Addressing Variation at Scale in Historical Document Collections

Eric Alexander, Deidre Stuffer, and Michael Gleicher, *Member, IEEE*

Abstract—In the digital humanities, variation in the data is both a blessing and a curse. On one hand, variation represents change that can provide us with broad-scale historical insight (e.g., "How is what people talked about in the 18th century different from what they talked about in the 17th century?"). On the other hand, some kinds of variation can actually *obscure* others (e.g., did people actually stop talking about a particular word in the 17th century, or did the spelling just change?). As DH research scales up to more and more documents covering ever longer spans of history, variation of both kinds is necessarily unavoidable and must be explicitly accounted for. In this paper, we discuss the challenges concerning variation that we have encountered in a multi-year, collaborative project focused on a collection of print documents from 1470-1800. We have addressed this variation through a combination of data standardization and task-driven visualization design.

Index Terms—Visualization, early modern literature, digital humanities

1 Introduction

In the digital humanities, variation in the data is both a blessing and a curse. Some kinds of variation are interesting—in fact, they are precisely the thing that we are looking for as researchers. They help us answer questions like "Is discourse more dominated by religion in the 17th century than in the 18th?" or "How does this genre compare with that one?" On the other hand, some kinds of variation can actually *obscure* others. For example, we might wonder if people stopped using a given word in the 17th century, or if the spelling of that word just changed. Which kinds of variation are actually "interesting" will depend completely upon the person asking the question, but the overlapping nature of this variation has the potential to lead any researcher astray, regardless of their subject interest.

As we scale up our research to include more and more documents, variation becomes more interesting, in part because more data allows us to make stronger statistical claims about what we find. However, it also becomes harder (to the point of impossibility) to manually filter away all the instances of "uninteresting" variation. If our corpus of documents is on the order of the works of Shakespeare, then perhaps we can very carefully curate each individual word. Such is no longer the case if we let our corpus grow to the scale of thousands, or tens of thousands of books.

As such, when working with historical documents, we must be careful to understand the varied sources of variation in the data and do our best to account for them in both our curation and analysis. In this paper, we describe our experiences in dealing with variation within the Visualizing English Print (VEP) project. This project has been a multi-year collaboration of computer scientists and humanities scholars focused on bringing the techniques of visualization and statistical analysis to bear on a collection of early-modern print documents from the years 1470-1800. Over the course of the project, we have learned to account for different sources of variation through a combination of transparent data standardization and task-driven visualization design.

2 The Project

For most of its existence, literature scholarship has been limited by how much any one person can read. While access to documents has often been restricted, even scholars with entire libraries at their fingertips could never hope to read more than a fraction of what was available. The vast majority of scholarship, therefore, has been centered around the **canon**: a small subset of works thought by experts to have had the greatest historical impact and cultural influence. While this focus has ensured that researchers have a common ground upon which to make arguments, it nonetheless reflects but a tiny sample of what was produced during the time period. As such, it is difficult to say whether arguments based upon it are representative of the period as a whole.

One of the primary goals of the VEP project was to move beyond these limitations and enable literature scholars to answer questions that can only be asked at scale, such as:

- What were people writing about during the early modern era?
- How did language and topics of discussion change over time?
- Is it possible to track the evolution of particular genres?
- Is our concept of "genre" itself an accurate reflection of the types of works that were created?
- What attributes make texts similar or dissimilar from one another?

Some of these questions do not even make sense without considering more documents than any one person could ever hope to read, and so they require computational methods to address them. Scholars have posed canon-centric theories about others of these questions; in such instances, our ability to scale up presented an opportunity to *scrutinize* these theories. Shifting our perspective to documents and trends that arise from statistical models of the entire collection of early modern print gave us a chance to "de-privilege" the documents that were hand-picked for the canon. We could potentially learn new things about the era that would sidestep centuries of bias that had been passed on and augmented over the years by small, select groups of experts.

The main collection, or set of collections, that we focused on was the Text Collection Partnership (TCP), which provides richly annotated SGML/XML editions of early printed books for the public domain. The TCP produced corpora hand-keyed from digitized microfilm images of select works from three databases: Early English Books Online (EEBO), Eighteenth Century Collections Online (ECCO), and Evans American Imprints (Evans). The TCP has released approximately 61,000 digital texts that were originally published between 1470 and 1800 in English speaking locales. Altogether, the result is a dataset of vast heterogeneity including not only fiction but scientific treatises, language books, sheet music, maps, and political cartoons.

In the early stages of planning for the VEP project, we encountered many types of variation in this dataset, some of which we anticipated and others which we did not. The years between 1470 and 1800 covered a time of great change for both printing and the English language

- *Eric Alexander is with Department of Computer Sciences at the University of Wisconsin-Madison. E-mail: ealexand@cs.wisc.edu.*
- *Deidre Stuffer is with the Department of English at the University of Wisconsin-Madison. E-mail: stuffer@wisc.edu.*
- *Michael Gleicher is with the Department of Computer Sciences at the University of Wisconsin-Madison. E-mail: gleicher@cs.wisc.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

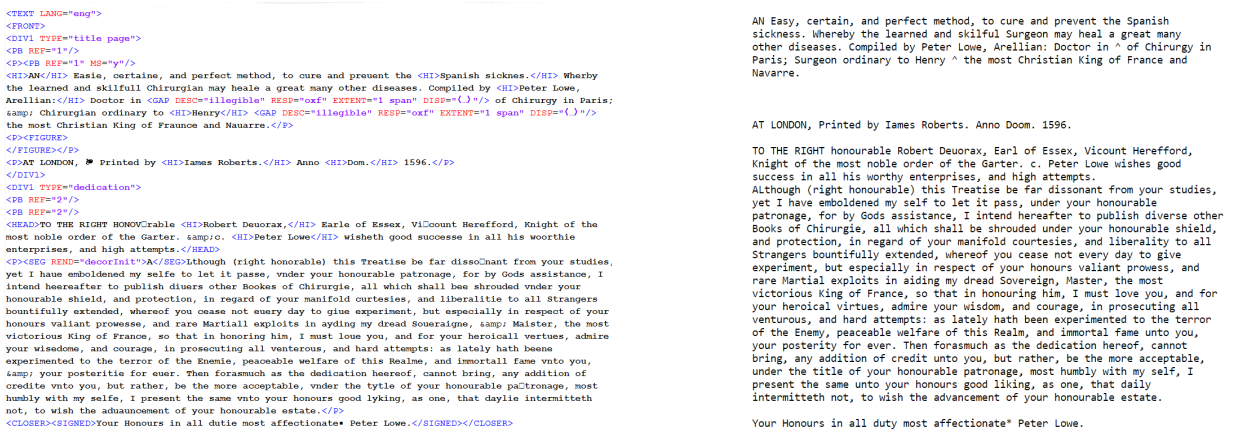

Fig. 1. Here, we see a side-by-side comparison of a document in its original TEI markup version and a version that has been converted to our SimpleText format. While SimpleText gets rid of some of the rich annotation of the original XML, its output is directly machine-actionable by a variety of natural language tools and techniques.

itself, as neither uniform printing practices nor even spelling was yet standardized. Beyond that lie all the things that can happen to a document in between it being printed and it appearing in one of our models. It took appreciation of all the sources of variation and a strategy for filtering them to be able to answer questions like those posed above.

3 Types of Variation

The particular challenge of a historical dataset like the one with which we work is the centuries-long history that each document has gone through—and all of the things that happen to a document along the way. Each text started with a glimmer of thought, resulting in *something* being written down—perhaps by the author or perhaps transcribed by someone else. This was then printed by a printmaker, operating under an entirely different set of constraints. The printed document was then subjected to the ravages of time, be they wind, weather, overuse, or neglect. The physical document was then microfilmed and subsequently digitized (by hand-keying in the case of the TCP, or by optical character recognition for some other collections). And finally, this digitized document was then read in and processed by any one (or more!) of a variety of computational libraries and tools.

At every step along the way, there may have been variation introduced that has nothing to do with the original content. It is tempting to separate these changes into "interesting" variation and "uninteresting" variation, though of course that distinction can change drastically depending on who is asking the question. Perhaps more problematic than which of these steps is worthy of study is the issue of determining *where* in this process any piece of observed variation occurred.

Suppose that in some document, we encounter the word "naïve" spelled with an **i** rather than an **ï** (with an umlaut). Where did this variation in spelling come from? Did the author intend to spell it this way, indicating something about their own vocabulary? Did the printer change it in typesetting, either because they thought that was correct or perhaps because they ran out of umlauts? Did one of the dots get smudged away from the physical copy over the years? Did the transcriber or OCR engine misread it? Does our file reading library even *handle* umlauts properly? Ideally, we would know where the change originated. Unfortunately, for many such cases, it is impossible to be sure.

4 Handling Variation: The Task-Driven Approach

In the process of obtaining, curating, and standardizing the documents, we encountered each of the types of variation described above. We determined that the *wrong* approach to handle all these variations was to try to create or rely upon a format of the documents that maintained every bit of information that we had. This would have required too much hand-curation of the documents, and would not scale to the full corpora. The irony of this statement is that 61,000 documents is not particularly immense in the current era of "big data." Computational analysis been used to great effect on corpora containing many millions of documents from sources like the New York Times, Wikipedia, and Twitter. However, these models (with the possible exception of the internet-speak on Twitter) benefit from operating on standardized language that has generally been created directly in the digital medium.

Unfortunately, there are some kinds of questions that we cannot answer. For instance, it is difficult for us to make claims about rare words. Small, impossible-to-identify sources of variation can have too large of an effect on such words for us to have confidence in any trends we might find. However, even if we *could* answer every kind of question we might want to, preparing the data to answer one question can get in the way of being able to answer another. Spelling is a good example of such a tradeoff. Are we interested in how spelling changed over the 17th century? If so, we should likely strive to keep all words in their original spelling (or as close as we can get) so as to be able to observe changes. Unfortunately, this gets in the way of more *content-based* inquiries. For instance, if a researcher is building a topic model of the same time period, they likely want it to recognize that a 17th century author writing "accompt" and an 18th century author writing "account" actually mean the same thing.

We took a task-driven approach to navigate these tradeoffs, focusing on what our collaborators actually wanted to do with the data. Rather than trying to capture everything about the documents, we honed in on the questions that they were most interested in asking. By targeting these questions, we were able to decide which types of variation were important to preserve, and which would only serve to confound our investigation. This allowed us to properly direct both our data curation process and the design of the visualization tools that provide our collaborators with a window into the dataset.

4.1 Curation and Standardization

Deciding which sorts of variation we cared about also meant deciding which we did not. There were a number of types of variation that we explicitly tried to remove—not because they are not potentially interesting, but because they were less relevant to our particular goals. We were not interested in exploring questions of typography, for example, nor were we going to worry about variations in spelling. We decided upon a set of *content-focused* tradeoffs that would result in a corpus of machine-actionable texts that would let us use a variety of digital tools to investigate variations in content.

When deciding upon the format that the instantiations of these choices would ultimately take, we opted to create something that was as simple as possible. There is much to be gained from complex XML markups like those produced by the TCP, which pack a wealth of infor-

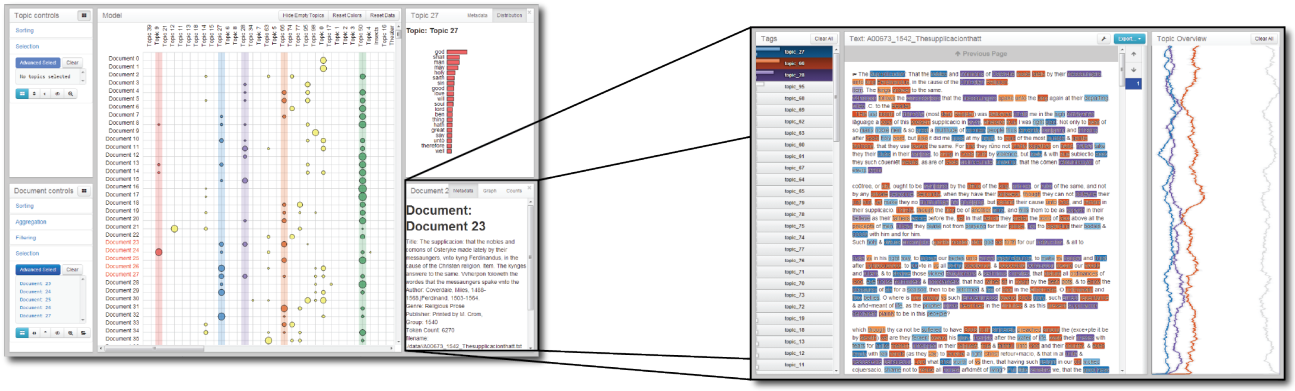

Fig. 2. Serendip helps scholars observe variation in topic content across large text collections by enabling high- and low-level exploration of topic models [3]. It uses a reorderable matrix (left) to help readers find corpus level trends and then lets them drill into individual passages of tagged text (right) to provide exemplars of these trends for closer study.

mation into a single file. However, this complexity also necessitates a complicated processing step in which the relevant tokens must be extracted before inputting them into other tools and libraries for analysis. This processing step is sensitive to differences in XML encodings, making it difficult to use and build tools that operate on a wide variety of corpora. On the other hand, just about any program can read text files, from Microsoft Word to the MALLET topic modeling suite. Though it would mean losing some of the richness of the encoded data, we wanted something that would work for the lowest common denominator.

As such, we created a format that we call SimpleText. SimpleText is just that: a simplified text format that sacrifices the rich structure and metadata that can be contained in XML formats so as to better facilitate algorithmic approaches to content analysis. SimpleText has a number of key features:

1. It substitutes UTF and Unicode characters for their closest counterparts in ASCII. This allows the text to be handled by a wider variety of digital tools, many of which are not well-equipped to handle Unicode.

2. It does not include any metadata annotations, favoring to store those in separate metadata-specific files. This also facilitates direct integration into digital tools without needing an intermediary processing step.

3. It does not preserve physical aspects of document layout or typography, but merely strives to maintain line breaks.

4. It employs simple, dictionary-based spelling standardization.

We have created multiple scripts that form a pipeline for converting TCP documents into SimpleText, covering character-cleaning, metadata-stripping, and other kinds of filtering. For this paper, however, we will focus primarily upon the process of spelling standardization. Standardization is difficult for early modern texts given the enormous amount of variation. (For example, versions of the word "diverse" can be found spelled as "diuerse," "diuers," "dyuers," "divers," etc.) Standardization is distinct from the related process of *modernization*. Its goal is not to replace each word with its modern equivalent, but to ensure as best we can that strings representing the same word-forms are all the same. Initially, we used VARD [4], a system that employs algorithmic techniques to *learn* which strings are representing the same word-forms. Though VARD tended to give good results, they were not perfect. More problematic (as nothing can be perfect) was that the machine learning approach to standardization made it an unfortunately opaque and non-determinisitic process. This served to introduce yet *another* source of enigmatic variation into the documents.

The version of spelling standardization that we ended up creating for converting TCP documents to SimpleText is, by design, easy to understand. By extensively evaluating a set of changes made by VARD on the TCP corpora and exhaustively inspecting each word that appears more than 2,000 times, we have come up with a dictionary of replacements that are applied uniformly across each text. (This is one of the reasons why we cannot make any promises about rare words, especially those appearing fewer than 2,000 times. Over 5.5 million words appear ten times or less in the TCP corpora, and manually looking at each of these would simply be impractical.) Our process of standardization is simply one of going through a non-standardized document token by token, comparing them to n-grams in our dictionary, and replacing them when there is a corresponding match.

This process is far from perfect, and there are many instances of standardization that we get wrong. However, we have prioritized certain types of "wrong" answer as being worse than others. In particular, we could fail to standardize a word that should have been changed, or we could standardize a word incorrectly. In collecting our set of replacements, we have carefully chosen the ones that would create the fewest errors without horribly altering the meaning, erring on the side of leaving words unstandardized when there is a significant conflict. Unfortunately, given that words often have multiple meanings, instances of both types are unavoidable with such a simple standardization process. We have decided that it is more important to be consistent and transparent than to be correct 100% of the time.

Take, for instance, the token "bee." This spelling, quite frequent within the TCP documents, can refer to either the insect or the verb. However, in the first 1,000 instances of the token in a subset of dramatic texts from the TCP, we found only 17 instances that referred to the insect. We felt that 1.7% frequency was not enough to skew results, and so decided to standardize "bee" to the verb form "be."

All of our standardization decisions are made public online. This transparency is crucial, because combined with the simplicity of the replacement process, it makes it easy for scholars using these documents to perform checks on individual findings. The ability to scrutinize findings is essential, both because of the types of variation that we have explicitly filtered out as well as types of variation of which we may still be unaware. Ultimately, we cannot account for everything, and so all analysis of this type must be performed with a critical eye. Through our task-driven standardization process, we seek to expose our pipeline to such critique.

4.2 Visualization Design

Throughout the VEP project, we have focused on visualization as a crucial method for putting statistical analysis in the hands of humanities scholars. As our data curation was performed with an eye towards the types of variation we most wanted to understand, so were our visualization techniques designed in a task-driven way, each attempting to help our collaborators understand a specific subset of variation within the corpora. From time to time, this process actually worked in reverse:

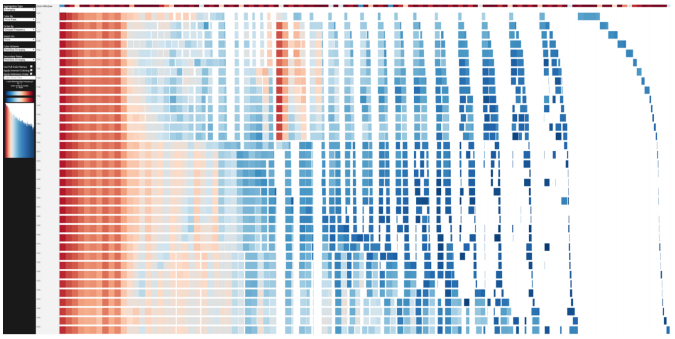

Fig. 3. TextDNA, a tool adapted from a gene-sequencing visualization tool, is used for identifying variation in word usage [1]. Word frequencies within each decade are represented by rows of colormappings. These encodings can be adjusted to highlight the trends in which researchers are most interested.

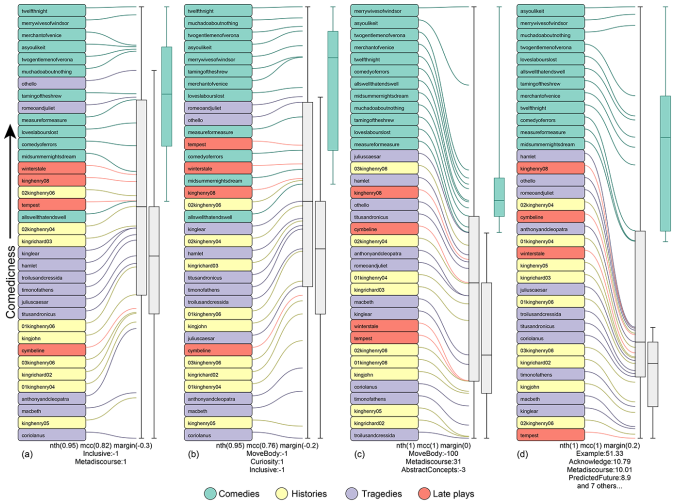

Fig. 4. Our work on Explainers allows researchers to align documents on simple, comprehensible axes that can be used to *explain* variations in metadata [6]. Here, we see four such axes that show the tradeoff between accuracy of classification and simplicity of the function.

sometimes a tool would expose new variation and help us decide how to act upon it. For instance, our techniques for tracking word usage (see Section 4.2.2) exposed certain kinds of spelling variation, and our topic model explorations (see Section 4.2.3) alerted us to a number of non-English documents and passages within the documents. In general, however, each tool was built to address a particular task.

4.2.1 Annotating close reading

An early task of our collaborators was to view algorithmically generated tags as they had been applied to words in individual documents. They had been using a system called Docuscope [7] to convert their documents into vectors of linguistic categories appearing within them, but wanted to be able to apply their methods of close reading to the individual documents, observing precisely how the high-level patterns they saw in models of the vectorized documents manifested themselves in the low-level passages. To address this need, we created a tool called TextViewer that used tagged text to show which words fell into which categories, as well as a line graph visualization to help researchers navigate to passages of particularly high variation [5]. Over the course of the project, this method for enabling close reading within statistical models of text has been adapted in a number of other tools [3].

4.2.2 Tracking word usage

A large source of variation within historical data is changes in word usage over time. Even after accounting for spelling variation, word usage can shift dramatically over long time spans. To help our collaborators investigate this phenomenon, we adapted a tool that we had created for gene-sequencing visualization to be applied to sequences of n-grams from large document collections (see Figure 3) [1]. Our collaborators' efforts with this tool uncovered a number of interesting vernacular changes (such as usage of the word "women" overtaking that of the word "wife" around the time of the women's suffrage movement) as well as a number of spelling standardizations that we had not initially accounted for (including variations of the long 's').

4.2.3 Exploring topic models

To be able to track changes in content across large corpora, our collaborators were interested in the ability to build and explore topic models of the documents. Important to this exploration task was the ability to connect high-level trends to low-level exemplary passages. To observe variation across multiple levels of abstraction, we created a tool called Serendip that merged the methods of close and distant reading (see Figure 2) [3]. At a "zoomed-out" level, the tool employs a reorderable matrix to allow readers to explore documents for interesting similarities and patterns of interest. At a "zoomed-in" level, tagged text visualizations allow readers to perform close reading with the added information of how each word fits into the larger context of the topic model. In between, line graph visualizations display document-wide trends and direct readers to passages of high density for individual topics. When combined, these views afford a workflow that easily connects high-level findings to low-level examples, helping the users build greater understanding of what they observed at a high level and to communicate those findings to their peers.

4.2.4 Comparing topic models

Though Serendip did much for enabling researchers to explore topic models after they had already been built and trained, any individual topic model is a single point in a vast parameter space. To help researchers navigate this parameter space, we devised a number of visual techniques for *comparing* different topic models [2]. These techniques were focused on the primary tasks for which we observed topic models generally being used: understanding topics, understanding similarity, and understanding change.

4.2.5 Explaining distinctions

Another source of variation lies in the *metadata* of the documents: author, genre, place-of-origin, etc. One of the questions our collaborators often ask is "What makes this group of documents different?" It is relatively simple to train machine learning classifiers to tell the difference between a comedy and a tragedy, for example, but if such a classifier does not easily convey *how* it makes its choices, this is not necessarily useful for a scholar trying to build new hypotheses about what makes such a distinction important. To help scholars in hypothesis formation, we developed a technique for projecting documents onto axes defined by metadata (e.g., "comedicness") that balance the tradeoff between accuracy and comprehensibility (see Figure 4).

5 Conclusion

For those seeking to understand large collections of historical documents at scale, variation in the data is both friend and foe. Given the many, often indiscernible sources of variation in such documents, it is important that they be accounted for. We have taken the approach of explicitly choosing the types of variation in which we are most interested, and fitting both our data curation and the tools we build to those types. In doing so in an transparent and documented manner, we hope that our techniques afford large-scale inquiry into these documents that can be conscientiously critiqued and verified.

Acknowledgments

This work was supported in part by NSF award IIS-1162037 and a grant from the Andrew W. Mellon Foundation.

REFERENCES

[1] D. Albers Szafir, D. Stuffer, Y. Sohail, and M. Gleicher. Textdna: Visualizing word usage with configurable colorfields. *Computer Graphics Forum*, 35(3):421–430, Jun 2016. doi: 10.1111/cgf.12918

[2] E. Alexander and M. Gleicher. Task-driven comparison of topic models. *IEEE Transactions on Visualization and Computer Graphics*, December 2015. doi: 10.1109/TVCG.2015.2467618

[3] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 173–182. IEEE, 2014.

[4] A. Baron, P. Rayson, and D. Archer. Quantifying early modern english spelling variation: change over time and genre. In *Conf. New Methods in Historical Corpora*, 2011.

[5] M. Correll, M. Witmore, and M. Gleicher. Exploring collections of tagged text for literary scholarship. *Computer Graphics Forum*, 30(3):731–740, jun 2011. doi: 10.1111/j.1467-8659.2011.01922.x

[6] M. Gleicher. Explainers: Expert explorations with crafted projections. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2042–2051, dec 2013. Proceedings VAST 2013. doi: 10.1109/TVCG.2013.157

[7] S. Ishizaki and D. Kaufer. Computer-aided rhetorical analysis. *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*, pp. 276–296, 2011.