STYLOMETRIC COMPARISON OF PROFESSIONALLY GHOST-WRITTEN AND STUDENT-WRITTEN ASSIGNMENTS

Robin Crockett, Kirstie Best

Abstract: We report a stylometric investigation of a portfolio of 20 assignments submitted by an individual student over two consecutive academic years. This investigation followed a formal disciplinary investigation which had identified that eight of the assignments had been ghost-written, with seven of those showing explicit ghost-writer ID information and three of those showing ID information from the same commercial provider. The stylometric investigation involved a conventional word and bigram frequency analysis and a prototype word complexity analysis. The word and bigram analysis identified four consistent groups of assignments, which associate other assignments with the eight known to have been ghost-written, indicating that those were probably also ghost-written. One of those groups comprises the three assignments from the same provider, plus another assignment, implying that the provider has a 'house style' and that the other assignment also came from that provider. The prototype analysis clearly categorised the core members of two of those same groups, including the group from the identified provider, adding further weight those associations. More generally, this investigation shows that it is possible to categorise assignments according to aspects of writing style: we would have obtained the same groups even if we had not possessed the ghost-writer ID information. Where such consistent groups are identified it implies, on balance of probabilities, multiple authorship of assignments and that the student concerned cannot have written all the submitted assignments and that some were ghost-written.

Key words: 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 an identified, well known and long established UK-based commercial provider of academic assignments (sometimes referred to colloquially as an 'essay-mill', see *e.g.* Cambridge English Dictionary online) 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 due to its anomalously high quality in terms of both the subject content and the written English. Those circumstances led to a detailed investigation of Student's submitted assignments over the final year of their Bachelor's degree and entire Master's 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 formal investigation and all consequent University processes, the authors decided to investigate the portfolio of assignments in more

DOI: https://doi.org/10.11118/978-80-7509-772-9-0035

detail as that portfolio offered the opportunity to investigate differences between known 'professionally' ghost-written[1] work and (possible) student-written work, across a consistent 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 (a) provide a coherent statement of honest endeavour that explained the evidence or (b) offer any information regarding any of the author names revealed in the document properties metadata. The headline document properties are summarised in Table 1.

In Table 1 and associated text, Bachelors and Masters assignments are labelled 'L6' and 'L7' respectively, the two letters following the underscore 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 suffixed to the assignment identifier. Three assignments showing different creating-authors associated with the identified provider are suffixed 'A1', 'A2', 'A3'; four assignments showing other ghost-writer identifiers are suffixed 'B' (two assignments), 'C' and 'D'. The Masters dissertation, identified by subject tutors as ghost-written, is suffixed 'X'. Two other assignments showing strong evidence are suffixed 'Y' (different page size and hacked core.xml document properties file) and 'Z' (different page size and basic presentation including the unusual, at least for assignments submitted at the University of Northampton, Canadian English language setting). In light of Student's inability to identify any of the author names in the files he submitted, it has been concluded that none of the names indicate a borrowed computer and all are associated with ghost-writers. It should be noted that until the Masters dissertation, Student had been careful to commission assignments at second-class honours and equivalent grades, consistent with their higher grades in previous years and thus not attracting tutors' suspicions at the time.

The aim of this investigation was 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 authorship and provenance are grouped with known ghost-written assignments, raising the possibility that those assignments were also ghost-written;
- whether stylometric analysis identifies consistent differences between ghost-written and student-written assignments, possibly identifying similarities among ghost-writer styles.

[1]By 'professional', we mean mature writers with significant knowledge and expertise within the subject-areas in which they write, *i.e.* writers as might be retained by upmarket providers (essay-mills) that advertise, for example, that they offer support and mentoring to their writers.

Table 1
Key document properties from the formal investigation

Assignment Identifier	Ghost-Written	Ghost-Writer Suffix	Doc. Type	English Variant	Page Size
L6_EL1	Yes: provider ID	A3	docx	Australian	A4
L6_EL2	Yes: author name	B	docx	UK	A4
L6_ET1	Yes: provider 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 core.xml file	Y	docx	UK	US Letter
L6_LL1	Yes: provider 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

Context and Literature Review

Contract-cheating, commissioning and ghost-writing

Contract-cheating (Clarke & Lancaster, 2006), also known as commissioning, is the variant of plagiarism where a student commissions a third-party, *i.e.* a 'ghost-writer', to write all or part of an assignment for them and then submits that assignment as their own work for assessment. The ghost-writer might or might not receive a reward, financial or otherwise, for their endeavour: that is immaterial in the academic-integrity context and it is the student's submission of the commissioned assignment as their own work that is the act of academic misconduct, *i.e.* plagiarism.

An 'essay-mill' (sometimes 'paper-mill', 'assignment-mill' etc.) is a business that acts as an intermediary between students and ghost-writers who might be its employees but, more generally (as based on perusal of numerous websites), are freelance writers who register with it (Medway et al., 2018; Rogerson, 2014; Ellis et al., 2018). In the context of academic misconduct and plagiarism, ghost-writers can work as independent freelance writers, possibly advertising via social media, but often register with one or more essay-mills (Sivasubramaniam et al., 2016). According to their online advertisements, some (upmarket) essay-mills require writers to demonstrate their qualifications, expertise and experience, offer forms of support and mentoring to their writers, and provide quality control on the purchased assignments. However, other

essay-mills make no such claims and act solely as intermediaries providing no other services (Sutherland-Smith & Dullaghan, 2019; Newton, 2018).

Some essay-mills operate multiple 'shop-fronts', *i.e.* websites offering different combinations of subjects or assignment-types, to more clearly advertise the full range of the services they offer to (prospective) customers. Some essay-mills assert that they only provide exemplars for tutorial purposes, with disclaimers that students who purchase such assignments should not submit them, others do not. However, it appears that such disclaimers are of little, if any, use in deterring students who are intent on dishonestly submitting purchased assignments. Some essay-mills and ghost-writers offer 'plagiarism free' guarantees, others do not – and it must be observed that however successful such guarantees might be in attracting custom, all such guarantees become meaningless in the event that students submit the commissioned assignments as their own work.

With regard to the detection of contract-cheated (commissioned) ghost-written assignments, on occasion a student is careless and leaves a tell-tale in the assignment which is readily identifiable by the assessing tutor(s), *e.g.* an identifier of some sort in the filename or an 'insert official assignment code here' place-holder in the text. However, this is often not the case and, where it is not the case, considerable effort on the part of tutors and investigators can be required to assemble a consistent body of evidence (Clarke and Lancaster, 2007). Also, it can be necessary to investigate and compare a student's entire portfolio of submitted assignments in order to establish a body of evidence that indicates that, according to a probability threshold, *e.g.* 'balance of probabilities' (QAA, 2017), one or more assignments were not written by the student in question. In general, whatever the nature and extent of the investigation, initial identification is often dependent on the alertness and awareness of individual assessing tutors (Bretag & Mahmud, 2009; Dawson & Sutherland-Smith, 2017; Lancaster & Clarke, 2007; Rogerson, 2017).

Stylometric analysis

Stylometry can be defined as the statistical, or quantitative, analysis of writing style. A common use of stylometry is for attribution of authorship within a collection (corpus) of texts, as is the case here (Klaussner et al., 2015; Stamatos, 2008). Given a collection of texts comprising samples of known and unknown and/or disputed authorship, it can be possible to group texts according to, for example, most frequent words or phrases (*n*-grams), numbers of words per clause or sentence as defined by punctuation, use of spelling conventions (*e.g.* in English, UK/British and US/American spellings) and a variety of relationships between stop-words and content-words (Kulig et al., 2017).

To a great extent, stylometry has evolved for the analysis of relatively long texts, such as plays and novels, and where the writers are experienced with developed individual writing styles (see, *e.g.* Juola, 2013). This is not generally the case with student assignments: assignments are often short, *e.g.* up to a few thousand words in length, and written to assignment briefs which can dictate basic writing styles that should be used by the students (or, indeed, ghost-writers) such as (expository) essays or (technical) reports. Thus, the relative shortness of student assignments can limit the

manifestation of subtler aspects of individual writing style, which might anyway be inhibited by the style constraints of some assignment briefs (Brocardo et al., 2013). For an overview of stylometric approaches in the identification of contract-cheated work see, *e.g.* Juola (2017).

Under some such circumstances, *e.g.* experienced writers with their own preferred vocabularies, it can be useful to remove stop-words and focus on content-words: stop-words effectively being regarded as 'noise' and thus carrying no information. However, this is generally not the case with student assignments and observation of written work at a variety of levels indicates that students, who are learners, often write less concisely and efficiently than mature experienced writers with subject expertise. Thus, stop-word usage can be important in differentiating student-written work from that of professional ghost-writers. Similarly, although much word-frequency analysis is performed on corpora in which all upper-case letters have been converted to lower-case, observation of written work at a variety of levels indicates that experienced writers tend to use capitalised words and proper nouns more consistently and correctly than students and learners.

Lastly, and at risk of oversimplification, in the context of student assignments essays often favour more discursive writing styles with longer, more complex sentences whereas reports (and taught-course dissertations) often favour less discursive writing styles, with shorter less complex sentences and features such as (sub-) headings and bullet-points (as documented in many 'how-to' guides see, for example, McMillan & Weyers, 2011; Greetham, 2014). Over the duration of their studies, students can be required to write assignments encompassing a wide-range of formats, each of which can influence the immediate writing style, which reinforces the focus of a stylometric analysis on shorter features, *e.g.* words and phrases rather than complex clauses and sentences.

Methodology

Preparation of the corpora of assignments

The assignments were prepared by removing footnotes and reference lists from the submitted assignment files and redacting all personal information such as student name and ID and modules codes, and saving as Unix-format Unicode (UTF-8) 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.

Stylometric analysis

The stylometric analysis was performed using the R open-source statistical computing software using the Stylo, Sylcount and Cluster library packages. The prepared text files were imported and processed into a single-word case-preserved corpus from which bigram (*i.e.* 2-gram, word-pair, two-word phrase), trigram (*i.e.* 3-gram, word-triple, three-word phrase), and higher up to 10-gram corpora were derived, in both case-preserved and lower-case versions. Stop-words were not removed on the basis of previous observations which indicated that experienced writers tend to write more

concisely and accurately than student-writers. Similarly, both lower-case and case-preserved corpora were analysed on the basis of previous observations which indicated that experienced writers tend to capitalise words more consistently and correctly than students and learners.

Results and Discussion

Tutors' evidence

Paraphrasing the tutors' evidence [redacted], as collated for the referral of the Masters dissertation for suspected contract-cheating:

- '*... 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 the chapters* [...] *runs counter to the approach that* [Student] *would have been advised to take throughout* [Student's] *studies (UG and PG).*';
- '*The final section* [...] *is quite unlike the approach that* [Student] *has taken with other work of* [Student's] *with which I am familiar, and not an approach that would be suggested.*';
- '*I confirm your suspicions/concerns about the authorship of this piece of work. The use of language, flow, connection of ideas is very good, as is the interwoven use of sources. This is not like previous work submitted by this student.*'.

The tutors' suspicions regarding the authorship of the Masters dissertation are clear, in particular the observations regarding the low Turnitin similarity score, as would align with an essay-mill or ghost-writer offering a 'plagiarism free' guarantee, and the different 'non-standard' approach and structure and use of language.

During the course of the formal contract-cheating investigation, subject tutors were contacted for comment with regard to other assignments. In particular, these (redacted) comments were received from a tutor who taught a Masters module that covered a similar subject to the Masters dissertation:

- With regard to L7_DI1 (initial Masters dissertation proposal): '*... some very minor issues as to structure and proof-reading but on the whole a clear, mature analysis. Well researched and showing similar qualities to the* [Masters] *dissertation i.e. neat and appropriate written presentation, appropriate research and referencing. The written 'voice' is similar to that for the* [Masters] *dissertation in terms of fluency and clarity of expression.*';
- With regard to L7_HR1: *'It covers a very different topic to* [Masters dissertation]. *It is of a different standard to* [Masters dissertation]*because the writer has not sufficiently followed the assignment brief and also misses the point as to the* [...] *issues raised by the problem analysed. Interestingly, reading the* [L7_HR1]*and* [L7_DI1]*assignments straight after each other, there are clear similarities in style.*

> [L7_HR1] *is well structured and clearly written and there is a turn-of-phrase which echoes that of the other work, and language which I wouldn't necessarily associate with this student. The skills reflection is also interesting as the writer comments on* [Student's] *involvement in class, which certainly does not correlate with my recollection of* [Student] *being pleasant but very quiet in classes.'.*

These comments are also revealing. In particular, the observation regarding the 'voice' of the Masters dissertation proposal which suggests that assignment might also have been ghost-written (possibly by the same ghost-writer as the Masters dissertation). Also, the observations regarding L7_HR1 in containing a discordant 'skills reflection' and missing the focus of the assignment brief, which could align with a (ghost-) writer adapting a previously written document, possibly in response to a 'contract' in which the commissioning student describes the assignment insufficiently precisely.

Stylometric evidence

The main stylometric analysis consists of multivariate analysis of word and bigram frequencies, and is presented as consensus-tree plots summarising the results of cluster analysis (Eder, 2012). Some analysis of trigram frequencies, and initial analysis of higher-order *n*-gram frequencies, was performed but it was observed that trigrams are the highest-order *n*-grams that revealed useful information and, for trigram and higher-order *n*-gram frequencies, increasingly (with order of *n*-gram) the assignment groupings were identified less clearly and consistently and showed more sensitivity to subject-specific phrasings. This accords to some extent with Lopez-Escobedo *et al*. (2013) who restricted their analysis to words, bigrams and trigrams, with the proviso that they analysed Spanish texts.

In addition, a prototype analysis of word complexity (length), as in the frequency distributions (histograms) of numbers of syllables per word, is reported. This analysis was instigated following preliminary analyses which suggested that professional writers use bigger vocabularies (and more complex words) even if 'dumbing-down' to mimic, say, an inexperienced student writing style. Number of syllables per word, with a range from single-syllable words (many stop-words, assumed low information) to multi-syllable words (generally content-words, assumed high information), was used as an intermediate between readability (Gillam, 2013) and more abstract measures of information content such as self-information and information entropy (Shannon, 1948; Guerrero, 2009).

Word and bigram frequencies

Figures 1, 2 and 3 show consensus-tree representations of the groupings of the assignments according to word and bigram frequencies. The consensus trees summarise the most consistent clustering (*cf*. correlation) among the text files 'averaged' over different samples of most-frequent words/bigrams to show the typical groupings. In essence, the further along the arms the group-members diverge, the more similar they are. Figures 1 and 2 are colour-coded according the groups in Figure 3 (see below).

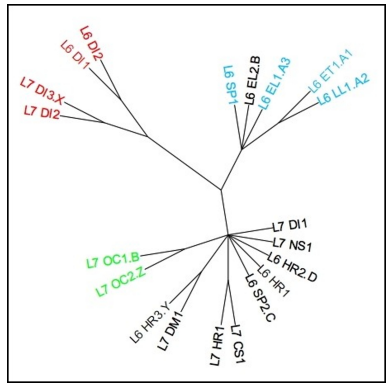

Figure 1. Assignment grouping according to lower-case single-word frequencies

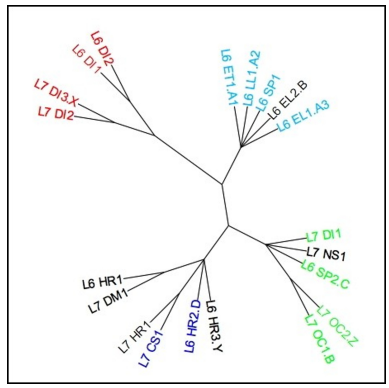

Figure 2. Assignment grouping according to case-preserved single-word frequencies

The importance of analysing both case-preserved and lower-case corpora, to reveal differences in the use of proper nouns etc., is revealed in Figures 1 and 2. Figure 1 shows the most typical grouping according to frequencies of lower-case single words, over 70–300 most-frequent words. Two distinct groups are revealed, coloured red and light blue, plus another proto-group coloured green. Figure 2 shows the most typical grouping according to frequencies of case-preserved single words, over 70–300 most-frequent words. The same two distinct groups as in Figure 1, coloured red and light blue, are revealed. However, the proto-group in Figure 1, coloured green, is more distinctly revealed as a group and a further proto-group, coloured blue, is revealed.

Many variants of these groups (as coloured red, light blue, green and blue in Figures 1 and 2) were observed as different numbers of most-frequent words and bigrams (and trigrams) were analysed but four core groupings were consistently observed, and these are shown in Figure 3.

Figure 3 specifically shows the most typical grouping for lower-case bigrams, over 155–300 most-frequent bigrams. The four typical groups are as follows:

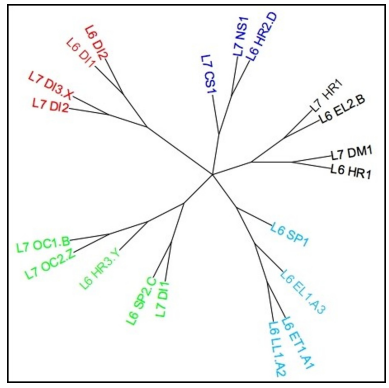

Figure 3. Typical assignment grouping (bigram frequencies)

i. L6_DI1, L6_DI2, L7_DI2 and L7_DI3.X, *i.e.* Bachelors and Masters dissertation draft and final versions (coloured red);
ii. L6_ET1.A1, L6_LL1.A2, L6_EL1.A3 and L6_SP1, *i.e.* the three assignments known to have been commissioned from the identified provider plus a further assignment (coloured light-blue);
iii. L6_HR2.D, L7_CS1 and L7_NS1, *i.e.* two Masters assignments with a ghost-written Bachelors assignment (coloured blue);
iv. L6_SP2.C, L6_HR3.Y, L7_DI1, L7_OC1.B and L7_OC2.Z, *i.e.* two ghost-written assignments, two suspected ghost-written assignments and the Masters dissertation proposal (coloured green);
v. As well as the four typical groups, four assignments, L6_HR1, L7_DM1, L7_HR1 and L6_EL2.B (noting its possible inclusion in group (ii), see Figures 1, 2) show little or no consistent grouping with other assignments, no grouping as consistent as groups (i)–(iv) and as such form a fifth (unassociated) group (uncoloured, black).

The first group is interesting because it is the most consistent group and shows a high degree of consistency of writing style across the Bachelors and Masters dissertations. It would be reasonable to expect that Bachelors draft and final versions would be similar, and also the Masters draft and final versions, but the grouping of the two pairs was unexpected due to the different subjects and qualities (unlike the Masters dissertation, the Bachelors dissertation did not stand out due to anomalously high quality). This implies that the Bachelors dissertation was also ghost-written, by a ghost-writer with a similar writing style to ghost-writer 'X', possibly by ghost-writer 'X' writing to two different standards but with underlying similarity of style.

The second group is interesting for two reasons. First, the similarity among the three assignments known to have been commissioned from the identified provider suggests a 'house-style' for writers who have had access to mentoring and guidance regarding style from that provider. Second, the inclusion of at least one other assignment, L6_SP1, strongly suggests that it, too, was written by a ghost-writer registered with the

identified provider. Also, there is the possibility that a second assignment, L6_EL2.B, grouped with the four other assignments, as shown in Figures 1 and 2, although much less consistently suggests that ghost-writer 'B' might be (or have been) registered with the identified provider (and see below).

The third group suggests that two Masters assignments might have been written by ghost-writer 'D'. However, it is possible that Student used the ghost-written Bachelors assignment, L6_HR2.D, as a starting point for the Masters assignments although only one of the two Masters assignments is in the same subject.

The fourth group is more complicated. First, other than some generic (sub-) headings and bullet-point-type sentences, there are no similarities between the two assignments that show ghost-writer 'B' as author. This suggests either that ghost-writer 'B' is, in fact, two writers using the same ID (whether by deliberate choice or because they used the same computer) or that the Bachelors assignment written by ghost-writer 'B' was used as a starting point for the later Masters assignment despite being on a different subject. Second, there are consistent similarities between assignments L7_OC1.B and L7_OC2.Z which are attributed to authorship. While similarity of subject matter cannot be completely discounted as a reason for similarities between these two assignments, analysis of trigrams and higher-order *n*-grams does not support that explanation. Third, the similarity between the two strongly suspected assignments, L6_HR3.Y and L7_OC2.Z, with at least one ghost-written assignment is further evidence supporting the document-properties evidence that these were commissioned. Overall, the inclusion of the Masters dissertation proposal, L7_DI1, in a group comprised of at least one ghost-written assignment plus two strongly suspected of having been ghost written, suggests that this assignment was also ghost-written.

Word-complexity analysis

The similarities among the word-length distributions were analysed using hierarchical cluster analysis: this is a more straightforward analysis than word (or *n*-gram) frequencies and produces a single clustering (noting that there are other clustering algorithms). Figure 4 shows the cluster dendrogram: in essence (a) the longer the vertical distance from where a cluster diverges from other clusters, the more distinct the cluster and (b) the shorter the vertical distances within a cluster from where individual members diverge, the more similar the cluster-members. Figure 4 is colour-coded according to Figure 3, revealing that this straightforward analysis shows some similar groupings to those revealed by the primary analysis, particularly the core elements of the two strongest clusters, *i.e.* the three identified-provider assignments (group (ii), suffixed 'A') and the two strongly-suspected assignments with a known ghost-written assignment (group (iv), suffixed 'B', 'Y' and 'Z').

Discussion

The word and bigram frequency analysis demonstrates that similarities and differences of writing style according to known authorship can be identified with high degrees of consistency. As well as similarities among individual assignments, three known to have been commissioned from the identified provider show similarities that might indicate

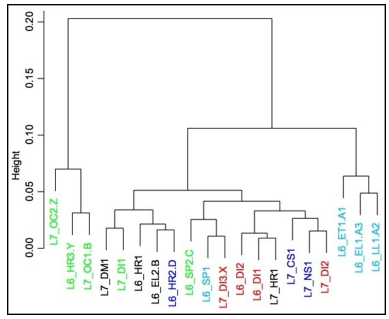

Figure 4. Similarities of word-lengths distribution

a 'house style'. In addition to the eight assignments known/identified as ghost-written as a result of the formal investigation (L6_ET1.A1, L6_LL1.A2, L6_EL1.A3, L6_EL2.B, L6_SP2.C, L6_HR2.D, L7_OC1.B and L7_DI3.X), a further seven have been identified as very probably ghost-written: *i.e.* dissertation assignments L6_DI1, L6_DI2 and L7_DI2 with L7_DI3.X; L6_SP1 as purchased from the identified provider; L6_HR3.Y and L7_OC2.Z with ghost-writer 'B'; and L7_DI1 with ghost-writer 'C'.

Under other circumstances, it might be necessary and appropriate to include trigrams or higher-order *n*-grams, depending on the portfolio of assignments under investigation. However, the reported investigation demonstrates that it is possible to use conventional word and *n*-gram frequency analysis on actual portfolios of student-submitted assignments to categorise assignments according to aspects of writing style.

The more straightforward and significantly less time-consuming analysis of word length (number of syllables per word) shows potential as an analysis for triaging a portfolio of assignments for similarity of writing style. This is a more abstract measure of information content than word or *n*-gram frequencies and is potentially less susceptible to variations in writing style associated with variations of assignment type than readability indices. A document's readability index is a measure of the complexity of its structure, syntax and vocabulary, and readability approaches have been considered for authorship attribution (Gillam, 2013). While such indices include measures of single-syllable and/or multi-syllable words, these also include measures of clauses and sentences, which can vary with assignment-dependent writing style rather than personal writing style. This suggests that such indices, which were not designed for authorship attribution purposes, might be less robust in this context than simpler measures such as word complexity or more abstract entropy-related measures (Shannon, 1948; Guerrero, 2009).

Generalising – the wider context

Under hypothetical circumstances where the University received the same advice from the Police but the seven ghost-written assignments did not contain the author ID information, then the same clusterings and groups would be observed. From there, it would be safe to deduce, on balance of probabilities, that if one group (as in

Subsection 3.2.1) comprised assignments written by Student then the other groups must comprise assignments not written by Student and, therefore, comprise ghost-written assignments. If it is allowed that the Masters dissertation would still be independently identified by tutors as having been ghost-written, then that would mark three other assignments as ghost-written, *i.e.* group (i) as in Subsection 3.2.1. The assignments with different page-sizes, language settings and document formats would raise further suspicions, implicating the assignments in group (iii) as in Subsection 3.2.1. Also, the tutors' observations (as in Section 3.1) would further implicate and associate other assignments although, in practice under circumstances such as these, the investigation would necessitate that additional subject tutors comprehensively re-examine Student's submitted assignments.

Also, it should be noted that the prototype analysis identified two groups of assignments (corresponding to cores of groups (ii) and (iv) as in Subsection 3.2.1), distinct from each other and distinct from the other assignments. That itself is sufficient to indicate that, on balance of probabilities, at least some of the assignments were probably written by someone other than Student.

This investigation demonstrates that by considering a portfolio of assignments submitted by a student, even without specific ghost-writer ID information, it is possible to use stylometric analysis to consistently categorise assignments into distinct groups. From an appropriately robust and consistent categorisation such as this, it is possible to deduce on balance of probabilities that not all of the assignments in the portfolio can have been written by the student in question and, therefore, conclude that some must have been ghost-written. In the reported investigation, correct and consistent capitalisation of proper nouns was a significant reader-observable stylistic feature that contributed to the categorisation. Under other circumstances it could be spelling convention, degree of formality of language or use of contracted/abbreviated forms etc. and, possibly, aspects of presentation and formatting (not considered by stylometric analysis) that provide sets of reader-observable distinguishing features that align with a stylometric analysis and provide a body of evidence.

Further Research and Recommendations

It is intended to pursue and develop this research. The main recommendation at this early stage is that research into stylometric techniques specifically for determining differences between the writing styles of students who, in assignments, are often encountering specific subject material and terminology for the first time, and the styles of more experienced writers for whom the specific subject matter and associated terminologies are familiar. Within this, there is a need to develop stylometric techniques specifically for short documents, such as student assignments. The prototype word-complexity analysis reported herein illustrates that information-theory based techniques show potential and should be investigated.

It is also recommended that, wherever possible and appropriate, basic stylometric analysis of a student's portfolio of submitted assignments is undertaken to assess whether any files of uncertain authorship can, on balance of probabilities, be either

differentiated from assignments known to have been written by the student in question or related to ghost-written assignments submitted by the student in question.

Conclusions

The immediate conclusion of this research is that in addition to the eight assignments known to have been contract-cheated, the stylometric analysis:

i. confirms the strong document-property evidence of ghost-writing initially identified in two assignments and, also,

ii. strongly implicates a further five assignments that did not show specific document-property evidence.

This is statistical evidence, *i.e.* evidence to be considered on 'balance of probabilities' along with other evidence, and is not proof 'beyond reasonable doubt'. However, it does serve to give a fuller picture of the likely scale of Student's commissioning/contract-cheating activity.

More generally, this (ongoing) research demonstrates that ghost-writer writing styles can be distinctly and detectably different to an individual student's writing style. Also, it demonstrates that, in some circumstances at least, ghost-writers can be linked according to provider house-styles. In summary, this research has revealed that professional ghost-writers tend to use English more correctly, consistently, concisely and precisely than the students who commission them, even if 'dumbing-down' to imitate relatively weak student presentational styles.

Ethical Statement

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

Software

R. R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Stylo. Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal, 8(1): 107–121.

Sylcount. Schmidt, D. (2019). Syllable counting and readability measurements.

Cluster. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). Cluster analysis basics and extensions.

References

Bretag, T., & Mahmud, S. (2009). A model for determining student plagiarism: Electronic detection and academic judgement. Paper presented at the *4th Asia Pacific Conference on Education Integrity (4APCEI)*, Wollongong 28–30 September 2009. *Journal of University Teaching and Learning Practice*, 6, 1, 49–60. `http://ro.uow.edu.au/jutlp`

Brocardo, M., 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, 7–8 May 2013.

Cambridge English Dictionary. (20/02/2020). *Cambridge University Press*, dictionary.cambridge.org/dictionary/english/essay-mill

Clarke, R., & Lancaster, T., (2006). Eliminating the successor to plagiarism? Identifying the use of contract cheating sites. Proc. *2nd International Plagiarism Conference*, Gateshead, UK, 19–21 June 2006. Learning Press.

Clarke, R., & Lancaster, T. (2007). Establishing a Systematic Six-Stage Process for Detecting Contract Cheating. Presented at the 2007 *2nd International Conference on Pervasive Computing and Applications*, Birmingham, UK. 26–27 July 2007.

Dawson, P., & Sutherland-Smith, W. (2017). Can markers detect contract cheating? Results from a pilot study. *Assessment & Evaluation in Higher Education*, 1–8. doi:10.1080/02602938.2017.1336746

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*, 155–170. WSF Press, Wroclaw.

Ellis, C., Zucker, I., & Randall, D. (2018). The infernal business of contract cheating: understanding the business processes and models of academic custom writing sites. *International Journal for Educational Integrity*, 14(1), 1–21. Springer. doi:10.1007/s40979-017-0024-3.

Gillam, L. (2013). Readability for author profiling? Notebook for PAN at CLEF 2013. *Proc. Int. Conference and Labs of the Evaluation Forum (CLEF) Notebook PAN*. 23–26 September 2013, Valencia, Spain.

Greetham, B. (2014). *How to Write Your Undergraduate Dissertation*. 2nd edition. Palgrave McMillan, UK.

Guerrero, F. (2009). A new look at the classical entropy of written English. *arXiv* preprint arXiv:0911.2284, 2009. www.arxiv.org

Juola, P. (2013). How a Computer Program Helped Show J. K. Rowling wrote A Cuckoo's Calling. *Scientific American*, 20 August 2013, Springer Nature.

Juola, P. (2017). Detecting Contract Cheating via Stylometric Methods. *Proc. Plagiarism across Europe and Beyond*. Brno, Czech Republic. 24–26 May 2017. 187–198.

Kulig, A., Kwapien, J., Stanisz, T., & Drozdz, S. (2017). In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, 375, 98–113. Elsevier. doi:10.1016/j.ins.2016.09.051.

Lancaster, T., & Clarke, R. (2007). Assessing contract cheating through auction sites – a computing perspective. *Proc. 8th annual conference for information and computer sciences*, University of Southampton, 28–30 August 2007.

Lopez-Escobedo, F., Mendez-Cruz, C.-F., Sierra, G., & Solorzano-Soto, J. (2013). Analysis of Stylometric Variables in Long and Short Texts. *Procedia – Social and Behavioral Sciences*, 95, 604–611. Elsevier. doi:10.1016/j.sbspro.2013.10.688.

McMillan, K., & Weyers, J. (2011). *How to Write Essays & Assignments*. 2nd edition. Prentice-Hall, USA.

Newton, P. (2018). How Common Is Commercial Contract Cheating in Higher Education and Is It Increasing? A Systematic Review. *Frontiers in Education*, 3. doi:10.3389/feduc.2018.00067.

QAA. (2017). Contracting to Cheat in Higher Education. `www.qaa.ac.uk/about-us/what-we-do/academic-integrity/publications-and-guidance\#`

Rogerson, A. (2014). Detecting the work of essay mills and file swapping sites: some clues they leave behind. *Proc. 6th International Integrity & Plagiarism Conference*, 1–9. Newcastle-on-Tyne, UK. 16–18 June 2014.

Rogerson, A. (2017). Detecting contract cheating in essay and report submissions: process, patterns, clues and conversations. *International Journal for Educational Integrity*, 13:10. Springer. doi:10.1007/s40979-017-0021-6.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. 27 (3): 379–423.

Sivasubramaniam, S., Kostelidou, K., & Ramachandran, S. (2016). A close encounter with ghost-writers: an initial exploration study on background, strategies and attitudes of independent essay providers. *International Journal for Educational Integrity*, 12(1) 1:14. Springer. doi:10.1007/s40979-016-0007-9.

Stamatatos, E. (2008). A Survey of Modern Authorship Attribution Method. *Journal of the American Society for Information Science and Technology*, 60(3) 538–556. doi:10.1002/asi.21001.

Authors

Robin Crockett, University of Northampton, University Drive, NN1 5PH Northampton, UK, e-mail: robin.crockett@northampton.ac.uk

Kirstie Best, University of Northampton, United Kingdom e-mail: Kirstie.Best@northampton.ac.uk