PAEB_114

Stylometric Comparison of Professionally Ghost-Written and Student-Written Assignments

Robin Crockett, University of Northampton, United Kingdom;
Kirstie Best, University of Northampton, United Kingdom

Keywords: *Commissioning; Contract-cheating; Essay-Mill; Ghost-Writer; Stylometry*

Introduction

In September 2018 the University of Northampton was advised by the Police that a student (hereinafter 'Student') had made a succession of payments to a named, well known and long established UK-based essay-mill during calendar year 2017, the period they were investigating for unrelated reasons. That period encompassed the latter part of the final year of Student's Bachelors (undergraduate) degree and first part of their Masters (postgraduate) degree. Also in September 2018, subject tutors were independently referring Student's Masters dissertation for suspected contract-cheating (Clark & Lancaster, 2006) due to its anomalously high quality in terms of both subject content and written English. Those circumstances led to a detailed investigation of Student's submitted assignments over the final year of their Bachelors degree and entire Masters degree. The outcome of that investigation and the associated formal hearings was that Student had committed extensive commissioning/contract-cheating over the two-academic-year period.

Following the conclusion of that investigation and all consequent University processes, the authors decided to investigate the portfolio of assignments in more detail as that portfolio offered the opportunity to investigate differences between known professionally ghostwritten work and (possible) student-written work, across the same range of subjects, in detail. The formal investigation had not required this depth of investigation owing to the sufficiency of 'headline' evidence and Student's inability to counter that evidence: the headline document properties are summarised in Table 1.

In Table 1 and associated text, Batchelors and Masters assignments are labelled 'L6' and 'L7' respectively, the two letters indicate the core subject matter and the number refers to first, second or third assignment as appropriate. Where known or strongly suspected, a ghost-writer flag is appended to the assignment identifier. Three assignments with essaymill identifiers are 'A1', 'A2', 'A3'; four assignments with other ghost-writer identifiers are 'B' (two assignments), 'C' and 'D'. The Masters dissertation, identified by subject tutors as ghost-written, is 'X'; and two other assignments showing strong evidence are 'Y' and 'Z'. It should be noted that Student was unable to offer any information regarding any of the names revealed in the document properties metadata and so it has been concluded that none of the names indicate a borrowed computer and all are associated with ghost-writers. It should also be noted that until the Masters dissertation, Student had been careful to commission assignments at second-class and equivalent grades, consistent with their higher

grades in previous years and thus not attracting tutors' attention at the time.

The aim of this investigation is to determine whether there are consistent stylometric differences between the use of English in professionally ghost-written and student-written assignments that could assist in evidence gathering in future contract-cheating investigations (Klaussner et al., 2015). The main objectives are to determine:

- whether any assignments of unknown provenance are grouped with known ghostwritten assignments, raising the possibility those assignments were also ghost-written;
- whether stylometric analysis identifies consistent differences between ghost-written and student-written assignments.

Method

The stylometric analysis (Eder, 2012) was performed using the R open-source statistical computing software (https://www.r-project.org/) using the Stylo (Eder et al., 2016), Sylcount and Cluster library packages. The corpora were prepared by removing footnotes and reference lists from the submitted assignment files and exporting redacted (e.g. student name and ID) plain-text files. All investigated assignments were between ca. 1,000 – 5,000 words in length except for the Masters dissertation (L7_DI3), at ca. 15,000 words.

Results and discussion

Paraphrasing the tutors' evidence (redacted) regarding the Masters dissertation:

- in terms of % similarity [Turnitin], there is very little in the text of the work other than common terms;
- the language/expression used is not that which I associate with the student;
- it has a fluency and maturity which is above that in [Student's] other work;
- the approach of not providing substantive introductions and conclusions to thechapters [...] runs counter to the approach that they would have been advised to take throughout their studies (UG and PG);
- the final section [...] is quite unlike the approach that they have taken with other work
- of theirs with which I am familiar, and not an approach that would be suggested.

The stylometric similarities among the assignments are summarised in Figure 1. Figure 1a is a consensus tree summarising clustering (cf. correlations) between frequencies of wordpairs (2-grams, 2-word phrases): in essence, the further along the arms the group-members diverge, the more similar they are. Analysis of frequencies of words and n-grams (cf. multiword phrases) was the primary analysis used, and is a widely used stylometric technique. The groupings are colour-coded and the three groupings in red, light green and light blue show the core consistent groupings revealed by this analysis. First (red), the two Bachelors dissertation assignments and the two main Masters dissertation assignments (i.e. not the proposal), implying authorship-in-common (ghost-writer 'X') for these assignments. Second (light green), two known ghost-written assignments (suffixed 'B', 'C') are grouped

with the two strongly-suspected assignments (suffixed ‘Y’, Z’), and a further assignment (L7_DI1, Masters dissertation proposal), implying a possible association between the two ghost-writers and an association between either (more probably ‘B’, see Fig. 1b) or both ghost-writers and the other three assignments. Third (light blue), the three known essaymill assignments (suffixed ‘A’) are grouped together with a fourth assignment (L6_SP1), which implies an essay-mill ‘house-style’ and also that the fourth assignment was very probably commissioned from the same essay-mill. The other two groupings (dark green, dark blue) are less consistent but it is probable that if any assignments are Student’s own work then they fall into these groupings and, within that, more probably L7_CS1, L7_DM1 and L6_HR1.

Figure 1b shows the cluster dendrogram of assignments according to word-length (number of syllables) distributions, and is a more straightforward analysis than word (or n-gram) frequencies. In essence (a) the longer the vertical distance from where a cluster diverges from other clusters, the more distinct and (b) the shorter the vertical distances within a cluster from where individual members diverge, the more similar the cluster-members. Figure 1b is colour-coded according to Figure 1a, revealing that this straightforward analysis also shows many of the groupings revealed by the primary analysis, particularly the core elements of the two strongest clusters, i.e. the three identified essay-mill assignments (suffixed ‘A’) and the two strongly-suspected assignments (suffixed ‘Y’, ‘Z’). Analysis of this type is more straightforward than readability-type analysis and can be less sensitive to ‘pseudo sentences’ such as headings and bullet-points associated with different assignment types.

The reason for the lack of association between the two assignments with author-name ‘B’ in either analysis is unknown but is tentatively attributed to two ghost-writers using the same computer or same ID.

Conclusion

The immediate conclusion of this research is that in addition to the eight assignments known to have been contract-cheated, the stylometric analysis confirms the strong evidence initially identified in two assignments and also strongly implicates further assignments. This is statistical evidence, not proof, but does serve to give a fuller picture of the likely scale of Student’s commissioning/contract-cheating activity.

More generally, this (ongoing) research has demonstrated that, in some circumstances at least, ghost-writers can be linked according to essay-mill house-styles. Also that ghostwriter writing styles can be distinctly different to a student’s writing style. In summary, this research has revealed that professional writers use English more correctly, consistently, concisely and precisely than the students who commission them, as might be reasonably expected, even if ‘dumbing-down’ to imitate relatively weak student presentational style.

References

Eder, M. (2012). Computational stylistics and biblical translation: how reliable can a dendrogram be? In Piotrowski, T. Grabowski, L. editors, *The Translator and the Computer*, pages 155–170. WSF Press, Wroclaw.

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107-121. https://doi.org/10.32614/RJ-2016-007.

Clark, R. & Lancaster, T., (2006). Eliminating the successor to plagiarism? Identifying the use of contract cheating sites. *Proc. 2nd International Plagiarism Conference*, Gateshead, UK, Northumbria. Learning Press.

Brocardo, M. L., Traore, I., Saad, S. & Woungang, I. (2013). Authorship Verification for Short Messages Using Stylometry, *Proc. IEEE Intl. Conference on Computer, Information and Telecommunication Systems* (CITS 2013), Athens, Greece, May 7-8, 2013.

Ethical Statement

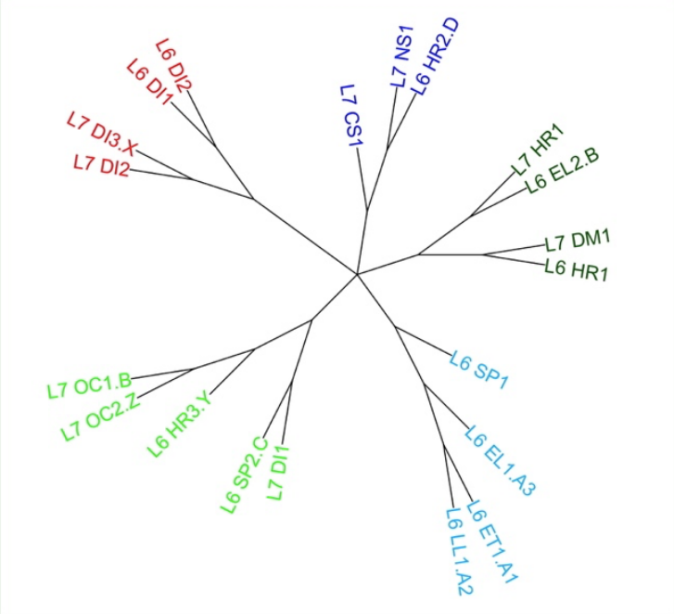
The module and assignment names and codes, and the names of Student, the essay-mill and ghost-writers have been redacted in accordance with the ethical approval given by the University of Northampton Research Ethics Committee in September 2018.

Table 1. Key Document Properties from Formal Investigation.

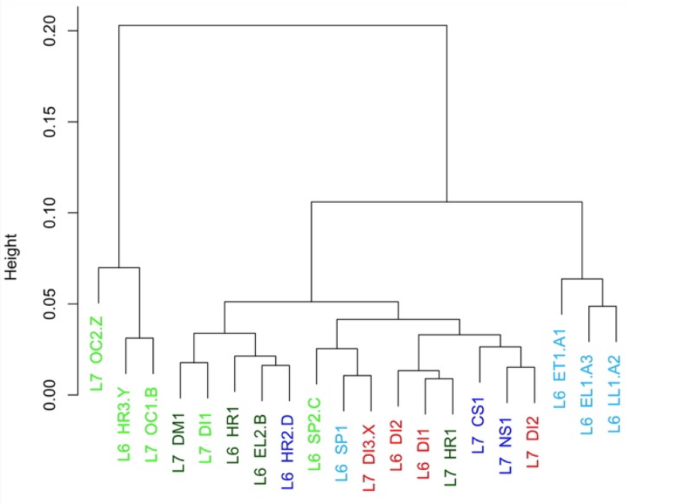
Assignment Identifier	*Ghost-Written*	*Ghost-Writer Flag*	*Doc. Type*	*English Variant*	*Page Size*
L6_EL1	*Yes: essay-mill ID*	*A3*	*docx*	*Australian*	*A4*
L6_EL2	*Yes: author name*	*B*	*docx*	*UK*	*A4*
L6_ET1	*Yes: essay-mill name*	*A1*	*docx*	*UK*	*A4*
L6_HR1	*Unknown*		*docx*	*UK*	*A4*
L6_HR2	*Yes: author name*	*D*	*docx*	*US*	*A4*
L6_HR3	*Strong evidence: hacked xml doc. props.*	*Y*	*docx*	*UK*	*US Letter*
L6_LL1	*Yes: essay-mill ID*	*A2*	*docx*	*US*	*A4*
L6_SP1	*Unknown*		*docx*	*US*	*A4*
L6_SP2	*Yes: author name*	*C*	*docx*	*UK*	*A4*
L6_DI1, L6_DI2	*Unknown*		*docx*	*UK*	*A4*
L7_CS1	*Unknown*		*docx*	*UK*	*A4*
L7_DM1	*Unknown*		*docx*	*UK*	*A4*
L7_HR1	*Unknown*		*docx*	*UK*	*A4*
L7_NS1	*Unknown*		*docx*	*UK*	*A4*
L7_OC1	*Yes: author name*	*B*	*odt*	*UK*	*A4*

L7_OC2	*Strong evidence: basic presentation*	*Z*	*docx*	*Canadian*	*US Letter*
L7_DI1, L7_DI2	*Unknown*		*docx*	*UK*	*A4*
L7_DI3	*Yes: tutor identified*	*X*	*docx*	*UK*	*A4*

Figure 1: Cluster-Analysis – Stylometric Groupings of Assignments.



(a) Grouping according to 2-gram frequencies.



(b) Grouping according to word-length distributions.