Digital Studies / Le champ numérique

Manning, Theodore Daniel, Eugenia Lukin, Ross Klein, and Patrick Juola. 2023. "Zesting Up Stylometry with MapLemon: A Corpus for Stylometric Demographic Identification," in "DH Unbound 2022, Selected Papers," ed. Barbara Bordalejo, Roopika Risam, and Emmanuel Château-Dutier, special issue. *Digital Studies/Le champ numérique* 13(3): 1–11. DOI: https://doi.org/10.16995/dscn.9665.

Open Library of Humanities

Zesting Up Stylometry with MapLemon: A Corpus for Stylometric Demographic Identification

Theodore Daniel Manning, City University of New York Graduate Center, US, tmanning@gradcenter.cuny.edu
Eugenia Lukin, Harvard University, US, citizengenia@gmail.com
Ross Klein, University of Pittsburgh, US, rak169@pitt.edu
Patrick Juola, Duquesne University, US, juola@mathcs.duq.edu

MapLemon is a corpus in its second iteration that was created to obtain a baseline corpus for linguistic variation among English-speaking North Americans. The MapLemon corpus currently houses upwards of 21,000 words across 185 participants, 10+ linguistic backgrounds, and 40+ US states and Canadian provinces. MapLemon also houses writing from 91 transgender and non-binary individuals. MapLemon presents a unique method for data collection in the virtual written medium and a corpus that has proven useful for identifying demographic information via writing style, otherwise known as stylometry.

MapLemon est un corpus en sa deuxième itération qui a été créé pour obtenir un corpus de référence des variations linguistiques parmi les anglophones d'Amérique du Nord. Le corpus MapLemon contient actuellement plus de 21 000 mots provenant de 185 participants de plus de 10 origines linguistiques et de plus de 40 États américains et provinces canadiennes. MapLemon contient également les écrits de 91 personnes transgenres et non binaires. MapLemon présente une méthode unique de collecte de données dans le domaine de l'écriture virtuelle et un corpus qui s'est avéré utile pour identifier des informations démographiques par le biais du style d'écriture, également connu sous le nom de stylométrie.

1. Introduction

MapLemon is a corpus in its second iteration that was created to obtain a baseline corpus for linguistic variation among English-speaking North Americans. The MapLemon corpus currently houses upwards of 21,000 words across 185 participants, 10+ linguistic backgrounds, and 40+ US states and Canadian provinces. It presents a unique method for linguistic data collection, as the HCRC Map Task Corpus once attempted this (University of Edinburgh 1993), however, not for the written medium. MapLemon additionally houses responses from 91 transgender and non-binary individuals, making it a fantastic resource for analyzing naturally elicited Queer writing.

We propose that MapLemon presents a unique form of data collection for the virtual written medium and a corpus that has proven useful for demographic identification via writing (stylometry) when the text is analyzed via the Java Graphical Authorship Attribution Program (JGAAP).

1.1 Scope

This paper will outline the current stage project MapLemon is in, suggest use-cases and hypothesize about results, and analyze the current data that has been gathered using the corpus. It will not examine the ethics surrounding demographic identification (including that of Queer people), nor will it defend the efficacy of stylometry (for more on the ethics of demographic identification, see Tomas, Dodier, and Demarchi 2022; for stylometry, see Neal et al. 2018). It, additionally, will not go into extensive detail about JGAAP; however, it will briefly be touched upon. Finally, it will not discuss Queer Theory at length.

1.2 Background

Stylometry uses parameters such as word choice and terminology to identify individual variation. Through this individual variation, one may be able to identify a person or a group of people based on their writing style. MapLemon was created for precisely this purpose; however, it was not created to be Queer-specific, despite many of the current responses being from Queer people. With MapLemon, we originally intended to do nationality and state identification at the very least, but when the responses came in and, through analysis in JGAAP, we found incidental evidence that transgender people may write more like their gender rather than their sex assigned at birth, we decided to pursue that line further and get an entire set of participants just to prove (or disprove) that hypothesis. We decided to create a new corpus rather than using an existing one because nothing existing really served our purpose; there was no existing corpus that contained naturally elicited electronic writing that contained elicitations for words

with common linguistic variation for all of North American English. The idea for a map and a recipe (hence "MapLemon") was created due to the terms used in both of these tasks being widely understood and also having a lot of terms with potential variations.

When this project was first presented, little analysis had been done on the corpus, which was still in its infancy at only 91 participants. No data had been purposely gathered from marginalized identities. The only analysis that had been carried out at the time of presentation was part-of-speech analysis in NLTK showing that cisgender men wrote slightly more than cisgender women (**Table 1**), and the identification of a naturally occurring participant of unknown nationality in JGAAP (which will be discussed later on in this article). It is the latter result which keyed us in to using JGAAP for further analysis of the corpus.

Total Number POSs			
Parts of Speech	**Gender**		
	Men	**Women**	**Non-Binary**
Adjectives	294	315	42
Nouns	1119	1569	150
Prepositions	541	743	72
Proper Nouns	77	95	9
Personal Pronouns	236	399	52
Possessive Pronouns	64	89	8
Adverbs	215	285	51
Verbs	745	995	140
Chads	5	14	2
Averages POSs			
Parts of Speech	**Gender**		
	Men	**Women**	**Non-Binary**
Adjectives	10.14	6.64	8.40
Nouns	38.83	33.49	30.00
Prepositions	18.62	15.81	14.40
Proper Nouns	2.69	2.00	1.80
Personal Pronouns	8.14	8.49	10.40
Possessive Pronouns	2.21	1.89	1.60
Adverbs	7.45	6.15	10.20
Verbs	25.48	21.17	28.00
Chads	0.17	0.32	0.40

(Contd.)

Variances POSs			
Parts of Speech	**Gender**		
	Men	**Women**	**Non-Binary**
Adjectives	140.77	20.76	61.80
Nouns	1299.36	434.12	285.50
Prepositions	336.17	126.81	66.30
Proper Nouns	9.44	2.74	4.70
Personal Pronouns	74.05	55.82	24.30
Possessive Pronouns	10.03	4.84	1.30
Adverbs	81.40	15.83	43.70
Verbs	609.76	214.41	268.00
Chads	0.15	0.57	0.30

Table 1: Showing the Parts of Speech counts for cisgender men, cisgender women, and non-binary people. These counts are from version one of MapLemon and are not up to date.

2. Experimental methods

In order to examine the relationship between specialized terminology and demographic characteristics, we collected a data corpus focused upon geographic features (Experiment I) and food terms (Experiment II). Participants were recruited in two rounds. In the first round, participants were primarily collected through word of mouth and Reddit posts, then later from the survey site Prolific, and in the second round, participants were recruited exclusively via Prolific. Prolific was used in order to ensure quality of response and control factors such as gender imbalance and regional variation. The survey was conducted via online responses in a Google form. The survey ultimately yielded 85 responses in the first round and 100 in the second round from participants of diverse linguistic, socio-economic, and geographic backgrounds. We chose to restrict the survey respondent location to North America, and linguistic background to speakers of English as primary language in order to control for linguistic variation. The experiment was designed to elicit as much natural writing as possible—we wanted to prevent participants from feeling as though they were taking a survey (to whatever extent possible) and thereby potentially writing in a different style than usual. Additionally, we wanted to divert participants' attention from the demographic characteristics collected. Thus, the demographic survey was placed after the experimental questions.

In Experiment I, the participants were asked to describe a path through an illustrated map, guiding the fictional Chad LemonLover to his destination: a lemonade stand. Participants were asked to be as detailed as possible and to use whatever direction indicators they wished (e.g., landmarks, cardinal directions, street names, etc.). The illustrated map is available as Figure 1 in Appendix I of this paper.

In Experiment II, the same individuals were asked to provide detailed instructions for making lemonade.

The participants then filled out a questionnaire collecting demographic data, including age, gender identity, sex assigned at birth, home city and state, level of education, profession, ethnic background, race, first language, most familiar language, and bilingual status. This detailed demographic data was collected to better understand the effects that educational, cultural, regional, and socioeconomic background, etc. have on the responses gathered. Participants received a small compensation, the equivalent of $5 USD. See Appendix I, Figure 2 for experimental questions.

3. Analysis

All analysis was conducted using the Java Graphical Authorship Attribution Program, with the following criteria for processing:

1. Canonicizers: Punctuation separator, normalize whitespace, unify case
2. Feature set: Stanford Part of Speech Ngrams 2–4 (meaning 2–4 parts of speech)
3. Event culling: None
4. Analysis: K-Nearest Neighbor (K = 1) with Metric Cosine Distance

3.1 JGAAP

The Java Graphical Authorship Attribution Program (JGAAP) has been proven useful in identifying unknown authors via stylometric analysis (Wang, Riddell, and Juola 2021). JGAAP works via user-driven document upload (uploading "unknown" authors, then known authors to compare with), then uses a three-phase model of authorship attribution to analyze documents: 1) canonicizing, or pre-processing, in which distracting or otherwise uninformative details of the document are neutralized by stripping data of unnecessary whitespace, variations in capitalization, variations in punctuation, etc.; 2) feature set (also called event set) generation, where the document is then broken into a series of "events" (most relevantly, one such event set is part-of-speech [POS] tagging); 3) event culling, where certain events are taken out of the data set and not analyzed; 4) analysis, where the events are then analyzed using many different classification methods such as nearest neighbour, where distances are calculated between a pair of documents, and documents of unknown authorship are assigned to the author of the closest document with a known author (Juola 2009).

4. Results

All the following results should be read starting with the top column, then moving down. When using nearest neighbour analysis, smaller numbers are more significant. That is

to say, the smaller a number becomes, the closer the unknown document is stylistically to the known document. Documents are generated by separating demographics into separate PDF files for JGAAP to analyze, for example, separating out all responses from Canadian respondents to make one large Canadian "author" text base. In our case, the tables presented in this section will always have the "unknown" document at the top, and the comparison authors/known documents are listed on the left side of the table. Documents generally range from 0–2 points of similarity, 0 being the exact same document, and 2 being extremely dissimilar. The tables in this section express results up to five significant figures.

Results from conducting stylometric analysis using K-Nearest Neighbor and part-of-speech tagging in the Java Graphical Authorship Attribution Program indicated that a naturally occurring unknown nationality in our responses was Canadian (when compared to American and Canadian authors from the same corpus). The respondent later confirmed they are Canadian, showing that MapLemon can be used to disambiguate region (**Table 2**). The amount of words in the American corpus was 5,550, and the Canadian corpus had 5,549.

Comparison in JGAAP	Mystery Nationality
Canadian	1.25
American	1.5

Table 2: Showing the values JGAAP produced when the unknown author was compared to American and Canadian respondents.

Using the same analysis methods, MapLemon seems to indicate that transgender respondents (noted in Table 3: Transgender Men as "Female to Male [FTM]" and Transgender Women as "Male to Female [MTF]," indicating the "direction" of their transition) write like their gender identity rather than their sex assigned at birth (**Table 3**).

Comparison in JGAAP	Transgender Men (FTM)	Transgender Women (MTF)
Transgender Men (Female to Male [FTM])	--	1.125
Transgender Women (Male to Female [MTF])	1.125	--
Cisgender Assigned Female at Birth	1.5	1.25
Cisgender Assigned Male at Birth	1.25	1.5

Table 3: Showing the results of gender comparisons, including transgender participants in JGAAP.

The word counts and total participants for the following results are 6,826 for Cisgender Assigned Female at Birth (45 participants), 5,322 for Cisgender Assigned

Male at Birth (32 participants), 2,444 for Transgender Men (Female to Male—15 participants), and 1,125 for Transgender Women (Male to Female—10 participants). It should be noted that word/participant counts for transgender people will naturally be lower due to minority status, particularly in the case of Transgender Women (MTF), as well that our corpus so far includes results from more non-binary people than binary transgender people.

Table 3 shows that, when compared to each other, Transgender Men (FTM) and Transgender Women (MTF) are the same distance apart in writing similarity, which seems to be indicative of a "transgender accent" that will be elaborated upon later in this paper.

Furthermore, Transgender Men (FTM), when compared to Cisgender people Assigned Female at Birth (AFAB; cisgender women), are a distance of 1.5 apart, and are 1.25 apart from Cisgender people Assigned Male at Birth (AMAB; cisgender men). That is to say, Transgender Men (FTM) write most similarly first to Transgender Women (MTF), then to cisgender men, then finally to cisgender women. So, we can see from these results that Transgender Men (FTM) write, aside from other transgender people, most closely to their gender rather than to their sex assigned at birth.

For Transgender Women (MTF), the same is true—they are a distance of 1.25 apart from cisgender women, and a distance of 1.5 from cisgender men.

Currently, research is being done on our non-binary respondents, results for which tentatively show that, unlike their binary transgender counterparts, text data on non-binary people works much better when non-binary people are lumped together as a dataset rather than separated into their sex assigned at birth (as in, "non-binary AFAB" or "non-binary AMAB"). Results for analysis against other participant sets is shown in **Table 4**. These results seem to indicate that non-binary respondents are, naturally and firstly, most similar to transgender people in their writing before they're similar to cisgender people. Why they are more similar to Transgender Women (MTF) than Transgender Men (FTM) is uncertain at present; however, the current hypothesis is that it's due to sample size. The word counts and participants for **Table 4** remain the same as mentioned above, aside from non-binary people, which contains 9,760 words across 66 participants.

Comparison in JGAAP	Non-Binary
Transgender Men (FTM)	1.125
Transgender Women (MTF)	1.0625
Cisgender Women	1.25
Cisgender Men	1.5

Table 4: Showing the results of the comparison of non-binary people to the other four binary genders present in the corpus.

5. Hypotheses

It is our belief that the collective results of analysis of transgender and non-binary responses indicate that transgender people may have an "accent" of their own. Some work on this has been done previously that corroborates this idea (see Zimman 2020).

Interestingly, the fact that cisgender men, generally speaking, write most differently from other genders seems to corroborate standing Queer Theory, which states that masculinity is the most "exclusive" gender presentation, whereas femininity is more inclusive; that is to say, masculinity is harder to conform to and has more rigid standards, but femininity has fewer of those same types of standards. This idea is expounded upon in Jean Bobby Noble's 2004 book *Masculinities Without Men?* (Noble 2004), and the classic 1990 piece by Judith Butler, *Gender Trouble* (Butler 1990).

6. Future plans

For the future, pending funding, we plan to firstly gather more responses for MapLemon, as the corpus is still in its infancy and is therefore still fairly small. We plan to analyze more data of our non-binary respondents, perhaps using different analysis methods within JGAAP, to get a better understanding of why they work best in a group rather than separated by sex assigned at birth (this, of course, is socially a good thing, but from a linguistic perspective still must be understood); furthermore, we intend to do so in a way that prevents binarism. We also plan on analyzing more state-related demographics, as well as beginning analysis on ethnicity and race, with the potential to gather another round of participants of specific ethnic or racial backgrounds, depending on the results of our forthcoming analysis, to understand better how these backgrounds may be influencing the gender differences observed in writing. As well, we plan on gathering data from Transgender Women (MTF) to increase their presence in our corpus. Finally, we plan on making this corpus publicly available once it's in a ready enough state.

7. Conclusion

MapLemon shows significant promise in being used as a tool for stylometric demographic identification. MapLemon is also uniquely positioned to have data of those from minority backgrounds be gathered and analyzed by people from those same backgrounds, so that the data is properly handled and understood—a concept that is currently rising in popularity due to outcry from, for example, Queer communities and Native American communities. MapLemon could also theoretically assist in the proving of prevalent Queer Theory ideas, as well as furthering the field of Trans Linguistics as a whole.

Appendix I

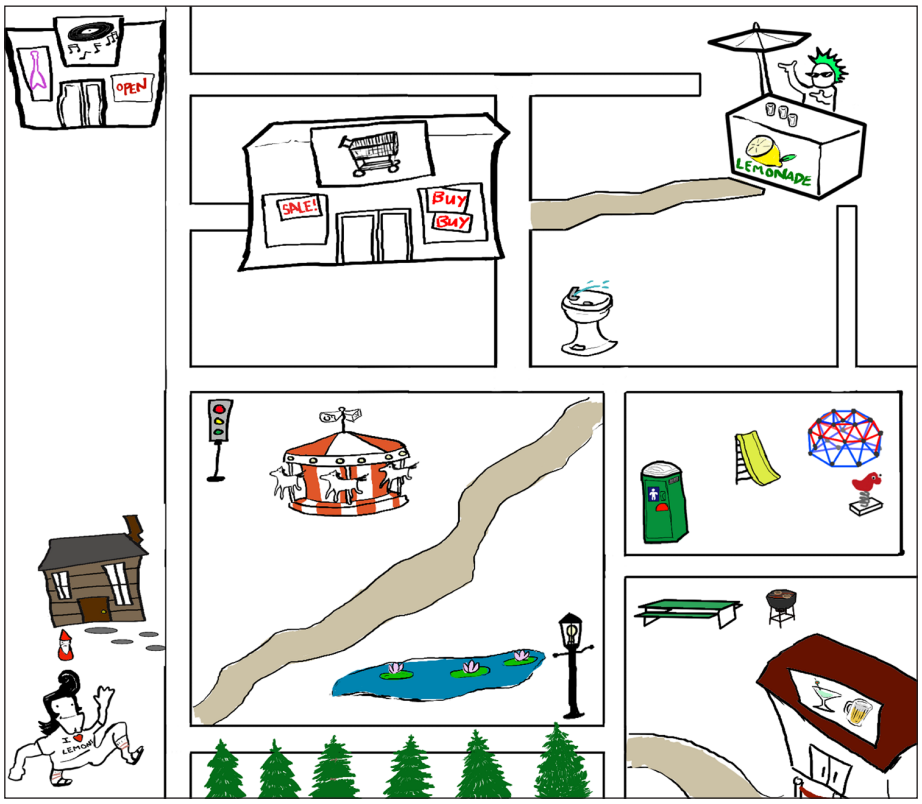

Figure 1: Experiment Map.

This is Chad LemonLover, who is desperate for a lemonade. He wants to get to the lemonade stand, but he's lost and doesn't know the way. Please help our Chad get to the lemonade stand by giving him detailed directions. Chad isn't very smart, and he's going to need you to be as detailed and as descriptive as you possibly can. Please use landmarks, cardinal directions, relative directions, and anything else you feel like, to help Chad find his lemonade. Explain like you really would to a friend or a passer-by. *

Long answer text

Chad is sick and tired of getting lost when getting lemonade. He decided to try and make lemonade at home, all by himself. Please help Chad by giving him your personal recipe of how to make lemonade. Remember, Chad is not very smart, so please be as detailed as possible. Start with some lemons... *

Long answer text

Figure 2: Experiment Questions.

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial

Authorship is alphabetical after the drafting author and principal technical lead. Author contributions, described using the CASRAI CredIT typology, are as follows:

Author name and initials:

Theodore Manning (TM)
Eugenia Lukin (EL)
Ross Klein (RK)
Patrick Juola (PJ)

Authors are listed in descending order by significance of contribution. The corresponding author is TM.

Conceptualization: TM, PJ, EL, RK
Methodology: PJ, EL, TM, RK
Formal Analysis: TM
Investigation: TM
Data Curation: TM
Writing – Original Draft Preparation: TM
Writing – Review & Editing: TM

Editorial

Special Issue Editors

Roopika Risam, Dartmouth College, United States
Barbara Bordalejo, University of Lethbridge, Canada
Emmanuel Château-Dutier, University of Montreal, Canada

Copy Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

References

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Juola, Patrick. 2009. "JGAAP: A System for Comparative Evaluation of Authorship Attribution." *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(1): 1–5. DOI: https://doi.org/10.6082/M1N29V4Z.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2018. "Surveying Stylometry Techniques and Applications." *ACM Computing Surveys*, 50(6): 1–36. DOI: https://doi.org/10.1145/3132039.

Noble, Jean Bobby. 2004. *Masculinities without Men? Female Masculinity in Twentieth-Century Fictions*. Sexuality Studies Series. Vancouver: UBC Press.

Tomas, Frédéric, Olivier Dodier, and Samuel Demarchi. 2022. "Computational Measures of Deceptive Language: Prospects and Issues." *Frontiers in Communication 7*. https://www.frontiersin.org/articles/10.3389/fcomm.2022.792378.

University of Edinburgh. 1993. "HCRC Map Task Corpus." Linguistic Data Consortium. DOI: https://doi.org/10.35111/9GE9-6C05.

Wang, Haining, Allen Riddell, and Patrick Juola. 2021. "Mode Effects' Challenge to Authorship Attribution." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, edited by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, 1146–1155. Online: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2021.eacl-main.97.

Zimman, Lal. 2020. "Transgender Language, Transgender Moment: Toward a Trans Linguistics." In *The Oxford Handbook of Language and Sexuality*, edited by Kira Hall and Rusty Barrett, 7–10. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780190212926.013.45.