



**Universität
Zürich^{UZH}**

Visualization of Narrative Structures

Master's thesis in Computer Science

Presented by

Katrin Affolter

Bergweg 2, 4142 Münchenstein

Student-ID 10-917-250

Prepared at

Institute of Computational Linguistics

University of Zurich

Prof. Dr. Martin Volk

Supervisor: Dr. Noah Bubenhofer

Date of submission: August 25, 2016

Abstract

Social behavior affects various aspects of our life and is said to be mirrored in typical linguistic usage, which contains statistically significant patterns of a speaker. In my thesis I use a data-driven approach to identify the narrative patterns in a corpus. I developed an interactive visualization based on visualization theory to analyze those extracted patterns. Furthermore, I gathered two different corpora, which I used to show the possibilities for analyzation opened up by my visualization.

Zusammenfassung

Soziales Handeln wirkt sich auf verschiedene Aspekte unseres Lebens aus und spiegelt sich im typischen Sprachgebrauch wieder, welcher statistisch auffällige Muster enthält. In meiner Arbeit verwende ich einen datengeleiteten Ansatz, um narrative Muster in einem Korpus zu identifizieren. Um die extrahierten Muster zu analysieren, habe ich eine interaktive Visualisierung, basierend auf der Visualisierungstheorie, entwickelt. Des Weiteren habe ich zwei verschiedene Korpora zusammengestellt, welche ich verwendet habe um die Analysemöglichkeiten meiner Visualisierung aufzudecken.

Acknowledgment

I want to express my gratitude to Prof. Dr. Martin Volk and Dr. Noah Bubenhofer, not only for giving me the chance to work on a very interesting project as my master's thesis, but also for their support during my time as a student at the University of Zurich. I would also like to thank Klaus Rothenhäusler and Danica Pajović for their amicable and productive team-work throughout the past year.

Contents

1. Introduction	1
2. Narrative Pattern	4
2.1. Estimation of complex- n -grams	5
2.2. Relationship between complex- n -grams	6
2.3. Position of complex- n -grams	7
3. Theoretical Background	8
3.1. Graph Theory	8
3.2. Visualization Theory	9
4. Corpora	20
4.1. Mountaineering Reports	20
4.2. Coming-Out Stories	26
5. NarrViz	29
5.1. Data Structure	30
5.2. Layout	30
5.3. Filter Functions	36
5.4. Customization	36
5.5. Exemplary Analysis	38
6. Conclusion	44
Bibliography	48
A. Appendix	50

List of Figures

3.1.	The preattentive attributes. [source: Meirelles, 2013, p. 23]	11
3.2.	The analytical patterns by Taylor (2014).	12
3.3.	Examples of the influence of whitespace on our perception. [source: Meirelles (2013, p. 19) and Graham (2008)]	13
3.4.	Examples of the influence of similarity on our perception. [source: Meirelles (2013, p. 51) and Graham (2008)]	14
3.5.	Examples of the influence of closure on our perception. [source: Ware (2012, p. 195) and Graham (2008)]	14
3.6.	Examples of the influence of continuity on our perception. [source: Meirelles (2013, p. 58) and Ware (2012, p. 192)]	15
3.7.	Example for the influence of symmetry on our perception. [source: Ware (2012, p. 193)]	16
3.8.	Examples of the influence of figure & ground on our perception. [source: Meirelles (2013, p. 126)]	16
3.9.	Examples of the influence of connection on our perception. [source: Ware (2012, p. 191)]	17
4.1.	Different versions of the boundary problem (white = R149; blue = R151).	22
4.2.	Distribution of the length of all mountaineering reports.	24
5.1.	Screenshot of the visualization with the mountaineering reports displayed.	29
5.2.	Example graph with 6 nodes, 4 links and 3 sections.	31
5.3.	Example hover of the node <i>E</i> from Figure 5.2	33
5.4.	Box plot of a stable (left) and an unstable (right) <i>n</i> -gram.	34
5.5.	Example of the rectangle representation of the coming-out corpus.	35
5.6.	Filter options	36
5.7.	Layout options	37
5.8.	Tooltip options	37

5.9. Coming-out corpus with five sections.	38
5.10. Overview of the mountaineering reports corpus.	39
5.11. Segment of the first section of the mountaineering reports.	39
5.12. Segment of the sections between 11 and 15 with maximal distribution difference = 14.	40
5.13. Overview of the coming-out corpus.	41
5.14. First section of the coming-out stories.	41
5.15. Middle sections of the coming-out stories (label type: show example).	42
5.16. End sections of the coming-out stories.	43
5.17. Rectangle representation of the node ', dass PPER ADJA sein ,' (Engl. ', that PPER be ADJA ,').	43

List of Tables

4.1. Number of articles I was able to label (and how many of them are mountaineering reports) with different approaches. Each approach is based on the previous one.	21
--	----

List of Acronyms

CWB	Corpus W ork b ench
JSON	Java S cript O bject N otation
NER	Named E ntity R ecognition
OCR	Optical C haracter R ecognition
POS	Part-Of- S peech
SAC	Swiss A lpine C lub
XML	e X tensible M arkup L anguage

1. Introduction

Social behavior affects various aspects of our life and is said to be mirrored in typical linguistic usage, which contains statistically significant patterns of a speaker (or a group of speakers). Therefore, the social organization of the world can be studied by observing the typical linguistic usage of individuals or groups of people (Bubenhofer, 2009, p. 2-3). Language usage patterns are indicators for discourses. This enables us to analyze the language data for their exemplariness and derive inductive discourse descriptions (Bubenhofer, 2009, p. 6).

My thesis is based on the paper of Bubenhofer, Müller, and Scharloth (2013). They show a data-driven approach to identify narrative patterns in a corpus consisting of reports about people's first sexual experiences. Apart from showing that it is possible to identify narrative patterns and visualize them with a directed graph, Bubenhofer et al. (2013) also mention some drawbacks of their approach: the visualization is only a static graph and the n -grams consist of tokens. If we want to look at stories independent of the author's gender, the drawback of token- n -grams in their corpus is, that the stories are written by men and women. If they refer to the other gender, they use *he* or *she*, which is not grouped into one n -gram. For example, the n -gram '*fragte sie mich ob ich*' (Engl. '*she asked me if I*') and the n -gram '*fragte er mich ob ich*' (Engl. '*he asked me if I*') are treated separately because of the *sie* (Engl. *she*) and *er* (Engl. *he*). The drawback of a static graph is the missing possibility of filtering and there is no possibility to get additional information on demand. This means that the graph either is overcrowded with information in the overview or it lacks information for a better analysis of the graph. For instance, if we look at the labels of the nodes: If they are always displayed they would overlap and most of them would be unreadable in the overview. But if they are hidden, we would miss the information while zoomed in, when the labels could be displayed without overlapping.

Bubenhofer et al. (2013) analyze different visualization methods: tables, collo-

cation and hierarchical graphs. They conclude that the best representation is a hierarchical graph because it shows the collocation profile and the position in the text as the hierarchical element. It combines the two important compounds of narrative patterns in a single visualization: *position* and *relationship* (collocation). Their implementation is a static graph that only gives a first overview or only shows a small segment of the data. One component of an exploratory graphic (cf. p. 10) is the interaction between user and visualization. Following the *Visual Information Seeking Mantra* by Shneiderman (1996) (cf. p. 19), we can expand the graph with functions for *zoom and filter* and *details-on-demand*, thereby offering methods to interact with the data and explore it in different ways. To do so, I will implement a web application with filter functions and the possibility to customize the visualization.

In the methodic part Bubenhofer et al. (2013) also address that the n -gram extraction can be expanded to complex- n -grams. They consist of a combination of different linguistic units, instead of only a specific one. I will use complex- n -grams, which consist of a combination of part-of-speech tags and lemmata (cf. p. 5).

In chapter 2, I will describe the characteristics of narrative patterns. Furthermore, I will describe the necessary steps to extract (complex-) n -grams, their positions, relationships and how to use this information to identify narrative patterns in a corpus.

In chapter 3, I will introduce the principles of graph (3.1) and visualization theory (3.2). Graph theory describes the data structure for the narrative patterns, while visualization theory describes different concepts how to design useful visualizations.

In chapter 4, I will introduce two different corpora, which I will later use to show the possibilities for analyzation of my visualization. The first corpus I use consists of mountaineering reports from the "Text + Berg" corpus (Bubenhofer, Volk, Leuenberger, & Wüest, 2015). The second corpus consists of coming-out stories which I gathered from the website *dbna - das Magazin für schwule Jungs!* (1997 - 2016).

In the last chapter, I will introduce my visualization for narrative patterns. It consists of the data structure, which I use for the complex- n -grams and their relationships, the basic functionality of the visualization, the possibilities of customization and finally I will use the corpora described in chapter 4 to show how my visualization

can be conduct various analyses.

2. Narrative Pattern

Narratives have a seriality and consist of their sum of typical pledges (Germ. *Ver-satzstücke*). The exemplary sequence of linguistic means of expression is interpreted as evidence of narrative patterns. Therefore, typical narratives follow narrative patterns, which in turn consist of certain linguistic patterns. Narrative structures are characterized by definable episodes, which make up causalities by their arrangement and thus constitute the storyline. Narratives are socially accepted patterns of interpretation. They shape our perceptions and representations of relationships at the same time (Bubenhofer et al., 2013).

Narrative patterns allow the identification of specific manners of language use, representing specific preferences for a possible expression variant (Bubenhofer, 2009, p. 4). This creates a knowledge base about the speakable and non-speakable. Collective concepts and values are actively constituted and updated, which shapes the conception of the world and hierarchies of values. For this reason, there is interest in the analysis of typical pledges, components of stories and their corresponding assembly and association (Bubenhofer et al., 2013).

To identify narrative patterns we are not interested in individual documents, but in the exemplariness in a document collection (Bubenhofer & Scharloth, 2013). The corpus pragmatically indicates significantly frequently occurring linguistic patterns as a result of recurrent speech acts of the texts contained in the corpus (Bubenhofer & Scharloth, 2011). The work on narrative patterns is therefore an empirical one. This means that trends in real data are detected, but can only be noticed if we take a step back and look at the whole picture. Each individual document contributes something to this picture (Bubenhofer & Scharloth, 2013).

In terms of method, the calculation of n -grams is suitable for the inductive analysis (Bubenhofer & Scharloth, 2013). N -grams consist of n sequential units, such as of words. This means that typical n -grams and their typical sequences are detected in

the corpus (Bubenhofer et al., 2013).

2.1. Estimation of complex- n -grams

The typical unit used to express n -grams are tokens or lemmata. Complex- n -grams are an extension of n -grams, where the n -gram can consist of combinations of different linguistic units instead of only one specific unit. For example, we can use part-of-speech (POS) tags and lemmata within the same n -gram. Complex- n -grams contain more information than the ordinary token- n -grams. They can combine different variants of the same underlying morphosyntactic pattern (Bubenhofer & Scharloth, 2011).

We do not need to know all possible complex- n -grams. We want to find those complex- n -grams for a specific corpus, which occur with a statistically relevant frequency. Therefore, it is necessary to compare the corpus with a reference corpus (Bubenhofer & Scharloth, 2011).

In my analysis I calculated the complex- n -grams by combining lemmata, POS tags, named entity information (if available) and with at least five units per n -gram. I used the following POS tags for the reduction¹: numbers (CARD), article (ART), adjective (ADJA & ADJD), personal pronoun (PRF & PPER), relative pronoun (PRELS & PRELAT), demonstrative pronoun (PDAT & PDS) and possessive pronouns (PPOSS & PPOSAT). The reduction of numbers is self-explanatory. The reduction of pronouns is most important for the mountaineering reports. Some authors write in plural (first person plural), others write more directly from their perspective (first person singular). The idea behind the reduction of adjectives is based on the empurpled writing style in the mountaineering reports. As reference corpus I used the "Text + Berg" corpus.

This leads to complex- n -grams like '*PPER CARD Jahr ADJD sein*' (Engl. '*PPER be CARD year ADJA*') which consists of token- n -grams like '*ich 14 Jahre alt war*' (Engl. '*I was 14 years old*'), '*ich 15 Jahre alt war*' (Engl. '*I was 15 years old*'), and so on. The drawback of this approach is, for example, the complex- n -gram '*sein PPER in ART ADJA*' (Engl. '*be PPER in ART ADJA*') which, for instance, consists of the token- n -grams '*bin in der ganzen*' (Engl. '*am in the whole*'), '*bin in*

¹If a word is labeled with an other POS tag, the POS tag will be ignored.

einer schlimmen' (Engl. 'am in a bad') and '*war in der 7.*' (Engl. 'was in the 7th'). The meaning of those token- n -grams are not the same, but are put together in one complex- n -gram. The reduction of adjectives is still a good approach because we receive complex- n -grams like '*PPER sein so ADJA*' (Engl. '*PPER be so ADJA*') where *ADJA* consists of *happy*, *glad* and *proud*.

To calculate the complex- n -grams, I used the program *cwb-n-grams*², which is an extension to the Open Corpus Workbench (Evert, 2010). The significance was tested with log-likelihood. The result was a list of all significant complex- n -grams with their level of significance and frequency.

I am also interested in the original, token-based occurrence of the complex- n -grams in the corpus. Those token- n -grams are saved as the *examples* of the complex- n -grams.

2.2. Relationship between complex- n -grams

As mentioned above, narratives do not only consist of typical n -grams, but also of the relationship between them. This defines the narrative patterns. To estimate those relationships the analysis of collocations is used in linguistics. This means, that for each complex- n -gram pair (c_i, c_j) I calculate whether c_i (called *basis*) and c_j occur together significantly often, compared to all other possible bases of c_j .

To extract all relevant complex- n -gram pairs (c_i, c_j) , the complex- n -grams in a context range of 30% of the text length³ are used. In contrast to Bubenhofer et al. (2013) who investigated at the left and the right side of c_i , I only considered the right side. By doing so, each collocation has a direction corresponding to the occurrence in the text. Furthermore, for the significance test I used the Fisher's exact test⁴.

This resulted in a directed graph with complex- n -grams as nodes and the collocation relationships as links (more about graph theory in section 3.1).

²Research team: *semtracks*

³based on Bubenhofer et al. (2013)

⁴Python implementation: Haibao Tang & Brent Pedersen
(https://github.com/brentp/fishers_exact_test/)

2.3. Position of complex- n -grams

The seriality is not only given through the relationship between the complex- n -grams, but also through the possible positions in the text. To get the typical position, it makes sense to calculate all possible positions first and afterwards compute the median. The advantage of the median is, that it reduces the influence of outliers. Another possible choice would be to take the position in which the complex- n -gram occurs most often. Furthermore segmentation into sections can be done on token level, sentence level and others more. I decided to use the token level.

The number of sections used to segment the text, can be chosen depending on the goal. If we choose a small number, we will get a more overall view of the data. If we choose a higher number, we get a more detailed look at the narrative patterns.

3. Theoretical Background

My visualization is based on the conclusion of Bubenhofer et al. (2013) that the best visualization for narrative patterns would be a hierarchical graph with the position in the text as the hierarchical element. I will use aspects from graph and visualization theory to improve the simple and static hierarchical graph. Graph theory is important to understand the data structure and the corresponding features. Visualization theory is the fundament for the creation of good and useful visualizations. Furthermore, it includes guidelines to create comprehensible layouts.

3.1. Graph Theory

A *graph* G is a data structure. It consists of a set of *nodes* (or *vertices*) V and a set of *edges* E . Formally, a graph is defined as a set $G = (V, E)$, where $V = v_1, \dots, v_m$ and each edge in E is defined as $e_{i,j} = (v_i, v_j)$ (Rada Mihalcea & Radev, 2011, p. 11; Voloshin, 2009, p. 2). In a visualization nodes are represented as points and edges as lines, which connect two points (Voloshin, 2009, p. 1).

A graph can be either *directed* or *undirected*. In a directed graph (or digraph) every edge has a defined direction of travel. This means that there is a distinction between the edge (v_i, v_j) and the edge (v_j, v_i) . The node v_i of the edge (v_i, v_j) is called the *tail* and the node v_j is called the *head* (Rada Mihalcea & Radev, 2011, p. 11). In a directed graph it is possible to identify a hierarchy and display them appropriately (Wilson, 1996, p. 56).

Two nodes v_i and v_j are called adjacent, if an edge (v_i, v_j) or (v_j, v_i) exists. If there is no such edge, they are called disjoint. Those adjacent nodes of v_i are called *neighbors*, and all the neighbors together are called *neighborhood* of v_i . The number of neighbors is called *degree*, denoted by $\delta(v_i)$. In a directed graph, the degree is split into the *in-degree* $\delta^-(v_i)$, defined as the number of edges that come from an

other node to v_i and into the *out-degree* $\delta^+(v_i)$, defined as the number of edges that go from v_i to an other node (Rada Mihalcea & Radev, 2011, p. 12; Voloshin, 2009, p. 2). A node v_i with $\delta^-(v_i) = 0$ is called a *source*. Similarly, a node with $\delta^+(v_i) = 0$ is called a *sink* (Voloshin, 2009, p. 123).

The *maximum degree* of G is the maximum degree over all nodes, denoted by $\Delta(G)$. The *average degree* of a graph is defined as $\frac{1}{m} \sum_{i=1}^m \delta(v_i)$, where m is the number of nodes. If all nodes have the same degree, the graph is called *regular* (Rada Mihalcea & Radev, 2011, p. 13; Voloshin, 2009, p. 2).

A sequence of nodes that are connected with edges, is called *path*. In a directed graph the direction has to be considered. If any two nodes are connected by a path, the graph is called *connected*. The *shortest path problem* is the problem of finding the shortest path in graph G from the node v_i to any other node in the graph. In a unweighted graph, the length is calculated as the number of edges. In a weighted graph, where each edge has a cost or a weight, we can use algorithms such as the *Dijkstra* or *Floyd-Warshall* algorithm (Rada Mihalcea & Radev, 2011, p. 13).

3.2. Visualization Theory

"A picture is worth a thousand words"

This English idiom can be extended to visual presentations, which are extremely easier to understand than a textual description (or spoken report) of the same task (Shneiderman, 1996). The idea behind visualization of abstract information is to reveal patterns, clusters, gaps or outliers in the data (Shneiderman, 1996). Moreover, *visual analytics* is a process which combines automated analysis techniques with interaction between the user and the data, leading to a better understanding, reasoning and decision making on the basis of very large and complex data (Keim, Kohlhammer, Ellis, & Mansmann, 2010, p. 7).

Explanatory and Exploratory Graphics

One possible way to characterize visualizations is to classify them into explanatory and exploratory graphics:

- **Explanatory Graphics** or Presentation Graphics

The purpose of explanatory graphics is to explain and present the data. The focus is on solving a specific problem, i.e. the Figure 4.2 can be used to show the distribution of the mountaineering reports length. Therefore these graphics are usually static and are drawn to summarize the data to be presented (Chen, Härdle, & Unwin, 2008, p. 4). As Taylor (2014) describes, this type of visualization can be used for the following tasks: answer a question, support a decision, communicate information or increase efficiency.

- **Exploratory Graphics**

The purpose of exploratory graphics is to support the user in looking for new results (Chen et al., 2008, p. 5), to explore the visualization and ask questions along the way (Taylor, 2014). It is possible to find many insights from a single graphic, the interaction with it provides an understanding rather than a specific answer to a problem (Taylor, 2014). Therefore, words, on the graphic and around, should tell the user how to read the design and not what to read in it (Tufte, 1983).

However, it is important to note that a clear distinction between explanatory and exploratory graphics is not always possible.

Preattentive Attributes

Attributes, which are used for rapid extraction of basic visual features (see Figure 3.1), were termed preattentive attributes by *Colin Ware*. We observe them very fast (usually in less than 10 milliseconds) (Meirelles, 2013, p. 21; Taylor, 2014) and in parallel (Meirelles, 2013, p. 21), even before we consciously notice them (Taylor, 2014). They are used to increase the performance of the following tasks: target detection, boundary detection, region tracking, counting and estimation (Meirelles, 2013, p. 22). Therefore, designers can use them to mark relevant information in a way, that they will literally "pop out" (Meirelles, 2013, p. 21).

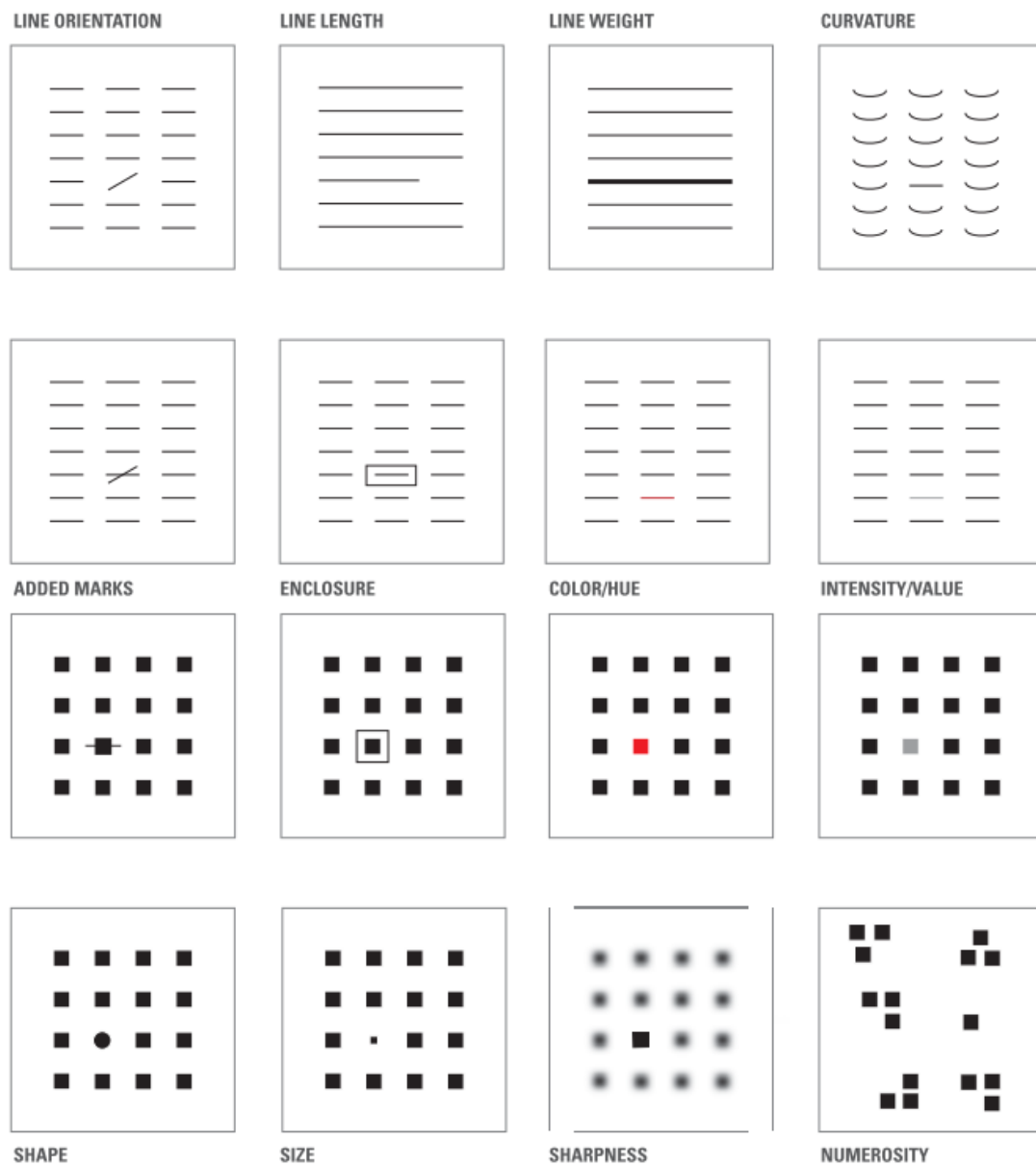


Figure 3.1.: *The preattentive attributes.* [source: Meirelles, 2013, p. 23]

Analytical Patterns

Taylor (2014) describes how we can combine different preattentive attributes to get analytical patterns. The basic analytical patterns found in diagrams are shown in Figure 3.2.













Pattern	Example	Pattern	Example
High, low and in between		Non-intersecting and intersecting	
Going up, going down and remaining flat		Symmetrical and skewed	
Steep and gradual		Wide and narrow	
Steady and fluctuating		Clusters and gaps	
Random and repeating		Tightly and loosely distributed	
Straight and curved		Normal and abnormal	

Figure 3.2.: *The analytical patterns by Taylor (2014).*

Gestalt Laws

In 1912 a group of German psychologists, among them *Christian von Ehrenfels*, *May Wertheimer*, *Kurt Koffka* and *Wolfgang Kohler*, made the first serious attempt to understand the patterns of perception. They established what is today known as the Gestalt school of psychology (Ware, 2012, p. 189). The German word *Gestalt* is translated to *configuration* by Graham (2008) and *pattern* by Ware (2012, p. 189). Gestalt theory influenced not only psychology but also other disciplines (Graham, 2008).

Today Gestalt theory is still valued because it provides a clear description of many basic perceptual phenomena (Ware, 2012, p. 189). It also describes the way we detect patterns and how we integrate individual units into a consistent perception (Meirelles, 2013, p. 22). Therefore we can use them to highlight the important parts and downplay the other parts of a visualization (Taylor, 2014) and to guide the attention in visual displays in order to help reason through the data (Keim et al., 2010, p. 112).

The Gestalt laws are often summarized as '*The whole is more than the sum of its parts*' (Keim et al., 2010, p. 112).

- **Proximity**

Proximity is the simplest and most powerful way to bring out the relationship between different data entities (Ware, 2012, p. 190). This means that we tend to group visual elements that are close together into perceptual units (Meirelles, 2013, p. 19; Ware, 2012, p. 189).

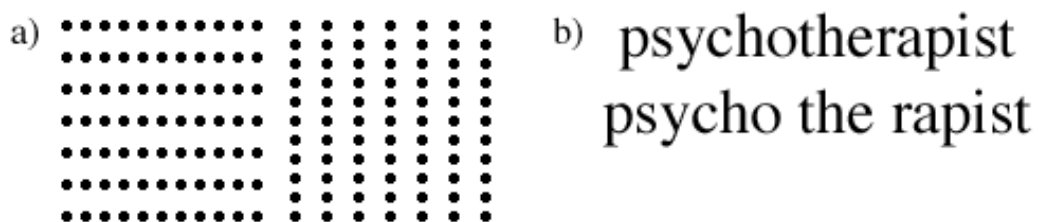


Figure 3.3.: *Examples of the influence of whitespace on our perception. [source: Meirelles (2013, p. 19) and Graham (2008)]*

Figure 3.3 shows two examples of the influence of whitespace. The only difference between the two graphics in (a) is that one is rotated 90 degrees. Otherwise, they are identical, but the space between the dots make us group the dots differently. We perceive rows in the first one and columns in the sec-

ond (Meirelles, 2013, p. 19). In (b) we see how the spacing can dramatically change meaning (Graham, 2008). In this case the graphic consists only of words, but the same applies to visualizations.

- **Similarity**

Similarity is used to group similar objects together with non-local characteristics, i.e. color, shape and texture (Meirelles, 2013, p. 51; Taylor, 2014; Ware, 2012, p. 190). This means, that we can group objects together that are separated in space (Graham, 2008).

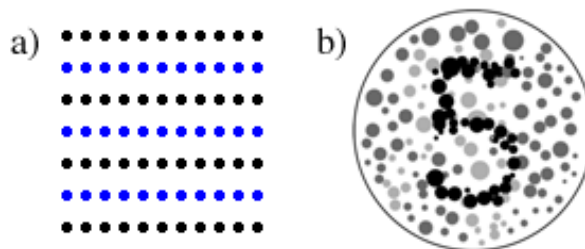


Figure 3.4.: *Examples of the influence of similarity on our perception. [source: Meirelles (2013, p. 51) and Graham (2008)]*

Figure 3.4 shows two examples of the influence of similarity. If we compare the graphic (a) of Figure 3.4 and Figure 3.3, we can now not only see rows, but also two types of rows: black and blue ones. Graphic (b) of Figure 3.4 is influenced by the color vision test. The test is used to determine if someone is colorblind. This is possible because we group objects with the same color together and can therefore see the number 5 in this example.

- **Closure**

Closure describes the tendency to see bounded visual elements as wholes and therefore as a single object (Meirelles, 2013, p.33; Ware, 2012, p.194). Moreover, the tendency exists to close forms to make them look more stable. We tend to fill in gaps with a familiar line, tone or pattern to complete the form if information is missing (Graham, 2008).

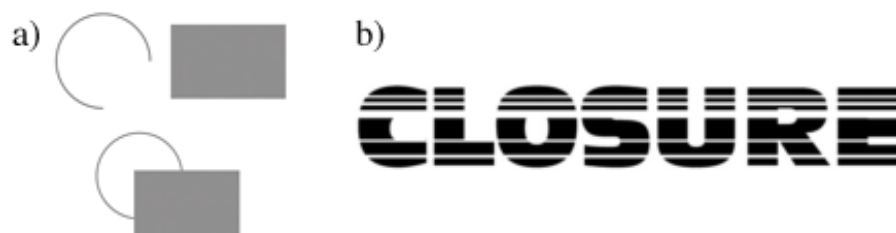


Figure 3.5.: *Examples of the influence of closure on our perception. [source: Ware (2012, p. 195) and Graham (2008)]*

Wherever we see a closed contour, there is a very strong perceptual tendency

to divide regions of space into *inside* and *outside*. This is extremely important to segment the screen into window-based interfaces (Ware, 2012, pp. 195-196).

Figure 3.5 shows two examples of the influence of missing information. We interpret the two graphics in (a) differently although they consists of exactly the same objects. At the top, we see an open circle (not a line) and a rectangle. At the bottom, we see a closed circle which lies behind the rectangle. In graphic (b) we can read the logotype "closure", although it just consists of single horizontal lines.

- **Continuity**

Continuity occurs when our eye follows along a line or sequence of shapes, even if they cross over other shapes (Graham, 2008). It means also that it is more likely to construct visual entities out of visual objects that are smooth and continuous, rather than ones that contain abrupt changes in direction (cf. Figure 3.6) (Ware, 2012, p. 191). It is also affected by the law of proximity, which means that entities closer together are more likely to be perceived as related and therefore continuous (Graham, 2008).



Figure 3.6.: *Examples of the influence of continuity on our perception. [source: Meirelles (2013, p. 58) and Ware (2012, p. 192)]*

Figure 3.6 shows two examples of the influence of continuity. In graphic (a) we can see that this principle can be applied to the problem of drawing networks. It is easier to perceive smooth continuous lines than lines with abrupt changes in direction (Ware, 2012, p. 191; Meirelles, 2013, p. 58). In graphic (b) we can see that the sub-graphic (top) could be perceived as a curved line and a rectangle (bottom-left) or as a line with two abrupt changes of direction and a rectangle with a missing piece (bottom-right). It is more likely that we perceive the left graphic with a curved line and the rectangle.

- **Symmetry**

Symmetry provides a powerful organizing principle (Taylor, 2014; Ware, 2012, p. 192). In Figure 3.7 we see in the left graphic the black object as a cross,

in the right graphic we "move" the black object away and see an alternative object. In the left graphic we suggest the cross, because we assume symmetry.

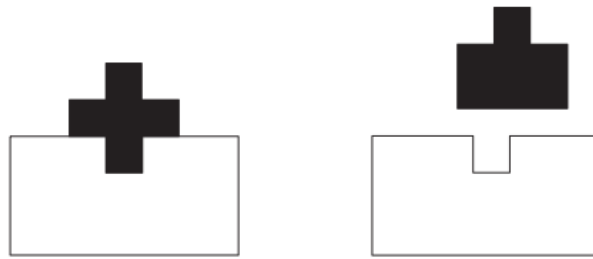


Figure 3.7.: *Example for the influence of symmetry on our perception. [source: Ware (2012, p. 193)]*

• Figure & Ground

A *figure* is an object that is perceived as being in the foreground. Therefore, everything that lies behind the figure is the *ground* (Ware, 2012, p. 196). Detecting the boundaries and making this distinction is the fundamental perceptual act of identifying objects (Ware, 2012, p. 196; Meirelles, 2013, p. 126).

This text itself is an example for the gestalt law: the text is the figure and is visible because of the contrast to the page which is the ground (Graham, 2008).



Figure 3.8.: *Examples of the influence of figure & ground on our perception. [source: Meirelles (2013, p. 126)]*

Figure 3.8 shows two examples of the usage of figure & ground. In graphic (a) we see that scale has an influence on how we perceive objects: the small shape will be viewed as the figure (Meirelles, 2013, p. 126). In graphic (b) we see the classic Rubin's Vase figure: we can see two white faces, noses to noses, or a grey vase in the center (Taylor, 2014; Ware, 2012, p. 197; Meirelles, 2013, p. 126). Therefore, we can say that we have an ambiguity, and it is not possible to see both versions simultaneously (Meirelles, 2013, p. 126).

- **Connection**

Connection is a powerful way to express relationship between different graphical objects by lines (Ware, 2012, p. 191; Taylor, 2014).

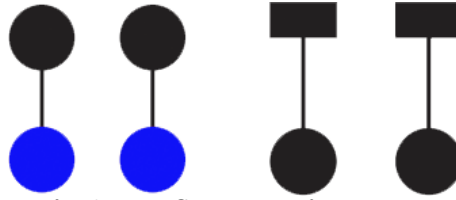


Figure 3.9.: *Examples of the influence of connection on our perception.*
[source: Ware (2012, p. 191)]

Figure 3.9 shows that connectedness is more powerful in grouping than color (left) and shape (right). It is also more powerful than proximity and size (Ware, 2012, p. 191).

Diagrammatic

The concept of diagrammatic is introduced by Krämer (2013). It is a theoretical background which describes and defines the characteristics of diagrams. According to Krämer (2013), parameters which belong to the grammar of diagrammatic are the following:

- **Planarity**

Planarity enables the viewer to gain an overview of the data. Through the bird's eye view, the complexity of the data is reduced to a 2-dimensional plain, an abstract representation. The real surface changes to a virtual plane, which gives the possibility and potential for inscriptions.

- **Graphism**

Graphism is the interaction of points, lines and areas. Points and lines are the root of the drawing and the script. By using them as connection or separation we assign them a meaning.

- **Relation**

Relation is an aspect of diagrams. It is important *how* the diagrams show relations, not the fact *that* they show relations. The diagram is a medium of topographical arrangements, in which spatial and non-spatial, arbitrary and regular correlations are made visible and thus also manageable.

- **Phenomenological Embodiment**

Phenomenological embodiment additionally gives the relation between two points a direction. This direction depends on the viewer and his embodiment. The order of the diagrams are correlated to the conditions of the viewer.

- **Syntheses of Picture and Text**

Syntheses of picture and text is more than the text inside the diagram, it is the text around the diagram. Schematic drawings and the written words are associated.

- **Usefulness**

Usefulness is the main characteristic of diagrams. By making the diagram, we gain something that, without diagram, would be difficult or impossible to have.

- **Different Function Value of Lines**

Different function value of lines means that two lines in the same diagram do not need to have the same function value. For example, one line could have the function value "time intersection", and another line could have the function value "correlation".

- **Translatability**

Translatability means that the diagrams are a picture which can be translated and transformed into each other without losing information. It is not limited to transforming one graphically form into an other. In fact we can overcome the ontological difference with the translatability of diagrams.

- **Generation of Knowledge**

Generation of knowledge through the diagram is bound on three conditions. (i) schematism: A predefined construction rule can be repeated as the order of action steps in the visualization for an unlimited number of times. (ii) surplus: The spatial configuration brings insights that are not already included in the construction rule. (iii) aspect change: The new insight is caused by a change in the function value of an element.

Visual Information Seeking Mantra

The *Visual Information Seeking Mantra* by Shneiderman (1996) is:

"Overview first, zoom and filter, then details-on-demand"

It is a guide to visually explore data and describes how the visualizations should present data on a screen (Keim et al., 2010, p. 11). According to Shneiderman (1996), it consists of the following tasks:

- **Overview**

It is important to give the user an overview of the entire collection.

- **Zoom**

Give the user the possibility to zoom in on times of interest. Because of that, we need navigation tools to pan or scroll through the data collection.

- **Filter**

Removing unwanted items allow the user to focus on their interests. This can be done with sliders, buttons or other control widgets, and should be coupled with rapid display update.

- **Details-on-demand**

The user should easily be able to browse the details about the filtered group or individual items.

Keim et al. (2010, p. 11) extend Shneiderman's visualization mantra in the context of visual analytics to *"Analyse first, show the important, zoom/filter, analyze further, details-on-demand"*. This indicates that it is necessary to analyze the data according to its value of interest.

4. Corpora

I decided to use two different corpora to illustrate how my visualization can be used and which features it provides.

4.1. Mountaineering Reports

The first corpus I used consists of mountaineering reports. Those reports are extracted from the "Text + Berg" corpus (Bubenhofer et al., 2015), which contains yearbooks of the Swiss Alpine Club (SAC) from 1864 till 2015. Thanks to *Patricia Scheurer*, these reports were already selected manually. She identified 2'146 mountaineering reports in the period from 1864 to 2013 in the "Text + Berg" release 149 (*R149*).

Data Gathering

I decided to use the newer, revised release 151 (*R151*). Therefore, I first had to identify the articles from the old release 149 in the new release 151. I could not use the article ID's to match them, because in one-third of the yearbooks, the number of articles changed between the two releases. As shown in example (1) and (2), we can find an article in release 149 which starts with the same words as in release 151, but it incorrectly continues with a second article (blue) in release 149 (compare Figure 4.1.b). This is the reason for the change in the article number between the two releases.

- (1) Aus dem Avers Von A. Wäber [...] nächsten Tagen nach Hause . *R149*
IL Freie Fahrten . Von August Stadler [...]
- (2) Aus dem Avers Von A. Wäber [...] nächsten Tagen nach Hause . *R151*

This means, that I needed to decide for each article in release 151 to which of the categories (*mountaineering reports* or *other articles*) it belongs, using the older release 149 as the reference. The release 149 contains observations for which category membership is known. Furthermore it is known that the original text of each yearbook has not changed, except for some revisions (see below).

As mentioned before, the number of articles in one-third of the yearbooks has changed. This means that at least in those yearbooks also the article boundaries have changed. However, I could not assume that in the other yearbooks, with the same number of articles, the article boundaries are still the same. Therefore I decided to verify in a range of ± 3 articles¹.

	<i>plain text</i>	<i>cleaned text</i>	<i>Levenshtein</i>	<i>boundaries</i>
m. reports	1'071	1'345	2'044	2'133
labeled	7'830	10'326	12'915	13'091
unknown	5'518	3'022	433	257

Table 4.1.: *Number of articles I was able to label (and how many of them are mountaineering reports) with different approaches. Each approach is based on the previous one.*

As shown in Table 4.1, I successfully labeled 7'830 articles (of which 1'071 articles are mountaineering reports) with a plain text approach. If I looked at the unknown articles, I could see, that not only the article boundaries changed, but also the tokenization. For example:

(3) [...] „ 12 n r > Gletscher und Eiszeit statt Gletscher-Eiszeit . [...] *R149*

(4) [...] „ 12 n r > Gletscher und Eiszeit statt Gletscher-Eiszeit . [...] *R151*

To overcome the tokenization problem, I "cleaned" the text from all whitespaces. With this approach I ignored the tokenization. For the purpose of finding the same article, this approach was sufficient. Table 4.1 shows that with this approach I

¹There is no sense in checking every article, because the position in the original yearbook did not change. Therefore an article that was in the beginning in one release will not be the end of the other release.

had an improvement of one-third in labeling articles. From these 10'326 labeled articles, 1'345 articles were labeled positive. This was approximately half of all mountaineering reports. There were a total of 3'022 articles left to be labeled.

In addition to the boundary problem mentioned above, Bubenhofer et al. (2015) also reviewed the yearbooks from 1864 till 1899 with a crowd-sourcing project to find and resolve optical character recognition (OCR) errors. Therefore, I also had to deal with differences between articles on the character level. As we can see in example (5) and (6) the OCR error **IL** has been corrected to **II**.

(5) [...] Bern , im Januar 1864 . A. R. **IL** Fahrten im Clubgebiet . [...] *R149*

(6) [...] Bern , im Januar 1864 . A. R. **II** . Fahrten im Clubgebiet . [...] *R151*

In order to match articles with removed OCR errors I used the Levenshtein distance. This distance calculates the minimal number of operations which are needed to transform *string 1* into *string 2*. In my case transforming the new article in the old article. I set the maximum operations to 5% of the text length. Thereby I could first check if the length of the old article was in the range of the new articles length $\pm 5\%$. If this was the case, I calculated the Levenshtein distance. If the Levenshtein distance was smaller then 5%, I assumed that the article was a match. As shown in Table 4.1, this gave an improvement of 25% to totally 12'915 labeled articles.

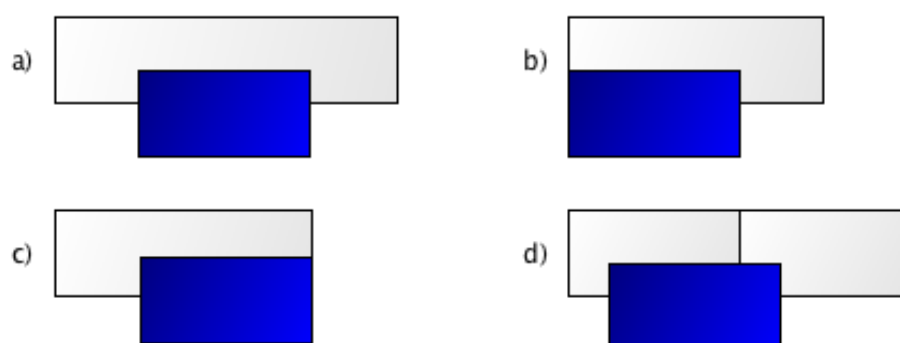


Figure 4.1.: *Different versions of the boundary problem (white = R149; blue = R151).*

Finally, I tried to solve the boundary problem. Figure 4.1 shows the four different types of the boundary problem, comparing the boundaries of the article of the release 149 (white) with the release 151 (blue). The blue article could completely be nested

within the white article (a). The next two scenarios are, that the blue article and the white article share only one boundary (b & c). It is also possible that the text of the blue article is spread over two white articles (d).

To solve the problem a), b) and c) from Figure 4.1, I took the first 100 characters of the article in release 151 and searched for the first occurrence from the left side in the release 149 article². In addition, to find the end of the release 151 article, I searched for the last 100 characters of the article from the right side in the release 149 article. If I could find both, I calculated the Levenshtein distance only for this specific part of the release 151 article. The drawback of this approach was, that the first 100 and last 100 characters had to be a full match (compare example (7) and (8)). With this approach I could label additional 176 articles (see Table 4.1).

(7) 3 . Die Clariden . ” Von E. Frey-Gessner . [...] *R149*

(8) 3 . Die Clariden . Von E. Frey-Gessner . [...] *R151*

In addition, I manually verified the yearbook from 1864, the 20 articles with the highest Levenshtein distance and the 20 articles, where the article numbers are different in the two releases. I could not find an error in those article pairs (*R149* and *R151*). In other words, I assumed that my matching algorithm did not make any mismatches. Therefore I concluded that my algorithm labeled 98% of the articles and I found approximately 99% of the mountaineering reports (I missed 13 articles).

Statistic

I succeeded in labeling 2'133 mountaineering reports from 1864 to 2013. Seven of those reports were not in German, but for my analysis I had to have the reports to be in the same language.

Figure 4.2 shows the distribution of the mountaineering report length. The range of each bar is 200 tokens (starting with 0). The median (C) is the bucket between 2'400 and 2'600 tokens. The first quartile (B) is the bucket between 1'600 and 1'800 tokens. The third quartile (D) is the bucket between 4'400 and 4'600 tokens. The lower whisker (A) is 0. The upper whisker (E) is the bucket between 8'600 and 8'800

²I only applied it on the articles that were not labeled so far.

tokens. The outliers (F) are above 12'800 tokens.

The shortest article with 3 tokens³ is from 1925 with the title (and text) ”*Die Walliser Fiescherhörner*” by *Oskar Hug* and has in addition seven pictures. The longest article with 30'225 tokens is from 1914 with the title ”*Aus dem Valsertal im Bündner Oberland*” by *W. Derichsweiler* and consists of several parts.

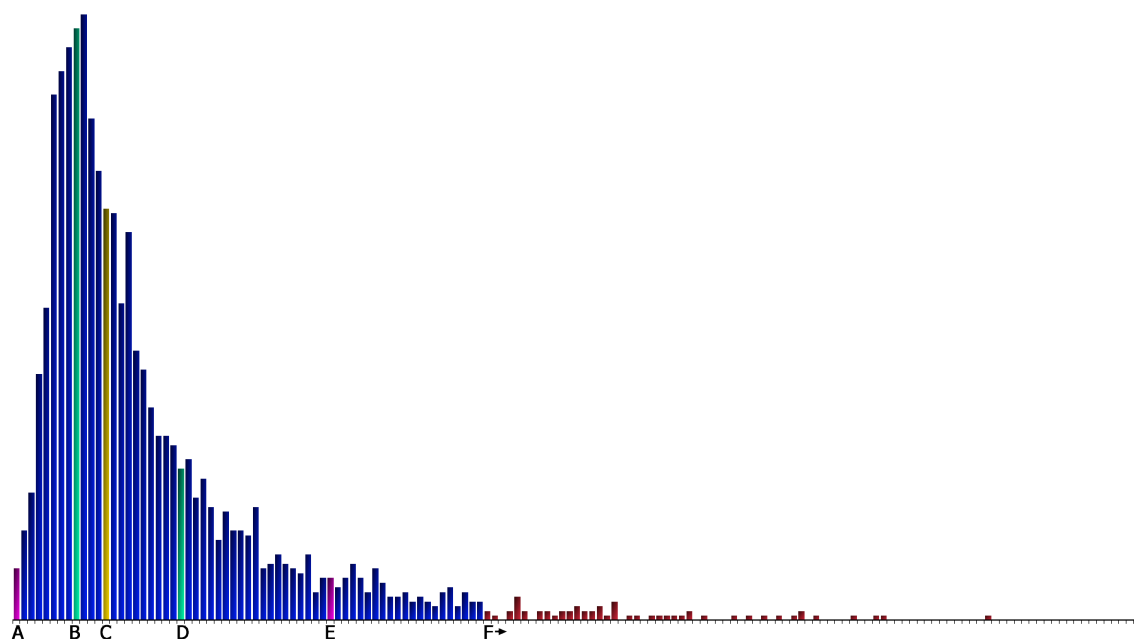


Figure 4.2.: *Distribution of the length of all mountaineering reports.*

To achieve a homogenous corpus, I used only those reports with a token length between 1'000 and 2'800. Therefore, I had a corpus with 976 mountaineering reports, consisting of a total of 1'803'128 tokens and 88'097 different lemmata.

Insight

The reports are about the experience of the author during a mountain ascent. The authors write not only about the used route but also about the surrounding area. Certain authors also write about interesting facts of the surrounding area.

All authors have an empurpled writing style with many adjectives when they write about the surrounding area. *Ruedi Horber*, 1990, describes the boulders as *wild* and the pine forest as *gloomy*. *Marianne Hodel-Gisin*, 1970, writes about the *ascendent* and *ice-infiltrated* north wall.

³It was labeled from *Patrica Scheurer* as a mountaineering report.

”[...] die Szenerie ist eindrucksvoll: wilde Felsblöcke noch und noch, düsterer Tannenwald, dazwischen das Geplätscher des Bergbachs.”

(Engl. ”[...] the scenery is impressive: wild boulders everywhere, gloomy pine forest, in between the murmur of the mountain stream.”)

’Skitour auf dem Damen-Montblanc’ von Ruedi Horber, 1990

” Wächtenüberspannt zieht sich der Grat zum Breitlauihorn, getragen von der aufsteigenden, eisdurchsetzten Nordwand.”

(Engl. ”Cornices-spanned runs the crest of Breitlauihorn, carried by the ascending, ice-infiltrated north wall.”)

’Dem Alltag entrückt’ von Marianne Hodel-Gisin, 1970

The beginning of the stories varies from author to author. For example, *Hermann Etter*, 1945, writes about his colleagues and their characteristics. In contrast *Gottlieb Studer*, 1864, describes where to find the area in the atlas.

” Arthur Spöhel, mein bewährter Seilgefährte auf manch schwieriger Fahrt, Ruedi Dietrich, ein Bergkamerad, wie man ihn seiner stets guten Laune und seines Optimismus wegen besser nicht wünschen könnte, und ich.”

(Engl. ”Arthur Spöhel, my proven rope companion on many difficult journeys, Ruedi Dietrich, one cannot wish someone for better than him because of his goofy humor and optimism, and myself.”)

’Doldenhorn-Südgrat’ von Hermann Etter, 1945

” Wer das Blatt XVIII des eidgen. Atlas zur Hand nimmt und sich auf demselben die mächtige, mit ewigem Firn überdeckte und reich umgletscherte Kette der Fletschhörner [...]”

(Engl. ”Who takes the sheet XVIII of the federal Atlas at hand and sees on it the mighty, covered with eternal firn and rich glaciated chain of Fletschhörner [...]”)

’Das Mattwaldhorn und sein Panorama’ von Gottlieb Studer, 1864

In the middle part they write about the route and the surrounding area. The reports end in various ways. For example, *Ruedi Horber*, 1990, describes his plan for the next year. In contrast *Hermann Etter*, 1945, reports only the climbing to the top of the mountain. In addition the article ’*Die Eisfrau - Begegnung in den vergletscherten Alpen*’ from *Gisula Tschärner*, 2000, ends with phone numbers and email addresses.

”Damit hätten wir uns für nächstes Jahr bereits ein neues Tourenziel ausgewählt: [...]”

(Engl. ”Thus, we would have already selected a new tour destination for next year: [...]”)

’Skitour auf dem Damen-Montblanc’ von Ruedi Horber, 1990

”Im Eiltempo treten wir über die Normalroute den Rückweg an, gelöst und beglückt vom neuen Bergerleben.”

(Engl. ”In a hurry we take the normal route back, relaxed and delighted by the new mountain experience.”)

’Doldenhorn-Südgrat’ von Hermann Etter, 1945

4.2. Coming-Out Stories

The second corpus I used consists of coming-out stories. Those stories were extracted from the website *dbna - das Magazin für schwule Jungs!* (1997 - 2016).

To collect the coming-out stories from the website, I used a web-crawler. Afterwards, I used a *natural language processing* pipeline consisting of tokenization, lemmatization and part-of-speech tagging.

I gathered 447 stories with a total of 295’306 tokens. They consists of 13’853 different tokens and 10’587 different lemmata. The median token length per story is 493 tokens. The smallest number of tokens in a story is 70 tokens. The highest number is 2’658 tokens.

The age of the authors is between 13 and 30. There is no data on the publication date. The stories include some explicit dates such as the newest story from *Joshua*, 18, where he writes about the 15.9.2014 when he finally told everybody via Facebook about his sexuality. The oldest story is from *Matthias*, 16, he writes that he finally became 14. This could mean that this story is 2 years old and the given author’s age is calculated with his birthday. Otherwise, *Manuel*, 15, writes that he was around 14, and that his coming-out was in July 2002. This implies that the author’s age is the age from the moment he wrote the article.

”Vor etwa 2-3 Monaten (da war ich noch 13) [...] Vor knapp einer Woche bin ich dann endlich 14 geworden [...]”

(Engl. "About 2-3 months ago (when I was still 13) [...] Just about a week ago, I finally became 14 [...]")

Matthias, 16

"Genauer gesagt im Juli '02. [...] ich wollte es mir aber bis ich 14 bin nicht eingestehen."

(Engl. "More specifically in July '02 [...] but I did not wanted to admit it to myself until I was 14.")

Manuel, 15

Insight

The stories are about the experience of the authors with their own coming-out. It is about how they found out that they were gay, about their struggle to tell their friends and family and how relieved they were after their coming-out.

The beginning of the stories is about their self-awareness: since when they knew it and how they found out. For example, *Julius*, 16, writes that he knew since he was a little boy and played marrying his teddy. In contrast *Kevin*, 18, thought it was normal and each boy would make some experiments with other boys.

"Ich wusste es schon in der Grundschule [...] Ich liebte es, mich als Braut zu verkleiden und meinen grossen Teddy zu heiraten."

(Engl. "I knew it already in primary school [...] I loved to dress up as a bride and marry my big teddy bear.")

Julius, 16

"Naja, also angefangen hat alles eigentlich mit 13! [...] Ich dachte es sei normal, da ich es auch immer wieder in Jugendzeitschriften las, dass Jungs ein wenig mit anderen Jungs experimentieren."

(Engl. "Well, everything has actually started when I was 13! [...] I thought it was normal, also because I read it all over in youth magazines that guys experiment with other guys.")

Kevin, 18

The middle part of the coming-out stories follows their struggle with themselves about telling friends and family. Furthermore, this text passage also deals with the reactions they got after they told someone about it. For instance, *Kevin*, 18, (not

the same as above) writes about how he could not find enough courage to tell a friend. In addition *Manuel*, 15, writes that his best friend was honored to be the first to know.

”Den Mut es einer Freundin oder einem Kumpel zu erzählen, hatte ich jedoch trotzdem nicht.”

(Engl. ”However, the courage to tell a friend or a mate, I had not.”)

Kevin, 18

”Sie fühlte sich total geehrt, dass sie es als Erste erfahren durfte.”

(Engl. ”She felt absolutely honored to be the first who learnt about it.”)

Manuel, 15

The last sections are about how relieved they are after their coming-out. For example, *Joshua*, 18, writes that he is happy with his coming-out and he never got a negative comment. Furthermore *Manuel*, 15, aims directly at the reader.

”Als Fazit kann ich sagen, dass ich sehr froh bin solche tollen Freunde und Verwandte zu haben, ich habe nach wie vor keine negativen Kommentare oder sonst was abbekommen und bin echt dankbar, dass das so ist.”

(Engl. ”In conclusion I can say that I am very happy to have such great friends and relatives, till now I have not received any negative comment or anything else and I am really grateful that this is so.”)

Joshua, 18

”Mein Tipp: Sagt es als ersten einem Girl, zu dem ihr eine sehr gute Freundschaft habt.”

(Engl. ”My tip: Tell it first a girl, with whom you share a very good friendship.”)

Manuel, 15

5. NarrViz

I decided to implement my visualization of narrative patterns as a web application. It allows the user a simple way to use the visualization for their own datasets. It also was important to me, that the visualization is an exploratory graphic. Therefore, I implemented different filter functions and the possibility to change the settings. Furthermore, I wanted to ensure that my visualization is as flexible as possible regarding the input data.

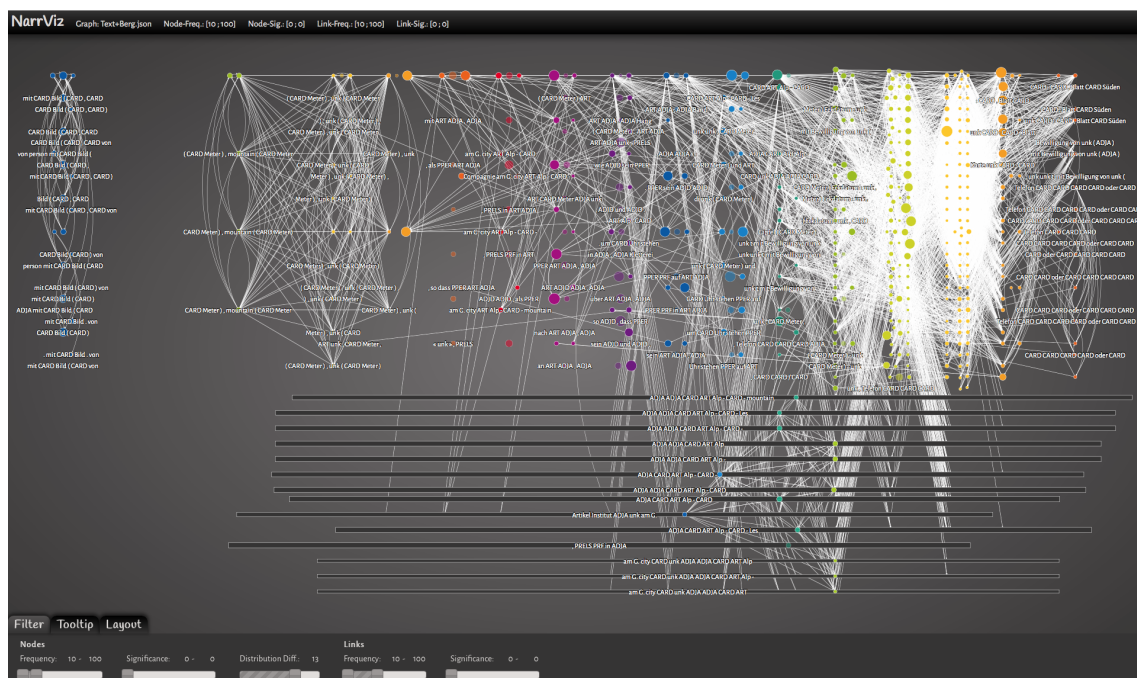


Figure 5.1.: Screenshot of the visualization with the mountaineering reports displayed.

Figure 5.1 shows my visualization with the mountaineering reports loaded. We can see that the first section is cut off from the other sections. Moreover, we can see that the next three sections have many connections between each other, but only a few to the following. In the last sections there are many connections, which implies a huge variation. The rectangle visualization of the high distribution nodes shows,

for example, that their median lies at the beginning of the high variation sections. Further explanation on how to read the visualization is provided in section 5.2.

5.1. Data Structure

There are many different data formats to store and access the data. Extensible Markup Language (XML) and JavaScript Object Notation (JSON) are both easy to read and write for humans. The most important advantage of JSON over XML is the easy handling, because JSON itself is valid JavaScript.

For the visualization I need at least the following information: the (complex-) n -grams, the position and the relationship between the (complex-) n -grams (cf. chapter 2). This leads to a directed graph (cf. section 3.1). I store this information in a JSON file (for the schema see Listing A.1). The n -grams (including their positions) are stored in the array *nodes*. The relationships between the n -grams are stored in the array *links*. For each n -gram a unique identifier (*id*), the section in which the n -gram occurs (*section*) and the n -gram (*ngram*) is saved. In addition, it is possible to save the frequency (*freq*), the significance level (*sig*) and all sections in which the n -gram occurs (*sections*). If the n -grams are complex, the token- n -grams can be stored in *examples*, if available with their frequency. For each relationship the unique identifier of the *source* and *target* node is saved. In addition, it is possible to save the frequency (*freq*) and the significance level (*sig*) of it.

5.2. Layout

The narrative patterns consist of n -gram positions and the relationships between them. The layout should help the user to *identify* and *explore* the narrative patterns. The data leads into a representation with a directed graph. For the hierarchical structure I do not use the typical links between the nodes, but the position of the nodes within the articles.

As mentioned in section 2.3 the seriality is not only given by the relationship between the n -grams, but also by the position in the article. To support this direction in the data, I use the *phenomenological embodiment* as described by Krämer (2013). Because the positions in the text are similar to a timeline, the position scale

is set on the x-axis. Furthermore, I use *proximity* and *similarity* of the Gestalt laws (see section 3.2) to group the different n -grams into sections. The space between different sections is based on *proximity*. Furthermore, space can also be seen as an element to receive *closure* as described by the following Gestalt law: the space between the different sections is larger than between the nodes of a section, therefore we group the sections together and can identify different sections. The space can also be referred to the *enclosure* of the preattentive attributes: the space around the sections separates them from each other. In Figure 5.2 there are three sections with the colors purple, blue and green. Node C is seen as part of the section *blue* despite the fact that the distance to the other nodes of this section is as large as the distance between the other sections. The reason for this is the *color* of the preattentive attributes and *connection* of the Gestalt law.

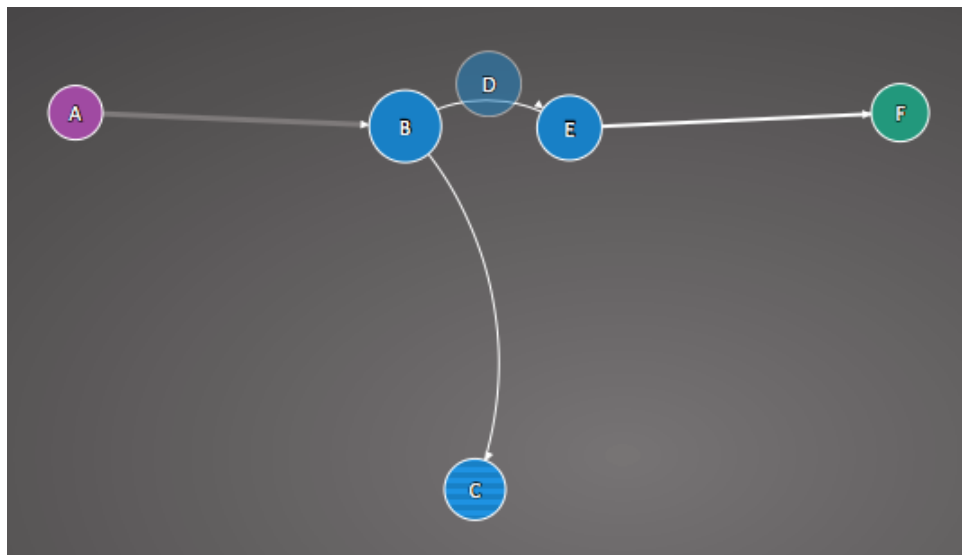


Figure 5.2.: Example graph with 6 nodes, 4 links and 3 sections.

The position on the y-axis is ordered by the degree δ of the node. Inside each section there is again an order in x-axis position:

- if the node is in the first section, in-degree $\delta^+ = 0$ (sink) and out-degree $\delta^- > 0$, it is positioned on the left side of the section (node A in Figure 5.2).
- if in-degree $\delta^+ > 0$ and out-degree $\delta^- = 0$ (source), it is positioned on the right side of the section (node C & F in Figure 5.2).
- if in-degree δ^+ of cross-section-links is bigger than 0, the node is positioned on the left side of the section (node B in Figure 5.2).
- if out-degree δ^- of the cross-section-links is bigger than 0, the node is posi-

tioned on the right side of the section (node E in Figure 5.2).

- all other nodes are positioned in the middle (node D in Figure 5.2).

Each inner-section (left - middle - right) is based on *proximity* of the Gestalt law and on the *phenomenological embodiment* by Krämer (2013). Furthermore, there is a *symmetry* between the sections: each section is constructed in the same way, except the first section.

The size of the nodes is based on their frequency. For instance, in Figure 5.2 node B has the highest frequency and node A the smallest. This corresponds to the preattentive attribute *size*. It can also be seen as *normal and abnormal* and/or *high, low and in between* of the analytical patterns. Furthermore, if a node has a degree δ equal to 0 it is shown with a reduced opacity (node D in Figure 5.2). This is part of the preattentive attribute *intensity/value*.

To support the interaction and exploration, it is possible to drag and drop the nodes. To do so, it is necessary to *unchain* the node first within the context menu. To put it back in its original position there is the *chain* function. Unchained nodes are visualized with a striped pattern (node C in Figure 5.2). This is based on the analytical pattern *normal and abnormal, color/hue* and *added marks* of preattentive attributes.

The links in the graph represent the collocation profile. Therefore they create an affiliation between the nodes which is part of the diagrammatic *graphism*. Similar to the size of the nodes, also the width of the links depends on the frequency. In Figure 5.2 the link from node A to B has a higher frequency than the link between the node B and E . This is based on the *line weight* of the preattentive attributes. The links between two nodes of the same section are *curved* (link between B and C in Figure 5.2), the links over different sections are *straight* (link between A and B in Figure 5.2). This is part of the preattentive attribute *curvature*, the analytical patterns *straight and curved* and part of the diagrammatic *Different Function Value of Lines and Relation*. There are no abrupt changes in the direction, which is part of the Gestalt law *continuity*. As the Gestalt law *connection* describes the links between two nodes is more powerful than the proximity and the size of the nodes. Furthermore, the intensity of the lines depends on the significance level. For example, in Figure 5.2 the link between node A and B has a lower significance level than the others. This is based on the preattentive attribute *intensity/value*.

The specific information of a node is revealed, once the cursor hovers over it - or when *node information* is selected in the context menu. In the graph, the neighborhood of the node is visualized in an other color. In Figure 5.3 the dark red nodes are the neighborhood of node *E*. In this case, the *similarity* is the connection with the node. In addition, the most probable path is colorized and shown in the tooltip. In Figure 5.3 the most probable path is colorized in red. Both follow the preattentive attribute *color/hue* and the Gestalt law of *similarity*.

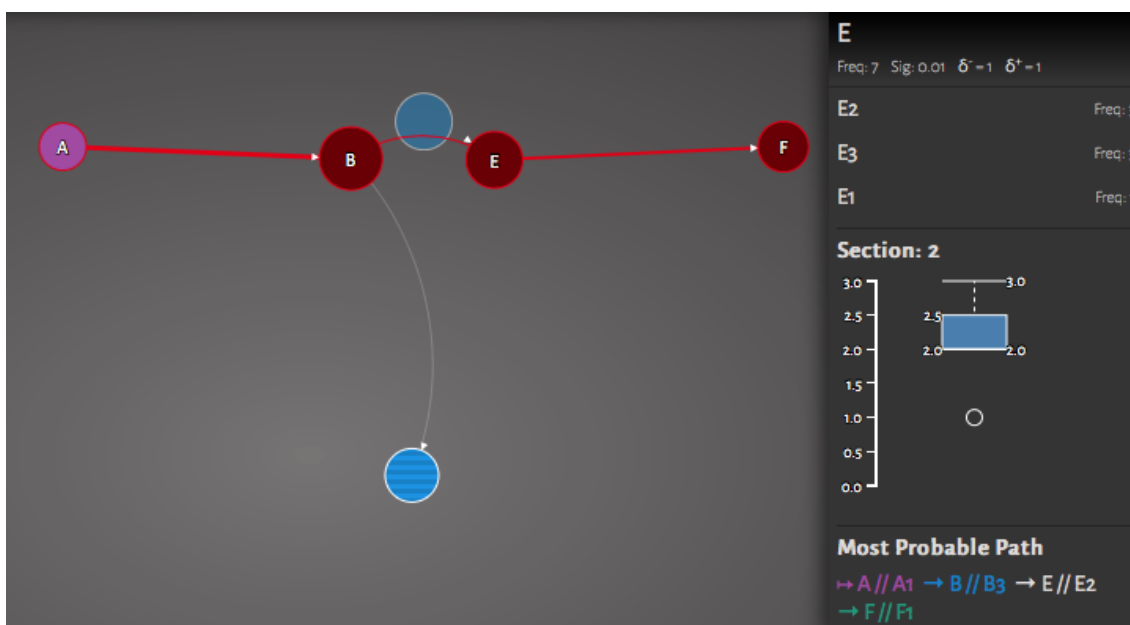


Figure 5.3.: Example hover of the node *E* from Figure 5.2

The tooltip (and the node information box) contains the information about the frequency, significance level, the degree (split in in-degree and out-degree), examples (if available), sections (if available) and the most probable path. If the item *sections* are filled, they are visualized as a box plot (see Figure 5.3). It consists of the information about the median, the first and third quartile, the whisker and the outliers. A box plot immediately gives us an idea about the range in which the data is located and how it is distributed over this range. Furthermore it shows if the *n*-grams are stable at a position or not. In Figure 5.4 we can see that the left box plot belongs to a stable *n*-gram in the first sections. On the other hand the right side of the figure belongs to the unstable *n*-gram ‘, dass PPER ADJD sein’ (Engl. ‘, that PPER be ADJD’). This *n*-gram is unstable because it can occur in the beginning, where the authors write about their self-awareness, in the middle, where they tell their friends and family and also in the end, where they are ”happy”

about telling that they are gay. If the item *examples* is available, they are displayed and sorted by frequency (if given). In the tooltip there is only space for a few, but in the node information box all examples are shown (cf. Figure 5.3).

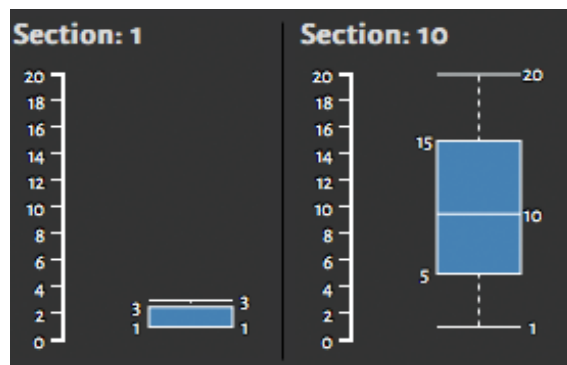


Figure 5.4.: Box plot of a stable (left) and an unstable (right) n -gram.

There is a filter function called *Distribution Diff.*, which will transform unstable n -grams from the *circle*-only representation into a *rectangle* version below the circle graph (compare preattentive attribute *shape*). I defined the distribution difference as the difference between the third and first quartile. Figure 5.5 shows the *coming-out* corpus with a maximal distribution difference set to 13.5. The new representation of the n -grams immediately generates knowledge (compare diagrammatic by Krämer (2013)) about the distribution without looking into the tooltips. The height of the rectangle is the same as the circle. This leads to the perception as a single object (compare Gestalt law of *closure* and *continuity*) which again leads to the preattentive attribute *enclosure* and the diagrammatic of *graphism*. The rectangles are sorted on the y-axis by their distribution difference and the section of the circle. Therefore it has a "direction" on the y-axis which can be seen as a part of the diagrammatic *phenomenological embodiment*. It is possible to follow the links although they are interrupted by other rectangles, which is described in the Gestalt laws *proximity* and *continuity*. This applies both to curved and straight lines. The color difference between the circle and the rectangle is part of the preattentive attribute *color*. This leads to the perception as *figure & ground* of the Gestalt law where the circle is the figure and the rectangle corresponds to the ground. Depending on the selected circle position in the preprocessing, the rectangle representation corresponds to the box plot (compare Gestalt law of *similarity*): the left side mirrors the first quartile, the right side mirrors the third quartile and the circle remains on the position it has in the original circle view - in the Figure 5.5 it is the median. This allows the user also to conclude about the skewness of the n -gram. Furthermore, the *similarity* is also

given by the position of the circle on the x-axis which does not change between the different representations and therefore corresponds to the original position in the circle graph. The *similarity* between the rectangles to each other allows the user to easily perceive differences in the distribution.

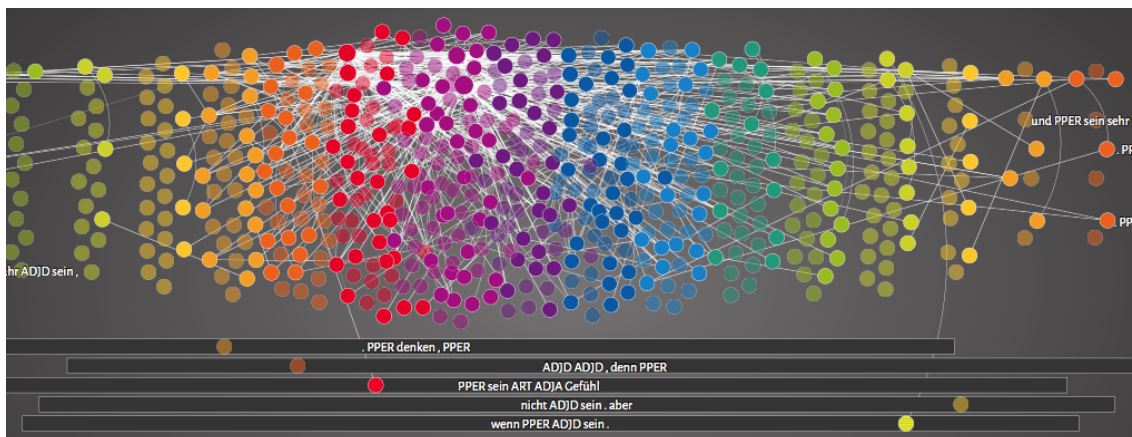


Figure 5.5.: *Example of the rectangle representation of the coming-out corpus.*

If the node information box is displayed the most probable path and neighborhood are fixed. This means the coloring does not disappear if the hover ends. If the mouse is over another node, the tooltip will open, the node will be highlighted, but not the neighborhood and the most probable path of this node. If the information box is closed, the coloring goes back to the default behavior.

In the context menu it is also possible to select *graph information*. In the graph information box the selected preferences are shown, including the number of remaining nodes and links (cf. section 5.4). The statistics of the given dataset are displayed. It consists of the graph statistics with the number of nodes, the range of the n -gram length, the range of the frequency and the range of the significance level. The same applies for the links. There is also a statistics about the degree, that consists of minimal, maximal for in-degree δ^+ and out-degree δ^- and average degree δ of the graph.

5.3. Filter Functions

The number of the nodes and links is often huge, for example, the mountaineering reports corpus consist of 8'181 nodes and 861'768 links. To help the user to explore the narrative patterns, I provide filter functions. The filter function is part of the *Visual Information Seeking Mantra* from Shneiderman (1996). Under the tab *Filter* (see Figure 5.6) the range of frequency and significance level of the nodes and the links can be selected. As mentioned above, there is an option to set the maximal distribution difference.

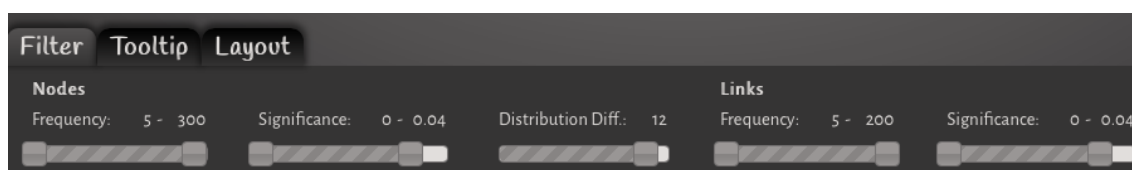


Figure 5.6.: *Filter options*

5.4. Customization

In addition to the filter functions, there are many possibilities to change the layout and style of the graph (see Figure 5.7). The radius size (in pixels) can be defined by the user. It applies on the selected frequency range of the nodes. If there is no frequency given, then all radii are set to the minimal radius size. The same applies for the stroke width for the links: it only applies to the selected frequency range of the links. Furthermore it is possible to change the color of each section and the color for the rectangle view.

The user has the possibility to change the scaling of the axis. The number of nodes above each other on the y-axis can be changed with *y-Axis Scale*. With *Section Distance* the distance between adjacent sections can be set.

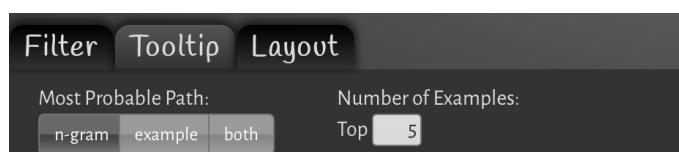
The used cost function for the calculation of the most probable path can be selected with *Most Probable Path*. The cost function can be either the significance level, the frequency or simply be based on the distance (number of nodes). The algorithm is based on Dijkstra¹.

¹JavaScript implementation: Andrew Hayward (<https://github.com/andrewhayward/dijkstra>)

Figure 5.7.: *Layout options*

The label type of the nodes can be either the n -gram or the most frequent example. Furthermore, it can be selected when the label should be shown. With the option *smart* the labels are only visible, if they do not overlap other labels and nodes. If the tooltip or the node information box is open, only the labels of this node and the highlighted part is visible. There is a similar function for the links called *Link Visibility*. If *hover* is selected, only paths and neighborhood are shown on hover. If *always* is selected all the links always are shown.

In addition, the user can set the color and the size of the font. For the links it is possible to change the default color, the coloring of the neighborhood and of the most probable path.

Figure 5.8.: *Tooltip options*

The tooltip can also be customized (see Figure 5.8). It is possible to choose the number of examples of the node that should be displayed in it. Furthermore, it is possible to decide in which way the most probable path is displayed: as the n -grams, as the most frequent example or both together.

5.5. Exemplary Analysis

To give a better understanding of the visualization, I would like to illustrate in this section how the visualization works with the corpora described in chapter 4.

At the first glimpse of the visualization it is possible to draw conclusions over the corpus. In addition the user can detect whether the preprocessing needs optimization. For example, Figure 5.9 shows the coming-out corpus with five sections. Particularly noticeable is the large number of curved links in section three and four. Those occur most frequently at the boundaries of the sections. This implies a narrative pattern within the section which is difficult to recognize because the corresponding nodes are in the same section. In contrast, many connections over several sections may mean that a smaller number of sections would be enough and result in a more intuitive visualization.

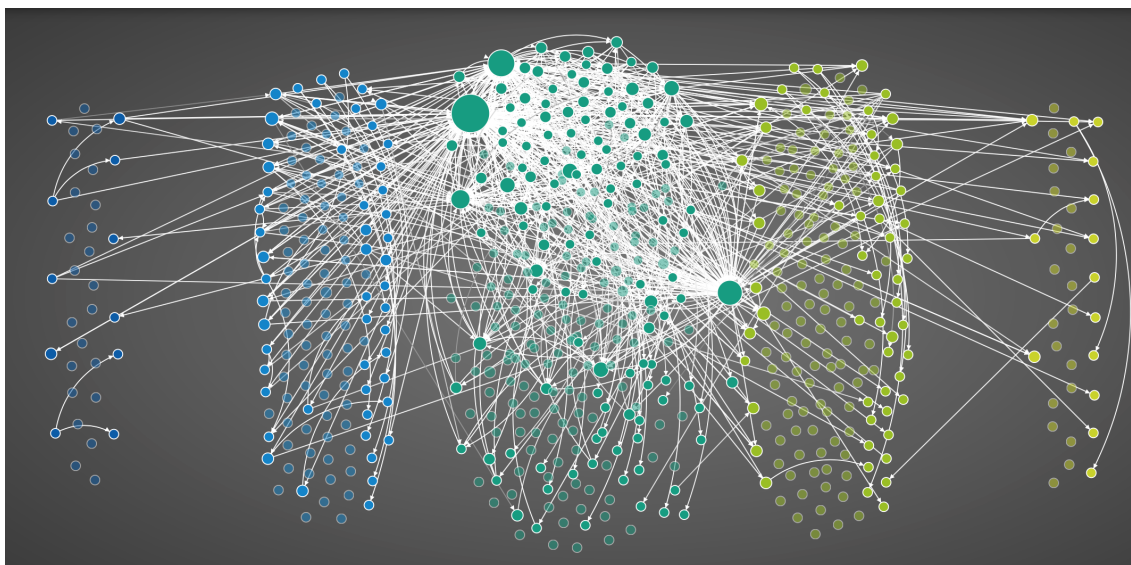


Figure 5.9.: *Coming-out corpus with five sections.*

If the number of sections is optimized, it is possible to draw some preliminary conclusions about narrative patterns. For example, as shown in Figure 5.13, the coming-out corpus consists of stories which start and end in similar ways (only a few different complex- n -grams). However, in the middle sections there are many nodes and links which imply a big variation in the stories. In contrast the mountaineering reports corpus shown in Figure 5.10 has a separated section without cross-section links at the beginning followed by two empty sections without any n -grams. This implies that in those two sections the variation is too large. In the later sections

the number of nodes and links are increasing. Section 16 and 17 have the highest number. In contrast to the coming-out corpus this implies an "open" ending of the reports, but the middle follows a certain standard.

Mountaineering Reports

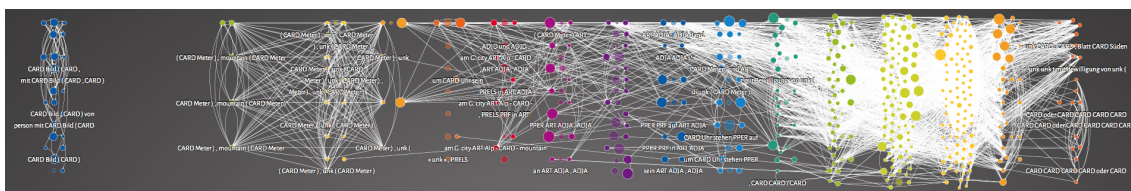


Figure 5.10.: *Overview of the mountaineering reports corpus.*

In this corpus the interesting patterns are revealed once we select the less frequent nodes and links. The high frequent nodes consist mainly of phone numbers and alike. In the mountaineering reports most of the patterns are rather strongly scattered throughout the sections. This can be seen in the box plots. A stable pattern, for example, consists of the title followed by the author and the specification of the number of images in the article. In Figure 5.11 we can see a segment of this pattern, for instance, in the middle is the complex- n -gram 'von person mit CARD Bild (' (Engl. 'from person with CARD image (') which refers to the author and the number of images.

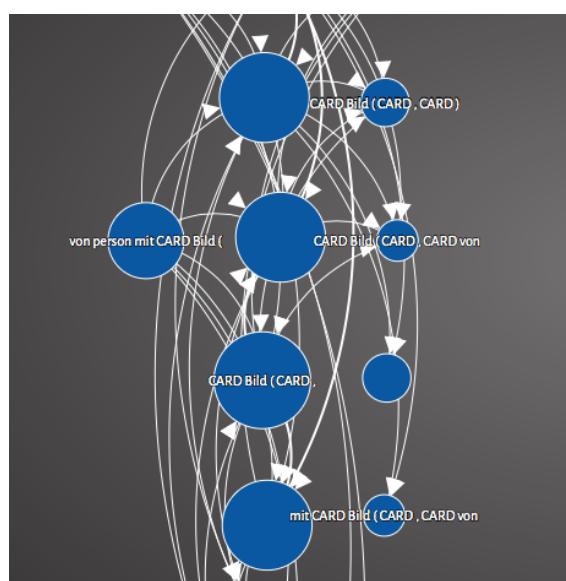


Figure 5.11.: *Segment of the first section of the mountaineering reports.*

The narrative itself is constructed by adjective sequences and mountain names

(and the hight). If we look at the nodes in other than the first section, we can see that they are all unstable. The reason could be that the authors write about the same - their experience with the mountains - but how they write about it is different.

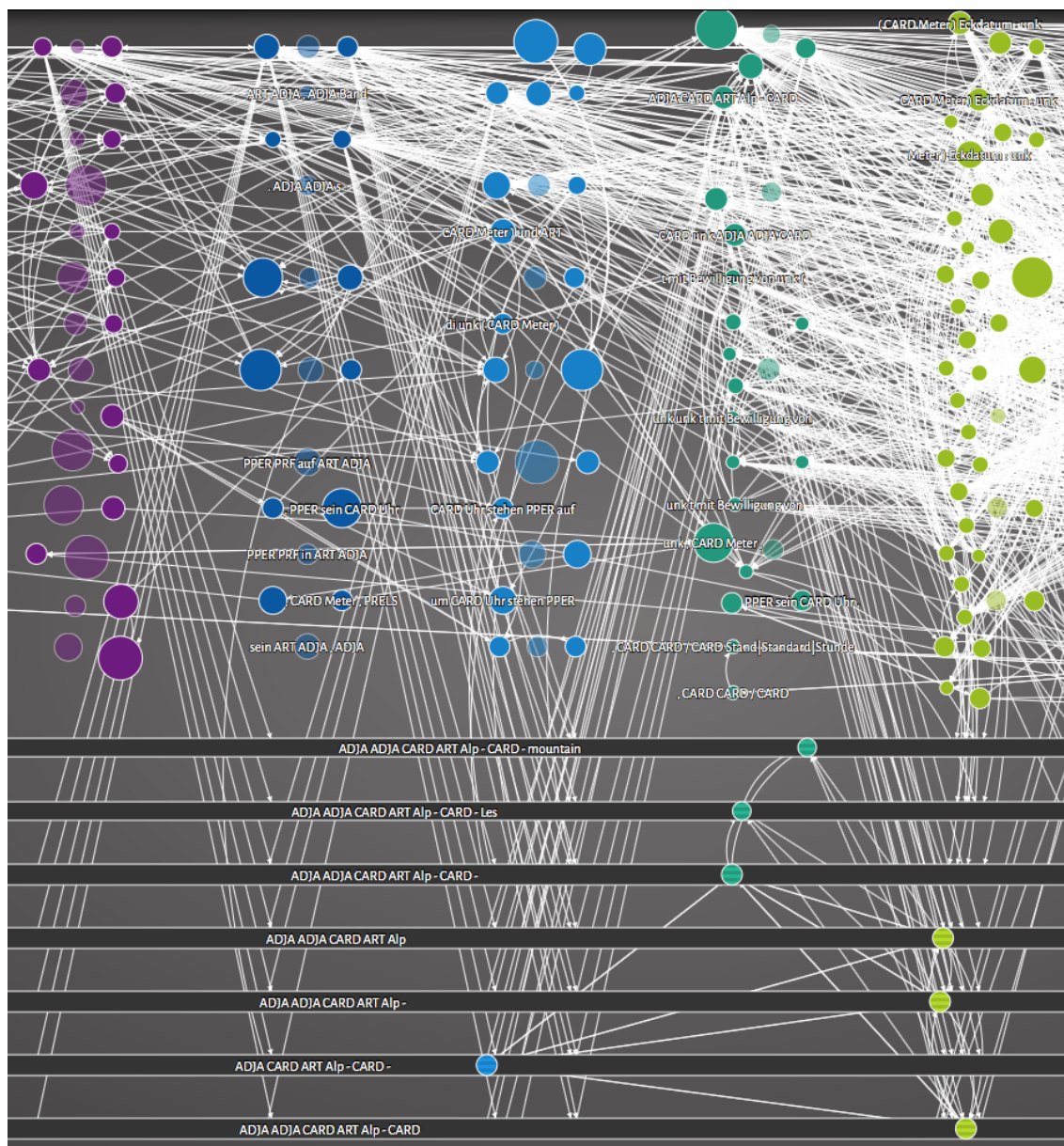


Figure 5.12.: Segment of the sections between 11 and 15 with maximal distribution difference = 14.

Figure 5.12 shows the influence of the filter function *Distribution Diff*. With the new perspective on the data it is possible to identify a pattern which was hidden in the data until now. First it is conspicuous that all the transformed complex- n -grams consists of 'ART Alp' (Engl. 'ART alp') in the middle and that everyone of them has a link to and from the complex- n -gram 'CARD ART Alp - CARD' (Engl.

'CARD ART alp - CARD'). Moreover, there are patterns with other n -grams in two not-directly adjacent sections, where the transformed n -gram occupies a placeholder position. In the visualization this can be seen as a jagged pattern.

Coming-Out Stories

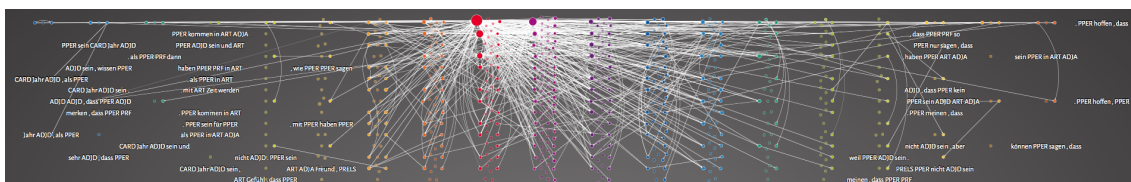


Figure 5.13.: Overview of the coming-out corpus.

In this corpus additional narrative patterns will be shown. Figure 5.14 shows the complete first section. We can see that the authors start their story with their age. This corresponds with my insight in the corpus (cf. section 4.2). For example, the node 'als PPER CARD Jahr ADJA' (Engl. 'when PPER CARD year ADJA'). Furthermore, this node has links into other sections.

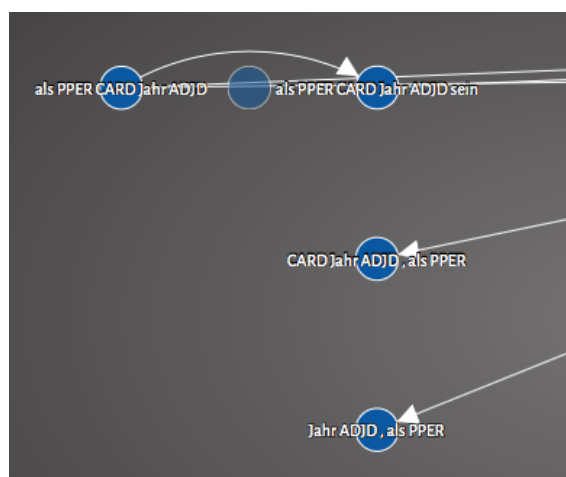


Figure 5.14.: First section of the coming-out stories.

For the next sections I changed the label type to *examples* to achieve a better understanding of the nodes.

The nodes with the highest frequency are placed in the middle sections (cf. Figure 5.13). These nodes contain their confession that they are gay. For example, the most frequent node is ', dass ich schwul bin' (Engl. ', that I am gay'). Furthermore, we can see in the middle of the red section in Figure 5.15 a node which refers to

the question "whether": ', ob ich schwul sei' (Engl. ', whether I am gay'). This corresponds again to my insight in this corpus.

In the middle they write about their coming-out to friends and family. Therefore it makes sense to find here the node ', dass ich schwul bin' (Engl. ', that I am gay') which refers to the coming-out and telling the truth about their sexuality.

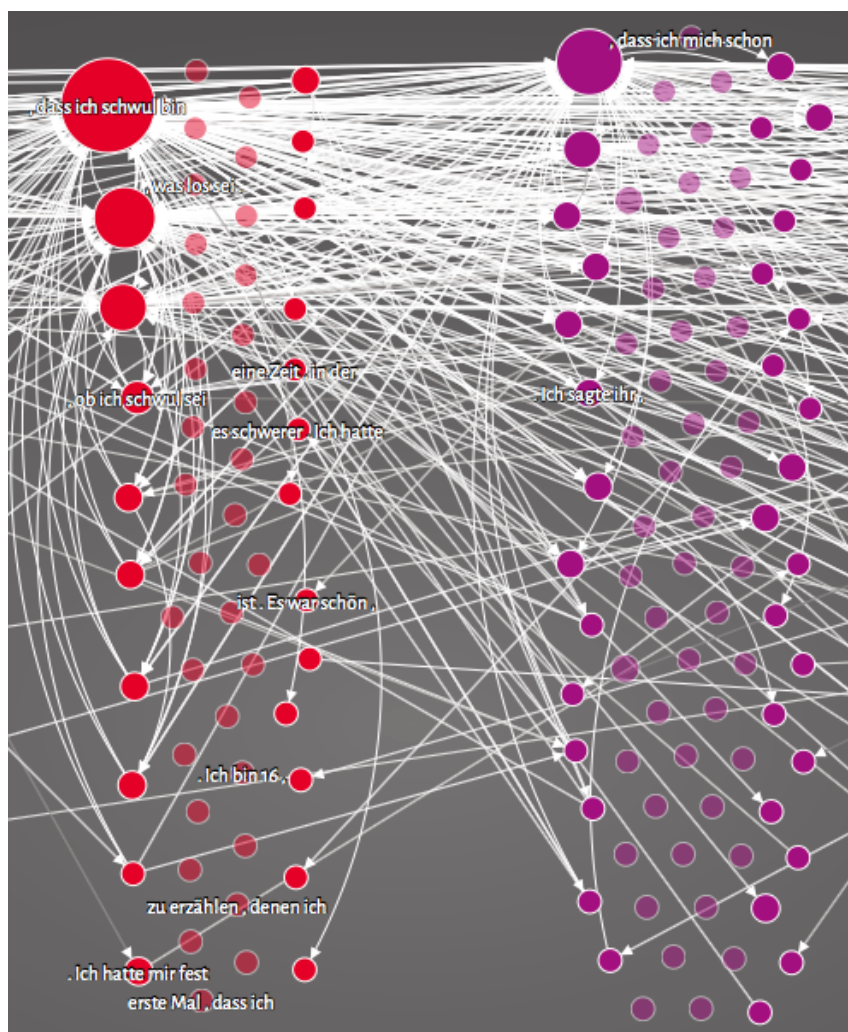


Figure 5.15.: Middle sections of the coming-out stories (label type: show example).

Figure 5.16 summarizes the last two sections. The penultimate section covers their feelings after coming-out. For example, there is the node 'PPER sein so ADJA , dass PPER' (Engl. 'PPER be so ADJA , that'). The adjective stands for *glücklich* (Engl. *happy*), *froh* (Engl. *glad*) and *stolz* (Engl. *proud*) in the original stories. Which again corresponds to my observation. For example, *Joshua*, 18, (p. 28) wrote "dass ich sehr froh bin" (Engl. "that I am very happy") which expresses his feelings.

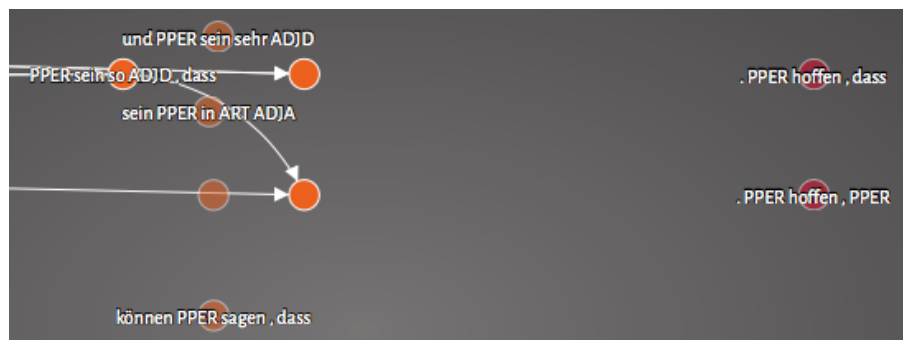


Figure 5.16.: *End sections of the coming-out stories.*

Figure 5.17 shows a segment of the rectangle representation with the complex- n -gram ‘, dass PPER ADJA sein ,’ (Engl. ‘, that PPER be ADJA ,’ highlighted. It displays a similar pattern as described in the mountaineering reports. It is possible to conclude that the statement about “being gay” of the author appears in various sections and in the context of other significant n -grams.

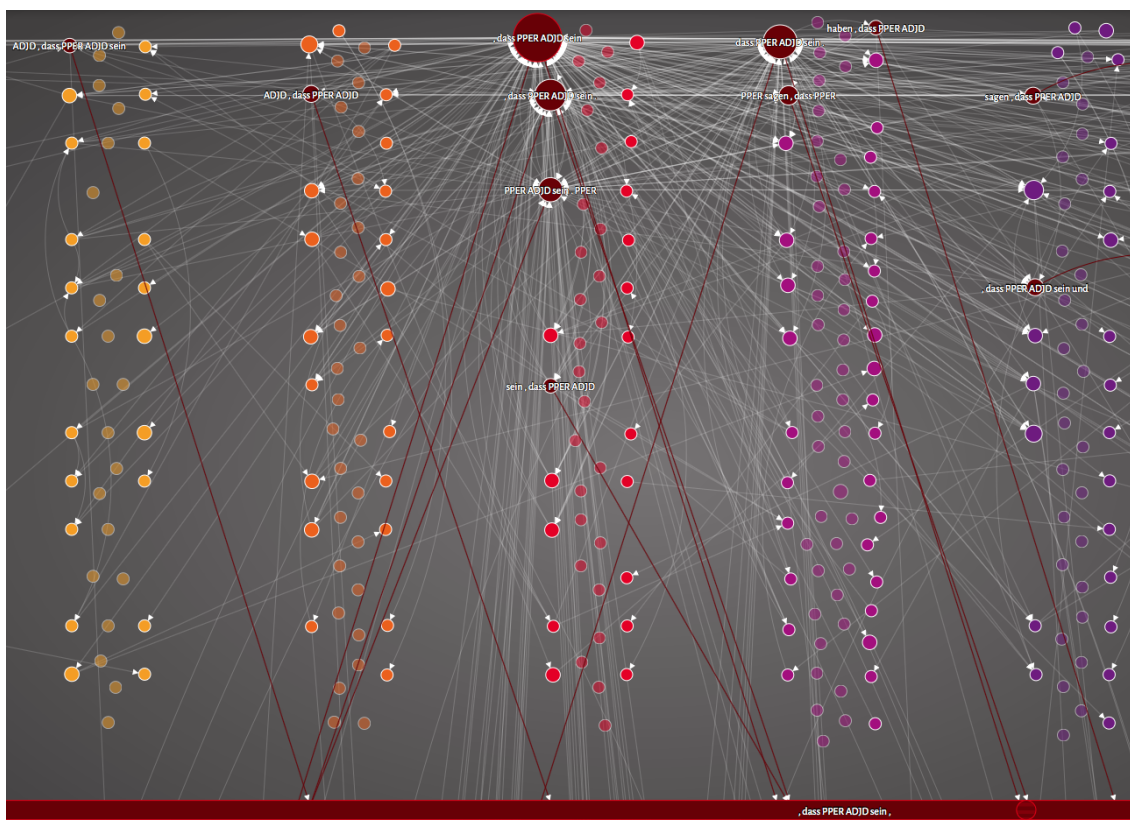


Figure 5.17.: *Rectangle representation of the node ‘, dass PPER ADJA sein ,’ (Engl. ‘, that PPER be ADJA ,’).*

6. Conclusion

In my thesis I developed a visualization for narrative patterns. It is based on the research of Bubenhofer et al. (2013). Therefore I used their best visualization approach as a template. Furthermore I applied their suggestion to use complex- n -grams with a combination of lemmata and POS tags instead of simple n -grams. This means that the tokens will either be reduced to their lemma or to their POS tag. For example, in the coming-out stories the token sequence '*19 Jahre alt*' (Engl. '*19 years old*') will be reduced to '*CARD Jahr ADJA*' (Engl. '*CARD year ADJA*'). This is a significant distinction because now the two token sequences '*19 Jahre alt*' and '*18 Jahre alt*' which have a similar meaning (the age of the person), will be reduced to one complex- n -gram. The drawback of the reduction are complex- n -grams like '*sein PPER in ART ADJA*' (Engl. '*be PPER in ART ADJA*') which consists of '*bin in einer schlimmen*' (Engl. '*am in a bad*') and '*war in der 7.*' (Engl. '*was in the 7th*') which do not have a similar meaning (cf. section 2.1).

A possible improvement would be to use synsets and/or sentiment analysis. For example, in the coming-out stories the authors write about the reactions they experienced during their coming-out. The reactions are described as '*keine negativen Reaktionen*' (Engl. '*no negative reactions*'), '*keine schlechten Reaktionen*' (Engl. '*no bad reactions*'), '*keine einzige negative Reaktionen*' (Engl. '*not a single negative reaction*'), '*nie negative Reaktionen*' (Engl. '*never negative reactions*'), '*positive Reaktionen*' (Engl. '*positive reactions*'), '*nur positive Reaktionen*' (Engl. '*only positive reactions*'), '*gute Reaktionen*' (Engl. '*good reactions*') and so on. All those statements refer to the same meaning of positive reactions, but with lemmata and POS tags it is impossible to group those statements together in a single complex- n -gram.

Moreover the calculation of the typical n -grams is based on the comparison between the corpus and a reference corpus. For both corpora I used the "Text + Berg" corpus as the reference. A different reference corpus can lead to different typical

n -grams as a result. Furthermore it is also possible to calculate the typical n -grams for each section on its own. In this scenario the other sections could be used as reference corpus. This leads to the specific n -grams for each section, compared to the rest of the texts. This would result in losing n -grams with a large distribution over the text. To overcome this problem it would be conceivable to calculate the n -grams over all sections, but for each possible section to make a copy of this n -gram. This is possible because the identification is not the n -gram itself but an integer (cf. section 5.1). It depends on the research question and the data which approach should be used.

The underlying calculation of n -grams is irrelevant for the visualization. The visualization receives a n -gram which will be used as labels of the node. In addition it can handle examples which will be displayed in the tooltip (cf. section 5.2). I used the complex- n -grams as label and the examples consist of n -grams based on tokens. The user decides in the preprocessing step of the corpus which kind of n -grams he wants to use in the visualization. It would also be possible to work with complex- n -grams as labels but to use lemma- n -grams instead of token- n -grams as examples. This allows the user to select the kind of n -grams that are relevant to his research question.

Not only the n -gram itself is important, but also the position of the n -gram in the text. In my approach I sliced every text into 20 sections based on the number of tokens. The section for each n -gram is based on the position of the center. For the complex- n -gram I used the median of all sections of the examples as the position in the visualization. It is again something that is calculated in the preprocessing of the corpus. The user is free to define boundaries for the sections and number of sections. For example, it would be possible to use sentences as boundaries instead of tokens. The visualization needs at least one section to position the n -gram in the graph. It is also possible to forward all sections in which the n -gram occurs. In the visualization they will be displayed as a box plot in the tooltip (cf. section 5.2).

In order to detect narrative sequences, the relationships between the n -grams are needed. Based on the approach of Bubenhofer et al. (2013) I used the analysis of collocations. They explored on the left and right side of the context for each n -gram in a 30% range of the text length. In contrast I only considered on the right side of the context. This is again a preprocessing step of the data which can be defined by the user. Therefore the visualization displays the relationship between two n -grams

as a directed link (cf. section 5.2).

Those settings lead to connections between sections which are not neighbors. For example, in the coming-out stories the n -gram 'als PPER CARD Jahr ADJA' (Engl. 'when PPER CARD year ADJA') in section 1 has a connection to the n -gram ', dass PPER ADJA sein' (Engl. ', that PPER be ADJA') in section 10 (see Figure 5.14 and 5.15). The reason is an interaction of the large distribution of the n -gram in section 10 and the combination of context length for the collocation analysis and the number of sections. Those two nodes have a similar distribution to those shown in Figure 5.4. The box plot shows how the distribution of the unstable node (right side) consist of examples which are located in section 1. Therefore it is possible to have a collocation between the two nodes. Furthermore in the context range of 30% text length, 6 sections are included (if we dived each text in 20 sections).

In addition there are connection between n -grams which complement each other. For example, 'CARD Bild (CARD , ' (Engl. 'CARD image (CARD , ') has a connection to 'CARD Bild (CARD , CARD' (Engl. 'CARD image (CARD , CARD'). A simple solution is to allow connection only between n -grams with different sections but this just works for stable patterns. A better solution would be to cluster the n -grams based on their similarity. In that case the visualization would present complex- n -grams clusters.

The drawback of the typical n -gram extraction is that we lose "outlier" stories. This is the case because of the chosen type of analysis which calculates the overview with the help of every single story. This is something that happens in the preprocessing step. In the visualization itself it is possible to filter for low frequent n -grams and/or links. It is also possible to filter within the significance level (cf. section 5.3).

The developed visualization fulfills the criteria for an exploratory graphic, since it allows the user to explore the data and supports different versions of questions. On the one hand it has filter functions, on the other hand there is the possibility to change the preprocessing. This allows, for example, to look only into low frequency nodes or in contrast only into high frequency nodes. This also corresponds to the *Visual Information Seeking Mantra* by Shneiderman (1996) which in addition requires the possibility of *zoom* and *details-on-demand*. *Details-on-demand* is implemented with the behavior on mouse over the node (tooltip & highlight). In addition, the diagrammatic by Krämer (2013), Gestalt laws, preattentive attributes and analytical patterns were considered during the process of the layout creation

(cf. chapter 5).

A possible point of criticism on the visualization is, that the user can not get to the original document. It would be possible to use a web-based graphical user interface of the Corpus Workbench (CWB). Since the preprocessing does not have to be done with CWB it would not be an optimal solution. However, the visualization provides enough information to the user such that he can identify the original document at his own.

Finally, it can be said that the visualization supports the user to find narrative patterns. The presented concepts of visualization theory were used during the layout implementation. The described graph theory was applied to the implementation -especially the definition of neighborhood - as well as to the shortest path for each node. With the example corpora I was able to show *how* the visualization can be used to find narrative patterns. Moreover it is possible to make conclusion by using the visualization, such as that many adjectives are used in the mountaineering reports (cf. section 5.5). When using the visualization it is obvious that the complex- n -grams are better suited than simple n -grams, as proposed by Bubenhofer et al. (2013). The added value of my visualization compared to theirs is given in particular by the dynamic filter function and the *details-on-demand* from the *Visual Information Seeking Mantra* by Shneiderman (1996).

Bibliography

- Bubenhofer, N. (2009). *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Walter de Gruyter.
- Bubenhofer, N., Müller, N., & Scharloth, J. (2013). Narrative Muster und Diskursanalyse. Ein datengeleiteter Ansatz. *Zeitschrift für Semiotik*, 419–444.
- Bubenhofer, N., & Scharloth, J. (2011). Korpuspragmatische Analysen alpinistischer Literatur. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 55, 241–259.
- Bubenhofer, N., & Scharloth, J. (2013). Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In *Diskurslinguistik im Spannungsfeld von Deskription und Kritik* (pp. 147–168). Warnke, Ingo H./Meinhof, Ulrike/Reisigl, Martin.
- Bubenhofer, N., Volk, M., Leuenberger, F., & Wüest, D. (Eds.). (2015). *Text+Berg-Korpus (Release 151 v01)*. XML-Format. (Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924, Die Alpen, Les Alpes, Le Alpi 1925-2014, The Alpine Journal 1969-2008)
- Chen, C.-h., Härdle, W., & Unwin, A. (2008). *Handbook of Data Visualization*. Springer-Verlag. doi: 10.1017/CBO9781107415324.004
- dbna - das Magazin für schwule Jungs!* (1997 - 2016). Retrieved 2016-05-15, from <http://www.dbna.de/comingout/erfahrungen/>
- Evert, S. (2010). *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*.
- Graham, L. (2008). Gestalt Theory in Interactive Media Design. *Journal of Humanities & Social Sciences*, 2(1), 1–12.
- Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the Information Age Solving Problems with Visual Analytics*. doi: 10.1017/CBO9781107415324.004
- Krämer, S. (2013). Diagrammatisch. Glossar. Grundbegriffe des Bildes. *Rheinsprung* 11(05), 162–174.
- Meirelles, I. (2013). *Design for Information* (Vol. 5). Rockport Publishers. doi:

10.1017/CBO9781107415324.004

Rada Mihalcea, & Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press, Cambridge. doi: 10.1017/CBO9781107415324.004

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. doi: 10.1109/VL.1996.545307

Taylor, T. (2014). *Principles of Data Visualization - What We See in a Visual*. Retrieved 2016-03-22, from <http://www.fusioncharts.com/whitepapers/downloads/Principles-of-Data-Visualization.pdf>

Tufte, E. (1983). Aesthetics and Technique in Data Graphical Design. In *The Visual Display of Quantitative Information*. (Vol. 2nd ed., pp. 177–190). Graphics press Cheshire, CT.

Voloshin, V. I. (2009). *Introduction to Graph Theory*. Nova Science Publishers, In. doi: 10.1017/CBO9781107415324.004

Ware, C. (2012). *Information Visualization: Perception for Design*. Elsevier. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Wilson, R. J. (1996). *Introduction to graph theory* (4th ed.). Prentice Hall Press. doi: 10.1017/CBO9781107415324.004

A. Appendix

```
1 {
2   "$schema": "http://json-schema.org/draft-04/schema#",
3   "type": "object",
4   "properties": {
5     "nodes": {
6       "type": "array",
7       "items": {
8         "type": "object",
9         "properties": {
10          "id": {
11            "description": "The unique identifier for a node",
12            "type": "integer"
13          },
14          "section": {
15            "type": "integer"
16          },
17          "sections": {
18            "type": "array",
19            "items": {
20              "type": "integer"
21            }
22          },
23          "ngram": {
24            "type": "array",
25            "items": {
26              "type": "string"
27            }
28          },
29          "examples": {
30            "type": "array",
31            "items": {
32              "type": "object",
33              "properties": {
```

```

34         "ngram": {
35             "type": "array",
36             "items": {
37                 "type": "string"
38             }
39         },
40         "freq": {
41             "type": "integer"
42         }
43     },
44     "required": ["ngram"]
45 }
46 },
47 "freq": {
48     "type": "integer"
49 },
50 "sig": {
51     "type": "number"
52 }
53 },
54 "required": ["id", "section", "ngram"]
55 }
56 },
57 "links": {
58     "type": "array",
59     "items": {
60         "type": "object",
61         "properties": {
62             "source": {
63                 "description": "The unique identifier of the
64                     source node",
65                 "type": "integer"
66             },
67             "target": {
68                 "description": "The unique identifier of the
69                     target node",
70                 "type": "integer"
71             },
72             "freq": {
73                 "type": "integer"
74             },
75             "sig": {
76                 "type": "number"
77             }
78         }
79     }
80 }

```

```
75     }
76   },
77   "required": ["source", "target"]
78 }
79 }
80 },
81 "required": ["nodes", "links"]
82 }
```

Listing A.1: *JSON Schema*