TRAViz: A Visualization for Variant Graphs

Stefan Jänicke
Leipzig University, Germany

Annette Geßner
Göttingen Centre for Digital Humanities (GCDH), Germany

Greta Franzini, Melissa Terras and Simon Mahony
UCL Centre for Digital Humanities (UCLDH), UK

Gerik Scheuermann
Leipzig University, Germany

Abstract

This article describes the development and application of an innovative tool, Text Re-use Alignment Visualization (TRAViz), whose aim is to visualize variation between editions of both historical and modern texts. Reading different editions of a text empowers research in literary studies and linguistics, where one can study a text's reception or follow the development of its language over time. One of the purposes of a text edition is to trace or reconstruct a possible archetype or something that might be considered to be an original version of the text in order to better understand its evolution over time. To do so, the textual scholar examines and records the similarities and the differences between a number of exemplars in what is known as a 'critical apparatus'. The result of this variant analysis can be visually represented as a 'Variant Graph', where the relationships between these exemplars can be more easily studied. Variant Graphs can be, in turn, visualized in order to facilitate reading and interaction with the source data. Borrowing from existing digital tools, TRAViz assists the scholar in the collation process by specifically focusing on design and user engagement, concurrently seeking to simplify interaction as a means of encouraging humanists to adopt the tool. The article will describe the needs and rationale behind the creation of TRAViz by exploring existing research, describing its functionality through examples, and by finally discussing how its application can influence future development of this tool in particular and of the field in general.

Correspondence:
Stefan Jänicke, Leipzig University, Leipzig, Germany.
E-mail: stjaenicke@informatik.uni-leipzig.de

1 Introduction

This research bears on Textual Criticism, a field of the humanities whose task it is to study the creation, distribution, and dissemination of textual heritage. The standard publications in this field are critical editions of literary (or non-literary) works. One of the purposes of a text edition is to trace or reconstruct the archetype or original version of the text in order to better understand its evolution over time. To do so, the textual scholar examines and records the similarities and the differences between a number

Digital Scholarship in the Humanities, Vol. 30, Supplement 1, 2015.

doi:10.1093/llc/fqv049 Advance Access published on 9 October 2015

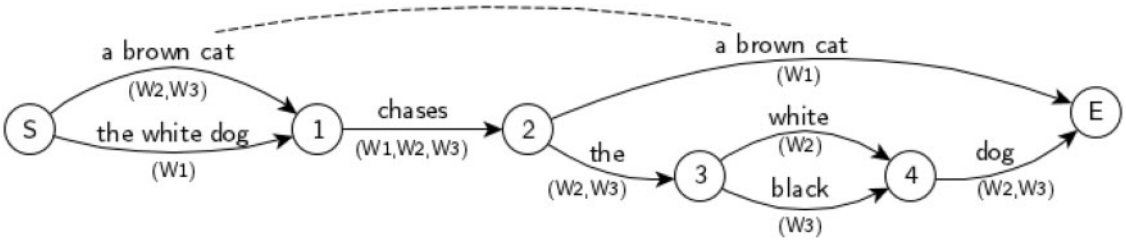

Fig. 1 Example Variant Graph

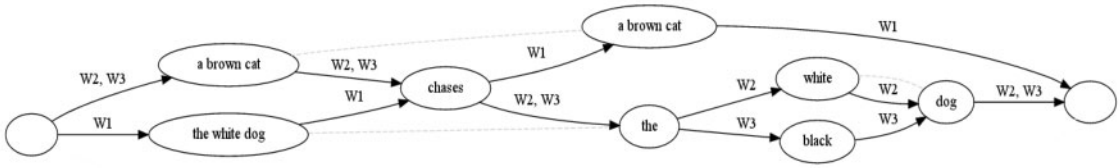

Fig. 2 Example Variant Graph with CollateX

of exemplars, a practice known in the field as collation (Tanselle, 1992). Traditionally, variations are recorded in the edition's critical apparatus, a textually dense area at the bottom of the page mostly, if not only, intelligible to experts (Szpiech, 2014).

In the process of collation, scholars select a number of witnesses they wish to compare, arrange these side by side, manually transcribe each exemplar (assuming no transcription already exists), and annotate variation between these. The more manuscripts one compares, the more complex and laborious the task becomes.

Modern, digital methods are beginning to address this time-consuming, error-prone task in the form of semi-automatic transcription[1] and automatic collation. The development of Text Re-use Alignment Visualization (TRAViz) falls under the latter effort but requires transcriptions in order to operate.

In 2009, Schmidt proposed the Variant Graph to represent multiple versions of a digital text (Schmidt and Colomb, 2009). A Variant Graph is a directed acyclic graph[2] capable of modeling the similarities and differences among various editions of a text. Fig. 1 shows an example of a Variant Graph in the design Schmidt uses for the three example variants: 'the white dog chases a brown cat' (W1), 'a brown cat chases the white dog' (W2), and 'a brown cat chases the black dog' (W3).

The reading direction of the above graph is sinistrodextral (Left-To-Right). All variants start at the S(tart) vertex and end at the E(nd) vertex. The 'vertices' or circles in the graph number and connect subsequent text snippets; each variant appears as a label identifying an arrow or 'edge'; the edition or variant identifiers are displayed in brackets. Additionally, transpositions of word groups ('a brown cat') are highlighted in the form of dashed connections.

While informative, from a graph-drawing standpoint, this Variant Graph is not particularly easy to read: the textual information (text and edition identifiers) borne by the edges puts a strain on comprehension.

In 2011, Dekker introduced CollateX,[3] a web-based collation framework that generates Variant Graphs and facilitates work with electronic editions in the browser (Dekker and Middell, 2011). Unlike Schmidt, Dekker focused on improving the alignment structure between the various text editions. Along with output formats such as XML or JSON, CollateX uses the GraphViz[4] library, which computes non-interactive visualizations of the resultant Variant Graphs. Consequently, the readability of the graph in Fig. 2 is made easier inasmuch, as text and

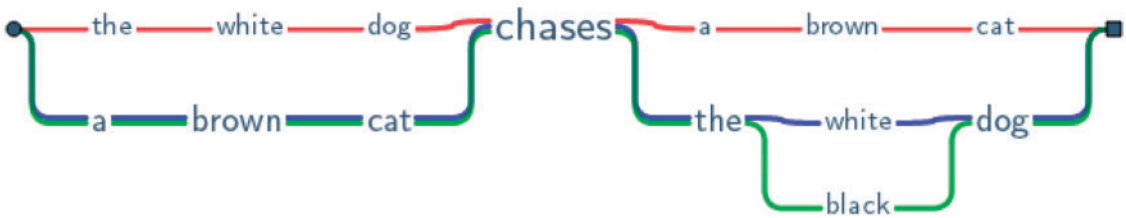

Fig. 3 Example Variant Graph with TRAViz

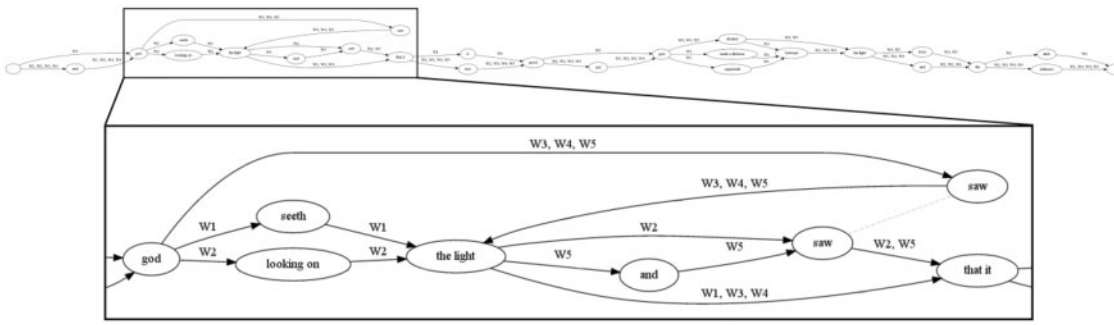

Fig. 4 CollateX: Variant Graph (with zoom on opening segment) illustrating the relationships between five different versions of *Genesis 1:4*

edition information are now split between vertices and edges.

The tool Stemmaweb (Andrews and Macé, 2013a) aims to support analyses of Variant Graphs. It extends the CollateX graph to enable user-driven annotation and modification of the graph structure (e.g. the merging and splitting of vertices). But for all its merits, Stemmaweb's straightforward adoption of the GraphViz visualization affects readability.

TRAViz,[5] a web-based Open Source library, addresses this issue by implementing a set of design rules aimed at styling both vertices and edges, and thus supporting an intuitive reading of the collation. By extending the choice of interaction possibilities, TRAViz enhances support and allows users, for instance, to tweak the visualization to meet specific research questions (Section 3). Fig. 3 is a TRAViz reproduction of the above example:

TRAViz's introduction of color, one per each witness, the use of word-sizing, the linear alignment, and the removal of unnecessary visuals (circular shapes) improve the readability of the Variant Graph. What follows is a description of each of these features with concrete applications to existing text-based projects.

2 Graph Layout Design

The improvements illustrated in Fig. 3 are the result of design rules TRAViz defined following a study of the visualizations generated by CollateX. CollateX was chosen as reference tool not only because it is one of the standard tools in the Digital Humanities, but also because it underpins many web-based extensions, including Stemmaweb.

A CollateX Variant Graph representing five English translations of *Genesis 1:4* (see Fig. 4), reveals a number of issues that hamper the readability of the collation.

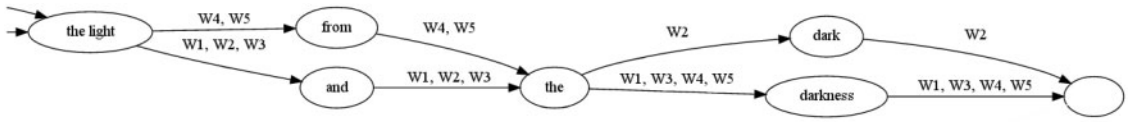

Fig. 5 CollateX: large edge labels in the final part of *Genesis 1:4*

For this reason, we propose five Variant Graph design rules based on related work in information visualization and on guidelines for drawing graphs.

The first rule concerns the 'label size of a vertex'. When looking at the CollateX graph in Fig. 4, it is hard to determine the frequency of a word across all editions. Hence, it is difficult to study the use of synonyms and the occurrence of specific text patterns. One workaround to this problem is to count the edition identifiers listed by the labels of the incoming edges. We can easily exploit this information by displaying vertex labels in varying font sizes, a common practice in tag clouds such as Wordle (Viégas *et al.*, 2009) and a feature in 'Word Tree' (Wattenberg and Viégas, 2008), a visualization similar to TRAViz.[6] In TRAViz, font size reflects the occurrence of individual text fragments, thus contributing to the clear identification of their frequency in the graph.

The second rule 'eliminates backward edges'.[7] Instinct pushes us to draw the edges of a directed graph as arrows. But why do so if we recognize the reading direction of the text? Introducing bi-directional cues makes the graph counterintuitive. In graph theory, the accepted layout for a directed acyclic graph is the layered graph drawing (Sugiyama *et al.*, 1981), where all edges point in the same direction. When we turn the Variant Graph layout into a layered drawing, we reduce the cognitive load of the visualization by replacing the arrows with undirected edges aligned to the writing direction.

The third design rule avoids 'labeling edges'. In CollateX, edges are labeled with edition identifiers. This leads to two problems. Firstly, edge labels interfere with the vertex labels (text fragments), forcing the reader to visually separate this information. Secondly, if many editions pass an edge or if long edition identifiers are used, the corresponding edge labels significantly increase in size (Fig. 5). As a consequence, adjacent vertices drift apart in order to gain enough horizontal space to accommodate the edge labels, thus forcing the width of the graph to the point where the reader rapidly loses the context of a text fragment. To avoid this outcome, TRAViz does not label edges when visualizing Variant Graphs. Instead, it draws an edge for each edition in a different color. As the human ability to distinguish colors is limited,[8] this solution only works well with a small number of editions (less than ten). However, following the suggestions made by Ware (Ware, 2013) and Harrower (Harrower and Brewer, 2003), a set of varying color hues is defined so as to increase this number (twenty-four). Hence, a legend is required to map the colors (or in the case of CollateX, the edition identifiers) to the corresponding edition. Nevertheless, TRAViz is also capable of displaying edge labels upon interaction (see Section 3).

The fourth rule 'bundles major edges'. When analyzing and comparing text editions, the user is often interested in those editions that deviate from the 'standard reading'. In Stemmaweb, edges that are passed by most editions are labeled with a 'majority' tag and are accordingly bundled. As per the third rule, users are presented with multiple lines, which they bundle as major edges. Resultant edge types—bundled and unbundled—are highlighted differently: unbundled edges are color-coded with inviting saturated hues, whereas bundled edges appear as thick gray strokes. By doing so, any deviations from the standard reading can be more easily detected. As per the present and third rules, TRAViz is able to reduce the number of edges—and, therefore, the cognitive load of the user's approach—to a minimum.

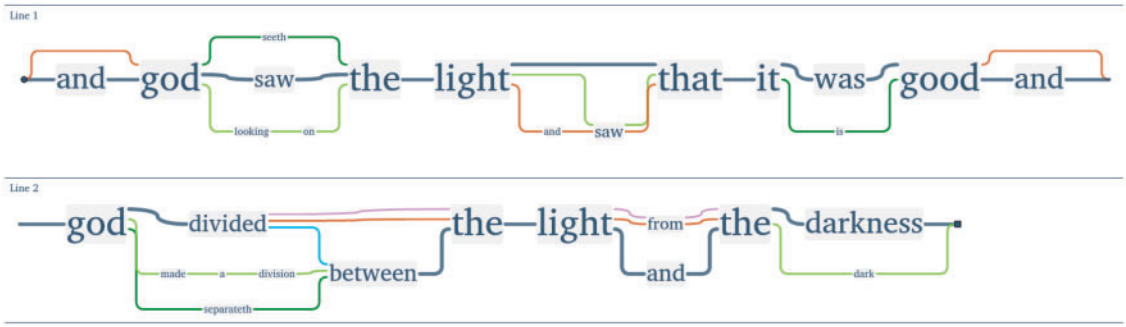

Fig. 6 TRAViz: Variant Graph illustrating the relationships between five different versions of Genesis 1:4

The fifth and final rule 'inserts line breaks'. One of the major problems emerging from the application of CollateX and TRAViz Variant Graphs to large portions of text was the enforced horizontal scrolling. In scrolling, especially long texts, one might lose context and struggle to track the distribution of editions in the graph. Moreover, the small screen space occupied by the graph translates into an increase in white-space. The outcome of a survey performed by the TAdER Project[9] on browser horizontal scrolling underpins our hypothesis that the user is accustomed to scrolling vertically. As a result, TRAViz inserts line breaks by splitting the Variant Graph when the width of the browser window is reached, thereupon mimicking the text layout of a book. As per the third rule, the user is presented with different-colored edges or edge bundles at the end of each line, so that all paths are visually identifiable at the beginning of the next line. Additionally, line numbers support vertical orientation. The insertion of line breaks helps the user navigate large graphs, concurrently preserving context.

Fig. 6 shows the TRAViz Variant Graph for *Genesis 1:4* in the aforementioned design. Unlike the CollateX Variant Graph, with TRAViz, it is possible to evaluate the level of variation: the central string of larger consecutive tokens connected by thick edges bundles analogous translations, and the two texts containing the highest degree of variation are the *Bible in Basic English* and *Young's Literal Translation*.

3 Means of Interaction

At the 2013 Digital Humanities Conference, Andrews stressed the importance for humanists to be able to interact with a Variant Graph and to modify its structure (Andrews and Van Zundert, 2013b). TRAViz responds to this requirement by providing a wide range of interaction possibilities, whereby users can tweak the visualization in order to suit their particular needs.[10]

Users may, for instance, prefer to visualize synonyms over spelling variation. In general, TRAViz only aligns exact word matches. This approach is useful, especially if the users want to focus on all types of variation (e.g. orthography). But if the users are not interested in orthographical variation, they are free to impose the alignment in order to better understand the correlation between the two words *A* and *B*. To this end, we define the 'Relative Edit Distance' *RED(A, B)* based upon the Levenshtein Distance *LevDist* (Levenshtein, 1966) as:

$$RED(A, B) = \frac{2^{*} LevDist(A, B)}{|A| + |B|}$$

where $RED(A, B) = 0$ is an exact match, $RED(A, B) \leq 0.2$ allows for smaller variations (e.g. 'beginnings' and 'beginning'), and $RED(A, B) \leq 0.5$ allows for greater variations (e.g. 'beginnings' and 'bigynnyng'). Higher values should not be used as the probability that unrelated words cluster progressively increases. If various versions are aligned, we use the most frequent version as the label for the

corresponding vertex in the graph. In the Bible usage scenario (see Section 4.1), the user can set the desired Relative Edit Distance by adjusting the slider provided.

The second means of interaction falls back to 'customary mouse behavior'. For a thorough analysis of the visualized Variant Graph, TRAViz employs the customary mouse hover-and-click gestures. Hovering over a vertex in the graph highlights all editions passing through it and hides all information and connections pertaining to other versions. This helps investigate the graph distribution of a subset of potentially similar translations and the exploration of the similarities and differences among them. Furthermore, this interaction singles out those editions forming majority edges. Two scenarios support the user in mapping colors to their corresponding editions (see Fig. 11), a particularly important functionality if working with a large number of editions. On the one hand, a mouseover on an edge shows the contributing editions in a tooltip; on the other, clicking on a vertex displays a pop-up window listing all editions and their corresponding tokens in the assigned colors.

Users may not always agree with machine-generated alignments. As stated by Andrews, humanists want to be able to modify graph structure and are more likely to adopt a tool that helps them achieve their desired alignment. For this reason, TRAViz offers the option to 'split and merge vertices'. To detach words, the user clicks on the corresponding button in the vertex pop-up window and creates a new branch in the graph. Merge operations are carried out through the customary Drag-and-Drop mouse functionality. As Variant Graphs are acyclic; two vertices can only be merged if this does not produce a cycle in the graph structure. When the user superimposes two vertices, the system calculates the feasibility of the merge: both words are highlighted in green if the merge is allowed or in red if not. If a merge is not possible in the first instance, a sequence of prior split operations can help to avoid a potential cycle.

Next, a user may want to explore those editions containing a higher degree of variation. By default, TRAViz follows the Stemmaweb concept and draws a majority edge—not individual edges per edition—if the connection between the corresponding vertices is passed by at least half of the editions. But depending on the research question and the data set to be examined, the 'definition of majority' may vary. For this reason, TRAViz allows the users to reduce the majority value if the overall variation among the texts is high. For editions sharing little variation, this threshold can be increased. In both cases, editions that do not make the majority group earn more visual presence.

Finally, TRAViz offers the option to 'visualize potential transpositions'. As Schmidt points out, the algorithmic detection of transposed text passages is extremely complex (Schmidt, 2009) and thus hard to manage in the web-browser. Nevertheless, TRAViz attempts to provide leverage points for potential transpositions by visually connecting related vertices. Two vertices are related if they share at least one word. Especially for large graphs, the number of potential transpositions is high. To avoid unnecessary clutter, the potential transpositions are only displayed when the users hover over a vertex for which transpositions have already been established. To visually separate transpositions from the graph's connections, we repurpose the concept of CollateX and visualize transpositions in the form of dotted black lines. This means of interaction is particularly useful when employing the relative edit distance merging various versions into one vertex. All of these vertices are labeled with only the most frequent version, so that transpositions are not always visible at first glance.

4 Usage Scenarios

The following case studies serve to demonstrate the flexible application and advantages of using TRAViz, with a view to emphasize its potential in the field of textual criticism in particular and of Digital Humanities in general.

4.1 Various English translations of the Bible

The first application of TRAViz involves the visualization of twenty-four English translations of the Bible. The Bible editions used for this example are listed in Table 1.

Table 1 Used Bible editions in chronological order[11]

1380 Wycliffe Bible	1885 English Revised Version
1535 Coverdale Bible	1890 Darby Bible
1537 Matthew Bible	1901 American Standard Version
1539 Great Bible	1949 Bible In Basic English
1560 Geneva Bible	1995 God's Word Translation
1568 Bishop's Bible	1999 American King James Version
1610 Douay-Rheims Bible	2000 Updated King James Version
1611 King James Version	2000 King James 2000 Version
1749 Douay-Rheims-Challoner Bible	2000 World English Bible
1833 Webster's Revision	2004 A Voice in the Wilderness
1863 Young's Literal Translation	2009 Catholic Public Domain Version
1876 Smith's Literal Translation	2009 Lighthouse Bible

Fig. 7 shows a web interface embedding TRAViz. The top panel lists the different Bible versions with their assigned colors. Sets of translations can be analyzed at once by checking or unchecking their corresponding boxes in the panel. Two sliders can be used to further modify the graph structure, either by defining the majority value or by re-aligning the underlying verses—dependent on a relative edit distance—so as to highlight the most obvious differences. The graph in Fig. 8 is the resulting output of the joint application of a majority of at least six editions and of a relative edit distance of 40%, aligning various versions of 'beginning', 'heaven', and 'earth' into one vertex each.

The following are three examples illustrating how this visualization supports distinct research questions.

The first example concerns itself with the sixth commandment (*Exodus 20:13* and *Deuteronomy 5:17*), commonly recited as 'You shall not kill'. But, there exist numerous different translations of this verse, which can be crucial for its interpretation. The Variant Graph in Fig. 9 displays the sixth commandment with a majority of four: the thick gray lines chain synonymous terms, which have been used by the translations at least four times; the larger font size of some of the words expresses higher usage. The variation of the verb is of particular interest. Twelve translations choose the word 'kill', and four editions its older form 'kyll'; five authors write 'murder', one 'sle', and another 'put anyone to death without cause'. The latter variation belongs to the '*Bible In Basic English*', a non-literal translation, and is easily distinguishable by its length. Another version, significantly shorter, is that contained in '*God's Word Translation*': 'Never Murder'.

Users not interested in orthographical differences may cluster those variants by manually dragging and dropping vertices, or by automatically enabling a relative edit distance of 40%. Not only does the latter approach save time, but the users learn to quickly define the optimal percentage for the relative edit distance inasmuch as the results are simultaneously displayed. The higher the percentage, the more the words tend to cluster. For the chosen examples, a relative edit distance of no more than 50% seems to yield the best results without clustering unwanted words. A higher percentage is also possible if the users unravel this cluster by manually splitting unrelated words into separate vertices. In the example above, the orthographic variations 'kill/kyll' and 'shalt/shall/schalt' are clustered, thus drawing more attention to the semantic dissimilarities. The sixth commandment as described in *Deuteronomy 5:17* (Fig. 10) showcases more variance compared to the *Exodus 20:13* version.

By selecting a relative edit distance of 50%, we notice a high frequency of the verb 'slay' and its variants 'slaye/sle/slea' (Fig. 11). Clicking the corresponding vertex reveals no semantic dissimilarities between the contributing editions.

A second example (Fig. 12) displays an output scenario for *Luke 2:1*, generated by selecting a majority of four, a relative edit distance of 40%, and by manually merging words and splitting vertices.

The result brings out synonyms and matching expressions in the different editions. For instance, the opening expression has many variations, including 'it came to pass(e)/about', and 'it happened/chanced/fortuned/was (don)'. The *God's Word Translation* does not translate this expression at all. Moreover, the translations vary between '(all) the (whole/habitable) world/globe' (the term 'globe' only appears in *Smith's Literal Translation*) and 'the roman empire', an interpretative variant of the *God's Word Translation.* Different expressions are also used to describe the emperor's decree, namely those stacked at the end of the verse stating that everybody should be 'taxed', 'enrolled', 'discyrned', or 'registered'. Whilst 'misplaced', variations containing the words

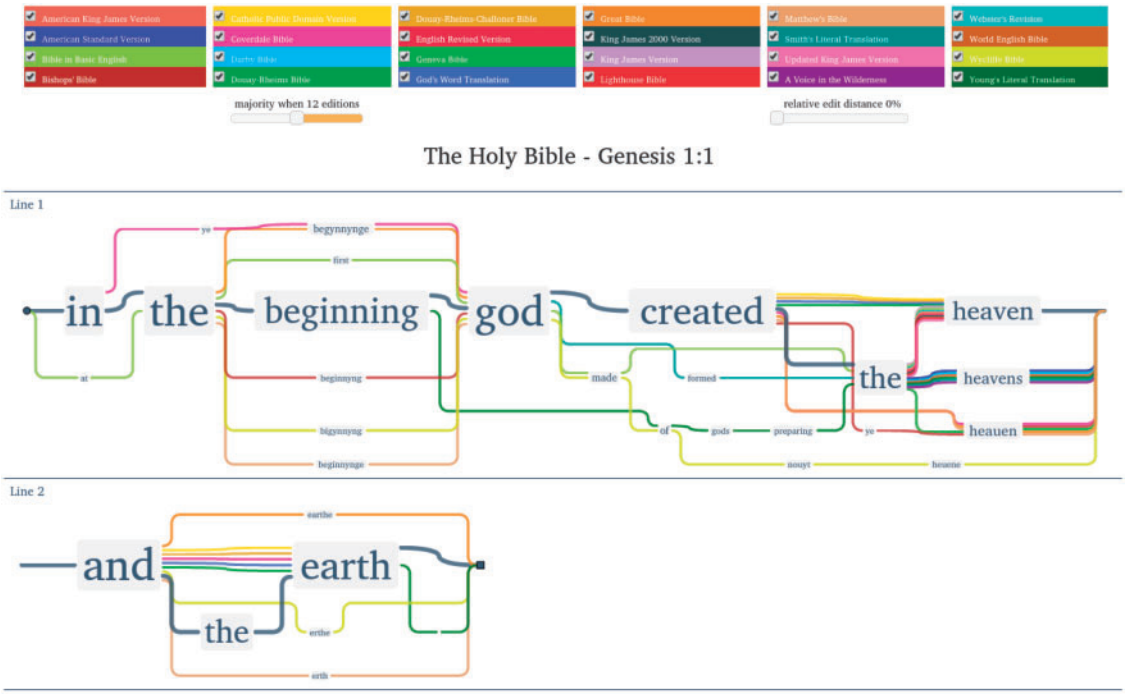

Fig. 7 *Genesis 1:1* in twenty-four English translations

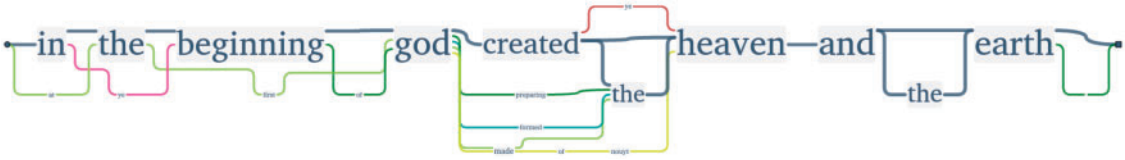

Fig. 8 *Genesis 1:1* in twenty-four English translations with a majority of four and $RED = 0.4$

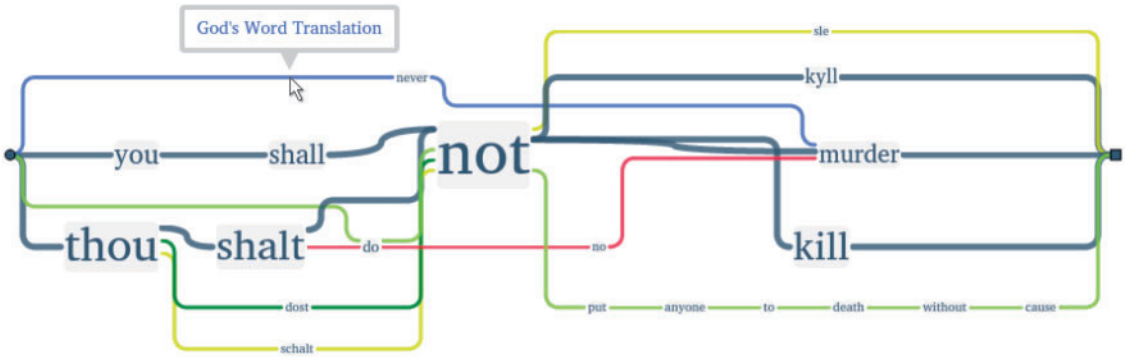

Fig. 9 *Exodus 20:13* in twenty-four English translations with a majority of four

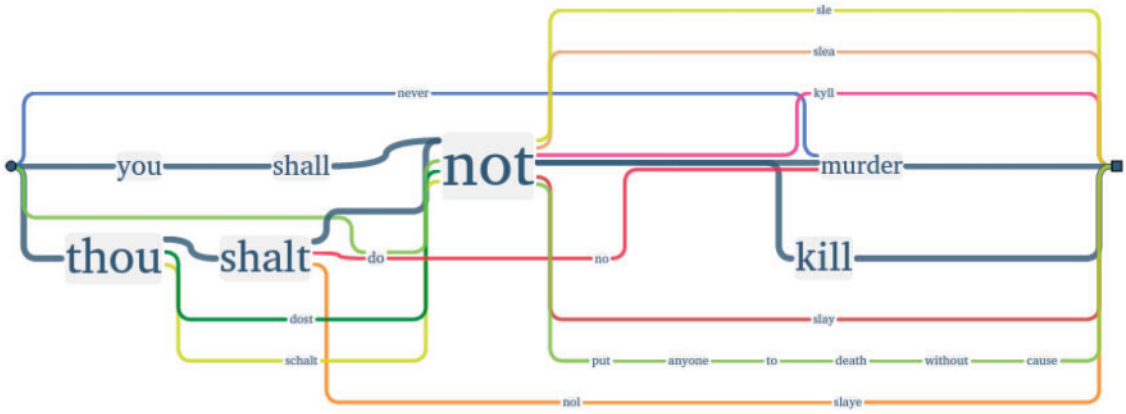

Fig. 10 *Deuteronomy 5:17* in twenty-four English translations with a majority of four and *RED* = 0

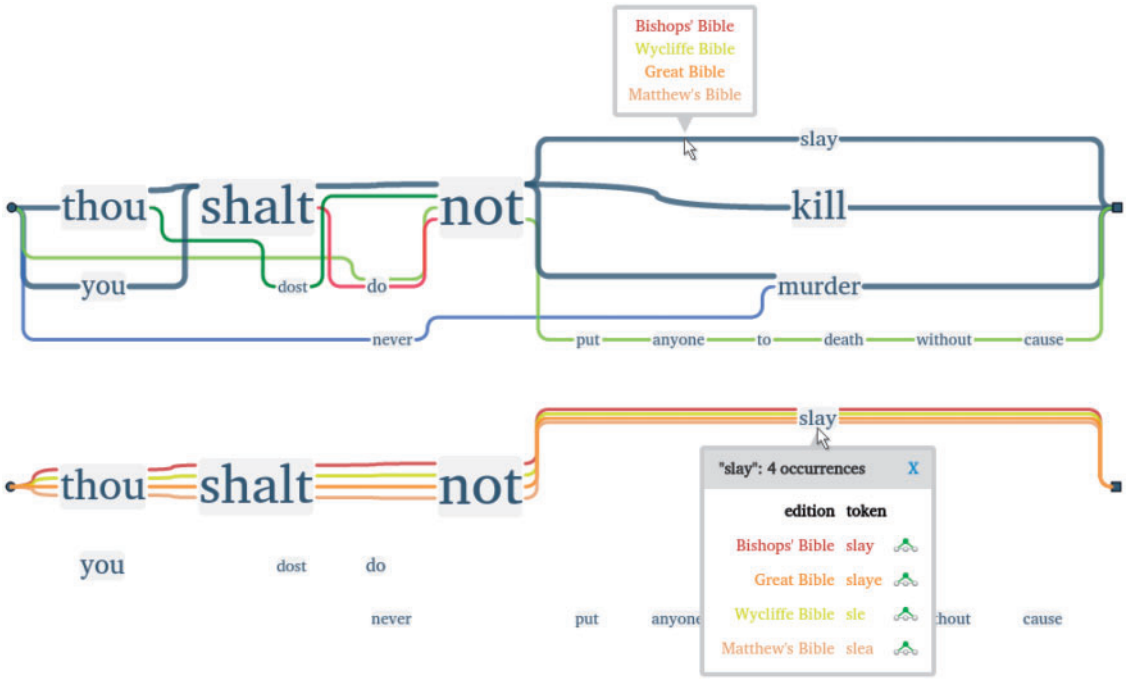

Fig. 11 *Deuteronomy 5:17* in twenty-four English translations with a majority of four and *RED* = 0.4. Hovering an edge reveals corresponding edition labels (top), clicking a vertex additionally lists orthographic variations (bottom)

'census' and 'numbering' are nevertheless clearly discernible owing to their divergent branch structure. When brought together, all of these variant readings help convey the meaning of the decree as understood by various translators.

Differences in word order can also be better understood by enabling the transposition option (see Fig. 13): the position of the word 'decree', for instance, changes between translations, from 'a decree went out' to 'went out a decree'.

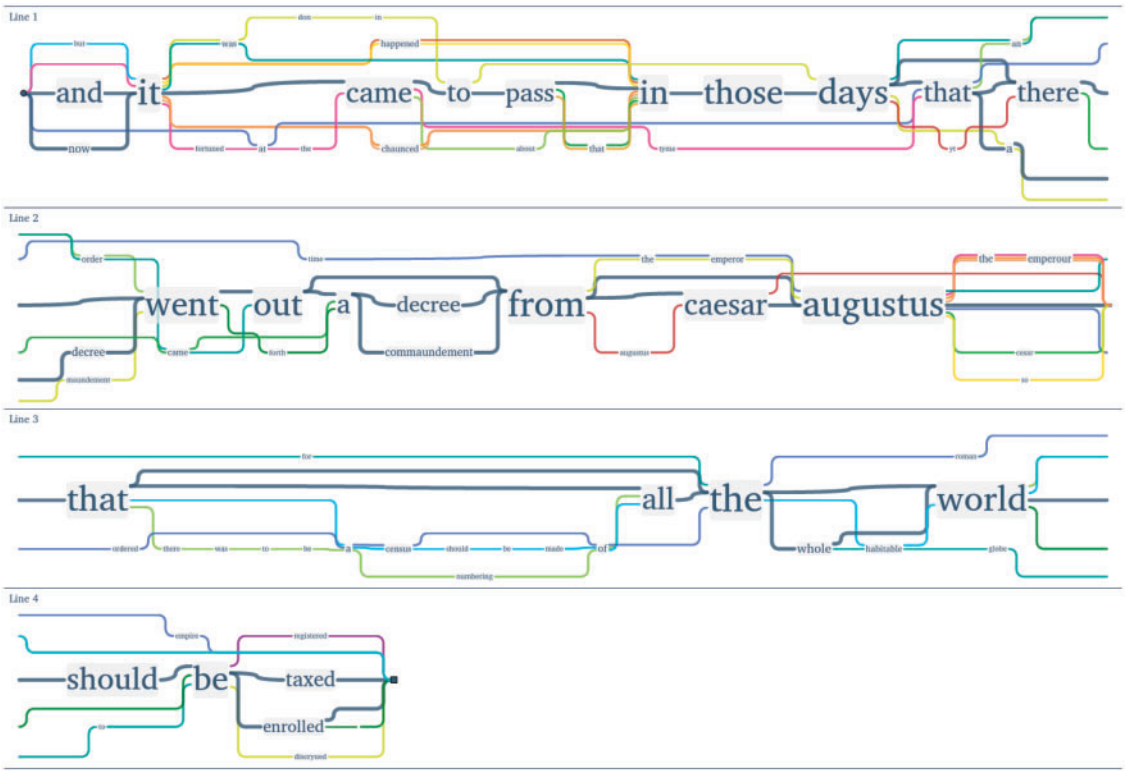

Fig. 12 *Luke 2:1* in twenty-four English translations with *RED* = 0.4

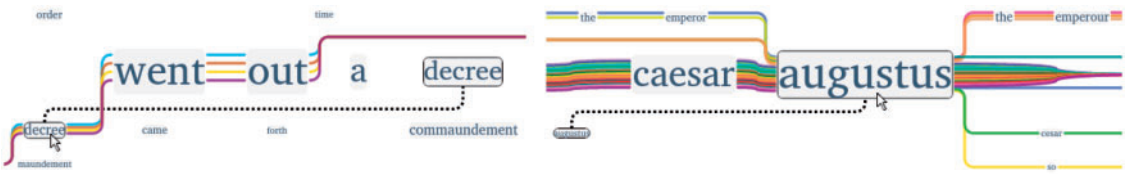

Fig. 13 Transpositions in *Luke 2:1*

Hovering the cursor over the term 'augustus' displays two transpositions, 'caesar augustus' and 'augustus caesar'. This detection leads to yet another transposition pertaining to the Roman emperor stemming from the word 'augustus'. Due to different orthography, the terms 'augustus cesar', '(the) emperor augustus', and 'augustus the emperour' are not shown as variations, but can easily be identified, because they are marked as transpositions.

The third example scrutinizes the most influential English translation of the Bible, the *King James Bible* (Ryken, 2011). In order to understand the development of biblical variants before the *King James Bible* became so important, eight translations dating between 1500 and 1800 are chosen as a representative sample: the *Coverdale Bible*, the *Matthew Bible*, the *Great Bible*, the *Geneva Bible*, *Bishop's Bible*, the *Douay-Rheims Bible*, the *King James* version, and the *Douay-Rheims-Challoner Bible*. Fig. 14 shows the resulting Variant Graph applied to *Genesis 1:28*.

From an orthographical standpoint, one notices the emergence of clusters. Unsurprisingly

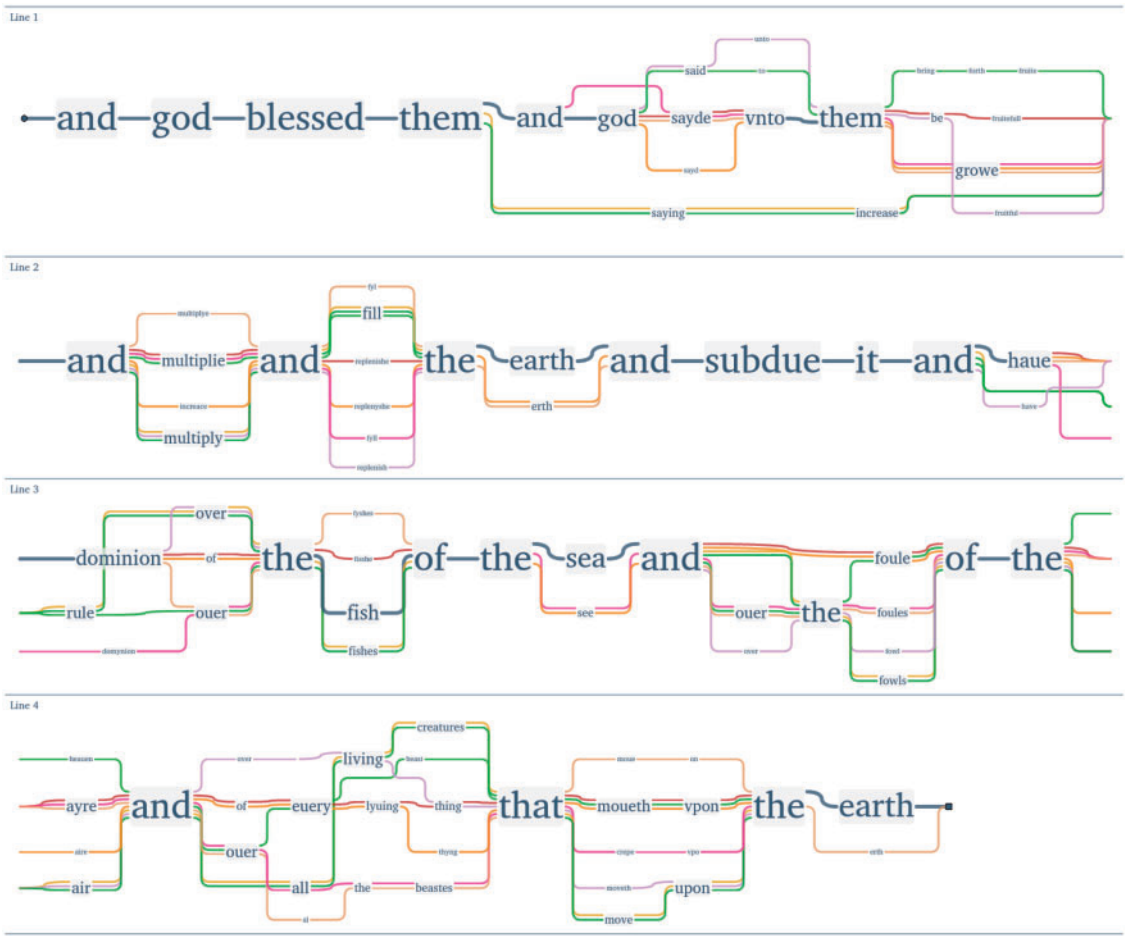

Fig. 14 Eight English translations for *Genesis 1:28*

Douay-Rheims- and the *Douay-Rheims-Challoner-Bible* form a cluster. It is evident, that those two editions tend to deviate from the majority. Although the *Douay-Rheims-Challoner Bible* is said to have been influenced strongly by the *King James Version* (Newman, 1859), a specific closeness to this edition is not visible, at least not in this verse. The *Bishop's Bible* has an independent writing style and as a revision of the *Great Bible*, it tends to keep its content-based variations (e.g. 'of the fish' instead of 'over the fish') but uses a more contemporary orthography. But hovering for instance over the word 'thing', which only these two Bible translations share, shows how similar the *Bishop's Bible* and the *King James Version* are in translating this verse.

Fig. 15 juxtaposes the Variant Graphs of old (1500–1800) and modern (after 1800) Bible translations, both including the *King James Version*, whose words appear in the vertical center of the graphs. We applied a majority of five and a relative edit distance of 50%. Comparing the results of this particular verse with other verses corroborates, the impression that the majority of both modern and older translations employed similar if not identical terminology to that used by the *King James Version.* The graph shows a strong line of transmission among older translations (Fig. 15, left), and even the deviations from it seem to be done in a similar way. Here, the more modern editions (Fig. 15, right) deviate significantly from the *King James Version*, especially the *Bible in Basic*

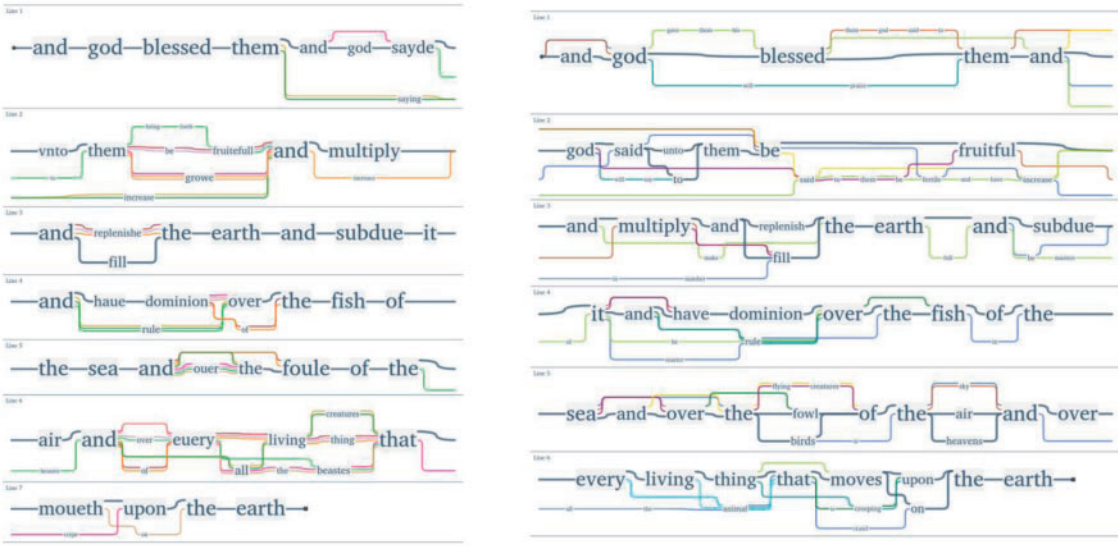

Fig. 15 Old (left) and modern (right) English translations of *Genesis 1:28* compared to the *King James Version* visualized with a majority of five and $RED = 0.5$

English, the *God's Word Translation*, *Smith's Literal Translation*, and the *World English Bible*.

4.2 A variation study of the earliest manuscript witnesses of St. Augustine's De Civitate Dei

TRAViz is also being used in a doctoral project that seeks to create the first digital edition of the oldest surviving manuscript of St. Augustine's *De Civitate Dei*.[12] To this end, the PhD candidate has recently deployed TRAViz to visualize her editorial work: whilst her collation is currently limited to the comparison of this manuscript to a standard edition of the text (*Corpus Scriptorum Ecclesiasticorum Latinorum*, Vol. 40), a postdoctoral enhancement envisages the addition of other witnesses in order to better understand the large textual tradition of *De Civitate Dei*. Fig. 16 illustrates a first collation experiment.

This sample visualization brings together the author's transcriptions[13] of the aforementioned standard edition and of four manuscripts of *De Civitate Dei* from four different digital repositories[14]: MS XXVIII(26)[15], Cod. Sang. 178[16], BSB-Hss Clm 6267[17], Cod. Lat. 121[18] and CSEL 40[19]. As per previous examples, TRAViz provides information about word frequency and enlarges those words which are shared by the majority of the witnesses. With a tradition and textual heritage spanning over 1500 years, this visualization has the potential to include numerous witnesses.

In terms of benefits, TRAViz relieves the author of substantial collation work by, effectively, transforming her transcriptions into a 'variorum' edition. In the example above, TRAViz allows both the editor and the reader to easily map and investigate errors (for instance, the misspelling of *ac* in Cod. Lat. 121 turns the conjunction into *hac*, a Latin demonstrative pronoun and/or adverb), alternative spellings, word order, and variant readings (*eisque* instead of *de his quae*). Current limitations are bound to palaeographical conventions, whereby special characters are used to abbreviate words. Such conventions are not unusual occurrences in ancient and medieval manuscripts and are commonly digitally represented in accordance with Unicode specifications.[20] TRAViz fully supports Unicode but is limited to those characters for which Unicode provides digital equivalents. Seeing as Unicode does not yet (at the time of writing) cater for all palaeographical conventions and that TRAViz depends on this availability,

Fig. 16 Excerpt of book 11, chapter 3 of *De Civitate Dei*

TRAViz visualizations of ancient and medieval documents may contain orthographical inaccuracies or adaptations.

TRAViz has proved to be an innovative working method for the study of the textual variation and transmission of Augustine's 'magnum opus'. Not only does it support—and partly automate—philological, palaeographical, and critical enquiries, but it also allows the author to easily identify mistakes produced during the error-prone transcription process.

4.3 Comparing German Shakespeare translations

The year 2014 marked the 450th anniversary of William Shakespeare's birth and the 300th anniversary of the *German Shakespeare Society*.[21] The joint celebration was accompanied by a multidisciplinary Symposium—held at the Mainz Academy of Sciences and Literature—on the German reception of Shakespeare's works. To complement contributions from the literary and cultural sciences, the Leipzig Max Planck Institute for Mathematics in the Sciences and the Natural Language Processing Group of Leipzig University were invited to present newly developed language analysis methods and their applications to Shakespearean works (Efer *et al.*, 2015).

TRAViz was presented as a means of visualizing linguistic and structural aspects expressing the dramatic progression in Shakespeare's plays, and of highlighting differences and similarities between various German translations of his works. Fig. 17 shows how varied the language can be across twenty-two German editions of the same Othello excerpt. Along the vertical axis, one notices word substitutions and preferred alternatives—constituting the so-called 'paradigmatic level'—such as 'brief' and 'schreiben' for 'letter', or 'galeeren' (galleys) and 'schiffe' (ships). Interestingly, the original 'hundred and seven galleys' was changed to '106' nine times. This circumstance is most likely attributable to a smaller syllable count and to the softer pronunciation of 'ga-lee-ren'. With regards to word position and order, the so-called 'syntagmatic level' bears fixed and grammar-related expressions, such as the plural 'meine briefe' (my letters) versus the singular 'mein brief' (my letter), or the passive construction 'mir wird gemeldet' (is reported to me) versus the active 'nennt mir' (tells me), which can be visually tracked along the horizontal axis. The overall syntax of the visualized Othello scene is relatively stable with the exception of the transposition of 'galeeren'. The corresponding edition by Schaller (1959) chooses 'a hundred and seven galleys say my letters' over the accepted 'my letter says a hundred and seven galleys'.

4.4 Exploring the multiple meanings of a term in ancient Greek texts

In the Digital Humanities *eXChange* project,[22] TRAViz is being used to align and visualize Ancient Greek text snippets containing specific terms. The humanists on the team explore the various meanings

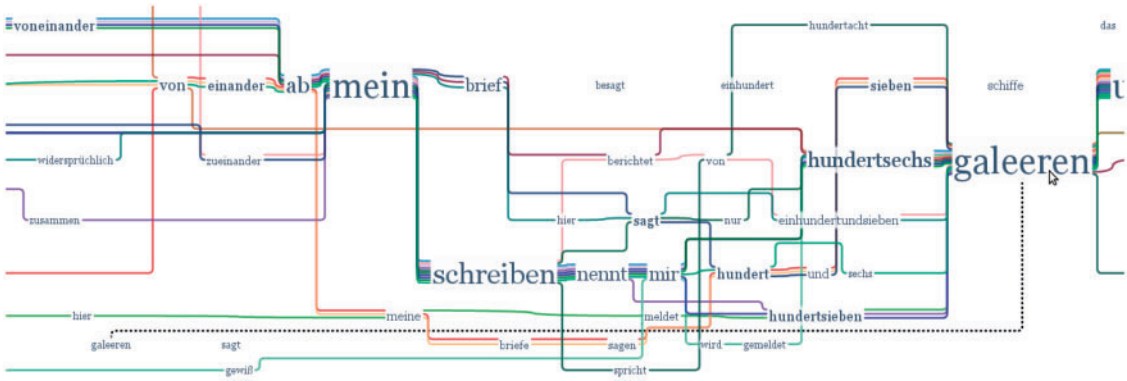

Fig. 17 Part of Shakespeare's 'Othello, Act 1, Scene 3'

of the given term defined by a set of descriptors. Their intent is to learn how these meanings were transmitted and how they changed over time.

Eva Wöckener-Gade, a classical philologist, is interested in the multiple meanings of the term *φάρμακον*. A visualization of text snippets containing the truncated form *φάρμακ-* yields clusters representing the different meanings, all easily traceable by hovering over vertices and highlighting branches that include related words (Fig. 18). The word *ἕλκει* (wound) and the variations of *ἐκέσματα* (cure) interpret *φάρμακον* as 'medicine', whereas the word *ἐθέλχθης* (aorist form of the verb 'to enchant') supports the meaning 'magical cure'. Both meanings are to be found in Ancient Greek poetry. By selecting a relative edit distance of 30%, TRAViz bundles up various versions of the verb 'to apply' (*ἔπασσεν*, *ἔπασσον*, *ἔπασσε*), a term often used in the context of medication. Similarly, *πίων* suggests the magical cure was drunk.

Contrarily to the traditional methods of analyzing the contexts of a given term's occurrence, this visualization facilitates a rapid comprehension of the term's different meanings by clustering the related tokens that define them.

4.5 Syllable stress

Dr. Michael Cade-Stewart, a British Academy Postdoctoral fellow at Kings College London whose main interest lies within the digital exploration of Poetic Rhythm between 1800 and 1970, analyzes and visualizes different ways of orally performing poems. An example of six different possibilities for a pentameter line (Shakespeare's Sonnet 18, line 1) is shown in Fig. 19. The stressed syllables are in upper-case, the un-stressed in lower. The likelihood that a syllable will be stressed in performance is indicated by the size of the word, and by the number of colored strands that run through it. The words 'a summer's day' are performed equally in all versions, and can therefore be regarded as relatively objective; the other syllables are more subjective. If one wanted to look at a feature in the stressed syllables, one might restrict one's focus to the more objective variants or, in this case, weigh them more strongly.

5 Conclusion

The present article introduced the web-based library TRAViz, which implements a novel design for Variant Graphs. Furthermore, it demonstrated its various means of interaction, e.g. the relative edit distance, to facilitate an on-demand modification of the underlying alignment, and the graph layout, strongly dependent on the user's research question. Finally, the use cases discussed have served to prove the intuitiveness of the design and the applicability of the tool to distinct studies in the humanities.

In the ongoing development phase, humanists working with TRAViz systematically evaluate design and interaction capabilities in order to meet the needs of the inexperienced, and perhaps sceptical, user. This iterative evaluation process has proven to be invaluable in the development of

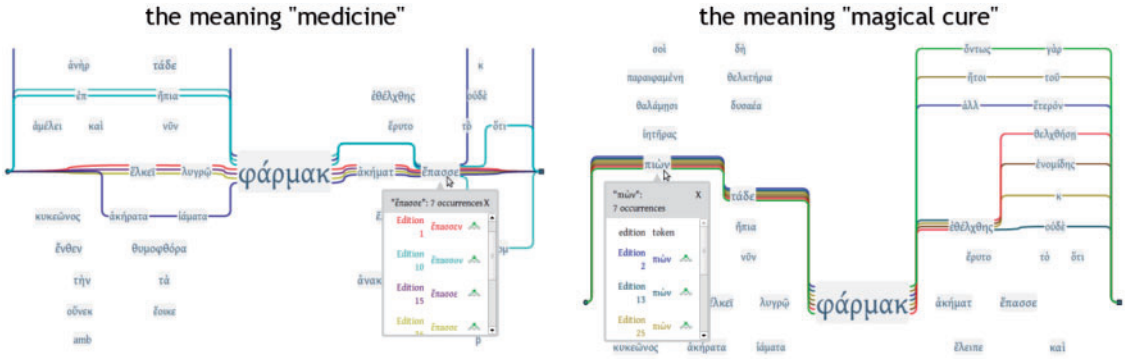

Fig. 18 The two meanings of *φάρμακον*

Fig. 19 Six ways of stressing 'Shakespeare's Sonnet 18, line 1'

TRAViz and is, therefore, highly recommended to scholars engaged in analogous projects.

By and large, the Variant Graph model works well on smaller text entities. The Bible use case adequately fits this model inasmuch as the highly structured text facilitates verse by verse collation. In addition, transpositions can be analyzed with tools such as TRAViz or CollateX. But when the text passages are long, the probability of transpositions drifting apart and of unaligned text snippets appearing rapidly increases. While the structure of the Variant Graph still works, an exhaustive visualization and analysis is almost impossible. One solution would be to modify the model so that it accepts cycles, thus ensuring that similar patterns retain alignment independently of their order in the corresponding source texts. As this approach is only applicable to moderately larger text entities, we instead suggest methods based upon Text Re-use detection.[23]

In the future, we will direct our attention to developing a distant reading visualization for Variant Graphs. The Bible scenario has demonstrated that TRAViz can be adequately used to compare multiple editions of a certain verse. But a distant view of other hierarchy levels such as chapter, book, or of the whole Bible is not yet possible.

It has been shown, that TRAViz is suited for working with various text corpora. An approach to work interlingually should prove very interesting as well, aligning words of similar meaning in different languages, for example translations of the same text (e.g. the Bible) into different languages. This could enrich dictionary databases, show different translation approaches as well as reception, or even be useful for language acquisition.

Acknowledgements

The authors thank Thomas Efer, Eva Wöckener-Gade, and Dr Michael Cade-Stewart for using TRAViz in their research and for providing the use cases discussed in this article. The authors are also indebted to Marco Büchler for fruitful discussions about TRAViz, to David Joseph Wrisley for suggesting the inclusion of the edit distance, and to the Baker Publishing Group for permission to include the *God's Word Translation* in our database.

Funding

This research was funded by the German Federal Ministry of Education and Research.

References

Andrews, T. L. and Macé, C. (2013a). Beyond the tree of texts: building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, 28(4): 504–21.

Andrews, T. L. and Van Zundert, J. J. (2013b). An interactive interface for text variant graph models. In *Proceedings of the Digital Humanities* 2013, Lincoln, United States.

Dekker, R. H. and Middell, G. (2011). Computer-supported collation with CollateX: managing textual variance in an environment with varying requirements. In *Supporting Digital Humanities*. Denmark: University of Copenhagen.

Efer, T., Heyer, G. and Jost, J. (2015). Text Mining am Beispiel der Dramen Shakespeares. In Jansohn, C., Habicht, W., Mehl, D. and Redl, P. (eds), *Shakespeare unter den Deutschen.* Stuttgart: Akademie der Wissenschaften und der Literatur, Mainz, Franz Steiner Verlag (in Vorbereitung).

Gibbs, F. and Owens, T. (2012). Building Better Digital Humanities Tools: toward broader audiences and user-centered designs. *Digital Humanities Quarterly,* 6(2).

Harrower, M. and Brewer, C. A. (2003). ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1): 27–37.

Jänicke, S., Efer, T., Büchler, M. and Scheuermann, G. (2015). Designing close and distant reading visualizations for text re-use. To appear in *Computer Vision, Imaging and Computer Graphics. Theory and Application Communications in Computer and Information Science.*

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10: 707.

Newman, J. H. (1859). The Text of the Rheims and Douay version of Holy Scripture. *The Rambler*, Vol. I, New Series, Part II.

Ryken, L. (2011). *The Legacy of the King James Bible: Celebrating 400 Years of the Most Influential English Translation*. Crossway, Wheaton, Illinois.

Schmidt, D. (2009). Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11–14, 2009. In Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009).

Schmidt, D. and Colomb, R. (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67(6): 497–514.

Sugiyama, K., Tagawa, S. and Toda, M. (1981). Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man and Cybernetics*, 11(2): 109–25.

Szpiech, R. (2014). Cracking the code: reflections on manuscripts in the age of digital books. *Digital Philology: A Journal of Medieval Cultures*, 3(1): 76–301.

Taliaferro, B. B. (2013). *Encyclopedia of English Language Bible Versions*. Jefferson, NC: McFarland.

Tanselle, G. T. (1992). *A Rationale of Textual Criticism*. University of Pennsylvania Press, Philadelphia, Pennsylvania.

Viégas, F. B., Wattenberg, M. and Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6): 1137–44.

Ware, C. (2013). *Information Visualization: Perception for Design*. Elsevier, Amsterdam, Netherlands.

Wattenberg, M. and Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6): 1221–8.

Notes

1 Abbyy FineReader OCR (Optical Character Recognition) and ICR (Intelligent Character Recognition) technologies, for instance, are currently being developed to support computer recognition of handwritten script.

2 A directed acyclic graph is composed of vertices and directed edges (arrows) and does not contain directed cycles. That means, it is not possible to start at an arbitrary vertex v of the graph and follow a sequence of edges that eventually loops back to v again.

3 CollateX was designed as the successor to Peter Robinson's Collate software (see http://www.sd-editions.com/), which offered a primarily textual representation of variation. The development of CollateX was guided by the Interedition consortium (http://www.interedition.eu/).

4 http://www.graphviz.org/

5 TRAViz (http://www.traviz.vizcovery.org/) primarily focuses on the design of Variant Graphs. In correspondence to the Gothenburg model (http://wiki.tei-c.org/index.php/Textual_Variance), the TRAViz pipeline consists of tokenization, normalization, alignment (details can be found in Jänicke *et al.*, 2015), analysis (e.g., a heuristic of transpositions at word level), and visualization.

6 A 'Word Tree' highlights the different endings of sentences that share the same beginning.
7 In Fig. 4, the direction of the edge from the word 'saw' (right-top) to 'the light' (bottom-left) is dextro-sinistral (Right-To-Left). This is called a ‚backward edge'.
8 http://www.w3.org/WAI/WCAG20/quickref/#qr-visual-audio-contrast-without-color
9 http://www.tader.info/scrolling.html
10 Gibbs and Owens call for better and more user-friendly digital tools to support humanities scholarship (Gibbs and Owens, 2012). TRAViz's focus on design and transparent documentation can be understood as a response to Gibbs' and Owens' recommendations.
11 Dates taken from the *Encyclopedia of English Language Bible Versions* (Taliaferro, 2013).
12 For more information about the project, please visit www.gretafranzini.com and https://sites.google.com/site/gretafranzini/home
13 These were produced for the sake of demonstration; the author has not produced complete transcriptions of the manuscripts cited.
14 For the purpose of this illustration, ligatures and abbreviations have been expanded. However, TRAViz also supports special characters. Additional details can be found under http://www.traviz.vizcovery.org/tutorial.html#htmlcodes.
15 Early 5th century; today in Verona, Italy.
16 Mid 9th century; today in St. Gallen, Switzerland. Permanent URL: http://www.e-codices.unifr.ch/en/list/one/csg/0178
17 9th century; today in Munich, Germany. Permanent URL: http://daten.digitale-sammlungen.de/~db/0003/bsb00039815/images/index.html?id=00039815&fip=ewqqrseayaxdsydxdsydeayaxdsydxsxdsydxdsydfsdr&no=&seite=4
18 Late 15th century; today in Budapest, Hungary. Permanent URL: http://www.corvina.oszk.hu/corvinas-html/hub1codlat121.htm
19 Standard edition, 1899-1900. PDF available to download from the Internet Archive: Part 1: https://archive.org/details/corpusscriptoru20wiengoog; part 2: http://archive.org/stream/corpusscriptoru07wiengoog#page/n7/mode/2up)
20 See http://unicode.org/ (Accessed: 26 February 2015)
21 Deutsche Shakespeare-Gesellschaft, http://shakespeare-gesellschaft.de/
22 http://exchange-projekt.de/
23 When applying Text Re-use detection methods, the source texts are segmented into small entities (e.g. sentences), and similarities among them are determined. A visualization of the results of such an approach is the Text Re-use Browser (Jänicke *et al.*, 2015). The source texts are juxtaposed and intervening lines connect related text entities. A regular pattern between text editions appears as a set of parallel lines, whereas transpositions stand out as intersections.