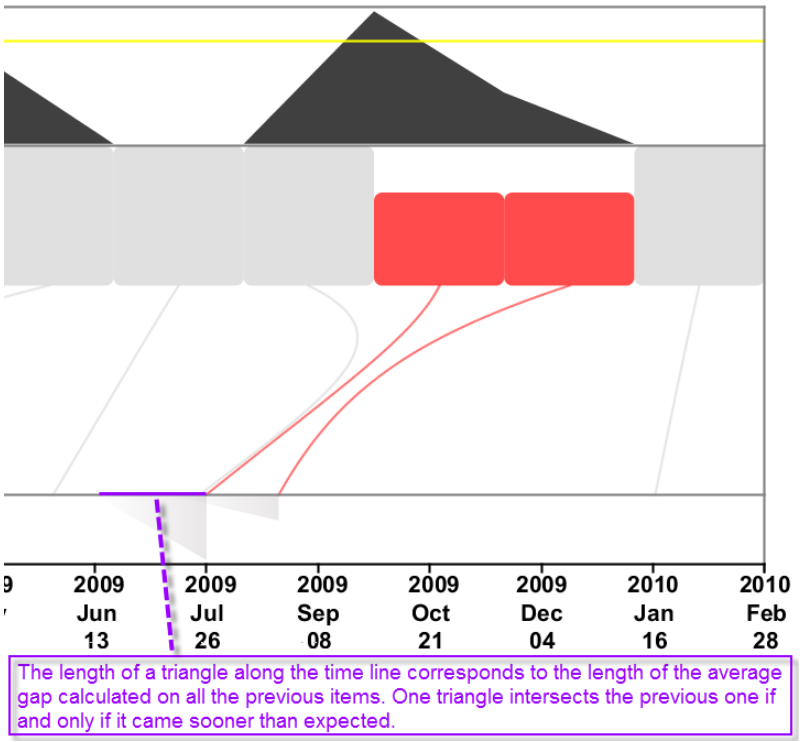


Figure 6.5: Example of the new visual representation with triangles (on the bottom). In order to provide a better understanding we coordinate the triangle visualization with the time density plots introduced in Chapter 5.2. The width of a triangle to the left reflects the *avg_gap* between pairs of items that have the same event type.

will overlap if and only if they are closer in time than in average. The height of the triangle depends on the documents' relative time distance to the previous document, see Figure 6.6. The height correlates with the interestingness from the analytic point of view: The closer an item is to its predecessor, the more interesting it is, and consequently the larger in height it will be scaled. The height corresponds to the time density value, i.e. it is maximal when two documents occur at the same point in time. Yet, if the gap to the previous document is larger than the interestingness threshold (time density = 0), it will be represented by one pixel only.

Initially, the lengths of the triangles along the time line may vary and only after a while become more stable. Such an effect is observable in the time lines of the features *air bag* and *air pack* discovered in the customer feedback stream, see Figure 6.7. Interestingly, both synonyms where not contained in the first part of the data and therefore the *avg_gap* initially is large and decreases over

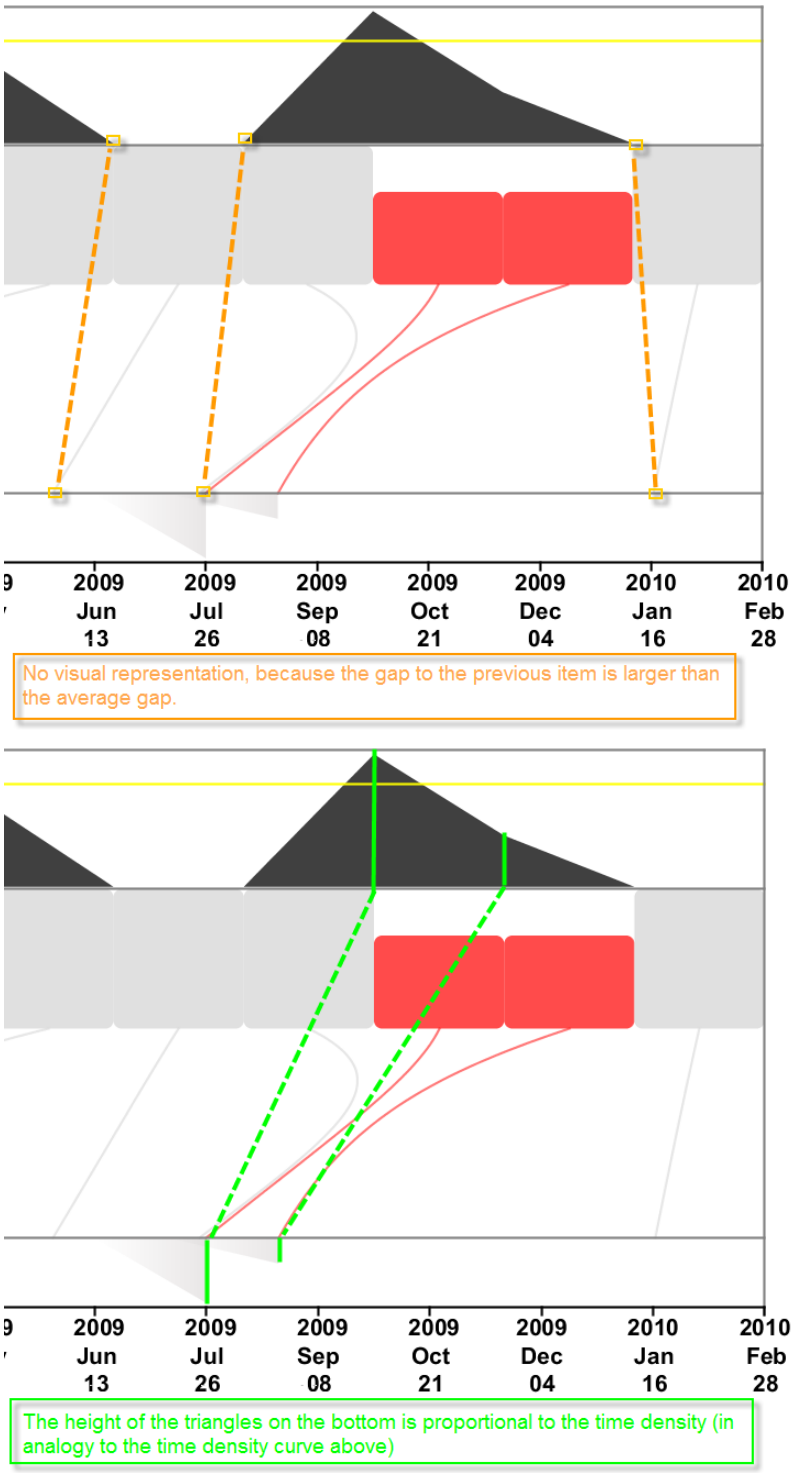


Figure 6.6: Example of the new visual representation with triangles (on the bottom). In order to provide a better understanding we coordinate the triangle visualization with the time density plots introduced in Chapter 5.2. The height of a triangle depends on the time density, i.e. the relative time distance to the previous item.

time.

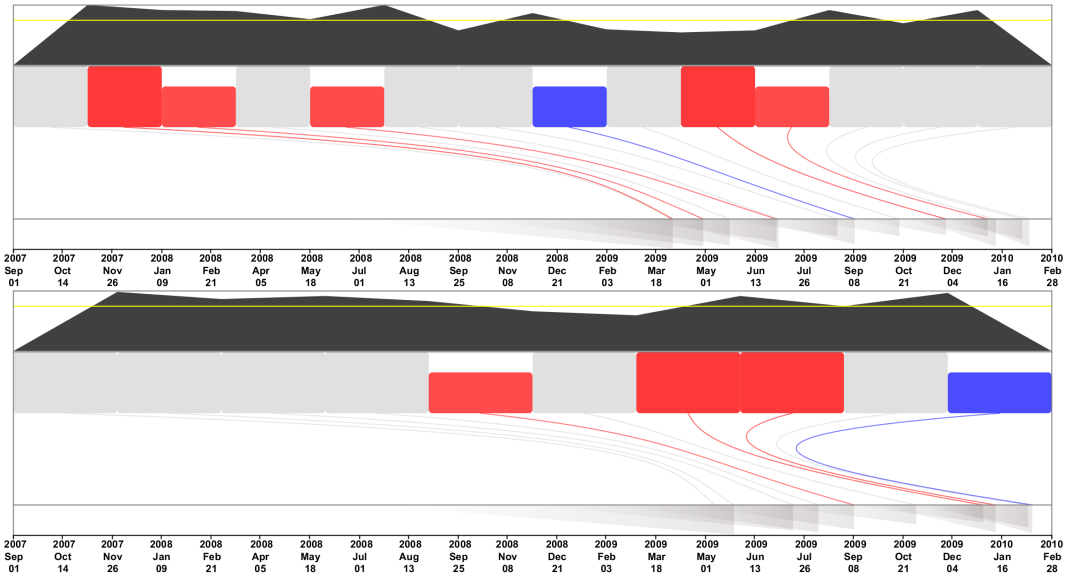


Figure 6.7: Example for two cases (top: *air bag*, bottom: *air pack*) where the *avg_gap* and with it the width of triangles along the time line decreases over time. For illustration purposes we coordinate the triangle visualization with the time density plots introduced in Chapter 5.2.

Visual Aggregation: In this approach again we implement the concept of visual aggregation as introduced in Section 5.1. The triangle shape leads to an increased overlap, the closer two items are. In order to achieve an aggregation effect, triangles are plotted in a semi-transparent grey color and lose opacity towards the outer end. Consequently, the higher the overlap at a certain point in time, the stronger the color. If a triangle overlaps with several predecessor triangles its color hue shifts from black to red. The more triangles it overlaps with, the more red it becomes, see Figure 6.8 for examples. The strength of this effect depends on one parameter that can be defined by the user.

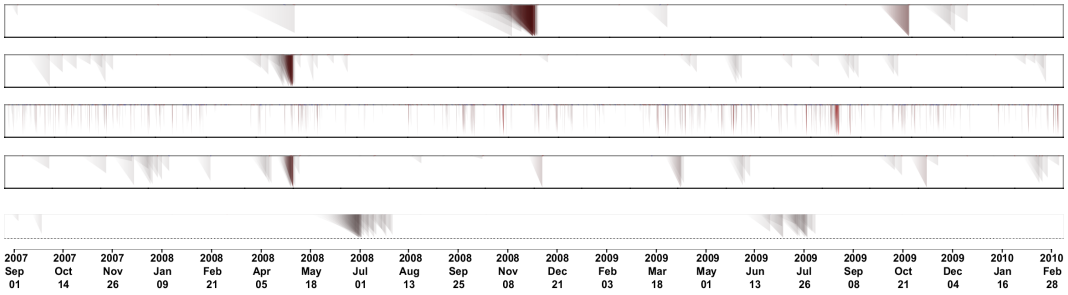


Figure 6.8: The top five issues detected within the customer feedback using the live scoring. From top to bottom they belong to the nouns *customs*, *packing list*, *coupon*, *packing slip*, and *july*.

On demand as user can additionally display a time density plot for a selected area along the time line or the whole time range as in the example of Figure 6.7.

Interactive Analysis Use Case

This use case on real data, the customer feedback stream, shall demonstrate how the incremental display is used to explore issues and enables analytical reasoning. The top pair of related event types, according to the content-based similarity measurement is *packing list* and *packing slip*, see Figure 6.11. The top pair of related event types, according to the co-occurrence-based similarity measurement is *survey* and *email*, see Figure 6.9.

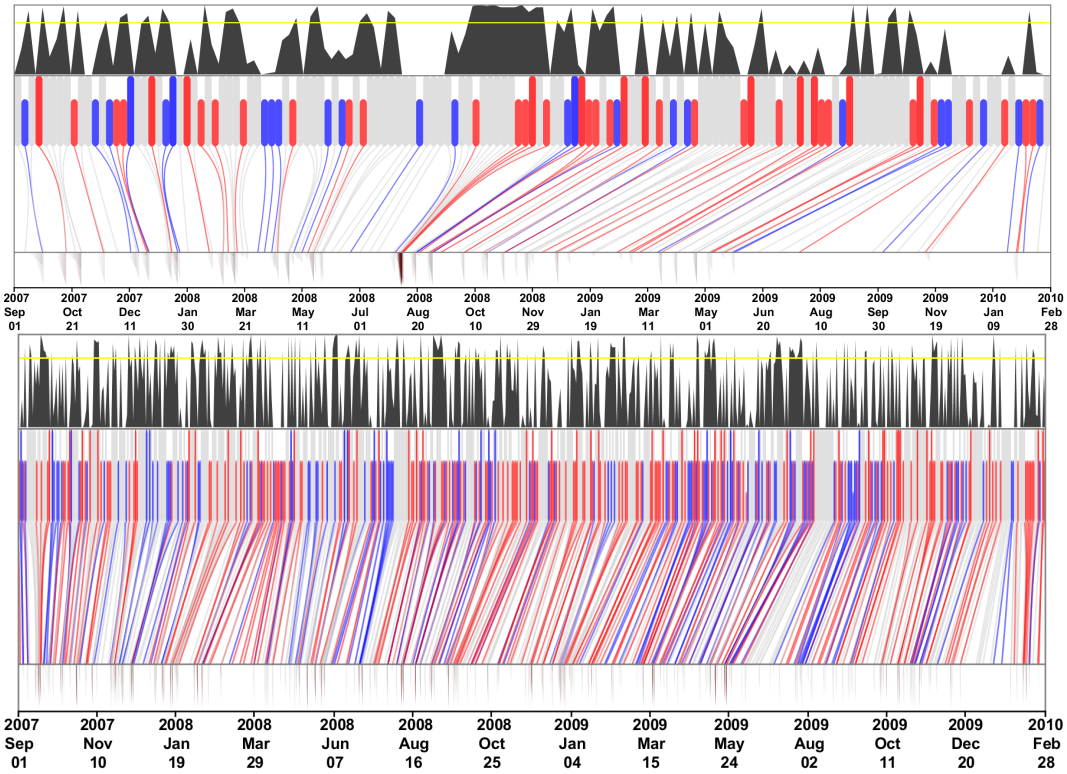


Figure 6.9: The two event types for *survey* and *email* aligned with a timeline. Customers complain about receiving several emails with a survey invitation at once.

The topic modeling can be performed at any given point in time based on the documents kept in memory at that time point. For our examples we have picked two points in time where two event types have been automatically identified as being related. Table 6.2 shows the topics that are generated as soon as *survey* and *email* have been detected to be related.

Topic Rank	Topic Strength	Descriptive Terms
1	1.29	survey, to send, email, copy, product, to use
2	0.99	survey, email, store, printer, not higher, to spam

Table 6.2: The topics generated based on all documents active on October 19th, 2009.

The topics have been ranked according to how strongly they are represented in the data, which is revealed by a parameter that the topic modeling returns. Both topics contain *survey* and *email* and further terms which indicate that

customers felt *spammed*, because they were *sent copies* of the same email several times.

Table 6.3 shows the topics that are generated as soon as *packing list* and *packing slip* have been detected to be related.

Topic Rank	Topic Strength	Descriptive Terms
1	1.28	packing list, amount, order, next, wrong, to show, error
2	1.06	packing list, charge, day air, not to put, to show, same, not to buy
3	0.94	to charge, incorrect, packing list, order, problem, cost, packing slip

Table 6.3: The topics generated based on all documents active on May 7th, 2008.

May 7th was the first day when there were enough documents active that three topics were requested in the topic modeling phase. All three topics relate to the packing list issue, where topic three is the most specific one containing both *packing list* and *packing slip*. The further words of the topic point to the issue: Packing lists/slips of *orders* were wrong *charging* an *incorrect cost* which was a *problem* for the customers.

One interesting observation was that the number of documents used for topic modeling, and with it the number of topics, was also a quite good indicator on whether there was currently an ongoing issue. Only once during the monitoring 4 topics popped up, and that was during the *customs* issue already described in Section 5.2.

6.2.6 Case Studies

For both the customer feedback stream and the microblog stream a basic ground truth is available. The customer feedback stream used here is an extended version of the dataset used in Section 5.2. Here we had app. 87,000 feedback messages at our disposal extending the app. 50,000 we used previously and for which we are aware of 17 issues that should be detected. These issues were provided by a data manager in [144]. For the microblog stream some issues were provided as part of the VAST Challenge 2011 Mini Challenge

1 solution.[10] Actually, for the microblog messages of the VAST Challenge also geo-spatial coordinates were provided and certain issues only occur locally. In our case we ignore this additional information.

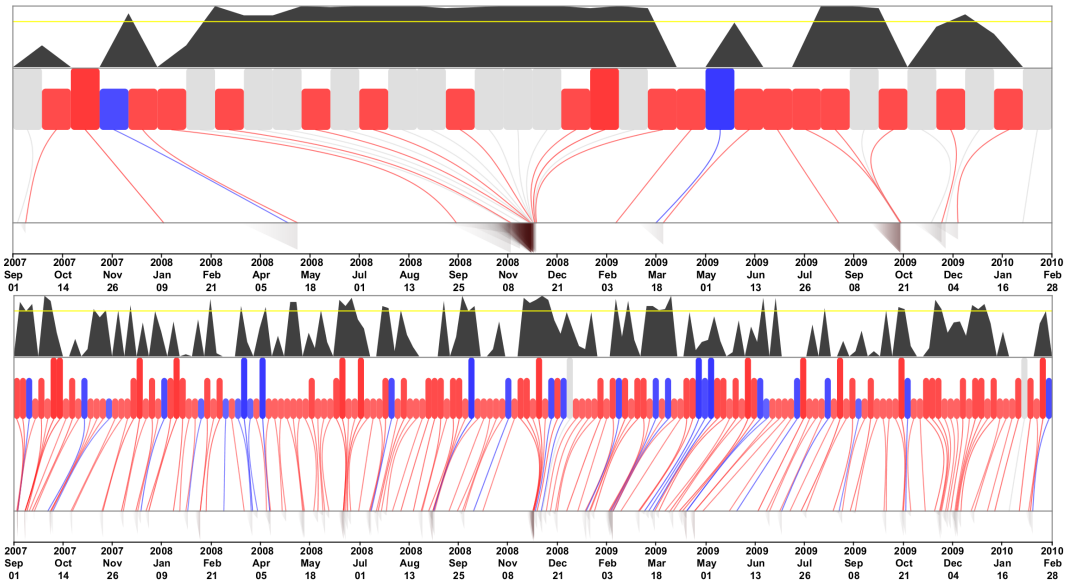


Figure 6.10: The event types *customs* (rank 1) and *delay* (rank 35) point to the same real-world issue. However, it is much more salient in the first case.

Customer Feedback Stream

For the evaluation of the event episode detection and scoring we saved all 140 event episodes detected during the simulated live monitoring of the stream. Figure 6.13 shows that the scores follow a long-tail distribution. Seven out of the ten top event episodes with the highest scores were meaningful and interesting. Some examples are shown in the Figures 6.10, 6.11, and 6.12.

Of course, even though an event episode might have a rather low rank among all episodes of the overall time range, it still might have been the top-scored episode within the specific time interval of its existence.

[10] http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/solution/index last revised on May 3rd, 2013.

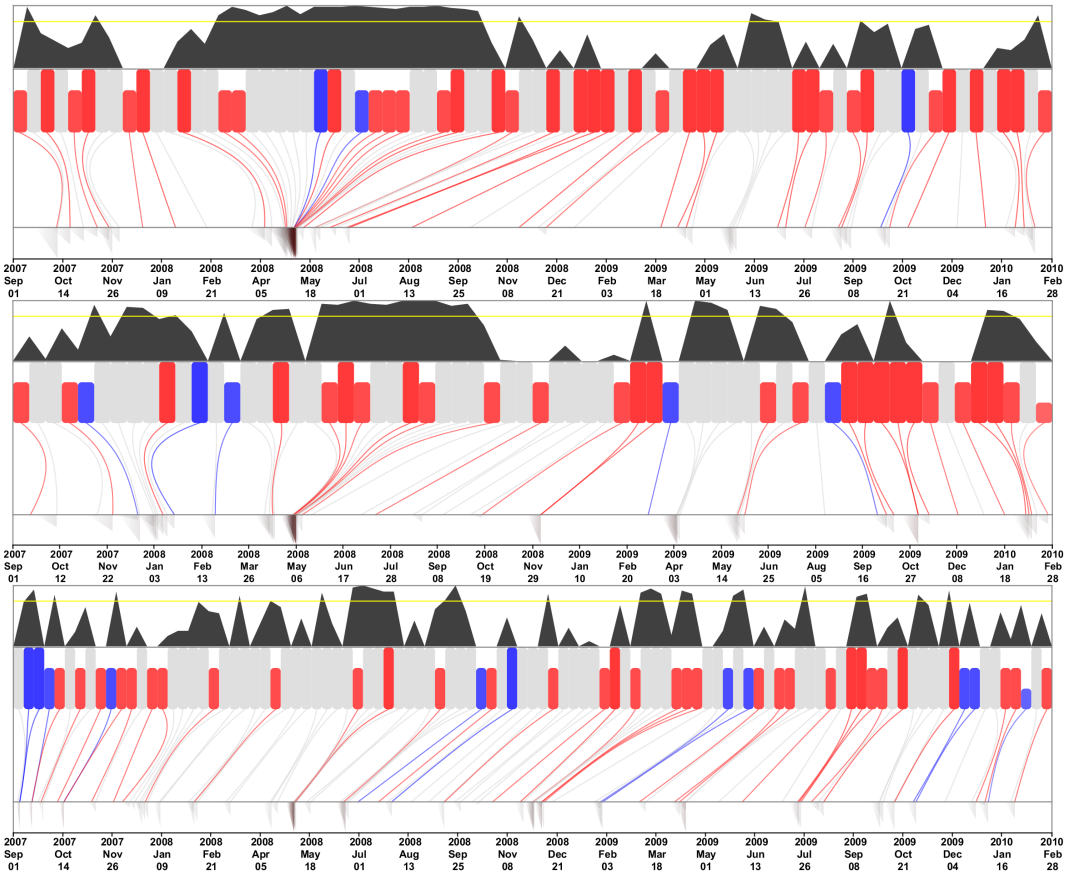


Figure 6.11: The event types *packing list* (rank 2), *packing slip* (rank 4), and *list* (rank 119) belong to the same real-world issue. However, it is much more salient in the first two cases.

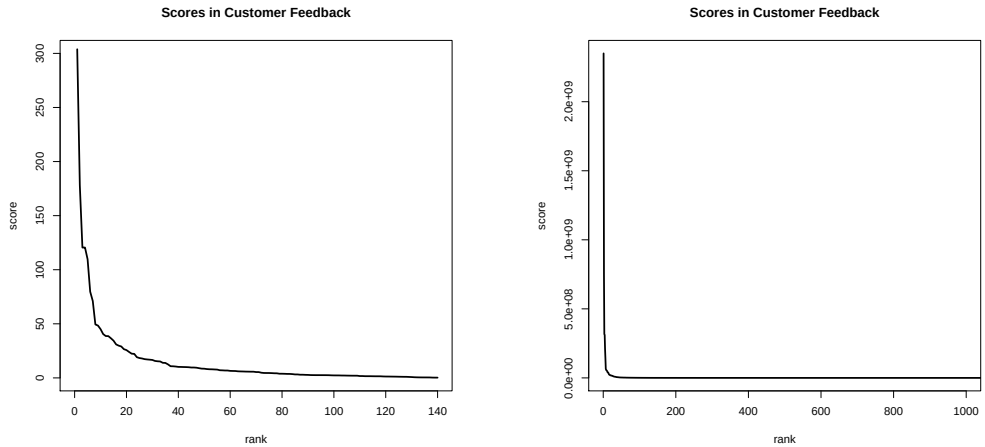


Figure 6.13: Distribution of event episode scores in the customer feedback stream (left) and the microblog stream (right).

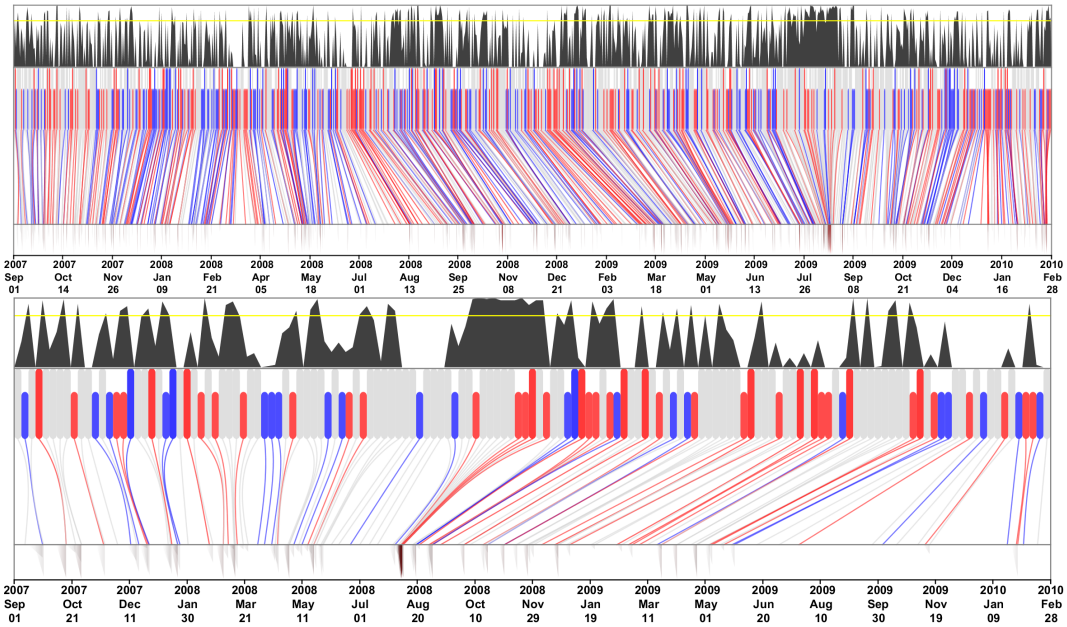


Figure 6.12: Event types pointing to issues: *coupon* (rank 3) and *survey* (rank 12).

Microblog stream

We took the dataset provided by the VAST Challenge 2011 Mini Challenge 1, which "contains microblog messages collected from various devices with GPS capabilities."[11], and goes from April 30th to May 20th 2011. All in all, 1.023.056 messages are contained having time stamps on minute-basis. The solution provided on the Challenge Web page contains the following ground truth events:

1. On May 17th, a truck accident occurs on the I-610 Bridge on the VAST River, which is the cause of a disease.

2. Flu-like symptoms disperse airborne on May 18th.

3. Gastrointestinal symptoms disperse waterborne on May 19th.

The distribution of the event episode scores for the dataset again is a long-tail distribution, see Figure 6.13. Almost all of the top issues relate to different sickness symptoms.

[11] http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/taskdesc/index last revised on June 24th, 2013

The top 15 highest scored event episodes detected with our real-time algorithm relate to the event types *flu*, *fever*, *plane*, *pneumonia*, *sleeeeeeeeep*, *sickness*, *chest pain*, *sweat*, *shock*, *temp*, *heartburn*, *sleeep*, *case*, *fatigue*, and *chill*. When plotting the event episodes over time (see Figures 6.14, 6.15, and 6.16)[12] it becomes evident that most of the top-scored episodes show up towards the end of the time period, from May 18th on, and indicate sickness symptoms. Directly before the disease outbreak, the *truck* accident sticks out. In addition, several further negative events can be seen, for example a *car accident*, a *plane crash*, a fire in the *capital building*, and a *bomb threat*. The different salient event types are visualized with the incremental visualization in Figure 6.17. Different data features can easily be recognized:

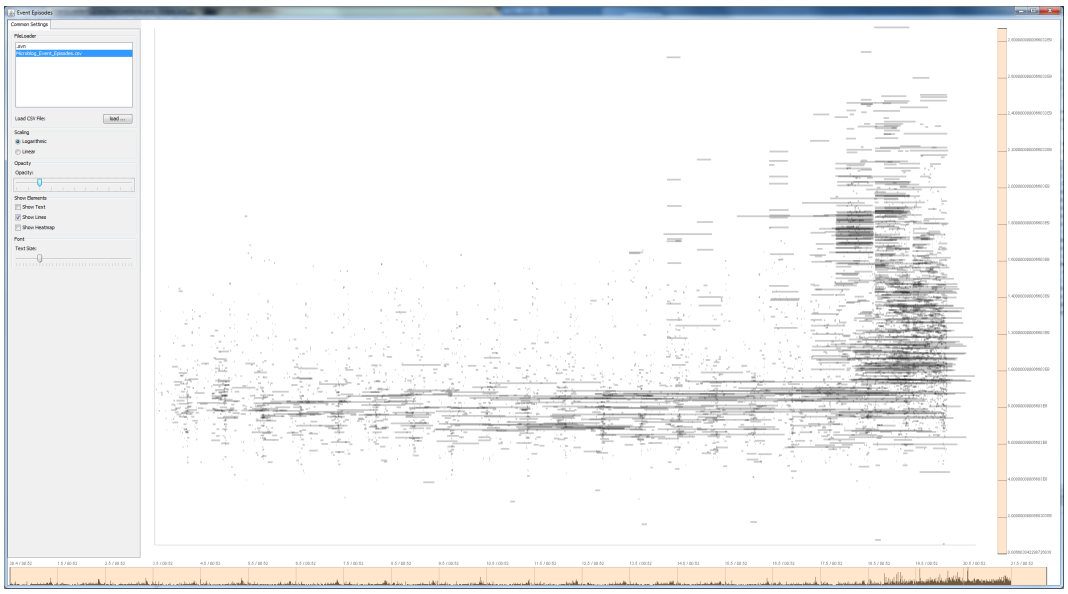


Figure 6.14: The event episodes are plotted as lines, where the x axis is the time axis and the y axis corresponds to the episode score with logarithmic scaling. Most event episodes have a comparatively low score and few event episodes stick out.

1. There is one periodically bursting word *apple*. This artifact in the data is most probably due to the data generation strategy of the contest organizers.

2. There is one repeatedly bursting words *car accident*, apparently there are several such accidents reported in the data.

[12] These visualizations were implemented by Dominik Jäckle.

3. Some aligned bursts indicate several events. First, a *plane crash* on may 13th (*plane*, *dirt*, *wasteland*, *dammage*[13]). Second, a fire on May 14th (*capital building*, *fire men*, *emergency service*, *hope noone*[14]). Third, a *bomb threat* on May 16th (*threat*, *rumor*, *bomb crew*, *bom*,[15] *city*, *drill*, *danger everyone*,[16]. Finally, a truck accident and explosion on May 17th (*destruction*, *truck*, *truck driver*, *shake*, *smog town*, *noon*).

4. After the disease outbreak the sequence of symptoms is clearly visible. First there are the flu-like symptoms like *fever*, *headache*, *sweat* and *fatigue* and later the gastrointestinal symptoms like *diarrhea*.

We did not participate in the VAST Challenge because it had already passed, when we performed the analysis. Yet, all three ground truth events, the truck accident and the two waves of diseases, could be detected with our methodology. This is remarkable because we did not include the geo-coordinates of the microblog messages in our analysis. Exploiting the additional geo-spatial hints would have facilitated the detection of events.

[13]The repeated use of this typo is probably an artifact due to the data generation strategy of the contest organizers.

[14]Part of the frequent phrase "hope noone got hurt" wrongly identified as a compound noun in the preprocessing

[15]From the context it becomes clear that the word should be *bomb*. The repeated use of this typo is probably an artifact due to the data generation strategy of the contest organizers.

[16]Wrongly identified as a compound noun in the preprocessing

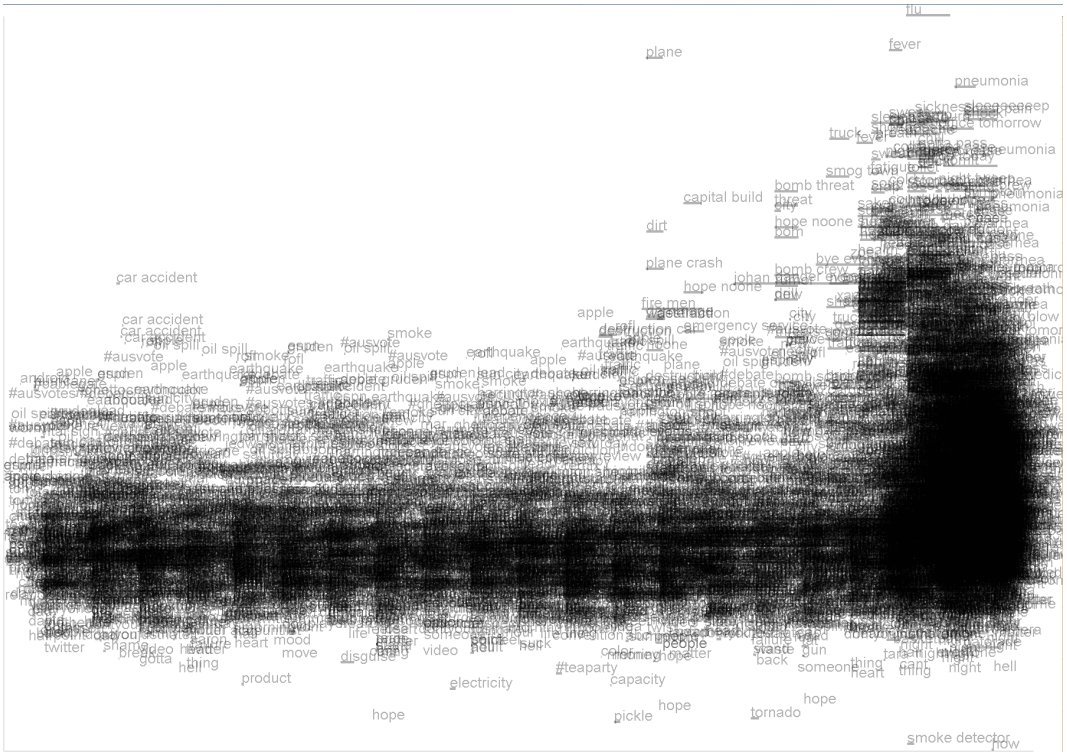


Figure 6.15: Enlarged view of the episodes from Figure 6.14. The event types are provided as labels.

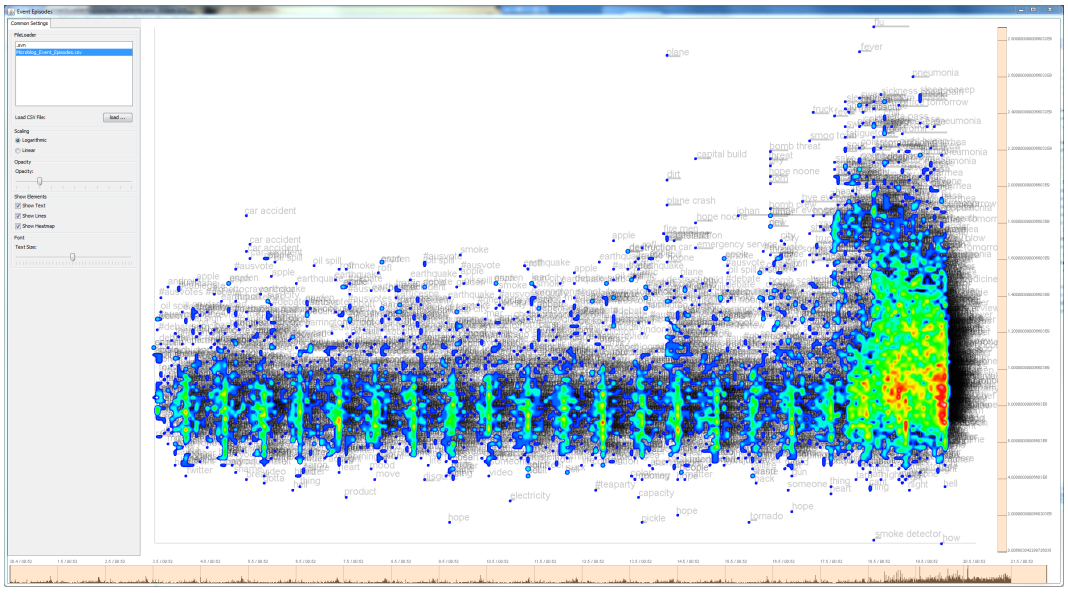


Figure 6.16: The event episodes are plotted as lines, where the x axis is the time axis and the y axis corresponds to the episode score with logarithmic scaling. A heatmap shows the density of event episodes. Daily peaks can be seen as well as the episode outbreak.

Figure 6.17: The salient issues from Figure 6.15 using the incremental triangle plots. The top 25 episodes ending before the disease outbreak and the top 25 episodes ending after the disease outbreak are shown. After the disease outbreak obviously the scores are much higher.

6.2.7 Performance Evaluation

For the evaluation we use historical records of both the customer feedback stream (about 87.000 messages over more than 2 years) and the microblog stream (more than 1.000.0000 messages over 21 days) mentioned earlier. While for the customer feedback the available time stamps have only day resolution, for the microblogs we have the exact minute of publication. Both streams are simulated based on the historical data, therefore we complement the customer feedback dates with random hours and minutes in order to make it more realistic. It has to be remarked that the tested datasets can be streamed through our current implementation on a single node computer, orders of magnitude faster than real-time. With some optimizations the methodology is able to process the 10% Twitter Stream (more than on million tweets per hour) in real-time on a single workspace computer (see [82, 98]). For evaluation we first analyze in detail the storage requirements of the real-time detection and scoring of event episodes in Section 6.2.7.

Storage Requirements

One important aspect of the live analysis is that only the necessary data will be stored to be readily accessible. In our case we have to store both the event types (nouns) and the content words (nouns, adjectives, verbs) together with a global counter for each of them. For each event type, we have to store in addition an average gap (here: $incr\_avg\_gap$) and the time stamp, when it was last seen. In order to lower the storage costs we do not consider every noun to be an event type. First, we filter out those nouns that have less than 3 characters or contain punctuation marks within their string. From all nouns that pass the noise filter we consider only those that have appeared at least twice during the monitoring process. That is, from its second appearance on a noun is an event type. Apart from that, at any point in time, only those documents have to be stored that do or potentially still could contribute to an event episode. A document contains full text, a bag of content words, a list of event types with sentiments, and a time stamp. Thus, documents are the most expensive objects to be stored.
We can show for the two datasets used that the number of documents to be

stored is constantly low and the number of event types and content words to be stored grows in a sublinear manner with respect to the number of documents processed (see Figure 6.18). While the number of different nouns initially grows strongly, adjectives, verbs, and event types show only a very limited increase over time. It is remarkable that in the case of the microblog data the number of different nouns grows more heavily. We could observe that this is due to the fact that the part-of-speech tagger tends to assign the tag *noun* to many noisy strings. The fact that the number of event types grows only very slowly reveals that most of the strings tagged as *noun* either are deleted by our noise filter or only occur once.

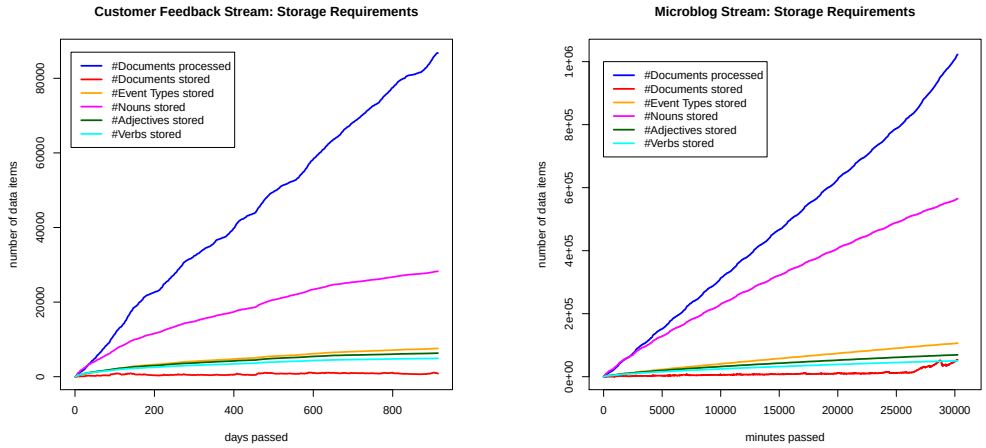


Figure 6.18: Detection and scoring of event episodes in the customer feedback stream (left) and in the microblog stream (right): The number of documents processed over time in comparison to the number of data items stored. In this case the gap between two events of the same event type had to be smaller than avg_gap in order to form an event episode, which is the least strict criterion.

6.2.8 Discussion and Conclusion

In this section we have demonstrated that the detection and visual analysis of event episodes of word occurrences is a practicable and beneficial way of monitoring document streams in real-time. We provided empirical evidence for the usefulness of our methods applying them to different data sources and different analysis tasks. We showed that detected event episodes point to interesting findings and sometimes even critical issues in different domains such

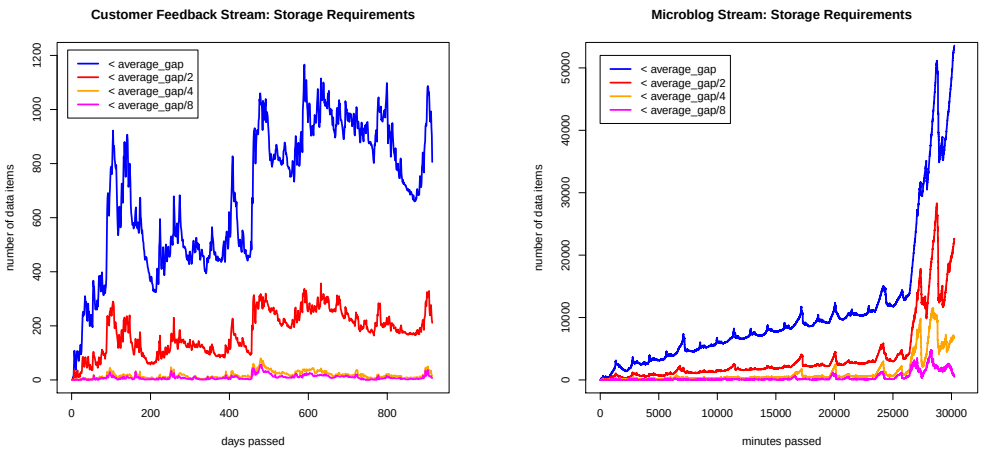


Figure 6.19: Detection and scoring of event episodes in the customer feedback stream (left) and in the microblog stream (right): The number of documents stored at any point in time depending on different thresholds. The time gap between two events of the same event type had to be smaller than avg_gap, $\frac{1}{2} \cdot avg\_gap$, $\frac{1}{4} \cdot avg\_gap$, and $\frac{1}{8} \cdot avg\_gap$ in order to form an event episode.

as customer feedback, or microblogs. The automatic detection and scoring of event episodes works in an incremental manner and thus enables a real-time processing of large text streams with very limited storage requirements. Almost all issues that we had in our ground truths could be found skimming through a relatively small number of top-scored event episodes. However, also some meaningless event episodes received high scores. That means, while we barely miss issues, the human analyst is still required to explore a certain amount of meaningless episodes before s/he can dismiss them. In our future work we aim to improve the scoring with respect to the precision. We were able to show event episode patterns for event types of the whole frequency spectrum. While some event types are very descriptive by themselves, like *car accident* or *bomb threat*, others have to be made sense of providing context through topic modeling and interactive exploration. Our approach readily enables the analysis of issues both with respect to their temporal context and topical context and brings the intelligent interactive exploration of critical issues in document streams to real-time. Another challenge for future work is to design a whole visual analytics system to support analysts in achieving situational awareness. The presented algorithms for text stream processing and event episode detection, with their capability of information filtering and

structuring, are a good foundation for such a system. The incremental triangle visualization with the integration of time density plots, however, so far is a stand-alone prototype implementation and will have to be complemented with further visualizations. While some of the automatic processing challenges, outlined in Section 6.1, are partly solved by our approach, on the visualization side there is yet a lot of further research to do.

Chapter 7

Concluding Remarks and Perspectives

Contents

In this final chapter I would like to sum up the contents of this thesis (Section 7.1), discuss contributions as well as strong and weak points (Section 7.2), and finally give a conclusion and an outlook on future research topics (Section 7.3).

7.1 Summary

This thesis presents novel **visual analytics** methods for the detection and understanding of diverse phenomena of **change** that can be observed either within **natural language text** or based on it.

In Chapter 1 I motivated the overall topic of this thesis. In Chapter 2 I described the role that visual analytics methods currently play in linguistic research and in time-oriented text mining. In both cases I give an overview of

the state of the art, open research issues, the as yet untapped potential for the integration of visual analytics into linguistic research methods, and the role and goals of this thesis within the respective research fields.

The first part of the thesis, Chapter 3 and 4, deals with visual data analysis for linguistic research. In these chapters, the thesis aims to provide a broad range of examples on how visual analytics can support different linguistic tasks. The approaches have been designed in close collaboration with domain experts in order to assure that they are effective and easy-to-understand for linguists. While the automatic text processing steps are mostly innovative, it turned out that in some cases for linguistic research, visualizations that had already been proposed could be used. In other cases novel visualizations and novel combinations of established visualization concepts were designed. The technical contribution of this thesis thus lies mainly in the design of the whole analysis process, which transforms raw data into actionable knowledge in form of easy-to-interpret visualizations. A special challenge here is the communication and collaboration across domains, which initially requires much time and effort. Thus, beyond providing technical contributions, the goal is to connect two disciplines, linguistics and visual analytics, which so far have neither had a tight integration nor shared a joint methodology. The challenge addressed in this thesis is to investigate which subfields of linguistics can profit from visual analytics, how such interdisciplinary research can become successful, and of course to provide examples of novel useful visual analytics methods that support the linguist's tasks. The main contribution here is that I do interdisciplinary pioneering work, provide role models for interdisciplinary research, and shape and advance a new research field.

The second part of the thesis, Chapter 5 and 6, deals with the analysis of time-oriented text data. In contrast to the first part of the thesis, this is a field where quite a bit of research has been conducted over the last few years. Here, the content of this thesis is focused on filling a gap in the state-of-the-art: Visual Analytics of unexpected changes and anomalies in text content, mostly related to target word distributions over time as well as target word contexts and sentiments. The corresponding chapters contain innovative methods both for the automatic and visual analysis, with a high technical depth. My research has contributed to the filing of seven patents, one out of which has

been issued to date. Among the innovations in text mining, one example is the detection of term bursts in arbitrary, not predefined time intervals. The novel concept of time-density based data modeling makes the method readily applicable to terms of different overall frequencies. An example for the innovations in visualization is the integration of a sequence view with a linear time line, which enables a detailed analysis on record level. In addition, the demonstrated real-time capability of the methods goes far beyond what other previous approaches are able to offer.

7.2 Discussion

The number of visual variables and primitives is limited and the space of options for meaningful combinations is limited as well. Visualization techniques with a big impact that have become part of the standard repertoire of visualizers, such as for example Parallel Coordinates [84] and Treemaps [87], are often innovations that date back further in time. During my intensive collaboration with other researchers in very diverse visual analytics projects, I could observe that nowadays innovation is mostly triggered by choosing innovative mappings from data to established visualization concepts, combining existing visualizations in intelligent ways, and most importantly, by facing the challenges of bringing visual analytics to new application domains. In contrast to the constant number of design options, the number of new data sources, the overall amount of data, and the number people that could benefit from visual analytics methods in their daily work and life is constantly growing in our knowledge-based societies and economies. This growing demand requires visual analytics researchers to solve new data processing challenges and come up with new task-driven designs. In any such system it is important to leave space for the individual customization of analysis displays and to offer different options for adapting the original analytic goal in order to solve follow-up tasks. In a sense, the challenges are to extract the important variables and structures out of the raw data and to create visualizations that help to solve a specific set of specialized task, while being general enough to foster the creativity of the analyst during the analytic process.

For such a complex data type as natural language, it is especially challenging

to process and parse the raw data, extract actionable knowledge, and prepare it for visualization and thus for human exploration.

This thesis introduces quite a number of innovative novel options for extracting both linguistic and business intelligence knowledge from text data as part of new visual analytics methods and systems. In most of the presented approaches complex models are derived from the data or complex statistical analyses are performed in order to extract features and measures that describe abstract phenomena, similar to what Oelke named *quasi-semantic properties* [90, 134]. Sometimes the visualization just represents the automatically computed patterns/results, as in the exploration of word sense developments in Section 4.1. In other cases, the data processing is very basic and it is the visualization that makes interesting patterns emerge as for example in the RSS feeds analysis in Section 5.1.

The visualizations in this thesis are highly data and task-oriented and the mapping of data attributes to visual attributes is done in a way that fosters the emergence of visual patterns that in turn will point the analyst to patterns in the data. In contrast to many approaches from the field of information visualization, the focus here is not so much on complex layouting strategies and aesthetics, but has a more pragmatical task-driven emphasis. Several layouting strategies, however, have proven to be very beneficial in different parts of the thesis. For example, the meaningful sorting and ordering of data points and data attributes. As shown in the different case studies of Chapter 3 many patterns and peculiarities become only visible, when an appropriate sorting strategy has been applied. In other cases like the time density edge bundling in Section 6.2 a careful layouting fosters the emergence of visual structures. In this thesis, the goal is to enable analysts to quickly detect visual patterns and peculiarities, to understand why a certain visual pattern has emerged, and to judge whether this could just be an artifact or truly relevant. The key point is not to overload analysts by providing too much information in one visual display or too complex visual mappings, hindering the detection and interpretation of patterns. At the same time, it is important to grant interactive access to the underlying data and offer easy-to-understand options for manipulating it. While interactive design is a key component of most visual analytics systems, it is especially relevant when dealing with such complex and ambiguous

data as natural language texts. As the accuracy of automated data processing steps in many cases is rather limited, in the end, the human analyst has to be enabled to verify the meaningfulness of visual peculiarities and back up or reject hypotheses that are triggered through the visual representation.
It also turned out to be quite important to keep visualizations general enough to let analysts reason for which of their further tasks the same visualization could also be applied beneficially. Domain experts know their tasks best and sometimes become inspired by one method, like in Section 3.2, where the matrix display was applied to several different tasks. After having been created for the analysis of vowel patterns, domain experts came up with the idea to test the methodology on different data sources for different tasks and actually could make interesting findings.

Two additional points that I would like to discuss separately are lessons I learned about interdisciplinary visual analytics research in the course of preparing this thesis (see Section 7.2.1) and the role of evaluation (see Section 6.2.7).

7.2.1 Interdisciplinary Visual Analytics Research

An important part of this thesis was the collaboration with domain experts. For a successful interdisciplinary research it is important that both visual analytics researchers and domain experts develop a mutual understanding of the basic terminology, scope, and best practices in both of their research fields. For the development of a novel visual analytics solution to domain experts' research tasks and problems it is necessary to iteratively elaborate prototypes and intermediate solutions. I would name this an iterative dialog-based interdisciplinary development process. Starting from first preliminary visualizations both domain experts and visual analytics researchers will gain a better understanding about both the data and about the points which are relevant for their counterpart. The fact that both can jointly look at a visualization, or an interactive system resulting in a visualization, supports the dialog between them. The abstract conversation about data, phenomena, and analyses, becomes much more straightforward when it can be centered around and externalization consisting in concrete visual instances. If this dialog goes back and forth, solutions keep on refining and become more elaborate. Sometimes

it may even turn out that a visualization inspires the domain expert to ask for a solution for another different task as s/he started understanding the space of possibilities. At the same time, as a data analysis and visualization expert starts understanding the domain, with its terminology, research approaches, and constraints, s/he will be able to become a lot more targeted in her/his work. That is at least the experience I have made within the process of doing the research presented in this thesis. I am convinced that becoming aware of the necessity for such an iterative dialog-based development process and pursuing strategies to formalize it may be highly beneficial for interdisciplinary visual analytics research. In my opinion, this is especially true for collaborations with research fields that have not been closely linked to visual analytics in the past, such as humanities. Joint research in this area is currently fostered within the upcoming field of *digital* or *enhanced humanities*, where I envision visualization to play an important role.

7.2.2 Evaluation

It is known that systems for visual analytics and information visualization are hard to evaluate properly because it is usually not possible to quantify the knowledge gain obtained through the use of a visual analytics system and it is often not obvious how meaningful benchmarks should be created. A visual analytics system is usually the result of a number of interdependent design decisions, each of which has alternatives and can be argued about. At the same time, it is typically infeasible to properly insulate single design decisions in order to evaluate them separately. The usefulness and meaningfulness of a component, in the end, has to be judged in the context of the whole system and in relation to the tasks that should be solved with this system. Case studies and use cases can provide empirical evidence for the good performance of a system and user studies may back up single design decisions when regarded as insulated parts, as for example in [57]. The lack of tangible objective evaluations of a whole system, however, is a weakness that most of the application oriented research in information visualization and visual analytics shares and that can not be easily overcome. As an attempt to provide a way to validate visual analytics systems as a whole, a systematic nested four-level model for visualization design and evaluation was suggested by Munzner [125].

According to Munzner, the points to be addressed are validations of (a) the "domain problem and data characterization", (b) the "operation and data type abstraction", (c) the "visual encoding and interaction design" and (d) the "algorithm design". In this thesis all of these points are addressed in order to give evidence for the good quality of the suggested solutions:

(a) Domain problem and data characterization: The domain of the target audience, their data, and tasks have to be clearly understood by the designer of the system.
As the approaches presented in this thesis are a product of a close cooperation between domain experts from linguistics or business intelligence and me, I spent a lot of time with the domain experts discussing their data, tasks, and iterations of possible solutions. Thus, I consider this requirement to be most widely fulfilled.

(b) Operation and data type abstraction: the right data type has to be derived from the initial data set so that it can be addressed by visualization techniques and general operations have to be defined that shall help to solve the specific tasks of the domain.
As in most cases text data cannot be directly visualized for analysis, the data type abstraction is important in this thesis and thoroughly discussed within the sections describing approaches. Concerning operations, rather generic tasks were addressed. The user shall be enabled to find correlations, patterns, and anomalies in the data, inspect possible causes, and generate and validate hypotheses. In addition, the special requirements of text analysis tasks were considered: The analyst has to be enabled to drill down to the original text sources in order to achieve a deeper understanding and certainty for sense-making or decision-making.

(c) Visual encoding and interaction design: the visualization has to be effective in conveying abstract features of the phenomena under investigation. In order to demonstrate that our solutions fulfill this requirement, we performed extensive case studies as described in the respective sections. In some cases the visual encoding was straightforward and in other cases I thoroughly discussed the reasoning behind the choices, as for example in Section 3.1.

(d) Algorithm design: the underlying algorithms must "carry out the visual encoding and interaction designs automatically" [125] and have to be acceptable in runtime and memory complexity. In the course of our experiments, we did not encounter any related problems. The computing of all of the approaches presented in this thesis works fully automatically. The approaches presented in Chapter 3, 4, and 5 in almost all cases have *not* been optimized for runtime and memory performance, as the implemented research prototypes work in interactive time as they are, and the datasets are small enough to keep them in the main memory. One exception is the sorting of languages within the extended Sunburst display of Section 3.1. Here, I implemented an optimized algorithm suggested in the literature. A different case is also Chapter 6, where the algorithms have been conceptually designed and optimized for real-time processing. A possible enhancement that I did not investigate so far is the extension of the algorithms for parallel processing.

Apart from the evaluation of visual analytics systems as a whole, of course, often there are some design decisions to be made that could potentially be evaluated individually. Especially in the Sections 5.2 and 6.2 we did a lot of informal trial-and-error evaluations in order to select one out of different options, e.g. for parameter settings. However, in most cases it was not possible to define an optimal setting, because the analysis results were of a semantic nature, that is they were not measurable formally, but had to be inspected manually. Sometimes, it became quite obvious to us that one solution was better than the other even though it would have been hard to formally proof this impression. Inspecting solutions manually also limits the number of tests that can be made. In some cases user studies might have helped, but would have required resources that go beyond the scope of this thesis. Possibly, some interesting aspects might be worth digging into more carefully in future work.

7.3 Conclusion & Perspectives

The novel methods and concepts introduced within this thesis contain several contributions to the fields of visual analytics for linguistic research and visual analytics of time-oriented data. Several new lines of research have begun to be

explored, for some of which this thesis can be considered as just an initial but groundbreaking step. In particular, the first part (Chapter 3 and 4) aims at demonstrating the opportunities that exist for developing new visual analytics methods and systems in order to support linguistic research. In future work, more in-depth research should complement the content of this thesis, which I would like to discuss separately for the different chapters:

Traces of Change: Cross-Linguistic Visual Analytics for Language Comparison: The *world's languages explorer* (Section 3.1) is a promising system, that is already in use by different researchers from linguistics. I hope to get feedback on the usability and potential improvements, before the system will be released as open-source software to the public. For example, there is potential for improving the geo-spatial display, which is a data type that is not in the focus of this thesis.
The visualization of sequences is an important task in visual analytics of linguistic data, because sequences play a role at different levels of natural language. The *matrix display* presented in Section 3.2 is a first step towards the exploration of binary sequences. One charming advantage of this method is that the *absence* of data, for example an unexpectedly low frequency of a certain sequence, receives an explicit visual representation. This is important for the analysis and often not done in visualization. When representing only *present* (in contrast to absent) data, longer sequences can also be displayed with existing methods as shown in the case study using the *droplet maps*. For languages with clear patterns this method is quite insightful, for other languages, however, the display becomes cluttered.

Visual Analytics of Diachronic Change in Lexical Semantics: The usage of topic models on word contexts for word sense induction from diachronic corpora is a small, but significant contribution. In contrast to previous related approaches we do not only enable the discovery of changes in word meaning, but also reveal concrete word senses. The visualization shows how the senses have changed over time and is especially useful to detect the emergence of new senses. As the research on topic models is a very dynamic field, it is to be ex-

pected that some drawbacks of our method can soon be overcome using novel advanced topic modeling methods. For example, while we had to manually predefine a number of topics to be retrieved, novel methods will potentially be able to estimate this parameter as well as further parameters of the method automatically in order to compute an optimized model. Hierarchical topic models might also be useful to give insight into the relations between different topics.
The investigation on the *-gate coinages* contains no clear-cut technical contribution, but is a good example on how the exploration of massive data with easy-to-understand visualizations can give an overview on the spread of a linguistic phenomenon that is otherwise hard to grasp. It could be revealed that the coinage and usage of words with *-gate* suffix seems to be a piece of language change that is also rhetorically motivated. With its pragmatic component the *-gate* suffix is an example for the relation that exists between the analysis of linguistic phenomena and the analysis of text content and sentiments, i.e. the relation between the first and second part of this thesis.

Visual Analytics of Diachronic Change in Text Content: The research in this chapter, especially in Section 5.2, has a high technical depth and contains contributions both with respect to the automatic and visual exploration. We could demonstrate that the innovations are useful in real-world business intelligence analyses for customer relationship management. With the help of the provided system domain experts were able to identify time-related issues in customer complaints that they had not been aware of before and that would have been very hard to find even when reading large sets of comments and performing standard text mining. We introduced a novel way of representing item-based time series integrating a sequence view with a time density display and a linear time line. The automatic detection of critical time-related issues in text documents was tailored to the application in target-based sentiment analysis. Different features were taken into account, like the time density, context coherence, sentiment, or the certainty of the analysis. While each of these components has only a limited expressiveness, the combination yields good results. The more components that are involved, the less error-prone and more

accurate becomes the analysis. In the future work, it would be interesting to experiment with further components and adapt the text analysis to the detection of more general events in other domains and tasks.
Section 5.3 shows that the exploration of term associations gives further insight. A challenge for future research will be to come up with methods that enable the exploration of developments of term associations over time, with the goal to generate an analytical added value. Taking into consideration that a certain minimal amount of data is needed in order to derive meaningful statistics about associations, and that the results also heavily depend on the choice of suitable time intervals for which to investigate associations, this is a challenging issue for future research. A further direction that has already been started to explore in two initial publications, is that of assessing and visualizing geo-term-associations and patterns of unexpected coincidences in text, time, and geo-space. This area offers an enormous potential for future research with relevance for different real-world applications.

Real-time Analytics and Visualization of Critical Event Episodes in Document Streams: This chapter has a conceptual contribution in Section 6.1, where different applications, challenges, and open issues for the real-time analytics and visualization of text document streams are extensively discussed. In Section 6.2 I extend methods from Section 5.2 in order to be applicable for real-time analysis. For the automatic detection I show that at any point in time only a limited amount of data has to be stored in main memory. I also suggest a novel incremental item-based visualization that integrates the *time density plots* from Section 5.2. However, so far this visualizations stands on its own and is not yet embedded into a system for real-time analysis where it would be coordinated with further views. For sure, this would be an important next step to give analysts further context and insight. How this could be achieved is an open issue that involves several of the raised challenges about dynamic visualizations for real-time sense-making. For sure, there is a potential to build on the output of the relevance-based topic modeling that was suggested.

All in all, the different lines of research suggested and pursued in this thesis offer a wide range of possible extensions and have the potential to - and hopefully will - inspire future research in the described subfields of visual analytics. I invite the interested reader to visit my publicly accessible Google Scholar profile[1] in the future and see whether such extensions will have been realized as part of my own future work or by others citing my research.

[1] http://scholar.google.de/citations?user=WComJHIAAAAJ&hl=de&oi=ao last revised on February 15th

Appendix

New Coinages with Suffix -gate List of the app. 700 supposedly new coinages with *-gate* suffix as extracted from app. 11 million online news articles in English, French, and German, used in Section 4.2.5.

Abbas-gate, Adamugate, Afghan-gate, Africagate, Agliottigate, Aid-gate, Airportgate, Alicante-gate, Alinghigate, Altai-gate, Altargate, Alugate, Alu-gate, Amazonasgate, Amazongate, Amosgate, Anelka-gate, Angolagate, Angola-gate, Angologate, Antennagate, Antenna-gate, Antennegate, Apple-gate, Apprentice-Gate, Apuestagate, Arrivalsgate, Arsmgate, Asiagate, Asia-gate, Assange-Gate, Atomgate, Babygate, Baligate, Ballgate, Ballsgate, Bananagate, Bandargate, Bannergate, Bari-gate, Batterygate, Battery-gate, Beckgate, Bee-gate, Bees-gate, Belenegate, Bench-gate, Berlingate, Bertiegate, Betsygate, Bettencourtgate, Bettencourt-Gate, Biffogate, Bigotgate, Bigot-gate, Billinsgate, Biscuitgate, Biscuit-gate, Bittergate, Blackberry-Gate, Blackjack-gate, Blackoutgate, Blackwatergate, Blattergate, Bloggergate, Bloodgate, Blue-gate, Bondage-gate, Bonusgate, Boobgate, Boob-Gate, Boo-Gate, Boozegate, Bostitch-Gate, Bottomgate, Boubagate, Boulogne-gate, Bourgigate, Bra-Gate, Breadgate, Breakfast-gate, Bribery-gate, Bridgegate, Broad-gate, Brook-gate, Browgate, Bruneigate, Buggygate, Bullygate, Buloggate, Bulog-gate, Bumpgate, Bunkergate, Butlergate, Buttongate, Buwog-Gate, Cablegate, Cable-gate, Cablegate-Gate, Caddie-gate, Caddygate, Cadmangate, Caldergate, Callistagate, Camerongate, Camillagate, Camilla-gate, Cannonsgate, Cargate, Carpetgate, Cashgate, Casinogate, Casino-gate, Casoria-Gate, Castle-gate, Catgate, Catgate, Cattlegate, Cementgate, Census-gate, Centralgate, Centurygate, Chaingate, Champagnegate, Cheriegate, Cherie-gate, Cherylgate, Chickengate, Chinagate, Chogm-gate, Choppergate, Christalmightygate, Cimategate, Cingapuragate, Cleavagate, Clementgate, Climagate, Climategate, Climate-gate, Climatgate, Coconutgate, Coffingate, Coingate, Colagate, Contragate, Coptergate, Copygate, Copy-Gate, Corfugate, Corfu-gate, Corngate, Cougar-gate, Cough-Gate, Cowengate, Crashergate, Crashgate, Crash-Gate, Cretingate, Cristianogate, Crotchgate, Crouchgate, Crowngate, Crown-gate, Cyclonegate, Dategate, Defragate, Deutschbankgate, Diarygate, Dijongate, Dikko-gate, Dildo-Gate, Dirndlgate, Disastergate, Dogbegate, Dollygate, Donnygate, Donorgate, Donygate, Dreamergate, Dubaigate, Dubai-Gate, Dudusgate, Duffygate, Duffy-Gate, Dunkgate, Dwarfgate, Edisongate, Elmo-gate, Elyséegate, Emailtheftgate, Embassy-gate, Emmygate, Erenicegate, Ericsson-gate, Escortgate, Esseku-gate, Ettehgate, Everton-gate, Expensegate, Expensesgate, Eyafjöllgate, Facebook-Gate, Factory-Gate, Fag-Gate, Fakelakegate, Fal-

congate, Fatahgate, Fatah-Gate, Fear-Gate, Fergiegate, Filegate, Fingergate, Finger-Gate, Fishergate, Fishersgate, Fivebargate, Flaggate, Flag-gate, Flakegate, Floating-gate, Fluoridegate, Footballgate, Foreclosure-gate, Fornigate, Fortisgate, Foxgate, Frock-gate, Froyogate, Fruitbatgate, Gabblegate, Gaga-gate, Galantgate, Gallagher-gate, Gamu-gate, Garbage-gate, Garglegate, Gargle-gate, Gatecrashgate, Gatesgate, Gecko-gate, Gelsen-Gate, Genèvegate, Genevagate, Gerba-gate, Ghouirgate, Gibbs-gate, Giggsgate, Gillet-tegate, Glaciergate, Glasgate, Glassgate, Glogogate, Glogo-Gate, Glovegate, Goggle-gate, Golfgate, Golf-gate, Gongadzegate, Gong-gate, Gonogate, Googlegate, Gordon-gate, Goregate, Gore-gate, Gorillagate, Gove-gate, Grannygate, Graygate, Greasegate, Greygate, Gropegate, Gurkha-gate, Guttengate, Hackergate, Hackgate, Hack-gate, Haka-gate, Hampshire-Gate, Handygate, Hangargate, Hansiegate, Hanssongate, Hansson-gate, Hanukkah-gate, Harigate, Harpergate, Hecklegate, Hedgegate, Heimdalsgate, Henry-gate, Himalayagate, Himalaya-Gate, Hobnobgate, Hollywoodgate, Hookergate, Hoop-gate, Hotelgate, Housegate, Huntgate, Iguanagate, Iguana-gate, Iguangate, India-Gate, Inkathagate, Irangate, Iraqgate, Irisgate, Iris-gate, Isaiahgate, Islam-gate, Ivygate, Jack-Gate, Jakartagate, J-gate, Jokegate, Joostgate, Juetengate, Justiagate, Künigl-Gate, Kaffee-Gate, Kandikalgate, Kanyegate, Karachigate, Karachi-gate, Karashigate, Karatschi-gate, Karatschi-Gate, Karichigate, Karlsruhegate, Katiagate, Katyagate, Kaudergate, Kazakhgate, Kazakh-gate, King-gate, Kiwigate, Klimagate, Klima-Gate, Knappik-Gate, Knuckles-gate, Kochigate, Kohlgate, Konstantingate, Konstatingate, Koreagate, Kostitsyn-gate, Krawatten-Gate, Kuchmagate, Kundusgate, Lake-Gate, Lance-gate, Landman-gate, Landrieu-gate, Lasagne-Gate, Leaflet-gate, Leakage-gate, Leakgate, Leipsigate, Liargate, Liegate, Lie-gate, Lightbulbgate, Lightergate, Lingam-gate, Liquorgate, Lob-bygate, Lobby-Gate, Locationgate, Lockergate, Loliondogate, Lolitagate, Lolita-Gate, Lotterygate, Lucindagate, Lumbergate, Luxurygate, Lylegate, Lyongate, Macaca-gate, Madagate, Mafoliegate, Maischberger-Gate, Mamogate, Manuelgate, Mariahilfgate, Mar-rakechgate, Marrgate, Maultaschen-Gate, Meangate, Meemogate, Meersäuli-Gate, Mehran-gate, Meischberger-Gate, Memegate, Memogate, Memo-gate, Merkelgate, Merkel-Gate, Minigate, Minougate, Mirrorgate, Mismah-gate, Mistra-gate, Mitre-gate, Mittalgate, Moatgate, Modigate, Mollygate, Mondegate, Mong-gate, Monicagate, Monica-gate, Mon-keygate, Monkey-gate, Moomoogate, Morgan-gate, Moringate, Moting-gate, Mounchipougate, Muffin-gate, Mukatagate, Munchgate, Muntakagate, Muntaka-gate, Murdochgate, Murdoch-Gate, Mustard-Gate, Nannygate, Nanny-gate, Naomigate, Nappygate, Nautilisgate, Nchin-dogate, Nestlégate, Ngota-Gate, Nigergate, Night-Gate, Nippelgate, Nipplegate, Nipple-

Gate, Noballgate, Noemigate, Noemi-gate, Nuttgate, Nutt-gate, Obbligate, Oda-gate, Ofergate, Officegate, Oilgate, Ornament-Gate, Pachaurigate, Paintergate, Pakigate, Paksagate, Palm-gate, Pandagate, Panda-gate, Papigate, Para-gate, Parliamentgate, Passgate, Pentagate, Pepsigate, Permitgate, Petrogate, Phonegate, Pianogate, Piegate, Piggygate, Pinot-gate, Pizzagate, Plamegate, Planegate, Polarbeargate, Poppygate, Poppygate, Pornogate, Porschegate, Porsche-gate, Portersgate, Pottergate, Prachandagate, Prattgate, Prettyblondespyinabikinigate, Pretzel-gate, Prive-gate, Puppygate, Pussygate, Quartergate, Queengate, Quoragate, Quotagate, Quotasgate, Rüttgers-Gate, Radiagate, Radio-gate, Raegate, Rafagate, Raitt-gate, Raja-gate, Refudiategate, Renogate, Ribérygate, Rihanna-gate, Rinkagate, Roethlisberger-gate, Rooneygate, Rubbygate, Rubygate, Ruby-Gate, Ruddockgate, Ruddock-gate, Rudigate, Rulon-Gate, Russiagate, Russiangate, Rwego-gate, Séguragate, Sachgate, Sachsgate, Sachs-gate, Sacksgate, Salamigate, Salami-gate, Sale-Gate, Saliva-gate, Salon-Gate, Sarkogate, Sarko-Gate, Satelitengate, Saunagate, Savoiagate, Schachbrett-Gate, Scheißegate, Schlaffgate, Schnapsgate, Schneidergate, Schummelgate, Schummel-Gate, Science-gate, Scootergate, Seguragate, Semplegate, Senkaku-gate, Servergate, Sexgate, Shaanikagate, Sharongate, Shawinigate, Sherrodgate, Shouldergate, Showgirl-Gate, Shrubgate, Silikongate, Singapurgate, Sizergate, Skategate, Skirtgate, Slapgate, Slickergate, Sluddengate, Sludden-gate, Smashitgate, Smeargate, Sockgate, Sokolgate, Solférigate, Sonygate, Sootygate, Sophiegate, Sorosgate, Sowetogate, Spillgate, Spitgate, Spitzel-Gate, Splattergate, Sprinklergate, Spygate, Spy-gate, Squatgate, Squidgygate, Squiffogate, Stormontgate, Stramongate, Strasser-Gate, Sudatgate, Suizagate, Suspension-gate, Swear-gate, Tabloid-gate, Tailsgate, Tanovgate, Tapegate, Tattoo-Gate, Tawke-gate, Taxigate, Tea-Gate, Teapotgate, Teapot-gate, Telekomgate, Temangalogate, Tequilagate, Terrygate, Tevezgate, Texasgate, Textgate, Thaksingate, Tharoorgate, Throttlegate, Tigergate, Tiger-gate, Timbergate, Tindallgate, Tipgate, Toagate, Toiletgate, Tommygate, Tony-Gate, Toothgate, Toyotagate, Trachtengate, Tractorgate, Tractor-Gate, Travelgate, Trimgate, Tripgate, Tritongate, Troopergate, Trousergate, Trouser-Gate, Tuansgate, Tunnelgate, Turkeygate, Twinbedsgate, Twittergate, Twitter-gate, Utegate, Ute-gate, Vatogate, Veggie-gate, Videogate, Vietgate, Virungagate, Visa-Gate, Vodafone-Gate, Volgagate, Volumegate, Votergate, Wafergate, Wagergate, Wag-gate, Waisselgate, Waltergate, Wargate, Warmergate, Watergate, Waterkantgate, Waterlilygate, Waterproofgate, Weegate, Weinergate, Weiner-gate, Weingergate, Wendygate, Wicketgate, Wikigate, Wiki-Gate, Williamsgate, Willogate, Willowgate, Wilsongate, Winegate, Wiregate, Woertgate, Woerthgate, Woerth-Gate,

Woethgate, Womengate, Woodsgate, Woolworthsgate, Worthgate, Woyomegate, Wulffgate, Yachtgate, Yachtsgate, Yeongpogate, Yeongpo-gate, Youngpo-gate, Yunusgate, Zahiagate, Zifagate, Zimbabwegate, Zinebgate, Zippergate, Ziscogate, Zorbagate, Zumagate

Bibliography

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.

[2] Amr Ahmed and Eric P. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. *CoRR*, abs/1203.3463, 2012.

[3] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. Human-Computer Interaction Series. Springer, 2011.

[4] Joshua Albrecht, Rebecca Hwa, and G. Elisabeta Marai. The Chinese Room: Visualization and Interaction to Understand and Correct Ambiguous Machine Translation. *Comput. Graph. Forum*, 28(3):1047–1054, 2009.

[5] Conrad Albrecht-Buehler, Benjamin Watson, and David A. Shamma. Visualizing Live Text Streams Using Motion and Temporal Pooling. *IEEE Computer Graphics and Applications*, 25:52–59, 2005.

[6] James Allan, Jaime G. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study Final Report, 1998. Computer Science Department. Paper 341. http://repository.cmu.edu/compsci/341.

[7] Jamal Alsakran, Yang Chen, Ye Zhao, Jing Yang, and Dongning Luo. STREAMIT: Dynamic visualization and interactive exploration of text

streams. In Giuseppe Di Battista, Jean-Daniel Fekete, and Huamin Qu, editors, *IEEE Pacific Visualization Symposium, PacificVis 2011, Hong Kong, China, 1-4 March, 2011*, pages 131–138. IEEE, 2011.

[8] Keith Andrews and Helmut Heidegger. Information Slices : Visualising and Exploring Large Hierarchies using Cascading Semi-Circular Discs. *Information Visualization*, pages 9–12, 1998.

[9] Daniel Angus, Andrew E. Smith, and Janet Wiles. Conceptual Recurrence Plots: Revealing Patterns in Human Discourse. *IEEE Trans. Vis. Comput. Graph.*, 18(6):988–997, 2012.

[10] Aleks Aris, Ben Shneiderman, Catherine Plaisant, Galit Shmueli, and Wolfgang Jank. Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration. In *INTERACT*, pages 835–846, 2005.

[11] Martin Atkinson and Erik Van der Goot. Near real time information mining in multilingual news. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1153–1154, 2009.

[12] R. Harald Baayen. On Frequency, Transparency, and Productivity. *Yearbook of Morphology*, pages 181–208, 1992.

[13] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and Issues in Data Stream Systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.

[14] L. Douglas Baker, Thomas Hofmann, Andrew K. Mccallum, and Yiming Yang. A Hierarchical Probabilistic Model for Novelty Detection in Text. Technical Report, 1999.

[15] Michael Balzer and Oliver Deussen. Voronoi Treemaps. In *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, page 7. IEEE Computer Society, 2005.

[16] Ziv Bar-Joseph, Erik D. Demaine, David K. Gifford, Nathan Srebro, Angèle M. Hamel, and Tommi Jaakkola. K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data. *Bioinformatics*, 19(9):1070–1078, 2003.

[17] Luc Beaudoin, Marc-Antoine Parent, and Louis C. Vroomen. Cheops: A Compact Explorer for Complex Hierarchies. In *IEEE Visualization*, pages 87–92, 1996.

[18] Benjamin B. Bederson, Ben Shneiderman, and Martin Wattenberg. Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. *ACM Trans. Graph.*, 21(4):833–854, 2002.

[19] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[21] Ulrik Brandes and Steven R. Corman. Visual Unrolling of Network Evolution and the Analysis of Dynamic Discourse. *Information Visualization*, 2(1):40–50, 2003.

[22] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.

[23] Samuel Brody and Mirella Lapata. Bayesian Word Sense Induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[24] Kevin Buchin, Bettina Speckmann, and Kevin Verbeek. Flow Map Layout via Spiral Trees. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2536–2544, 2011.

[25] Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S. Ebert, and Thomas Ertl. Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2012, Seattle, Washington, USA, 14-19 October 2012, part of VisWeek 2012*, pages 143–152. IEEE, 2012.

[26] Claudine Chamoreau and Isabelle Léglise. *A Multi-model Approach to Contact-induced Language Change*, volume 2 of *Language contact and bilingualism*. Berlin: De Gruyter Mouton, 2012.

[27] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey. *ACM Comput. Surv.*, 41(3), 2009.

[28] Chaomei Chen. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.

[29] George Chin, Mudita Singhal, Grant Nakamura, Vidhya Gurumoorthi, and Natalie Freeman-Cadoret. Visual Analysis of Dynamic Data Streams. *Information Visualization*, 8:212–229, June 2009.

[30] Christopher Collins. Interactive Visualizations of Natural Language. Ph.D. Dissertation, University of Toronto, 2010.

[31] Christopher Collins, Fernanda B. Viégas, and Martin Wattenberg. Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*, pages 91–98, 2009.

[32] Bernard Comrie. *Typology*, volume 6 of *Handbook of Pragmatics Highlights*. John Benjamins, 2010.

[33] Paul Cook and Suzanne Stevenson. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 28–34, Valletta, Malta, 2010.

[34] Terry Crowley and Claire Bowern. *An Introduction to Historical Linguistics - Fourth Edition*. Oxford University Press, 2010.

[35] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2412–2421, 2011.

[36] Weiwei Cui, Huamin Qu, Hong Zhou, Wenbin Zhang, and Steven Skiena. Watch the Story Unfold with TextWheel: Visualization of Large-Scale News Streams. *ACM TIST*, 3(2):20, 2012.

[37] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X. Zhou, and Huamin Qu. Context Preserving Dynamic Word Cloud Visualization. In *IEEE Pacific Visualization Symposium PacificVis 2010, Taipei, Taiwan, March 2-5, 2010*, pages 121–128, 2010.

[38] Michael Cysouw and Bernhard Wälchli. Parallel Texts: Using Translational Equivalents in Linguistic Typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99, 2007.

[39] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[40] Steve DeNeefe, Kevin Knight, and Hayward H. Chan. Interactively Exploring a Machine Translation Model. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*. The Association for Computer Linguistics, 2005.

[41] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, pages 115–122, 2010.

[42] Xiaowen Ding and Bing Liu. The Utility of Linguistic Rules in Opinion Mining. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 811–812. ACM, 2007.

[43] Xiaowen Ding, Bing Liu, and Philip S. Yu. A Holistic Lexicon-based Approach to Opinion Mining. In Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti, editors, *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 231–240. ACM, 2008.

[44] Bruce C. Donaldson. *A Grammar of Afrikaans*. Berlin: Mouton de Gruyter, 1993.

[45] Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson. VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14:1205–1212, November 2008.

[46] Marian Dörk, Daniel M. Gruen, Carey Williamson, and M. Sheelagh T. Carpendale. A Visual Backchannel for Large-Scale Events. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1129–1138, 2010.

[47] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. ParallelTopics: A Probabilistic Approach to Exploring Document Collections. In *2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23-28, 2011*, pages 231–240. IEEE, 2011.

[48] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle Zhou. LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2012, Seattle, Washington, USA, 14-19 October 2012, part of VisWeek 2012*, pages 93–102. IEEE, 2012.

[49] Matthew Dryer and Martin Haspelmath. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. `http://wals.info/`, 2011.

[50] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing Tags over Time. *ACM Transactions of the Web (TWEB)*, 1, August 2007.

[51] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[52] Tom Fawcett and Foster J. Provost. Activity Monitoring: Noticing Interesting Changes in Behavior. In *KDD*, pages 53–62, 1999.

[53] John R. Firth. *Papers in Linguistics 1934-1951*. London: Oxford University Press, 1957.

[54] Steven Roger Fischer. *A History of Writing*. Reaktion Books Ltd, 2001.

[55] Danyel Fisher, Aaron Hoff, George G. Robertson, and Matthew Hurst. Narratives: A visualization to track narrative events as they develop. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2008, Columbus, Ohio, USA, 19-24 October 2008*, pages 115–122, 2008.

[56] Mirjam Fried. *Introduction: From Instances of Change to Explanations of Change*, volume 6 of *Handbook of Pragmatics Highlights*. John Benjamins, 2010.

[57] Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2013.

[58] Mohammad Ghoniem, Dongning Luo, Jing Yang, and William Ribarsky. NewsLab: Exploratory Broadcast News Video Analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2007, Sacramento, California, USA, October 30-November 1, 2007*, pages 123–130, 2007.

[59] David Graddol. The Future of English? A Guide to Forecasting the Popularity of the English Language in the 21st Century. 1997.

[60] Tom Güldemann. "Sprachraum" and Geography. In *The Handbook of Language Mapping*, Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter, in press.

[61] Lars Kai Hansen, Sigurdur Sigurdsson, Thomas Kolenda, Finn Arup Nielsen, Ulrik Kjems, and Jan Larsen. Modeling Text With Generalizable Gaussian Mixtures. In *In Proceedings of ICASSP'2000*, pages 3494–3497. IEEE, 1999.

[62] M.C. Hao, U. Dayal, C. Rohrdantz, M. Hsu, M.E. Dekhil, and R. Ghosh. Selecting Sentiment Attributes for Visualization, November 26 2013. US Patent 8,595,151.

[63] Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Christian Rohrdantz. Patent Application US 2012/0060080: Visual Representation of a Cell-based Calendar Transparently Overlaid with Event Visual Indicators for Mining Data Records (http://www.google.com/patents/US20120060080), March 2012.

[64] Ming C. Hao, Umeshwar Dayal, and Christian Rohrdantz. Patent Application WO/2012/044305: Identification of Events of Interest (http://patentscope.wipo.int/search/en/WO2012044305), April 2012.

[65] Ming C. Hao, Umeshwar Dayal, Christian Rohrdantz, and Lars-Erik Haug. Patent Application US 2013/0054597: Constructing an Association Data Structure to Visualize Association among Co-occurring Terms (http://www.freepatentsonline.com/y2013/0054597.html), February 2013.

[66] Ming C. Hao, Umeshwar Dayal, Baoyao Zhou, Cheng Chang, Meichun Hsu, Mohamed E. Dekhil, Riddhiman Ghosh, and Christian Rohrdantz. Patent Application WO/2012/167399: Sentiment Trend Visualization Relating to an Event occurring in a Particular Geographic Region (http://patentscope.wipo.int/search/en/WO2012167399), December 2012.

[67] Ming C. Hao, Daniel A. Keim, Umeshwar Dayal, Daniela Oelke, and Chantal Tremblay. Density Displays for Data Stream Monitoring. *Comput. Graph. Forum*, 27(3):895–902, 2008.

[68] Ming C. Hao, Christian Rohrdantz, and Umeshwar Dayal. Patent Application US 2013/0046756: Visualizing Sentiment Results with Visual Indicators Representing User Sentiment and Level of Uncertainty (http://www.freepatentsonline.com/y2013/0046756.html), February 2013.

[69] Ming C. Hao, Christian Rohrdantz, Umeshwar Dayal, Daniel Keim, and Lars-Erik Haug. Patent Application US 2012/0109843: Visual Analysis of a Time Sequence of Events Using a Time Density Track (http://www.google.com/patents/US20120109843), May 2012.

[70] Ming C. Hao, Christian Rohrdantz, Halldor Janetzko, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Meichun Hsu. Integrating Sentiment Analysis and Term Associations with Geo-Temporal Visualizations on Customer Feedback Streams. In Pak Chung Wong, David L. Kao, Ming C. Hao, Chaomei Chen, Robert Kosara, Mark A. Livingston, Jinah Park, and Ian Roberts, editors, *SPIE 2012 Conference on Visualization and Data Analysis (VDA 2012)*, 2012.

[71] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, 8:9–20, 2002.

[72] Susan Havre, Elizabeth G. Hetzler, and Lucy T. Nowell. ThemeRiver: Visualizing Theme Changes over Time. In Jock D. Mackinlay, Steven F. Roth, and Daniel A. Keim, editors, *IEEE Symposium on Information Visualization 2000 (INFOVIS'00), Salt Lake City, Utah, USA, October 9-10, 2000*, pages 115–123. IEEE Computer Society, 2000.

[73] Jeffrey Heer, Stuart K. Card, and James A. Landay. Prefuse: A Toolkit for Interactive Information Visualization. In Gerrit C. van der Veer and Carolyn Gale, editors, *Proceedings of the 2005 Conference on Human*

Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005, pages 421–430. ACM, 2005.

[74] Elizabeth G. Hetzler, Vernon L. Crow, Deborah A. Payne, and Alan E. Turner. Turning the Bucket of Text into a Pipe. In *Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 89–94, Washington, DC, USA, 2005. IEEE Computer Society.

[75] Gerhard Heyer, Florian Holz, and Sven Teresniak. Change of Topics over Time and Tracking Topics by Their Change of Meaning. In Ana L. N. Fred, editor, *KDIR 2009: Proc. of Int. Conf. on Knowledge Discovery and Information Retrieval*. INSTICC Press, October 2009.

[76] Raymond Hickey. *Language Change*, volume 6 of *Handbook of Pragmatics Highlights*. John Benjamins, 2010.

[77] Hans Henrich Hock. *Principles of Historical Linguistics*. Berlin/New York: Mouton de Gruyter, second edition, 1991.

[78] Hans Henrich Hock and Brian D. Joseph. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin/New York: Mouton de Gruyter, second edition, 2009.

[79] Danny Holten. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Trans. Vis. Comput. Graph.*, 12(5):741–748, 2006.

[80] Florian Holz and Sven Teresniak. Towards automatic detection and tracking of topic change. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings*, volume 6008 of *Lecture Notes in Computer Science*, pages 327–339. Springer, 2010.

[81] Timo Honkela, V. Pulkki, and Teuvo Kohonen. Contextual Relations of Words in Grimm Tales, Analyzed by Self-organizing Map. In

F. Fogelman-Soulie and P. Gallinari, editors, *Proceedings of International Conference on Artificial Neural Networks (ICANN-95)*, volume 2, pages 3–7, 1995.

[82] Michael Hund. Real-time Event Detection for Emergency Response: Visually Analysing Twitter and Online News. Bachelor Thesis, University of Konstanz, October 2012.

[83] Indratmo, Julita Vassileva, and Carl Gutwin. Exploring Blog Archives with Interactive Visualization. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, pages 39–46, New York, NY, USA, 2008. ACM.

[84] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization*, pages 361–378, 1990.

[85] Yoshiharu Ishikawa and Mikine Hasegawa. T-Scroll: Visualizing Trends in a Time-Series of Documents for Interactive User Exploration. In László Kovács, Norbert Fuhr, and Carlo Meghini, editors, *Research and Advanced Technology for Digital Libraries*, volume 4675 of *Lecture Notes in Computer Science*, pages 235–246. Springer Berlin / Heidelberg, 2007.

[86] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[87] B. Johnson and Ben Shneiderman. Tree maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In *IEEE Visualization*, pages 284–291, 1991.

[88] Daniel A. Keim. Information Visualization and Visual Data Mining. *IEEE Trans. Vis. Comput. Graph.*, 8(1):1–8, 2002.

[89] Daniel A. Keim, Milos Krstajic, Christian Rohrdantz, and Tobias Schreck. Real-time visual analytics for text streams. *IEEE Computer*, 46(7):47–55, 2013.

[90] Daniel A. Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler. Visual Analytics: Combining Automated Discovery with Interactive Visualizations. In Jean-François Boulicaut, Michael R. Berthold, and Tamás Horváth, editors, *Discovery Science, 11th International Conference, DS 2008, Budapest, Hungary, October 13-16, 2008. Proceedings*, volume 5255 of *Lecture Notes in Computer Science*, pages 2–14. Springer, 2008.

[91] Soo-Min Kim and Eduard Hovy. Determining the Sentiment of Opinions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 1367–1373, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[92] Slava Kisilevich, Christian Rohrdantz, and Daniel A. Keim. "Beautiful picture of an ugly place". Exploring Photo Collections Using Opinion and Sentiment Analysis of User Comments. In *International Multiconference on Computer Science and Information Technology - IMCSIT 2010, Wisla, Poland, 18-20 October 2010, Proceedings*, pages 419–428, 2010.

[93] Teuvo Kohonen. The Self-organizing Map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[94] Milos Krstajic, Enrico Bertini, and Daniel A. Keim. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2432–2439, 2011.

[95] Milos Krstajic, Enrico Bertini, Florian Mansmann, and Daniel A. Keim. Visual Analysis of News Streams with Article Threads. In *StreamKDD '10: Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, pages 39–46, New York, NY, USA, 2010. ACM.

[96] Milos Krstajic, Florian Mansmann, Andreas Stoffel, Martin Atkinson, and Daniel A. Keim. Processing Online News Streams for Large-Scale Semantic Analysis. In *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, pages 215–220. IEEE, 2010.

[97] Milos Krstajic, Mohammad Najm-Araghi, Florian Mansmann, and Daniel Keim. Incremental Visual Text Analytics of News Story Development. In *SPIE 2012 Conference on Visualization and Data Analysis (VDA 2012)*, 2012.

[98] Milos Krstajic, Christian Rohrdantz, Michael Hund, and Andreas Weiler. Getting There First: Real-Time Detection of Real-World Incidents on Twitter. In *2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA*, 2012.

[99] J. B. Kruskal and J. M. Landwehr. Icicle Plots: Better Displays for Hierarchical Clustering. *The American Statistician*, 37(2):162–168, 1983.

[100] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. Beverly Hills and London: Sage, 1978.

[101] Ela Kumar. *Natural Language Processing*. I.K. International Publishing House Pvt. Ltd., 2011.

[102] John Lamping and Ramana Rao. Laying Out and Visualizing Large Trees Using a Hyperbolic Space. In *ACM Symposium on User Interface Software and Technology*, pages 13–14, 1994.

[103] John Lamping, Ramana Rao, and Peter Pirolli. A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In Irvin R. Katz, Robert L. Mack, Linn Marks, Mary Beth Rosson, and Jakob Nielsen, editors, *Human Factors in Computing Systems, CHI '95 Conference Proceedings, Denver, Colorado, USA, May 7-11, 1995*, pages 401–408. ACM/Addison-Wesley, 1995.

[104] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and M. Sheelagh T. Carpendale. SparkClouds: Visualizing Trends in Tag Clouds. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1182–1189, 2010.

[105] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John T. Stasko, and Haesun Park. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Comput. Graph. Forum*, 31(3):1155–1164, 2012.

[106] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[107] Bing Liu. Sentiment Analysis and Subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

[108] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In Allan Ellis and Tatsuya Hagino, editors, *Proceedings of the 14th International Conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 342–351. ACM, 2005.

[109] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Interactive, Topic-based Visual Text Summarization and Analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 543–552, New York, NY, USA, 2009. ACM.

[110] Anke Lüdeling and Stefan Evert. The Emergence of Productive Non-medical *-itis*. Corpus Evidence and Qualitative Analysis. In S. Kepser and M. Reis, editors, *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Berlin: Mouton de Gruyter, 2005.

[111] Dongning Luo, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel A. Keim. EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Trans. Vis. Comput. Graph.*, 18(1):93–105, 2012.

[112] Verena Lyding, Ekaterina Lapshinova-Koltunski, Stefania Degaetano-Ortlieb, Henrik Dittmann, and Christopher Culy. Visualising Linguistic Evolution in Academic Discourse. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 44–48, April 2012.

[113] Alan M. MacEachren, Anuj R. Jaiswal, Anthony C. Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. SensePlace2: GeoTwitter Analytics Support for Situational Awareness. In *2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23-28, 2011*, pages 181–190. IEEE, 2011.

[114] Jock D. Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. Graph.*, 5(2):110–141, 1986.

[115] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of Frequent Episodes in Event Sequences. *Data Min. Knowl. Discov.*, 1(3):259–289, 1997.

[116] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, June 1999.

[117] Florian Mansmann, Daniel A. Keim, Stephen C. North, Brian Rexroad, and Daniel Sheleheda. Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1105–1112, 2007.

[118] Markos Markou and Sameer Singh. Novelty Detection: A Review - Part 1: Statistical Approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[119] Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A. Keim. Visualizing Vowel Harmony. *Linguistic Issues in Language Technology*, 4(Issue 2):1–33, December 2010.

[120] Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. Consonant Co-Occurrence in Stems across Languages: Automatic Analysis and Visualization of a Phonotactic Constraint. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 70–78, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[121] Thomas Mayer, Christian Rohrdantz, Frans Plank, Miriam Butt, and Daniel A. Keim. A Quantitative Approach to the Contrast and Sta-

bility of Sounds. In *Proceedings of the 4th Conference on Quantitative Investigations in Theoretical Linguistics (QITL-4)*, pages 59–64, 2011.

[122] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[123] Qingliang Miao, Qiudan Li, and Ruwei Dai. AMAZING: A Sentiment Mining and Retrieval System. *Expert Syst. Appl.*, 36(3):7192–7198, 2009.

[124] D. Gary Miller. *Language Change and Linguistic Theory I - Approaches, Methodology, and Sound Change*. Oxford University Press, 2010.

[125] Tamara Munzner. A Nested Process Model for Visualization Design and Validation. *IEEE Trans. Vis. Comput. Graph.*, 15(6):921–928, 2009.

[126] David Nash. *Topics in Warlpiri Grammar*. Garland Publishing, New York, London, 1986.

[127] Roberto Navigli. Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69, 2009.

[128] Petra Neumann, Stefan Schlechtweg, and M. Sheelagh T. Carpendale. ArcTrees: Visualizing Relations in Hierarchical Data. In Ken Brodlie, David J. Duke, and Kenneth I. Joy, editors, *EuroVis05: Joint Eurographics - IEEE VGTC Symposium on Visualization, Leeds, United Kingdom, 1-3 June 2005*, pages 53–60. Eurographics Association, 2005.

[129] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics, 2006.

[130] Johanna Nichols. *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press, 1992.

[131] Irina Nikolaeva and Maria V. Tolskaya. *A Grammar of Udihe*. Berlin: Mouton de Gruyter, 2001.

[132] Hans J. Nissen, Peter Damerow, and Robert K. Englund. *Archaic Bookkeeping: Writing and Techniques of Economic Administration in the Ancient Near East*. Chicago: The University of Chicago Press, 1993.

[133] Marc Nunkesser and Daniel Sawitzki. Blockmodels. In Ulrik Brandes and Thomas Erlebach, editors, *Network Analysis: Methodological Foundations [outcome of a Dagstuhl seminar, 13-16 April 2004]*, volume 3418 of *Lecture Notes in Computer Science*, pages 253–292. Springer, 2005.

[134] Daniela Oelke. *Visual Document Analysis: Towards a Semantic Analysis of Large Document Collections*. PhD thesis, 2010.

[135] Daniela Oelke, Ming C. Hao, Christian Rohrdantz, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Halldór Janetzko. Visual Opinion Analysis of Customer Feedback Data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*, pages 187–194. IEEE, 2009.

[136] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[137] Doantam Phan, Ling Xiao, Ron B. Yeh, Pat Hanrahan, and Terry Winograd. Flow Map Layout. In *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, page 29. IEEE Computer Society, 2005.

[138] Christian Pich. MDSJ: Java Library for Multidimensional Scaling (Version 0.2). Algorithmics Group at University of Konstanz. http://www.inf.uni-konstanz.de/algo/software/mdsj/, 2009.

[139] Ana-Maria Popescu and Oren Etzioni. Extracting Product Features and Opinions from Reviews. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Van-*

couver, British Columbia, Canada. The Association for Computational Linguistics, 2005.

[140] Konstantin Pozdniakov and Guillaume Segerer. Similar Place Avoidance: A Statistical Universal. *Linguistic Typology*, 11(2):307–348, 2007.

[141] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding Domain Sentiment Lexicon through Double Propagation. In Craig Boutilier, editor, *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 1199–1204, 2009.

[142] Ellen Riloff and Janyce Wiebe. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[143] George G. Robertson, Jock D. Mackinlay, and Stuart K. Card. Cone Trees: Animated 3D Visualizations of Hierarchical Information. In Scott P. Robertson, Gary M. Olson, and Judith S. Olson, editors, *Conference on Human Factors in Computing Systems, CHI 1991, New Orleans, LA, USA, April 27 - May 2, 1991, Proceedings*, pages 189–194. ACM, 1991.

[144] Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. Feature-Based Visual Sentiment Analysis of Text Document Streams. *ACM TIST*, 3(2):26, 2012.

[145] Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel A. Keim. 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)*, pages 305–310, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[146] Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. The World's Languages Explorer: Visual Analysis

of Language Features in Genealogical and Areal Contexts. *Computer Graphics Forum*, 31(3):935–944, 2012.

[147] Christian Rohrdantz, Steffen Koch, Charles Jochim, Gerhard Heyer, Gerik Scheuermann, Thomas Ertl, Hinrich Schütze, and Daniel A. Keim. Visuelle Textanalyse - Interaktive Exploration von semantischen Inhalten. *Informatik Spektrum*, 33(6):601–611, 2010.

[148] Christian Rohrdantz, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel A. Keim. Comparative Visual Analysis of Cross-linguistic Features. In Joern Kohlhammer and Daniel A. Keim, editors, *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)*, pages 27–32, 2010.

[149] Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. Lexical Semantics and Distribution of Suffixes - A Visual Analysis. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 7–15, April 2012.

[150] Christian Rohrdantz, Daniela Oelke, Milos Krstajic, and Fabian Fischer. Real-Time Visualization of Streaming Text Data: Tasks and Challenges. In *Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek 2011*, 2011.

[151] Stuart J. Rose, Scott Butner, Wendy Cowley, Michelle L. Gregory, and Julia Walker. Describing Story Evolution from Dynamic Information Streams. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*, pages 99–106, 2009.

[152] Rudolph J. Rummel. *Applied Factor Analysis*, pages 298–299. Nortwestern Univ. Pr., 1970.

[153] Eyal Sagi, Stefan Kaufmann, and Brady Clark. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, 2009.

[154] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 851–860. ACM, 2010.

[155] Hinrich Schütze. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[156] Lei Shi, Furu Wei, Shixia Liu, Li Tan, Xiaoxiao Lian, and Michelle X. Zhou. Understanding Text Corpora with Multiple Facets. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, pages 99–106, 2010.

[157] Ben Shneiderman. Tree Visualization with Tree-Maps: 2-d Space-Filling Approach. *ACM Trans. Graph.*, 11(1):92–99, 1992.

[158] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *VL*, pages 336–343, 1996.

[159] Aidan Slingsby, Jason Dykes, and Jo Wood. Configuring Hierarchical Layouts to Address Research Questions. *IEEE Trans. Vis. Comput. Graph.*, 15(6):977–984, 2009.

[160] Ramakrishnan Srikant and Rakesh Agrawal. Mining Quantitative Association Rules in Large Relational Tables. In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 1–12. ACM Press, 1996.

[161] John T. Stasko and Eugene Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In Jock D. Mackinlay, Steven F. Roth, and Daniel A. Keim, editors, *IEEE Symposium on Information Visualization 2000 (INFOVIS'00), Salt Lake City, Utah, USA, October 9-10, 2000*, pages 57–65. IEEE Computer Society, 2000.

[162] Peter Stein. *Schriftkultur: Eine Geschichte des Schreibens und Lesens.* Wissenschaftliche Buchgesellschaft, Darmstadt, 2006.

[163] Boris V. Sukhotin. Méthode de déchiffrage, outil de recherche en linguistique. *T.A. Informations*, 2:1–43, 1973.

[164] Russell Swan and David Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80, 2000.

[165] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[166] Roberto Therón, Laura Fontanillo, Andrés Esteban, and Carlos Seguín. Visual analytics: A Novel Approach in Corpus Linguistics and the Nuevo Diccionario Histórico del Español. *III Congreso Internacional de Lingüística de Corpus*, 2011.

[167] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* National Visualization and Analytics Ctr, 2005.

[168] James J. Thomas and Kristin A. Cook. A Visual Analytics Agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[169] Frank van Ham. Using Multilevel Call Matrices in Large Software Projects. In *9th IEEE Symposium on Information Visualization (InfoVis 2003), 20-21 October 2003, Seattle, WA, USA.* IEEE Computer Society, 2003.

[170] Fernanda B. Viégas, Scott A. Golder, and Judith S. Donath. Visualizing Email Content: Portraying Relationships From Conversational Histories. In Rebecca E. Grinter, Tom Rodden, Paul M. Aoki, Edward Cutrell, Robin Jeffries, and Gary M. Olson, editors, *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, pages 979–988. ACM, 2006.

[171] Fernanda B. Viégas and Marc A. Smith. Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces. In *HICSS*, 2004.

[172] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying Cooperation and Conflict between Authors with *History Flow* Visualizations. In Elizabeth Dykstra-Erickson and Manfred Tscheligi, editors, *Proceedings of the 2004 Conference on Human Factors in Computing Systems, CHI 2004, Vienna, Austria, April 24 - 29, 2004*, pages 575–582. ACM, 2004.

[173] Roel Vliegen, Jarke J. van Wijk, and Erik-Jan van der Linden. Visualizing Business Data with Generalized Treemaps. *IEEE Trans. Vis. Comput. Graph.*, 12(5):789–796, 2006.

[174] Bernhard Wälchli. Indirect Measurement in Morphological Typology. In Andrea Ender, Adrian Leemann, and Bernhard Wälchli, editors, *Methods in Contemporary Linguistics*. Berlin: Mouton de Gruyter, 2012.

[175] Bernhard Wälchli. Algorithmic Typology, Aggregating Without Features and Going from Known to Similar Unknown Categories Within and Across Languages. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology and Typology*. Berlin: Mouton de Gruyter, Forth.

[176] Xiaoyu Wang, Wenwen Dou, Z. Ma, J. Villalobos, Yang Chen, T. Kraft, and William Ribarsky. *I-SI*: Scalable Architecture for Analyzing Latent Topical-Level Information From Social Media Data. *Comput. Graph. Forum*, 31(3):1275–1284, 2012.

[177] Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, and Daniel A. Keim. Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. In *Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)*, 2009.

[178] Martin Wattenberg and Fernanda B. Viégas. The Word Tree, an Interactive Visual Concordance. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1221–1228, 2008.

[179] Andreas Weiler, Marc H. Scholl, Franz Wanner, and Christian Rohrdantz. Event identification for local areas using social media streaming data. In Kristen LeFevre, Ashwin Machanavajjhala, and Adam Silberstein, editors, *Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks, DBSocial 2013, New York, NY, USA, June, 23, 2013*, pages 1–6. ACM, 2013.

[180] Søren Wichmann, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck, and Helen Geyer. The ASJP Database (version 13). URL: `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`, 2010.

[181] Pak Chung Wong, Harlan Foote, Dan Adams, Wendy Cowley, and Jim Thomas. Dynamic Visualization of Transient Data Streams. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, InfoVis'03, pages 97–104, Washington, DC, USA, 2003. IEEE Computer Society.

[182] Jo Wood and Jason Dykes. Spatially Ordered Treemaps. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1348–1355, 2008.

[183] Yingcai Wu, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma. Semantic-Preserving Word Clouds by Seam Carving. *Comput. Graph. Forum*, 30(3):741–750, 2011.

[184] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. OpinionSeer: Interactive Visualization of Hotel Customer Feedback. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1109–1118, 2010.

[185] Xuchen Yao and Benjamin Van Durme. Nonparametric Bayesian Word Sense Induction. In *Proceedings of the 2011 Workshop on Graph-based*

Methods for Natural Language Processing, TextGraphs-6, June 23, 2011, Portland, Oregon, USA, pages 10–14. The Association for Computer Linguistics, 2011.

[186] Ypsilanti. A Digital Library of Language Relationships. Ypsilanti, MI: Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. `http://multitree.org/`, 2009.

[187] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1079–1088, New York, NY, USA, 2010. ACM.

[188] Jian Zhao, Fanny Chevalier, Christopher Collins, and Ravin Balakrishnan. Facilitating Discourse Analysis with Interactive Visualization. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2639–2648, 2012.

[189] Jürgen Ziegler, Christoph Kunz, and Veit Botsch. Matrix Browser: Visualizing and Exploring Large Networked Information Spaces. In Loren G. Terveen and Dennis R. Wixon, editors, *CHI Extended Abstracts*, pages 602–603. ACM, 2002.

[190] George Kingsley Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.* Addison-Wesley, 1949.