HOW MUCH OVERLAP MEANS PLAGIARISM? A CONTROLLED TEST CORPUS

Patrick Juola[1]

[1]*Duquesne University, United States of America*

Abstract

The easiest way to find plagiarism is to see if two people used the same words to describe the same thing. But there are only so many ways to talk about something. How much word overlap must we see before we assume we found plagiarism?

In this paper, we analyze a newly developed corpus, the MapLemon corpus (Manning, et al., 2022). This corpus contains 91 pairs of English language essays written by experimental online participants in late 2021. Participants were asked to write and submit essays on very specific topics. In the first topic, participants were presented with an illustrated map and asked to give directions from one specific point to another. In the second, participants were asked to provide instructions for making lemonade. All writers were asked to be as explicit as possible to allow for collecting a larger number of tokens. On average, each participant wrote 63.40 words on the map subtask and 86.84 words on the lemonade subtask.

Within each subcorpus, we preprocessed all responses by converting data to lower case, stripping out all punctuation, and tokenizing by breaking at whitespace. We then analyzed all essay on the same topic in pairs, calculating the Jaccard similarity coefficient (the number of word types appearing in both essays divided by the number of word types that appear in either essay) for each pair. With 91 participants, this created 4095 essay pairs under each condition. The Jaccard coefficient varies between a maximum of 1.0 (when every word that appears in the first essay also appears in the second, and vice versa) and a minimum of 0.0 (when there are no words in common between either essay) and thus can be read as a proportion of word types that overlap. Our expectation is that this provides a reasonable estimate of the degree of lexical overlap that will be created when two people write brief passages [in English] on the same topic under the same conditions.

On the map corpus, the Jaccard coefficient ranged from 0.52 to 0.03, with a mean of 0.2100 +/- 0.0694. On the lemon subcorpus, the Jaccard coefficient ranged from 0.57 to 0.00 [exactly], with a similar mean but slightly greater variance (0.1906 +/ 0.0702). The 0.00 indicates that a small set of lemonade recipe essay pairs had literally no words in common, a surprising finding easily explained by observing that a typical essay in such a pair was extremely short and atypical in content. For example, one "recipe" simply said "Go to this supermarket" (presumably to buy prepared commercial lemonade), contained only four words, and notably did not mention any of the typical ingredients, processes, or even common function words like "the," "a," "and," and so forth. The median similarities are very close to the mean similarities (map: 0.2105; lemon: 0.1944) suggesting that these outliers did not have a significant effect on the overall averages. Finally, the correlation between the Jaccard coefficients of the map and lemon pairs by the same writers was 0.2892, indicating that there appears to be a strong effect of individual writing styles, and that people who use similar

vocabularies in giving map directions also use similar vocabularies when writing recipes (likewise for dissimilar vocabularies).

This paper thus provides an empirical and quantitative confirmation of the heuristic that too much lexical overlap indicates non-independent writing. Some overlap is expected due to topic similarity, and some will arise from the structure of English itself, but a student whose recipe or instructions overlapped with 60% of another person's lemonade recipe would be noteworthy and probably involve some sort of academic integrity violation.

We hope to extend this analysis both to investigate longer phrases and to investigate the expected degree of overlap between cross-topic essays to determine comparative effect sizes. We are also interested in replicating this study in other languages or other varieties of English.

References

Manning, T.D., et al. (2022, May 17-19). Construction and Analysis of a Stylometric Map-Based Corpus for Tracking Individual Token Use and Demographic Characteristic Identification. [Poster presentation.] In Risam, R. et al., EDS. *DH Unbound* 2022. Virtual. https://dhunbound2022.ach.org/