GutenTag: A User-Friendly, Open-Access, Open-Source System for Reproducible Large-Scale Computational Literary Analysis

Adam Hammond
adam.hammond@utoronto.ca
University of Toronto, Canada

Julian Brooke
julian.brooke@unimelb.edu.au
University of Melbourne, Australia

Introduction

GutenTag is a cutting-edge resource that allows literary researchers of all levels of technical expertise to perform large-scale computational literary analysis. It allows users to build large, clean, highly customized worksets and then either analyse them in-system or export them as plain text or richly-encoded TEI. It has been built from the ground up by literary scholars for literary scholars: rather than relying on off-the-shelf tools poorly suited to the domain of literature, we have developed many of the components ourselves based on the specific demands of literary research. GutenTag is fully open-source, its analyses are based on entirely open corpora, and researchers can save and distribute all the parameters of their analyses, allowing for unprecedented reproducibility of research in a field plagued by siloed corpora. GutenTag is easy to use, permitting casual non-programmers to perform complex computational literary analysis via an online interface, while offering additional offline customization options to more advanced users. Although GutenTag was initially designed to facilitate our own research in polyvocality and dialogism, we show here that it can be leveraged to intervene in pressing debates unrelated to our specific research, such as the discussion surrounding Matthew Jockers's analysis of gender in *Macroanalysis*.

Overview of GutenTag

The system has grown considerably since our initial proposal, presented to an audience of computer scientists (Brooke et al., 2015). Below, we review the main features of the software with particular emphasis on recent improvements.

Interface: GutenTag is primarily accessed through an HTML GUI, accessible via the web or as a downloadable tool (both can be accessed from http://www.projectgutentag.org). In offline mode, the configuration files can be saved and loaded, and additional lexicons and other lists used for analysis can be specified by the user. A Python API is also included.

Corpora: The original version supported only the 2010 image of Project Gutenberg USA, but we have expanded support to all texts from Project Gutenberg USA as well as Project Gutenberg Canada and Australia, which include many additional texts published after 1922 and still under copyright in the USA.

Metadata: Document collections of interest can be defined using a variety of metadata tags. These include metadata provided by Project Gutenberg (title, author, author birth, author death, and, for some texts, Library of Congress classification and subjects). We have added genre (fiction, non-fiction, poetry, drama), determined using a sophisticated machine classifier, as well as author and text information (author gender, author nationality, publication date, publication country, single work or collection, etc.) derived from (mostly) unstructured resources including Wikipedia and the texts themselves.

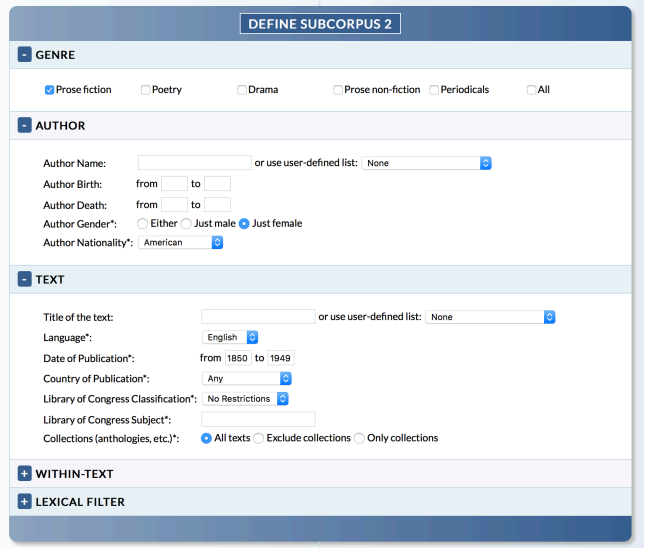

Figure 1: The GutenTag interface, showing the creation of a workset based on advanced metadata (Genre, Author Sex, Author Nationality, Date of Publication)

Text cleaning and tokenization: Sophisticated regex-based heuristics are applied to remove meta-text elements related to Project Gutenberg before, after, and sometimes within the text boundaries. Literature-specific tokenization is provided, preserving important information needed for downstream analysis.

Structural Tagging: This module identifies the main structural elements of the texts. First, heuristics are used to identify the likely boundaries between front matter, body, and back matter. Identification of structure within the main text is driven primarily by the identification of headers, and fully supports recursive structures including entire embedded texts which can have their own front and back matter separate from that of the anthology. Structural tagging is sensitive to genre: in the context of fiction, we identify parts, chapters, and speech; for poetry, we identify poems, cantos, stanzas, and lines; for drama, we identify acts, scenes, speakers, speech, and stage directions.

Lexical tagging: GutenTag includes lemmatization and POS tagging. There are several built-in lexicons which capture semantic and stylistic distinctions, and users can define their own lexicons, including multiword lexicons. Most recently, and most relevant to our case study below, we have added our own state-of-the-art literature-specific named entity recognition system (LitNER) which bootstraps from context-based clustering of common named entities to distinguish previously unseen people and locations from other named entities (Brooke et al. 2016b). For fiction, we group individual person names into collections of characters, and then assign speech events to these characters in the vicinity, using efficient, rule-based logic inspired by work in He et al. (2013). We identify the indicated sex of these characters primarily using large lists of names and titles; when a name does not appear on our list, we fall back to matching common sex-indicative character n-grams automatically derived from those lists (e.g. names ending with "a" tend to be female).

TEI output: When corpus output is required, we use XML-based TEI format as the default output format when structure (rather than simply tokens) is requested.

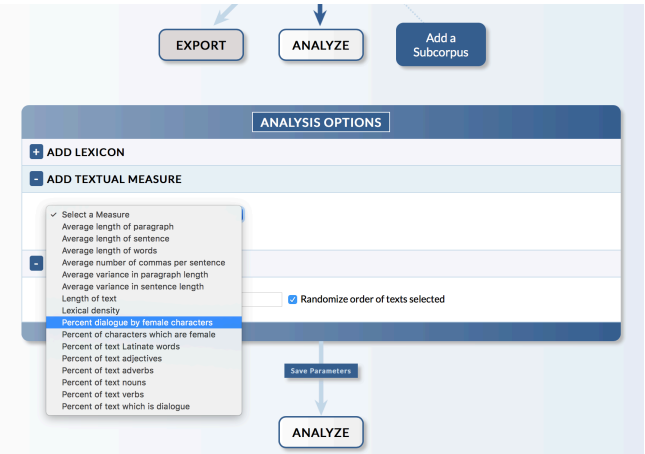

Figure 2: The GutenTag interface, showing in-system options for analysis via textual measure

Analysis: In addition to building corpora for exporting, GutenTag users can directly compare the distribution of relevant lexical tags or other textual metrics across multiple corpora as defined in the metadata filtering phase. The latest version includes a selection of standard textual metrics (e.g. average sentence length), part-of-speech based metrics such as lexical density, and metrics that rely on structural/lexical tagging, such as the amount of dialogue and the amount of dialogue that has been assigned to female characters. Advanced users can easily define their own textual metrics using Python; these then become available through the main interface. We also welcome requests for metrics from the DH community.

Research Applications

GutenTag was initially developed to facilitate our own research in literary dialogism (Hammond et al. 2016, Brooke et al. 2016a). GutenTag allows us to perform three crucial steps in our research process: first, to build customized corpora (a set of novels published from 1880-1950, for which it yields 4,088 results); second, to identify passages of character speech in each novel and assign a unique character to each passage of speech; and third, to calculate a measure of dialogism for each text using an algorithm based on our six-style approach (Brooke et al. 2016a). Further, GutenTag allows us to save our workflow in a parameter file so that it can be reproduced by other researchers.

GutenTag is designed as a general system, however — not merely as a vehicle for our specialized research. We thus present an example of how it can be employed

(by a non-programmer) to investigate a prominent debate in Digital Literary Studies, Matthew L. Jockers's discussion of gender and authorship in Macroanalysis. Jockers argues that female authorship can be predicted reliably through topic modelling, based on the presence of themes that "correspond rather closely to our expectations and our stereotypes" such as "Affection and Happiness," "Female Fashion," and "Infants" (Jockers 2013). A reader might respond to Jockers's analysis by querying his assumptions about literary authorship; specifically, his failure to distinguish between authors and characters. Suppose that female characters were just as likely to discuss "Female Fashion" in novels written by men as those written by women, but that female authors tended to include more female character speech in their novels, as Muzny et al. (2016) suggest. If this were so, Jockers's findings would not confirm stereotypes about female authorship, but simply reveal the tendency of female authors to include more female voices in their texts than men.

GutenTag is uniquely suited to investigating such a question. Its advanced metadata and sophisticated lexical tagging allow it to easily and rapidly analyze the question of female character speech in a large corpus of English-language novels.

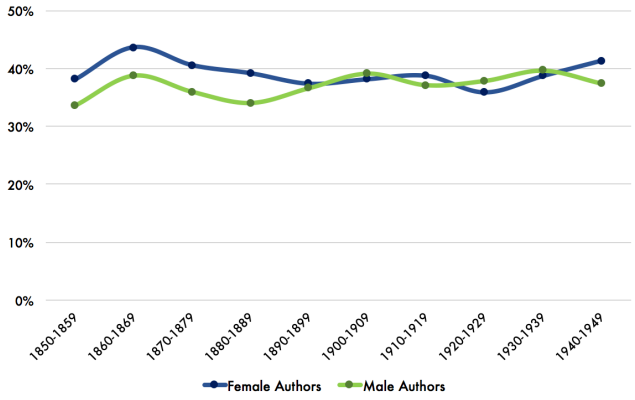

Figure 3: Mean proportion of text which Is dialogue in prose fiction, female vs. male authors, 1850-1949.

Sample sizes as follow, in number of texts. 1850-1859: 53 female, 97 male. 1860-1869: 86 female, 128 male. 1870-1879: 110 female, 137 male. 1880-1889: 122 female, 262 male. 1890-1899: 221 female, 583 male. 1900-1909: 299 female, 975 male. 1910-1919: 354 female, 960 male. 1920-1929: 148 female, 656 male. 1930-1939: 77 female, 413 male. 1940-1949: 52 female, 135 male.

Figure 3 shows that female authors in the twentieth century included approximately the same amount of dialogue as a proportion of total text length as male authors, but that in the latter half nineteenth century, they included approximately 5% more than men. Since Jockers focuses on the nineteenth century, this finding alone might impact his conclusions.

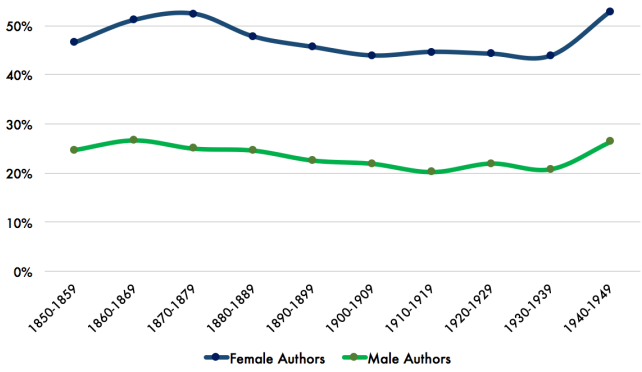

Figure 4: Mean proportion of dialogue allotted to female characters in prose fiction, female vs. male authors, 1850-1949

Sample sizes as follow, in number of texts. 1850-1859: 53 female, 97 male. 1860-1869: 88 female, 128 male. 1870-1879: 110 female, 137 male. 1880-1889: 122 female, 261 male. 1890-1899: 220 female, 583 male. 1900-1909: 300 female, 795 male. 1910-1919: 354 female, 960 male. 1920-1929: 148 female, 655 male. 1930-1939: 77 female, 413 male. 1940-1949: 54 female, 135 male.

As Figure 4 shows, GutenTag supports Muzny et al.'s contention that female novelists incorporate far more (approximately twice as much) female dialogue compared with male novelists. The finding that the proportion of female dialogue decreased from the late nineteenth to the mid-twentieth century, in both female and male authors, is one that bears further investigation — particularly in relation to the emergence in that period of popular genres, such as children's literature, Westerns, and romance novels.

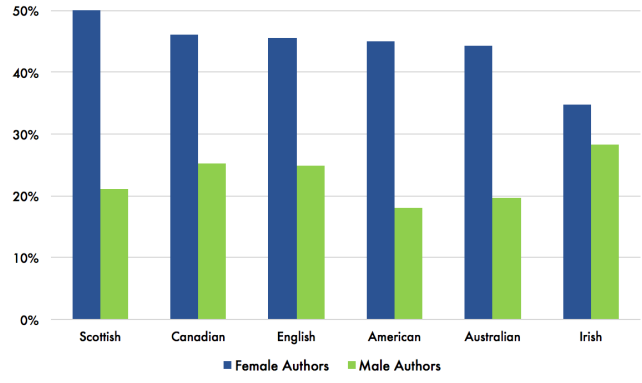

Figure 5: Mean proportion of dialogue allotted to female characters in prose fiction, female vs. male authors, by nationality, 1850-1949

Sample sizes as follow, in number of texts. Scottish: 31 female, 80 male. Canadian: 49 female, 78 male. English: 339 female, 1308 male. American: 572 female, 1545 male. Australian: 38 female, 104 male. Irish: 21 female, 92 male.

In Figure 5, we employ GutenTag's ability to filter results by author nationality. The marked discrepancy between proportion of female dialogue in male authors from England and the United States again suggests the need for an further investigation of genre; for instance, whether the American preference for male-centred genres like the Western might explain the result. Looking at GutenTag's fine-grained outputs, we observe that the texts with the lowest proportion of female dialogue are those directed at a young male audience (especially adventure fiction for boys) while those with the highest proportion consist largely of fiction for young women (L. M. Montgomery's Anne of Green Gables devotes over 90% of its dialogue to female characters). These findings might prompt our hypothetical researcher to engage in a smaller-scale study of the representation of gender in children's literature. Because all texts in GutenTag are accessible to users, it easily accommodates such movements from large-scale analysis to close reading.

Conclusion

GutenTag allows researchers of all levels of technical expertise to perform advanced large-scale literary analysis, as well as to independently test the hypotheses and conclusions of prominent research in the field. Our case study further shows how the integrated, end-to-end GutenTag system allows users to raise new research questions in the course of their analyses (such as the correlation between the emergence of children's fiction and the proportion of female dialogue) and then, since all its corpora are accessible, to shift scales and explore these questions through close reading.

Bibliography

Brooke, J., Hammond, A., and Hirst, G. (2016a). Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. *Digital Scholarship in the Humanities*, 2(2): 1–17.

Brooke, J., Hammond, A., and Baldwin, T. (2016b). Bootstrapped Text-level Named Entity Recognition for Literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL '16).

Brooke, J., Hammond, A., and Hirst, G. (2015). GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. *Workshop on Computational Linguistics for Literature*. Denver: NAACL, pp. 1–6.

Hammond, A., Brooke, J. (2016). Project Dialogism: Toward a Computational History of Vocal Diversity in English-Language Literature. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 543-544.

He, H., Barbosa, D. and Kondrak, G. (2013). Identification of Speakers in Novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL '13).

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History.* Urbana-Champaign: University of Illinois Press.

Muzny, G., Algee-Hewitt, M., Jurafsky, D. (2016). The Dialogic Turn and the Performance of Gender: the English Canon 1782-2011. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 296-299.