Visualizing Translation Variation: Shakespeare's *Othello*

Zhao Geng[1], Robert S.Laramee[1], Tom Cheesman[2], Alison Ehrmann[2], and David M. Berry[2]

[1] Visual Computing Group, Computer Science Department, Swansea University, UK,
{cszg,r.s.laramee}@swansea.ac.uk
[2] College of Arts and Humanities, Swansea University, UK,
t.cheesman@swansea.ac.uk, alison.ehrmann@t-online.de, d.m.berry@swansea.ac.uk

Abstract. Recognized as great works of world literature, Shakespeare's poems and plays have been translated into dozens of languages for over 300 years. Also, there are many re-translations into the same language, for example, there are more than 60 translations of *Othello* into German. Every translation is a different interpretation of the play. These large quantities of translations reflect changing culture and express individual thought by the authors. They demonstrate wide connections between different world regions today, and reveal a retrospective view of their cultural, intercultural, and linguistic histories. Researchers from Arts and Humanities at Swansea University are collecting a large number of translations of William Shakespeare's *Othello*. In this paper, we have developed an interactive visualization system to present, analyze and explore the variations among these different translations. Our system is composed of two parts: the structure-aware Treemap for document selection and meta data analysis, and Focus + Context parallel coordinates for in-depth document comparison and exploration. In particular, we want to learn more about which content varies highly with each translation, and which content remains stable. We also want to form hypotheses as to the implications behind these variations. Our visualization is evaluated by the domain experts from Arts and Humanities.

1   Introduction

William Shakespeare is widely regarded as one of the greatest writers and his plays have been translated into every major living language. This is a historical and contemporary phenomenon. In German, the first translation of one play, *Othello*, was produced in 1766. By now there are over 60 translations including 7 new translations of this play which have been produced since the year 2000. Questions about these translations are seldom asked in the Anglophone world, because interpreting them is difficult without specialist linguistic and cultural knowledge. The original Shakespeare's work in English is normally considered more important than any translations. But with increasing awareness of global cultural interconnections, more Arts and Humanities researchers recognize the significance of translations and are investigating them.

The interpretation of Shakespeare's work in translation is always influenced by the translator's own culture, customs and conventions. Therefore, each translation is a product of changing culture as well as an expression of each translator's individual thought

within that culture. Also, each translation is a reply to received ideas about what Shakespeare's work means. Semantic and textual variations between translations in the corpus carry relational cultural significance. Normally, researchers from Arts and Humanities read and compare cultural text in its raw form and this makes the analysis of the multiple translations difficult. In addition, interesting patterns are often associated with text metadata, such as historical period, place, text genre or translator profession.

Up until now, researchers from Arts and Humanities have collected more than 50 different versions of German translations of Shakespeare's play, *Othello*. Our general goal is to identify similarities and differences among these translations. Compared to traditional text mining, text visualization incorporates the visual metaphors and interactive design to facilitate in-depth exploratory data analysis.

In this paper, we aim to develop an interactive visualization system to help the researchers from Arts and Humanities perceive and understand their collected German translations in new ways. In order to do so, we collect a large amount of metadata associated with the original documents and extract semantic features from the document contents. Based on such extracted information, various visualizations can be applied. We propose a structure-aware Treemap for metadata analysis and document selection. Once a group of documents are selected, they can be further analyzed by our Focus + Context parallel coordinates.

The rest of this paper is organized as follows: In Section 2, we review the previous work on text visualization. In Section 3, we describe our source data. In Section 4, we explain how are the original documents processed before being input to the visualization. In Section 5, we illustrate our structure-aware Treemap for meta data analysis. In Section 6, we present the Focus + Context parallel coordinates for translation variation exploration. In Section 7, we report the feedback from the domain experts. Section 8 wraps up with the conclusion.

2 Related Work

Since 2005, from the major visualization conferences, we can observe a rapid increase in the number of text visualization prototypes being developed. As a result, various visual representations for text streams and documents are proposed to effectively present and explore the text features.

A large number of visualizations have been developed for presenting the global patterns of individual document or overviews of multiple documents. These visualizations are able to depict word or sentence frequencies, such as Tag Clouds [1], Wordle [2], WordTree [3], or relationships between different terms in a text, such as PhraseNet [4], TextArc [5] and DocBurst [6]. The standard Tag Clouds [1] is a popular text visualization for depicting term frequencies. Tags are usually listed alphabetically and the importance of each tag is shown with font size or color. Wordle [2] is a more artistically arranged version of a text which can give a more personal feel to a document. ManiWordle [7] provides flexible control such that the user can directly manipulate the original Wordle to change the layout and color of the visualization. Word Tree [3] is a visualization of the traditional keyword-in-context method. It is a visual search tool for unstructured text. Phrase Nets [4] illustrates the relationships between different words

used in a text. It uses a simple form of pattern matching to provide multiple views of the concepts contained in a book, speech, or poem. A TextArc [5] is a visual representation of an entire text on a single page. It provides animation to keep track of variations in the relationship between different words, phrases and sentences. DocuBurst [6] uses a radial, space-filling layout to depict the document content by visualizing the structured text. The structured text in this visualization refers to the is-kind-of or is-type-of relationship. These visualizations offer an effective overview of the individual document features, but they cannot provide a comparative analysis for multiple documents.

In contrast to single document visualizations, there are relatively few attempts to differentiate features among multiple documents. Noticeable exceptions include TagLine Generator [8], Parallel Tag Clouds [9], ThemeRiver [10] and SparkClouds [11]. Tagline Generator [8] generates chronological tag clouds from multiple documents without manual tagging of the data entries. Because the TagLine Generator can only display one document at a time, it is unable to reveal the relationships among multiple documents. A much better visualization for this purpose is Parallel Tag Clouds [9]. This visualization combines parallel coordinates and tag clouds to provide a rich overview of a document collection. Each vertical axis represents a document. The words in each document are summarized in the form of tag clouds along the vertical axis. When clicking on a word, the same word appearing in other vertical axes is connected. Several filters can be defined to reduce the amount of text displayed in each document. One disadvantage of this visualization is its incapability to display groups of words which are missing in one document but frequently appear in the others. When we explore the variations among the *Othello* translations, the domain experts would like to know groups of words which a particular author never uses but which frequently appear in other authors' work. Also, brushing multiple words in different documents might introduce clutter due to the crossing lines in parallel tag clouds.

We also observe some interesting visualizations which can depict time trends over different documents. SparkClouds [11] integrates sparklines into a tag cloud to convey trends between multiple tag clouds over time. Results of a controlled study that compares SparkClouds with traditional trend visualizations, such as multiple line graphs, stacked bar charts and Parallel Tag Clouds, show that SparkClouds is more effective at showing trends over time. The ThemeRiver [10] visualization depicts thematic variations over time within a large collection of documents. The thematic changes are shown in the context of a time line and corresponding external events. This is the first work, to our knowledge, that compares multiple translations of a single play.

3 Background Data Description

The domain experts from Arts and Humanities have collected 57 different German translations of Shakespeare's play, *Othello*. For each translation, metadata recorded includes the author name, publication date, country, title of the play and impact index. The translations were written between 1766 and 2006 in seven different countries defined including Germany (pre-1949), East Germany (1949-1989), West Germany (1949-1989), FRG (Germany since 1989), Austria, Switzerland and England. The impact index refers to each translator's productivity and reputation. it includes the re-publication figures or
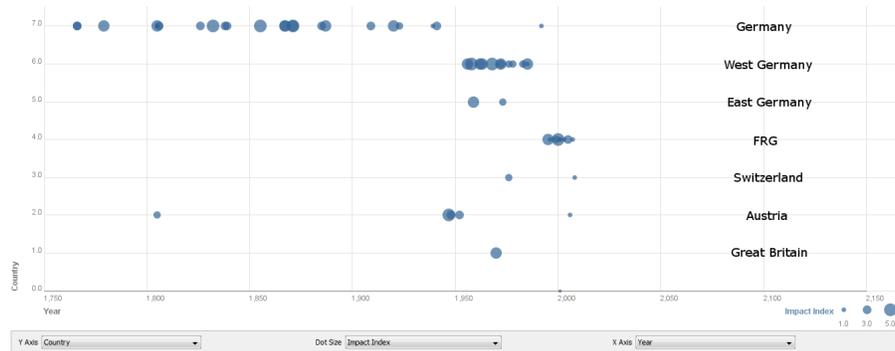
Fig. 1. This image illustrates the distribution of our collected German Othello translations. The X-axis is mapped to the publication date and Y-axis to seven different countries. The dot size is mapped to the impact index. A larger radius depicts a translation with higher re-publishing figures.

each *Othello* translation. Figures were derived from the standard bibliography of Shakespeare in German [12]. The index has five levels ranging from 1 to 5, where 1 means that the translator is not listed in the bibliography and 5 means that more than 50 publications and re-publications by the translator are listed in the bibliography. Figure 1 shows the chronological distribution of our collected documents. The X-axis is mapped to the publication date and Y-axis to the different countries. The ellipse radius is mapped to impact index.

4 Text Preprocessing

Before the original translation can be analyzed within our visualizations, we need to generate various features from the textual information and transform them into numerical vectors. In this work, we process our original text in five steps, namely document standardization, tokenization, stemming, vector generation and similarity calculation. The major outputs include making concordance of each document and computing their similarity.

Since the *Othello* translations are collected from various sources (some PDF, some archival typescripts, mostly books), we firstly transform and integrate them into a standard XML format. Next, document tokenization breaks the stream of text into a list of individual words or tokens. During this process, common words carrying little meaning which are not of interest to domain experts, such as "der" (the), "da" (that) etc, are eliminated from the token list. Furthermore, stemming reduces all of the tokens to their root forms. Based on this cleaned and standardized token list, we are able to generate a concordance table for each document by counting the frequency of every unique token.

For in-depth document comparison, we also need an objective document similarity measure. The domain experts from Arts and Humanities suggest a list of high-frequency keywords as a search query. This keyword list can be extracted from multiple interesting documents. The similarity between our collected translations can then be measured

using the LSI (Latent Semantic Index) model [13]. This model is widely used in information retrieval where the list of terms associated with their weight is treated as the document vectors. The weight of each term indicates its importance in a document, and is given by $Tf \times Idf$. We use Tf (Term Frequency) to refer to the number of times a term occurs in a given document, which measures the importance of a word in a given document. Idf (Inverse Document Frequency), as its name implies, is the inverse of the Document Frequency. The Document Frequency is the number of documents in which a word occurs within the collection of documents.

Thus the weight of a term i in document j can be defined as:

$$w_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times log\frac{N}{df_i}$$

where N is the total number of documents in the corpus, df is the document frequency and idf is the inverse document frequency. Large values of $w_{i,j}$ imply term i is an important word in document j but not common in all documents N.

Then a document j can be represented as a vector with each dimension replaced by the term weight:

$$D_j = (w(0, j), w(1, j), ..., w(n, j))^{\mathbf{T}}$$

A large number of words in the search query might lead to extremely high-dimensional document vector, so we use the SVD (Singular Vector Decomposition) to perform a dimension reduction. Then the similarity between the two documents j and k can be measured by the angle between these two vectors:

$$\cos\ Sim(D_j, D_k) = \frac{D_j \cdot D_k}{|D_j||D_k|}$$

Such similarity measures are generated for all of our *Othello* translations. This information is featured in our treemap and parallel coordinates.

5   Structure-aware Treemap

As discussed in Section 3, metadata of each document includes author name, play title, date, place of publication and impact index. The scatterplot in Figure 1 is able to present the overall historical distribution, but it cannot provide an aggregation of the data. For example, if the user wants to explore or rank the total number of translations, or the total number of re-publications in any century, decade or country in our document collection, the scatterplot is unable to convey an answer. Next to this, we observe that the meta data can be arranged in a hierarchical structure. For example, each century breaks down into several decades. In each decade a few translations are published in several countries. In each country several authors published their work. For each author his translations have the impact index. Given this structure, we are able to generate a Treemap [14, 15] visualization.

The traditional treemap is able to compare the node values in any tree level. But it lacks the ability to show the entire tree structure intuitively. For tracing the treemap
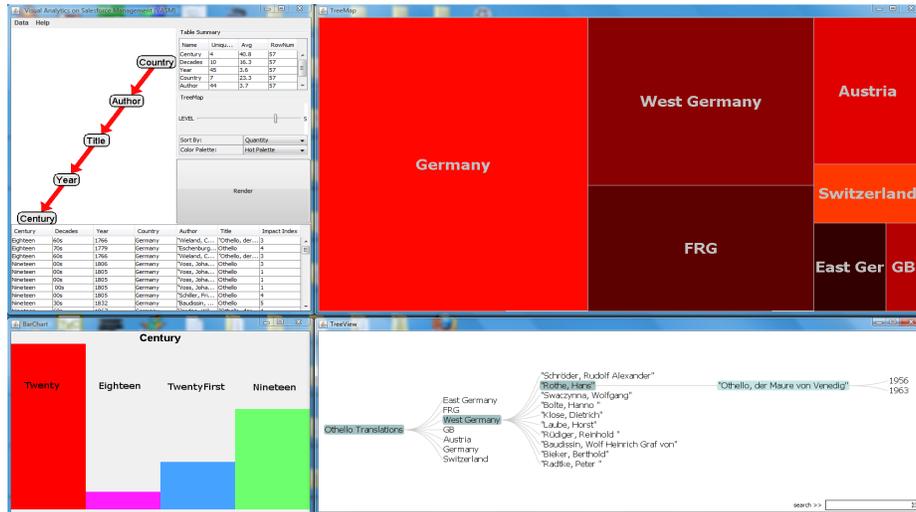
Fig. 2. This image illustrates the interface of our structure-aware treemap. The left part shows the control panel by which the user is able to manipulate the tree hierarchy, compare the values in each hierarchy via a bar chart and set up the configuration for the visualization. Also the user is able to select their interesting documents from the spreadsheet. The right part shows the treemap and DOI-tree. The area of the leaf node is mapped to the quantity. As we drill down and up to different tree levels, the DOI-Tree keeps track of the structure. Also, the DOI-tree could initiate a searching task.

hierarchy, it's necessary to only list the relevant substructure which shows the ancestor and descendants of the interested node. The Degree-of-Interest tree [16] provides a clear hierarchy at a low cost of screen space by changing the viewpoint and filtering out the uninteresting tree nodes. In addition, it offers instant readability of the node labels. Therefore, we adopt linked views using both DOI tree and treemap to enable structure tracing. Our system is composed of two parts, namely the control panel and structure-aware treemap. The control panel is shown on the left half of Figure 2. It extracts the ontological hierarchy information from the input data sets and sets up the configuration for the visualization. The user is able to change the order of hierarchy or reduce the number of hierarchies by moving the graph nodes. The right half of Figure 2 is a structure-aware hierarchical visualization, containing the coordinated views of the squarified treemap and DOI tree [16]. As we traverse back and forth between the intermediate levels of the treemap, the DOI tree view clearly keeps track of how each selected node is derived from its ancestors.

The area of the leaf node can be either mapped to the impact index, the similarity measure or the quantity. In Figure 2, from the bar chart, we learn that most of our collected translations were published in the twentieth century. During this century, most translations are published in the 1940s and 1970s. For the domain specialists, this raises questions about possible correlations with comparable datasets (translations of other or all Shakespeare plays), and about possible correlations between periods in German history, and specific interest in Othello. By changing the hierarchy, we also learn that
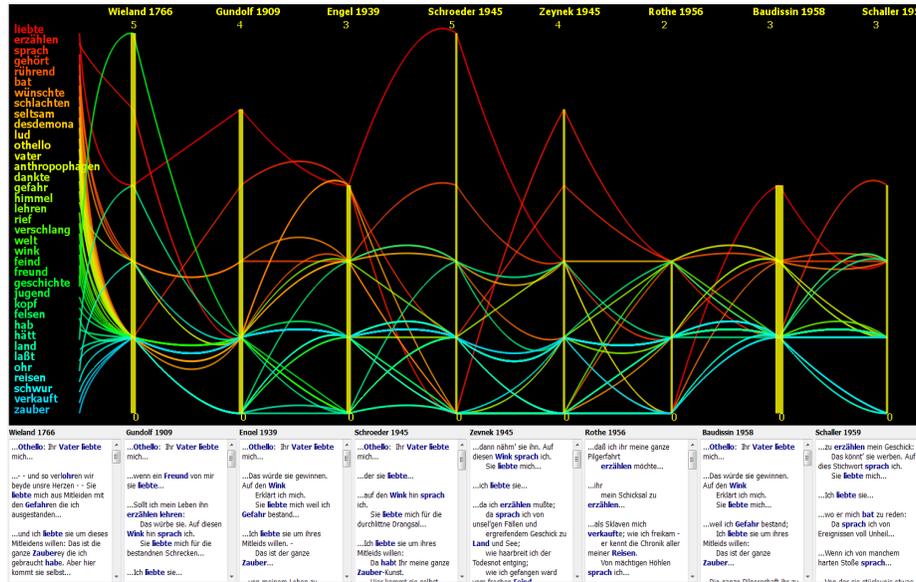
Fig. 3. This image shows an overview of our visualization. The parallel coordinates illustrates a focus view of the term frequency. The text boxes below the parallel coordinates show the context views. They present the entire sentences from the original text where each keyword appears.

although the documents are all translations of *Othello*, they have different titles: the commonest titles of the translations are "Othello" or "Othello, der Mohr von Venedig", some authors use the title "Die Tragdie von Othello, dem Mohren von Venedig", two use "Othello, der Maure von Venedig" and one author uses the title "Othello, Venedigs Neger". These outliers are particular interest to the domain experts.

Our treemap system helps users manage their documents, such as ranking the documents according to different criteria, analyzing the global features of the metadata and selecting the interesting documents. It can be scaled up to include new datasets such as translations of other works by Shakespeare and enable users to explore common patterns in the metadata. The DOI tree can initiate the searching task by which a user is able to search terms in any hierarchy. Since the collection of our German translations is still expanding, our treemap will play an increasingly important role in the meta data analysis.

6 Focus+Context Parallel Coordinates

Parallel coordinates, introduced by Inselberg and Dimsdale [17, 18] is a widely used visualization technique for exploring large, multidimensional data sets. It is powerful in revealing a wide range of data characteristics such as different data distributions and functional dependencies [19]. As discussed in Section 4, the textual information of each document can be transformed into a vector. In our parallel coordinates, we encode the document dimensions as term frequencies.
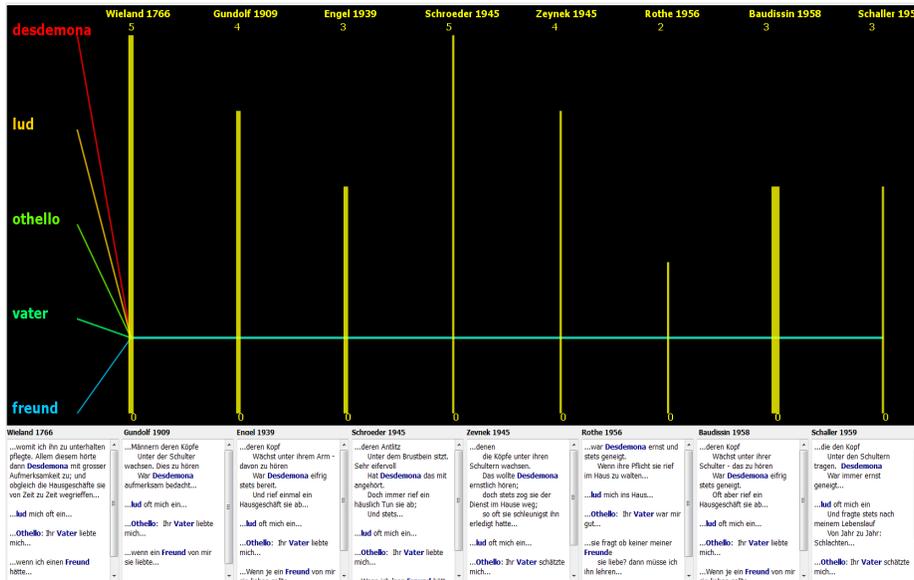
Fig. 4. In this image, we obtain five keywords which only appear once in all documents.

Domain experts from Arts and Humanities selected eight interesting translations according to their similarity score. For initial analysis, we chose a significant passage from the play, Othello's big speech to the Venetian Senate in Act1, Scene3: the longest single speech in the play (about 300 words in Shakespeare's text). Figure 3 shows an overview of our visualization. The column on the far left displays a list of selected keywords: these are most frequently occurring significant words in the document corpus. The parallel coordinates present a focused view of keyword frequencies. Each document is represented by a vertical axis. In order to maintain a unified scale, the height of each vertical axis is made proportional to the range between each document's minimal and maximal word frequencies. Zero frequency simply means that a keyword has not occurred in that document. The thickness of each vertical axis is mapped to the document's similarity with others in terms of LSI score: a thicker line means a higher similarity value. The number of occurrences of each keyword in each document is connected by a polyline. Each polyline is rendered in a different color to enable visual discrimination. The text boxes below the parallel coordinates provide context views for keywords selected by the user. Each text box represents an individual document and shows the entire sentences from the original text where each selected keyword occurs. We also apply the edge bundling to enhance the visual clustering and user is able to control the curvature of the edge [20]. Curves with the least curvature become a straight line.

We provide various interaction support, such as selection, brushing and linking. As the user selects individual or multiple keywords, the corresponding polylines are rendered. The user can also select various frequency levels in any document and the
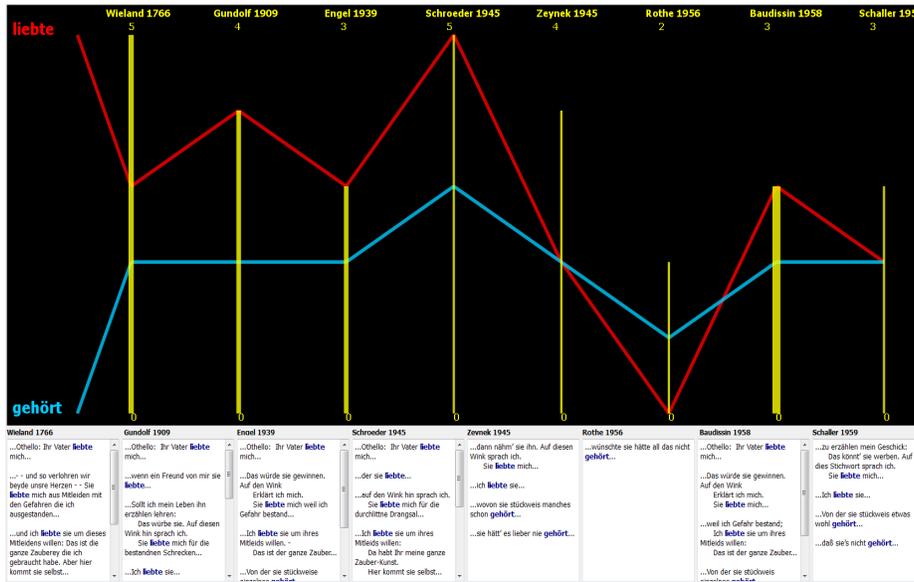
Fig. 5. In this image, there are two keywords showing a strong correlation.

corresponding keywords having that frequency are displayed. Along with the selection and brushing, the text boxes which show the context views keep updating.

Our system also supports composite brushing such as an AND-bush or OR-bush [21]. We can use the AND-Brush to obtain all keywords which occur in every document: words used by all translators regardless of the translators' reputations and impact. If we brush the keywords which do not appear in document "Baudissin 1958", we learn that this document contains all the keywords except "fand". This helps to explain why this document has the highest similarity score. The domain experts indicates that this finding is surprising and interesting. As shown in Figure 4, we observe five keywords which appear just once in all the documents. From the context views, the sentences containing these two words are almost the same in every translation. As shown in Figure 5, there are two keywords showing a strong correlation. Both findings raise interesting questions for the domain experts.

7 Domain Expert Reviews

The focus+context parallel coordinates permits comparative visualization and exploration of concordances. A concordance is normally displayed as a simple list of words in a vertical column (in order of frequency or alphabetically). Standard concordance software also offers the option to display contexts of use for a particular word (i.e. the different word strings in which a word appears). This tool successfully combines a concordance-derived keyword list and context views with display of frequencies of words across multiple, comparable versions, in the form of parallel coordinates. This is a promising way of exploring texts through their different uses of meaningful words. In

the display of parallel coordinates, the composite brushing enables us filter for any correlations between word-uses, positive or negative: pairs/groups of words which appear together, or never appear together. The similarity of each document tells us an objective measure of how similar each document is to the keyword lists. In this particular case, the visualization tells us that Baudissin's translation-which is the standard, most often republished and performed German translation of the play - contains the most keywords in this speech which are common to most of the other translations. Since other translations are produced and marketed as "alternatives" to Baudissin, this high degree of apparent dependency on the standard translation is surprising, and it demands further investigation.

Our current corpus of German *Othello* translations is relatively small (under 60 documents), but we envisage it growing: in respect of other works (Shakespeare's many other plays, and poems; and potentially works by other writers) and also in respect of other languages of translation (at least one of Shakespeare's works exists in about 100 languages). Hence, the flexible metadata overview offered by the structure-aware Treemap visualization will become increasingly valuable in managing the dataset, exploring its various dimensions and selecting subsets of translations for further analysis.

8 Conclusion

In this paper, we describe an interactive visualization system for presenting, analyzing and exploring the variation among different German translations of Shakespeare's play, *Othello*. A structure-aware treemap is developed for metadata analysis and the focus + context parallel coordinates is developed to investigate the variations among the translations. Our parallel coordinates incorporate an objective similarity measure for each document using LSI model. Also, various interaction supports are realized to facilitate the information seeking mantra: overview first, zoom and filter and detail on demand. Our visualization is evaluated by the domain experts from Arts and Humanities. Because it is just the beginning of our project, in the future, we would like to add more advanced features to the parallel coordinates, such as visual clustering. Also, we will keep on collecting more translations. Further statistical and linguistic analysis will be implemented.

Acknowledgments

References

1. B.Scott., G.Carl., N.Miguel.: Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. In: HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, New York, NY, USA, ACM (2008) 193–202

2. Viegas, F.B., Wattenberg, M., Feinberg, J.: Participatory Visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics **15** (2009) 1137–1144
3. Wattenberg, M., B.Viegas, F.: The Word Tree, an Interactive Visual Concordance. IEEE Transactions on Visualization and Computer Graphics **14** (2008) 1221–1228
4. van Ham, F., Wattenberg, M., Viégas, F.B.: Mapping Text with Phrase Nets. IEEE Transactions on Visualization and Computer Graphics **15** (2009) 1169–1176
5. Paley, W.B.: TextArc: An Alternative Way to View Text (2002) `http://www.textarc.org/`, Last Access Date: 2011-2-18.
6. Collins, C., Carpendale, M.S.T., Penn, G.: DocuBurst: Visualizing Document Content using Language Structure. Computer Graphics Forum **28** (2009) 1039–1046
7. Koh, K., Lee, B., Kim, B.H., Seo, J.: ManiWordle: Providing Flexible Control over Wordle. IEEE Transactions on Visualization and Computer Graphics **16** (2010) 1190–1197
8. Mehta, C.: Tagline Generator - Timeline-based Tag Clouds (2006) `http://chir.ag/projects/tagline/`, Last Access Date: 2011-2-18.
9. Collins, C., B.Viegas, F., Wattenberg, M.: Parallel Tag Clouds to Explore and Analyze Facted Text Corpora. In: IEEE Symposium on Visual Analytics Science and Technology, IEEE Computer Society (2009) 91–98
10. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Computer Graphics **8** (2002) 9–20
11. Lee, B., Riche, N.H., Karlson, A.K., Carpendale, M.S.T.: SparkClouds: Visualizing Trends in Tag Clouds. IEEE Transactions on Visualization and Computer Graphics **16** (2010) 1182–1189
12. H., B., Schmidt, W.: Shakespeare deutsch: Bibliographie der bersetzungen und Bearbeitungen. (2003)
13. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41** (1990)
14. Johnson, B., Shneiderman, B.: Tree Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In: IEEE Visualization. (1991) 284–291
15. Shneiderman, B.: Tree Visualization With Treemaps: a 2-d Space-filling Approach. ACM Transactions on Graphics **11** (1992) 92–99
16. Card, S.K., Nation, D.: Degree-of-Interest Trees: A Component of an Attention-Reactive User Interface. In: WORKING CONFERENCE ON ADVANCED VISUAL INTERFACES (AVI). (2002) 231–245
17. Inselberg, A., Dimsdale, B.: Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In: In proceedings of IEEE Visualization. (1990) 361–378
18. Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer (2009)
19. Keim, D.A., Kriegel, H.P.: Visualization techniques for Mining Large Databases: A Comparison. IEEE Transactions on Knowledge and Data Engineering **8** (1996) 923–938
20. Zhou, H., Yuan, X., Qu, H., Cui, W., Chen, B.: Visual Clustering in Parallel Coordinates. Computer Graphics Forum **27** (2008) 1047–1054
21. Hauser, H., Ledermann, F., Doleisch, H.: Angular Brushing of Extended Parallel Coordinates. In: In proceedings of IEEE Symposium on Information Visualization, IEEE Computer Society (2002) 127–130