The Story of One: Humanity scholarship with visualization and text analysis

Tanya Clement*, Catherine Plaisant#, Romain Vuillemot#
*Dept. of English; #Human-Computer Interaction Laboratory
University of Maryland, College Park, USA
tclement@umd.edu; plaisant@cs.umd.edu; romain.vuillemot@insa-lyon.fr

Most critiques of *The Making of Americas* (Paris 1925) by Gertrude Stein contend that the text deconstructs the role narrative plays in determining identity by using indeterminacy to challenge readerly subjectivity. The current perception of *Making* as a postmodern text relies on the notion that there is a tension created by frustrated expectations that result from the text's progressive disbandment of story and plot as the narrative unweaves into seemingly chaotic, meaningless rounds of repetitive words and phrases. Yet, a new perspective that is facilitated by digital tools and based on the highly structured nature of the text suggests that these instabilities can be resolved by the same seemingly nonsensical, non-narrative structures. Seeing the manner in which the structure of the text makes meaning *in conversation with* narrative alleviates perceived instabilities in the discourse. The discourse about identity formation is engaged—not dissolved in indeterminacy—to the extent that the reader can read the composition.

One method for reading the composition of the text without relying on what becomes a non-existent framework based on plot is to view the progression of words according to a different framework, a framework that relies on comparative associations based on word usage. Using *WordHoard*,[i] we compared word usage between texts and text parts by calculating the log-likelihood ratio, which describes the size and significance of the difference between word frequencies in a base text versus a reference text.[ii] In this analysis, we measured word usage in *The Making of Americans* in comparison to two different sets of reference texts with more traditional narrative structures: (a) a set of 19th century novels written by Jane Austen, Charles Dickens, George Eliot, and George Meredith; [iii] (b) between the first and second half of *Making*, which it has been argued also represents different narrative trends (Clement 2008). Visualizing this information in *Wordle*[iv] —a word cloud application (Wattenberg & Viegas, 2008)—is useful primarily because it provides a visual overview of word frequencies that is easy to understand and to publish for reference. The *Wordle* application facilitates this kind of analysis by visualizing the list of words in a cloud that maximizes the space utilization on a computer screen by sizing the words by their relative frequencies. The more frequently a word occurs in a particular text (relative to another text) the larger the word appears. In the set of visualizations that accompany this discussion, [v] each cloud serves to visualize words that are more or less frequent in any given comparison (see examples in Fig. 1). What becomes immediately evident in comparing these visualizations is the prominence of a particular word that consistently scores a high value in terms of discrepancy between *Making* and the reference texts: *one*.

The word *one* appears consistently across every cloud that marks the words that are more common in *Making* and less common in the sample of nineteenth century texts and words that are more common in the second half rather than the first half of the text. We then compared the relative frequencies of multiple pronouns, which revealed that the frequency of *one* surges by the end of the text (Fig. 2). What is most interesting about this graph is that the high frequency of *one* is the result of the confusion accomplished by the word's schizophrenic nature. Words here are represented according to occurrence, not to type of occurrence. Thus, the word *one* —unlike

he, *she*, *I*, *we*, or even *you* or *it*—which plays many positions in the text, in the role of a pronoun or an adjective and in the subject or object position, surges in frequency.

To better understand how the word *one* is used in the text, we created another set of visualizations prepared using a prototype we developed called *PosViz.*[vi] *PosViz* allowed us to compare word usage based on parts of speech in individual chapters from *Making* to the whole text. These comparisons allowed us to isolate and analyze words used more and less frequently throughout the course of the novel and to measure how and if these patterns change within the text itself. In these visualizations, each part of speech for each word is treated as a separate word instance and each instance is placed according to its appearance in the text, color-coded according to its part of speech, and sized according to its overall frequency. Thus, by using *PosViz*, the progression of the manner in which the word *one* is used in terms of different parts of speech is documented, allowing us to see that the use of *one* appears to change as the text progresses. For example, a relatively small *one* appears three times in the chapter 1 cloud (Fig. 3). This visualization indicates that the occurrence of the word *one* has little variance in terms of how it is used (its part of speech) and occurs relatively infrequently in occurrences that are localized to the beginning paragraphs of the chapter.[vii] By chapter 9, however, *one* dominates the discourse both in terms of its frequency and in terms of its multiple uses (Fig. 4).

By identifying the manner in which word usage changes in correlation to the presence and absence of narrative both in comparison to other novels and within the text itself, these comparisons enable a new perspective on the meaning-making processes of the text's composition. For example, these visualizations illustrate the nature of the word *one* as it is used to heighten the word's propensity for different reading possibilities. This lends to a reading in which *one* may represent a singular subject position or multiple subject positions at once. With this information, a further argument can be made that the discourse about identity formation is engaged in this multiplicity, not dissolved in indeterminacy. Thus, employing composition in her representation of identity formation in *The Making of Americans* becomes the method by which Stein seeks to represent identity, but if and how the reader is able to recognize and interpret this endeavor is predicated by her ability to see it.

This work with *The Making of Americans* is part of research and development within the MONK (Metadata Offer New Knowledge) project, a Mellon-funded collaborative seeking to develop text mining and visualization software in order to explore patterns across large-scale text collections. Stein's text was a productive text for analysis during the beginning phases of the MONK project since its many and complicated repetitions could be processed and visualized.[viii] This presentation focuses on how the process of determining decision criteria for text mining led to the discovery that various textual features (n-grams, parts-of-speech, and log-likelihood ratios) and various visualizations (*FeatureLens*, *Spotfire*, *Wordle*, and *PosViz*) ultimately facilitated an iterative discovery process and a new reading of Gertrude Stein's *The Making of Americans*.

Appendix:

A

B

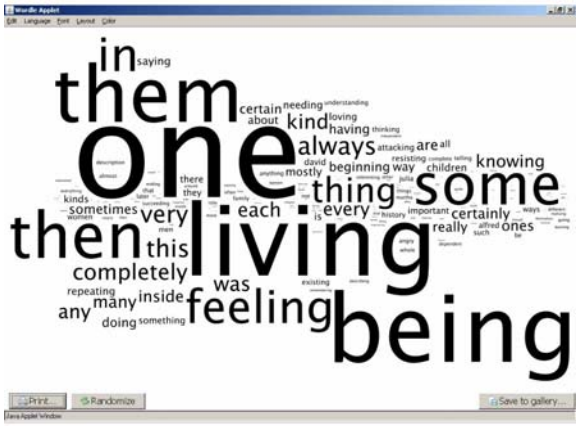

Fig. 1: Comparsions between *Making* and novels by George Meredith, using log-likelihood ratios from *WordHoard* and visualizations produced in *Wordle*; A: words that are more common in Meredith; B: words that are more common in *Making*.

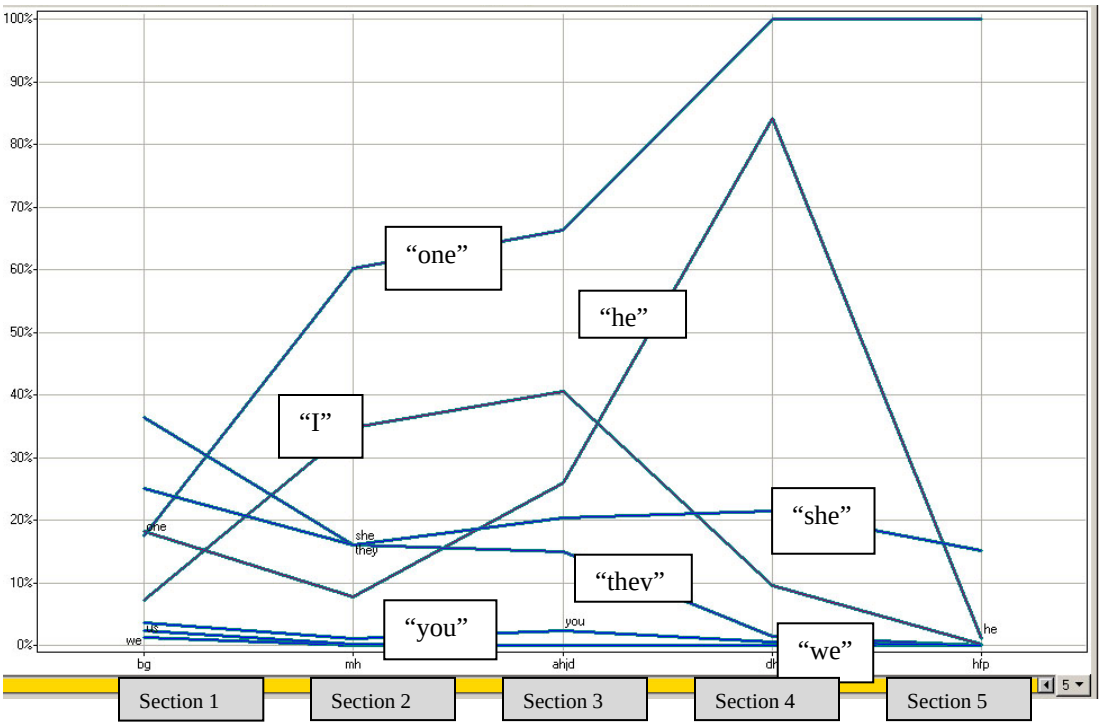

Fig. 2: The frequency of "one," "I," "you," "we," "he," "she," and "they" are mapped across the five sections of the text in *Spotfire*

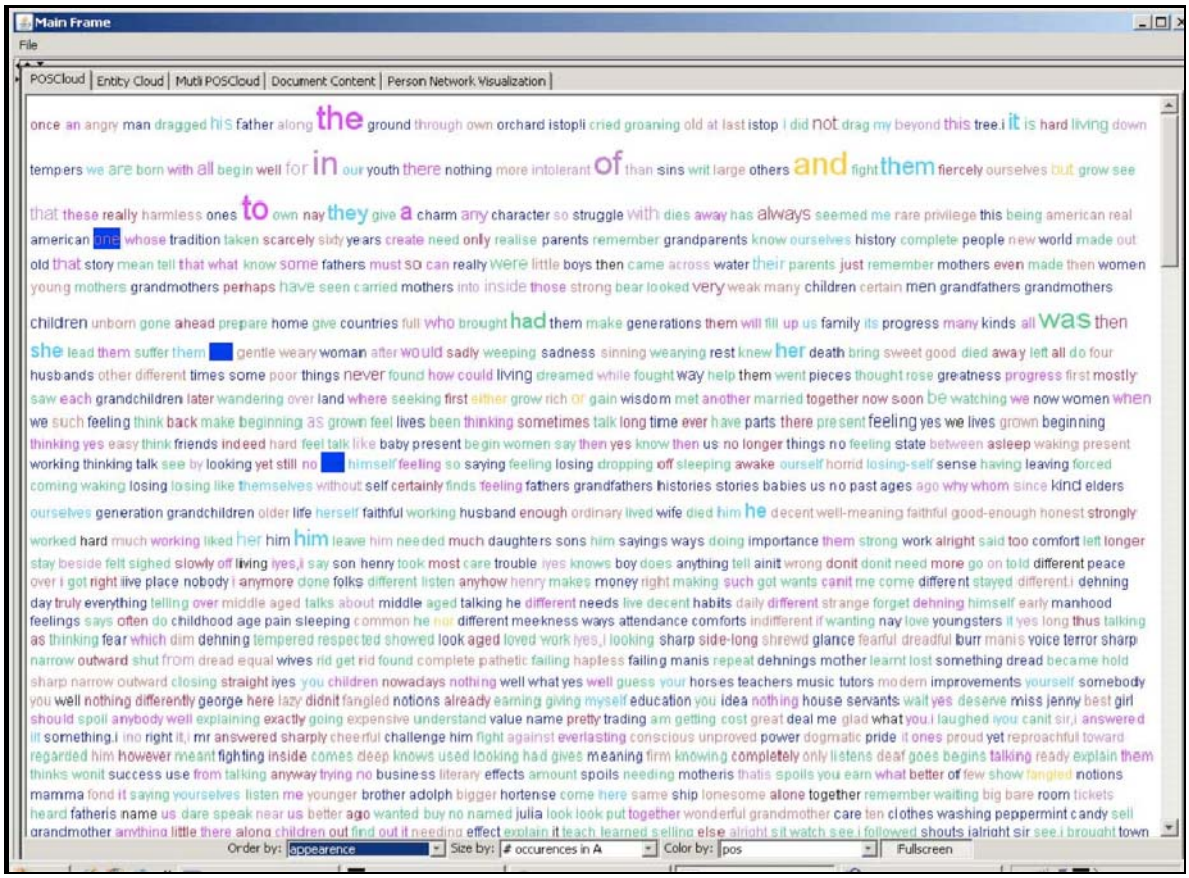

Fig. 3: Chapter 1, the word "one" highlighted in *PosViz* visualization

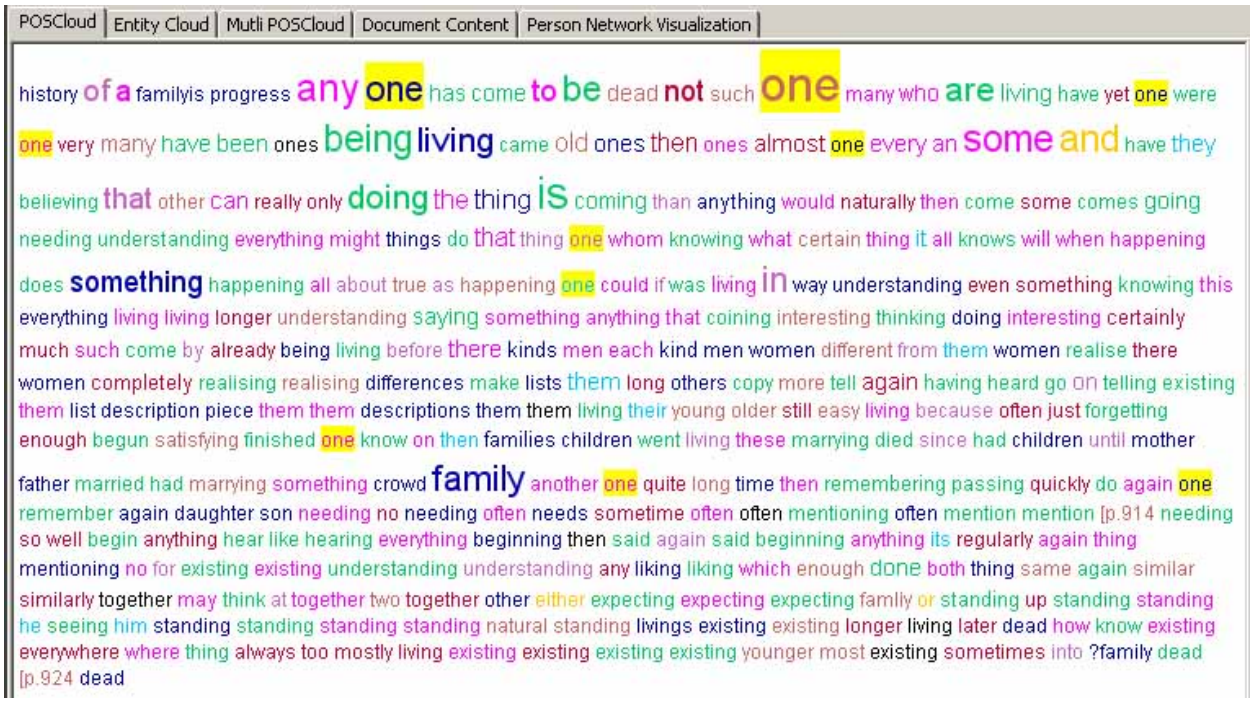

Fig. 4: Chapter 9, the word "one" highlighted in *PosViz* visualization

---

Notes

[i] Please see http://wordhoard.northwestern.edu/userman/.
[ii] This analysis is based on Dunning's log-likelihood analysis. Please see http://wordhoard.northwestern.edu/userman/analysis-comparewords.html#loglike.
[iii] The books used in this study are those available in the *WordHoard* application. They are listed at http://terpconnect.umd.edu/~tclement/less-more.htm. These authors were chosen because Stein repeatedly compares her text to their novels. See Stein 1990, p. 506.
[iv] Please see http://wordle.net/.
[v] These visualizations are pictured in an online appendix entitled "Visual Comparisons of Gertrude Stein's *The Making of Americans* using *WordHoard*, *Wordle*, and *PosViz*" at http://terpconnect.umd.edu/~tclement/less-more.htm. In these slides, the comparisons have been visualized with four sets of data for each data set, all of which are set in comparison to the base text *The Making of Americans* and include and exclude 'common words' such as articles, conjunctions, and pronouns. The full list of these 'stop-words' is unavailable, but the creator Jonathan Feinberg has indicated in an email that 'I have modified them by hand over time. The English one came from the Snowball stemmer project' at http://snowball.tartarus.org/algorithms/english/stop.txt (personal correspondence).
[vi] Currently, *PosViz* does not have a web presence.
[vii] The analysis program used to label these uses is part of the SEASR (Software Environment for the Advancement of Scholarly Research) analytic routines (http://seasr.org/). Though imperfect, the system is consistent—it labels the same behaviors the same way each time. Thus each occurrence represents the *perception* of a different use of the word *one*.
[viii] This work is published in two articles: Don et al., 2007 and Clement, 2008.

References

Clement, T. (2008). "'A thing not beginning or ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*." *Literary and Linguistic Computing*, 23.3: 361-382.
Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C. (2007). "Discovering interesting usage patterns in text collections: integrating text mining with visualization." *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 213-222.
Stein, G. (1990). "Transatlantic Interview 1946." In *The Gender of Modernism*. Bonnie Kime Scott and Mary Lynn Broe (eds). Bloomington: Indiana University Press, pp. 502-516.
Wattenberg, M. and Viegas, F. (2008). "Tag clouds and the case for vernacular visualization." *Interactions*, 15.4: 49-52.
Stein, G. (1995). *The Making of Americans: Being a History of a Family's Progress*. Normal, IL: Dalkey Archive Press.
Weiss, Sholom M. *et al.* (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer, pp. 85-86.