

<https://helda.helsinki.fi>

Information visualization for corpus linguistics: Towards interactive tools

Siirtola, Harri

2010

Siirtola , H , Räihä , K-J , Säily , T & Nevalainen , T 2010 , Information visualization for corpus linguistics: Towards interactive tools . in S Liu , M X Zhou , G Carenini & H Qu (eds) , First International Workshop on Intelligent Visual Interfaces for Text Analysis : Proceedings . ACM , New York , pp. 33-36 , First International Workshop on Intelligent Visual Interfaces for Text Analysis (IVITA) , Hong Kong , China , 07/02/2010 . <https://doi.org/10.1145/2002353.2002365>

<http://hdl.handle.net/10138/136254>
10.1145/2002353.2002365

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Information Visualization for Corpus Linguistics: Towards Interactive Tools

Harri Siirtola, Kari-Jouko Räihä
TAUCHI

Department of Computer Sciences
University of Tampere
{harri.siirtola,kari-jouko.raihä}@cs.uta.fi

Tanja Säily, Terttu Nevalainen
VARIENG

Department of English
University of Helsinki
{tanja.saily,terttu.nevalainen}@helsinki.fi

ABSTRACT

In this paper linguists and researchers of visual data analysis outline the requirements and benefits of an information visualization approach for corpus linguistics. Over the years, the information visualization community has come up with a number of methods to visualize text, but the majority of these techniques do not serve the needs of the linguistic community. This is evident in the over-simplification of the linguistic problems and generally caused by a poor understanding of the domain. We started a joint research effort with linguists, data miners, and information visualizers to design and produce better data analysis tools for corpus linguistics. This work is still in its early stages, but we have a shared vision of what needs to be done.

Author Keywords

corpus linguistics, information visualization

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

INTRODUCTION

Corpus linguistics is the study of language use by means of large electronic text collections, or corpora [2]. These are carefully compiled to ensure representativeness across desired features such as time, genre and the social status of writers/speakers. Nowadays corpora are often annotated for, e.g., part of speech or sentence structure; however, there is a lack of sophisticated tools for visualizing and analyzing such tagged and parsed data.

Information visualization is about using external tools to amplify cognition. Often these external tools are visual as more information is acquired through vision than via all the other senses combined [15, p. 2]. Visual and interactive representations of data improve problem solving and acquisition of insight.

© ACM, 2010. This is the author’s version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in Proceedings of the First International Workshop on Intelligent Visual Interfaces for Text Analysis, <http://doi.acm.org/10.1145/2002353.2002365>

Text as data is more challenging to visualize than numerical, nominal or categorical data. Text is high-dimensional and, e.g., equality tests are complicated because of multiple meanings and complex relations. Often the non-linguistically motivated text visualizations take shortcuts by ignoring the ordering relationships within the text, and by stemming the words (i.e., reducing them into their roots).

Here we will discuss our linguistically motivated visualizations for corpus linguistics. Although general-purpose visualization techniques provide a good starting point, techniques that dig deeper into the structure of the documents in the corpus, and work bottom-up from the texts, are needed to gain insight into linguistic variation and change.

Our work is based on the part-of-speech or POS-tagged version of the *Parsed Corpus of Early English Correspondence* (PCEEC) [9]. It is used as a running example in this paper whenever the copyright allows. The corpus consists of 4,968 letters written between the years 1415 and 1681, and has 2,155,446 words. The part-of-speech tagged text has each word marked up according to its definition and context, e.g., ‘Hopkins_NPR’ denotes that ‘Hopkins’ is a ‘proper noun’. Although PCEEC is relatively small as a corpus, it is challenging to analyze because of the variations over time. In cases where the PCEEC copyright would be compromised, the freely available plain text version of *The Adventures of Sherlock Holmes* [3] by Sir Arthur Conan Doyle is used as example material.

TEXT VISUALIZATION

Text data comes in many forms: articles, books, novels, letters, web pages and blogs, just to name a few. In addition to texts created by a human author there are many computer-generated text genres as well, such as log files and other outputs from computer programs.

Text visualization is popular, both as an object of research and among the ‘consumers of visualizations’. About one third of the user-created visualizations on the *Many Eyes* [8, 4] collaborative visualization service are related to text visualization, and media both in print and on the web routinely use such techniques as tag clouds and thematic maps to illustrate their texts (see, e.g., [12]).

The Many Eyes service has four text visualization modules in its selection (Figures 1 to 3 below). They provide a rep-

representative sample of how the visualization community generates insight into text documents. Figure 1 shows a *Wordle* [5] visualization that illustrates the occurrence of words in PCEEC. In a *Wordle* visualization, the size of a word is determined simply by its frequency, and the placement of a word does not convey any additional information.



Figure 1. Wordle visualization of the two million words of the *Parsed Corpus of Early English Correspondence*.

The *Wordle* visualization in Figure 1 is impressive in that it manages to summarize almost 5,000 letters in just one picture. Although it is thought-provoking and entertaining, its worth from the linguistic point of view is questionable. As the designer of *Wordle* notes [14], a significant number of *Wordle* users do not even understand what the graphics are encoding, and a user might seek explanation for the proximity of certain words when they are just put together randomly. This is not to be taken as criticism of *Wordle*, as it does exactly what it was designed to do, and there are interesting applications for it. It has been used in the domain of text corpus visualization as well, to get an overview of Shakespeare's sonnets and plays. It might also be interesting to compare *Wordle* visualizations of PCEEC material per letter or per author, as in Feinberg's comparison of the inaugural addresses of the presidents of the United States [6]. In addition to *Wordle*, *Tag Cloud* is another word frequency visualization from the Many Eyes service. In a tag cloud the words of a text are laid out in alphabetical order and their size is in proportion to their frequency.

Figure 2 displays another text visualization, *Word Tree*, from the Many Eyes service which is a visual version of a traditional concordance (a list of words in their context). Instead of a traditional list view of such tools, the data is displayed as a tree. The primary difference besides the representation is that the tool displays only the right hand side of the current word's environment. However, selecting a word will permit exploration of the left hand side environment as well.

In a *Phrase Net* visualization, as seen in Figure 3, the user selects one of the pre-defined 'bigrams' or two-word patterns or creates a new one. The patterns define the elements that should appear between two words, and the tool creates a graph of all word pairs that match the pattern. Size is again used to encode frequency.

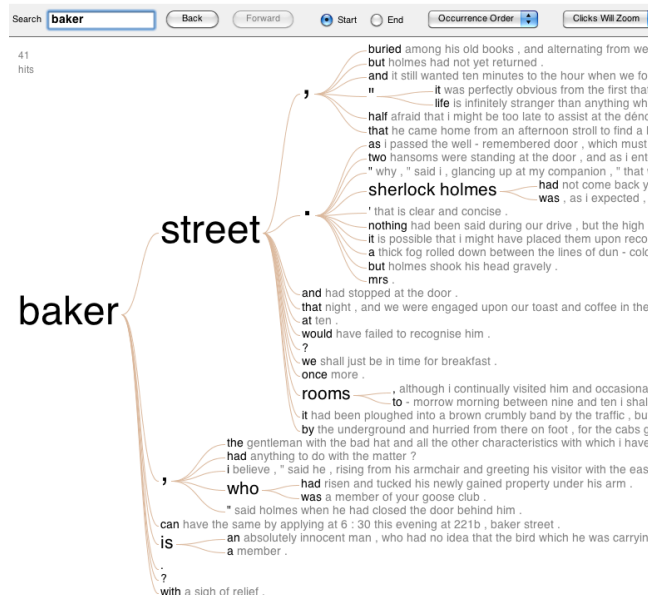


Figure 2. Word Tree visualization of *The Adventures of Sherlock Holmes* with focus on the word 'baker'.

SCENARIO

While Many Eyes can be a useful toolset for corpus linguistics, it is not designed to utilize annotation in corpora. This section presents a scenario of how the information visualization approach might be applied to solving a linguistically motivated problem in a POS-tagged corpus. Two open-source and freely available software tools are employed: the general statistical data-visualization system *Mondrian* [13] and the statistical system *R* [10].

Suppose that the aim of the study is to investigate if the 'nouniness' of language is affected by the sociolinguistic background variables as manifested in the PCEEC corpus. Nouniness, or the proportion of nouns in a text, can be determined in a part-of-speech tagged corpus by computing the percentage of words tagged as nouns against the whole corpus. The question of which of the tags are regarded as nouns is a matter of definition and open to some debate.

The first concern in a study of a POS-tagged text is data integrity, to make sure that the phenomenon under study is correctly encoded in the corpus. In this case, it means checking that the words tagged as nouns are really nouns, and that tokenization (how the text is segmented into words) is handled correctly. With a historical corpus, this step may involve many manual operations and computer scripts that 'prune' the corpus, and the result is a refined version of the corpus. Suppose the result from this step is in the following form (only 6 out of 2,154,210 lines are shown):

	word	tag
1	Mr.	NPR
2	Hopkins	NPR
3	yow	PRO
4	discourse	VBP
5	wisely	ADV
6	and	CONJ

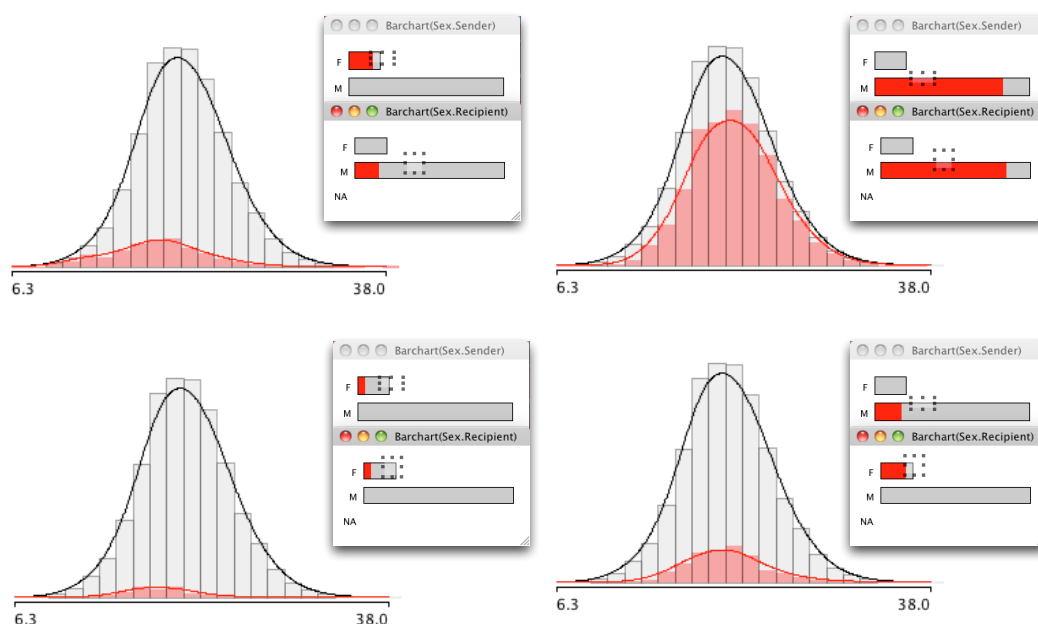


Figure 4. The effect of gender of letter sender/recipient on the ‘nouniness’ of text explored with the interactive data visualization system Mondrian.

REFERENCES

1. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3):321–346, 2003.
2. D. Biber, S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.
3. A. Conan Doyle, Sir. *The Adventures of Sherlock Holmes*, volume 1661 of *Project Gutenberg*. Project Gutenberg, 1999 [George Newnes Ltd, 1892].
4. C. M. Danis, F. B. Viegas, M. Wattenberg, and J. Kriss. Your place or mine? Visualization as a community component. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI Conference on Human factors in Computing Systems*, pages 275–284. ACM, 2008.
5. J. Feinberg, *Beautiful Word Clouds*, 2009. <http://www.wordle.net/>.
6. J. Feinberg, *Inaugural Addresses*, 2009. <http://www.research.ibm.com/visual/inaugurals/>.
7. S. T. Gries. *Quantitative Corpus Linguistics with R*. Routledge (Taylor and Francis), New York, 2009.
8. IBM Visual Communication Lab, *Many Eyes*, 2009. <http://manyeyes.alphaworks.ibm.com/manyeyes/>.
9. A. Nurmi, A. Taylor, A. Warner, S. Pintzuk, and T. Nevalainen. Parsed Corpus of Early English Correspondence (PCEEC), tagged version. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive, 2006.
10. R Development Core Team. R: A language and environment for statistical computing. <http://www.R-project.org/>, 2009.
11. P. Rayson, G. Leech, and M. Hodges. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.
12. The New York Times Visualization Lab, 2009. <http://vizlab.nytimes.com/>.
13. M. Theus and S. Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2008.
14. F. B. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1146, 2009.
15. C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufman, San Francisco, CA, second edition, 2004.
16. H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), November 2007.
17. G. Wills. Selection: 524,288 ways to say “this is interesting”. In *InfoVis'96: Proceedings of the IEEE Symposium on Information Visualization 1996*, pages 54–61. IEEE Computer Society, 1996.