Text X-Ray: An Interactive Text Visualization Tool for Corpus-based Language Teaching and Learning

Xi He, Ying Zhu
Department of Computer Science
Georgia State University
Email: xhe8@student.gsu.edu, yzhu@gsu.edu

Eric Friginal
Department of Applied Linguistics and ESL
Georgia State University
Email: efriginal@gsu.edu

***Abstract*—Corpus and corpus tools have become an important part of language teaching and learning. And yet text visualization is rarely used in this area. In this paper, we present Text X-Ray, a Web tool for corpus-based language teaching and learning. Interactive text visualizations in Text X-Ray allow users to quickly examine a corpus or corpora at different levels of details: articles, paragraphs, sentences, and words. Users can use the text visualizations to quickly compare the complexity of multiple articles by visually checking sentence lengths, sentence grammar trees, and word patterns. Traditional linguistic analyses such as readability indices or lexical density provide useful information, but they do not provide enough detail. Text visualizations fill that gap. The text visualizations are synchronized with the textual display and are designed for easy transition between textual and visual analysis. Text X-Ray also allows teachers or students to load their own articles or corpora and compare them with other corpora.**

I. Introduction

A corpus is a large collection of texts. In general, corpora are not meant to be read, but to be processed and analyzed by computer programs, which are called corpus tools. For example, Google's Ngram Viewer is a corpus tool that allows users to search the Google Books corpus and visualizes the use of selected English words over many years. Corpus tools have been developed for many areas such as biomedical research, linguistic research, intelligence analysis, literature critic, marketing, etc. The focus of this project is on corpus tools for English language teaching and learning.

The value and effectiveness of corpus-based language learning and teaching have been established in previous studies [1]–[3]. For example, teachers can guide students to use a concordancer to study a corpus and build their knowledge about language use. Although many corpus tools have been developed for studying English language, they were mostly designed for linguistic research, not for language teaching and learning. For example, many corpus tools are designed for specific corpora and don't allow users to import their own corpus. In addition, the user interfaces of many corpus tools are designed for language specialists. Students and teachers often find them unhelpful in analyze their own writings.

To address this issue, we developed a corpus tool, Text X-Ray, that helps teachers and students to analyze their own corpus and compare their own corpus with other corpora. Text X-Ray consists of a natural language processing engine, a visualization engine, a corpus analysis engine, corpus databases, and a visualization interface. The Web based visual interface provides visualizations of a corpus at every levels of detail: articles, paragraphs, sentences, and words. The goal of this tool is to make it easier for users to see the structures, patterns, and relationships that are not easily recognized in plain English text. This is why it is called Text X-Ray.

Comparing with other related corpus tools, Text X-Ray introduces a much more visual approach to language teaching and learning. Traditional corpus tools provide statistical analysis of the texts, which are useful but do not give the full picture. Text visualizations can fill that gap. For example, two articles may have similar readability indices but quite different writing styles. One article may use longer sentences but simpler words, while the other one uses simpler sentences but more difficult words. Such differences are readily visible in Text X-Ray's visualizations.

Text X-Ray is also more user friendly and versatile. Linguistic researchers can use it to analyze and compare articles in a corpus. Students can use it to compare their own writings with a corpus. To help teachers use it in classroom settings, Text X-Ray also supports corpus management and user management.

II. Related Work

The most popular corpus tools for linguistic studies include the BYU corpora [4], AntConc, WordSmith, Sketch Engine, Sarah, Monoconc Pro, WMatrix, etc. [5]. There are generally two types of corpus tools: (1) corpus tools developed for a particular corpus or particular corpora; (2) corpus tools that can process different corpora. Our tool, Text X-Ray, belongs to the second category.

The main issue with the most existing corpus tools is that they are not designed for language learning. In a recent review of corpus tools, Laurence Anthony [5], author of the popular AntConc program [5], pointed out that the current corpus tools are "... generally researcher-centric in that they do not always lend themselves to easy use in the classroom with learners." He also points out that students "need a corpus tool that gives them access to a corpus in an easy and intuitive way. They also need a tool to show them results that are immediately applicable to a given learning task ..."

A number of corpus tools, such as Compleat Lexical Tutor [6], AntConc [5], Word and Phrase [7], have attempted to address this issue by making them more useful for teachers and students. For example, Compleat Lexical Tutor [6], a Web based corpus tool, allows users to enter their own texts or

a corpus and then make a concordance for all the words in the text. AntConc [5] also provides similar functions. Wordandphrase.info [7] allows users to load their own texts and then highlight all of the medium and lower-frequency words, based on selected corpora. This helps students focusing on learning the new, low frequency words.

The main difference between Text X-Ray and the above corpus tools is the visualization interface. The existing corpus tools display statistical analysis and highlight texts, but with limited user interactivity. Text X-Ray retains important functions found in most corpus tools but provides a much more visual and interactive user experience. In particular, Text X-Ray's interactive sentence bar and parse tree visualization are new features that are not seen in existing corpus tools.

Although many text visualization techniques have been developed [8], text visualization is rarely used in linguistic research and education. Siirtola, et al. [9] pointed out that "There are very few exploratory visualization and analysis tools that are linguistically motivated ..., and even fewer that allow rapid exploration of linguistic parameters." We examined recent surveys [8]–[10] on text visualization techniques and found only a few text visualization techniques for linguistic research [9], and no comparable work on language learning and teaching. This is an area that hasn't been sufficiently explored, and our work is an attempt to address this issue.

III. Characteristics of text visualization in language teaching and learning

Text visualization for language teaching and learning needs to meet particular requirements. These requirements are not unique to language teaching and learning, but they are emphasized more in this area than in others.

- The texts should be displayed along side the visualizations. In many text visualization techniques, the visualizations replace the texts. The original texts are often not displayed in the interface. But in language teaching and learning, the texts need to be examined at all times. Visualizations should support the examination of the original text, not replacing it. Text visualizations should be linked and synchronized with the text display.
- For language teaching and learning, highly innovative and abstract data visualization should be introduced with great care. Because users often switch between the original text and the visualizations, the form of the visualizations should be closer to the conventional textual display for easier mental transition and connection. Some text visualization techniques are difficult to use because they require too much mental adjustment from one form of display to another. Therefore the complexity and abstractness of the visualization needs to be controlled.

These two principles are the main guidelines for our interface design.

IV. System

A. Overview

Text X-Ray [11] is an online interactive text visualization system that can accept user-uploaded corpora, processing them, and display them in both textual and visual forms. It contains a server-side subsystem and a browser-side subsystem. The server-side subsystem is designed to handle natural language processing that requires intense computation. It contains seven components:

- Web Server
- Natural Language Processing (NLP) Engine
- Corpus Analysis Engine
- Corpus Cache
- Corpus Management System
- Corpus Database
- Corpus API

The browser-side subsystem consists of Visualization Engine and Visualization Interface, which provides user interactions and data visualizations. The server-side and browser-side subsystems communicate via HTTP request/respond messages.

Figure 1 shows the main components and the workflow in Text X-Ray. On the server side, there are three major pipelines. The first pipeline is for processing corpora. Once corpus processing requests are received, Web Server will check whether the requests have been processed before. If yes, then the requests will be sent to Corpus Cache, which will send the processed corpus data back to the browser. Otherwise, NLP Engine takes the requests and retrieves the requested corpus from the Corpus Database, and parses it. The results are delivered to Corpus Analysis Engine for further text analysis. Finally the processed corpus is cached in Corpus Cache and also sent back to the browser. The second pipeline is designed to allow corpus management. After receiving requests from Web Browser, Corpus Management System performs Add/Modify operations on Corpus Database. The third pipeline provides API for corpus based computation. Corpus API contains a set of functions that manage Corpus Database.

A requested corpus with plain text is processed in the server-side subsystem, transformed into a hierarchical object and sent to the browser-side subsystem. The hierarchical object not only retains the original full-text information, but also maintains identifiable information for paragraphs and sentences. Associated statistics for separate paragraphs and sentences, and sentence parsed trees are also contained in the hierarchical objects.

On the browser side, users can choose a corpus and explore the articles in that corpus. Visualization Engine will load the entire processed data for a specific corpus from the server side before users can interact with Text X-Ray. Such design is to guarantee that the user interaction is smooth and not affected by network transmission. Once corpus data is loaded, Visualization Engine constructs data visualization and is ready to respond to user inputs from Visualization Interface.

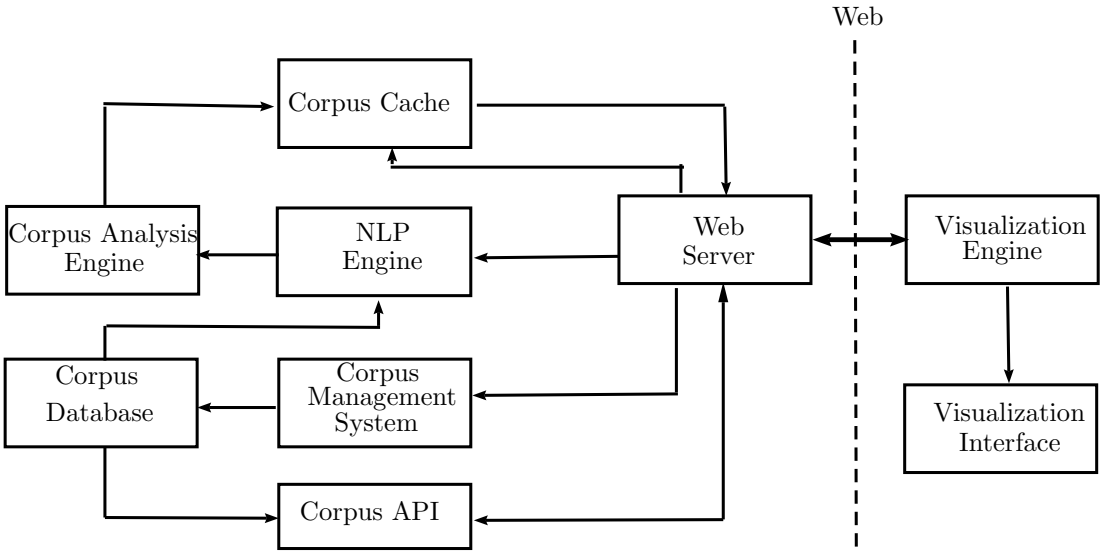

Fig. 1: Workflow in Text X-Ray

The server-side subsystem is written using Java servlet with Stanford Natural Language Parser [12] and Google Gson libraries. The browser-side subsystem is developed with jQuery UI [13] and D3 [14] library.

B. Natural Language Processing (NLP) Engine

The NLP Engine is built on top of Stanford Natural Language Parser [12]. The Stanford Parser is a probabilistic parser that uses the knowledge from human parsed sentences to analyze the structure of new sentences. The Stanford Parser provides Java APIs that can analyze and construct the grammatical structure of sentences. In Text X-Ray, we use the Stanford Parser to construct a grammar tree for each sentence. We also use Stanford Parser as a Part-Of-Speech (POS) tagger. The POS tags are stored in the grammar trees. The grammar trees are stored in an internal data structure. The text analysis and text visualization, particularly the sentence tree visualization, are based on these grammar trees.

For each article in a corpus, Text X-Ray extracts paragraphs and then divide it into sentences. It then uses Stanford Parser to process each sentence and construct a grammar tree. These grammar trees are forwarded to Corpus Analysis Engine.

The main concern in developing the NLP engine is the speed of parsing. Natural language processing is quite time consuming, especially for a large corpus consisting of thousands of sentences. To speed up the NLP engine, we have developed a multi-threading NLP engine using the Java executor framework. The NLP engine set up a thread pool with configurable number of threads in it. Then sentences are evenly distributed to the thread for processing. When done, the parsed sentences are collected and reorganized in the original order. On a 1.9 GHz quad-core processor, the engine can process an 8,000-word New Yorker article in about 5 seconds.

C. Corpus Analysis Engine

Corpus Analysis Engine has two primary functionalities. The first one is to compute linguistic statistics of the corpora. It contains a collection of functions that implement various linguistic analysis algorithms. Currently implemented functions include:

- Identify thirty-one Part-Of-Speech (POS) items such as nouns, verbs, adjectives, etc.
- Identify long words
- Identify long sentences
- Calculate readability indices
- Calculate lexical density
- Calculate word frequency

The second functionality of the corpus analysis engine is to prepare corpora for text visualization. Texts in corpora do not provide enough information for text visualizations. Visualization Engine needs to be able to identify paragraphs, sentences and words in corpora, and obtain information such as the length of a sentence or the POS tag of a word in order to visualize them. The solution in Corpus Analysis Engine is to transform a corpus into a hierarchical data structure of article objects, paragraph objects, sentence objects, and word objects, making the navigation of these objects efficient.

D. Visualization Engine and Visualization Interface

Visualization Engine and Visualization Interface constitute a standalone browser application. Visualization Interface is responsible for text visualization and user interaction while Visualization Engine controls the work flow and maintains visual system status.

Text X-Ray is divided into multiple panels: text panel, visualization panel, linguistic analysis panel, and control panels (Fig. 2). The text panel displays the currently selected article. The visualization panel, always parallel to the text

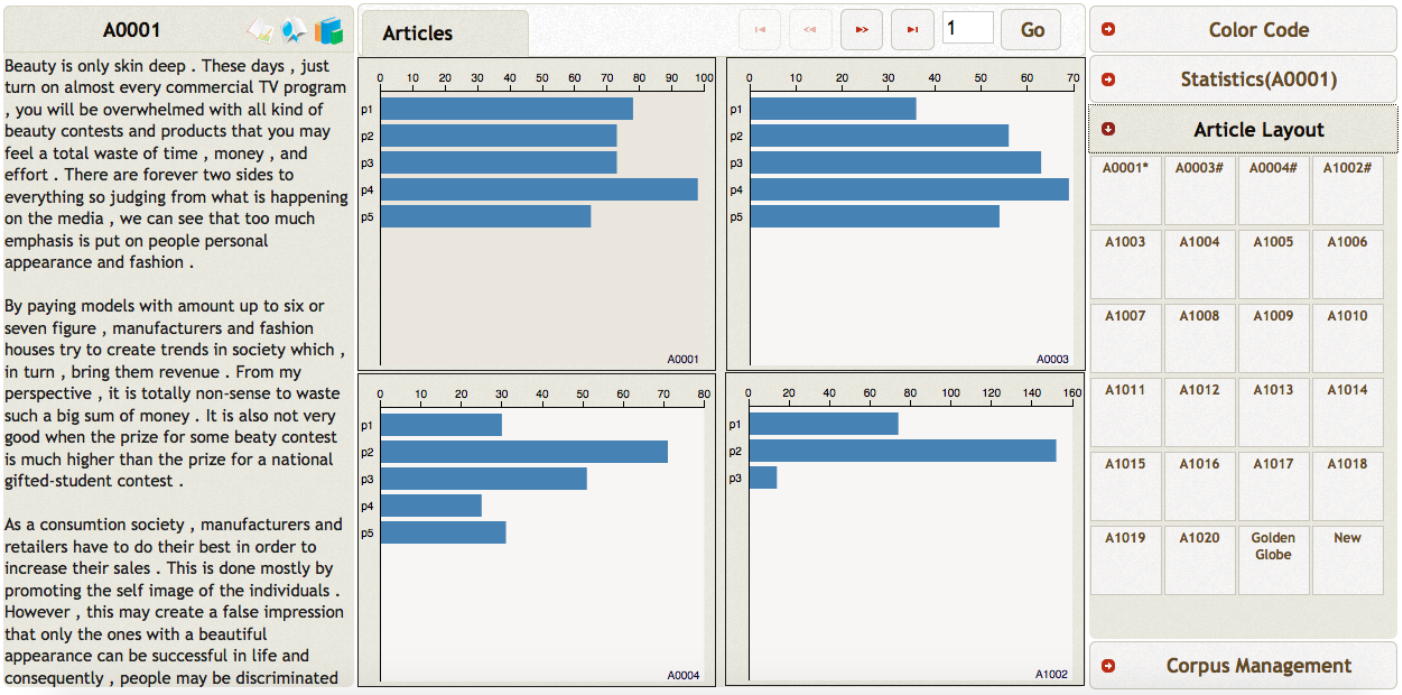

Fig. 2: Overview of Visualization Interface

panel, enable users to analyze the texts in five levels of detail: corpus, articles, paragraphs, sentences, and words. The linguistic analysis panel (Fig. 3) shows the output of linguistic analysis of the corpus or an article, such as readability indices, lexical density, etc. The control panel (Fig. 4) allows users to adjust the visualization settings and manage corpora or users.

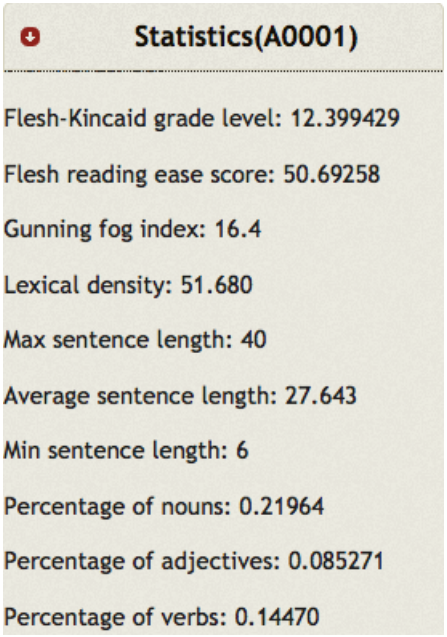

Fig. 3: Linguistic analysis panel

As mentioned in the previous chapter, the text panel and

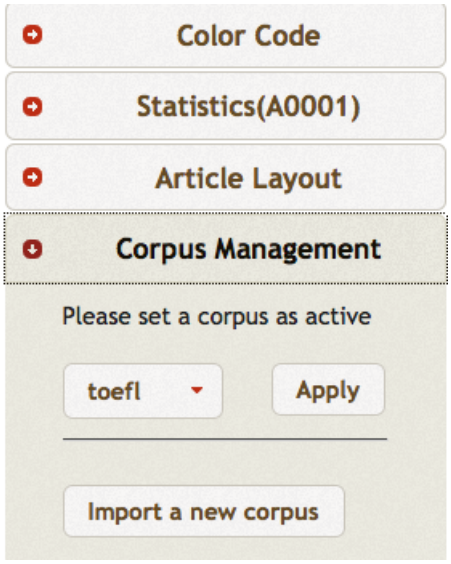

Fig. 4: Control panel

visualization panel are displayed next to each other so that users can easily switch between reading visualization and reading the original text. The text and text visualization are also linked. For example, clicking on a sentence bar or a word block in the text visualization, the corresponding sentence or word is highlighted in the text window.

The corpus view allows users to have an overview of multiple articles in a corpus. Users can visually compare different articles for the length of the article and the length

of the paragraphs. The paragraphs are visualized as horizontal bars.

For further details, a user can click on a paragraph bar, and the sentence view is displayed. In the sentence view, sentences are visualized in two forms: horizontal sentence bars and horizontal sentence grammar trees. The sentence bars (Fig. 5) allow users to quickly compare the lengths of the sentences, something they cannot easily do with plain text. For even more details, a user can click on a sentence bar to visualize words as color blocks (Fig. 6). Users can choose to have the color coding being based on either POS tags or word frequency.

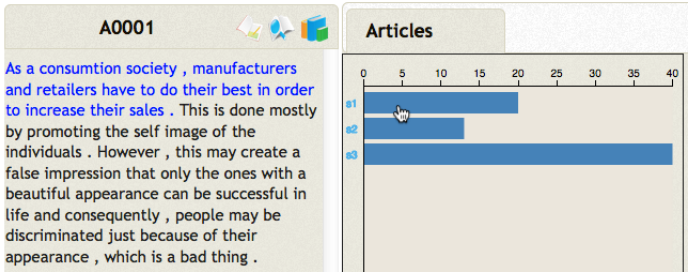

Fig. 5: Sentence bars

Fig. 6: Sentence bars (with color coded word blocks)

The sentence bars, word blocks, and sentence trees (discussed below) give users a new way of analyzing the reading complexity of an article. Traditionally the complexity of an article (or reading difficulty) is measured and presented in various readability index numbers or lexical density number. (These index numbers are presented in the linguistic analysis panel.) Reading difficulty is usually calculated by counting long sentences, low frequency words, and complex sentence. But a single readability index does not convey enough information. For example, two articles may have similar readability indices but one article may have some very difficult paragraphs and some easy paragraphs, while the other article may have a more evenly distributed reading difficulty. A readability index may not distinguish these two articles, but our text visualization will. With our text visualizations a user (or author) can quickly see that certain parts of an article are more complex than other parts.

Our sentence tree visualizations (Fig. 7) show the grammatical structure of the sentences as parsed by the Stanford Parser. This visualization addresses a major weakness in the traditional reading difficulty analysis. In a plain English text, the grammatical complexity of the sentences cannot be quickly identified. Traditional readability indices only measure the complexity of a sentence by its length, not by its syntactic structure. The sentence tree visualizations makes it much easier for viewers to see and compare the structure of the sentence and its complexity. We also add color coding to the sentence tree visualization. By color coding words (i.e. leaf nodes) by their frequency or syllable count, the sentence tree visualization can show sentence length, sentence structure, and word difficulty in one view. To the best of our knowledge, none of the existing readability measures can achieve this.

The sentence tree visualization is implemented using our own Indented Level-base Tree drawing algorithm [15], which is based on a classic tree drawing algorithm [16] and implemented using d3.js and jQuery. The conventional way of drawing a grammar tree is to draw it vertically, with the root on top. In our Indented Level-base Tree drawing algorithm [15], the grammar tree is drawn horizontally from left to right. There are two benefits of drawing the tree horizontally. First, we read English texts horizontally from left to right. It's mentally easier for readers to switch between reading the text and reading the grammar tree that is displayed horizontally. Second, a horizontal sentence grammar tree fits the wide 16:9 aspect ratio of current computer displays. In our program, viewers can choose to use the traditional, vertical level-based tree drawing algorithm or our horizontal, indented level-based tree drawing algorithm.

The sentence tree visualization is highly interactive. Users can expand or fold any node on the tree and the tree will be automatically re-drawn. This allows users to simplify certain part of the grammar tree and focus on the other parts, thus working on multiple levels of detail. This kind of interactive grammar tree visualization is an innovation in text analysis and is not found in any existing corpus tool.

A more complete example is shown in Fig. 9 and 10.

V. Application of Text X-Ray in Teaching and Learning

Explorations into the usefulness of corpus tools for pedagogy are expanding with online technology and the availability of internet-based resources that are freely distributed to teachers and students. In general, vocabulary learning and the analysis of grammatical structures for different proficiency levels of language learners have been the more dominant foci of many online tools intended for classroom use [17]. Studies highlighting academic vocabulary in content-based language instruction [18], [19] have demonstrated the feasibility and potential utility of corpus techniques in intensive English programs especially in the United States (U.S.). These and related areas in teaching are directly addressed by the design and applications of Text X-Ray.

For our initial pilot studies in the classroom, we have given access to Text X-Ray primarily to graduate students and writing instructors working with adult English as a second language (ESL) learners. These users were asked to explore

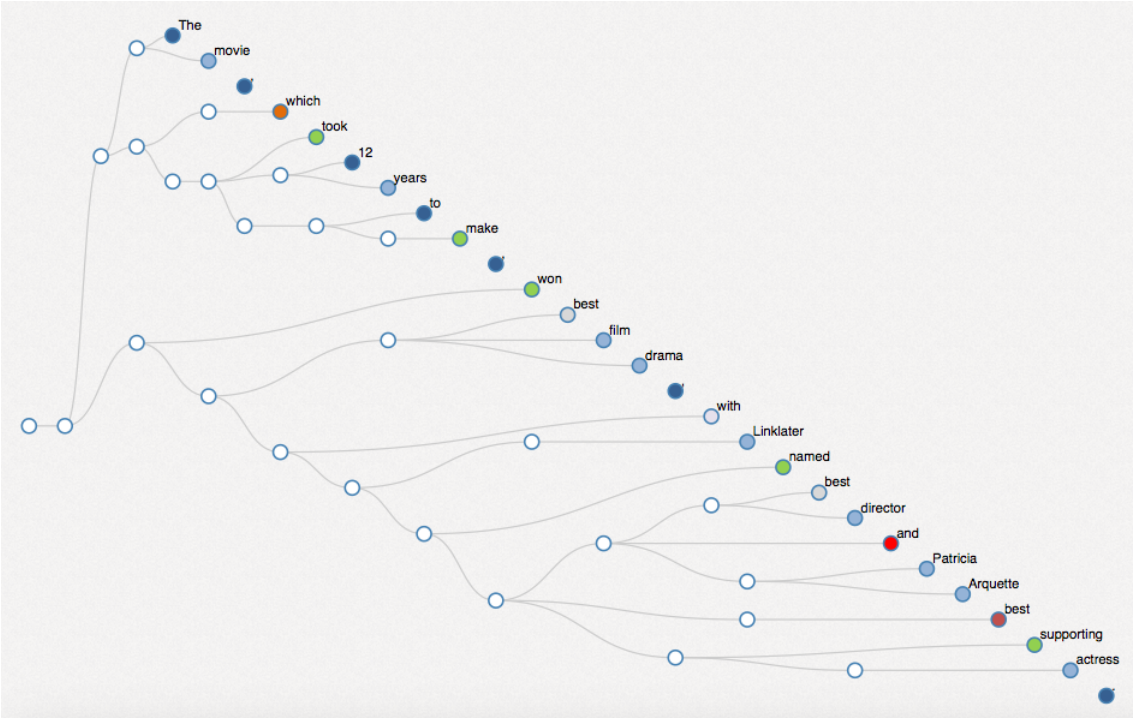

Fig. 7: A sentence tree

the software and its various applications in their teaching. Our pilot activities include participants from various universities in the U.S. and also international users from the Philippines, Korea, and China. Pilot users report that Text X-Ray contributes a wealth of data and information for teachers and learners by providing an effective visualization and description of written texts across academic genres. The software, in its most basic application, can show ESL learners the actual use and context of particular parts of speech (e.g., nouns, verbs, adjectives, and adverbs) in a text. If a specific objective in a course, perhaps a course in second language grammar for international students in the U.S., focuses on the use of a certain feature such as 'existential there', Text X-ray can provide a quick access to many examples from the corpus for students to analyze (i.e., data-driven learning applications). This feature can build awareness of a form's construction and typical placement at sentence-level, paragraph-level, and even entire composition-level writing. This color-coded visualizer helps users to focus on these 'tagged' features easily within the same text or group of texts. A sample lesson or activity in using Text X-Ray as a tool for students to assess the level of academic complexity in their writing is briefly described below.

Students can use Text X-Ray as part of a peer review activity to explore and evaluate each other's work or the academic essays already pre-loaded in the software. Text X-Ray can help ESL learners develop their academic writing skills by identifying common linguistic patterns and allowing opportunity for growth in complexity at word and sentence-level. Instructors may develop a set of evaluation criteria to guide students in their peer review activity. For example (steps for teachers):

- Choose appropriate context addressing the overall purpose of the lesson.
- Evaluation guides should include which tool the students should use and what they are looking for.
- Students may explore essays loaded on Text X-Ray and evaluate structures based on criteria outlined by instructor.

Readability Tool
Check for academic complexity by highlighting complex words and sentences in text visualization. (Here the instructor would provide the amount of syllables and member of words to be searched)

Parts of Speech Tool		
Highlight the personal pronouns. Are personal pronouns appropriate for academic writing?	Highlight the verbs Is the correct tense being used? Agreement?	Highlight the articles Are they being used correctly?

Fig. 8: Text X-Ray as readability tools and POS tools

A teacher following similar instructions as above may also focus on the use of the Academic Word List (AWL) to encourage learners to produce more academic language and to monitor their use of words from the AWL in their

writing. Learners will be encouraged to "play" around more with the Readability and Visualization functions (Fig. 2-8), allowing them to see the amount of complex words and sentences typical in academic writing. Other ways highlighting grammatical categories could also be used to tailor an activity to what the teacher is working on in class at a particular point in the semester. For example, a teacher might want to encourage learners to compose more compound sentences using coordinating conjunctions, to proofread for misuse of prepositions, etc. Being able to easily track their use of these categories throughout their writing would be valuable to students.

We are conducting a user study to evaluate the effectiveness of Text X-Ray in classroom teaching. Some of the preliminary user feedback are shown below.

CM, Doctoral Student

"I played with this program quite a lot and found many things that I liked about it. It's user-friendly and straightforward. It goes beyond POS, giving information about readability and offering an opportunity to compare the user's sample with other corpora."

JX ,Visiting Scholar - China

"Using Text X-Ray can highlight how native speakers of English use certain language forms, vocabulary items, and expressions. It offers students the use of authentic and real-life examples when learning writing which are better than examples that are made up by the teacher. It allow students to learn useful phrases and typical collocations they might use themselves as well as language features in context which means that students learn language in context and not in isolation. And it can help students get a broader view of language by comparison. By doing so, students become aware of lexical chunks that are useful when it comes to completing writing tasks. It helps teachers to demonstrate how vocabulary, grammar, idiomatic expressions and pragmatic constraints with real-life language."

JH, ESL Teacher – Korea

"Comparing with other programs, it is VERY user-friendly. I thought that I could use concordancers only when I prepare the class, but I thought I won't recommend students to use this kind of program before I saw the text x-ray. However, this test x-ray changed my thoughts. It is colorful and it is very easy to use. Without tagging, if students can find the nouns, verbs, and adjectives, I could use it when I teach verb valency to students. Since my interest is teaching grammar using corpus, I mainly thought of the methodology that I can use for grammar teaching. Because of this visual recognition on the screen of Text x-ray, I think students' learning will last than simple rote memory."

VI. Conclusion

Text X-Ray is a Web-based text visualization program for Corpus-based English language learning and teaching. Students can use Text X-Ray to explore a corpus for language usage patterns. They can also compare their own work with similar articles in a corpus. A teacher can use Text X-Ray for corpus-based language teaching. She can also use Text X-Ray to analyze a corpus of her students' writings. For example, she can also compare her students' works with another corpus of students' writings.

The main difference between Text X-Ray and other corpus tools is the text visualization interface. The text visualizations allow users to quickly examine a corpus at multiple levels of detail. Users can quickly see the complexity of texts by comparing sentence length, sentence grammar tree, word length, word frequency, and part of speech (POS) distribution. These visualizations, combined with traditional linguistic analysis, give users a more complete picture of the texts.

Text visualization is rarely used corpus tools and this work is an attempt to fill this gap. In the near future, we plan to enhance Text X-ray with more features, such as a concordancer and better filtering functions. We are also experimenting with new techniques for visualizing text complexity.

References

[1] U. Römer, "Corpus research applications in second language teaching," *Annual Review of Applied Linguistics*, vol. 31, pp. 205–225, 2011.

[2] H. Nesi and S. Gardner, *Genres across the disciplines: Student writing in higher education*. Cambridge University Press, 2012.

[3] A. O'keeffe, M. McCarthy, and R. Carter, *From corpus to classroom: Language use and language teaching*. Cambridge University Press, 2007.

[4] "BYU," http://corpus.byu.edu, accessed: 2015-01.

[5] L. Anthony, P. Crosthwaite, T. Kim, T. Marchand, S. Yoon, S.-Y. Cho, E. Oh, N.-Y. Ryu, S.-H. Hong, H.-K. Lee *et al.*, "A critical look at software tools in corpus linguistics," *Linguistic Research*, vol. 30, no. 2, pp. 141–161, 2013.

[6] "Compleat Lexical Tutor," http://www.lextutor.ca/, accessed: 2015-01.

[7] "WORD AND PHRASE," http://www.wordandphrase.info/, accessed: 2015-01.

[8] "Text visualization browser," http://textvis.lnu.se/, accessed: 2015-01.

[9] H. Siirtola, T. Säily, T. Nevalainen, and K.-J. Räihä, "Text Variation Explorer: Towards interactive visualization tools for corpus linguistics," *International Journal of Corpus Linguistics*, vol. 19, no. 3, pp. 417–429, 2014.

[10] J. Nualart-Vilaplana, M. Pérez-Montoro, and M. Whitelaw, "How we draw texts: A review of approaches to text visualization and exploration," *El profesional de la información*, vol. 23, no. 3, pp. 221–235, 2014.

[11] "Text X-Ray," http://textvis.gsu.edu:8080/NLPVis/html/index.html, accessed: 2015-01.

[12] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," *In Proceedings of the ACL conference*, 2013.

[13] "jQuery UI," http://jqueryui.com/, accessed: 2015-01.

[14] "D3.js," http://d3js.org/, accessed: 2015-01.

[15] X. He and Y. Zhu, "An indented level-based tree drawing algorithm for text visualization," *submitted to Graphics Interface Conference*, 2015.

[16] E. M. Reingold and J. S. Tilford, "Tidier drawings of trees," *IEEE Transactions on Software Engineering*, no. 2, pp. 223–228, 1981.

[17] J. Bloch, "A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers," *Journal of writing research*, vol. 2, no. 2, pp. 219–244, 2010.

[18] N. Nesselhauf, "The use of collocations by advanced learners of english and some implications for teaching," *Applied linguistics*, vol. 24, no. 2, pp. 223–242, 2003.

[19] T. Salsbury and C. Crummer, "Using teacher-developed corpora in the cbi classroom," *English Teaching Forum*, vol. 46, no. 2, 2008, pp. 28–37.

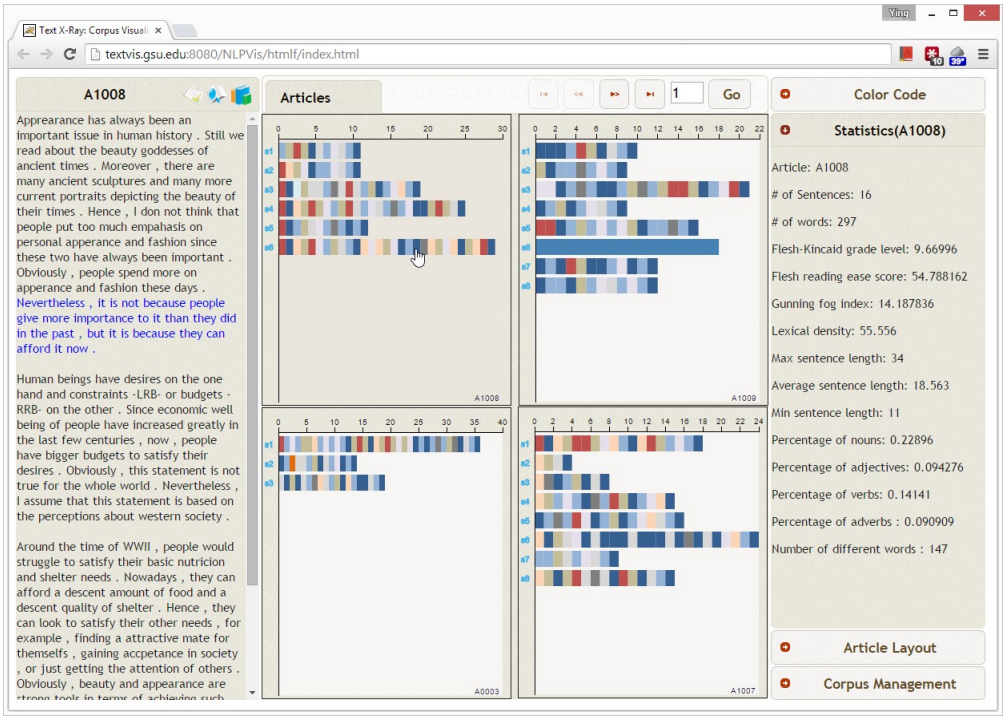

Fig. 9: Text X-Ray user interface. Users can quickly compare sentence lengths and word usage in the visualization. The selected sentence bar is highlighted in the original text.

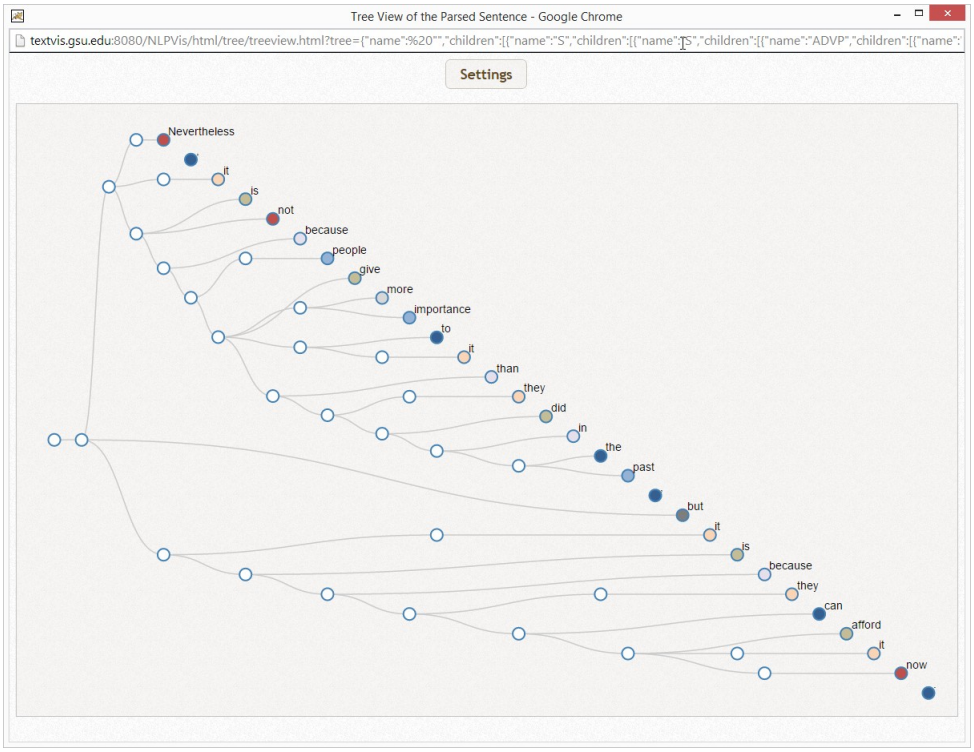

Fig. 10: The selected sentence grammar tree is visualized. The leaf nodes in the tree are color coded in the same way as the word blocks.