VisLR:
Visualization as Added Value in the Development, Use and Evaluation of Language Resources

Workshop Programme

09:00 – 10:30 – Morning Session, Part I

09:00 – 09:10 – Introduction

09:10 – 09:40
Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, *An Interactive Visualization of Crosslinguistic Colexification Patterns*

09:40 – 10:00
Roberto Theron and Eveline Wandl-Vogt, *The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually*

10:00 – 10:30
Pierrick Bruneau, Olivier Parisot, Amir Mohammadi, Cenk Demiroğlu, Mohammad Ghoniem and Thomas Tamisier, *Finding Relevant Features for Statistical Speech Synthesis Adaptation*

10:30 – 11:00 Coffee break

11:00 – 13:00 – Morning Session, Part II

11:00 – 11:30
Florian Stoffel, Dominik Jäckle and Daniel A. Keim, *Enhanced News-reading: Interactive and Visual Integration of Social Media Information*

11:30 – 11:50
Markus John, Florian Heimerl, Andreas Müller and Steffen Koch, *A Visual Focus+Context Approach for Text Comparison Tasks*

11:50 – 12:10
Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe and Daniel A. Keim, *V1 in Icelandic: A Multifactorical Visualization of Historical Data*

12:10 – 12:40
Stefan Jänicke, Marco Büchler, Gerik Scheuermann, *Improving the Layout for Text Variant Graphs*

12:40 – 13:00 – Closing Discussion

Editors

Annette Hautli-Janisz	University of Konstanz
Verena Lyding	European Academy of Bolzano/Bozen
Christian Rohrdantz	University of Konstanz

Workshop Organizers

Annette Hautli-Janisz	University of Konstanz
Verena Lyding	European Academy of Bolzano/Bozen
Christian Rohrdantz	University of Konstanz

Workshop Programme Committee

Noah Bubenhofer	Dresden University of Technology
Miriam Butt	University of Konstanz
Chris Culy	University of Tübingen
Christopher Collins	University of Ontario Institute of Technology
Annette Hautli-Janisz	University of Konstanz
Gerhard Heyer	Leipzig University
Kris Heylen	University of Leuven
Daniel Keim	University of Konstanz
Steffen Koch	University of Stuttgart
Verena Lyding	European Academy of Bolzano/Bozen
Thomas Mayer	Philipps-Universität Marburg
Daniela Oelke	DIPF Frankfurt
Christian Rohrdantz	University of Konstanz

Table of contents

Author Index

Preface to the VisLR Workshop Proceedings

The VisLR workshop aims at bringing together people from visual analytics and computational linguistics to discuss the potentials and the challenges related to visualizing language data and in particular language resources. Linguistics has a long tradition of visually representing language patterns, from tree representations in syntax to spectograms in phonetics. However, the large amounts and ever-increasing complexity of today's resources call for new ways of visually encoding a multitude of abstract information on language in order to assure and enhance the quality and usability of these language resources.

We invited submissions on research demonstrating the development, use and evaluation of visualization techniques for language resources. This includes work applying existing visualization techniques to language resources as well as research on new visualization techniques that are specifically targeted to the needs of language resources.

The workshop contributions comprise visualization approaches for lexicographic data, text resources as well as speech data. Mayer et al. and Theron & Wandl-Vogt present two visualizations for facilitating the interactive exploration of lexicographic data. Bruneau et al. show how to make use of visual tools for analyzing high-dimensional models for speech synthesis adaptation. The visualizations for text and corpus data propose visual approaches that can be applied to support enhanced news-reading (Stoffel et al.), distant reading (John et al.) and the study of language change (Butt et al.) as well as the comparison of different editions of a text (Jänicke et al.).

An Interactive Visualization of Crosslinguistic Colexification Patterns

Thomas Mayer[1], **Johann-Mattis List**[1], **Anselm Terhalle**[2], **Matthias Urban**[3]

[1]Philipps-Universität Marburg, [2]Heinrich-Heine-Universität Düsseldorf, [3]Universiteit Leiden
thomas.mayer@uni-marburg.de, mattis.list@uni-marburg.de,
terhalle@phil.uni-duesseldorf.de, m.urban@hum.leidenuniv.nl

Abstract

In this paper, we present an interactive web-based visualization for the CLICS database, an online resource for synchronic lexical associations *(colexification patterns)* in over 200 language varieties. The associations cover 1,288 concepts and represent the tendency for concepts to be expressed by the same words in the same languages and language varieties of the world. The complexity of the network structure in the CLICS database calls for a visualization component that makes it easier for researchers to explore the patterns of crosslinguistic colexifications. The network is represented as a force-directed graph and features a number of interactive components that allow the user to get an overview of the overall structure while at the same time providing an opportunity to look into the data in more detail. An integral part of the visualization is an interactive listing of all languages that contribute to the strength of a given pattern of colexification. Each language in the list is thereby attributed a different color depending on its genealogical or areal affiliation. In this way, given associations can be inspected for genealogical or areal bias.

Keywords: interactive visualization, colexification, crosslinguistic database

1. Introduction

What does 'good' have to do with 'beautiful'? Logically, not everything that is good is beautiful, and not everything that is beautiful is good. However, people seem to associate these concepts quite strongly, as linguistic data suggest: 'good' and 'beautiful' are expressed as identical words in at least 27 languages from 8 different language families. To assess the cognitive, linguistic, and cultural implications of this fact correctly, additional information would be useful: Where are these languages located on the globe? How are they distributed among the 8 families? Which other concepts are verbalized by the same form as 'good' and 'beautiful'?

Synchronic lexical association or *colexification*, i.e., the verbalization of two or more concepts by means of the same form in a given language, for instance, the words expressing both the notions 'good' and 'beautiful' mentioned above, is an important source of information for investigations in cognitive linguistics, linguistic typology, and historical semantics. In this paper, we present an interactive web-based visualization for CLICS, an online database that contains a large crosslinguistic data set on colexifications worldwide. We provide a brief outline of CLICS, describe the functionalities of the visualization, and show that the visualization is an indispensable tool that provides researchers an overview of the data and allows them to concisely plan further quantitative analyses according to their needs.

Concept	IDS key	Families	Languages
money	11.43	15	33
coin	11.44	9	13
iron	9.67	3	3
gold	9.64	2	2
tin, tinplate	9.69	2	2
white	15.64	2	2
blunt, dull	15.79	1	1
bright	15.57	1	1
chest	4.40	1	1
clock, timepiece	14.53	1	1
copper, bronze	9.66	1	1
earring	6.77	1	1
hammer	9.49	1	1
helmet	20.33	1	1
jewel	6.72	1	1
lead (noun)	9.68	1	1
price	11.87	1	1
razor	6.93	1	1
saw	9.48	1	1

Table 1: Common colexifications involving the concept 'silver' in CLICS. Concepts which are expressed by the same word form in more than one language family are shaded gray. In order to browse the table on the CLICS website, use the following URL: `http://clics.lingpy.org/all.php?gloss=silver`.

2. CLICS

CLICS (***C**rossl**i**nguistic **C**olexifications*, `http://clics.lingpy.org`) is an online database of synchronic lexico-semantic associations in 221 languages and language varieties of the world. CLICS exploits already existing large online lexical databases, but has the advantage that it makes visible the relationships between meanings and forms in the object languages, something which is not easily possible using the interfaces of its sources themselves. Table 1 gives an example of the basic structure of the data in CLICS.

2.1. Homonymy, polysemy, and colexification

A well-known concept from lexical semantic analysis is that of *polysemy*. It refers to the situation in which a lexical item possesses more than one identifiable sense be-

tween which there is a conceptual relation. A number of tests are available to distinguish polysemy from *semantic vagueness,* in which a division into distinct senses is not warranted. From an analytical perspective, polysemy has to be further distinguished from homonymy and contextual variation. *Homonymy* refers to the "accidental" verbalization of at least two meanings by the same sound chain, without any conceptual relation that is more than coincidental. *Contextual variation* designates the adaption of a lexicalized meaning to contextual factors in an utterance. Although historical and synchronic criteria have been proposed to distinguish polysemy from homonymy, and contextual variation can be tested by resorting to categorization (Blank, 1997), the differentiation depends on the individual analysis of every single word and is not entirely objective. Hence, it is difficult for quantitative investigations to provide this differentiation in advance. Here, we use the term *colexification* (originally from François, 2008) to refer to the situation in which two or more of the meanings in our sources correspond to the same lexical item in one of the languages. For instance, we would say that in Wayuu (an Arawak language of Venezuela), *anasɨ* colexifies 'good' and 'beautiful.'[1] Colexification is thus a deliberately ambiguous label that allows us to avoid making a commitment in each case as to the adequate lexical semantic analysis. Roughly speaking, colexification can correspond either to polysemy or contextual variation in lexical semantic analyses. Since CLICS is not based on such analyses that would allow us to further discriminate between the two, we chose colexification as a label that deliberately does not make a commitment with regard to this distinction. However, as we will show below, quantitative approaches are available to rule out effects of accidental homonymy.

2.2. CLICS' data and sources

CLICS (Version 1.0) offers information on colexification in 221 different language varieties covering 64 different language families.[2] All language varieties in our sample comprise a total of 301,498 words covering 1,280 different concepts.[3] Using a strictly automatic procedure, we identified 45,667 cases of colexification that correspond to 16,239 different links between the 1,280 concepts covered by our data.

At present, four sources feed into CLICS: (1) The *Intercontinental Dictionary Series* (IDS, Key and Comrie, 2007), offering lexical data for 233 languages and language varieties. Ideally, datasets for each language contain 1,310 entries, though coverage differs in completeness for individual languages. Of all 233 languages in IDS, 178 were automatically cleaned and included in CLICS. (2) The *World Loanword Database* (WOLD, Haspelmath and Tadmor, 2009), the main goal of which has to do with identifying lexical borrowings, but which nevertheless also provides general lexical data for 41 languages. The vocabularies for the individual languages differ somewhat in their size, ranging somewhere between 1,000 and 2,000 items. 33 of the 41 vocabularies are included in CLICS. (3) Data for four languages represented neither in IDS nor WOLD were added from the *LOGOS* dictionary (`http://www.logosdictionary.org`), a multilingual online dictionary. (4) Additional data for six Himalayan languages was taken from the *Språkbanken* project (University of Gothenburg, `http://spraakbanken.gu.se`).[4]

2.3. Network modeling of CLICS

As mentioned above, there is no guarantee that lexical associations within CLICS reflect conceptual associations. For example, there are three attested links between the concepts 'arm' and 'poor' in the current version of CLICS, which are due to homonymy in some Germanic languages (German, Dutch, and Yiddish).

In order to distinguish strong association tendencies from spuriously occurring associations and to rule out cases of accidental homonymy, List et al. (2013) model cross-linguistic colexification data as a weighted network in which nodes represent concepts and weighted edges between the nodes represent the number of attested colexifications in the data. With the help of *community detection analyses*, strongly interconnected regions in the colexification network can be identified. Communities are groups of nodes in a network 'within which the connections are dense but between which they are sparser' (Newman, 2004). List et al. (2013) apply a weighted version of the community detection algorithm by Girvan and Newman (2002) to a cross-linguistic colexification network consisting of 1,252 concepts translated into 195 languages covering 44 language families. Their analysis yielded a total of 337 communities, with 104 communities consisting of 5 and more nodes and covering 68% of all concepts. A qualitative survey of the largest communities showed that most of them constitute meaningful units and accidental homologies were successfully excluded.

2.4. Limitations and caveats

The data structure in CLICS directly mirrors the data structure of the sources we used. We did not manipulate or reanalyze the data in any way, to the effect that the reliability of CLICS is greatly dependent on the reliability of its sources. Additionally, it should be pointed out that we also cannot rule out the possibility of artifacts arising from automatic data cleaning in cases where textual coding of the data was inconsistent. As for its actual application, it also must be borne in mind that CLICS reflects a certain bias regarding the geographical locations of the languages included in its sources: IDS features many languages from South America and the Caucasus, while WOLD includes a disproportionate percentage of European languages. Hence, the sheer frequency of instances of a particular colexification pattern

[1] See also the example of 'money' and 'silver' in the case study in Section 3.5. below.

[2] This count includes 12 language isolates, and 3 unclassified languages, according to the classification schema of Ethnologue (Lewis and Fennig, 2013).

[3] Since some concepts are expressed by more than one word in the respective languages, the number of words is higher than the expected one (282,880) if multiple synonyms per concept were not allowed.

[4] In all cases, we ignored protolanguages and archaic languages (like Latin and Ancient Greek), and those languages which did not have enough coverage in terms of lexical items.

in CLICS may be misleading insofar as a pattern may seem very robust crosslinguistically, but actually is so only in certain regions of the world. We have not implemented any computational method in CLICS to balance out the picture *a posteriori*. Since we nevertheless want to present potential users of CLICS with the possibility to assess possible areal patterns in the data, we include a powerful visualization that enables them to detect areal imbalances in colexification patterns in individual cases themselves.

3. Visualization

The CLICS database is available online at `http://clics.lingpy.org` and offers its users a search interface to all concepts and crosslinguistic colexifications between concepts. The wealth of information in the database and the various possibilities of exploring the colexifications in the network call for an additional component that makes potentially interesting observations more easily accessible to the researcher. The idea was to equip the database with a visualization component that provides various interactive functionalities and enables users to navigate through the networks of colexifications while at the same time providing more detailed information on the actual language data.

3.1. Web-based visualization

We opted for a web-based implementation of the CLICS visualization in JavaScript using the D3 library (Bostock et al., 2011). The main benefits of a web-based visualization are its platform independence and the fact that users can access it from any device with a browser supporting JavaScript. There is no need for the installation of additional software or for maintenance of the system on the part of the user (Murray, 2010). In addition, links to the descriptions of the external resources can easily be included to allow users to explore the CLICS data in more detail on demand.

3.2. Data preparation

In its current form, the data in CLICS yields a *small world network* in which all nodes are densely connected. Browsing such a dense network is very confusing and provides few insights for the user (see Figure 1). In order to break down the complexity inherent in CLICS, we employed two different strategies to present the data from two different perspectives. According to our first strategy, we decided to split the data into *communities* first. Starting from 1,280 concepts in CLICS which were connected to at least one other concept, we applied the *Infomap* algorithm by Rosvall and Bergstrom (2008) to cluster all concepts into communities. The *Infomap* algorithm requires that weights are defined for the edges of the network. Here we used the number of attested language families per colexification as edge weights. Following a suggestion by Dellert (2014) we further normalized the number of attested language families with the help of Formula 1:

$$W = \frac{C^2}{O_A + O_B - C}, \quad (1)$$

where C is the number of attested language families for the colexification of concept A and concept B, O_A is the number of language families in which concept A occurs, and O_B is the number of language families in which concept B occurs. The *Infomap* algorithm was chosen because of its remarkable performance on the community detection task, both in terms of computation time and quality of results (Lancichinetti and Fortunato, 2009). With the help of this analysis, it was possible to subdivide the 1,280 concepts into 271 communities. Of these communities, 118 are *large*, containing more than five nodes. The large communities cover 65% (828) of all nodes in the original network (1,280). In order to enable the user to quickly identify communities of specific interest, we labeled all communities by taking the concept with the highest degree as representative. The communities do not differ much in size, ranging from 2 ('men's house') to 16 concepts ('fur') with an average of 4.72 concepts per community.

The advantage of the community perspective on CLICS is that it provides an independent automatic clustering of the data. The disadvantage is that this preselection deprives the user of finding alternative, possibly interesting connections between the concepts. Community detection methods are not error-proof, and their performance varies depending on the algorithms being used and the data being analyzed. Even more importantly, most community detection methods are based on restricted decisions when clustering the nodes of a given network into groups. Every node is assigned to one community only. No transitions between communities are possible. In order to offer a less biased perspective on the network, we decided to extract subgraphs for each concept in the data showing its strongest connections up to a certain depth of resolution. The individual subgraphs were constructed as follows: starting from a given concept, we first searched for all of its direct neighbors with which it shared at least 5 colexifications in five different language families. The same procedure was applied to all neighbors which were added to the subgraph in the first run and repeated two times. If the resulting graph for a given concept was too small, a relaxed threshold of 4 colexifications in four different language families was used. In the resulting subgraphs, the longest possible path length between the selected concept and all other concepts is 3. Having excluded 280 subgraphs in which no stronger links between the selected concept and any other concept could be found, this procedure left us with 1,000 individual subgraphs of individual size and structure, ranging from 56 nodes (selected concept 'take') to 2 (selected concept 'north'). When browsing the network representation of the data, the user can select between both perspectives on the data in CLICS, the community perspective, and the perspective of the strongest connections.

3.3. Interactive functionalities

The visualization features various interactive functionalities that are designed to enhance the exploration of the CLICS data at the level of communities or strongest connections. The main component is a flexible force-directed graph layout that displays the concepts as nodes and the crosslinguistic colexifications as edges (see Figure 2). The strength of the force in the edges of the graph is dependent on the number of language families that can be attested to

having lexical associations for the respective concepts that are linked through the edge. As mentioned in the previous section, we decided to provide two different views on the network, one for separate communities obtained from the *Infomap* analysis (Rosvall and Bergstrom, 2008), and one in which the subgraph containing the strongest links for each concept is displayed. With the help of the web interface for network browsing (`http://clics.lingpy.org/browse.php`) the user can look up the respective subgraphs by selecting the concept of interest and specifying the desired 'view.'

The force-directed graph layout ensures that all concepts are neatly arranged according to their similarity as defined by the number of crosslinguistic colexifications. As a result, concepts that are highly connected are located close to each other. To make it easier for users to explore the network that is depicted in the graph, concepts can be dragged to different positions where there is less overlap. The dragging behavior of a concept is activated when mousing over the respective node in the graph (when the cursor symbol turns into a cross hair).

As mentioned above, the edges of the graph represent the number of cases of crosslinguistic colexifications for the linked concepts. For a more detailed view on which languages contribute to the strength of the connections, the user can mouse over the links in the graph to see the forms in the individual languages responsible for the associative link (Figure 3). The list includes additional information on the languages such as their ISO 639-3 language code and family. Furthermore, each entry in the list provides a hyperlink to the original source from where the information is taken.

Each language in the list is attributed a different background color depending on its language family or location

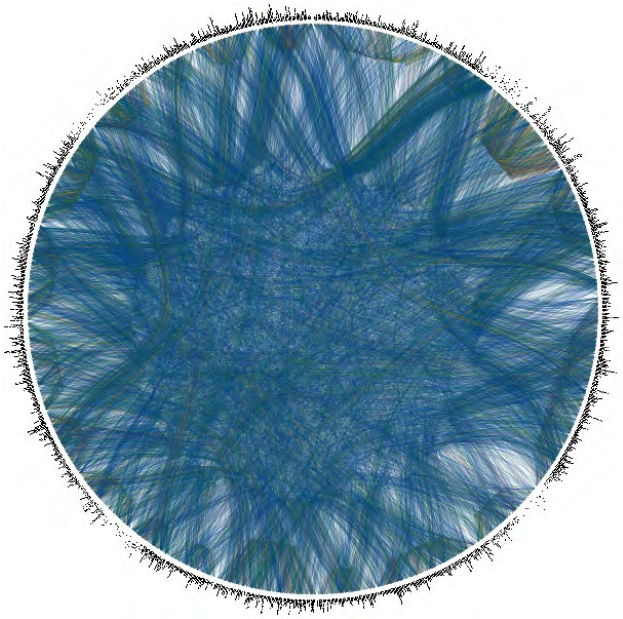


Figure 1: Full network of all 1,288 concepts in CLICS (outer circle) together with their connections. The strength of the connections is marked in different colors, with very strong links represented in red

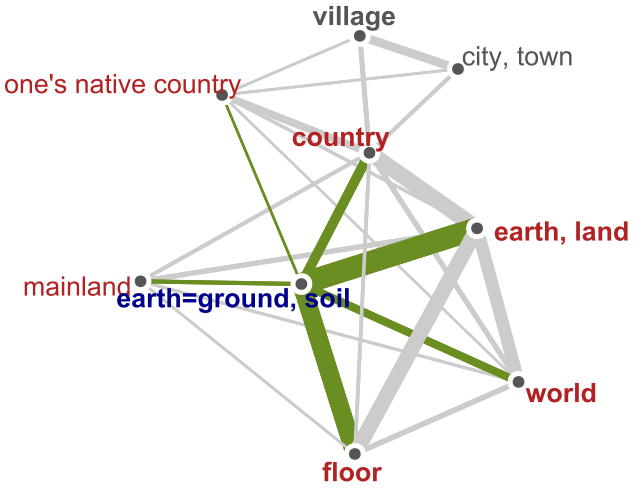


Figure 2: Force-directed graph with mouse-over functionalities highlighting all connected concepts. In order to browse the graph on the CLICS website, use the following URL: `http://clics.lingpy.org/browse.php?gloss=earth,%20land`.

in order to allow for an at-a-glance overview of all languages in the list. The user can choose from a drop-down menu whether to include the genealogical or areal information as the background color. For the genealogical information, all language families are attributed a different color value. Languages belonging to the same language families are therefore given the same background color. Moreover, the list is sorted according to language families. In this way, the user can immediately see how many languages of a given family contribute to the overall strength for the connection at hand.

As to the areal information, the world map is provided with a color gradient as shown in Figure 4. To this end, each position in the world map is attributed a color value using the L*a*b* color space. The color hue thereby indicates the position on the map in terms of the longitude (specifying the east-west position) whereas the lightness of the color represents the position in terms of the latitude information (specifying the north-south position).[5] The mapping from geolocation to color values allows for an easier evaluation of areal patterns in the selected connection. In this regard, users can directly detect whether a certain pattern of colexification is restricted to a certain region of the world or constitutes a more widespread colexification pattern (see the case study in Section 3.5. below). In addition, all languages in the list are displayed with their geographical location on a world map (see Figure 3). Hence, areal patterns can be directly compared to the genealogical information in the list (if the first option is chosen).

In addition to the interactive functionalities described above, the visualization also features a variety of further components that allow for an easier exploration of the database. The graph layout is equipped with panning and zooming functionality that enables the user to navigate through the network graph. Panning is enabled when the

[5] See Mayer et al. (2014) for a different approach of a linguistically informed color gradient of the world map.

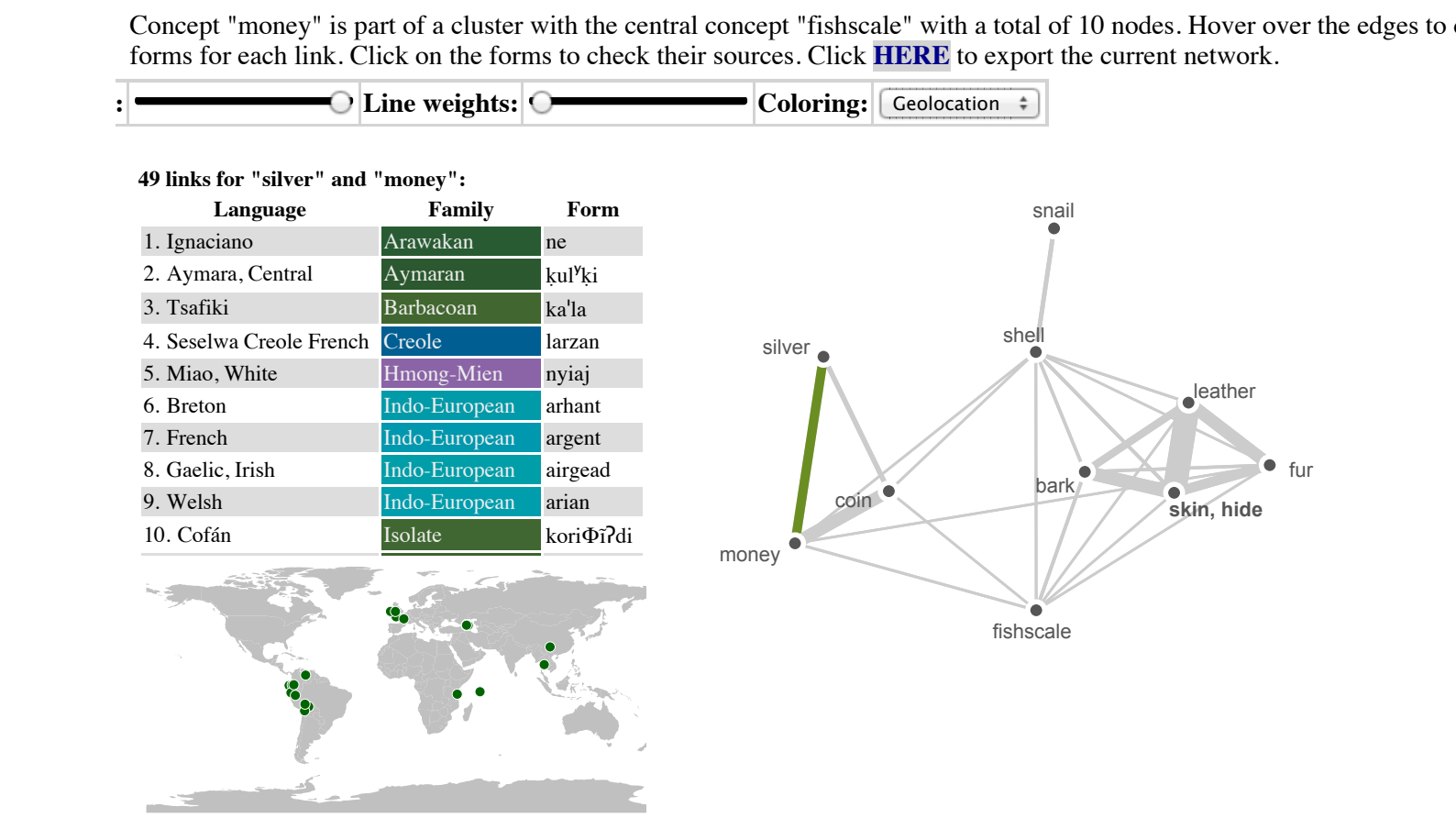


Figure 3: Force-directed graph with mouse-over functionalities showing the strongest links of the concept 'money.' The entries have different background colors depending on their location in the world map (cf. Figure 4). In order to browse the graph on the CLICS website, use the following URL: `http://clics.lingpy.org/browse.php?gloss=money&view=part`.

Figure 4: World map with color gradient. The color represents the location where a language is spoken (see Figure 3 where the background color identifies the location of the language)

cursor changes into a hand symbol when mousing over a link of the graph. The whole graph can then be dragged to a new position. The zooming behavior is activated with the scroll wheel. When mousing over a concept (node) in the graph all connected links and concepts are highlighted in order to provide a better overview of the connectivity of certain concepts (see Figure 2). The control panel of the visualization also includes a slider button that allows the user to show only those edges in the graph with a minimum number of crosslinguistic colexifications.

3.4. Implementation

The visualization is implemented in JavaScript using the D3 library (Bostock et al., 2011).[6] The force-directed graph is generated with the `force()` function from the `d3.layout` module. The layout implementation uses position Verlet integration for simple constraints (Dwyer, 2009).[7] In order to ensure that the concept labels are located close to the concept nodes, a second force layout (with a static weight of 1) is set up for each concept link to the node.

The color values for the world map gradient scale are computed from the two-dimensional geographical coordinates that are given as an input. The latitude [-90;90] and longitude [-180;180] values are thereby normalized between [0;1] and serve as the input for the function `cl2pix`.[8]

```
function cl2pix(c,l){
    var TAU = 6.2831853
    var L = l*0.61 + 0.09;
    var angle = TAU/6.0 - c*TAU;
    var r = l*0.311 + 0.125
    var a = Math.sin(angle)*r;
    var b = Math.cos(angle)*r;
    return [L,a,b];
};
```

[6] `http://d3js.org`

[7] See `https://github.com/mbostock/d3/wiki/Force-Layout` for a description of the implementation.

[8] The code was adapted from the GNU C code by David Dalrymple (`http://davidad.net/colorviz/`, accessed on January 25, 2014) and translated into JavaScript.

The actual HTML color code is generated with the function `d3.lab` from the D3 library, which takes the three values for `[L,a,b]` as input. The main reason for choosing the L*a*b* color space is a smoother transition between different color hues without any visible boundaries. As can be seen in Figure 5, the color gradient in the L*a*b* color space exhibits a much smoother perceptual transition between the color hues on the y-axis. For the coloring of the language families, the background colors are generated with the categorical scale functions of the `d3.scale` module.

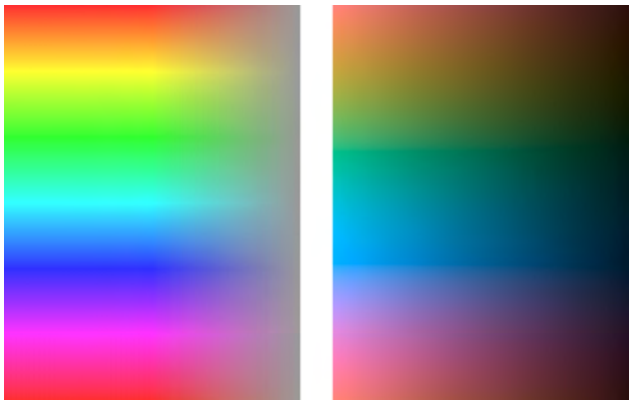


Figure 5: Comparison between two-dimensional color gradients in the L*a*b* (right) and HSV (left) color space.

The dragging and panning functionalities of the graph are implemented with the `drag()` function from the `d3.behavior` module and the SVG `transform` and `translate` attributes. The interactive world map is generated with the `topojson` package and makes use of the `d3.geo` projection module.

3.5. Case studies

In order to illustrate the usefulness of the visualization for the purposes of exploring the database, consider the graph in Figure 3. Among other things, it contains the connection between the concepts 'money' and 'silver.' A subset of the languages and words contributing to this connection are shown on the left where the background color represents the location of the languages. For instance, French contributes to the crosslinguistic colexification because both concepts are realized by the same word (viz. *argent*) in that language. When looking at the areal distribution of the languages, a clear pattern emerges at a glance (see Figure 6 for the full list of languages showing this colexification pattern). Most of the languages contributing to the colexification are from two major regions: Caucasus (marked in blue) and South America (marked in green). However, as mentioned in Section 2.4., this distribution might be an artifact of the general bias for languages of the Caucasus and South America in the underlying databases. In any case, the visualization directly points the attention to this pattern. As the aim of the visualization component is not to replace linguistic research but to guide it, such patterns have to be looked at in more detail by checking the actual data.

Another example deals with the colexification of the concepts 'wheel' and 'foot.' In contrast to the case of 'money'

Figure 6: Languages and words contributing to the connections of lexical associations for the concepts 'money' and 'silver'

and 'silver' above, these concepts at first glance may not immediately suggest a close association. Yet such cases do exist as the link in Figure 7 reveals. The connection links two bigger communities of nodes, including spherical objects on the one hand and parts of the lower body on the other. The list of languages for the connection 'wheel' and 'foot' in Figure 7 clearly shows that the association is restricted to languages of South America. This geographical restriction may reflect semantic borrowing among South American languages, but since the distribution within South America is rather erratic, independent innovation is also a possibility. At any rate, the color coding in the visualization immediately draws the researcher's attention to the potentially interesting geographical patterning.

4. Conclusions and future work

The size and complexity of today's language resources call for a data preparation pipeline that enables researchers to find meaningful patterns among the multitude of different factors that can be taken into consideration. In our view, such a data preparation pipeline necessarily consists of two major parts, both of which are illustrated in the present paper. On the one hand, methods and techniques from data mining or computational linguistics help to detect basic trends or groups of similar objects in the search space. On

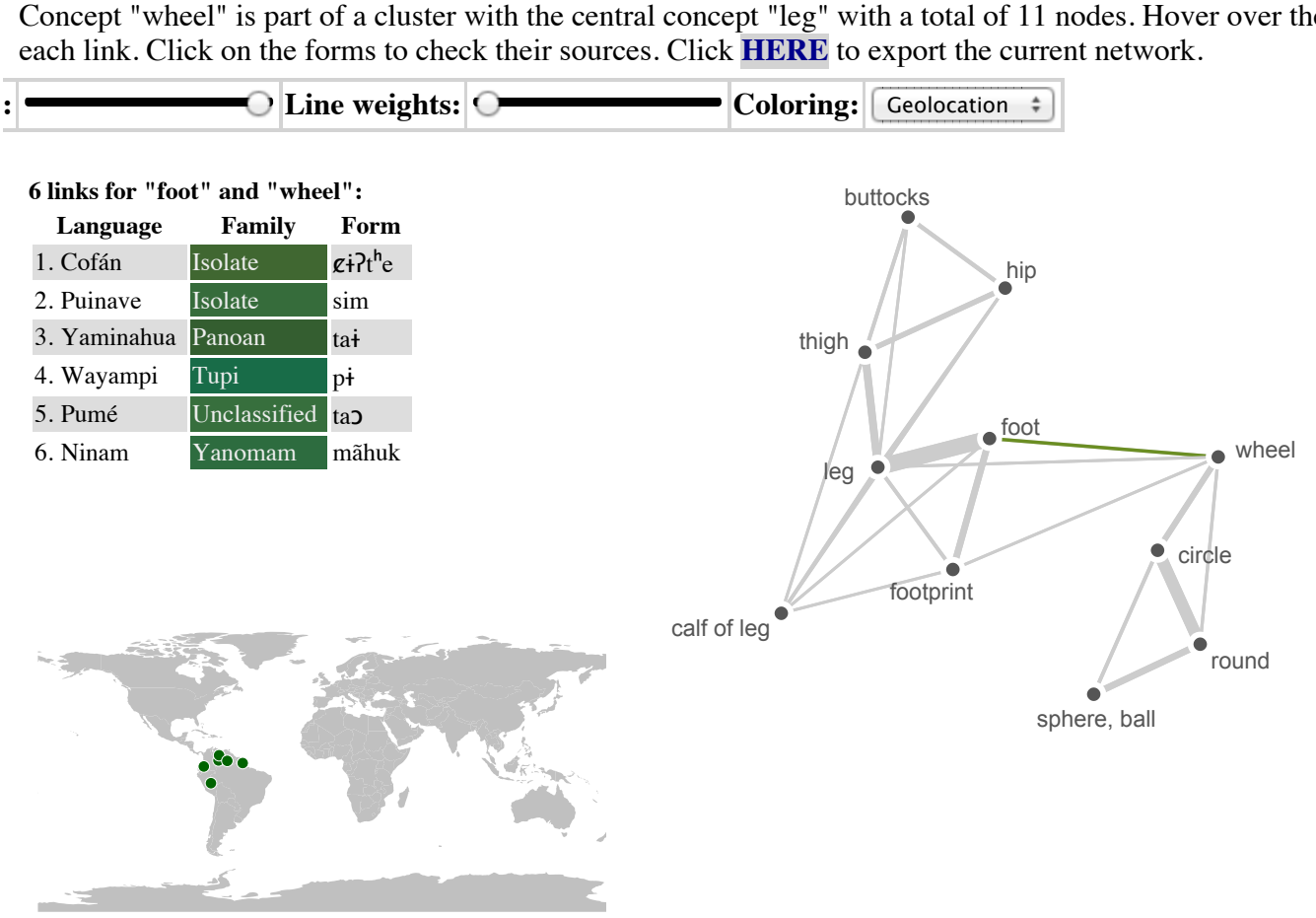


Figure 7: Force-directed graph with areal distribution for the concepts ‘wheel’ and ‘foot.’ In order to browse the graph on the CLICS website, enter the following URL: `http://clics.lingpy.org/browse.php?gloss=wheel&view=part`.

the other hand, the resulting groups or trends are mapped to visual variables in order to make interesting observations readily accessible to human perception.

The CLICS database contains a wealth of information about colexification patterns in the languages of the world. Manually inspecting the large amount of connections in the database, however, is a laborious and time-consuming task that allows for a detailed exploration of individual links but does not capture overall trends in the data. This paper presents an attempt to combine the advantages of human inspection with the strength of a computational approach (Keim et al., 2008).

The CLICS visualization features an automatic preprocessing of the colexification links into so-called communities on the graph, groups of highly connected nodes that reveal a meaningful overall trend in the worldwide patterns of lexical associations. The communities are then graphically represented in a force-directed graph that shows all connections within the various concepts that are included. Interactive components in the visualization allow for a more detailed view of associations at the level of the languages that contribute to the colexification. Mapping the genealogical and areal information on individual languages to colors enables an at-a-glance evaluation of potentially interesting trends in individual colexifications (see the case studies in Section 3.5.). In this way, users can get an overview of the general trends in the data and at the same time have the opportunity to directly inspect the lexical associations.

In future work, we plan to enhance the visualization tool with further interactive components that allow for a better overview of the complete network of colexifications (shown in Figure 1) and facilitate the detection of genealogical or areal trends in the database. The idea is to integrate a sunburst visualization (Stasko and Zhang, 2000) of the genealogical information in order to enable a better overview of the language families that are involved in a given colexification pattern.[9] In addition, we intend to equip the user interface with further interactive components that allow users to explore the database from different perspectives (e.g., compare individual languages in terms of shared lexical associations). All components will be made publicly available online for the (linguistic) research community.

5. Acknowledgements

This work was supported by the German Research Foundation (DFG project ‘Algorithmic corpus-based approaches to typological comparison’ and DFG CRC 991 ‘The structure of representations in language, cognition, and science’) and the European Research Council (ERC starting grant 240816 ‘Quantitative modelling of historical-comparative linguistics,’ and ERC advanced grant 103891 ‘The linguistic past of Mesoamerica and the Andes’).

6. References

Blank, A. (1997). *Prinzipien des Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.

[9] See Mayer et al. (2014) for an example of using sunburst displays to represent the hierarchical structure of language families.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3: Data-driven documents. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)*, 17(12):2301–2309.

Dellert, J. (2014). Lifting a large multilingual dictionary to the level of concepts. Talk held at the "Workshop on historical and empirical evolutionary Linguistics" (February 15–16, 2014, Eberhard-Karls-Universität Tübingen).

Dwyer, T. (2009). Scalable, versatile and simple constrained graph layout. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'09, pages 991–1006, Aire-la-Ville, Switzerland. Eurographics Association.

François, A. (2008). Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In Vanhove, M., editor, *From polysemy to semantic change*, pages 163–215. Benjamins, Amsterdam.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Haspelmath, M. and Tadmor, U. (2009). *World Loanword Database*. Max Planck Digital Library, Munich.

Keim, D. A., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: definition, process, and challenges. In Kerren, Andreas, Stasko, John T., Fekete, Jean-Daniel, and North, Chris, editors, *Information Visualization*, pages 154–175. Berlin: Springer Verlag.

Key, M. R. and Comrie, B. (2007). *IDS – The Intercontinental Dictionary Series*. Springer, Berlin, URL: `http://lingweb.eva.mpg.de/ids/`.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5 Pt 2):056117.

Lewis, M. P. and Fennig, C. D., editors. (2013). *Ethnologue*. SIL International, Dallas, 17 edition. URL: `http://www.ethnologue.com`.

List, J.-M., Terhalle, A., and Urban, M. (2013). Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, pages 347–353, Stroudsburg. Association for Computational Linguistics.

Mayer, T., Wälchli, B., Rohrdantz, C., and Hund, M. (2014). From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. In *Language Processing and Grammars. The role of functionally oriented computational models*, pages 13–38. John Benjamins.

Murray, S. (2010). *Interactive Data Visualization for the Web*. O'Reilly Media, Inc.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, 70(5):056131.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.*, 105(4):1118–1123.

Stasko, J. and Zhang, E. (2000). Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 57–65, Washington, DC, USA. IEEE Computer Society.

The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually

Roberto Theron, Eveline Wandl-Vogt
Department of Computer Sciences and Automation (University of Salamanca)
Institute for Corpus Linguistics and Text Technology (ICLTT, Austrian Academy of Sciences)
theron@usal.es, eveline.wandl-vogt@oeaw.ac.at

Abstract

Historical dictionaries and non-standard language corpora can greatly benefit from a visual access in order to grasp the inherent tangled and complex nature of the knowledge encapsulated in them. Although visual analytics has been used to tackle a number of language and document related problems, most dictionaries are still reproducing the book metaphor in which Web pages substitute the paper and the user experience is only enhanced by means of hyperlinks. Although fields such as dialectology and dialectal lexicography have incorporated Geographic Information Systems and advanced computational linguistics features, spatio-temporal dynamics can be discovered or understood if appropriate visual analytics techniques are used to surpass the idea of these linguistic resources as alphabetically ordered lists. In this paper we present the work carried out in this direction for the Dictionary of Bavarian Dialects in Austria. By means of multiple-linked views an access that fosters the exploratory analysis of the data is enabled.

Keywords: visual navigation, variational linguistics, historical corpora

1. Introduction

Variation in language use is one of the main concerns in linguistics and its study has been greatly benefited by the use of computers. The early adoption of electronic corpora have paved the way to deeper understanding of the variations that can be found among speakers, associated with age, gender, geography, etc. Thus, electronic corpora have enabled approaching the *life cycle* of linguistic variability in order to shed light on the dynamics of a given language across time and space.

By combining the strengths of machine and human data processing, visual analytics science has established itself as a very successful approach to a great variety of problems across domains, in which the complex nature of the data and its volume have proven to be too great a challenge for previous efforts. The visual analytics community has emphasized the fact that spatio-temporal analysis is not exclusively needed and done by highly qualified specialists (Andrienko et al., 2010); on the contrary, any person concerned about the development of their communities, regions or countries would want to take advantage of the data being collected. This is the case for any language for which vast amounts of information have been recorded and made available through dictionaries and corpora.

Although different methodologies and techniques have been developed from corpus and computational linguistics, and in some cases those advances have been coupled with different forms of visual interactive aids (Lee and Kretzschmar, 1993; Masron et al., 2005), the majority of electronic dictionaries are still regarded as flat lists of entries. This paper presents our preliminary results in an effort aimed at providing a different access to historical dictionaries and corpora by means of visual analytics techniques, in which the exploratory analysis is fostered so the idea of a flat dictionary is surpassed.

In the past, there have been several visualization efforts mainly aimed at escaping the idea of dictionaries as flat lists of entries. During the past decade the navigation of the web of words by means of hyperlinks was enabled and these word relationships are often visualised using different types of graph drawing techniques (Visual Thesaurus[1]). While this is an advantage for dictionary users, there are still several aspects that can dramatically change the users' experience: an enhanced access to the wealth of sources related to dictionaries tailored to the particular needs and tasks of such users can indeed surpass this very notion of a flat dictionary.

The previous statement is especially relevant when non-standard data are considered. Space and time become the steering wheel of any inquiry. In this regard, there are few works that take full advantage of visual interactive access to non-standard dictionaries. In (Theron et al., 2011; Theron and Fontanillo, 2013) one of the first efforts in this direction can be found. Furthermore, in (de Vriend et al., 2011) it is suggested that a working environment offering full support for using visualization as a research tool could take dialect geography into the era of eScience.

The rest of the paper is organised as follows: in the next section, we briefly explain the case of the Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich [WBÖ], from its origin to its present form. In the third section we discuss how the WBÖ data sources have been enhanced thanks to GIS technology. The fourth section is devoted to show how the same data can be exploited from a visual analytics approach, providing a more comprehensive user experience. Finally, we expose the main conclusions and future works that we envision.

[1] Visual Thesaurus [www.visualthesaurus.com/] [accessed: 15.1.2014]

2. 100 Years of Variational Linguistics in a Nutshell

Dealing with non-standard language corpora concerning time (historical language) and space (variants, regional language, dialects) is a challenge, especially when focusing usability, navigation and access. To exemplify the challenges and suggested solutions, we picked the example of the Dictionary of Bavarian Dialects in Austria.

The main reasons for doing so are obviously: the dictionary and its material are digitally available, yet the project stands for the type of a traditional, sophisticated, so called territorial dictionary. The materials in the dictionary and those ones used for the dictionary are highly heterogeneous. The dictionary and its corpus offers historic (8th century -) as well as recent materials (- 1998). Furthermore, the corpus offers materials from written sources as well as from questionnaires of spoken language. Professionals and laypersons collected the data in a period for about 100 years in written and digital format.

The project started in 1911 to introduce novel standards in lexicography and give a complete and detailed overview of the Bavarian dialectal variants in the Austrian-Hungarian Monarchy and the Kingdom of Bavaria, by presenting the complete lexicons of these areas with detailed definitions and contextualized examples, by recording the authentic pronunciation, by defining the grammatical coding for each word entry, by tracing the etymology of each lexical item and by registering expert knowledge in the fields of rural techniques, traditional folk medicine and customs. From 1913-1932 109 questionnaires and 9 auxiliary questionnaires with approx. 24,000 detailed questions were sent to selected municipalities. In 1961 the Austrian and Bavarian Academy decided to publish two separate parts of the lexicon: Part I dealing with Austria (with the exception of the province of Vorarlberg) and the (former) German-speaking parts across the current borders of Italy, Slovenia, Slovakia, Hungary and the Czech Republic, and part II dealing with the Bavarian dialects in Germany. Part I, the WBÖ, has been published since 1963 (4 volumes, 8 parts: A-E).

In 1993 the project *Datenbank der bairischen Mundarten in Österreich (DBÖ)* / Database of Bavarian dialects in Austria was initiated aiming at the digitalisation of the material for the dictionary and of source material (e.g. paper slips, cited texts) as well as background information (e.g. cv of collectors and co-workers).

In 1998 a rationalisation concept was issued targeting the dictionary as a (virtual) unit consisting of the printed dictionary and the complementary database. In 2008 the first step in the development of the system *Datenbank der bairischen Mundarten in Österreich* / Database of Bavarian dialects in Austria electronically mapped (dbo@ema) was taken. Since 2010 (Wandl-Vogt, 2010), some thousand example entries are digitally available; furthermore, background data and registers are online and interactively accessible.

In 2013, first steps towards a machine-readable dictionary and connection with Linked (Open) Data were taken (Wandl-Vogt and Declerck, 2013).

3. Access to the dbo@ema

In its current form, dbo@ema can be explored in a traditional way: the user can access the data according to different categories (lemma, bibliography, person, location). For each of these categories, the user would need to access the desired information following a list of pages ordered alphabetically.

Figure 1: Access to the dbo@ema by means of alphabetical lists.

Figure 1 shows a screenshot of dbo@ema in which a list of lemmas is used to provide direct access to the desired lemma. Once the user has selected a particular item of information, a page containing relevant information is shown, e.g., type of lemma, multimedia sources, etc. (see Figure 2).

Figure 2: Details of the lemma 'Kufe' including multimedia sources.

The same procedure can be followed when searching for persons, regions, etc. In these cases, since most of the information is georeferenced, an interactive map is included, so the user can grasp the spatial influence of the searched item.

Figure 3 shows the details page for Werner Bauer, including personal information and a list of sources related to this author.

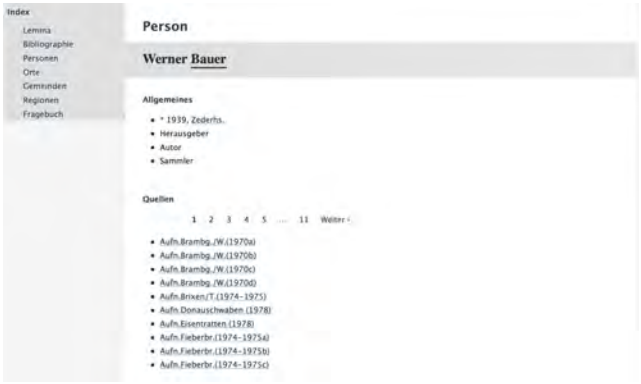


Figure 3: Details related to author Werner Bauer.

Accordingly, Figure 4 shows an interactive map of the spatial distribution of data related to the same author.

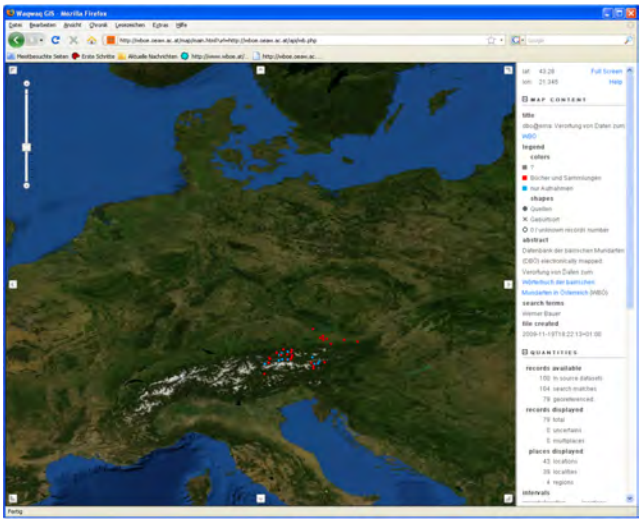


Figure 4: Spatial distribution of data related to the author Werner Bauer..

Finally, the user can go inspecting the details of each spatial point in the map, available by means of tooltips (Figure 5).

Although this provides a method of accessing the spatial distribution of the data stored in the corpus, it is only useful for directed searches, where the user actually knows what she is looking for. Exploratory tasks are not supported, since the context of information is lost once the user access the information of a given lemma, person, location, etc.

Furthermore, although details can be navigated by means of hyperlinks, maps are independent, so a click on a hyperlink directs the users to a new Web page; also the interactions in the map are limited to its current information and the user cannot access further information from the tooltips. Thus, we approached the same data sources from a different/complementary perspective; we introduced some visual analytics techniques in order to provide a more comprehensive access to the corpus data.

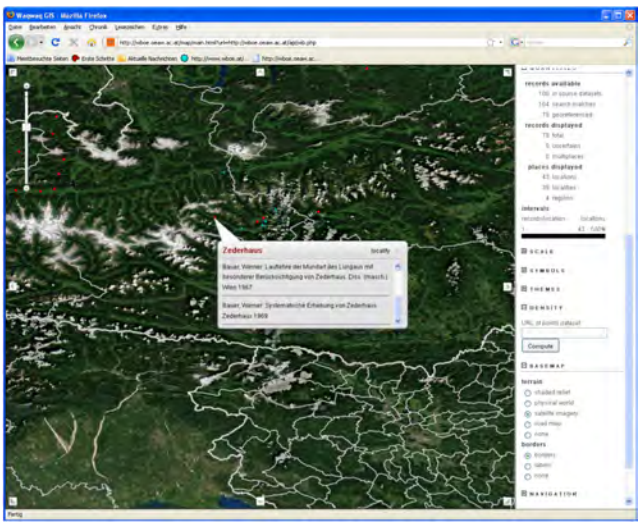


Figure 5: The user may explore the spatial distribution of the data and access individual information of a particular point..

4. Visual Analysis in the dbo@ema

One of the first issues that we wanted to address was providing an overview of the spatial distribution of the data that in turn could be further explored by means of user interaction. We wanted also to maintain the access to all the details originally available in the Web version.

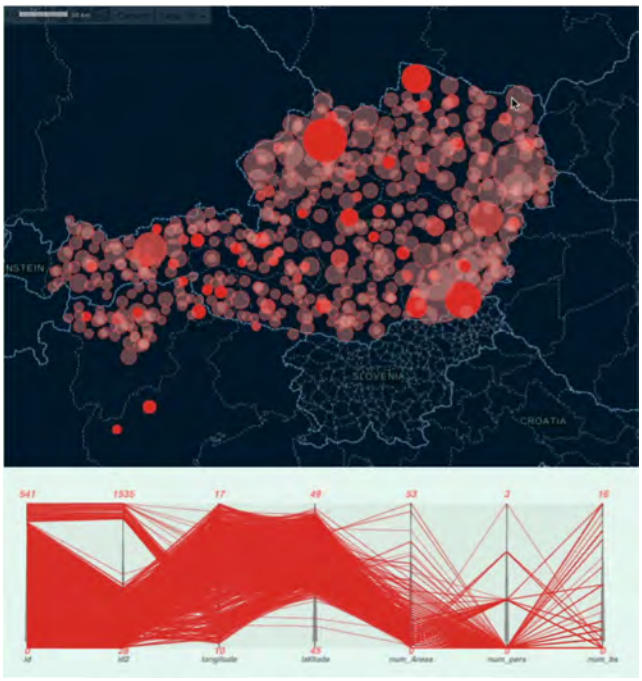


Figure 6: Graduated point symbol maps and parallel coordinates views that enable visual analysis in dbo@ema.

Figure 6 shows two views conveying quantitative information related to regions (it can be done for communities and localities similarly). For each region the number of subregions, persons, lemmas and documents is computed. On

top, a graduated point symbol (Flannery, 1971) or bubble map is shown, in which the size opacity of each bubble. In the case of figure 6, the bubble size depends on the number of sub regions, while the opacity depends on the number of documents related to that region. These visual encodings can be used to convey results of dialectometry studies such as maximum similarity indices.

At the bottom of figure 6, a parallel coordinates (Inselberg, 1985) plot (PCP) can be seen, which conveys the same information by means of polylines (each polyline is linked to a bubble). This way the user can explore the whole dataset and discover patterns related to distribution of sources, etc. The PCP is interactive, so the user can filter data according to different criteria (i.e., filter out the regions that have less than 3 persons related to them and within a particular geographic area (longitude and latitude values), and so on). Finally, once a particular region is on focus, the detailed information is shown, and further information can be retrieved (e.g., details of documents or persons) and shown upon user's request). Both views feature several ways of interaction that enable a typical sense-making loop, going from the overview, filtering data, to details on-demand (Keim et al., 2008).

Figure 7 shows an example in which the user has focused the analysis on the central area of the map, using filters for the latitude and longitude of the data points. The user is inspecting the region Kärnten with a high (16) number of related documents.

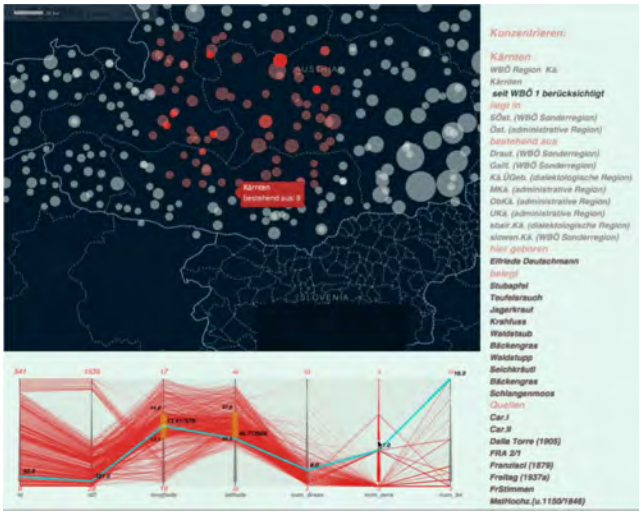


Figure 7: By means of interactions, the user can easily explore the spatial dynamics of dbo@ema.

5. Conclusion

The use of visual analytics can greatly enhance the experience of users, enabling exploratory analysis that contribute to the discovery of spatio-temporal dynamics that would remain hidden with traditional flat access to dictionaries and corpora. We have assessed the potential of exploration of georeferenced data for regular dictionary visitors.

In this first version of our prototype we have used mainly information available in the dbo@ema, but it has been designed from a visual analytics approach, to facilitate the addition of dialectometry studies such as maximum similarity indices.

However, the temporal dimension has not been fully exploited, and the visual analytics research has developed efficient techniques that can be integrated in our work.

6. Acknowledgements

This research was supported by the Spanish MINECO grant FI2010-1623.

7. References

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M. J., Schumann, H., and Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600.

de Vriend, F., Boves, L., Van Hout, R., and Swanenberg, J. (2011). Visualization as a research tool for dialect geography using a geo-browser. *Literary and Linguistic Computing*, 26(1):17–34.

Flannery, J. J. (1971). The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 8(2):96–109.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). *Visual analytics: Scope and challenges*. Springer.

Lee, J. and Kretzschmar, W. A. (1993). Spatial analysis of linguistic data with gis functions. *International Journal of Geographical Information Science*, 7(6):541–560.

Masron, T., Rainis, R., Ghazali, S., Lah, S. C., Sazali, A., Ghani, A. A., and Ghafor, S. A. (2005). Using gis to grasp dialectal variation. In *Proceedings of Map Asia 2005 Conference, Jakarta, Indonesia*, pages 1–10.

Theron, R. and Fontanillo, L. (2013). Diachronic-information visualization in historical dictionaries. *Information Visualization*. Published online, doi: 10.1177/1473871613495844.

Theron, R., Fontanillo, L., Esteban, A., and Seguín, C. (2011). Visual analytics: A novel approach in corpus linguistics and the nuevo diccionario histórico del español. In Carrió, M. L. and Candel, M. A., editors, *Actas del III Congreso Internacional de Lingüistica de Corpus. Las Tecnologías de la Información y las Comunicaciones: Presente y futuro en el Análisis de Corpus.*, pages 335–342.

Wandl-Vogt, E. and Declerck, T. (2013). Mapping a traditional dialectal dictionary with linked open data. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., and Tuulik, M., editors, *Proceedings of the eLex 2013 conference, Electronic lexicography in the 21st century: thinking outside the paper.*, pages 460–471.

Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis*, 2:35–53.

Finding Relevant Features for Statistical Speech Synthesis Adaptation

Pierrick Bruneau[*], Olivier Parisot[*], Amir Mohammadi [**], Cenk Demiroğlu[**], Mohammad Ghoniem[*] and Thomas Tamisier [*]

[*]Centre de Recherche Public - Gabriel Lippmann
Informatics, Systems and Collaboration Department
41, rue du Brill, L-4422 Belvaux (Luxembourg)
lastname@lippmann.lu

[**]Electrical and Electronics Engineering Department
Özyeğin University, Istanbul (Turkey)
amir.mohammadi@ozu.edu.tr, cenk.demiroglu@ozyegin.edu.tr

Abstract

Statistical speech synthesis (SSS) models typically lie in a very high-dimensional space. They can be used to allow speech synthesis on digital devices, using only few sentences of input by the user. However, the adaptation algorithms of such weakly trained models suffer from the high dimensionality of the feature space. Because creating new voices is easy with the SSS approach, thousands of voices can be trained and a Nearest-Neighbor (NN) algorithm can be used to obtain better speaker similarity in those limited-data cases. NN methods require good distance measures that correlate well with human perception. This paper investigates the problem of finding good low-cost metrics, i.e. simple functions of feature values that map with objective signal quality metrics. We show this is a ill-posed problem, and study its conversion to a tractable form. Tentative solutions are found using statistical analyzes. With a performance index improved by 36% w.r.t. a naive solution, while using only 0.77% of the respective amount of features, our results are promising. Deeper insights in our results are then unveiled using visual methods, namely high-dimensional data visualization and dimensionality reduction techniques. Perspectives on new adaptation algorithms, and tighter integration of data mining and visualization principles are eventually given.

Keywords: Speech Synthesis, Speaker Adaptation, Feature Selection, Visual Analytics

1. Introduction

Statistical speech synthesis (SSS) and unit selection based synthesis are two leading technologies in the speech synthesis field (Black et al., 2007). Despite its lower average quality compared to the unit selection approach, SSS has some advantages that make it attractive especially in consumer electronics applications. An important advantage of the SSS approach is the ability to adapt to a target speaker with a couple of minutes of adaptation data (Yamagishi and Kobayashi, 2007). In this approach, a speaker-independent (SI) model is generated first and the speaker-adapted (SA) voice model is found by statistically adapting the SI model to the target speaker using speech data collected from the target speaker. Thousands of voices have thus been generated with SSS using speech databases prepared for speaker-independent recognition systems (Yamagishi et al., 2010).

Even though new voice models can be easily generated with SSS, differences between those voices have only been investigated perceptually through listening tests and objectively using the Root-Mean-Square-Error (RMSE) criterion (Yamagishi et al., 2010). To date, mathematical and visual analysis of the voices have been very limited because of the huge dimensionality of the models. Moreover, considering the high number of model features w.r.t. the number of speaker models at hand, it can be reasonably hypothesized that the best adaptation paths lie on a lower dimensional manifold. Hence, exploration of the models with visualization and data mining can lead to a better understanding of the adaptation algorithms and point to which features or feature transformations are effective at identifying differences between speaker models after adaptation. Such insight can also enable distance measures that correlate well with the perception of speaker similarities in listening tests. Novel distance measures through analysis of the speaker-adapted voices can be helpful in developing adaptation algorithms that work with very small amounts of adaptation data. One way to achieve this is the use of Nearest-Neighbor methods that require good distance measures between statistical voice models to work well. When such distance measure is available, relying on nearest neighbors may significantly reduce the amount of adaptation data needed from the target speaker. Moreover, generating target-specific SI models using the nearest neighbors can also lead to more effective speaker adaptation (Ijima et al., 2013).

After a short review of the SSS approach in Section 2., we put a greater emphasis on the objective evaluation of the SA model quality, and derive a robust performance index in Section 3.. In the latter section, we show that the problem at hand is intractable, and thus reformulate it and propose a statistics-based heuristic for approximate solutions in Section 4.. Insight is gained from the obtained results using visual inspection tools and visual summary statistics, as shown in Section 5.. Our findings are summarized in Section 6., along with perspectives to this work.

2. Overview of the Statistical Speech Synthesis Approach

SSS systems work in several steps. First, parameters that represent short-time speech spectra are extracted. Distribution of those parameters are then modeled using Hidden Semi-Markov Models (Yamagishi and Kobayashi, 2007).

Those models are used to regenerate parameters during synthesis. Parameter extraction and modeling algorithms are summarized below.

2.1. Parameter Extraction and Modeling

During the training phase, both spectral features (such as Line Spectral Frequencies (LSFs) or Mel-Generalized Cepstral (MGC)) (Tokuda et al., 1994) and pitch (log fundamental frequency (LF0)) parameters are extracted. The spectral and pitch features are fused together to create the static feature vector which is then concatenated with delta and delta-delta features.

Phonemes are modeled with N-state Hidden Semi-Markov Models (HSMM) where the duration (d) of each state i, $\Pi_i(d)$, is modeled by a Gaussian distribution, $N(\mu_{d,i}, \Sigma_{d,i})$. Spectral features are typically modeled with a multivariate Gaussian distribution $N(\mu_{c,i}, \Sigma_{c,i})$. Similarly, pitch features are modeled with multivariate Gaussian distribution $N(\mu_{p,i}, \Sigma_{p,i})$ for voiced states. These acoustic model parameters are trained with a maximum likelihood method approach (Zen et al., 2007).

2.2. Decision Tree based State Clustering

Typically, there is not enough training data to train all states observed in the database. This problem is circumvented by clustering HMM states, and have the states in a given cluster share the same training data (Shinoda and Watanabe, 1997). Top-down clustering with decision trees is then generally used, with nodes splitting according to linguistic questions. When it is not possible to split a node because of an insufficient amount of data for the child nodes and/or an insufficient reduction of the within-child entropy, node splitting stops. The states clustered at the leaf nodes of the tree are then tied together.

In text-to-speech systems, it is critical to take into account the context of phonemes to synthesize them as close to natural as possible. In the SSS approach, phonemes have labels that contain information about the phone-level, syllable-level, word-level, phrase-level and utterance-level context information in addition to syntactic features such as part-of-speech tags and intonation tags (e.g. TOBI endtones for sentences). This contextual information is used at the nodes of the decision tree to cluster phoneme states that have different labels but are found to be acoustically similar.

3. Objective Evaluation of Speaker Similarity

3.1. Data Preprocessing

The experiments in this paper use the Wall Street Journal speech database (Paul and Baker, 1992). Two hours of data from four male speakers are used to train an average voice. The average voice is adapted using the linear regression algorithm in (Nakano et al., 2006) for 157 male speakers to create the SA models. 150 utterances from each speaker are used to create the voice models.

75-dimensional mean vectors (spectral energy + MGC features) are used to represent the short-time speech spectrum (Tokuda et al., 1994). Each voice model has 5 trees and each tree contains 919, 922, 1378, 1313, and 960 states respectively. This amounts to 5492 states in each model and each state has a 75 dimension mean vector. By concatenating all mean vectors of all states from all trees, each speaker is represented with a supervector that has a dimension of 411900 (75 $\times$ 5492).

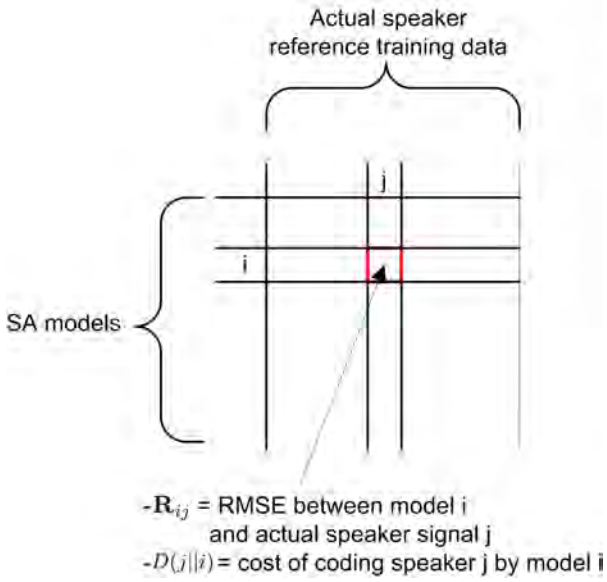


Figure 1: Interpretation of the RMSE matrix cells.

3.2. RMSE Criterion

The RMSE criterion, commonly utilized in the SSS field (Yamagishi et al., 2009), is used to evaluate the voice similarity between a SA voice model and an actual speaker signal. Speech signal is first synthesized following the SA model, and aligned to the actual signal, ensuring a meaningful comparison between the two samples. The RMSE between the L-dimensional aligned MGC features:

$$x(0), x(1), ..., x(L-1) \tag{1}$$

and:

$$y(0), y(1), ..., y(L-1) \tag{2}$$

is then computed as (Imai et al., 2009):

$$\sqrt{\sum_{m=0}^{L-1} (x(m) - y(m))^2 / L}. \tag{3}$$

3.3. Nearest-Neighbor Performance Index of SA Models

As stated in the introduction, this work focuses on the study of SA voice models, and specifically how their features may correlate with the RMSE metric. The set of SA models can be seen as the 157×411900 matrix $\mathbf{M}$, with $\mathbf{M}_{i.}$ the numeric vector defining the i^{th} SA model.

A RMSE matrix $\mathbf{R}$ is generated from this set of models: $\mathbf{R}_{ij}$ is the RMSE between speech that has been generated by $\mathbf{M}_{i.}$, and the actual speaker signal that was used to learn $\mathbf{M}_{j.}$. Let us note that this measure is dissymmetric: this

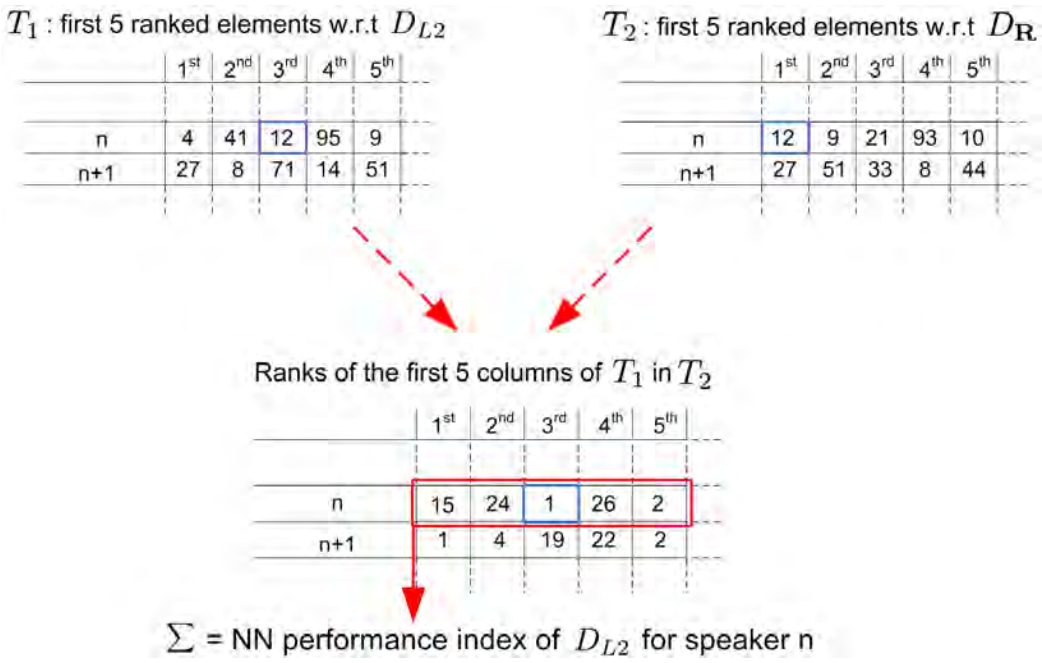


Figure 2: Illustration of the NN Performance index computation. A specific example is highlighted in blue.

observation led us to define the *divergence* between models $D(i||j)$ as the cost of coding model i by model j. As illustrated in Figure 1, this divergence should be read in the columns of the matrix $\mathbf{R}$, because the coded model, i.e. the actual signal, is mapped to the column indexes. We thus define $D_{\mathbf{R}}(i||j) = \mathbf{R}_{ji}$.

Likewise, let us define d_{L1}, d_{L2} and $d_{\cos}$ respectively as the L1, L2 and cosine distances (that are by extension valid divergences, respectively D_{L1}, D_{L2} and $D_{\cos}$) between rows in $\mathbf{M}$. A NN search amounts to computing a *ranking*, materialized by a permutation $p(D, i, \mathbf{M})$ of $1 \dots 157$. Its elements $\{p_j\}_{j=1\dots157}$ get their value according to the following constraint:

$$j < k \longleftrightarrow D(i||j) < D(i||k) \tag{4}$$

Using this formalism, we investigate to which extent D_{L1}, D_{L2} and $D_{\cos}$ are able to map $D_{\mathbf{R}}$. Additionally, we hypothesize that distances computed on the sole static features, without the energy feature (i.e. 24 features per state vector instead of 75) could be sufficient to perform this mapping. The exclusion of the energy feature arises from the observation that the microphone volume is not much likely to characterize the speaker.

Specifically, for every speaker, and for each divergence function, we computed the permutations according to equation (4). The permutations parametrized by $D_{\mathbf{R}}$ serve as a gold standard, and we observe how the first 5 cells of other permutations rank in them. Summing these relative ranks then gives a robust index of the performance of a given divergence in the NN task. This approach is summarized by Figure 2, and is further coined as the *NN performance index*. This index should be as low as possible. This sum is bounded from below by 2355, that is `157 * sum(1:5)` in the case of identical permutations. Results for candidate distance functions are reported in Table 1. In the remainder of this paper, performance indexes are implicitly computed against $D_{\mathbf{R}}$.

	d_{L1}	d_{L2}	$d_{\cos}$
all features	17856	18796	19974
selected features	14211	**13595**	18358

Table 1: NN Performance indexes for every possible setting.

We see that the best results are obtained with the L2 distance defined over the static subset of features. But, considering the margin between the lower bound and the best result, a significant progression margin remains.

However, we have no clue on the optimality of the above-mentioned static selection of features w.r.t. the performance index. The obvious path is a brute-force algorithm, trying all possible combinations of features, for all possible metrics (and assuming some acceptable solution lies in the set of such potential Minkowski and cosine distances) and ranking them according to their NN performance index. But the complexity of this approach is exponential in the number of features, and is tractable up to about 20 features, way below the dimensionality of the SA models. We thus face a difficult optimization problem, for which a heuristic has to be found.

4. Statistical Analysis for a Tractable Heuristic

This section discusses how the problem can be reformulated under a more tractable form, and describes a statistical approach for tentative solutions.

4.1. Distortion as an Objective Function

Distance-like (i.e. symmetric and 0 for self) metrics are mandatory for many theoretical frameworks, but unfortunately $D_{\mathbf{R}}$ does not fulfill this requirement. The most straightforward way to make it symmetric is to define $\mathbf{R}^+ = \mathbf{R} + \mathbf{R}^T$, and use $d_{\mathbf{R}^+}(.,.)$, the dissimilarity induced by $\mathbf{R}^+$. To enforce another property of distances, we force the diagonal of $\mathbf{R}^+$ with $\mathbf{0}$. But prior to any analysis, we have to evaluate how faithful to $D_{\mathbf{R}}$ is a NN search according to $d_{\mathbf{R}^+}$. The NN performance index (see Section 3.3.) is 4440 in the latter case, which indicates a strong similarity (albeit not perfect). We later on assume it is sufficient for our purpose, and justifies the use of $\mathbf{R}^+$ when symmetricity is required.

Having a reference pairwise symmetric dissimilarity matrix (i.e. $\mathbf{R}^+$), and a candidate distance matrix $\mathbf{D}$ on the same set of elements, it is tempting to state that the problems of maximizing the NN performance index, and the proximity (or matching) of $\mathbf{D}$ to $\mathbf{R}^+$, are equivalent. Though not trivial, the latter problem seems indeed much more affordable. Some rigorous approaches may be derived, for example by adapting the metric learning and kernel alignment literature (Xing et al., 2002; Cristianini et al., 2002; Song et al., 2007; Cortes et al., 2010) to the non-standard requirement of a ground truth similarity matrix (instead of an attribute for classification or regression, classically), constrained by the selection of a minimal subset of features. More heuristically, manifold computations and correlation analyses could also be conducted.

Such approaches would optimize the degree of agreement in two distance matrices. This can be measured by the distortion of a candidate with respect to the reference matrix. In a nutshell, we compute a percentage that indicates, for each pair of elements, to which extent its average distance deviates from the reference. For further details, the reader may consult (Aupetit, 2007).

The distortions specific to each SA model for d_{L2} are displayed in Figure 3. It shows models as points in a 2D space, obtained from the 2 first components of the kernel PCA projection that uses a Gaussian kernel function on d_{L2} (Schölkopf et al., 1998). In short, kernel PCA is a non-linear projection method, in which components are the orthogonal axes that explain the most the pattern of L2 distances in $\mathbf{M}$. The average distortion w.r.t. $D_{\mathbf{R}^+}$ associated to each model colors the Voronoi cell of the respective point in Figure 3 using a heat color scale, mapping low values to a light shade.

Two clear clusters noticeably stand out in Figure 3. This observation was not raised prior to this study, and understanding how they arise might give a supplementary cue for analysis. The view also highlights in which regions of the model data space the distortion w.r.t. the ground truth metric is the largest.

Unfortunately, there is no theoretical guarantee that optimizing distortion patterns monotonically increases the NN performance index. We illustrate this by noting that the median of the distortion values is 17.2% when using d_{L2} and the full set of features, whereas it is 18.2% when using the same distance function on the static subset of features. This somehow contradicts the results reported in Table 1. Distortion patterns can thus be occasionally employed, but care must be taken for their interpretation.

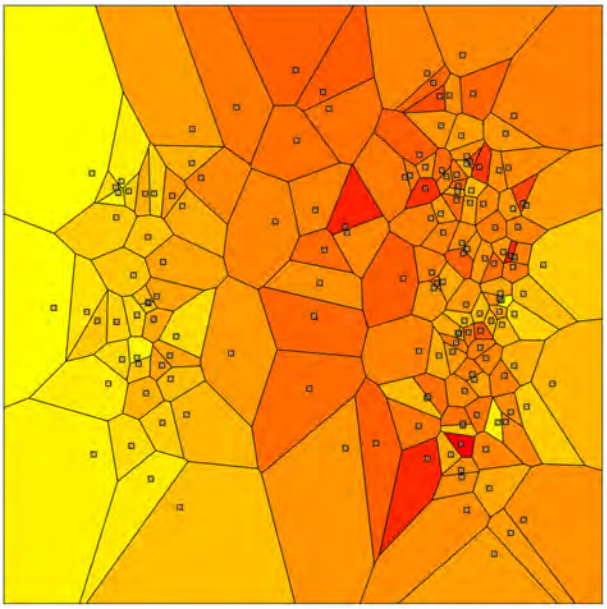


Figure 3: Kernel PCA projection the SA models using D_{L2}. Average distortions of D_{L2} with respect to $D_{\mathbf{R}^+}$ are mapped to a heat scale, and overlaid on the Voronoi cells of the respective SA model glyphs.

4.2. Using LASSO and Ridge Regression

In this section, we propose to simplify the problem, by approximating it as a regression problem with an excessive number of predictors. To this aim, with inspiration from the end of Section 4.1., $\mathbf{R}^+$ is converted to a set of potential target variables by computing its kernel PCA projection with a Gaussian kernel function, and retaining its components (i.e. the two major of which are used as plotting coordinates in Figure 3).

We thus came up with the following hypothesis: the relevant features are those that are able to jointly predict the value of a kernel PCA principal component. Regularized regression models, such as LASSO or ridge regression (Tibshirani, 1996; Friedman et al., 2008),[1] solve this problem by finding the smallest subset of features able to fit a target variable (i.e. a principal component).[2] Joint regression of several variables is theoretically possible, but the size of the problem precludes it. Instead, we choose the following approach:

- We perform the separate regressions of the 20 main kernel PCA principal component of $\mathbf{R}^+$ (with ei-

[1] We used the Elastic regression method (Friedman et al., 2008) parametrized close to, but not exactly equal to, pure ridge regression. The latter indeed uses L2 penalty, notably inefficient for feature selection.

[2] Regression practitioners would object that our problem is largely degenerate, i.e. the number of variables is so large that despite a perfect fit, the empirical risk for previously unseen predictions is ridiculously large. But our objective here is not to ensure a good generalization, but find the most parsimonious fit.

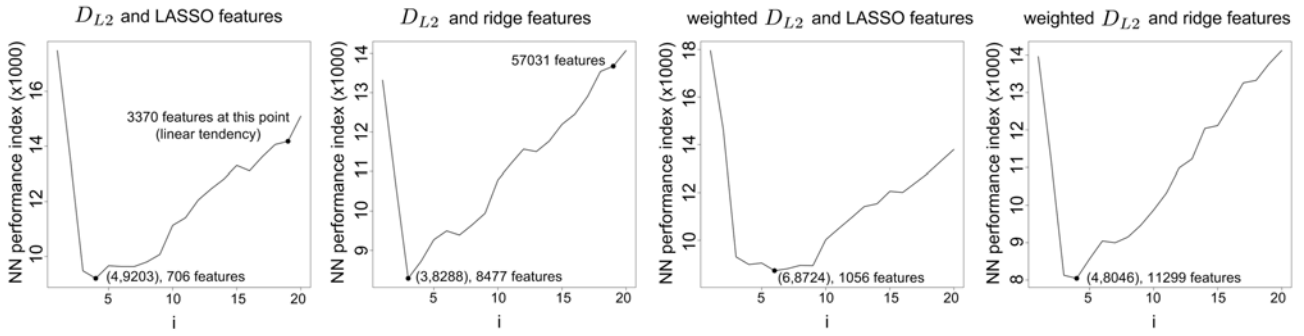


Figure 4: NN performance indexes of feature sets selected by a varying number of regression predictors. Empirical optima are highlighted for each type of predictor.

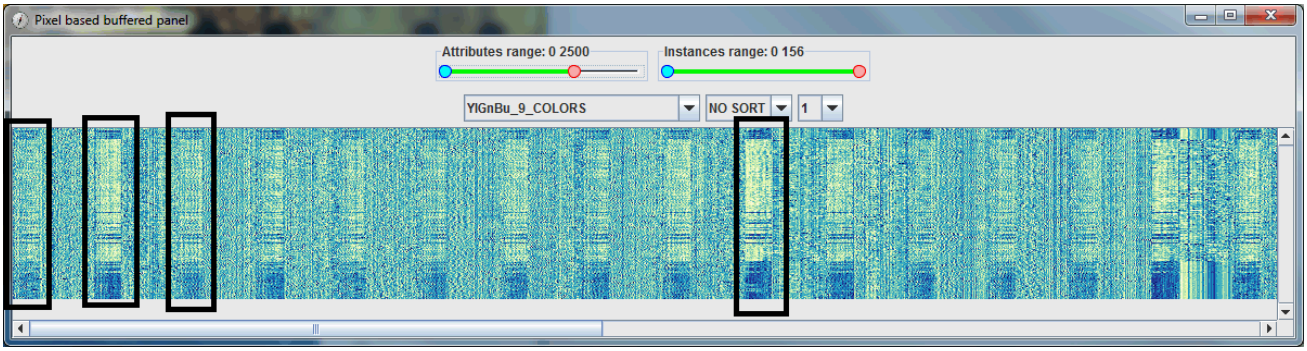


Figure 5: Pixel-based view the SA models, with a filter showing the first 2500 features. Blocks of static features are highlighted.

ther LASSO or ridge). The regression of the k-th component selects a subset of the model features (i.e. the regression model defines a subset of non-zero coefficients according to an expression akin to $y = \sum_{p=1}^{411900} a_p x_p$). Let S_k be this set.

- For $i \in 1 \ldots 20$, we form the sets $\cup_{k=1}^{i} S_k$, i.e. the union of features selected by the regression of main components up to rank i.

In Figure 4, we present the results of NN searches using L2 distances defined on these union sets, for a varying i. The optimal number of selected features and the associated number of regressors involved are indicated where relevant. In this figure, references to a *weighted* L2 distance are made. Formally it can be stated as:

$$L2_{\text{weighted}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{I} a_i (x_{1i} - x_{2i})^2}, \text{ with } a_i \geq 0. \quad (5)$$

The classical L2 distance is the special case where all a_i equal 1. Alternatively, we define the weighted L2 distance that uses the regression coefficients (after taking their absolute value, and rescaling them in $[0, 1]$).

First we can see that all the approaches proposed in this section largely outperform the naive solution. In particular, we see that the LASSO sets are very parsimonious solutions, with only 706 selected features, for a performance index of 9203. Ridge solutions lead to better performance indexes, but need approximately 10 times more features (i.e. still much less than the static features subset though). Weighting the L2 distance according to Equation (5) with the regression coefficients leads to better performance, but at the expense of largest selected feature sets. Thus no solution among these candidates is strictly better than the others, and a trade-off has to be chosen.

5. Visual Inspection and Interpretation of the Selected Feature Sets

For the sake of clarity, we focus on two feature sets presented previously:

- The one resulting from the LASSO analysis, using the weighted L2 distance (1056 features)
- The one resulting from the ridge analysis, also using the weighted L2 distance (11299 features)[3]

Further insight could be gained by visualizing these selections in the context of the full feature set. Most classical data visualizations, such as scatter-plots or parallel coordinates, cannot handle more than a dozen of dimensions, and are clearly not adapted to the SA models. Some techniques,

[3]The LASSO sets are imbricated, so the chosen set includes the solution obtained when using the classical L2 distance. Same observations hold for ridge solutions respectively.

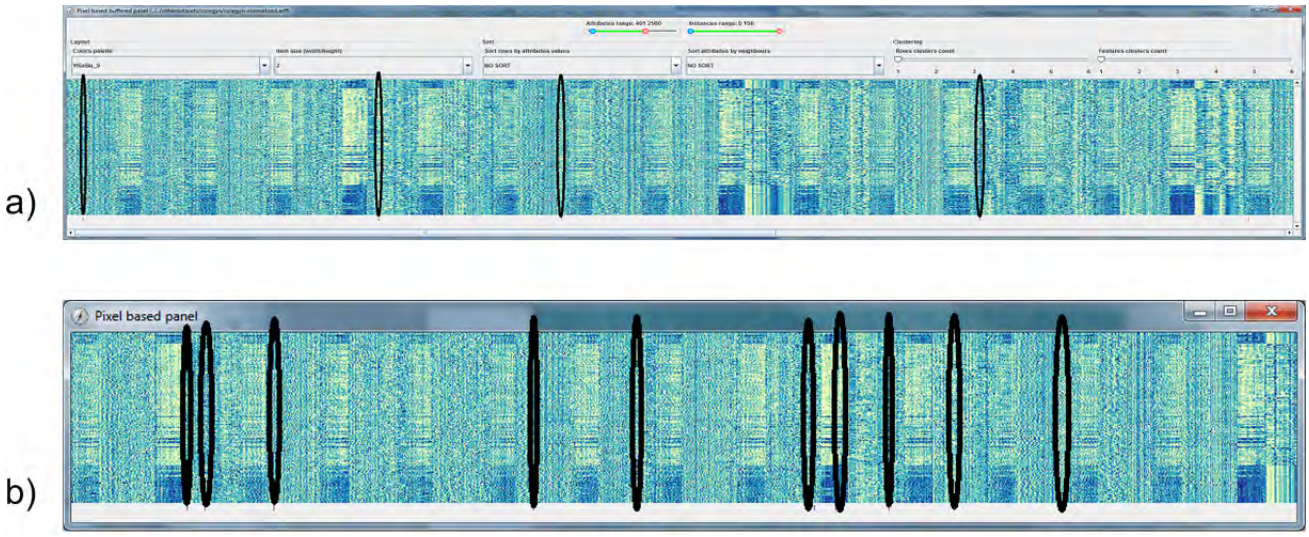


Figure 6: Highlight of the features selected by the LASSO (*a)*) and ridge (*b)*) methods, in the context of the full feature set.

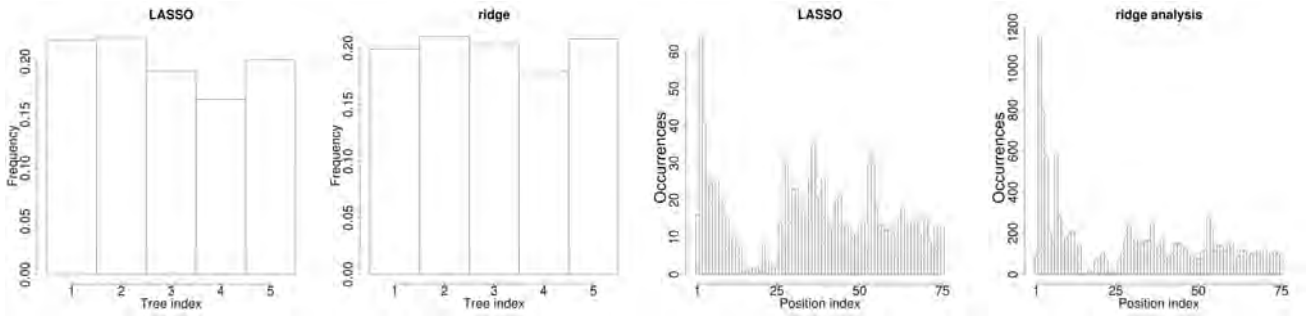


Figure 7: *Left:* distribution of selected features across the trees. Frequencies are normalized according to the tree size. *Right:* distribution of selected features according to the position in the 75-dimensional output state vector.

among which are pixel-based visualizations (Keim, 2000), were derived specifically to scale up to some orders of magnitude.

Pixel-based maps are graphical representations of data where the individual values contained in a matrix are represented as colored rectangles. Specifically, following state-of-the-art recommendations (Brewer, 1999), the numerical values of each feature are mapped to a heat color scale (low values map to light shades, respectively high values to dark shades).

As the whole dataset can not be shown in one single view, some techniques of *Visual Analytics* have to be applied in order to allow an efficient visual exploration of data (Keim et al., 2008). For instance, several classical functionalities have been added into this component, i.e. zooming, filtering, sorting rows according to a feature, sorting features according to the L2 distance with a reference picked by the user (Shneiderman, 1996).

This view has been used on a normalized version of the dataset (to obtain values between 0 and 1); then, filtering capabilities have been progressively applied to visualize parts of the dataset (Figure 5).

Some tendencies stand out when we look at the visualization. First of all, the leading 25 feature blocks in each 75-dimensional state vector are strikingly similar across states. They show a highly bi-modal pattern, very likely to be linked to the two clear clusters already seen in Figure 3. As the remaining features (i.e. delta and delta-delta) look much more evenly and independently distributed, the view suggests to focus the investigation about these clusters on the static features. Further inspection is possible by highlighting the features selected by the LASSO and ridge methods in the full feature set, such as shown in Figure 6.

As a complement, in Table 2 we show to which extent pairs of these subsets overlap, The static subset (i.e. the output state static features) serves as a reference. Only the lower triangle of the matrix is filled, as we are mostly interested in the proportion of the smallest set that lies in the largest (i.e. the reciprocal is not very insightful).

Interestingly, in Table 2, we see that LASSO features are picked at 28% in the static subset close to its 33% nominal proportion (i.e. the proportion of static features in the complete set). Ridge leads to selecting significantly more static features (44%): the amount of features taken in the static subset thus decreases according to the parsimony of the regression model. Also, LASSO and ridge are clearly emerging from a similar design choice (68% of LASSO features lie in the ridge selection).

	static	ridge	LASSO
static			
ridge	44%		
LASSO	28%	68%	

Table 2: Overlap percentages among selected feature sets. Should be read as: x% of the feature set *row* exists in the feature set *column*.

Figure 7 shows complementary metrics on the distribution of chosen features with respect to the SSS trees, and also with respect to the feature position in the 75-dimensional output state vector. We see that selected features are distributed almost evenly across the five trees, irrespectively of the method. The second tree (respectively the fourth tree) seems a little over-represented (respectively under-represented) though.
The distributions in the 75-dimensional state vectors are much more interesting. The following observations can be drawn:

- The selection in the static part is very skewed, for both methods: the first 15 features are frequently chosen (with peaks on the second and third), while the 10 last are almost never selected. This skewness is also somehow observed for the delta and delta-delta parts, but to a much lower extent.
- As already suggested in Table 2, both methods noticeably select delta and delta-delta features: this shows that the naive assumption should be overcome.

6. Conclusion

In this paper, we investigated how SSS model features could be correlated with the respective RMSE metric between models. We proposed a statistical formulation of the problem, described a tentative algorithm, and used visual tools for inspecting and interpreting the results. This is a first step though, as further work is needed to use these findings in the context of weak speaker adaptation where very limited amount of adaptation data is available.
To our knowledge, not much work exists in visual feature selection, especially for non-standard settings such as SA models, with no well-defined target variable. Visual metric learning approaches, such as Dis-function (Brown et al., 2012), have recently been proposed, but as discussed in Section 4.1., this is not the perspective needed in our context.
In this work, several statistical and visual tools have been used to understand the high-dimensional SA models, and derive a principled approach for selecting features that correlate with the RMSE-based performance index. However, as clear in Section 5., these tools are currently loosely coupled, and hardly qualify as a visual analytical method. Yet, we believe that their integration is a promising direction of research. For instance, semantic information about SSS tree states, overlaid on the pixel-based view of the selected features (see Section 5.) would help a posterior analysis, e.g. identifying redundant selections. A feature selection and evaluation tool could also report the NN performance index of interactively selected feature sets, possibly supported by statistical tools such as regression analysis or metric learning methods, or by re-ordering actions taken on the pixel-based view. The metrics reported in Section 5., such as distributions within trees and state vectors, could also be added as complementary views in this interactive analytical tool.

7. References

Aupetit, M. (2007). Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, pages 1304–1330.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4.

Brewer, C. A. (1999). Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pages 55–60.

Brown, E. T., Liu, J., Brodley, C. E., and Chang, R. (2012). Dis-function: Learning distance functions interactively. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 83–92.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2010). Two-stage learning kernel algorithms. In *International Conference on Machine Learning*.

Cristianini, N. Kandola, J., Elissedff, A., and Shawe-Taylor, J. (2002). On kernel target alignment. In *Advances in Neural Information Processing Systems*, volume 14.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, pages 1–22.

Ijima, Y., Miyazaki, N., and Mizuno, H. (2013). Statistical model training technique for speech synthesis based on speaker class. In *ISCA Speech Synthesis Workshop*, pages 141–145.

Imai, S., Kobayashi, T., Tokuda, K., Masuko, T., Koishida, K., Sako, S., and Zen, H. (2009). *Speech signal processing toolkit (SPTK), Version 3.3*.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G., (2008). *Information Visualization*, chapter Visual Analytics: Definition, Process, and Challenges, pages 154–175. Springer-Verlag.

Keim, D. (2000). Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*.

Nakano, Y., Tachibana, M., Yamagishi, J., and Kobayashi, T. (2006). Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing*, page 2286–2289.

Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, pages 1299–1319.

Shinoda, K. and Watanabe, T. (1997). Acoustic modeling based on the MDL criterion for speech recognition. In *Proceedings of Eurospeech*, volume 1, pages 99–102.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., and Bedo, J. (2007). Supervised feature selection via dependence estimation. In *International Conference on Machine Learning*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 267–288.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Proceedings of the International Conference on Spoken Language Processing*, volume 94, pages 18–22.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2002). Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*.

Yamagishi, J. and Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, 90(2):533–543.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):66–83.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y., et al. (2010). Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):984–1004.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *ISCA Speech Synthesis Workshop*, pages 294–299.

Enhanced News-reading: Interactive and Visual Integration of Social Media Information

Florian Stoffel, Dominik Jäckle, Daniel A. Keim
University of Konstanz
Department of Computer and Information Science
Box 78
78457 Konstanz, Germany
firstname.lastname@uni-konstanz.de

Abstract

Today, everyone has the possibility to acquire additional information sources as supplement to articles from newspapers or online news. The limitations of classical newspaper articles and restrictions of additional materials on online newsportals often lead to the situation where the reader demands additional news sources and more detailed information. When using the Internet, exploiting new information sources is a trivial task. Besides professionally administered information sources, like for example large newsportals such as cnn.com, there is a growing amount of user generated content available. Services like Twitter, Facebook or Reddit allow free discussion of any subject, giving everyone the possibility to participate. In this paper, we demonstrate an approach that combines professionally generated news content with user-generated data. This approach effectively enriches the information landscape and broadens the context of a given subject. For the presented system, we focus on Reddit, one of the biggest web portals for user-generated contents. Taking the general nature of user generated content into account, we exploit meta-data and apply *Natural Language Processing (NLP)* methods to allow users to filter additional information, which is also supported visually.

Keywords: Challenges in visualizing language data, Visualizations for language resources, Visualization infrastructures / visualization components

1. Introduction

Large parts of today's information gathering processes imply reading professionally written news articles. Depending on the medium and the kind of publication, the articles need to be limited in length and cannot naturally cover the whole context of the article's subject. Nevertheless, they often carry enough information to make readers curious about additional information; for example to be able to assess the background or specific aspects of the subject. Today, these additional information needs can easily be satisfied by visiting different other newsportals or news aggregators like Google News[1] or The Huffington Post[2].

Besides those additional information sources, there is an increasing number of websites available, which allow users to post comments on almost all subjects one can think of. Popular examples are Twitter with more than 200 million active users and over 400 million tweets per day (Wickre, 2013), or Facebook with 728 million daily users (Facebook, 2013). More like a general place for discussions and lesser a social network, Reddit[3] has more than 112 million registered users in January 2014 (Redddit, 2014). These huge numbers of participants lead to large data sets, covering a wide range of topics with different views and opinions. For those data sources, in terms of context and broader information for a specific subject, there is a high chance to find and gather additional information, deviating opinions, and even controversial discussions.

We propose to enrich professional news and articles with contents from such proposed websites. Existing approaches that establish a connection between those different content types (Hu et al., 2012; Gamon et al., 2008; Park et al., 2013; Meij et al., 2012), link news sites with blog posts, tweets, Wikipedia articles, and vice versa. We think, that Twitter is not fully suitable to provide context for given news, mainly because of the short length of posts of maximal 140 characters. Also, the users are using special terms and words to keep their posts short. Posts are in addition changing over time. This adds more complexity to machine-based feature extraction, and requires users, that are not familiar with this type of language, to learn it in order to fully understand a tweet.

To overcome named issues, we decide to use Reddit as one source for social media data. Besides the language issues, this choice also provides us a much richer data source in terms of text length, metadata, and the discourse structure. Apart from having a rating system for each user post, the discourses are already sorted in different so called *Subreddits*. Subreddits are usually dedicated to a specific topic, such as U.S. news, world news, sports, television food or general politics.

We present a user-driven method to retrieve broader context of actual news items. To do so, users can select and combine terms or phrases from news items to form the context of interest. This terms are then used as non-restricted query to retrieve matching Reddit posts whilst the user creates and modifies the context. Our contributions are: (1): A web based framework for linking user-generated social media contents with news, (2): A real-time social media analysis and visualization system, (3): The utilization of contents from Reddit as social media data.

[1] `http://news.google.com/`
[2] `http://huffingtonpost.com/`
[3] `http://www.reddit.com/`

2. Related Work

In the field of social media analysis already exists a huge amount of related research. Because of the vast amount and different types of data, the wide topic range, and a possibly large number of different users who generate data, social media data analysis is also a field with many research disciplines. It ranges from text analysis, development and application of graph algorithms up to text or metadata based visualization. Following, we focus on two fields, which are the most relevant to us. Namely, the classification of social media text, and the visualization of social media data. Besides this social media data oriented literature, we introduce relevant work in the field of linking different sources and data types.

Classification of Social Media Contents. An early work of Pang et al. (Pang et al., 2002) deals with the challenges of analyzing online contents. The authors classify sentiment for online contents using different machine learning techniques. These techniques are evaluated against the simpler method of topic-wise categorization. In (Kim and Hovy, 2004; Pak and Paroubek, 2010) the authors present techniques to automatically create a sentiment classifier based on human labeled test data and posts from the microblog plattform Twitter. A different goal is pursued by Yang et al. (Yang et al., 2007). In this approach, the authors perform emotion classification with the help of support vector machines and conditional random fields on blog data. The authors found out, that the last sentence describes a posts emotion best.

All these techniques concentrate on the classification of documents from online or social media sites using state-of-the-art machine learning techniques. A simpler approach is the usage of affective word lists as presented by Nielsen (2011). The author also elaborates on the differences between results which use his word list and an elaborated machine learning techniques from Thelwall et al. (2010). Surprisingly, the word list performs almost on par with machine learning techniques. An extensive report on state-of-the art technologies can be found in (Pang and Lee, 2008).

Since we want to build an interactive system that incorporates the classification of social media data, we decided to use the text classification presented by Nielsen (2011).

Visualization and Social Media Data. Hao et al. present a combination of sentiment analysis and social media contents in a streaming environment (Hao et al., 2011). The authors apply topic-based sentiment analysis to streaming data, which is used to generate two kinds of Twitter microblog visualization: a so called *Sentiment Calendar* and a *Sentiment Geo Map*. These two visualizations display data with temporal or geographical context.

A compact visualization of opinion analysis outcomes of user-contributed text data is presented by Oelke et al. (Oelke et al., 2009). To do so, the authors create a pipeline to extract application specific features, which are visualized compact to be able to provide reports out of the generated data. The combination of different techniques for sentiment analysis and the visualization of their outcome is the subject of Gamon et al. (Gamon et al., 2005). Starting with an application-dependent taxonomy, the authors describe the combination of classification, clustering, and a Tree Map visualization as part of an interactive system. Similar analysis goals are pursued in (Abbasi and Chen, 2007), where a system, capable of analysis of user-generated text data from different online sources, is shown. Having defined different features on three levels, namely style, topic and sentiment, the visualization is generated on text level and plotted into the document background.

Besides visualizing text features extracted from social media data, it is also possible to create visualizations for non-textual features. Smith et al. (Smith et al., 2009) present a system which visualizes the user base of a social media network with graph layouts and algorithms from graph research.

In excerpt from relevant literature, it becomes clear, that the actual visualization is application-dependent. If the task is to provide detailed insights into extracted features, a detailed visualization like the Ink Blots (Abbasi and Chen, 2007) is appropiate. Since we are aiming at overview or report style visualization, our approach is more like the work presented by Oelke et al. (Oelke et al., 2009). A high level overview of visual analysis of social media data can be found in (Schreck and Keim, 2013).

Linking News and Social Media. An investigation whether there exists a link between social media, in this case Twitter, and real world news is presented by Hu et al. (2012). The authors come to the conclusion that this is the case, and there can be even different types of users with respect to news spreading for the observed case. A tailored technique for merging political news with blog posts is proposed by Gamon et al. (2008). This technique can be described as supervised, as some domain knowledge is incorporated before linking takes place. Park et al. show a case study for political elections, where a link between real world news and Twitter messages is used in order to explore social media data regarding elections (Park et al., 2013). To link tweets and Wikipedia articles is the goal of the work of Meij et al. (2012). The authors extract concepts and different text and metadata features, and use a machine learning approach to establish a link between those different data sources. In (Tsagkias et al., 2011), the authors introduce a methodology to link news contents with social media contents in general, which also serves as the foundation for our work.

3. System and Process Pipeline

Today's news websites provide a huge amount of news articles and data, but they do not focus on the user's information needs. This can result in long lasting news browsing sessions to get additional insights or new information in general, which may also lead to sites outside the starting news source. In contrast, we present a user-centered system providing a link between social media data and a user-defined context generated from news articles. In this chapter, we motivate the approach and describe our process which is modeled along a pipeline.

Data Science and Visualization have already been brought to the web and are therefore widely spread. The creators

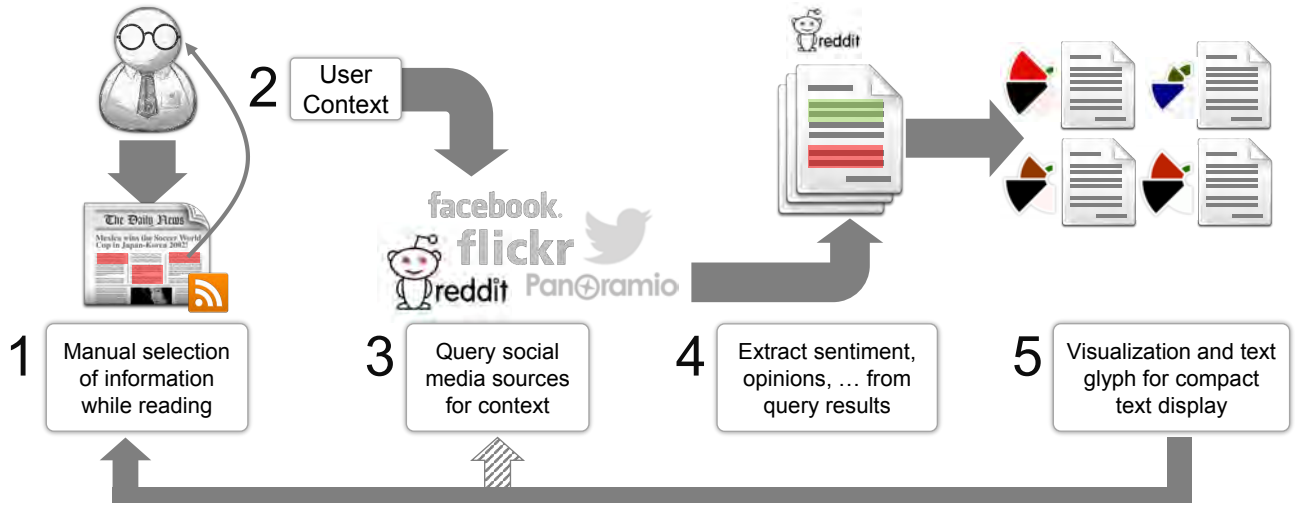


Figure 1: Five step pipeline. (1) The user reads different sets of news articles, possibly from different sources. To (2) create a context, the user either uses the integrated search functionality or manually selects words or phrases with the mouse. (3) Based on the selection, social media data providers are linked to right into the system. (4) Once these resources have been queried, we apply text analysis to extract scores from the retrieved data. (5) The retrieved data is visualized with a compact glyph, and some additional information is shown. After (5), the user can decide to go back to news reading, or to create links to other social media data providers.

and innovators in that field enjoy great reputation for their work, but today's web-based data visualizations are either static or contain interaction techniques that do not suit the user-driven, visualization supported analysis which is called Visual Analytics. This is a challenge for our framework, both, for the visualization and the text processing side. The web-, and as a consequence also the browser-based implementation of the framework has been chosen, because it opens possibilities of portability and accessibility not possible with locally running, native programs. There is also the advantage of being able to deploy the analysis infrastructure at a central location outside of the actual web interface, which makes it easy to get the system running on a smart phone or tablet device.

3.1. Process Pipeline

The framework follows the process pipeline shown in Figure 1, which defines the user interaction and the linking of a user-generated context with social media data sources. In the beginning, the user selects relevant news sources which are merged and displayed together (1). In order to create a context describing further information needs of the user, parts of the displayed news can be selected with the mouse (2). Besides that, a free text input is also available. In the next part of the pipeline, the link between the user-generated context and social media data providers is established (3). To do so, a provider specific query is generated and executed. After the query results have been retrieved, the data is analyzed and different scores are computed for each of the returned pieces of data (4). In order to provide a responsive interface, these analysis steps have to meet strict processing time and resource requirements. The extracted scores are visualized as glyph besides a compact display of the retrieved data (5). At this stage of the pipeline, there are two possibilities to continue. The first is to examine the retrieved data, which refers back to (1). To enhance this task, various filters, orderings and an adjustable scoring are available. Another possibility is to introduce a link between different social media data providers to broaden the retrieved results. Currently, this possibility is part of our pipeline, but not implemented in the framework.

3.2. Integration of Multiple Data Sources

There are two stages in the pipeline (Figure 1), which rely on the combination of multiple data sources. The first one happens in (1), where the content of user-defined news sources is merged in one display (Figure 2, (2)). The second part, where multiple data sources are integrated, is indicated by the feedback from (5) to (3), the shaded arrow in Figure 1. This is, where the linking from one external data source, i.e. not a news source, to another external social media source should take place. The combination of data sources requires the ability to extract and retrieve the same information and metadata, like title, text, date, tags and more. This is one of the biggest challenges, since those information extractors are specific to the each of the data providers. Most of them use similar techniques for retrieving the data (JSON-RPC, SOAP, or REST), but the returned data format can differ heavily.

3.3. User-Driven Knowledge Discovery

Users play a crucial role in all kinds of Visual Analytics applications. Therefore, it is worth to explicitly mention the steps of the pipeline in Figure 1 involving the user again. The user has to define the news feeds which should be presented (1). Additionally, a context to use for linking social media data sources has to be defined manually (2). The visualization had to be interpreted (5), and there is the optional filtering and ordering, which is completely user-driven (5) $\rightarrow$ (1), (5) $\rightarrow$ (3). By involving the user at those crucial steps of the pipeline, it is ensured that the results are

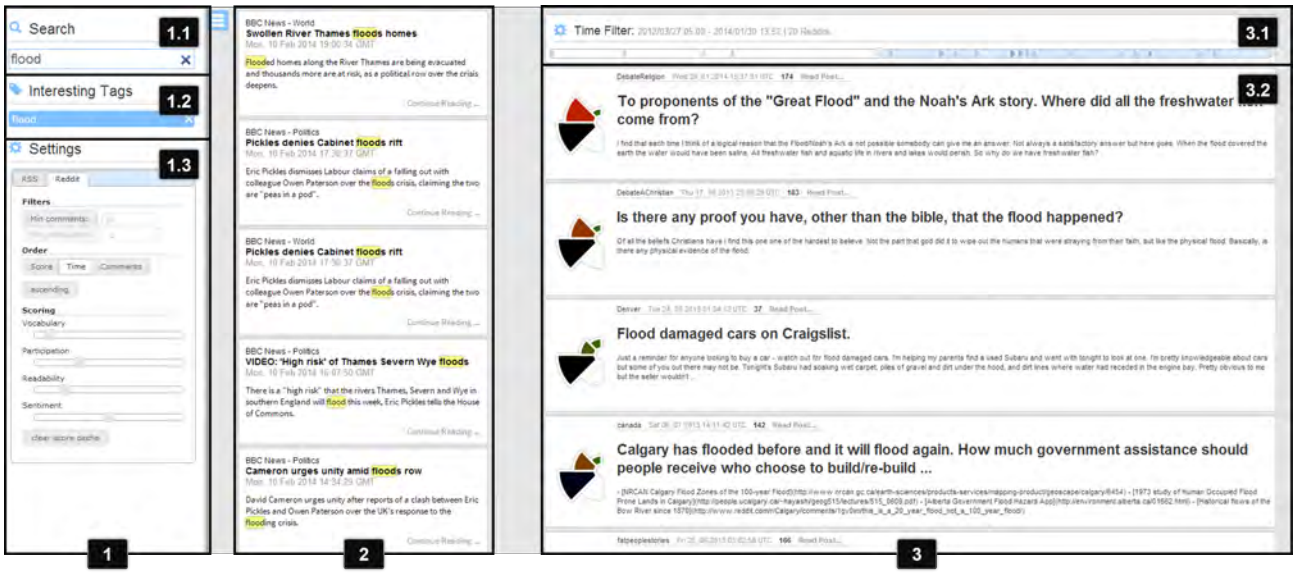


Figure 2: Our system consists of three main parts, placed next to each other from left to right. (1) System interactions: (1.1) Search for relevant terms within all displayed news articles, (1.2) manage tags used to create the context the user is interested in, and (1.3) allows further customization of the display in (3), for example the ordering or apply in-place filters. (2) The news data is presented in descending order over time. It changes automatically when the news sources are changed by the user. (3) Following the user's context, data from social media providers are presented. These can also be sorted by the user either (3.1) by time or (1.3) with the help of various ranking and filters.

tailored to the user's interests, and not by parameters from algorithms which are hard to understand.

3.4. System Design

Figure 2 gives an overview of the system implementing the pipeline. It is built out of three main elements: a sidebar offering various interaction and configuration possibilities (1), the news article display (2), and the augmented social media display (3). When the application is started, the user can adjust the active sources in (1.3), which are then merged in the combined news display (2). This list can be browsed through and used to build the context by selecting words or phrases with the mouse, which is displayed in (1.2). Additionally, there is a full text search feature available, which interactively filters the displayed news items (1.1). Every time the context is changed, the link to the social media data sources, currently only Reddit, is established and the results are retrieved asynchronously and displayed in (3), augmented with the glyph. The display can be adjusted by the user in terms of the order, and some in place filtering can be applied (1.3).

3.5. Text Analysis and Scores

To be able to judge whether a returned post is worth reading or of general interest, a text analysis is performed. The analysis has to meet strict resource and runtime limits. Otherwise, the system would slow down the process of knowledge discovery, which results in not fulfilling the intended task or a displeasing slow down in the interaction. Currently, these limits are met because the text analysis methods are not implemented by using any resource or computation intensive technologies. To compute text readability, the well established and efficiently computable Flesch Reading Ease measure is used (Kincaid et al., 1975). A measure for the complexity of the used vocabulary is introduced, to take into account not only the quantitative, but also the qualitative aspect of the text To do so, the share of words not contained in the 10,000 most frequently used English words, extracted from the Project Gutenberg corpus[4], is computed. To provide a sentiment score, the positive/negative wordlists from Hu et al. are used (Hu and Liu, 2004). The sentiment score is computed by subtracting the number of negative sentiment words from the number of positive sentiment words. To express the comprehensiveness of the discourse, which we assume is positively correlated with the number of participants in the discussion, the average number of posts per user is computed. Each of the described features is computed on sentence (readability, sentiment) or post level (non-basic vocabulary, number of posts per user) respectively. Those features have been chosen, because they give an insight in the quality of the text and discourse, which in our point of view are relevant when users decide if a text is worth reading or not.

3.6. Visualization of Analysis Results

The visualization for the text analysis output is shown in Figure 3. Each text feature is represented in a separate sector, which are equally sized when they are at their maximal extent. The feature value is double encoded in the size of each segment and the color value. For each of the segments, the feature value and color is modified separately to optimize its display. For example, the sentiment is mapped to colors ranging from red (negative) to white (neutral) to green (positive). A segment displaying a neutral sentiment score is covering roughly one fifth of the available, maximal segment area. If the score is negative or positive, the segment is

[4] http://www.gutenberg.org/

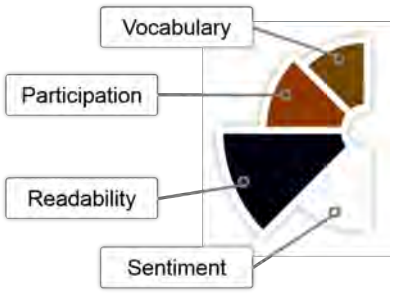


Figure 3: The glyph used to display the text analysis results, which are computed for each data item separately.

enlarged according to reach its maximal size when the text is classified as maximal negative or positive. Since different colors are assigned positive and negative sentiment, the actual sentiment is easy to determine. The readability score is mapped to a color scale from white (easy to read), blue (medium readability), black (hard to read). We chose non signal colors to not induce any ranking like a text is really hard to read, when the segment is colored red, since the actual readability is subjective. For the level of used words and the participation score, we use a color map ranging from green to red, because in those cases we want to support the user in his selection of text to read.

4. Application Example

This application example illustrates the process of broadening a given context and generating new possible exploration directions of a given subject.

At first, the user begins to browse a news stream retrieved from ESPN, a popular sports news site. In a news item, the term *Olympic* sticks out and stirs up the user's curiosity. To add this term to the current user context, the term can be selected with the mouse (Figure 4 (1)). Immediately after changing the current user context, the social media sources, in this case Reddit, is linked and posts regarding the context; in our case, the term *Olympic* is retrieved. Since only recent posts are of note, the time filter is applied to the results by only selecting results that begin with the middle of October 2013 (Figure 4 (4)). Right below, the generated overview from the social media data, enriched with the glyph, is presented. Additionally, the order of the displayed social media data is adjusted by the user-defined scoring, where the high user interaction is rated highest, followed by the sentiment, readability and vocabulary score (Figure 4, (3)). In the first post, the name *Sochi* sticks out, which attracts the users further interest.

To continue with the new context addition *Sochi*, the user applies the integrated full text search, Figure 5 (1), to filter the news items accordingly. Also, the term is added to the user context, Figure 5 (2). This causes an immediate update of the displayed social media posts, as presented in Figure 5 on the right hand side.

In the news, which are now filtered by the term *Sochi*, a shift of providers can be seen. The news displayed in Figure 4, originated from ESPN, as the user has decided to read sports news. In Figure 5, the displayed news item originate from CNN only. As it can be seen on the example Figure 5 (3), there new perspectives arise from the new user context, shifting from sports to politics.

5. Discussion

Currently, the system is designed to support one task: the extension or generation of social media context given a set of terms, manually selected by the user from news items. Not taking the whole news item, but only a set of user selected terms for context generation is the guarantee, that the system only retrieves contents which the user is interested in. Having that terms and the link to a social media data provider, one can think of additional tasks besides the context extension.

A number of other tasks can be supported by the system and the underlying technologies. When trying to describe them, four categories have to be considered separately: At first, users must be able to formulate the context they are interested in. Currently, there is the possibility to select terms that describe the context by clicking and dragging on the news item text. To allow terms which are not contained in the news item, but relevant for the context according to the user, a manual term input is available. This works reasonably well for describing the desired context with text, but for other data sources that do not provide text data other input methods must be available. The second category describes the technology used for linking a social media data source, since it must be possible to establish, query and retrieve the data type provided by the source. In our case, we use the search api from Reddit, which returns different types of data for one query in the same JSON based data format. The third category covers available interactions. It is obvious, that for exploration tasks, the system should be able to establish not only a link between user-defined context terms, but also between different social media sources, as it has been already formulated in the pipeline on Figure 1. This should be covered by the available interaction techniques and also the type of user interface. In the current state, we display the results from the social media provider as a list with a glyph describing each result visually. This list can not be used for further exploration of the context, because it doesn't allow direct selection of terms from the results. For an exploration task, this should be possible. The fourth category covers the user interface. To stick with the exploration task example, having the possibility of establish a link between different social media data sources, it should be possible to use that link in both directions, if a dead end of the exploration has been reached. A simple solution for linking in both directions is to extend the user interface by providing both ends of the link. In the current system, this would mean that two lists with postings are shown next to each other, and both provide the same interaction techniques.

6. Conclusion and Future Work

In this paper, we present a framework to link news and social media content. Our application example shows this combination for newsfeeds from ESPN and CNN, combined with the social media platform Reddit. We allow users to define the context they want to explore by themselves, thus not imposing any direction or things they are not interested

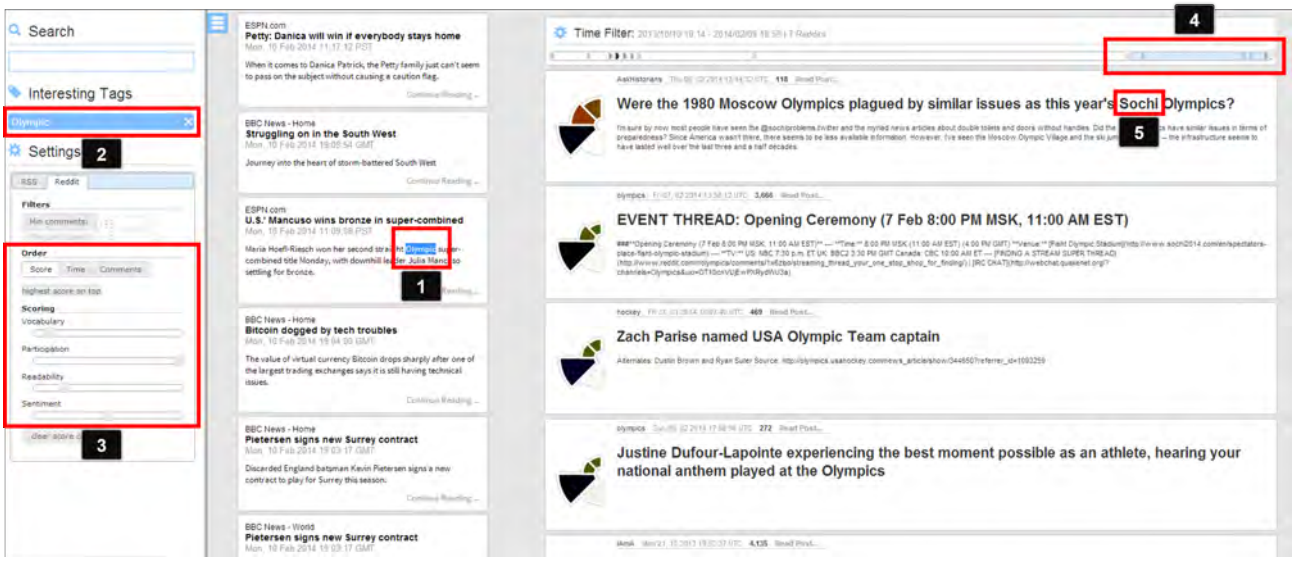


Figure 4: First step in the process towards enhanced news reading. When starting to use our application, the user is presented a temporal ordered news stream that consists of multiple news sources. ESPN sports-news-stream publishes articles about the Olympic games that take place in 2014. The user is interested in additional information and therefore (1) manually highlights the term *Olympic* within the article. The selected term is automatically transfered and added to the (2) list of interesting tags. Meanwhile, social media content matching the given keyword is retrieved. To rank the social media results according to the user's preferences, the user changes the (3) scoring order and applies the (4) time filter. The top ranked result reveals the (5) venue of the Olympics: *Sochi*.

in, but can be rated high when automatic methods decide which content to link.

We chose Reddit, because it provides much richer content than other social media data sources like Twitter. Also, accessing data is not restricted in any ways, at is can be on platforms like Facebook. It also has the advantage to be actively moderated, so the resulting content should be free of any off-topic contents, troll postings, or similar internet phenomena.

Compared to existing work, we bring the concept of linking news and social media directly to users. In our framework, we apply it to news reading in form of providing a broader context and information, which is usually not included in articles from news sites. Especially, when the integration of further data sources, different types and the linking among them has been added, the additional information provided by our framework must not only be presented as text, as it is the case now. For example it is then possible to represent additional context information as image tiles, which are associated with the corresponding social media contents. This would reduce the cognitive load required for users to actually generate new knowledge about the subject they are interested in.

6.1. Future Work

In the future, we want to expand the capabilities of the system to generate a broader social media context. To do so, there is the possibility of using the user-provided context as seed for a classical query expansion. Sources like WordNet or DBPedia can be used to expand the context before actual queries are created and sent to the social media data providers.

We are also interested in the possibility to add other data types than text. Currently, we have a compact, row like representation of social media text data, which is displayed with a title on the right, see Figure 2, 3.2. There are other social media data sources available, like Flickr, Panoramio, or Instagram, where annotated image data is available. Also, with the images, one could think of a dashboard like representation of the social media data, represented by associated images. In terms of the actuality of the retrieved data, it is without doubt that Twitter would also be an optional, valuable addition to the data sources.

The basic text processing, which has been chosen because of the desired real-time capabilities of the linking, could be extended by introducing a two staged analysis model. The first stage represents our current processing, capable of computing results almost instantly after the data had been retrieved. More elaborated methods, like the combination of machine learning techniques for identifying positive and negative arguments, or to analyze the type and structure of the discourse, act as second stage. This additional information can be used as another dimension in the glyph visualization, the data filtering capabilities, or to determine the order of the retrieved data in a fine grained way which is currently not possible.

It is also an open question, whether we can apply the same measurements and methods to different social media data sources. The first problem which can arise, is the type of language and the length of posts given two sources like Twitter and Reddit. An automatic evaluation of different methods and their result should make it clear, whether we need to employ different analysis techniques for different data sources.

The glyph that represents the classification score of each social media entry, helps users to rapidly identify valuable

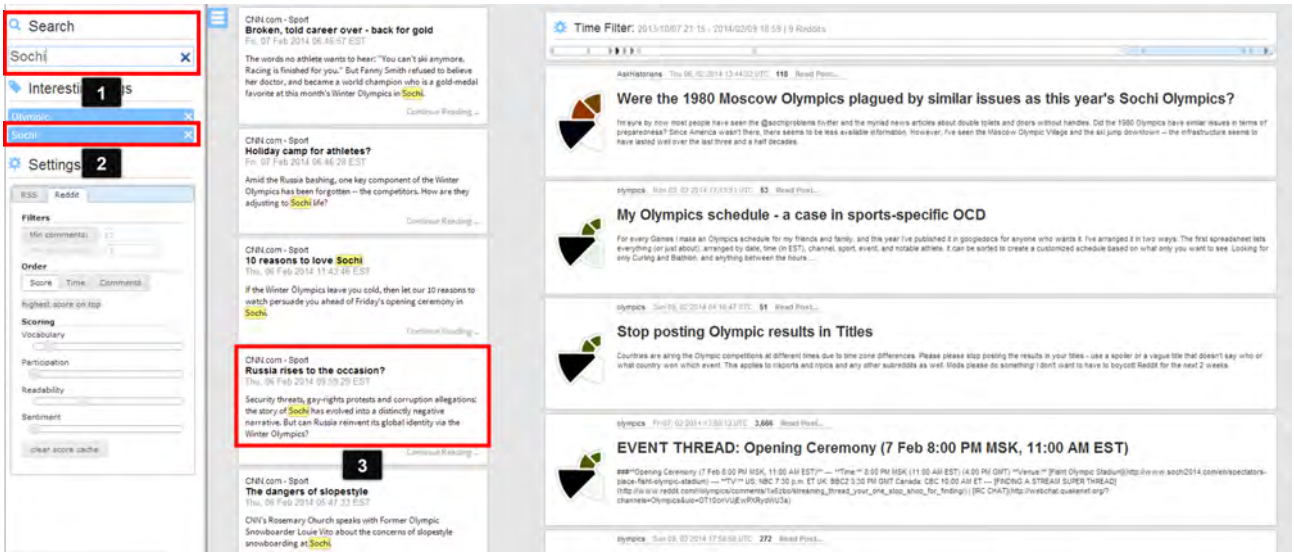


Figure 5: In the second step towards enhanced news reading, the user (1) searches the news for the keyword *Sochi* (determined in Figure 4). In accordance with the query, the news stream is filtered and presents articles that contain given keyword only. Additionally, the shortcut *shift+return* adds the keyword to (2) the list of interesting tags and refines all social media results. The filtered news stream reveals an (3) interesting article published by CNN, named: Russia rises to the occasion? An article about security threats, gay-rights, etc. which gives dissimilar impressions. That finding serves as entry point for further enhanced exploration.

entries, but it lacks of overview. It is obviously hard to visually group entries, especially with more than four categories. Hence, future work includes enhancements of the proposed glyph.

A formal task definition, the required components, analysis techniques and interaction possibilities can be a step towards making our technique more applicable in general. Currently, the supported task is to retrieve context information for a given news item. But given the underlying technology, it is worth exploring different possibilities when the link between a news item and social media data providers has been established. One possible task that builds on the current basis, is the generalization of the system for a broader investigation of a given subject. For this task, extensions of the supported data sources and the interactions, for example to effectively support browsing through retrieved data, need to be added. Nat at last, we want to conduct a controlled user study to see whether the tasks supported by our system are useful for real users.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) project “Feature-based Visualization and Analysis of Natural Language Documents” (VisADoc).

7. References

Finn Å. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In M. Rowe, M. Stankovic, A. Dadzie, and M. Hardey, editors, *#MSM*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98. CEUR-WS.org.

A. Abbasi and H. Chen. 2007. Categorization and Analysis of Text in Computer Mediated Communication Archives Using Visualization. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’07, pages 11–18, New York, NY, USA. ACM.

Facebook. 2013. Facebook Reports Fourth Quarter and Full Year 2013 Results. http://investor.fb.com/releasedetail.cfm?ReleaseID=821954. Online Report, accessed: 2014-02-08.

M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining Customer Opinions from Free Text. In *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA’05, pages 121–132, Berlin, Heidelberg. Springer-Verlag.

M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König. 2008. BLEWS: Using Blogs to Provide Context for News Articles. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, editors, *ICWSM*. The AAAI Press.

M. C. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L. Haug, and Mei-Chun Hsu. 2011. Visual sentiment analysis on twitter data streams. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 277–278, Oct.

M. Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’04, pages 168–177, New York, NY, USA. ACM.

M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. 2012. Breaking News on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’12, pages 2751–2754, New York, NY, USA. ACM.

S. Kim and E. Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING ’04, Stroudsburg, PA, USA. Association for Computational

Linguistics.

J. P. Kincaid, R. P Fishburne Jr, R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, DTIC Document.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding Semantics to Microblog Posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 563–572, New York, NY, USA. ACM.

D. Oelke, M. C. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L. Haug, and H. Janetzko. 2009. Visual opinion analysis of customer feedback data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 187–194, Oct.

A. Pak and P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

B. Pang and L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Souneil Park, Minsam Ko, Jaeung Lee, Aram Choi, and Junehwa Song. 2013. Challenges and Opportunities of Local Journalism: A Case Study of the 2012 Korean General Election. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 286–295, New York, NY, USA. ACM.

Redddit. 2014. About Reddit. http://www.reddit.com/about/. Online Web Page, accessed: 2014-02-08.

T. Schreck and D. A. Keim. 2013. Visual Analysis of Social Media Data. *Computer*, 46(5):68–75, May.

M. A.. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. 2009. Analyzing (Social Media) Networks with NodeXL. In *Proceedings of the Fourth International Conference on Communities and Technologies*, pages 255–264, New York, NY, USA. ACM.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment in Short Strength Detection Informal Text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December.

M. Tsagkias, M. de Rijke, and W. Weerkamp. 2011. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 565–574, New York, NY, USA. ACM.

K. Wickre. 2013. Celebrating #Twitter7. https://blog.twitter.com/2013/celebrating-twitter7. Online Blog Post, accessed: 2014-02-08.

C. Yang, K. Lin, and H. Chen. 2007. Emotion Classification Using Web Blog Corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 275–278, Washington, DC, USA. IEEE Computer Society.

A Visual Focus+Context Approach for Text Comparison Tasks

Markus John[1], **Florian Heimerl**[1], **Andreas Müller**[2], **Steffen Koch**[1]

[1]Institute for Visualization and Interactive Systems
University of Stuttgart
<firstname.lastname>@vis.uni-stuttgart.de

[2]Institute for Natural Language Processing
University of Stuttgart
Andreas.Mueller@ims.uni-stuttgart.de

Abstract

The concept of distant reading has become an important subject of digital humanities research. It describes a mode of textual work in which scholars are aided by automatic text analysis and visualization to directly find and access information relevant to their research questions in a large volume of text. While such techniques have proven to be effective in saving time and effort compared to extracting the information by linearly reading through the text, they introduce a new abstract level of analysis that masks the original source text. In this work, we present a flexible focus+context approach that facilitates scholarly textual work while at the same time supports efficient distant reading techniques. Users have full access to digital text sources for the perusal of a single text passage or the comparison of multiple ones. For each selected passage, an interactive visual summarization of its respective context allows users to effortlessly switch back and forth from close to distant reading. We demonstrate the capabilities of our approach with a usage scenario from the comparative study of poetics. The applicability and usefulness based on expert feedback is discussed afterwards.

Keywords: digital humanities, visual text analysis, text comparison, distant reading

1. Introduction

Over the last decade the digital humanities have emerged as a thriving new interdisciplinary research area that integrates scientific research methodologies and techniques with those from the humanities. Growing interest in this field from researchers in the humanities as well as the sciences has sparked the development of exciting new methods and concepts to facilitate humanistic research. One concept that has attracted a lot of attention is distant reading (Moretti, 2013), a mode of textual work that enables scholars to discern, interpret, and discuss information from texts without reading them. Distant reading techniques extract information from literary works using text processing, and present a summarization typically in a visual rather than a textual form.

We are working with a digitized collection of historic documents, containing 20 selected works from an overall collection of about 1000 German poetics from the years 1770 to 1960. This comprehensive collection has been initially explored and compiled by Richter (2010). Poetics are secondary sources that contain the results of scholarly inquiry about literature and art and as such they discuss, interpret, and classify authors and their works. One line of inquiry based on this corpus is to gain insight into different views and perceptions of scholars at different time periods. Working on such questions requires close scrutiny of text passages that discuss e.g. the use of a specific term by a poet, and the comparison of passages about similar aspects of this poets work from other sources.

The approach we present here was designed to offer assistance for this particular way of comparative textual analysis and provides many functionalities to expedite such work. It supports close perusal of single digital text passages and comparison of multiple ones from the same or from different sources through a visual interface. This interface presents multiple texts side by side using a focus+context technique that lets users select interesting parts of the texts to compare them in detail, while still providing an overview of the whole documents. To support users in comparing documents, or finding suitable passages for comparison, we offer a search function that highlights key terms and employ a straightforward method to identify common text passages across the documents. The respective passages are marked in the focus area and in the context visualization.

In this paper, we describe our approach in detail and present an initial implementation. Based on this implementation, we provide further insights by discussing a usage scenario that demonstrates the usefulness of our approach. Moreover, we report on an interview with an expert about the approach and its implementation that will serve as a basis for a discussion of its drawbacks and advantages.

2. Related Work

In recent years, visualization techniques were introduced to support literary scholars and experts from other domains, who are working with text material. A broad variety of visual methods such as word clouds, topological landscapes (Wise et al., 1995), pixel-based overview (Keim and Oelke, 2007), or clustering and classification techniques (Heimerl et al., 2012) were developed to support text workers in exploring and analyzing text corpora on abstract levels. Even if there is a recent trend in the digital humanities to develop visual methods facilitating distant reading (Moretti, 2005) by providing visual abstraction, it is obvious that humanities scholars still need to work with text resources directly. Many visual approaches address this problem by offering different perspectives for overview and

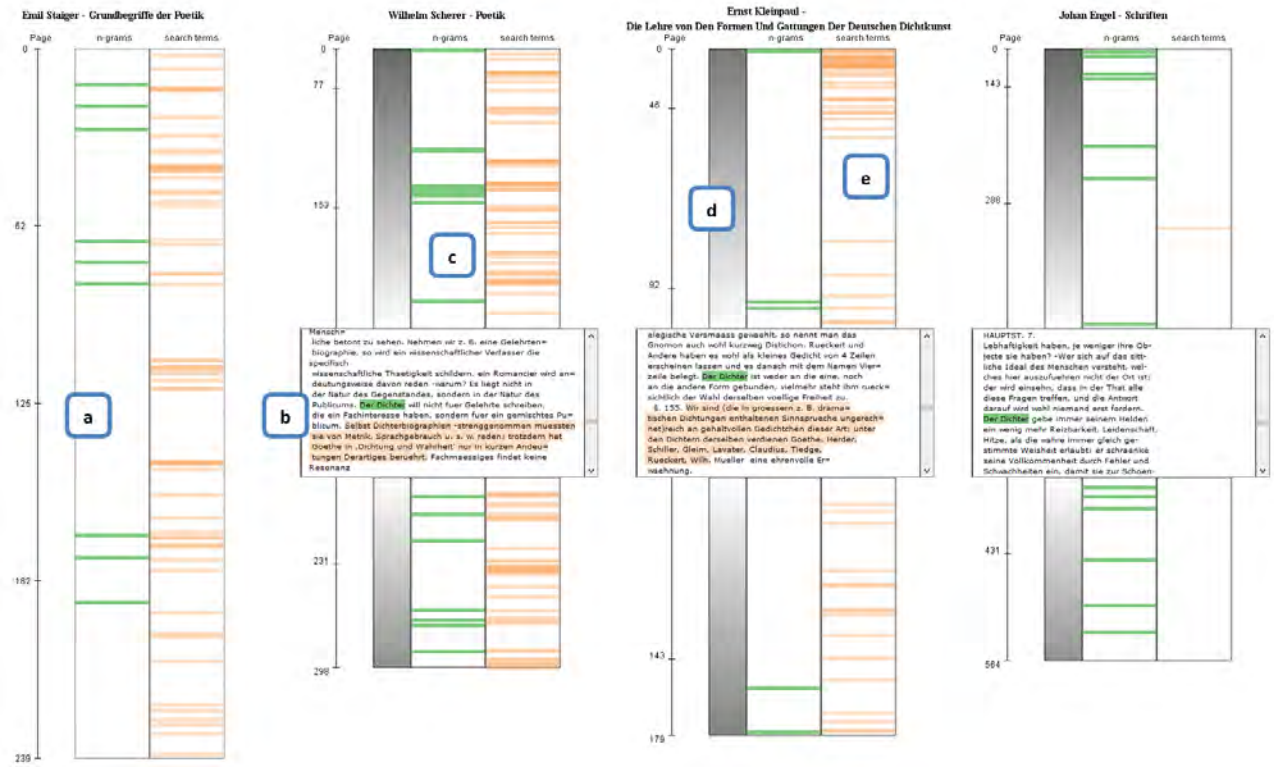


Figure 1: The image shows a screenshot of our implementation. Four selected text sources are displayed next to each other. Three of them are focused on a text passage, the left one is in overview mode. Each of the sources has a scale indicating the page number of the original documents to the left, a ribbon showing the position of the selected n-gram as green bars, and a ribbon showing the position of the search results as orange bars. The documents with an open text passage are equipped with a third ribbon indicating text density by color.

text details. Most often this is realized with a separate text view or window that helps to inspect or verify a "distant" finding in its source context, for example in (Don et al., 2007). Our approach focuses on highlighting text passages similar to Eick et al. (1992) that were identified in previous steps through search or automatic extraction mechanisms. At the same time, specific findings can be inspected in a focus+context manner within a text perspective that is directly integrated into the overview perspective. The idea of providing both, the distribution of findings and the possibility to inspect a specific text passage in detail has been pursued in other works as well, e.g. by indicating an overview of findings next to a text view's scroll bar (Correll et al., 2011). Additionally, focus+context interaction techniques were developed in order to improve reading or browsing text documents (Robertson and Mackinlay, 1993). The presented technique differs in that it shows the context of a document as a ribbon under the focused section and by presenting documents side by side to help users in comparing the overview and the details.

3. Approach

Our approach is designed to support scholars with their textual work by offering methods to access and compare digitized textual resources on different levels of abstraction. Figure 1 depicts the main work space of our implementation into which we have loaded four example documents from our poetics corpus. Additionally, a menu bar is available that provides a drop down list with all identified n-grams, a search box to enter terms including a menu to select the search mode, and a menu to choose a distortion mechanism for the focus+context visualization.

When the application is started, the text sources are fed into the linguistic processing pipeline that applies a tokenizer and a sentence splitter. We use the implementations of the Apache OpenNLP library (Apache Foundation, 2014) for this. After tokenization and sentence splitting, the pipeline identifies all token n-grams up to a length of four tokens that occur in more than one of the source texts. Sequences that merely contain stopwords and punctuation marks and no further useful information are filtered out, while the rest is retained as candidates for closer inspection by the user. We found n-gram identification to be a good starting point to find interesting text passages to propose to users as a starting point for closer analysis.

After this preprocessing, users can select one of the identified n-grams at any given time and explore its distribution across the documents with an interactive visualization. Additionally, each occurrence can be inspected in detail with a focus+context method. As depicted in Figure 1, the work space is organized into up to four columns, one for each of the text resources. Initially, each text source is displayed as a ribbon representing the entire document. Each occurrence of the selected n-gram is depicted as a green bar along this ribbon (Figure 1a) creating a visual abstraction for distant reading. Users can select text passages for analysis by

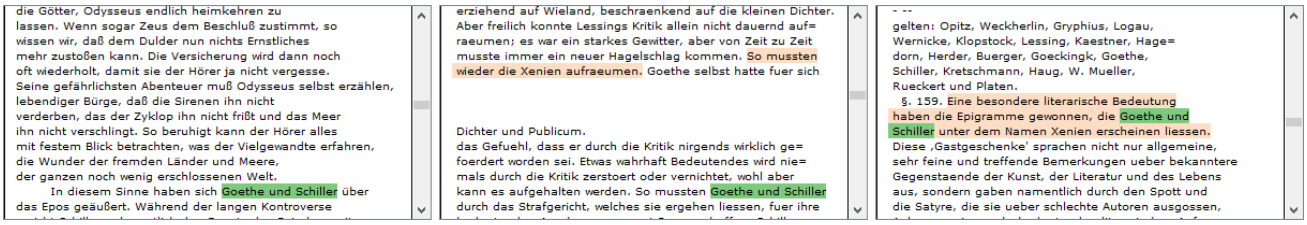


Figure 2: Passages from the three documents that contain “Goethe und Schiller”. The two documents on the right contain additional references to the “Xenien” which is highlighted as a search result.

clicking on them, which opens up a text box for the document at the corresponding position (Figure 1b). The selected passages are placed next to each other for easy comparison. As depicted in Figure 1b, the text boxes are positioned on the ribbon which summarizes the context before and after the selected passage. Users interact with it to inspect occurrences of the selected n-gram within the same document. Once one or multiple passages of text have been selected for closer analysis, users can read and compare the text within the boxes. The n-gram occurrences are marked in the text with a green background (Figure 1b). Users can influence the presentation of context by switching between a linear and a quadratic mapping. The latter emphasizes context closer to the currently selected passage to give users a higher resolution of detail around selected areas compared to areas further away. To illustrate the distortion at each position of the ribbons, we introduced the gradient ribbon (Figure 1d) that maps text density to luminance. Thumbnails of the scans of the source documents are displayed below each column, depicting the original page that contains the current passage. Those can be enlarged in a separate window for further analysis, giving users instant access to all nontextual information within the source documents.

As depicted in Figure 1 the user can scroll through the text document using the scrollbar besides the text box. This approach can be likened to moving an elastic text ribbon under a lens. N-grams that are scrolled out of the text box are displayed within the ribbon above or below the box, depending on the scrolling direction. The ribbon will stretch across all of the available vertical screen space on either side of the text box when the volume of text on the respective side is 50% or more of the total volume of text in the document. If more than 50% of the document accumulate at one side of the text box, the length of the ribbon on this side will remain constant, and its contents compressed accordingly. With less text, the respective side of the ribbon will be reduced in length proportional to the amount of text included. This helps to give users an impression of the relative position of the text box in the document.

For further analysis, users can execute searches for arbitrary terms, and choose between a conjunctive and a disjunctive search mode. The search engine identifies all sentences that match the search within all of the loaded text sources. Their position within the texts are depicted within the context visualization as bars in an additional ribbon (Figure 1e), allowing users to assess the density of matching sentences and their proximity to the selected n-gram across the texts. In conjunctive mode, only sentences that contain all of the terms are marked, while in disjunctive mode occurrences of each single term are marked with a distinctive color. As for n-grams, users can select search results to place the text box at the respective position. To handle overlapping bars, we use semi-transparency and alpha-blending to visually encode dense areas with higher opacity (Figure 1e).

4. Expert Feedback

To gain first insights into the usefulness and applicability of our approach, we asked a German literature PhD candidate to conduct an analysis session with the system. We aimed to get feedback about the advantages and drawbacks of the approach and of specific quirks of the implementation.

4.1. Process and Scenario

The expert selected four texts that we initially loaded into the system. We then explained its UI and asked the expert to use the system for a comparative text analysis session. During the session, we recorded notes of the expert's freely verbalized thoughts and comments including his analysis steps and usage patterns of the system according to the think-aloud methodology.

The expert starts the analysis session by skimming over the proposed n-grams to find a combination of terms that he thinks is interesting for further inquiry. The n-gram “Goethe und Schiller”, the names of two major German poets, is the first one to attract his attention. He selects it to overview their occurrences in the four selected documents in a distant reading manner, and learns that it occurs multiple times in three of them. The fourth document, although published later, was written before Goethe and Schiller's time. The expert recalls that Goethe and Schiller have jointly published a collections of poems called “Xenien”, and hypothesizes that this collection is discussed in conjunction with the occurrence of both poet's names. To test this hypothesis, he uses the search function to search for “Xenien”, resulting in the display of the term's occurrences in the respective ribbon. By exploring the occurrences, he finds that one of the documents distinguishes between “dramatic” and “epic poetry” and mentions Goethe and Schiller's essay on this distinction, but it does not mention the “Xenien”. The second document doesn't make this distinction and only mentions the “Xenien” in conjunction with Goethe and Schiller. The third one, however, does make the distinction and mentions the “Xenien”, but not Goethe and Schiller's essay. Figure 2 depicts passages from those three documents in the respective order from left to right. The fact that the third one doesn't mention the essay

but only the "Xenien" arouses the expert's interest, and he states that this finding is a starting point for further inquiry.

4.2. Feedback and Discussion

The feedback of the expert was mainly positive. He found the system easy to understand and intuitive to use. The main benefit he saw was the visual presentation of the context of the text passages, and the ease of navigation within the different texts. Both features speed up comparative text analysis tasks.

There were also shortcomings of the approach and the visual representations that the expert addressed in his feedback. His main point of criticism was that all documents are mapped to the same length within the work space, although their text lengths differ significantly. We visualize the documents this way to use the available screen space efficiently and thus give as much detail as possible about the occurrences of n-grams and search results. To alleviate this problem, we added page marks to the document ribbons as depicted in Figure 1. The expert additionally mentioned that the presentation of the n-grams as an alphabetically sorted list could be improved on. We agree that the present version lacks an overview visualization of the comparison candidates which are currently n-grams, but consider this problem out of the scope of this work.

The expert noted that the automatic n-gram matching across documents is helpful for identifying relations between documents with respect to the discussion of similar concepts. Especially during exploratory analysis it brought up new concepts mentioned in several sources that offered new possibilities of relating the texts to each other. However, he also stated the obvious drawbacks of this simple and quite coarse method to identify potentially interesting text passages. The search function was also mentioned, and praised, as it allows a flexible overview of occurrences of arbitrary terms in the documents. It can serve as a tool to compare contextual patterns of n-grams across documents.

The mostly positive feedback indicates that our approach is indeed effective and useful for comparative scholarly textual work. Additionally, the expert expressed his interest in the approach, and mentioned that he would like to use it to work on his research questions. This illustrates that methods that integrate distant and close reading within one interactive environment are useful for humanities research, and that digital humanities researchers should refine and advance such methods.

5. Conclusion and Future Work

In this work we presented a novel focus+context technique for comparative textual work that supports seamless switching between distant and close reading of scholarly sources. It provides access to digitized literary text sources by providing a view on the raw text that is embedded in an interactive visualization of its context within the text documents it is part of. Our approach includes natural language processing to identify potentially interesting text passages for perusal and comparison, and a flexible search mechanism offering powerful analytic capabilities. We presented an implementation of the approach, and provided a usage scenario and a discussion of the feedback of a literary analysis expert on our implementation. From the expert feedback we are able to discern that our approach is indeed effective and useful for comparative literary analysis.

The implementation we presented is currently work in progress, and we plan to develop it further, improve its usability, and address implementation issues mentioned by the expert. This includes a more sophisticated method to find pairs of text snippets as comparison candidates. We further aim to conduct a comprehensive analysis of our approach and other text analysis approaches for scholarly works to shed light on its advantages and drawbacks compared to other systems.

6. Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) as part of the 'ePoetics' project, and the German Science Foundation (DFG) as part of the priority program (SPP) 1335 'Scalable Visual Analytics'.

7. References

Apache Foundation. (2014). Apache opennlp. `http://opennlp.apache.org`. version 1.5.3.

Correll, M., Witmore, M., and Gleicher, M. (2011). Exploring collections of tagged text for literary scholarship. *Comp. Graph. Forum*, 30(3):731–740.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. (2007). Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proc. 16th ACM Conf. on Inf. and Know. Man.*, pages 213–222, New York, NY, USA. ACM.

Eick, S. G., Steffen, J. L., and Sumner, Jr., E. E. (1992). Seesoft-a tool for visualizing line oriented software statistics. *IEEE Trans. Softw. Eng.*, 18(11):957–968, November.

Heimerl, F., Koch, S., Bosch, H., and Ertl, T. (2012). Visual classifier training for text document retrieval. *IEEE Trans. Vis. and Comp. Graph.*, 18(12):2839–2848, Dec.

Keim, D. and Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. In *IEEE Symp. VAST*, pages 115–122, Oct.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract models for a literary history*. Verso Books.

Moretti, F. (2013). *Distant Reading*. Verso Books.

Richter, S. (2010). *A History of Poetics: German Scholarly Aesthetics and Poetics in International Context, 1770-1960*. De Gruyter.

Robertson, G. G. and Mackinlay, J. D. (1993). The document lens. In *ACM Symp. on User Int. Soft. and Tech.*, pages 101–108.

Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *IEEE Conf. on Inf. Vis.*, pages 51–58, Oct.

V1 in Icelandic: A Multifactorical Visualization of Historical Data

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle,
Christian Rohrdantz, Dominik Sacha, Nicole Dehé, Daniel A. Keim
University of Konstanz
78457 Konstanz, Germany
firstname.lastname@uni-konstanz.de

Abstract

We present an innovative visualization technique for the analysis of historical data. We illustrate our method with respect to a diachronic case study involving V1 word order in Icelandic. A number of interacting factors have been proposed by linguists as being determinative of matrix declarative V1. The significance of these factors in contributing to declarative V1 can be explored interactively via our multifactorial visualization within a given text, but also comparatively over time. We apply the visualization to a corpus study based on the IcePaHC historical corpus of Icelandic and show that new results emerge very clearly out of the visualization component and that the appearance of declarative V1 is not confined to the situations identified so far by linguists. We demonstrate that the multifactorial visualization opens up new avenues for the exploration of alternative explanations. The visualization can be applied to any linguistic problem studying an interaction between several factors across time.

Keywords: V1 declaratives, Icelandic, IcePaHC, language change, corpus, visualization

1. Introduction

We present an innovative method for the visualization of multifactorial linguistic data over time. We present our visualization with respect to a concrete corpus study of V1 matrix declaratives found in the recently released annotated Icelandic parsed historical corpus (IcePaHC) (Wallenberg et al., 2011; Rögnvaldsson et al., 2012). The visualization allows for an "at-a-glance" appraisal of the occurrence of V1 declaratives and we are able to adduce evidence that advances the existing linguistic understanding of V1 in Icelandic.

Each instance of a V1 declarative is coded as to what kind of subject it occurs with and what kind of verbal element is involved. We focus on these factors as they have been proposed by previous literature as being determinants for the occurrence of V1 in declaratives. Our multifactorial visualization also shows whether the occurrence of a given V1 with a certain type of subject and a certain type of verb is higher or lower than expected.

The visualization brings out several unexpected results. For one, contrary to what is suggested in the previous literature, V1 is not overly associated with subject-less sentences, nor does it occur primarily with certain kinds of lexical verbs. Furthermore, although V1 is generally thought to be relatively stable in the history of Icelandic, the visualization shows a marked decrease as of 1900 and an intermediary dip around 1550 CE. The latter can be attributed to genre, as V1 primarily occurs in narratives and the corpus around 1550 mainly comprises religious texts. The reasons for the former effect (decrease of V1 from about 1900) remains subject for further linguistic research.

In the discussion, we emphasize that the majority of our results could not have been obtained via "standard" corpus linguistic methods and that multifactorial visualizations such as the one presented here are poised to play a key role in furthering our understanding of diachronic data.

2. Theoretical Background

2.1. Word Order and Diachronic Change

Questions of word order change in languages have constituted a major point of inquiry within historical linguistics. For example, like most Indo-European languages, the Germanic languages are generally taken to have been SOV languages in their earlier manisfestations (e.g., Eythórsson (1995); Fleischer and Schallert (2011); Fischer et al. (2000)). Today, English is an SVO language and German is what is known as a V2 language. The SOV, SVO or V2 word order generally refers to the word order in basic declarative matrix clauses. As is also well known, word order can vary within a language according to the type of clause expressed. For example, imperatives in German and English are V1 (verb initial), as are yes-no questions. However, German also contains instances of V1 in matrix declaratives (Axel, 2009; Hinterhölzl and Petrova, 2010). Some of the questions that arise with respect to historical change and word order are: 1) what determines changes in word order over time; 2) what determines differences in word order at a synchronic level? Many accounts to date invoke purely syntactic reasons for word order. One major factor that has been implicated in a change from SOV to SVO is the loss of case marking (e.g., see Kiparsky (1997); Kiparsky (2001) for a theory of diachronic change involving an interacting system of case, agreement and word order), factors such as finiteness are also known to play a role. More recently, a perspective that focuses more on the role of pragmatics and information-structure is being articulated in some detail (e.g., see Petrova (2011); Hinterhölzl and Petrova (2010); Hinterhölzl and Petrova (2011)).

In this paper, we look at data from Icelandic. Icelandic is a Germanic language that is rich in case marking and is a V2 language with fairly fixed word order. Moreover, V1 in matrix declaratives has existed throughout the recorded history of Icelandic and continues to exist today. In this paper, we examine data involving V1 matrix declaratives with the aim of identifying factors that facilitate V1 word

order.

2.2. V1 in Icelandic

Previous studies on V1 in Icelandic (e.g., Sigurðsson (1990); Franco (2008)) have argued that declarative V1 is mainly confined to narrative inversion and is connected to the introduction of known referents. These studies present a syntactic account by which V1 in essence is actually a V2 (verb second) construction, just with a *pro* (empty) subject. Sigurðsson additionally presents a small corpus study comparing stages of Icelandic. He concludes that V1 is a fairly stable phenomenon in Icelandic, with only two changes, one being that Old Icelandic allowed for referential as well as non-referential *pro* while Modern Icelandic only allows non-referential *pro*. In addition, Modern Icelandic now allows for an expletive in initial position, thus rendering some of the old V1 structures effectively into V2 structures. He also notes that V1 is found mostly in narrative texts, a finding that is confirmed by our larger corpus study. However, we find that V1 is not primarily confined to narrative inversion. The purely syntactic account as promoted by Sigurðsson (1990) and Franco (2008) implies that only unaccusative verbs (including passives, etc.) are possible in V1 declaratives. However, our findings show that V1 is not restricted by lexical class. We also found only 448 instances of subjectless V1 sentences as opposed to 4893 sentences with overt subjects. This result casts serious doubt on the hypothesis that V1 is actually underlyingly a V2 construction, just with an empty subject.

Some typical examples of declarative matrix V1 as identified by our corpus study are shown in (1)–(5).

(1) Vil ég heldur rita þeim hinum fám
will I rather write the.DAT those.DAT few.DAT
sinnum er þarf d og s
times.DAT REL required d and s
'I shall rather write *d* and *s* on the few occasions where it [z] is needed.'
(1150, *First Grammatical Treatise*)

(2) gengur síðan á brott.
goes then to away
'Then (he) goes away.'
(1350, *Finnboga saga ramma*)

(3) Verður Ketill forviða fyrir atsókn þeirra
becomes Ketill.NOM astonished for attack their
'Ketill becomes surprised by their attack.'
(1650, *Illuga saga Tagldarbana*)

(4) Gnæfði gaflinn hátt yfir fjöruna...
towered gable.the.NOM high over beach.the.ACC
'The gable towered hight over the beach...'
(1907, Jón Trausti *Leysing*)

(5) Sýndi drottinn mikla miskunn
showed Lord.NOM great.ACC mercy.ACC
vin sínum sankti Georgíum...
friend.DAT his_own.DAT saint.DAT George.DAT
'The Lord showed his great mercy to his friend St. George...'
(1525, *Georgíus Saga*)

(1) shows an example with a modal verb in V1 position and a pronominal subject, (2) a V1 without an overt subject, (3) illustrates a 'become' verb, (4) a V1 with a definite NP subject. We could so far not identify any particular verb classes that V1 is confined to. We did find V1 with the verb types identified in previous literature, i.e., unaccusatives, verbs of motion, verba dicendi, passives, presentationals, existentials and inchoatives. Significantly, however, we also found agentive transitives as in (5), which are not predicted by the syntactic account.

A significant portion of the V1 structures involved modals or the verbs 'do', 'have', 'be' and 'become'. Previous analyses (e.g., Franco (2008)) have suggested that modals are a fringe phenomena with respect to V1. This could not be confirmed via our corpus study. Furthermore, all types of verbs, auxiliaries and modals are used overwhelmingly with overt subjects. These overt subjects could be definites, indefinites or pronouns.

An alternative approach to V1 is the information structural view of declarative V1 put forward by Hinterhölzl and Petrova (2010), Hinterhölzl and Petrova (2011) and Petrova (2011) for German. Hinterhölzl & Petrova argue that V2 in German is derivative of an original V1 structure in Old High German. They develop a theory of information structure by which V1 occurs in particular contexts. For one, V1 is used systematically for presentational clauses and existential constructions. These clauses lack a topic-comment structure and the entire clause is in the scope of the assertion (focus). For another, V1 is used in declaratives with a range of verbs that includes motion verbs, verbs of saying and transformative/inchoative verbs. Hinterhölzl & Petrova argue that the basic function of V1 clauses is to introduce new referents to the discourse and that the various types of V1 clauses can be subsumed under the characteristics of lacking a topic-comment structure and the entire clause being in the scope of assertion, i.e. the entire clause is in focus.

Our data indicate that an information structural approach to V1 in Icelandic is more promising than a purely syntactic one, though the precise factors identified by Hinterhölzl & Petrova for German are not found in Icelandic. We conclude that more in-depth linguistic analysis is necessary and present the possibilities offered by our multifactorial visualization as a promising avenue for further research.

3. Initial Corpus Study

One of the properties of visual analytic techniques is the ability to present large amounts of data compactly while at the same time allowing for an interactive exploratory analysis of individual data items. The IcePaHC consists of 60 texts dating from the 12th to the 21st century. It comprises about 1 million words. The texts come from different genres and are not representative across the centuries. For example, in some centuries texts from narrative Sagas are more numerous than in others.

IcePaHC is annotated according to the syntactic annotation scheme of the Penn Treebank (Marcus et al., 1993). The annotation scheme is also compatible with the Penn historical corpora (Kroch and Taylor, 2000).

The corpus is in principle ideal for an investigation of V1 matrix declaratives as differences between matrix embedded

clauses are clearly coded, as are null and expletive vs. overt subjects. Among overt subjects, one can further distinguish between pronominal subjects vs. indefinite vs. definite NPs.
We conducted an initial query to the corpus via the CorpusSearch tool.[1] This search tool is specialized for dealing with corpora in the Penn-Treebank format. The initial query conducted with the CorpusSearch tool yielded roughly 4400 matrix V1 sentences with overt subjects (from a total of 73014 sentences) across nine centuries.
An advantage of the IcePaHC annotation is that yes-no questions and imperatives could be excluded automatically via the initial query as these are annotated clearly. Questions and imperatives can be distinguished readily from the matrix V1 declaratives that were the object of our study. However, an examination of the results showed that the results obtained by the CorpusSearch query were not entirely "clean", but contained some false hits of V1 questions and imperatives. These false hits were identified and removed automatically via a script, leaving 3964 matrix declarative V1 sentences.[2]
A further very valuable feature of IcePaHC is that verbs are coded for verb types: BE 'be', HV 'have', DO 'do', RD 'become', MD modals, VB main verbs. Since previous literature had indicated that verb class played a role, we included the type of the verb involved in V1 as one of the factors to be investigated.
A disadvantage we discovered as part of launching the initial query via CorpusSearch lies in the annotation of V1 declaratives that do not contain an overt subject (cf. (2)). IcePaHC conveniently does provide an annotation for empty subjects (NP-SBJ *pro*) so that one can easily identify subjectless sentences.[3] However, IcePaHC places these empty subjects **before** the main verb, thus rendering surface V1 structures as V2 structures in the annotated form. This design decision is consonant with the syntactic approach to V1 outlined above (seeing V1 as being underlyingly V2, just without an overt subject); however, it meant that subjectless V1 constructions are not recognized as such by a query targeting V1 structures.
We thus launched a more sophisticated query and identified a total of 5341 V1 declaratives out of an overall set of 73014 sentences. Table 1 shows the distribution of V1 according to the type of verb involved. The division into time periods has been done according to Haugen (1984).
The results of the corpus query show that declarative V1 occurs with all kinds of verbs and most frequently with main verbs. Thus, rather than being a fringe phenomenon, the occurrence with modals is a regular part of the language (e.g., contra Franco (2008)).
Recall that referentiality of the subject and the presence or absence of an overt subject have also been implicated in the appearance of declarative V1 in previous literature. It would therefore be instructive to know how each of the verbal categories behave with respect to referentiality and overt presence or absence of the subject NP. For each verb type, we therefore calculated the number of times it occurs with the following: a pronominal subject, a definite NP, an indefinite NP, a pro-dropped (overtly absent) subject and a null expletive subject.
We then checked whether the observed correlation between each verb type and each subject type was statistically significant via the χ^2-test, which measures an observed distribution against an expected distribution. The distribution of the five subject types within each time period was taken as the 'expected' factor, while the distribution for each possible verb type in relation to the three subject constructions was taken as the 'observed' factor. That is, for each text, we calculated the overall occurrence of type of subject with type of verb and then checked whether the numbers found for just the V1 declaratives deviated from this expected pattern. We were able to determine several highly significant correlations via this test.
However, some of the highly significant results are based on a very small data set. This is illustrated with respect to the data for the time period from 1900 to the present, shown in Table 2. Here, the patterns found with respect to 'have', 'be', 'do' and 'become' all come out as being highly significant. However, a look at Table 1 shows that the data being calculated with are very small: 10 instances for 'have', 2 for 'do', and 4 for 'become'.
We did find interesting indications from the χ^2-test for some other correlations. However, the overall picture is not clear and the χ^2-test does not identify particular correlations between verb type and subject type — it only points to an overall pattern that deviates from the overall pattern found in the larger corpus. As a next step, we therefore turned to the question of whether methods from visual analytics could help with identifying determining factors for the appearance of V1 throughout the history of Icelandic.

4. Visual Analytics

4.1. Motivation

Statistical analyses provide just a very rough picture of the data and are suited to either confirm or reject previously expected or at least anticipated knowledge, i.e. they are limited to the investigation of more or less evident hypotheses on the data characteristics.
A visual analysis of a given data set on the other hand allows one to push beyond what is possible with traditional corpus analysis. Instead of merely (dis)confirming the expected, it can lead to unexpected insights by granting an interactive explorative access to the data. The structured statistical analysis process becomes open-ended, in that visualizations of the data may enable researchers to spot unexpected patterns or outliers that they had not been aware of before and that might be of high relevance to them. In particular, visualizations may reveal feature patterns that disappear when using conventional statistics, especially when the statistics require parameters to be fixed, e.g., fixed time epochs that are the basis for some calculations.
In order to support such a visual analytics process in a beneficial manner, carefully designed visual mappings are required (cf. Mayer et al. (2010), Rohrdantz et al. (2012)). Usually, the very first design idea will not be maintained, but

[1] http://corpussearch.sourceforge.net/CS.html

[2] The results also revealed some errors in the annotation in that the corpus provides the infinitive (citation) form of verbs for each finite verb encountered. However, some of these infinitive forms were incorrect. We manually corrected these.

[3] In fact, it distinguishes between two different kinds of empty subjects: expletives and pro-dropped subjects.

Time	have	do	be	become	main verb	modal	Total
upto 1350	67	23	336	50	1045	127	1648
upto 1550	47	10	185	42	944	81	1309
upto 1750	58	6	225	47	569	64	969
upto 1900	64	6	217	62	818	75	1242
upto present	10	2	23	4	121	13	173
Total	246	47	986	205	3497	360	5341

Table 1: Instances of V1 according to type of verbal element

		Type of Subject in %					
Time	Verb Type	Pronoun	Indef. NP	Def. NP	*pro*	expletive	χ^2
1900–present	have	90	10	0	0	0	88.20***
1900–present	be	43	22	9	9	17	61.65***
1900–present	do	100	0	0	0	0	118.39***
1900–present	modal	54	8	0	31	8	7.63
1900–present	become	50	25	0	25	0	42.65***
1900–present	main verb	41	6	5	44	4	4.52
expected	all types	46	9	5	35	6	

Table 2: Significant Correlations according to the χ^2-test. Highly significant correlations are indicated with "***".

it enables insights that feed back into the design process. Our initial visualization showed each V1 sentence in the data individually, but then we decided to aggregate the sentence characteristics to the document level for visualization. Similarly, further details were changed iteratively during the design process. This included the integration of derivations of the original data: sometimes it is not the original numbers contained in the data that provide the best insights, but further calculations based on the initial raw numbers. We rejected the idea of visually representing the absolute number of appearances of certain phenomena in a text because the text lengths are quite heterogeneous and a proper comparison would have been difficult. Instead, the visual representation was designed to show whether such a number of appearances was higher or lower than expected.

4.2. Design and Development Process

Even though the visualization does not depend on previous assumptions, it requires some crucial initial design decisions which subsequently frame the kinds of insights that will be enabled. The challenge is to come up with solutions to plot the concrete data for generic explorations of historical data. Of course, the design can be improved iteratively, but it is quite helpful to reason first about which parts of the data might in principle be relevant and interesting to convey visually. Visualization designers and subject matter experts need to work tightly together and try to understand as much about the background and perspective of their counterparts as possible. In this project we pursued such a strategy. As a result of this, we suggest a novel visualization in order to provide both an overview on the described language corpus as well as insights at different levels of detail.

From an information visualization perspective the data consists of several texts, each of them including meta data and a set of sentences. Additional meta data about the texts, include, for example, genre, year and number of sentences. For each individual sentence, linguistic features can be derived as well. As a next step these features can be mapped to so called visual variables. We use a glyph for the visualization of a text within the corpus. Glyph representations of documents are known from the Information Retrieval field, e.g. the TileBars technique, which provides a compact and informative iconic representation of the documents' contents with respect to certain query terms (Hearst, 1995).

Our glyph contains several visual variables, namely *position*, *shape* and *color* in order to encode sentence features aggregated on the document level. Each glyph holds some shapes that provide information about the occurrence of (V1-)sentence features for the corresponding text. We therefore transform all elements (texts/documents, sentences and related features) into our internal data model and offer aggregated features on different concept levels (e.g., sentence features are aggregated to text features based on the number of sentences). We chose the glyph representation because the number of features/dimensions are suitable for a glyph visualization and in addition we are able to integrate shapes into those glyphs. That is why we designed a novel glyph instead of using standard histograms, line/bar/pie-charts or glyphs (e.g., star glyphs). The glyphs are arranged in space for exploration. We encode the text age and the genre to x- and y-position as they are the most interesting variables. In general, our intent is to use redundant encodings for relevant features.

Our visualization offers several interaction techniques in order to drill down in to the data if desired. First, we added zooming and panning interactions in order to navigate within the visualization's viewport. Our visualization also enables analysts to get details on demand through tooltip operations.

Our system is built using the *Java Piccolo 2D*[4] framework as it matches our requirements and has basic visualizations (e.g., basic shapes) and interactions built in. Language processing scripts that are the basis for feature extraction are integrated into the visualization system.

4.3. Our Multifactorial Visualization

Figures 3 and 4 show the overall visualization we arrived at. The texts contained in the IcePaHC are arranged from oldest to newest with the oldest represented at the very top, the newest at the very bottom. Each text is visualized as one composed glyph. A horizontal line extends from the middle of each glyph towards the right. This horizontal line is bisected by a vertical line. The horizontal line indicates the time span covered by the corpus. The vertical line shows where in the time span the text represented by the glyph is to be situated, see Figure 2 for an example.

Each glyph consists of three main parts. A horizontal bar on the top represents the length of the text in comparison to the longest text in the corpus. The bar of the longest text covers the whole width of the glyph. At the positions of the sentences containing V1, vertical light gray stripes are drawn in this horizontal bar. The user can thus determine the position of the V1 occurrence in the text and can determine where in the narrative flow of the text the V1 declarative is situated.

The main part of the glyph is a matrix containing colored items. The matrix represents the interaction of verb types with subject types. The columns encode verb types while the rows encode the subject types. The different verb and subject types can be identified by their position alone. However, for better visibility, verb types have redundantly been encoded by color and subject types have been encoded redundantly by shape. If a certain matrix cell is empty, the corresponding combination did not occur in the given text.

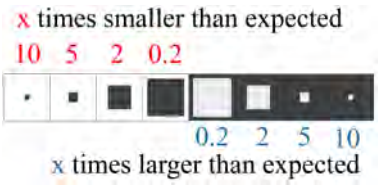


Figure 1: Different visual encodings depending on whether an observed frequency was higher or lower than expected. If the observed frequency is lower than expected, the outer part is white and the inner part is dyed in the color of the verb. If the observed frequency is higher than expected, the inner part is light gray and the outer part is dyed in the color of the verb. The ratio along one axis between both rectangle parts corresponds to the ratio between the observed and expected frequency.

The colored cell furthermore shows whether the combination of subject and verb type depicted by the cell occurred more or less frequently within the given text than expected. The deviation from the expected value is quantified via the scale shown in Figure 1.

[4]http://www.piccolo2d.org/

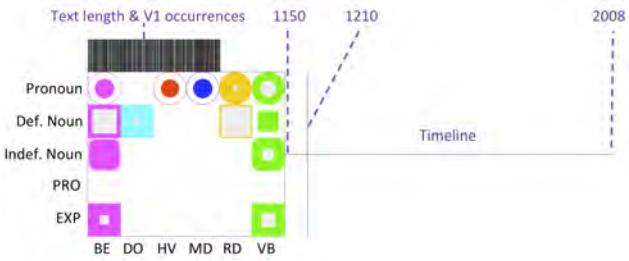


Figure 2: Glyph representation for a text consisting of three main parts: (1) A horizontal bar representing the text length and V1-occurrences, (2) A matrix containing colored shapes that belong to the occurrence of a particular verb and (3) a timeline on the right.

The visualization opens up new avenues of research for linguists. For one, it allows for a single, comprehensive view of the entire 60 texts of the corpus, as shown in Figures 3 and 4. Many details and possible influencing factors are displayed. For another, the visualization is interactive. After having identified potentially interesting patterns within the overall picture, the user can zoom into particular glyphs and examine the distribution of V1 declaratives within a particular text. The horizontal bar shows how long a text is and where in the text the particular V1 declarative is situated, allowing for an immediate assessment of the significance of the data. Furthermore, the visualization incorporates meta data on genre type. These can be accessed via mouse-over for each glyph. In Figures 3 and 4 all the Sagas have been arranged to appear on the very left of the visualization. The clump of texts in the middle (they appear around 1550) that are further to the right represent mainly religious texts (in the Figures, they appear at the bottom of Figure 3 and the top of Figure 4).

5. Insights and Discussion

The visualization shown in Figures 3 and 4 represents a powerful new tool for inquiries into diachronic change. Our findings to date show that proposals put forward to explain the occurrence of V1 declaratives in Icelandic so far do not match up with the actual patterns observed via the IcePaHC diachronic corpus, though we can confirm some of the findings of the previous literature.

Recall that Sigurðsson (1990) concluded that V1 is a relatively stable phenomenon in Icelandic, with the caveat that Modern Icelandic now allows for an expletive in initial position, thus rendering some of the old V1 structures effectively into V2 structures. Overall, this conclusion appears to be valid, as is his note that V1 is found mostly in narrative texts. As Figures 3 and 4 show, V1 is predominantly found in the Sagas (entirely narrative in nature). Our data also show that V1 shows a marked decrease as of 1900. This result is not entirely in keeping with Sigurðsson's observation. The innovation of an overt expletive by itself should not make such a huge difference as expletives overall did not constitute a large part of the V1 declarative occurrences (total of 239 of 5341 instances; 10 expletive V1 instances were found in the modern Icelandic texts).

Overall we note that a syntactic account which seeks to

understand declarative V1 primarily in terms of non-overt subjects (i.e., as underlying V2s) is not supported by our data. As shown in Figures 3 and 4, empty subjects (*pro*) and non-overt expletives are not generally found more frequently than expected. In absolute numbers, they make up a total of 448 of 5341 V1 declarative instances found in IcePaHC.

Figures 3 and 4 also show very clearly that all verb types occurred with V1, especially modals and lexical (main) verbs. V1 does not seem to have been restricted to only a subclass of verbal types. In particular, our data show that modals are not fringe phenomena with respect to V1, as suggested by the existing literature, but constitute a central part of the phenomenon.

Finally, the visualization of genre type in Figures 3 and 4 makes a set of data stand out. Although Icelandic is generally held to be stable with respect to V1 declaratives up until modern times, the texts used to represent the Icelandic of around 1550 show a comparative absence of V1. Since the texts are all not Sagas, but tend to be drawn from religious literature, we suspect that IcePaHC is characterized by a genre effect that was not originally intended by the developers of IcePaHC. We would assume that the Icelandic of around 1550 as according to IcePaHC will exhibit differences to the other parts of the corpus in other areas as well — this is a factor that could seriously impede diachronic studies. However, it is also a factor that is brought to the forefront of one's attention immediately by the visualization and can thus be factored into any analysis that is developed.

An issue that we have not addressed is whether the Icelandic data provide support for the alternative information structural approach to V1 articulated by Hinterhölzl & Petrova (2010, 2011) and Petrova (2011) for German.

The characteristics of our data set match their results in that the verb types involved in V1 are not confined to particular classes of verbs. However, V1 seems to appear with discourse old referents (pronouns, definite NPs) to a large degree, so if there is an information structural effect, it is not immediately obvious that it is the same one that was identified for German.

6. Conclusion

We present a powerful new visualization tool for the study of diachronic change. We have developed this tool with respect to a case study of V1 declaratives in Icelandic, however, the tool should be applicable to any diachronic study that seeks to understand a multifactorial interaction of the type exemplified by Icelandic declarative V1.

7. Acknowledgements

This work was partially funded by the German Research Foundation (DFG) under grant BU 1806/7-1 "Visual Analysis of Language Change and Use Patterns" and the Research Initiative *LingVisAnn* at the University of Konstanz, Germany.

8. References

Axel, K. (2009). The verb-second property in Old High German: Different ways of filling the prefield. In Hinterhölzl, R. and Petrova, S., editors, *Information Structure and Language Change: New Approaches to Word Order Variation in Germanic*, pages 17 – 43. Berlin: de Gruyter.

Eythórsson, T. (1995). *Verbal syntax in the Early Germanic languages*. Ph.D. thesis, Cornell University.

Fischer, O., van Kemenade, A., Koopman, W., and van der Wurff, W. (2000). *The Syntax of Early English*. Cambridge University Press, Cambridge.

Fleischer, J. and Schallert, O. (2011). *Historische Syntax des Deutschen: Eine Einführung*. Narr, Tübingen.

Franco, I. (2008). V1, V2 and criterial movement in Icelandic. *Studies in Linguistics*, 2:141 – 164.

Haugen, E. (1984). *Die skandinavischen Sprachen*. Hamburg: Buske.

Hearst, M. A. (1995). Tilebars: Visualization of term distribution information in full text information access. In *CHI*, pages 59–66.

Hinterhölzl, R. and Petrova, S. (2010). From V1 to V2 in West Germanic. *Lingua*, 120.2:315–328.

Hinterhölzl, R. and Petrova, S. (2011). Rhetorical Relations and Verb Placement in Old High German tatian. In *Salience. Multidisciplinary perspectives on its function in discourse*, pages 173–201. Berlin: de Gruyter.

Kiparsky, P. (1997). The rise of positional licensing. In van Kemenade, A. and Vincent, N., editors, *Parameters of Morphosyntactic Change*, pages 460–494. Cambridge University Press, Cambridge.

Kiparsky, P. (2001). Structural case in Finnish. *Lingua*, 111:315–376.

Kroch, A. and Taylor, A. (2000). Penn-Helsinki parsed corpus of Middle English. second edition.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

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

Petrova, S. (2011). Modeling word order variation in discourse: On the pragmatic properties of VS order in Old High German. *Oslo Studies in Language*, 3(3):209–228.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of LREC 2012*, pages 1978–1984.

Rohrdantz, C., Hund, M., Mayer, T., Wälchli, B., and Keim, D. A. (2012). The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Comput. Graph. Forum*, 31(3):935–944.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). *Icelandic Parced Historical Corpus (IcePaHC)*.

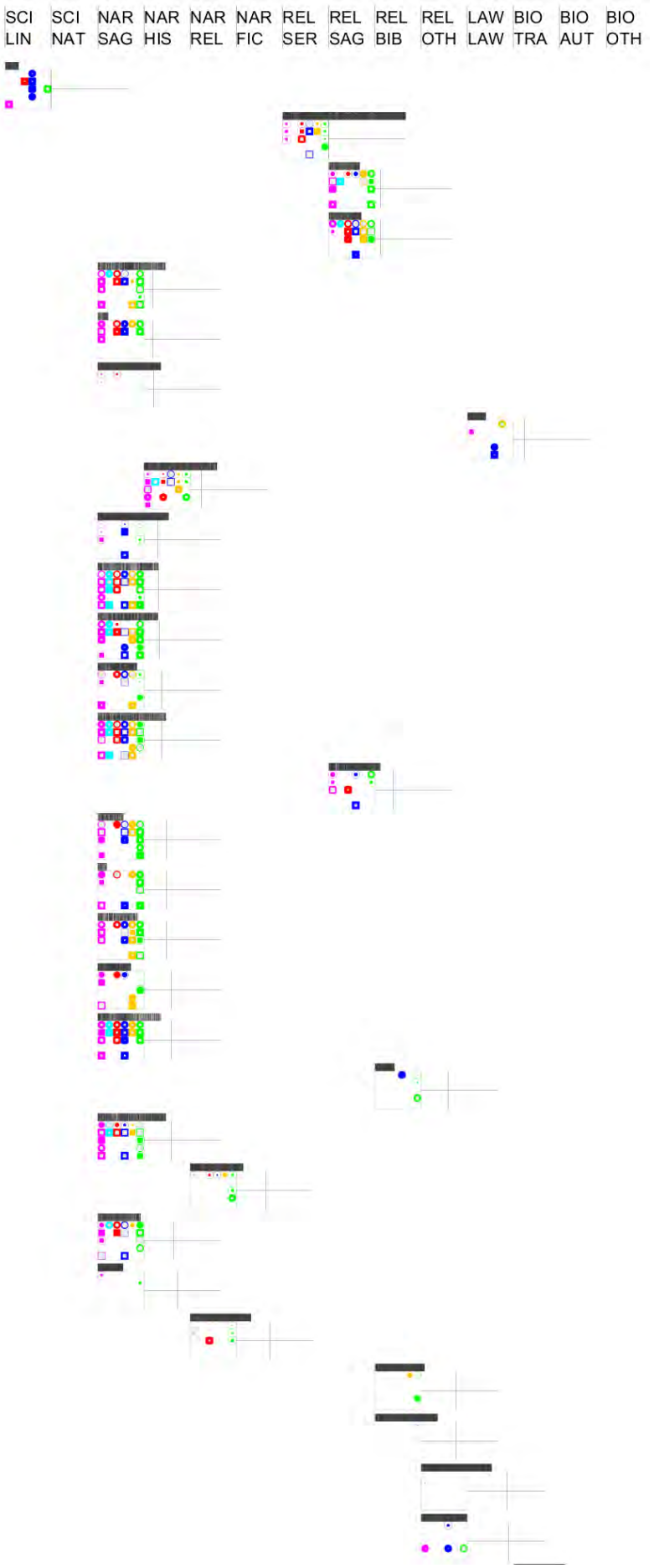


Figure 3: Top half of Visualization of V1 Occurrences in IcePaHC

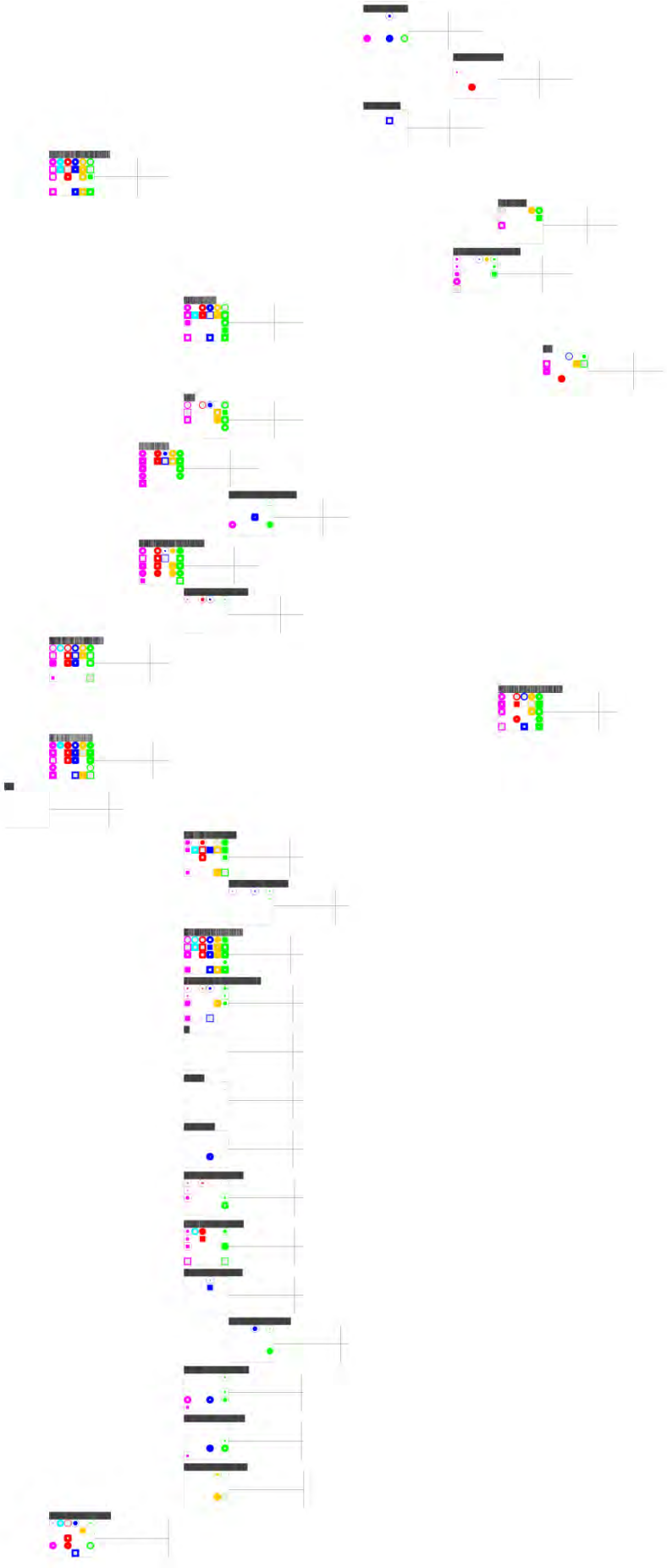


Figure 4: Bottom half of Visualization of V1 Occurrences in IcePaHC

Improving the Layout for Text Variant Graphs

Stefan Jänicke[1], **Marco Büchler**[2], **Gerik Scheuermann**[1]
[1] Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Leipzig
[2] Goettingen Centre for Digital Humanities, University of Goettingen, Goettingen
{stjaenicke,scheuermann}@informatik.uni-leipzig.de, mbuechler@gcdh.de

Abstract

Sentence Alignment Flows are visualizations for Text Variant Graphs that show the variations between different editions of texts. Although the resultant graph layouts are a substantial improvement to standard tools that are used in the corresponding Digital Humanities research field, the visualization is often cluttered due to large amounts of edge crossings and the occlusion of edges and vertices. In this paper, we present methods for the layering of vertices, the bundling of edges and the removal of overlaps between edges and vertices to reduce clutter, and therefore, to improve the readability for such graphs. Finally, we present the results of our survey with participants from the humanities and computer science, who had the task to compare the readability of Sentence Alignment Flows to the layouts generated by our improved method.

Keywords: Sentence Alignment Flows, Text Variant Graphs, Graph Layout, Directed Acyclic Graphs

1. Introduction

One of the substantial tasks in the field of textual criticism is called collation, which is the cautious comparison of various editions of one and the same text. The traditional approach of a humanities scholar is to put the texts of several editions next to each other and mark the differences among the textual entities (e.g., sentences, sections, chapters). Since this is an extremely laborious approach many Digital Humanities projects investigate tools that support the humanities scholars with computational methods.

To solve the collation task automatically, a great variety of algorithms was developed that compute the alignment between different text editions. CollateX (Dekker and Middell, 2011) is a standard tool used in the Digital Humanities that implements such alignment algorithms as well as provides a static visualization for Text Variant Graphs. The interactive interface Stemmaweb (Andrews and Macé, 2013) extends the CollateX graph to allow for user-driven annotation and modification of the graph structure. Both tools only provide plain horizontal layouts for Text Variant Graphs without highlighting its essential features. This makes it hard to visually follow subsequent tokens.

A recently published article proposes so called Sentence Alignment Flows (Jänicke et al., 2014) as a visualization solution for Text Variant Graphs. In the spirit of Wattenberg's Word Tree (Wattenberg and Viégas, 2008), it generates a proper overview that allows to follow how the words of an individual text edition disseminate in the graph. Therefore, it utilizes font size to highlight the number of occurrences of individual tokens and it uses horizontal links and vertical aligned splines to connect subsequent tokens. Figure 1 shows a Sentence Alignment Flow example. However, as discussed in Section 2.3. the layout algorithm is designed the way that occlusions of distinct splines and overlaps of splines and vertices often occur. Especially, when working with text editions that comprise lots of variation, these artifacts hamper the readability of the resultant Text Variant Graph layouts.

The purpose of this paper is to improve the readability for Text Variant Graphs based upon Sentence Alignment Flows. In particular, we provide the following methods to reduce visual clutter within the resultant graph layout:

- **Vertex Layering by Edition**: The goal is to place the subpaths of the Text Variant Graph that are passed by the same edition closely to each other to reduce the overall height of vertical connections to be drawn.
- **Improved Edge Routing**: This method aims to simplify the visual separation between different paths by bundling links with the same source or destination and removing occlusions of similarly routed links.
- **Overlap Removal**: In contrast to Sentence Alignment Flows, we suggest removing all occlusions between edges and vertices to avoid misinterpretations when observing the graph.

To obtain an objective evaluation, we finally conducted a survey with 53 participants. Their task was to compare the readability of the graph layouts computed with the Sentence Alignment Flow and our improved layout for Text Variant Graphs.

2. Related Work

For the purpose of modelling the differences and similarities between various editions of the same text, Schmidt et. al (Schmidt and Colomb, 2009) proposed so called Text Variant Graphs, which are directed acyclic graphs that emphasize such overlapping textual structures. Lots of research has been done in developing algorithms for directed

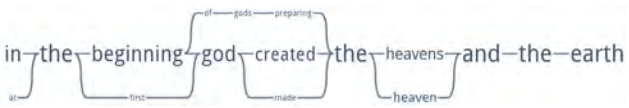


Figure 1: Sentence Alignment Flow for seven various English translations of the first Bible verse

acyclic graphs. In this section, we want to discuss traditional methods of the Graph Drawing community and the required steps in adopting some of the presented ideas for Text Variant Graphs. Furthermore, we want to consider methods developed for the dedicated research field in the Digital Humanities. Finally, we discuss Sentence Alignment Flows: a recently published layout algorithm for Text Variant Graphs.

2.1. Layout Algorithms for Directed Acyclic Graphs

Layered graph drawing as introduced by Sugiyama is the common drawing style used for directed acyclic graphs (Sugiyama et al., 1981). Typically, the vertices are placed on equally spaced horizontal (or vertical) layers and the edges are routed downwards (or rightwards) between the layers. Sugiyama's approach as well as many of its variations (Gansner et al., 1993; Cole, 2001; Utech et al., 1998; Eiglsperger et al., 2004) need to be adapted for the purpose of visualizing Text Variant Graphs, because only single vertices of one path are usually placed on one layer. This complicates the required vertical alignment of synonyms consistent of various amounts of tokens (e.g., "swarmed" and "brought forth abundantly" in *Genesis 1:21*). Additionally, the width of the vertices of a Text Variant Graph vary, so that a placing on vertical layers of equal width would further increase the distance between adjacent tokens.

To remove the occuring clutter for layered graph drawings with lots of edges, some approaches bundle edges to improve the readability of the resultant layouts (Eppstein et al., 2007; Pupyrev et al., 2011). Sentence Alignment Flows as well as its extension presented in this paper utilize this idea to compute well readable layouts for Text Variant Graphs.

2.2. Text Variant Graph Visualizations

The comparison of textual editions is a common task in textual criticism. Various research projects in the Digital Humanities focus on providing digitized editions of text to the collaborating humanities scholars.

Büchler proposed a horizontal alignment (Büchler et al., 2010) for the visualization of a Text Variant Graph for two similar text passages. One edition is used as a main branch and the variations to the second edition are highlighted in form of sub-branches in a certain color.

Several web-based tools were developed that also utilize the Text Variant Graph model to support the work with multiple digital text editions in the web browser (Dekker and Middell, 2011; Andrews and Macé, 2013). These tools compute horizontally aligned directed acyclic graphs with a plain design. The vertices are labeled with equally sized text tokens, which makes it hard to visually compare the number of occurrences of tokens. Directed edges labeled with the corresponding edition identifiers connect subsequent vertices. The layout often creates a very wide graph that contains edges routed opposed to the reading direction. These circumstances deteriorate the readability of the graph and make it hard for the observer to follow the route of a certain edition.

2.3. Sentence Alignment Flows

Within a Digital Humanities project, Sentence Alignment Flows were developed for visualizing Text Variant Graphs (Jänicke et al., 2014). In collaboration with humanities scholars experienced in the field of textual criticism, design principles were elaborated to facilitate the readability of the generated graphs, e.g., using vertex labels of different size to reflect the number of occurrences for a token, or horizontal layering of the vertices to improve the coherance of individual editions in the layout and the vertical alignment of synonyms. However, Sentence Alignment Flows produce well readable layouts for Text Variant Graphs if the extent of variation between the texts is limited. For examples with more complex variation structures, the resultant graph layouts are often cluttered due to the following incidents that are not treated by the algorithm:

Edge splines crossing text vertices: If connections are drawn between vertices of layers that are not adjacent to each other, occlusions between splines and vertices of the intermediate horizontal layers are possible (e.g. the edge between "helper" and "as" crosses the vertex "helpmate" in Figure 2). An improved vertex placement as described in Section 3.1. reduces the amount and length of vertical links, and therefore, the number of potential edge/vertex overlaps. A further strategy (Section 3.3.) is attached to remove occuring overlaps.

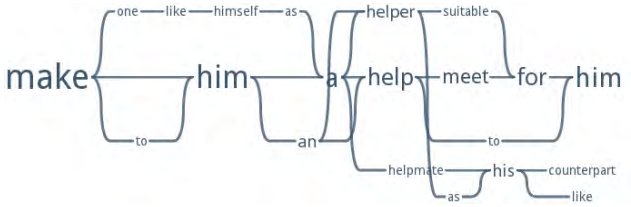


Figure 2: Edge/vertex overlaps in *Genesis 2:18*

Occlusion of similarly routed edges: An edge $\{t_1, t_2\}$ routed between text vertices of different layers consists of a spline starting from t_1, a horizontal line between the layers, and a spline ending at t_2. Thereby, it often happens that multiple horizontal lines occlude each other, so that it is hard to determine the destination of an edge. As shown in Figure 3, it is not clear if the vertex labeled "down" is linked to "as", to "like" or to both vertices. The method described in Section 3.2. solves this problem.

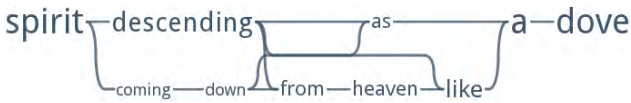


Figure 3: Edge overlaps in *John 1:32*

Occlusion of multiple splines: For strongly varying text passages lots of splines need to be drawn in the same area (see the token "adam" in Figure 4). This produces a large number of edge crossings and makes it difficult to identify the individual paths easily. The edge bundling approach proposed in Section 3.2. removes the crossings of a vertices' incoming and outgoing edges.

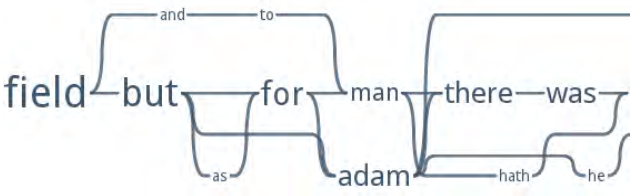


Figure 4: Edge spline occlusions in *Genesis 2:20*

3. Improved Text Variant Graph Layouts

Text Variant Graphs are constructed through various text editions transformed into paths, so that the graph basically contains multiple edges between the same vertices (like in Figure 5). Since these are routed side by side, it is sufficient to treat the readability issues mentioned in Section 2.3. for the "general case" with aggregate edges to be drawn.

3.1. Vertex Layering by Edition

When layers are assigned to vertices in the Sentence Alignment Flow algorithm, initially, the corresponding vertices for a dedicated edition e_0 are placed on the center layer 0. Iteratively, the shortest paths $\{t_1, ..., t_n\} \in G$ are determined with assigned layers for t_1 and t_n and missing layers for the vertices $\{t_2, ..., t_{n-1}\}$. The result is that the subpaths of an edition are treated at unpredictable iteration steps of the algorithm. Thus, the layers for consecutive subpaths of an edition may drift apart, so that some connections need to cross the intermediate layers. This can lead to frequent layer changes for individual editions, additional edge crossings and overlaps between vertices and edges.

We modify this process to place the subpaths of an edition as close as possible to each other by still achieving compactness of the graph. We also start with putting the vertices of e_0 on layer 0. Afterwards, we iteratively determine the next edition e_i with most vertices already assigned to layers. If multiple editions reach the same score, we choose the edition with vertices assigned to layers with lower absolute indices. For all subpaths of e_i to be inserted, we subsequently calculate the corresponding layer using the Sentence Alignment Flow method. This slightly modified approach treats the subpaths of e_i in a row, and thereby, keeps the extent of layer changes for e_i at a low level. An example can be seen in Figure 5. It uses the edition drawn in red color as e_0. The Sentence Alignment Flow algorithm (Figure 5 (a)) inserts the subpaths by increasing length in the order "so to jehova", "to yahweh your" and "to the lord your". Our proposed method (Figure 5 (b)) inserts the subpaths by edition in the order "so to jehova" (brown edition), "to the lord your" (green), and finally "to yahweh your" (orange). This reduces the total of layer changes among all editions.

3.2. Improved Edge Routing

To avoid cases such as shown in Figure 3, we insert a path layer p_i above each vertex layer l_i that reserves enough space to route horizontal links without overlaps. Furthermore, we remove overlaps of distinct vertical links. Between parallel routed edges, we require a minimum, configurable gap. The following four steps describe our improved edge routing procedure.

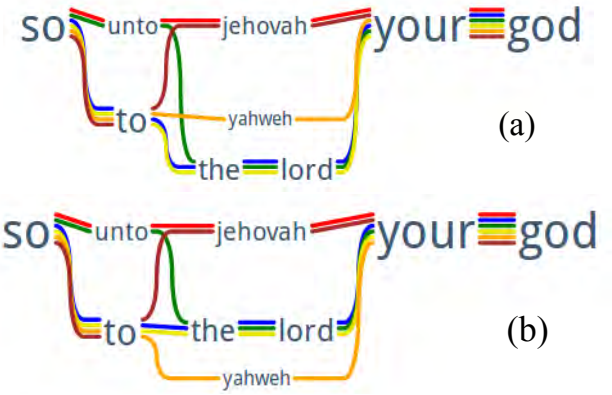


Figure 5: Vertex layering for part of *Deuteronomy 12:4*

3.2.1. Initialization of different edge types

We initialize the type for each edge $e = \{t_l, t_r\}$ dependent on the corresponding layers l_l and l_r of the connected vertices t_l and t_r. We separate three different edge types (see Figure 6):

type 0: If $l_l = l_r$ and there is no other vertex placed on l_l between t_l and t_r, e is drawn as a straight horizontal line.

type 1: If $l_l = l_r$ and there are vertices placed on l_l between t_l and t_r, a path is routed above l_l, consistent of an upward vertical link v_l, a horizontal link h on path layer p_l and a downward vertical link v_r.

type 2: If $l_l \neq l_r$, a path consistent of an upward (or downward) vertical link v_l, a horizontal link h on path layer p_l (or p_{l+1}) and an upward (or downward) vertical link v_r. We always put the horizontal link on the corresponding path layer with the higher absolute index.

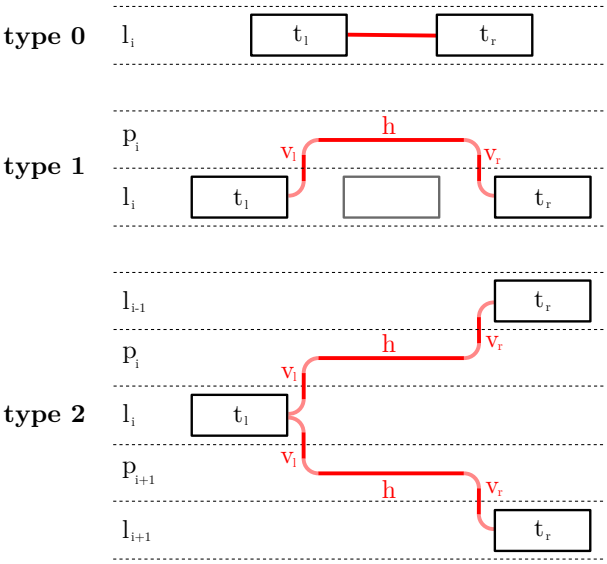


Figure 6: Different edge types

To smooth the graph layout, we connect each vertical link to its adjacent vertex and horizontal link using bends with radius r_b. Since four bends are required to draw edges of type 1 and 2, all adjacent vertices receive a minimum gap of $4 \cdot r_b$.

3.2.2. Bundling horizontal links

After all edges are initialized, we receive a list of horizontal links $h_1, \ldots, h_n$ with $h_i = \overline{t_l t_r}$ for each path layer p. We begin with constructing bundles $B = b_1, \ldots, b_n$ of horizontal links for all edges with the same left-hand vertex t_l and for all edges with the same right-hand vertex t_r. Thus, all horizontal links occur twice over all bundles. Afterwards, we sort B by decreasing number of horizontal links within the bundles. Iteratively, we insert the first bundle b_1 of B onto p. If b_1 overlaps with other already inserted bundles, we merge all these bundles into an overlap group. Then, we remove the duplicates of the horizontal links of b_1 from the remaining bundles of B and sort B again by decreasing number of horizontal links.

After placing all bundles onto p, we order and adjust the bundles of each overlap group parallel to each other. Thereby, we try to keep the number of edge crossings as minimal as possible. We perform this step iteratively by decreasing number of bundles in the overlap groups. Once a bundle that is part of multiple overlap groups is adjusted, it remains fixed for further ordering iterations. An example ordering is shown in Figure 7(a).

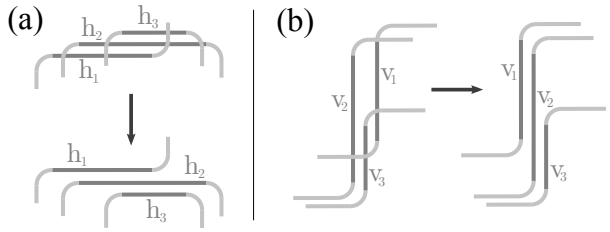


Figure 7: Ordering links of overlap groups

3.2.3. Bundling vertical links

For each vertex that is linked to neighbors with edges of type 1 and 2, we create a total of four bundles for its upward/downward incoming vertical links and its upward/downward outgoing vertical links. Since bundles of distinct vertices can be too close or even overlap each other, we perform the following two steps to keep the graph layout uncluttered. Firstly, we insert the bundles stepwise by increasing x-value into the graph layout. If a minimum gap to bundles that are already inserted is not given, we merge these bundles into overlap groups. Secondly, after all bundles are inserted, we order the vertical links of each overlap group, so that the number of edge crossings is minimal as it is shown in Figure 7(b). Finally, we test whether the required gaps between each vertical link v of the group and its subsequent glyph (right-hand vertical link v_r or vertex t_r) is still guaranteed. If this is not the case, we slightly shift all subsequent edges and vertices of v to the right so that the requirement is fulfilled.

3.2.4. Converting edges with type 2

To improve the readbility of the graph layout, we try to simplify edges of type 2 by removing one vertical link, and thereby, two of the four bends. Figure 8 illustrates an example of our approach. The upper connection $\{v_{l1}, h_1, v_{r1}\}$ between t_l and t_{r1} is replaced by $\{v_{new}, h_{new}\}$, because

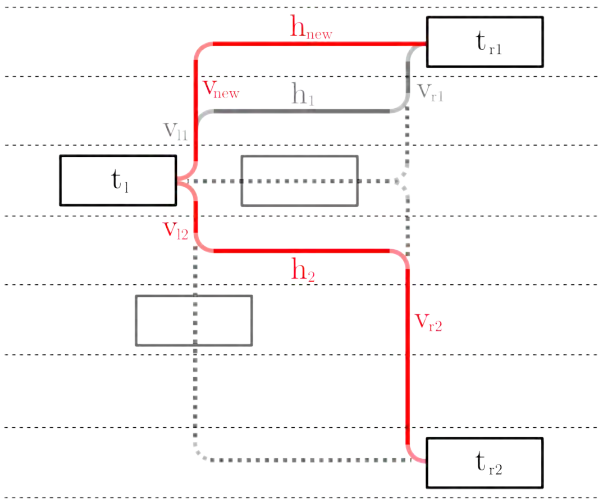


Figure 8: Converting edge type 2

neither v_{new} nor h_{new} cause an overlap with a vertex of the layout. Since this is the case for the lower connection $\{v_{l2}, h_2, v_{r2}\}$ between t_l and t_{r2}, it cannot be replaced.

There are two possibilities for each edge conversion, either the left hand vertical connection v_l gets removed and the right hand vertical connection v_r gets replaced by v_{new} or vice versa. If no overlaps are produced in both cases and $|l_l| > |l_r|$, we remove v_l and replace v_r. Otherwise, we remove v_r and replace v_l.

3.3. Removing overlaps between vertices and edges

Although the graph is designed the way the observer follows the spreading of an edition in horizontal direction, potential overlaps between a vertical link v that crosses an intermediate vertex layer l_m and a vertex t placed on l_m may hamper the readability of the graph. In such a case, we check if t can be moved horizontally without overlapping bends or other vertices by keeping the minimal required gaps to its neighbors.

An example can be seen in Figure 9(a). A leftward move-

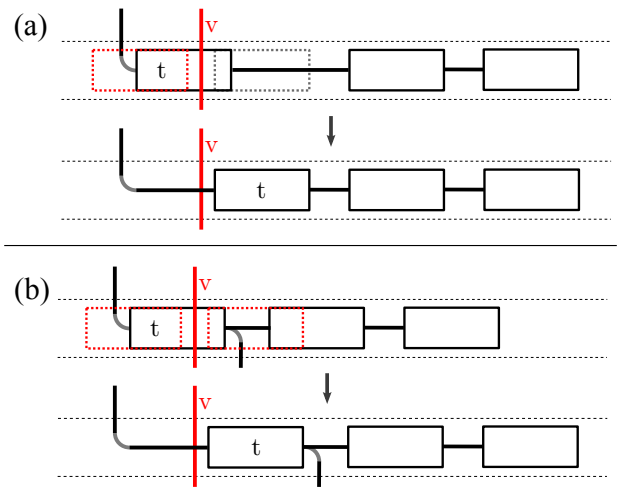


Figure 9: Overlap removal examples

ment of t is preferred, since the final position would be closer to its current position. Because this attempt fails, we move t to the right. If a horizontal movement of t is not possible (see Figure 9(b)), we move t and its subsequent edges and vertices, so that the overlap gets removed.

3.4. Improvement Validation

The benefit of the improved vertex layering method can be seen for the *Genesis 2:18* example. In comparison to the Sentence Alignment Flow layout in Figure 2, the vertical links in the modified layout (Figure 10) are shorter and rarely cross intermediate vertex layers. Furthermore, the overlap removal procedure moves the vertex labeled "helpmate" leftwards, so that the overlap with the connection between "helper" and "as" gets removed.

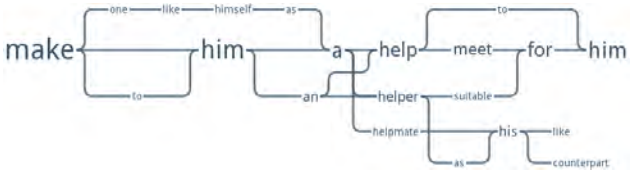


Figure 10: Removed edge/vertex overlaps in *Genesis 2:18*

Figure 11 shows the same example as Figure 3 with our new edge routing technique. The horizontal adjustment step unveils that the token "down" is connected to "as".

Figure 11: Removed edge overlaps in *John 1:32*

Finally, the advantage of bundling vertical links is visible in Figure 12. Compared to the splines drawn in Figure 4, the large number of crossings for the incoming and outgoing edges for the vertex "adam" is reduced.

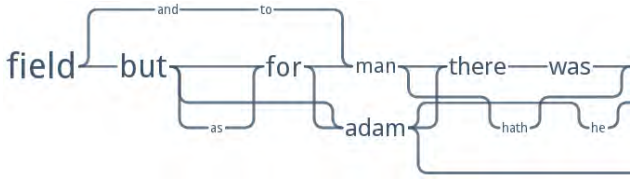


Figure 12: Removed edge spline occlusions in *Genesis 2:20*

4. Evaluation

To evaluate the benefit of the improvements for Text Variant Graph layouts described in Section 3., we conducted a survey with 53 participants. We decided against a mathematical comparison of both methods, because the resultant visualization for Text Variant Graphs needs to be understood and analyzed by humanities scholars working in textual criticism. Thus, our attitude was to let the target group decide whether our modifications are expedient or not. We further wanted to confirm our own expectations by computer scientists that are used to work with graphs. Dependent on the corresponding research fields, we could devide the participants roughly into three groups:

- **14 humanities scholars**, experienced in the field of textual criticism and partially involved in Digital Humanities projects
- **16 computer scientists**, working in Digital Humanities projects and partially dealing with issues in textual criticism
- **23 computer scientists**, most of them focused on research in visualization (partially experienced in graph visualization) and natural language processing

Within our project, we provide an interface to the humanities scholars that highlights the multiple colored edges on interaction (like in Figure 5), which further supports the understanding of the underlying structure. But for the purpose of discussing the graph readability issues, we confined the survey to the "general" case with aggregate edges.
The survey had a plain structure. The participants had to compare the readability of both layouts generated for six different examples. Each of the examples was an alignment of the same verse from seven various English translations of the Bible. The participants were asked to assess, whether they prefer Layout A (the resultant Sentence Alignment Flow), Layout B (our improved method), or if the readability of both layouts is similar. Furthermore, the reasons for the taken decisions were requested. Most of the participants had never seen any of both visualizations before the survey, so we could expect unbiased results.
We selected three of the Bible verses that contain the different structures with varying complexity that occur for typical Text Variant Graph use cases. Another three examples were randomly chosen from the remaining 28,629 verses. We wanted to assure that Text Variant Graphs benefit from our modifications generally, and not only a minor list of examples. The resultant layouts for the three selected verses are shown in Figure 14.
The results for all participants and the three different groups are juxtaposed in Figure 13. From a total of 318 comparisons, the Sentence Alignment Flow was preferred 36 times (11.3%), our improved layout 232 times (73%), and 49 times both layouts readability was assessed similarly (15.7%). When we take a look at the various groups, the

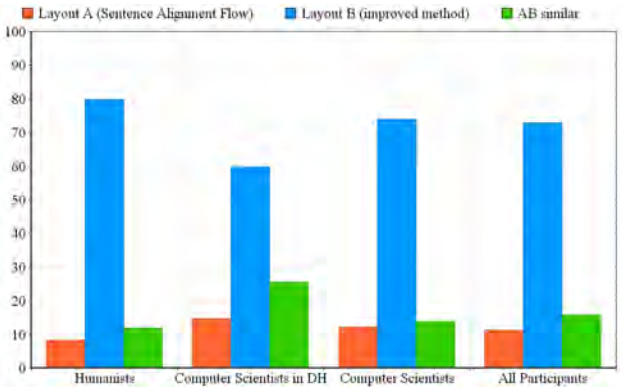
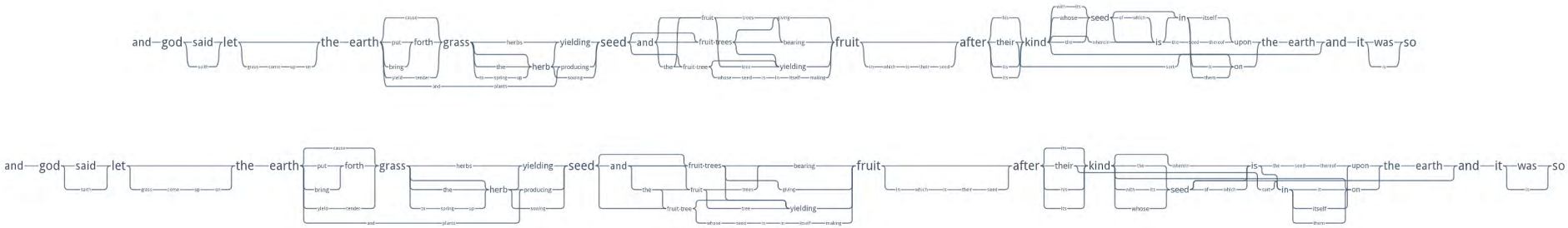
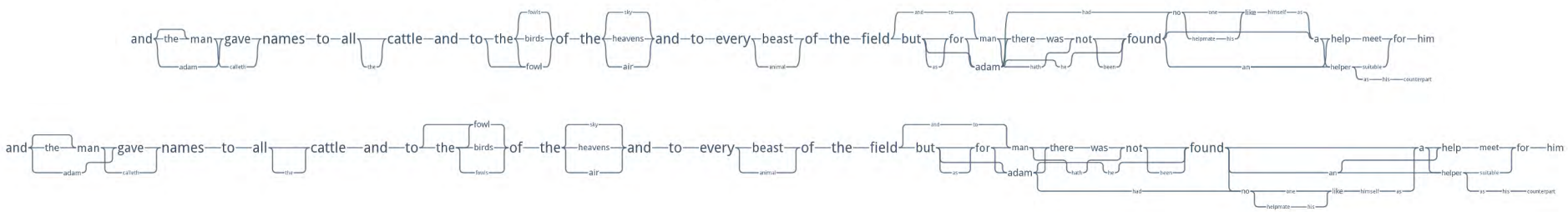


Figure 13: Preferred Layouts by participant group (in %)

Genesis 1:11



Genesis 2:20



Luke 17:1

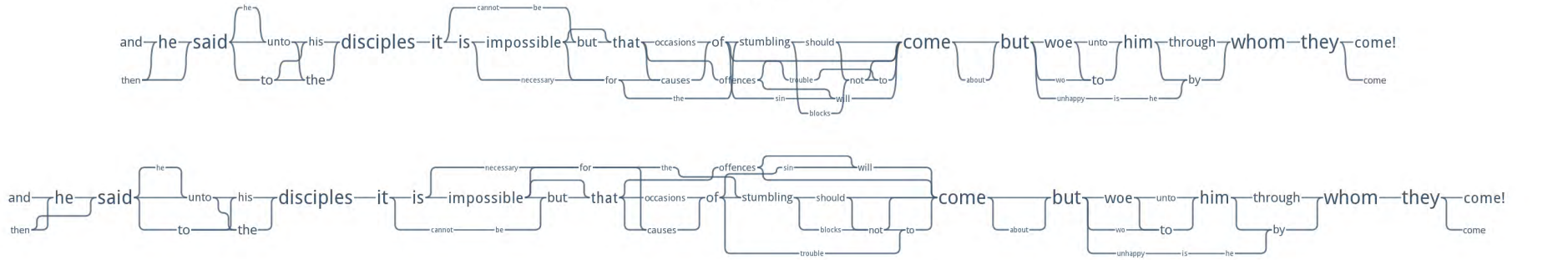


Figure 14: The three selected verses used within the survey. The upper layout is the result of the Sentence Alignment Flow, the bottom layout is the result of our method.

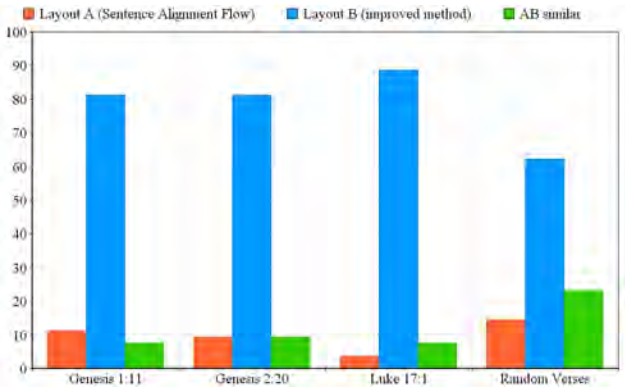


Figure 15: Preferred Layouts by verses (in %)

preferences vary only slightly. Especially satisfying was the evaluation of the dedicated target group from the humanities. In 79.8% of all cases (67 of the 84 presented layouts) the improved method of this paper was preferred over the Sentence Alignment Flow visualization.
Figure 15 shows the results for the individual examples. Our improved method turned out to be the participants' preference for the three selected verses. The reasons given for the decisions were congruent to the purpose of our modifications. Enclosed, some of the participants' comments:

- "Layout A seems to have more crossings and even lines through words which is disturbing."
- "Layout B has less clutter, no overlaps and crossings."
- "The line crossings in Layout B are orthogonal and fewer."

Particularly, one of the discussed issues of the Sentence Alignment Flow – the edge spline occlusions in *Genesis 2:20* starting from "adam" – was often perceived as problematic, e.g.:

- "In case of Layout A, the slightly vertical paths are harder to follow (e.g. at adam there is confusion of paths)."
- "At some point (e.g. after "adam"), it is easier to follow the text in example B."

For nearly one fourth (23%) of the randomly chosen verses the readability of both layouts was assessed as similar. Figure 17 shows one of the typical examples for these cases. Both layouts show less variation and clearly visible paths and the overall number of edge crossings is apparently reduced compared to the selected examples with a higher complexity. The Sentence Alignment Flow was favored in 14.5% of the cases over our new method. Often, the verses were short and properly visualized and with few or without edge crossings. An example is given in Figure 16. The participants preferring Sentence Alignment Flows argued like:

- "Compact Layout A for small sentence; easier to follow in one view."
- "Less spacing for Layout A."

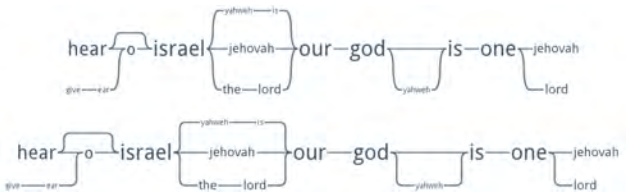


Figure 16: Sentence Alignment Flow preferred for *Deuteronomy 6:4*

- "... and Layout A is shorter."

Due to the various edge types and the required spacing between adjacent vertices, the width of the layouts computed by our method is constantly larger compared to the Sentence Alignment Flows. Although some participants favoring our method mentioned the too small spacings between subsequent vertices in Sentence Alignment Flows, the resultant compactness seems to improve the readability for some of the participants.
But our method was still mostly preferred (62%) also for the random verses. The slightly decreased percentage is attributable to the averagely lesser complexity of the random verses. Finally, some of the participants' comments justifying their decisions when favoring our improved method:

- "The lines are easier to follow – less twisty."
- "The circular edge bending looks better structured than the stretched/elliptical, which removes orthogonal feeling."
- "The word spacing and line curvature of the B layouts made them more readable to me."
- "I just like the angular style!"
- "Layout B seems to have less edge crossings and is more balanced (same amount of text above and below center line)."
- "Though the sentences in B are a lot longer/wider ... they are a lot clearer as there are no lines in the way of words while you are trying to read."

In conclusion, we determined that our method was constantly preferred when the corresponding examples contained lots of variation, and therefore, a larger amount of edge crossings. This fact confirms the benefit of the improvements proposed in this paper.

5. Conclusion

In this paper, we presented three methods to improve the readability for Text Variant Graphs compared to Sentence Alignment Flows that often contain clutter in form of edge crossings and occlusions between splines and text vertices. The first method improves the layering of the vertices that are inserted onto the layers in dependency to the corresponding editions, so that the resultant layering keeps vertices of one edition close to each other. This procedure shortens vertical links, and thereby, minimizes the amount of edge crossings and potential overlaps with other vertices.

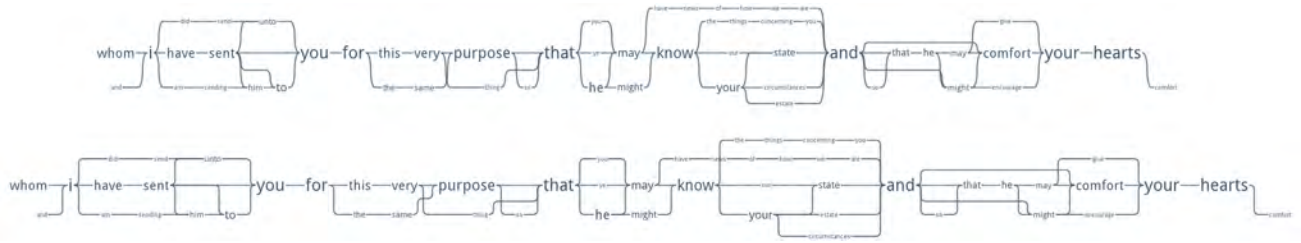


Figure 17: Similar readability of layouts for *Colossians 4:8*

As a second method, we proposed a four-step-approach to improve the routing of the edges, so that each path of the layout can be traced easily by the observer. Therefore, we separate different edge types dependent on the layers of the corresponding vertices, bundle links that share the same source or destination and route groups of overlapping links parallel to each other to avoid occlusions.

Finally, we provide a solution to remove all overlaps between vertical links and vertices. Either the corresponding vertex can be moved horizontally, or the vertex and all its successors are moved rightwards to avoid overlaps.

To evaluate our method, we conducted a survey with researchers from humanities and computer science. Their task was to compare the layouts for Text Variant Graphs produced by Sentence Alignment Flows and our method. For several examples, the participants could choose the preferred layout. Independent from the background, the majority judged the layout generated by our method as better readable, especially for examples with long texts, more complex variations and many edge crossings. For examples with minor variations and lesser edge crossings the readability of both layouts was often assessed similarly. In some cases, also the Sentence Alignment Flow was preferred.

One issue for the kind of visualization presented in this paper remains. When several editions vary the way that whole blocks of text are put in a different order, the width of the resultant graph increases rapidly, since only small parts of the texts can be aligned in form of a directed acyclic graph. To keep the graph layouts still compact and readable when highlighting such structures is one of the great challenges for the future developement of visualizations for Text Variant Graphs.

6. Acknowledgements

The authors like to thank Christian Heine for fruitful suggestions, Muhammad Faisal Cheema and Thomas Reimann for proof reading and the 53 participants of the survey for their time. This research was funded by the German Federal Ministry of Education and Research within the project eTRACES (project number: 01UA1101A).

7. References

Andrews, T. L. and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*.

Büchler, M., Geßner, A., Eckart, T., and Heyer, G. (2010). Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2).

Cole, R. (2001). Automated Layout of Concept Lattices Using Layered Diagrams and Additive Diagrams. In *Proceedings of the 24th Australasian Conference on Computer Science*, ACSC '01, pages 47–53, Washington, DC, USA. IEEE Computer Society.

Dekker, R. H. and Middell, G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. *Supporting Digital Humanities 2011*.

Eiglsperger, M., Siebenhaller, M., and Kaufmann, M. (2004). An Efficient Implementation of Sugiyama's Algorithm for Layered Graph Drawing. In *Proceedings of the 12th International Conference on Graph Drawing*, GD'04, pages 155–166, Berlin, Heidelberg. Springer-Verlag.

Eppstein, D., Goodrich, M. T., and Meng, J. Y. (2007). Confluent Layered Drawings. *Algorithmica*, 47(4):439–452.

Gansner, E. R., Koutsofios, E., North, S. C., and phong Vo, K. (1993). A Technique for Drawing Directed Graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230.

Jänicke, S., Geßner, A., Büchler, M., and Scheuermann, G. (2014). Visualizations for Text Re-use. In *GRAPP/IVAPP*, pages 59–70.

Pupyrev, S., Nachmanson, L., and Kaufmann, M. (2011). Improving Layered Graph Layouts with Edge Bundling. In Brandes, U. and Cornelsen, S., editors, *Graph Drawing*, volume 6502 of *Lecture Notes in Computer Science*, pages 329–340. Springer Berlin Heidelberg.

Schmidt, D. and Colomb, R. (2009). A Data Structure for Representing Multi-version Texts Online. *Int. J. Hum.-Comput. Stud.*, 67(6):497–514, June.

Sugiyama, K., Tagawa, S., and Toda, M. (1981). Methods for Visual Understanding of Hierarchical System Structures. *Systems, Man and Cybernetics, IEEE Transactions on*, 11(2):109–125, Feb.

Utech, J., Branke, J., Schmeck, H., and Eades, P. (1998). An Evolutionary Algorithm for Drawing Directed Graphs. In *Proceedings of the International Conference on Imaging Science, Systems and Technology*, pages 154–160. CSREA Press.

Wattenberg, M. and Viégas, F. B. (2008). The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, November.