A STUDY OF 'IDEA' PLAGIARISM IN TWO UNDERGRADUATE STUDENTS' EMAILS USING FORENSIC AUTHORSHIP ANALYSIS AND PLAGIARISM DETECTION METHODS

SUMAYYA MULLA (Forensic Linguistics)

Abstract

Plagiarism in student's writing has long been a prolific issue facing many academic institutions. With today's plagiarists becoming increasingly aware of the limitations of existing detection systems and 'idea' plagiarism now something of an academic conundrum, this study firstly seeks to relate linguistic patterns of plagiarism to the computerized textual features utilised in plagiarism detection tools and secondly employs stylometry from authorship analysis to generate a statistical analysis of the literary style. It begins with the premise that plagiarism has already occurred and aims to determine whether the stylistic, structural and semantic choices of writers can detect the direction of 'idea' plagiarism and to what extent this grammar-based method is effective. The study used 'known' plagiarism writings between two undergraduate students to test whether the original and plagiarised documents were identifiable through the use of both existing and novel methods within the discipline of forensic linguistics. Through qualitative and statistical data analysis, including correlation graphs, the conclusion was reached that ideas can be easily merged with a writer's own idiolect which renders 'idea' plagiarism almost impossible to identify with another writer's typical language features.

Keywords: Forensic Linguistics, Plagiarism, Originality, Authorship, Idiolect.

Introduction

To accurately and effectively measure the extent to which an original idea document can be distinguished from a plagiarised idea document, in the context of the academic community and email genre, the term 'idea plagiarism', including the distinction between literal and intelligent plagiarism, requires definition. Similarly, both forensic authorship analysis and existing plagiarism detection methods require elaboration in the novel area of known or established plagiarism. Providing a 'single and comprehensive' definition of plagiarism is problematic in light of its 'diversified nature as a practice' (Hussein 2014, 37). An educationally-based definition, however, can be founded on the concept of intellectual property and defined as 'an appropriation of ideas' (Gibbons & Turel 2008, 265) and 'a form of cheating' to 'lift an idea as one's own' (Hussein 2014, 37). Likewise, linguistic plagiarism differs from idea plagiarism with the latter being 'the most serious', 'more difficult for linguists to establish' (Gibbons & Turel 2008, 271) and includes both the plagiarised idea (content) and the language used (form) as relevant issues of consideration (Gibbons & Turel 2008, 266).

Similarly, the distinction between literal and intelligent plagiarism, the former denoting a simple 'copy and paste of text with few alterations' and the latter including 'text manipulation using lexical and syntactical paraphrasing, synonyms and sentence restructuring' (Salha et al. 2012, 135-6), renders idea plagiarism as a complex 'linguistic phenomena' (Pecorari 2010, 1) of 'sophisticated "borrowing"' (Coulthard & Johnson 2007, 187-8). The email genre, used 'increasingly for criminal purposes' (Calix et al. 2008, 1) is equally a linguistic phenomenon, in that it is a hybrid of both spoken and written languages which both have different organisation structures, lexical usage patterns and grammatical constructions (Coulthard 2005, 9). Therefore, with plagiarists becoming increasingly aware of the limitations of existing detection systems, such as the lack of paraphrase check on the academic TurnitIndetector (Maurer et al. 2006, 1074-7), together with idea plagiarism, which recently has become a prolific and real academic problem' (Salha et al. 2012, 134) facing many academic institutions (Gibbons &Turel 2008, 270), this essay, firstly, seeks to relate linguistic patterns of plagiarism with the computerized textual features utilised in plagiarism detection tools (Salha et al. 2012, 134), and secondly to use stylometry from authorship analysis to generate a statistical analysis of the literary style (Zheng 2006, 379). In relation to the legal context, linguistic markers and discourse strategies are 'decisive' in establishing a solid plagiarism detection case (Gibbons & Turel 2008, 265). Thus, this project aims to use 'known' plagiarism writings between two undergraduate students, to test whether the original and plagiarised documents can be identified through the use of both existing and novel methods within the discipline of forensic linguistics. Due to the limited scope of the analysis, factors of dialect, culture and 'translation plagiarism' will not be discussed.

Literature Review

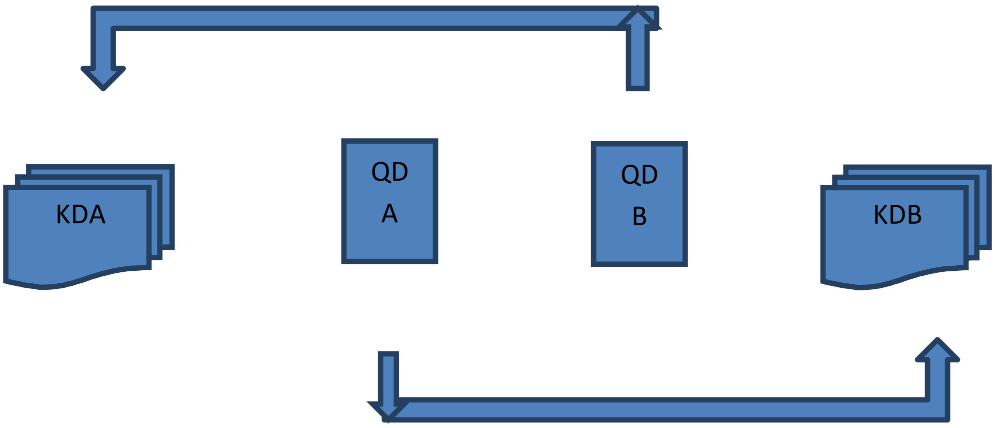
Previous research into idea plagiarism is limited to description and does not extend in detail to methods of identification, analysis or previous studies. Due to this limitation the present study will highlight the range of: methods, linguistic categories, and theoretical perspectives in order to conduct an encompassing analysis and to reach complete conclusions. Firstly, the methods of authorship analysis and plagiarism detection methods will be considered. Secondly, the categories, characteristics and combinations of linguistic features will be measured. Thirdly, theoretical perspectives will be considered under the terms of idiolect, variation, genre and style and finally limitations and problems will be deliberated to identify gaps in the research.

Methods of authorship analysis and plagiarism detection

The notion of a 'write-print', achieved by extracting patterns of language from question documents (QDs) to identify consistent and distinctive similarities with known documents (KDs), in an aim to identify an author (Iqbal et al. 2008a, 45), is merely applicable in the case of known plagiarism. It should be noted that a write-print differs from the concept of the linguistic fingerprint, in that the former identifies *similarities* in the *patterns* of language use, whereas the latter is an 'unhelpful, if not actually misleading metaphor' (Coulthard 2005, 6), suggesting that any linguistic data set contains all the necessary information for the identification of an individual, whereas in reality, even a large data set would provide only 'very partial information about its creator's idiolect' (ibid.). Thus, in regards to the write-print, only when similarities are identified between the different authors' work in the combinations illustrated in Figure 1 below, would it be possible to infer significant patterns to distinguish an original from a plagiarised idea.

Figure 1

KD (Known Document); QD (Question Document)



Likewise, linguistic stylistic analysis, which incorporates individual linguistic characteristics of an individual, known as their idiolect, and 'exploits the principal of inherent variability in language' that 'no two writers write in exactly the same way and that no individual writer writes consistently in an identical manner (McMenamin 2002, 163), is only partially true, as paraphrasing would disguise and conceal the attempts of authors to 'write in the same way'. Furthermore, attribution methods, in which there is typically a set of candidate authors and text samples of unknown authorship (Stamatatos 2009, 545) is, again, not applicable, as all the authors of the text samples in a known plagiarism case are already established. These limitations in the research and existing methods, however, may be partially overcome by overlapping the areas of authorship attribution and plagiarism to establish linguistic, grammatical and even pragmatic markers (Hussein 2014, 38), such as discourse, sentence and phrase structure. Thus the combination of traditional statistical data with a grammatical analysis would uniquely provide a set of grammar–specific identifying variables which are representative of both academic and email language genres.

Categories, characteristics and combinations of linguistic features

In regard to the methods, features, and concepts available for analysis, the following are used to identify plagiarism: plagiarism detection, write-print extraction, qualitative and quantitative methods, similarity detection and cluster or combination analysis. Similarly, the characteristics of linguistic features may be described under the headings of content-specific and application-specific, style, structure, syntax, semantic, character-based, lexical and functional features.

In meeting the definition of textual plagiarism there must be a set of repeated words, ideas or linguistic structures between the question documents that do not exist in the known documents, to demonstrate that the texts cannot be coincidental (Pecorari 2010, 6). Similarly, analysis may be conducted by either extrinsic plagiarism detection (Salha et al. 2012, 137), a word-to-word basis with the comparison made against a body of known documents (Maurer et al. 2006, 1056-7), or intrinsic plagiarism, an analyses of the query document in isolation (Salha et al. 2012, 137).

A write-print, defined as the combination of lexical, syntactical, structural and content-specific features that occur frequently in an individual's writing, is a concept borrowed from 'pattern mining' and 'frequency pattern', a technique for finding 'hidden patterns', and has been uniquely applied to authorship analysis through the method of filtering out the common frequent patterns to then identify the unique patterns that differentiate the writing style of a suspect (Iqbal et al. 2008a, 43).

Qualitative or descriptive analysis involves specifying the range of variation by describing the collective set of all variation and deviation at every linguistic level, for both the question and known documents (McMenamin 2002, 122), and is an effective method to appeal to the 'nonmathematical but structured sense of probability held by judges and juries' (McMenamin 2002, 129). Although qualitative evidence is 'more demonstrable' than quantitative evidence in court, the latter is essential to test whether any differences or similarities in the data are significant, and to evaluate the significance of any relationships between variables across the documents (McMenamin 2002, 137-8). While statistical tests range from: evaluating potential relationships among variables, evaluating the independence of variables, comparing two percentages, and finding the standard error of difference (proportion test) (McMenamin 2002, 138-45), it is 'frequency distribution', which is not test but a description of the data, that would best demonstrate relationships and correlations in the data or lack thereof (McMenamin 2002, 139). Similarly, the 'length of borrowed strings' and the 'linguistic dexterity' (Pecorari 2010, 21) of borrowed/copied ideas requires consideration.

Theoretical perspectives

Linguistic variation

Regarding linguistic variation, McMenamin states that '[t]he measurement of variation in written language is a powerful complement to its description' and therefore is important to a successful analysis and interpretation of style (2002, 137). Thus, variation studied from the 'bottom-up approach model searches for recurrent patterns, distributions, and forms of organization in the writing' in order to demonstrate evidence for 'the presence of units, existence of patterns, and formulation of rules' related to a writer's individual style as opposed to the 'top-down model of style analysis' which starts with a 'predetermined taxonomy of stylistic items' which would allow for 'the discrimination of writers within a certain speech community' (McMenamin 2002, 54).

Idiolect

The notion of idiolect, described both as the 'range of linguistic variation' (McMenamin 2002, 121) and the characteristic features in the writing style of an individual, is an integral concept to authorship analysis.

Genre and Style

The application of authorship analysis to online messages in recent years (Zheng 2006, 57) demonstrates that it is possible a writer's unique style can be 'reduced to a pattern' through the construction of measurements of various stylometric features from the written text itself (Calix et al. 2008, 1). Thus typological and character-based features are important variables to consider in the methodology. Similarly, style relates the language choices of writers to the variables of topic, purpose of interaction, writers' social characteristics and their education level (McMenamin 2002, 110). Style, therefore, is the most significant to demonstrate the different variant forms of language possible to express similar linguistic meaning. This is because a writer's linguistic style contains features independent of their will, and therefore cannot be consciously manipulated by the author.

Gaps in the research and problems to overcome

Many of the research questions asked and recommendations for future study put forward in the existing literature relate largely to traditional authorship analysis or developments for plagiarism detection tools. Although there is discussion on what 'idea' plagiarism constitutes, actual studies on 'idea' plagiarism are almost non-existent. Similarly, studies on cases of 'known' or established plagiarism, together with methods of authorship analysis or plagiarism detection, cannot be sourced in the existing literature.

The principal problems to overcome include: the time-consuming and labour intensive aspects of statistical analysis without computer software (McMenamin 2002,138), establishing a scale of how different or similar patterns must be to reach conclusions (Gibbons & Turel 2008, 274) to identify the 'write-print' or patterns of language across an extensive set of variables, which are required for accuracy (Iqbal et al. 2008a, 46), to choose only the methods which are 'feasible' in an academic environment' (Maurer et al. 2006, 1061) and to find solutions to detect paraphrasing and synonym usage which remain undetected within the traditional plagiarism software. Solutions to these problems are:

- to filter out any common frequent patterns (Iqbal et al. 2008b, 59)
- to correlate all textual features in graphs to create accurate and extensive write-prints or patterns of similarity
- to choose a grammar-based approach to frequency distribution to highlight differences and similarities in the academic ability of the writers
- to shed light on paraphrasing and synonym usage by incorporating the novel grammatical variables of phrase structure and semantic overlap.

Data and Methodology

Research questions and aims

The following research questions were addressed:

- Can novel methods (a grammar-based approach) shed light on effective approaches to 'idea' plagiarism in academia?
- To what extent is it possible to consistently and effectively identify patterns of similarity and difference across an extensive range of variables and methods?
- Is an encompassing analysis from the many perspectives of styles, structure, lexis and semantics, more effective in identifying creative plagiarism than the computerised variables to identify copied textual features?

Data experiment and variables

Once the question of known idea plagiarism between students' work was highlighted as an issue with need for further study, the task was to locate writing samples of equal variables. Thus the variables of educational degree type, level of study, formality and length of text were to be kept consistent. Within the original study, to analyse students' essays would prove too time-consuming and difficult to locate, hence, the genre of student email was adopted to re-create a more focused study of the same question.

Two students were recruited. Participant A was required to write a hypothetical formal email to a tutor about a project idea/proposal that he/she had thought of. This constituted the original email / question document (QDA). Participant B was asked to the read the original email and plagiarise the idea by also writing an email addressed to the same tutor with either an identical or very similar idea. This second email constitutes the plagiarised email (QDB).

Data collection

Participants were instructed to agree amongst themselves who would write the original email (QDA) and who would create the plagiarised version (QDB). Both emails were anonymised and forwarded to me from only one of the volunteers.

After these question documents were sent to me, a request was sent for ten formal emails of similar length and formality that had been sent to academics from each of the volunteer's university mailboxes. These ten emails, being known documents (KDs) would then be used as a comparison with the question documents (QDs).

Extraction of features

The following methods were combined to build a feature-based classification model: structural, syntactic, lexical and content specific features (Zheng 2006, 378). Likewise, 'statistically-supported' authorship attribution was conducted by measuring the frequency of textual features to facilitate the differentiation of texts written by different authors (Stamatatos 2009, 538). Similarly, when giving opinions in court, it has been proven that both qualitative and quantitative approaches to plagiarism detection are valid and complementary (Gibbons & Turel 2008, 265). Thus, both these methods were used to analyse the date into tables and correlation graphs.

The features used in the analysis were determined by a combination suggested in previous research and were used to provide a wide analysis spectrum on a small sample pool of data. The features of style, including comma usage, capitalisation, punctuation, coordination, subordination, embedding, determiners, pronouns and questions were used. In regard to features of structure and syntax, discourse, sentence and phrase structure were analysed, along with non-grammatical constructions. Thirdly, features of semantics, lexis and content-specific words were analysed primarily under content words and shared lexis. All features were identified, totalled in terms of frequency, and calculated into averages across all the data samples, both between and within the QDs and KDs.

Strengths and limitations

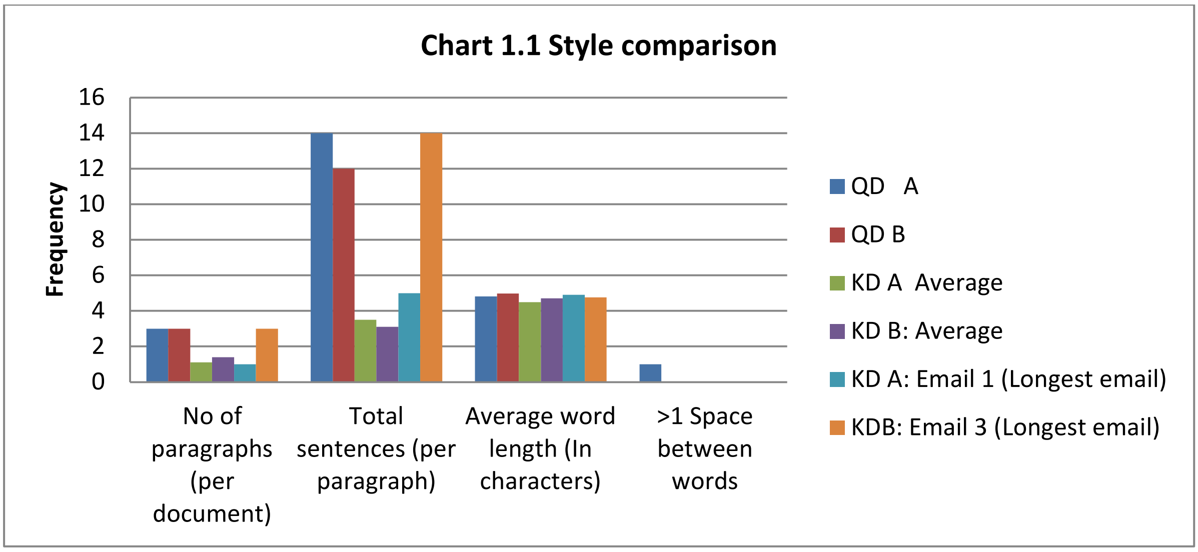
As far as the author is aware, this project is unique in that it begins with the understanding that plagiarism has already occurred. To maintain a focused study, the following variables were also kept consistent: document length, formality and register of the text, student's degree subject, and the student's level of study. The varying lengths of the emails initially posed a problem; the solution was to ensure that the total number of lines, rather than sentences, from the KDs, were equal to those of the QDs. The QDs were on average 30 lines in length and the KD is total were on average 31. To further remedy this problem, the longest email (most similar to the QDs in length) from the KDs of each participant was considered as a separate variable alongside the averages of all the KDs. This would provide insight into the extent that the genres of academia and email impact: language use, text length and structure, in spite of formality and education levels being kept consistent. Similarly, the text samples are request writings and thus do not wholly represent a natural plagiarism of ideas. For the purpose of this study, however, this factor could not be changed and the results had to account for this.

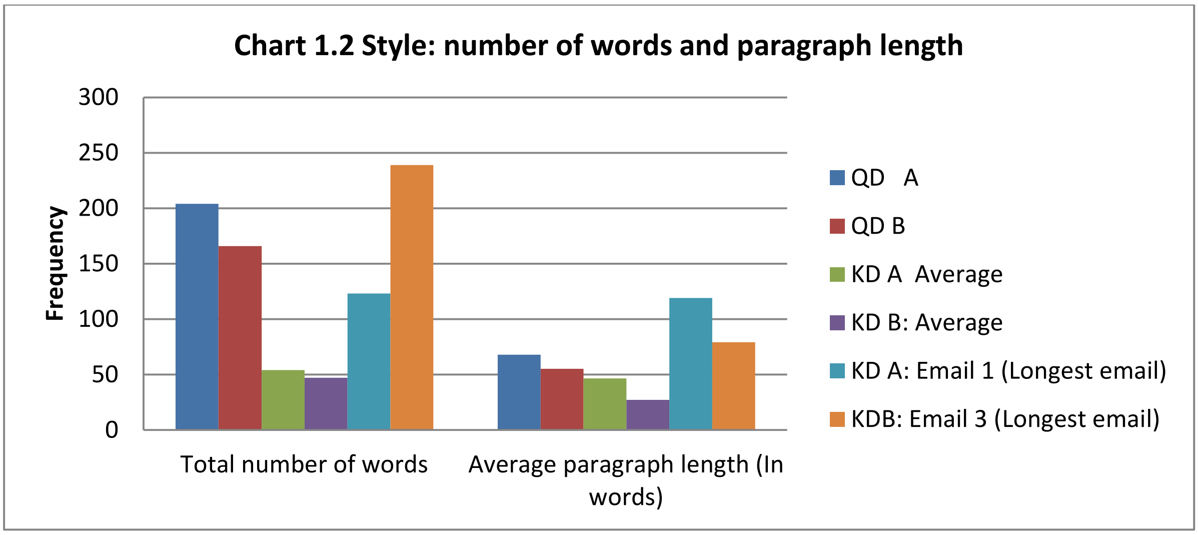
Results and Discussion

1. Stylistic

To a partial extent stylistic features can identify patterns which demonstrate the direction of 'idea' plagiarism. These are compared in Chart 1.1 below.

Many expected correlations, such as the similarities in QDB and KDB are shown here, but Chart 1.1 also shows correlations which, in reality, are unlikely to represent a distinctive similarity. For example the feature of [> 1 space between words] has a very low frequency and it would not constitute a distinctive feature. The same may be said of Chart 1.2, number of words and paragraph length, which demonstrates no distinctive correlation across the data set, but instead only shows individual frequency patterns.





When considering the mean line length (Chart 1.3 below) there is a stronger correlation between QDB, KDA Email 1 and KDA, than with KDB or KDB Email 3, suggesting the frequency of ‘mean line length’ from the question document is abnormal for participant B in relation to their known writings.

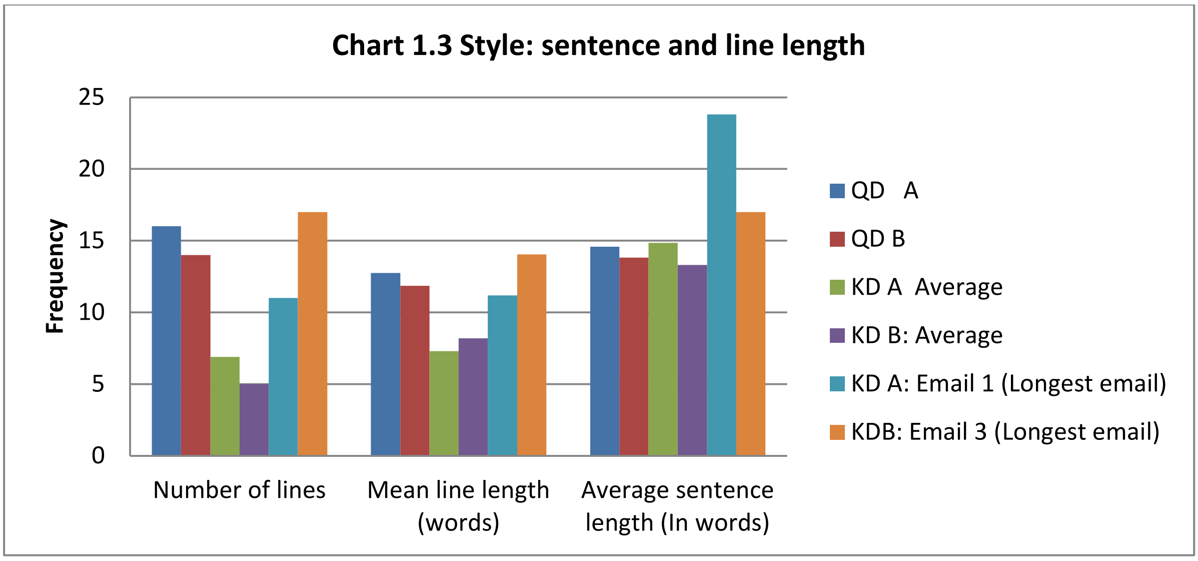


Chart 1.3 Style: sentence and line length

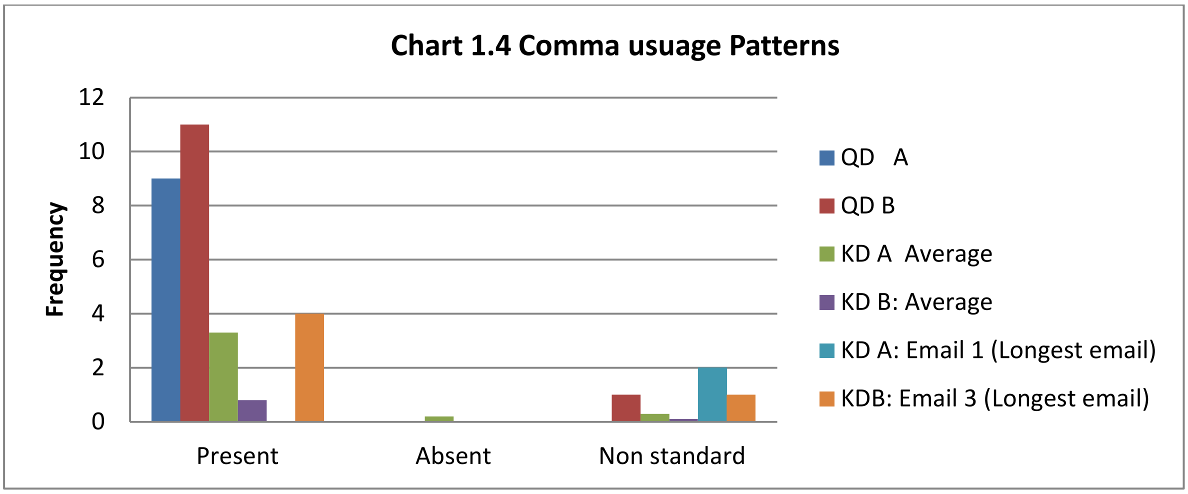
Similarly, this pattern is replicated in 'average sentence length' with QDB correlating with KD average more than the KB average or KB longest email, both of which lack consistency. Table 1 represents the statistical data used for the correlations.

Table 1

		QD A	QD B	KD A		KD B	
		N	N	Average	All 10emails	Average	All 10emails
Style							
1	Line spacing	1 space (4 x)	1 space (4x) 2 space (1x)	1 space (2)		1 space (2)	
2	No of paragraphs per document	3	3	1.1	11	1.4	14
3	Average paragraph length (In words)	68	55.33	46.45	464.5	27.36	273.7
4	Total sentences (per paragraph)	3,8,3 (14)	2,7,3 (12)	3.5	35	3.1	31
5	Average sentence length (In words)	14.57	13.83	14.86	148.6	13.34	133.4
6	Total number of words	204	166	54	540	47.3	473
7	Total characters Average word length (In characters)	983 4.81	828 4.98	248.3 4.49	2483 44.9	217.2 4.70	2172 47.03
8	>1 Space between words	2 Spaces x1	0	0	0	0	0
9	Number of lines Mean line length (words)	16 12.75	14 11.85	6.9 7.3	69 73	5 8.28	50 82.82

Commas

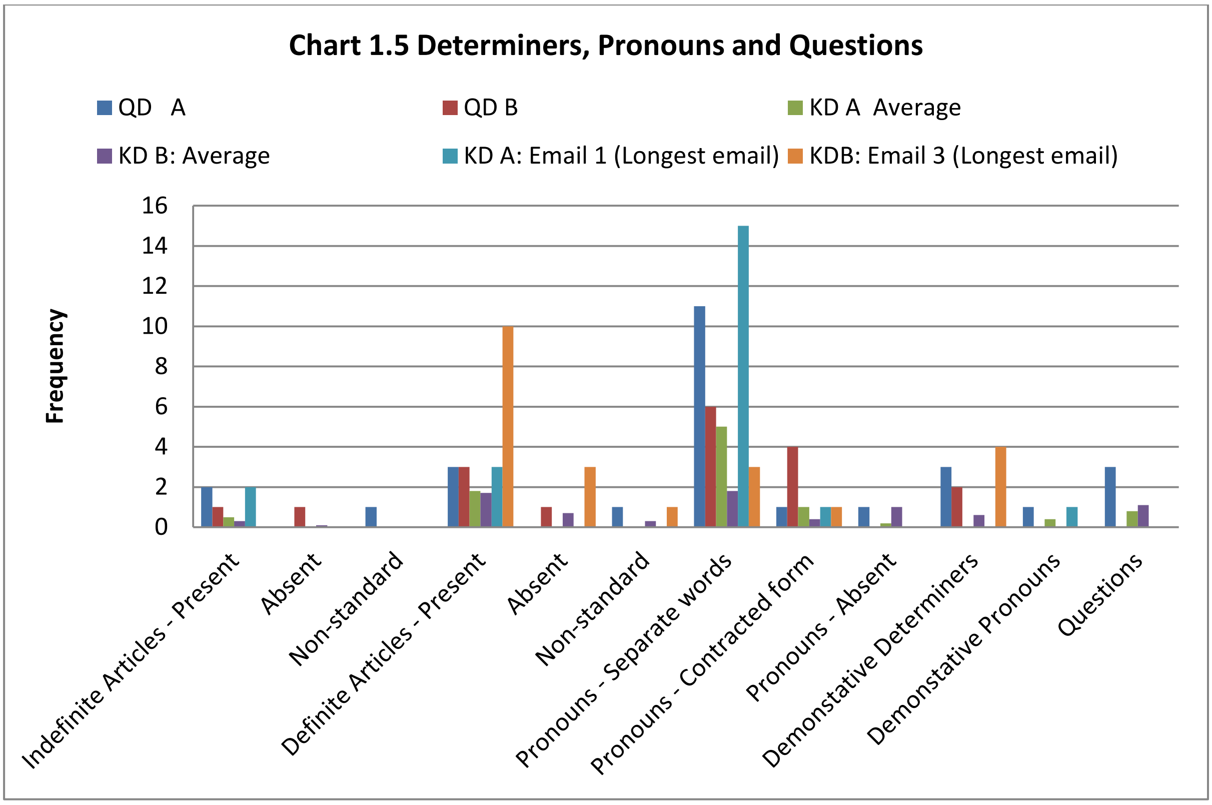
As shown in Chart 1.4, 'comma usage patterns' are not able to represent any significant correlations in the data, except the similarities that are expected between the writings of the same author. If qualitative data is consulted it becomes evident that although writers differ in their correct, incorrect and absent comma usage, the act of idea plagiarism will not hugely alter functional features. What is likely is that the idea being plagiarised would be merged with the language characteristics of that writer.



Capitalisation, punctuation, coordination, subordination and embedded clauses

Like comma usage, capitalisation, punctuation, coordination, subordination and embedded clauses were investigated, but no patterns were found among the data. Thus, it may be concluded that functional features of style, which are largely unconscious efforts by writers, are not valuable or useful variables to demonstrate the direction of idea plagiarism beyond what we expect.

Chart 1.5, below, concerning lexical items, should be more conducive to demonstrate correlations than functional features. In the feature of 'separate pronouns' QDB usage has again, a stronger correlation with QDA, KDA Average, and KDA Email 1 than with KDB Average. Similarly the high feature count of 'definite articles' of KDB Email 3, suggests the lengthy emails is not a consistent length but a rare frequency which does not tally with the pool of KDB's writings.



Stylistic summary

Stylistic features, to a partial extent, have pointed to some patterns linking QDB with the KDs of participant A. However, limitations far outweigh the strengths, as functional categories showed no significant or consistent correlations, suggesting that they cannot shed light on the direction that 'idea' plagiarism has occurred. Lexical features, however, have demonstrated a greater likelihood of identifying significant patterns and consistent conclusions.

2. Structural / Syntactic

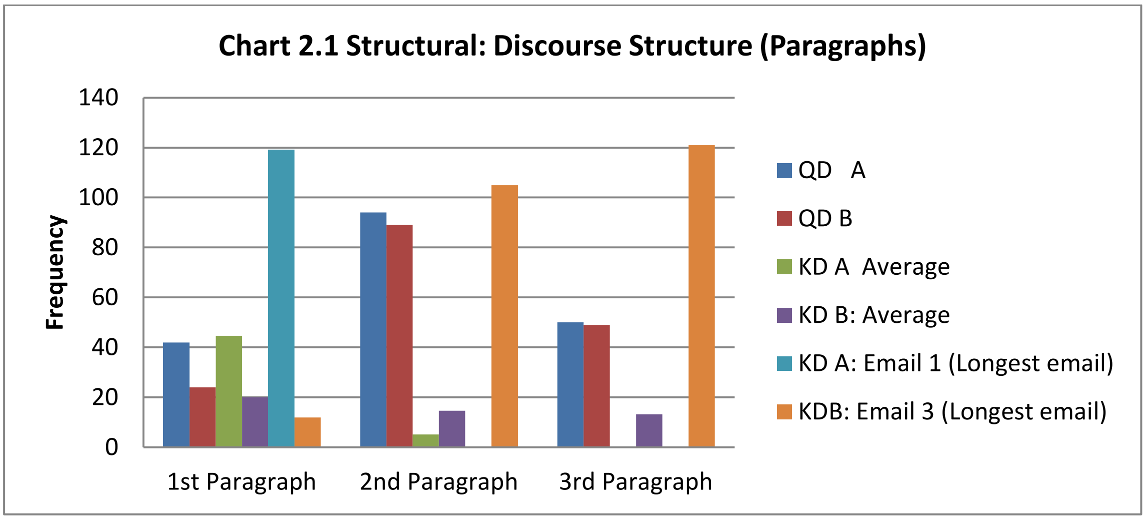
In the category of structural features the following categories will be considered to create an analysis based on the entire sample document: discourse structure, including paragraph structure, greetings and signature usage; sentence structure, including, main / subordinate clauses and sentence starters; phrase structure and non-grammatical constructions, including verb forms, noun phrase, propositional phrases, and conjunctions.

Discourse structure

Discourse structure was used to analyse the entirety of all the texts and provide an analysis beyond word or cluster and focus instead on the structure and organisation of the text.

Paragraphs

Chart 2.1 demonstrates a consistent similarity between QDA and QDB but suggests mixed findings along the number of paragraphs. For example the first paragraph indicates a small correlation between QDB and KDA, but when consulting the qualitative data, it is apparent that all the sample writings contain a first paragraph. Therefore this correlation cannot be considered as distinctive. The second paragraph does show a strong and distinctive correlation between the QDs and KDB Email 3, but this also cannot be labelled as consistent as only one email from a total of ten bears this pattern. Thus, once again, correlations which do suggest a relationship are either distinctive or consistent, they are not both. As these patterns cannot be labelled as both distinctive and consistent, they therefore cannot form solid evidential data.



Greeting and signature

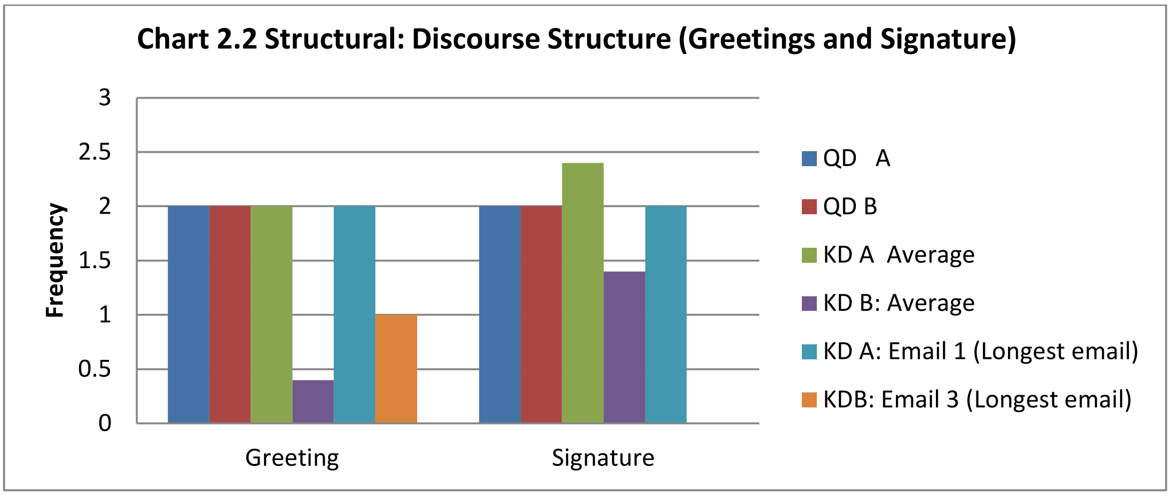


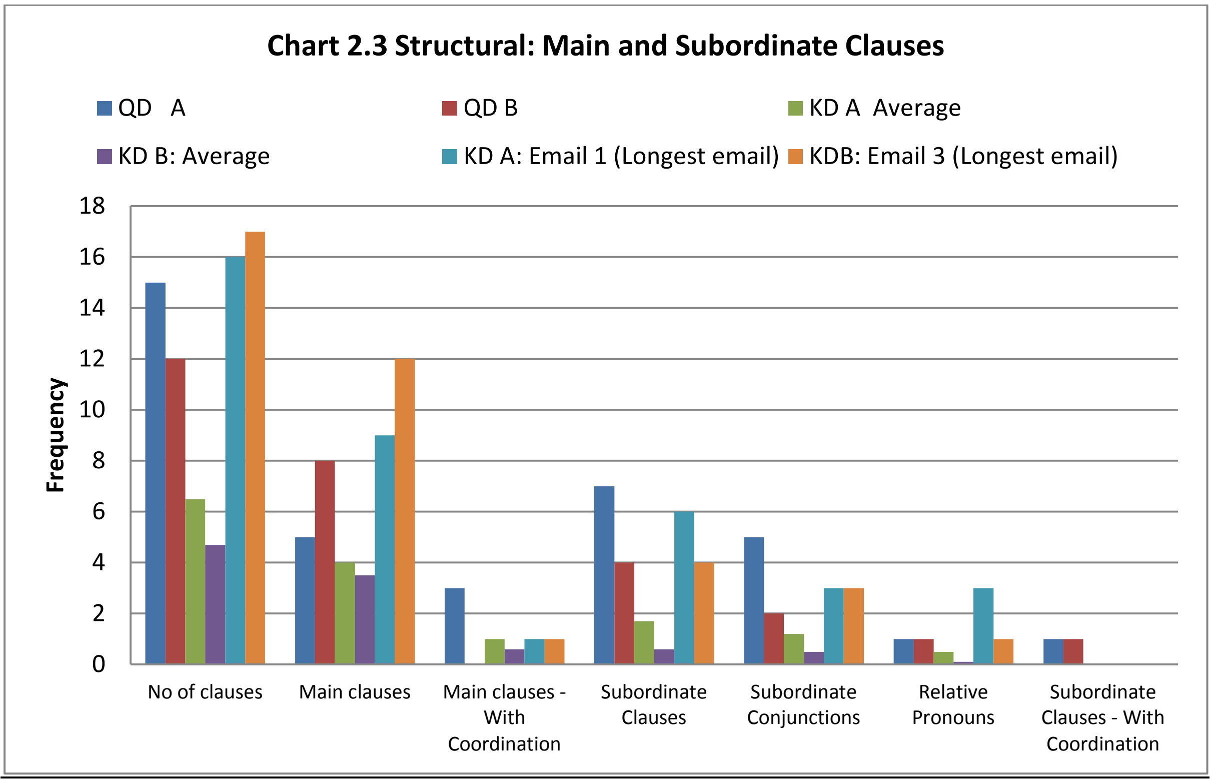
Chart 2.2, above, suggests a consistent and also distinctive correlation between the QDs , KDA Average and KDA Email 1 suggesting, in terms of greetings and signatures, that participant B consistently bears features and frequencies associated with the known and questioned writings of participant A. This evidence, however, needs to be taken into consideration alongside the other features being tested as one feature alone is not sufficient evidence.

Sentence structure

Sentence structure as a variable was used to identify language combinations, distinctive patterns and consistent reoccurrences.

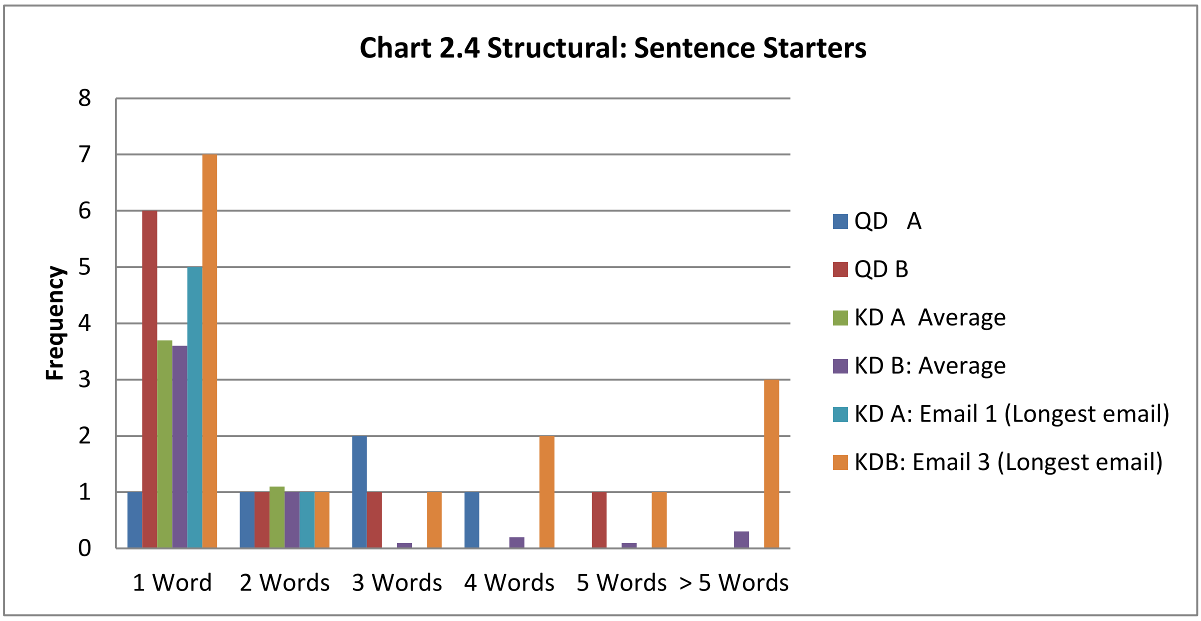
Main and subordinate clauses

Correlations between QDA and KDA Email 1, demonstrate a consistent pattern across the grammatical categories being tested. Also in the feature of 'number of clauses' the correlation between QDB and KDB Email 3 is not as strong and the KDB average has the lowest frequency suggesting this correlation is not consistent and further still, that it is rare and not the average usage pattern of participant B. Other significant correlations, once again, all point to expected patterns in the data. For example in the feature of subordinate clauses the following expected patterns are demonstrated: QDA and KDA Email 1/QDB and KDB Email 3.



Sentence starters

The pattern between QDB and KDA Email 1 in the frequency of ‘1 word’, again, shows a distinct correlation. In the other categories, however, it is difficult to identify any correlations with the frequencies being so low. When qualitative data is consulted it remains equally difficult to find consistent patterns in sentence starter usage. Like functional categories, a conclusion could be made that idea plagiarism does not included the plagiarism or copying of general language features, but only lexical content words or the combination of phrases which contain the idea being plagiarised.

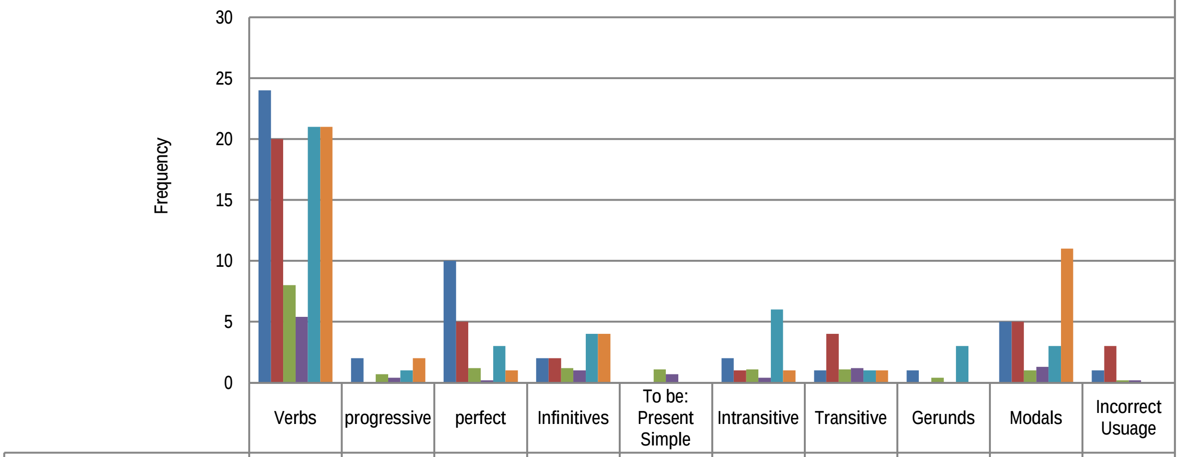


Phrase structure

Similar to discourse and sentence structure, phrase combination was used as a variable to establish the norm for patterns of language use of each participant and thereafter identify any deviations from the established patterns in their work.

There are correlation in the verbs forms, particularly in the features of general verbs, perfect aspect and modals, between QDB and KDA Email 1. These correlations are distinctive, but are not, however, consistent along all the verb forms. This limitation may be a natural, unconscious part of language use. As the data suggests, it is unlikely that idea plagiarism includes the imitation of verb types. The plagiarist is more likely instead to paraphrase or use synonyms, especially in the case where the concealment of plagiarism is creative and ‘intelligent’ and includes ‘text manipulation, lexical and syntactical paraphrasing, synonyms and sentence restructuring’ (Salha et al. 2012, 135-6).

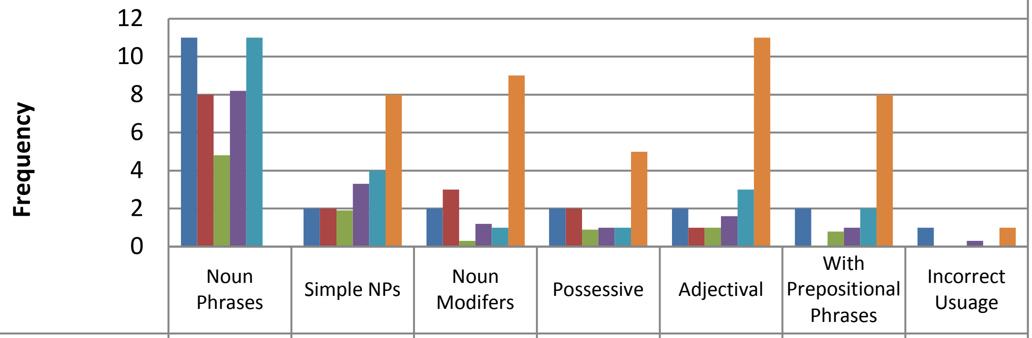
Chart 2.5 Phrase Structure: Verb Forms



	Verbs	progressive	perfect	Infinitives	To be: Present Simple	Intransitive	Transitive	Gerunds	Modals	Incorrect Usage
■ QD A	24	2	10	2	0	2	1	1	5	1
■ QD B	20	0	5	2	0	1	4	0	5	3
■ KD A Average	8	0.7	1.2	1.2	1.1	1.1	1.1	0.4	1	0.2
■ KD B: Average	5.4	0.4	0.2	1	0.7	0.4	1.2	0	1.3	0.2
■ KD A: Email 1 (Longest email)	21	1	3	4	0	6	1	3	3	0
■ KDB: Email 3 (Longest email)	21	2	1	4	0	1	1	0	11	0

A similar argument can be made for noun phrases; again, only expected correlations between the writers' own texts are demonstrated. One correlation which is distinctive, however, is the feature of 'simple NPs', as QDB and KDA Average are almost identical. This correlation is further strengthened as the correlation between QB and KDB is weak.

Chart 2.6 Phrase Structure: Noun Phrases



	Noun Phrases	Simple NPs	Noun Modifers	Possessive	Adjectival	With Prepositional Phrases	Incorrect Usage
■ QD A	11	2	2	2	2	2	1
■ QD B	8	2	3	2	1	0	0
■ KD A Average	4.8	1.9	0.3	0.9	1	0.8	0
■ KD B: Average	8.2	3.3	1.2	1	1.6	1	0.3
■ KD A: Email 1 (Longest email)	11	4	1	1	3	2	0
■ KDB: Email 3 (Longest email)	0	8	9	5	11	8	1

Prepositional phrases

The prepositional features ‘in’, between’ and ‘of’ demonstrate an identical and consistent correlation and usage pattern between QDA and KDB Email 3. This conclusion is dissimilar to the previous correlations of QDB with KDA. This implies, firstly, all conclusions must be taken together and not as separate variables when constructing the evidential data and, secondly, the methods of statistical and grammar-based analysis approach are proving inconsistent to demonstrate distinctive correlations between the data sets. This, however, may predominantly be due to the fact ‘idea’ plagiarism does not necessarily equate to the imitation of writing.

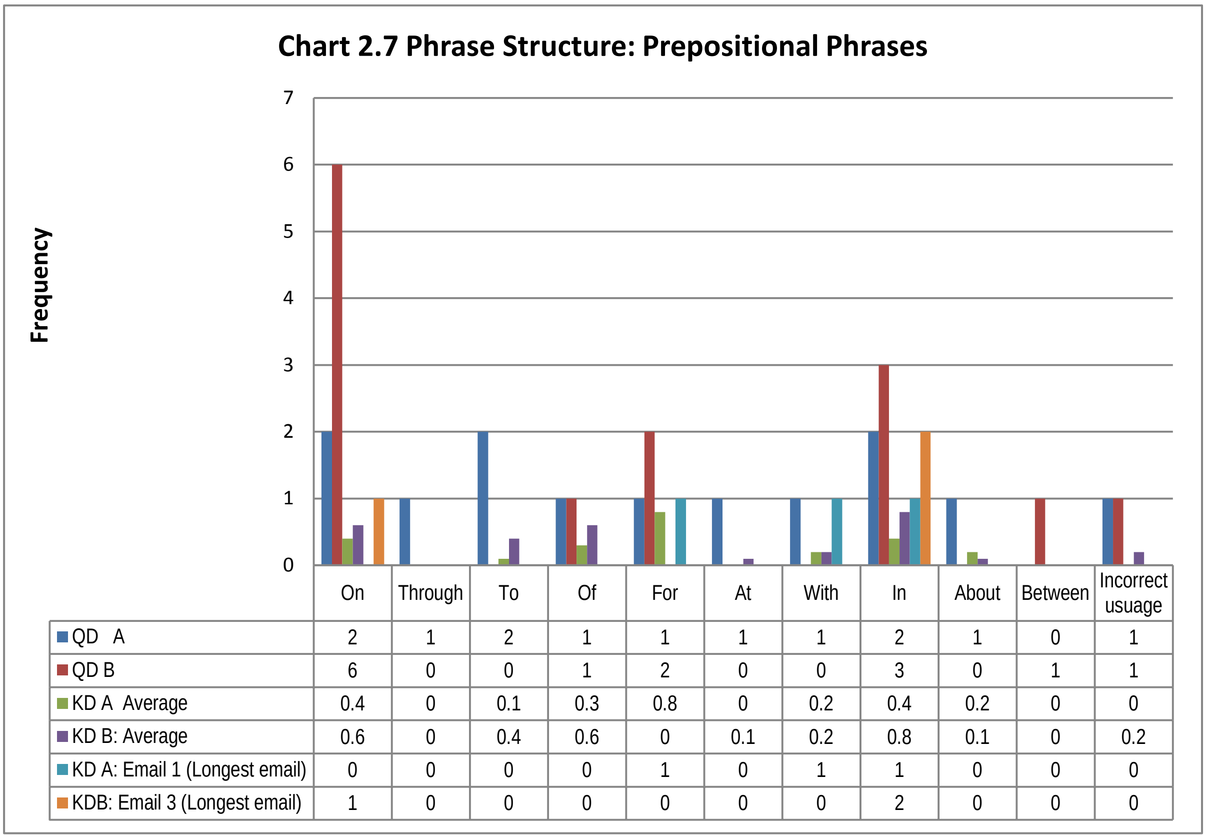


Chart 2.7 Phrase Structure: Prepositional Phrases

	On	Through	To	Of	For	At	With	In	About	Between	Incorrect usage
🟦 QD A	2	1	2	1	1	1	1	2	1	0	1
🟥 QD B	6	0	0	1	2	0	0	3	0	1	1
🟩 KD A Average	0.4	0	0.1	0.3	0.8	0	0.2	0.4	0.2	0	0
🟪 KD B: Average	0.6	0	0.4	0.6	0	0.1	0.2	0.8	0.1	0	0.2
🟦 KD A: Email 1 (Longest email)	0	0	0	0	1	0	1	1	0	0	0
🟧 KDB: Email 3 (Longest email)	1	0	0	0	0	0	0	2	0	0	0

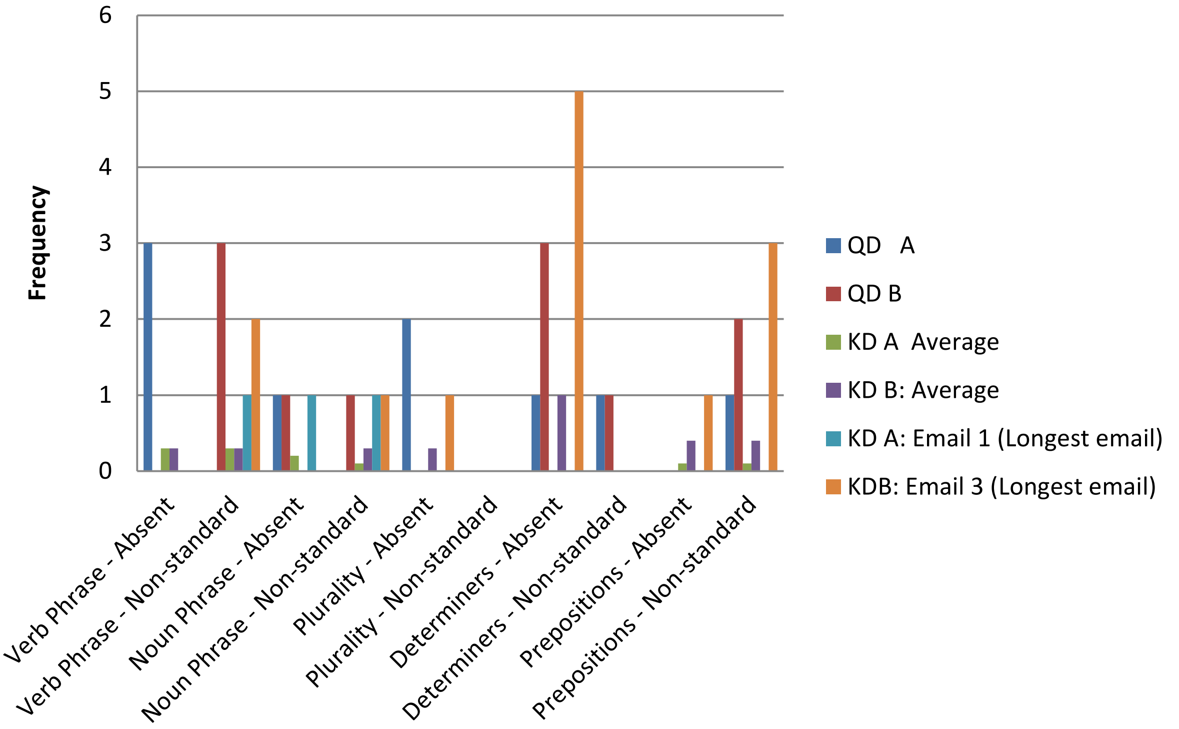
Conjunctions

The use of conjunctions, like the other functional features tested, demonstrates the same results: only the expected correlations between the participants’ own writings are highlighted.

Non-grammatical constructions

In the feature of ‘absent determiners’, a correlation is demonstrated between QDA and KDB Average especially as the correlation between QDB and KDB Average is very weak. Similarly, in the feature of 'absent noun phrases', QDB and KDA Email 1 are almost identical. With the general conclusion made, however, that functional categories generally cannot represent the direction of 'idea' plagiarism as the features themselves are unlikely to have been plagiarised, the analysis of these correlations can be put under scrutiny.

Chart. 2.8 Non-Grammatical Constructions



Structural / Syntactic Summary

To summarise, yes, to a certain extent, the direction of 'idea' plagiarism can be in the instances where the correlations match QDB to KDA and simultaneously show a weak correlation between QDB and KDB. Consistent and distinctive correlations, however, can be suggested or indicated but not definitely identified.

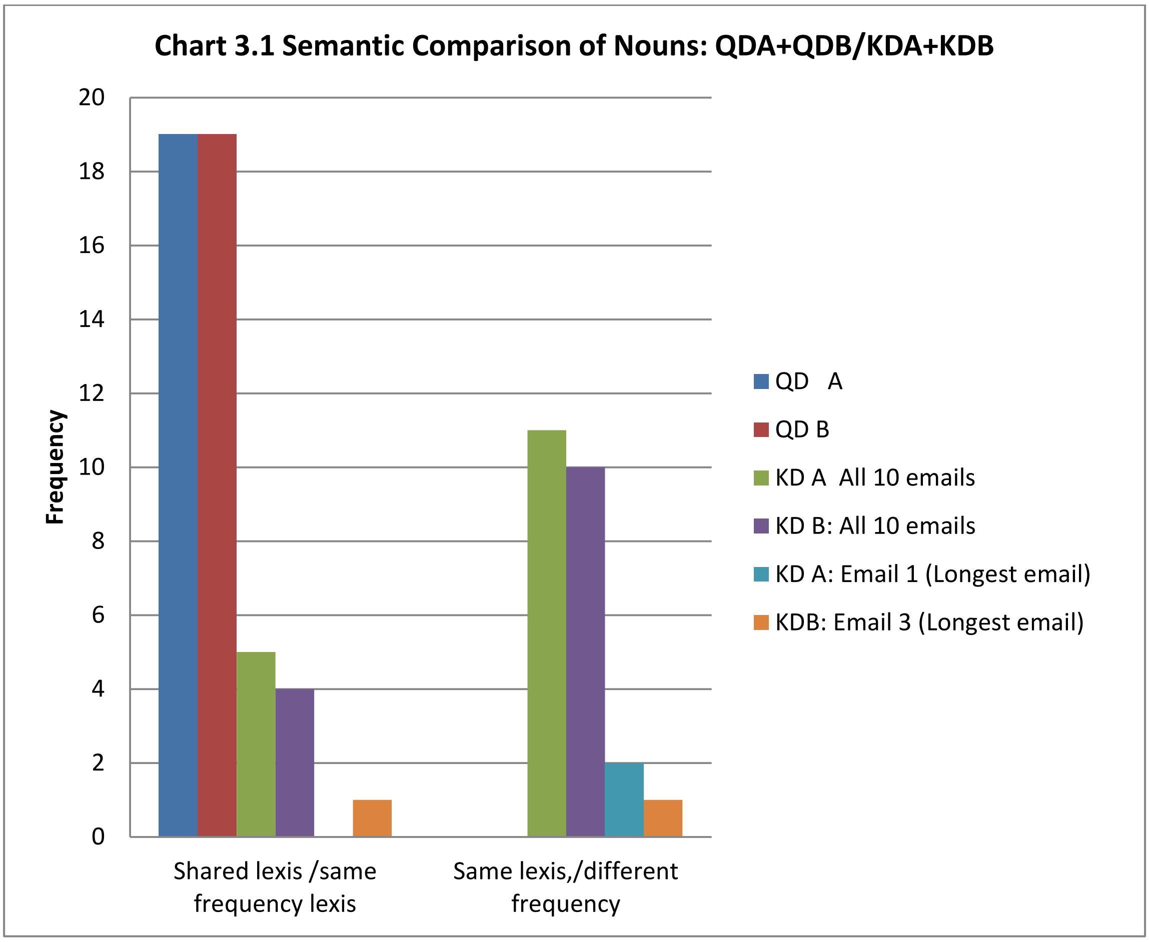
On the other hand, there have been several limitations in the consistency and distinctiveness of correlations which are demonstrated in the statistical and grammar-based data. The majority of significant correlations which have been demonstrated are expected patterns. This sheds light onto the difficulty, problems and limitations in demonstrating consistency in the direction of 'idea' plagiarism.

3. Semantic/lexical/content-specific

The purpose for a semantic analysis was to highlight the lexis which is: identical, of the same and different frequencies, and also lexis which is not replicated in either KDA or KDB. This was in order to create a narrow specification of variables to draw accurate correlations from.

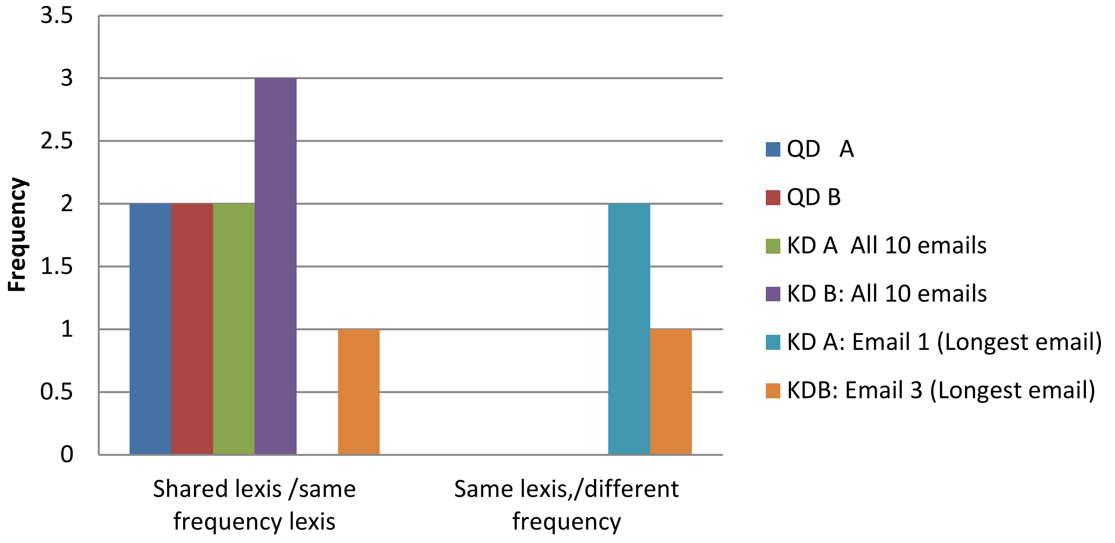
Nouns

The comparison between the QDs and KDs was to identify and later eliminate the correlations which may be coincidental or natural due to the language of the email genre and academic degree course. The correlations show the identical patterns between QDA and KDA, accurately acknowledging plagiarism has occurred, as particular lexis is identically repeated.



This cross-comparison analysis demonstrates a consistent correlation between QDB and KDA. This correlation is further solidified as the same correlation is also made in the category of 'same lexis/different frequency' between QDB and KDB, demonstrating accuracy in the patterns. On the surface this would suggest firstly that the direction of lexis is travelling between QDB and KDA, and secondly that the lexis in QDB is rare and distinctive, as it does not feature in the KDs of participant B but does feature in the KDs of participant A.

Chart 3.2 Semantic Comparison of Nouns: QDA+KDB/QDB+KDA



Semantic/lexical/content-specific summary

To summarise, the semantic comparison of lexis has demonstrated, like the stylistic and structural categories, the general trend that the features in question document B relate to the known documents of participant A. The frequency or consistency of these correlations, however, is not enough to conclude participant B created the plagiarised document. This is especially true, as opposite correlations have also been made to indicate participant A's question document has matched features with participant B's known writings.

Concluding summary of the study

To only a partial extent, the correlations created from statistical and grammar-based analysis do, to a small degree, indicate a direction of QDB to KDA, in all the categories of style, structure and semantics. Caution, however, should be taken, as these correlations are firstly only few in number, are not consistent, and correlations in the data have also suggested the opposite argument: that the direction of idea plagiarism may be QDA to KDB; however, these correlation patterns are in fact fewer in number.

Further study into qualitative data may perhaps prove useful to shed more light on methods which have the potential to act as effective tools to identify paraphrasing and synonym usage. This study has shown, however, idea plagiarism is not always possible to detect as actual writings, words and grammar do not necessarily need to be replicated, but instead can be creatively imitated and merged into the writer's own idiolect.

Overall conclusion

After completing the study and analysis, it was revealed that the plagiarised idea document was in fact QDB. It may be argued that it was the discourse structure, lexical items and semantics, rather than the functional features, which pointed partially towards this conclusion. However, these partial suggestions, even if correct, do not equate the methods of analysis used in this study to successfully and effectively identifying idea plagiarism, as consistency was not achieved in the correlations. Furthermore, with a larger set of data, analysis and the identification of significant correlations would prove very difficult and time consuming. It is recommended that a computerised grammar-based analysis method is used to tackle the problem of student plagiarism.

The principal strength of this analysis was in the grammar-based method, which focused on each feature in isolation. The predominant limitations of the study, however, were the time-consuming nature of data extraction and analysis, the problems of paraphrasing and synonym usage across a large set of data, and the fact that 'idea' plagiarism does not necessarily involve the re-production of language use, even in an 'intelligent' manner. Thus the conclusion can be made that ideas can be easily merged with a writer's own idiolect and thus become almost impossible to identify with another writer's typical language features.

References

Coulthard, M. 2005. 'Some Forensic Applications of Descriptive Linguistics', *Veredas – Journal of Linguistic Studies* 9:2, 9-28.

Coulthard, M. & Johnson, A. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Calix, K., et al. 2008. 'Stylometry for E-mail Author Identification and Authentication', *Proceedings of CSIS Research Day, Pace University.*

Gibbons, J. & Turel, T. 2008. *Dimensions of Forensic Linguistics.* Amsterdam: John Benjamins.

Hussein, K. 2014. 'An experiment in plagiarism detection in academic research articles using attributional techniques', *Research on Humanities and Social Sciences* 4:3.

Iqbal, F., et al. 2008a. 'A novel approach of mining write-prints for authorship attribution in e-mail forensics', *Digital Investigation* 5, 42-51.

Iqbal, F., et al. 2008b. 'Mining writeprints from anonymous e-mails for forensic investigation', *Digital Investigation* 7, 56-64.

Maurer, H., et al. 2006. 'Plagiarism - A Survey', *Journal of Universal Computer Science* 12:8, 1050-1084.

McMenamin, G. 2002. *Forensic Linguistics: Advances in Forensic Stylistics.* Boca Raton, Florida: CRC Press

Pecorari, D. 2010. Academic Writing and Plagiarism: A Linguistic Analysis. London: Continuum International Publishing

Salha, M., et al. 2012. 'Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods', *IEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews* 42:2, 133-149.

Stamatatos, E. 2009. 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology* 60:3, 538-556.

Zheng, R. 2006. ‘A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques', *Journal of the American Society for Information Science and Technology* 57:3, 378-393.