Visual Readability Analysis: How to Make Your Writings Easier to Read

Daniela Oelke* University of Konstanz
David Spretke† University of Konstanz
Andreas Stoffel‡ University of Konstanz
Daniel A. Keim§ University of Konstanz

ABSTRACT

We present a tool that is specifically designed to support a writer in revising a draft-version of a document. In addition to showing which paragraphs and sentences are difficult to read and understand, we assist the reader in understanding *why* this is the case. This requires features that are expressive predictors of readability, and are also semantically understandable. In the first part of the paper, we therefore discuss a semi-automatic feature selection approach that is used to choose appropriate measures from a collection of 141 candidate readability features. In the second part, we present the visual analysis tool *VisRA*, which allows the user to analyze the feature values across the text and within single sentences. The user can choose different visual representations accounting for differences in the size of the documents and the availability of information about the physical and logical layout of the documents. We put special emphasis on providing as much transparency as possible to ensure that the user can purposefully improve the readability of a sentence. Several case-studies are presented that show the wide range of applicability of our tool.

Index Terms: I.7.5 [Document and Text Processing]: Document Capture—Document Analysis; I.5.2 [Pattern Recognition]: Design Methodology—Feature evaluation and selection

1 MOTIVATION

A common challenge when producing a text, is to write it in a way that it is easy to read and understand by the target community. This includes aspects like ensuring contextual coherency, avoiding unknown vocabulary, difficult grammatical structures, or misspellings. In this paper, we are introducing the tool *VisRA* that is specifically designed for supporting the writer in the task of revising a text. After loading a text, VisRA gives the user detailed feedback about passages and sentences that may be difficult to read and understand. The feedback not only points to problematic sentences, but it also identifies and explains the reason(s) *why* this sentence may be difficult to read. This allows an efficient and effective revision of the written document.

There are several basic aspects of readability. Primarily these are problems in linguistics and content-wise difficulties. Consider for example the sentence "I think, therefore I am". It is not difficult to understand the sentence in terms of vocabulary or grammar, but content-wise, it requires some deeper thought. In addition, the readability of a document is also influenced by the contextual coherence[1] and consistency[2] as well as the print layout of a page.

In this paper, we concentrate on features that measure the first two aspects of readability (linguistic and content-wise appropriateness). A special challenge in our application scenario is issued by the need for features that are a) semantically understandable and b) at the same time allow for a detailed analysis of the text with respect to the reasons for the observed difficulties. Section 3 discusses how we find appropriate features from a large set of candidates using a semi-automatic feature selection approach.

Section 4 introduces the VisRA tool. The tool is designed in a way that it is easy to see the characteristics of the features across the document, while at the same time it identifies the single paragraphs and sentences that are most in need of being revised. Visualization techniques support the user in the analysis process and are employed to convey the information about why a sentence or paragraph is difficult to read and/or why it cannot be understood effectively. Finally, the case studies in section 5 show the wide range of applicability of our tool.

2 RELATED WORK

2.1 Readability Analysis

Several well known formulas to measure readability exist. Among the most popular ones are the Flesch-Kincaid Readability Test [19], Flesch Reading Ease [12], SMOG [21], the Coleman-Liau-Index [8], and Gunning Fog [13]. It is common to all these measures that they are solely based on statistical properties of the text, such as word length (either measured as the number of syllables or the number of characters), the number of words in a sentence / paragraph, and the number of easy and hard words. A word is considered as "hard", if it consists of three or more syllables or alternatively, if it is not contained in a list of easy words. The most severe disadvantage of these methods is that the calculated value does not allow the user to conclude what exactly has to be changed to improve the readability of the text.

Other approaches take more aspects of readability into account. For example, [17] and [23] consider the syntactic complexity with the help of features like the depth of the parse tree of a sentence or the number of sentences in passive voice. In both papers the vocabulary usage is taken into account with a statistical language model to avoid the need for a vocabulary list, same as [9] and [24] do. The difficult problem of measuring how coherent a text is, is tackled in [3]. Their approach is based on the assumption that the distribution of entities can be used to defer information about the local coherence of the text. Additionally, [22] takes discourse relations into account to measure text coherence and show that they are good predictors of readability (comparing them to several other readability features). However, their method requires the discourse annotation, since so far, it cannot be determined automatically. [5] analyzes if syntactical surface statistics are good predictors for sentence fluency.

In contrast to the above mentioned methods, we do not make assumptions about what features might be good predictors for readability. We prefer to start with a high number of features and let automatic algorithms decide what the best predictors are. Furthermore, our goal is to provide the user with a tool that guides the improvement of the text within the scope of special requirements, in which we need features that are semantically understandable.

*e-mail: oelke@inf.uni-konstanz.de
†e-mail: spretke@inf.uni-konstanz.de
‡e-mail: stoffela@inf.uni-konstanz.de
§e-mail: keim@inf.uni-konstanz.de

[1] "The coherence of a text is the degree to which the reader can describe the role of each individual sentence (or group of sentences) with respect to the text as a whole." [7]

[2] Consistency in this case can be interpreted as being in agreement or harmony with what has already been set as well as always following the same style.

2.2 Document Visualization

In our tool, we want to show the user not only which passages are difficult to read, but also demonstrate why they are less readable. We therefore need a visualization technique that permits a user to analyze a document in detail, rather than using approaches that were intended to support the browsing or summarization of large document collections. Related approaches that meet this requirement are reviewed below.

Literature Fingerprinting [18] is a technique that depicts each text unit with a single pixel and visually groups them into higher level text units. Color is mapped to a specific text feature allowing for a detailed analysis of the text. We use this technique in our tool in the overview and navigation panel. Closely related to the Literature Fingerprinting technique are the visualizations that were introduced in [2, 16, 10]. *Seesoft* [2] has been designed for the visual analysis of program code, depicting each line of code with a (proportionally scaled) line in the diagram. We employ the approach in one of our overview representations. The intention of *TileBars* [16] is to provide a compact and meaningful representation of Information Retrieval results, whereas the *FeatureLens* technique, presented in [10], was designed to explore interesting text patterns, find meaningful co-occurrences of them, and identify their temporal evolution.

Beyond that, the visualization techniques *Ink Blots* [1] and the system *Compus* [11] have to be mentioned as examples of detailed text visualizations. In contrast to the other techniques, both the Ink Blot technique and Compus visualize a multitude of features at once in a single visualization by accepting much overplotting. As a result, they cannot cope with features that provide values for each single text unit (at least not without giving up their claim to visualize multiple attributes at once).

A different visualization technique for documents are thumbnails of document pages. They are used to give an overview of the documents and to allow the user to navigate to a page or passage of interest. The *enhanced thumbnail* [28] is a combination of a plain thumbnail that preserves the gestalt of the page with keywords that describe the page content. A combination of the enhanced thumbnails with a detail view is presented in [26]. This combination allows efficient navigation in documents while having the details at hand. A different navigation approach is presented by the *space-filling thumbnail* system [6] that uses a space-filling placement of plain thumbnails for navigation and opens a detail view on demand. We incorporate the idea of the enhanced thumbnails in our tool, but instead of keywords we show the readability of passages to the user.

3 Finding Semantically Rich Readability Features

To provide the user with detailed feedback about *why* a passage in a document is difficult to read, we need a readability measure that is both semantically rich and expressive. Feature Selection can be considered a difficult problem in general. Sometimes common sense or expert knowledge is used to determine the right features. However, with such an approach it easily happens that features are ignored that do have a high expressiveness but are not commonly associated with the task. On the other hand, fully automatic feature selection techniques may end up with features that are semantically difficult to understand.

In this paper, we follow a semi-automatic approach. We start our search for text features that are expressive with respect to readability with a large initial feature set to ensure impartiality. First, automatic feature selection methods are applied to determine measures that are expressive with respect to readability. Second, redundancy between the features is detected with the help of correlation measures. The user is incorporated in the last step of the feature selection process. By manual inspection of the feature classes, it is much more likely that semantically meaningful features are selected.

Note that it is critical to start with a feature set that is as exhaustive as possible. Aspects that cannot be measured with the provided initial feature set will be missed in the process. To alleviate this problem, it is advisable to work in close collaboration with an expert who could eventually identify aspects which might trigger an iterative analysis process. The feature selection process can be considered as a one-time effort, although some features require adaptation to the target community (see section 3.1).

3.1 Initial set of text features

Our goal was to search in a manner that is as unbiased as possible for text features that are expressive with respect to readability. We therefore implemented 141 different text features which can be classified into the following categories:

- *Features that are based on word classes:* After a text has been part-of-speech tagged (using the Stanford POS Tagger [27]), the frequencies of the different word classes (such as nouns, verbs, pronouns, etc.) are calculated. Furthermore, the ratio between different word classes is taken into account.

- *Features that are based on word frequencies:* Large document collections such as the Project Gutenberg (`http://www.gutenberg.org/`) or Wikipedia (`http://www.wikipedia.com`) make it possible to calculate the average usage frequency of a word. We exploited those resources to determine how common the words of a text sample on average are. This was done on different granularity levels, taking the most frequent 50, 100, 500, 1000, or 2000 words into account. In some application scenarios, it is more appropriate to determine the most frequent terms on a domain-dependent collection. The rationale behind this is that even words that are difficult to understand in general may be well-known within a specific community and therefore appropriate to use in such a context. Since we analyze documents from the visual analytics community in two of our case studies, we additionally calculated term frequencies on a collection of VAST and InfoVis papers of previous years.

- *Features that analyze the sentence structure:* Besides measuring the sentence length, we implemented features that are based on the phrase structure tree[1] of a sentence as determined by the Stanford Parser [20]. Features such as the depth of the phrase structure tree, its branching factor or the position of the verb were implemented to take the grammatical structure of a sentence into account.

- *Others:* In addition to the aforementioned features, several other features were implemented, e.g. measuring the number of quotations in a text or the number of sentences in passive voice.

The selection of appropriate features is performed in a two step process. First, the feature set is reduced by removing all features that only show a low expressiveness with respect to the text property readability. Second, a set of semantically meaningful, non-redundant features is being determined.

3.2 Step 1: Removing features with low expressiveness with respect to readability

Using a ground-truth data set of text samples that include examples that are both very easy and very difficult to read, features that show no or only a very low expressiveness with respect to readability are filtered out. The necessary ground-truth data set is compiled of a collection of books for children (most of them are rated as

[1] A phrase structure tree is a hierarchical representation of a sentence that is build according to the nesting of its (sub)phrases.

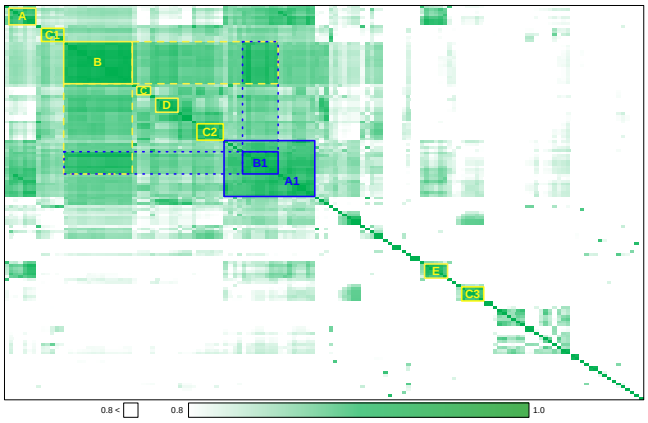


Figure 1: Correlation matrix of the features remaining after removing the ones with low expressiveness. As can be seen, some features are highly correlated to each other measuring the same aspect of readability.

being suitable for children aged 4 to 6) and the work program of the FP7 initiative[2]. Note that the two data sets were arbitrarily chosen. Because we only conservatively discard features in this step of the process, the choice of samples is not that critical as long as the two sets are clearly discriminating with respect to readability.

The aforementioned documents are split into text samples of about 1000 words each. Next, the 65 samples that are rated by the Flesch Reading Ease Measure [12] and the easiest and the most difficult ones are chosen to be a part of the training data set. For each of the 141 features and 130 text samples a normalized value between 0 and 1 is calculated, resulting in a 130 dimensional vector for each feature. To determine the discrimination power of each features, the Pearson Correlation Coefficient is calculated assuming that the ideal feature should rate all FP7 documents as 1 and the samples that are taken from children's literature as 0. Only features that score at least 0.7 in this test are kept (which is about 40% of all features).

3.3 Step 2: Selecting semantically meaningful, non-redundant features

After filtering out all features that show a low discrimination power with respect to the two classes, we select appropriate features that a) are semantically meaningful and b) are non-redundant (i.e. do not measure the same aspect of readability). Using again the Pearson Correlation Coefficient, the correlation factors between all possible feature pairs are calculated. To detect features that highly correlate with each other, we resort the rows and columns of the resulting correlation matrix with the help of a hierarchical clustering algorithm. Furthermore, the cells of the matrix are colored according to the value that they represent (starting with values ≥ 0.8, see color scale in figure 1). Next, the clusters are manually inspected to find out which semantic aspect they measure. For each cluster, one feature is chosen as a representative. If there is no common semantic aspect, the feature is chosen that is easiest to understand. "Easy to understand" in this case means that the feature must be consciously controllable when writing a text, allowing an analyst to improve the readability of a sentence with respect to this feature.

In figure 1, clusters from which features were chosen are marked in yellow. Cluster B_1 was dismissed because of its strong correlation to cluster B (see overlap area of dashed lines). The same is true for A_1 which correlates with A. Interestingly, the clusters C, C_1, C_2, and C_3 contain features that are semantically similar (different variants of measuring nominal forms), but despite this, no strong correlation can be perceived. Features that are not distinguishable on a semantic level do not help the user when refining a text. We therefore decided to choose one feature from each cluster but to present only the one with the highest score to the user. Cluster D summarizes features that measure how common the used vocabulary is (in comparison to a reference corpus). Finally, cluster E contains features that measure the sentence structure complexity.

3.4 Resulting Feature Set

Finally, the following features were selected:

- *Word Length* (cluster B): Measured as the average number of characters in a word.
- *Vocabulary Complexity* (cluster D): Measured as the percentage of terms that are not contained in a list of common terms. These terms are either defined as the 1000 most frequent terms in a large document collection of the specific language (the so-called basic vocabulary of the language)[3] or are determined from a set of documents of the specific domain (in this case VAST/InfoVis papers).
- *Nominal Forms* (clusters C-$C3$): This is a combined measure (see section 3.3) consisting of features that take the noun/verb ratio and the number of nominal forms (i.e. gerunds, nominalized words (ending with *ity, ness, etc.*) and nouns) into account.
- *Sentence Length* (cluster A): Measured as the number of words in a sentence.
- *Sentence Structure Complexity* (cluster E): Measured as the branching factor in the phrase structure tree of a sentence. This measure is related to the one proposed in [29]. It follows the assumption that the mental complexity of processing a sentence is increased if parts of the sentence are interrupted by subordinate sentences or parenthesis. In this case, the brain is forced to remember incomplete parts of the sentence.

All features are normalized with respect to sentence length and mapped between 0 and 1. We use the values that we observed for our ground-truth data set to determine the normalization factors for each feature. Figure 2 shows the three cases that are possible: (a) The values of the easy-to-read samples are clearly separated from the values of the difficult ones. (b) There is no separation at all between the two classes. (c) The observed values overlap each other, meaning that there is a range of values for which we cannot decide the class the text unit belongs to.

The features values are normalized in a way that the interval size for both classes is the same (e.g. one class between 0 and 0.4 and the other class between 0.6 and 1). The distance between the observed values of the two classes is accounted for by the size of the gap between the two intervals (see graphics and formulas in figure 2).

For the values of the easy-to-read samples a color scale from light blue (fairly easy) to dark blue (very easy) is used. Similarly, values in the interval of the difficult samples are colored in shades of red. Values in between the two intervals are colored in white if there is a clear separation between the two classes, and in grey if both classes overlap (see color scales in figure 2).

[2]FP7 stands for the *Seventh Framework Programme for Research and Technological Development* of the European Union, whose work programs are generally agreed on as being difficult to read.

[3]As an English word list we use [14] (based on Project Gutenberg), our German word list is [15] (calculated on a corpus of newsarticles).

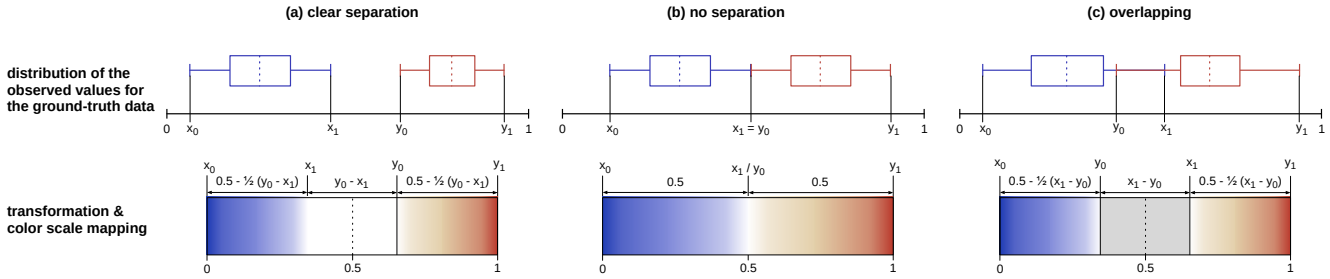


Figure 2: Normalization of the feature values is done relatively to the values that we observed for our ground-truth data set. The graphic shows the formulas and color scales for the 3 different cases that are possible.

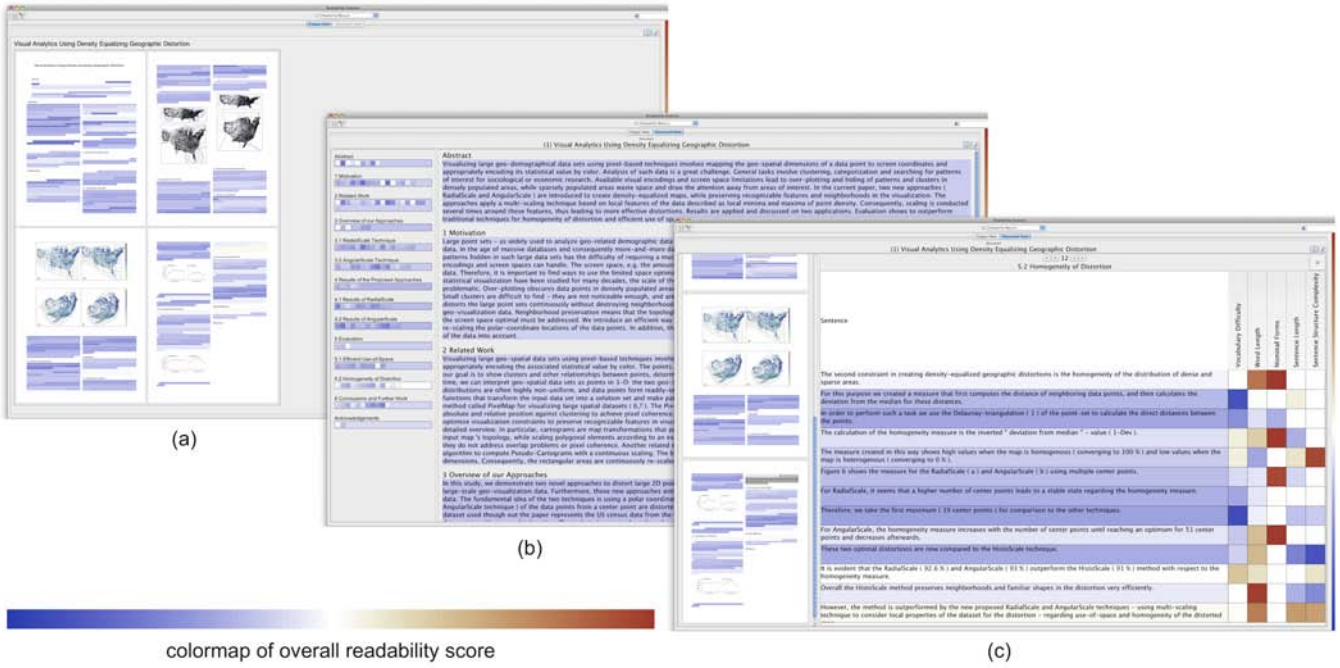


Figure 3: Screenshot of the VisRA tool on 3 different aggregation levels. (a) Corpus View (b) Block View (c) Detail View. To display single features, the colormap is generated as described in section 3.4 and figure 2.

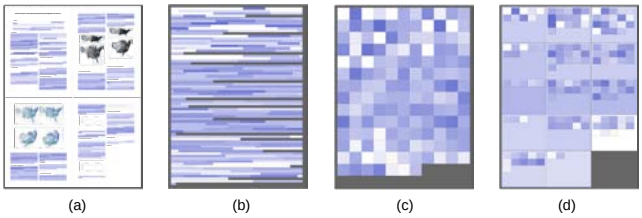


Figure 4: Embedded representations: (a) *Structure Thumbnails*, (b) *Seesoft representation* and (c, d) *Literature Fingerprinting representation*.

3.5 The Readability Measure

Central to the concept of our tool is to provide the user with a detailed view allowing him or her to determine why a specific sentence is difficult to read. However, in the overview representations we still need a single value for each sentence or paragraph that guides the user to the sections that need a closer inspection. We therefore calculate the average of the different features as an overall readability score.

4 VisRA - A Tool for Visual Readability Analysis

Figure 3 shows a screenshot of the VisRA tool. Three different views are available: The Corpus View (figure 3(a)), the Block View (figure 3(b)), and the Detail View (figure 3(c)).

4.1 The Corpus View

The corpus view (see figure 3(a)) serves as an overview representation. In this view, each document is represented by a rounded rectangle whose color is mapped to the overall document score. Within such a document thumbnail, the development of the feature values across the document is indicated by an embedded visualization. Some of these visualizations make use of the internal structure of the document (e.g. chapters and sections) and/or the physical layout of the pages. If no structure is available, the document is split into equal-sized blocks of text whose size may be determined by the user. Depending on the type of document (corpus) that is to

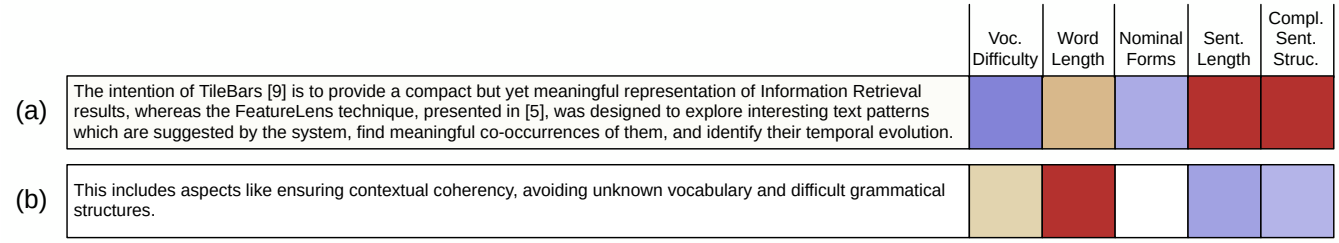


Figure 5: Two example sentences whose overall readability score is about the same. The detail view reveals the different reasons why the sentences are difficult to read.

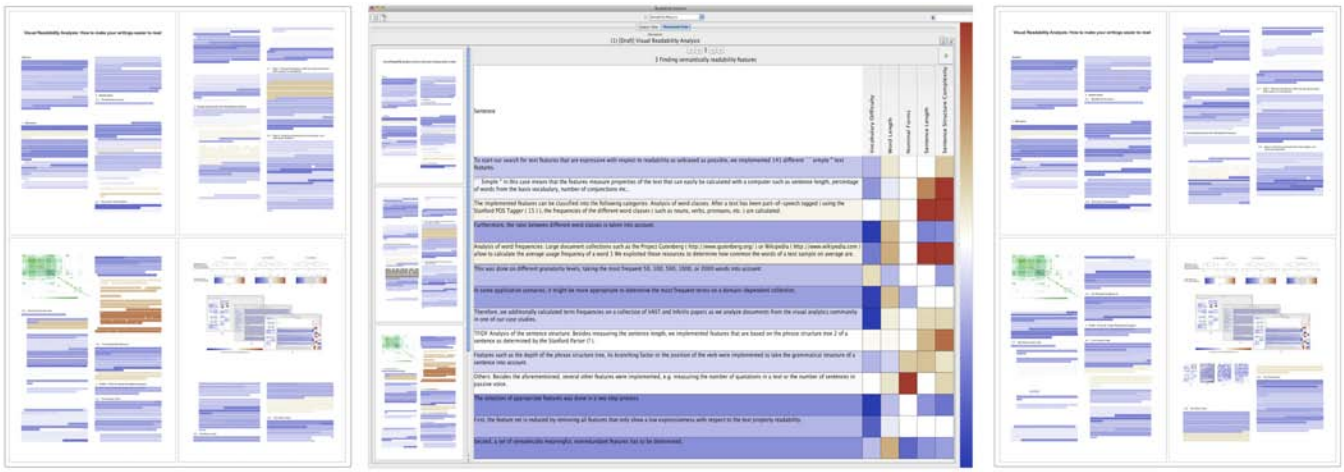


Figure 6: Revision of our own paper. (a) The first four pages of the paper as structure thumbnails before the revision. (b) Detail view for one of the sections. (c) Structure thumbnails of the same pages after the revision.

be analyzed, the user can choose between three different embedded representations:

- *Structure Thumbnails*: If the structure and the print layout of the document(s) are known, structure thumbnails can be employed (see figure 3(a) and 4(a)), including as many details as possible.
- *The Seesoft representation*: If the print layout is unknown, a representation like the one suggested in [2], which represents each sentence as a line whose length is proportional to the sentence length, may be suitable (figure 4(b)).
- *The Literature Fingerprinting representation*: As suggested in [18], each text unit (e.g. a section/block or a sentence) is represented by a single square that is colored according to the calculated feature value. The size of the squares is chosen in a way that the whole document can be displayed at once (see figure 4(c)). If enough space is available, big rectangles are used instead of squares to visualize the blocks and the sentence level is shown within them using small squares to depict a sentence (figure 4(d)). This technique is the most scalable one of the three, allowing to provide an overview even for large documents, respectively to show several documents at once on the screen.

4.2 The Block View

In this intermediate level, complete blocks or sections are displayed and are colored with the overall score of this section / block (see figure 3(b)). In contrast to the corpus view, the text is already readable in this view allowing the user to choose the section that is most in need of being revised. Both, the block view and the detail view offer a navigation panel at the left which can be used to locate the position of the displayed text in the document and to select a specific region for further analysis. Again, the user can choose between two different representations, the Structure Thumbnails (see figure 4(a)) and the Literature Fingerprinting technique (see figure 4(c)+(d)). Depending on the type of analysis task, the size of the document, and the available information about the physical and logical document structure (see section 4.1 for an explanation of the two techniques) either one of them is more suitable.

4.3 The Detail View

In the detail view, each sentence is displayed separately (see figure 3(c)). The background color of a sentence is set to its overall readability score, calculated as the average of the 5 feature values. Alternatively, the user can choose to have only one of the features displayed. Next to each sentence, the values for each feature are shown separately allowing the user to investigate the reasons why a sentence was classified as difficult. For this step, the color scales of figure 2 are used, meaning that colors are assigned relative to the values that were observed for the very easy and very difficult text samples in the ground-truth dataset. Note that the design of this representation is similar to TileBars [16]. Hovering over one of the cells, triggers the highlighting of the parts of the sentence that contribute to the feature value in the sentence. For example, for the feature *Vocabulary Difficulty* all the words that were classified as difficult are underlined. This supports the user in understanding the rating which is especially important for features that are a bit

unstable with respect to the length of the text unit. Additionally, the sentences of a section can be sorted according to the readability score or one of the features. This is very helpful if the user's task is to increase the readability of the document, because sentences that are most in need of being revised are presented first. To help the user to locate a sentence within the section after resorting, the position of the sentence within the document is highlighted with a blue border in the navigation panel as soon as the user hovers over a specific sentence.

5 CASE STUDIES

In the following, several case studies are presented that show the wide range of applicability of our tool.

5.1 Advantage of detailed insight over a single score

Figure 5 shows two example sentences whose overall readability score is about the same. Only the detail view reveals that there are different reasons why the sentences are difficult to read. In figure 5(a), our tool detects a complex sentence structure whereas in figure 5(b) the high percentage of gerunds (verbs acting as nouns) is complicating the sentence. This exemplifies that the details that our tool provides are a clear benefit in the refinement process.

5.2 Revising our own paper

We also used the tool to revise our own paper. Figure 6(a) shows the structure thumbnails of the first four pages of the paper. The physical and logical structure of the paper was automatically extracted using the technique described in [25]. Lines with meta-data, such as the names of the authors, their affiliations, keywords, etc., are automatically filtered out. (Section) titles are presented in the flow of the document but are excluded from the analysis. The remaining sentences are colored according to their overall readability score. As can be seen, the readability of the paper is already quite good, but some passages clearly need a revision. Figure 6(b) shows section 3 of the paper in the Detail view. The fifth sentence from the top seems to need some revision as it is colored in red (for an enlarged version see figure 7(a)). We find out that the difficulty of the sentence is primarily caused by the fact that we forgot to set a period after the inserted footnote. By hovering over the sentence, it is highlighted in blue in the navigation panel at the left, which makes it easier to find it in the paper.

Figure 7 shows some more examples for problems that can be found with the tool. (a) This is an enlarged version of the sentence with the missing period that we discuss above. (b) In this case, the sentence was too long and its structure too complex. We split it into several separate ones and dissolved the nested sentences. (c) The main difficulty of this sentence was that we had nominalized several verbs and adjectives. We reformulated the sentence in such a way that wherever possible the verb and adjective forms were used. Although this lengthens the sentence, it can be processed easier by the brain, because fewer words need to be transformed back into their original form [4]. (d) We found a comment in German that we forgot to delete. (e) Interestingly, only a few sentences could be found that are difficult with respect to the used vocabulary in previous VAST proceedings. This confirms that the VAST conference is the proper venue at which to present our research. In addition to pointing us to some sentences in German (sentences registered as using uncommon words compared to the previous VAST papers), one of the sentences in the related work section was highlighted. Since the average VAST paper does not talk about readability measures, it cannot be expected that the terms used are known by the respective community, which means that they should be introduced properly.

Figure 6(c) shows the first four pages of the paper after the revision.

5.3 Revising a large document

When revising a large document such as a book, our thumbnail representation would not be scalable enough. Consequently, several visualization techniques can be chosen on every level of the tool, depending on the size of the document and the availability of information about its logical and physical structure. The figure at the right shows a screenshot of four chapters of a new book on data visualization like it is shown in the navigation panel. A total of about 170 pages are displayed, whereby each of the pixels represents one sentence of the book. It is easy to see that the book is very well written with respect to readability. Only a few sentences stand out as being difficult to read. Further investigation revealed that some of those sentences talk about an application domain to which the introduced visualization was applied. Our vocabulary difficulty feature registers this as an accumulation of many words that are uncommon in the visualization community. Additionally, the tool revealed some long sentences that might have better been split into two sentences.

5.4 Analyzing a corpus with election agendas

The VisRA tool cannot only be used for refining single documents, but also for a comparative analysis of several documents with respect to the different aspects of readability. Figure 8 shows eight election agendas from the elections of the German parliament in 2009. As an embedded visualization, we chose the Literature Fingerprinting technique on sentence level. This allows us to display the large data set on one screen, while still providing the necessary details.

In Figure 8(a) the average readability score is mapped to color. It can easily be seen that two of the election agendas are significantly shorter and easier to read than the rest of the documents (first two documents in the first row). Those are special versions that are provided by the parties *SPD* and *Die Linke* for people that are less proficient in reading. Interestingly, the normal election agenda of *Die Linke* (third one in the last row) is the second most difficult one. At first, we were surprised to see that this agenda is rated as comparably difficult to read.

A more detailed analysis with respect to the different aspects of readability revealed some of the reasons for this. Figure 8(b) shows how the sentences are rated with respect to the vocabulary difficulty. To determine if a word is common, the dictionary of the University of Leipzig is employed. Frequencies in this dictionary are based on a large corpus of news articles. Closer analysis of the election agenda of *Die Linke* revealed that a high number of socialistic terms were used in the text. This terminology is not common in German newspapers. As mentioned earlier, two election agendas were intended to be easy to read. Strikingly, these two agendas also contain difficult vocabulary. The detail view reveals that in those documents long words are broken up by inserting a dash ("-"). These words are most often compound words and characteristic to the German language (e.g. in genitive constructions). They are often broken up by dashes or hyphens in order to allow for better comprehension. However, these words cannot be found in the list of most frequent terms (since they are spelled differently now from the words provided in the vocabulary list), they are classified by the algorithm as uncommon. Long words are avoided at all costs in the special election agendas that are written in a easy to read language. This fact is reflected by the visualization of the average word length that is depicted in figure 8(c). It also explains the significant differences between the easy-to-read election agendas and the more difficult ones.

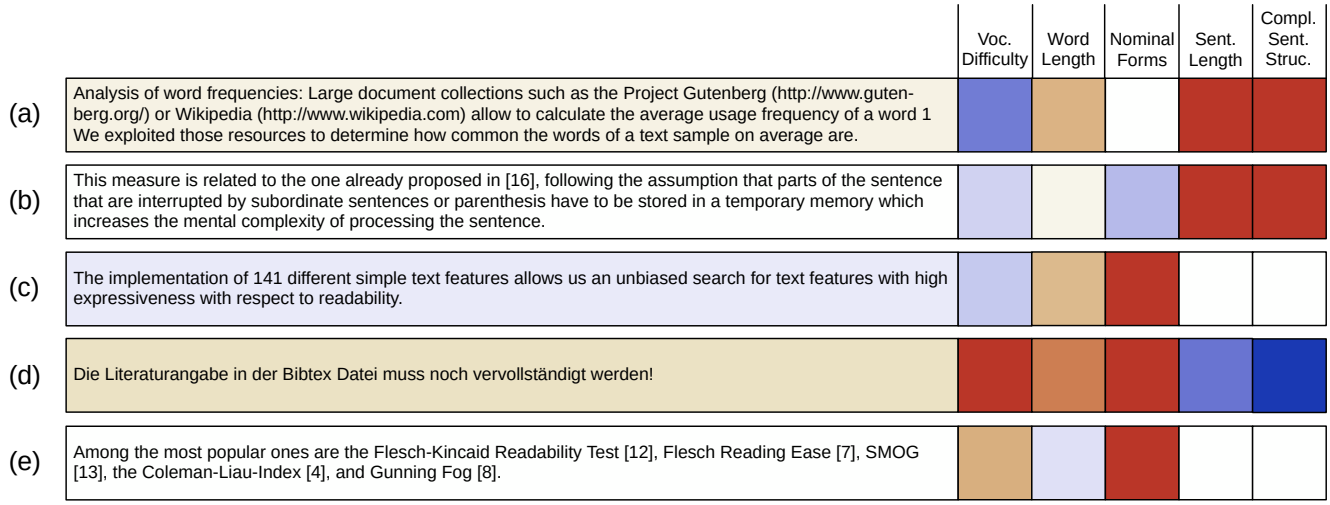


Figure 7: Examples for different reasons of difficulties that were found while revising our own paper with the VisRA tool. The detailed view reveals for each sentence what causes the difficulty. (a) A forgotten period. (b) Long and complex sentence structure. (c) Large number of nominal forms. (d) German comment that we forgot to delete. (e) Many terms that are uncommon in the VAST community.

Finally, figure 8(d) displays the feature sentence structure complexity. Obviously, all election agendas are well-formulated with respect to this property. Only single sentences are highlighted for which a revision might have been advisable.

6 Conclusions

In this paper, we introduced a tool for visual readability analysis that supports the writer in refining a document, and thereby to increase its readability. Special consideration was given to the selection of features that are non-redundant and semantically understandable. This is reflected in the design of the tool that provides insight into the data at several levels of detail. At the highest resolution, for every single sentence the values of the different features are displayed instead of only visualizing the average score. Several different overview representations account for differences in the size of the documents and the knowledge about the physical and logical structure of the document.

In the future, we plan to add additional features to the tool. For example, it might be interesting to include features that measure how appropriate the writing style of a document is or how well it is structured. Both measures are dependent on the domain or on the community, for which the document is written. Additionally, they would be asking for a calculation that compares the document to others in the same context. Furthermore, it would also be valuable to take measures into account that work on the discourse level and measure the consistency of the text. A user-study could be conducted to quantify the value of the different features. Finally, we envision enhancing the tool with natural language generation techniques to provide a written summary of the results.

Acknowledgements

This work has been partly funded by the German Research Society (DFG) under the grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, Konstanz.

References

[1] A. Abbasi and H. Chen. Categorization and analysis of text in computer mediated communication archives using visualization. In *JCDL '07: Proc. of the 2007 Conf. on Digital Libraries*, pages 11–18, 2007.

[2] T. Ball and S. G. Eick. Software Visualization in the Large. *IEEE Computer*, 29(4):33–43, 1996.

[3] R. Barzilay and M. Lapata. Modeling local coherence: an entity-based approach. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 141–148, 2005.

[4] M. Billig. The language of critical discourse analysis: the case of nominalization. *Discourse & Society*, 19(6):783–800, 2008.

[5] J. Chae and A. Nenkova. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *EACL '09: Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 139–147, 2009.

[6] A. Cockburn, C. Gutwin, and J. Alexander. Faster document navigation with space-filling thumbnails. In *CHI '06: Proc. of the SIGCHI Conf. on Human Factors in computing systems*, pages 1–10, 2006.

[7] Online resource of the Multilingual Information Processing Department at the University of Geneva, `http://www.issco.unige.ch/en/research/projects/isle/femti/html/182.html`, last accessed on 07/14/10.

[8] M. Coleman and T. Liau. A computer readabilty formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

[9] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proc. of HLT / NAACL 2004*, 2004.

[10] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *CIKM '07: Proceedings of the 16th ACM Conf. on Information and Knowledge Management*, pages 213–222. ACM, 2007.

[11] J.-D. Fekete and N. Dufournaud. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *DL '00: Proc. of the fifth ACM Conf. on Digital Libraries*, pages 47–55. ACM, 2000.

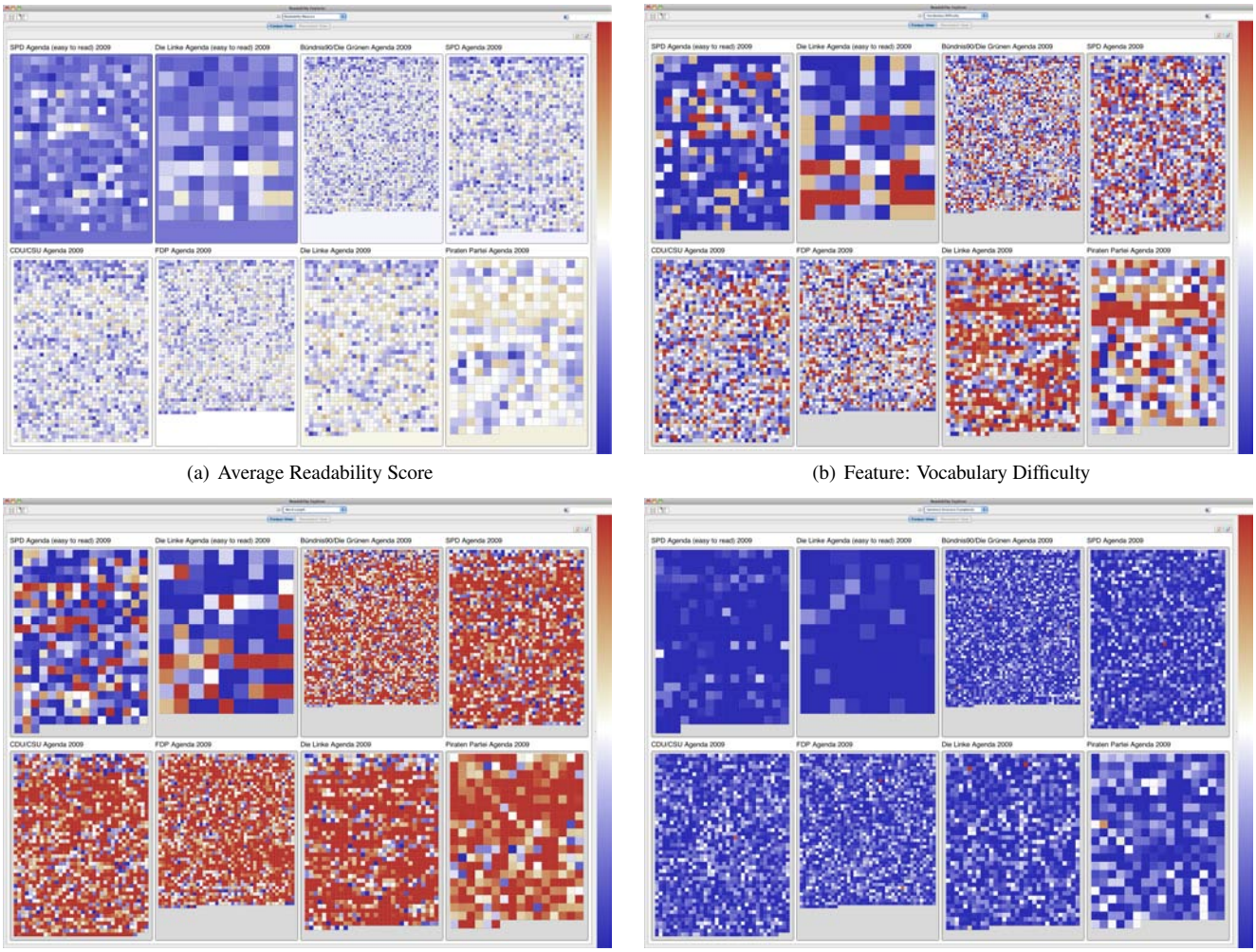
[12] R. F. Flesch. A New Readability Yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.

[13] R. Gunning. *The technique of clear writing*. McGraw-Hill, forth printing edition edition, 1952.

[14] Dictionary of the most frequent words in the Project Gutenberg, `http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000`, last accessed on 03/29/2010.

[15] Dictionary of the most frequent words in the Project Wortschatz Universität Leipzig, `http://wortschatz.uni-leipzig.de/html/wliste.html`, last accessed on 03/29/2010.

[16] M. A. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *CHI '95: Proc of the Conf.*



(a) Average Readability Score

(b) Feature: Vocabulary Difficulty

(c) Feature: Word Length

(d) Feature: Sentence Structure Complexity

Figure 8: Visual Analysis of 8 election agendas from the elections of the German parliament in 2009.

on Human Factors in Computing Systems, 1995.

[17] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *HLT-NAACL*, pages 460–467, 2007.

[18] D. A. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *VAST '07: Proc. of the IEEE Symposium on Visual Analytics and Technology*, pages 115–122, 2007.

[19] J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas for Navy Enlisted Personnel. Research branch report 8-75, Naval Air Station Memphis, 1975.

[20] D. Klein and C. D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003.

[21] H. G. McLaughlin. SMOG Grading - a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

[22] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *EMNLP '08: Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 186–195. ACL, 2008.

[23] S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *ACL '05: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. ACL, 2005.

[24] L. Si and J. Callan. A statistical model for scientific readability. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM, 2001.

[25] A. Stoffel, D. Spretke, H. Kinnemann, and D. Keim. Enhancing document structure analysis using visual analytics. In *Proc. of the ACM Symposium on Applied Computing 2010*, 2010.

[26] B. Suh, A. Woodruff, R. Rosenholtz, and A. Glass. Popout prism: adding perceptual principles to overview+detail document interfaces. In *CHI '02: Proc. of the SIGCHI Conf. on Human factors in computing systems*, pages 251–258. ACM, 2002.

[27] K. Toutanova and C. D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *EMNLP/VLC 00: Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70, 2000.

[28] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli. Using Thumbnails to Search the Web. In *CHI '01: Proc. of the SIGCHI Conf. on Human factors in computing systems*, pages 198–205. ACM, 2001.

[29] V. H. Yngve. A Model and an Hypothesis for Language Structure. In *Proc. of the American Philosophical Society*, volume 104, 1960.