Edited, paraphrased and still unique? Authorship analysis and uniqueness of encoding in undergraduate student essays

Vlad Mackevic

Aston University, UK
mackeviv@aston.ac.uk

Vlad Mackevic is a final year undergraduate student of International Relations and English Language at Aston University, Birmingham. His research interests include forensic linguistics, business English, computer-mediated communication and translation studies. He is also working as a writing mentor and as a freelance public service interpreter.

Abstract:

This paper describes an experiment in authorship analysis of undergraduate student essays, where quantitative methods (specifically ANOVA and T-Test) are employed to determine whether the frequencies of certain function words and hapax legomena can serve as authorship markers that cluster the same authors and discriminate between different ones. The findings show that discrimination success rates are much higher than those of clustering, that the one-sample T-Test is particularly good for discriminating, and that ANOVA, which was used for clustering only, clustered the texts better then T-Test, but ANOVA results would only be applicable in the civil court. Certain function words have been found to serve as very good authorship markers, even under very rigorous conditions, with results of T-Test possibly applicable in the criminal court, but not all markers are that efficient. The problem, however, could be not with the method, but with the data as different parts of the text are structured differently. This paper includes a post-experimental test that demonstrates greater uniformity of data when essay introductions and conclusions are excluded from the data set.

Keywords:

Authorship analysis, quantitative methods, forensic linguistics, ANOVA, T-test, statistics, function words, hapax legomena, authorship markers

Introduction

The concept of forensic linguistics is very recent compared to other strands of the discipline. Coulthard (2004) writes that despite it being 35 years since Svartvik published *The Evans Statement: a Case for Forensic Linguistics*, major developments began only in the mid-1990s. Although the issue of authorship analysis and attribution goes far into the past (for instance, disputed Shakespeare's plays or Biblical texts), it is only recently that these have begun to play a major role in court, and linguists are called to present their evidence as expert witnesses.

The issue of authorship attribution in a forensic context is of great importance: after all, the linguistic expert's evidence must serve justice, and therefore be precise 'beyond reasonable doubt' unless one is in a civil case where it is on the balance of probabilities and one has to raise legitimate doubt when working for the defence (Grant 2010). Although the UK justice system does not currently have any set criteria for evaluating the reliability of the expert witness' analysis except for the reliability of the expert (Grant 2007), this is going to change and something similar to the US system's Daubert Criteria is going to be introduced. The Daubert criteria are the following:

1. The theory must have been tested;
2. It must have been subjected to peer review and publication;
3. It must have a known error rate;
4. It must be generally accepted in the scientific community.

(Chaski 1997; Grant 2007; Tiersma and Solan 2002, cited in Coulthard, 2004)

The Daubert criteria prompted an increase in quantitative (in particular statistical) research involving the analysis of texts with a view to discovering trends that would cluster texts written by the same author and discriminate between different authors. Mathematicians and engineers have entered the field previously dominated by linguists (see, e.g. Burrows 2003; Holmes and Forsyth 1995). Moreover, linguists are shifting from traditional approaches, developing quantitative techniques that will enable them to express 'beyond reasonable doubt' in numbers (see, e.g. Chaski 2001; Chaski 2005; Grant 2007).

It could be said that all ideas behind various methods stem from one: the existence of idiolect. Defined by Chaski (1997) as the idiosyncratic use of dialect, idiolect is a way of speaking (and, consequently, writing) that is unique for each individual. This paper attempts to test a hypothesis that uniqueness in not only encoded in lexical or syntactic choices, (see Coulthard and Johnson, 2007 for Derek Bentley statement), spelling, punctuation and syntax, (see Chaski 2001; Chaski, 2005), but also in the use of function words.

The main aim of this study is to test:

a) The usability of ANOVA and T-Test statistical techniques for discriminating and clustering authors;
b) The presumption that the frequencies with which certain words are used in text are unique to each individual and could serve as authorship markers.

It examines extracts from academic assignments written by L1 English speakers and looks at markers that can be quantified, focussing predominantly on ratios of certain function words and hapax legomena (words occurring in the text only once).

Research Background

The existence of idiolect has been discussed extensively by Coulthard (2004) and Coulthard and Johnson (2007) – for example see their analysis of Robert Brown Appeal or Derek Bentley's statement. Moreover, the idea that 'language patterns and syntactic features are generated beyond an author's conscious control' is also expressed by de Vel (2001, p.3). He also gives function words – the focus of this study - as an example of such features. However, when idiolect must be demonstrated in court, it is very difficult to put forward an argument that the way the text is written is unique and it is highly probable that a certain person is the author; it is even harder to express the likelihood in numbers. Coulthard (2004, p. 476) writes that it is 'a challenge to the academic community to test the error rate and at the same time to fix an acceptable statistical equivalent for 'beyond reasonable doubt'. Chaski (2001, p. 2) adds that it is 'the linguist's responsibility to create theoretically sound hypotheses' and test them.

Rudman (1998) gives the discipline of authorship analysis and attribution a thorough critique, outlining key problems. A particularly acute problem is the lack of a complete theoretical framework, which, he says, leads to lack of consensus on methodology. However, in defence of the practitioners, it must be said that Forensic Linguistics (and particularly the part involving computing and statistical techniques) is a very young discipline. Moreover, attempts have been made to construct theory behind idiolect, and, just like the elaboration of methodology, it is work in progress. For example, Grant (2010) outlines two theoretical frameworks: one based on neuroscience and another one which claims that the author is influenced by the language he/she is exposed to. It could also be said that what Rudman (1998, p.351) calls 'flawed statistical techniques' is, in fact, experimenting and looking for the right technique and methodology. Moreover, Grant (2010) counters Rudman's claims, saying that 'simple detection of consistency and determination of distinctiveness' would be able to help practical authorship analysis more than even a strong theory.

Researchers have made efforts to test statistical techniques for authorship analysis. For example, Chaski's (2001; 2005) and Grant's (2007) research is very labour-intensive. Just as Rudman (1998, p.351) suggests, they treat each study as 'a unique, hard scientific experiment'. First of all, both Grant and Chaski take texts from the same register, arguing that register differences might be mistaken for idiolectal differences. This study takes the same approach, exclusively analysing undergraduate academic essays. Secondly, Chaski (2001) analyses texts in pairs, against each other, applying a set of criteria and later critically examines the results. In her 2005 study, Chaski uses software that she has designed specifically for quantitative analysis of authorship markers. Most importantly, both Grant and Chaski always critically assess their accuracy and error rates, showing what works and what does not. This study will attempt to do the same.

There are two studies that deal with function words in texts and their applicability as authorship markers – those by Holmes and Forsyth (1995) and Burrows (2003). However, they analyse the function words slightly differently from the way this study does. They take the 50 most common function words in texts and look at their frequencies for comparison. This study looks only at a small number of function words due to its dual purpose: to test if the markers will work for this particular study and to comment on the possible universal application of each marker. The second reason led to the fact that only function words which may have a semantic function in the sentence were analysed (e.g. to, of, in, that and as can all play different semantic roles).

Two particularly interesting aspects and also areas of concern in my research were text length and number of texts. As repeatedly pointed out by Coulthard (2004), Coulthard and Johnson (2007) and Chaski (2001), texts that forensic experts usually work with are very few and short: around 100-400 words. In her 2001 study, Chaski used texts that varied in length between 93 and 556 words. However, the most famous study that examined function words (Wallace and Mosteller's Federalist Papers, 1986, cited in Grant and Baker, 2001; also cited in Holmes and Forsyth, 1995) looked as 85 essays, each 900 to 3,500 words in length. The number of texts matters as much as their length does: e.g. Grant (2007) examines 63 texts (3 authors with 21 texts per author). This study, on the other hand, uses 16 texts (four each by four authors), varying between 220 and 1179 words in length (see Table 1 and the Methodology section for details). Therefore, the texts are both short and few. This may make the analysis more problematic as Grant's (2007) analysis breaks down when he reduces the number of texts per author. However, this study attempts to replicate, at least in part, a forensic experiment and research difficulties may be seen as the real-world challenges.

Table 1: Word Count in Texts by all Authors

	Author A	Author B	Author C	Author D
Extract 1	715	1031	220	511
Extract 2	1042	1179	298	318
Extract 3	1011	893	276	334
Extract 4	558	945	322	366

Mcmenamin (2002) talks about stylistic features like font and paragraph formatting. However, despite stylometry being important in forensic settings, for the purposes of this assignment they will be ignored. The reason is the environment in which academic assignments are written: many students keep default settings for paragraphs, font and other features; spellcheckers correct all typing errors. Also, texts have the potential to be copied from one document to another, and every computer may have its own settings, modified by the user.

In the light of the last paragraph, it must be said that Chaski's (2005) and Grant's (2007) studies are particularly relevant to this research because they examine digitally produced texts. Very few features remain when all formatting is removed and we are left with a plain .txt file. Function words are among these features.

Another problem with research in the field is that too much emphasis is placed on statistics as a flawless method of determining authorship. Rudman's (1998) quoted disagreements on methodology are all about statistics. Chaski (2001), Grant and Baker (2001), Coulthard (1998) and Coulthard and Johnson (2007) all denounce one technique or another, or cite someone who does. It must be noted in advance that the same methodology may not work for two different cases for a number of reasons: different genre, different register, insufficient text length, low presence or even absence of certain markers (an issue also encountered during the course of this study). It must not be forgotten that we are dealing with language and its unique structure – unique to every individual. Rudman (1998) claims that the number of style markers is finite, and this is true, but it is also an enormous number, and one must not try to cram all possible cases into one research paper. Time and effort should be devoted to a thorough study of the subject, and one must not forget that these are still pioneering experiments.

Methodology

The Data Set

This study examines authorship markers in undergraduate student essays using quantitative methods – namely Analysis of Variance (ANOVA) and T-Test. I have collected a number of undergraduate assignments from the public domain (http://www.gradeguru.com), a website where students can post their lecture notes and past assignments. The data set comprises four authors with a number of complete essays collected from each one's profile. Essays used for the analysis were written by students of business, politics, history and philosophy. Essays in literature proved to be unreliable during the analysis due to a large number of direct quotes that back up the students' arguments. The corpora for each author are diverse in terms of composition: for Authors A and B four extracts from three different essays have been taken; for Author C, one essay was used for all four extracts, and for Author D – two essays.

The timeframe within which the texts have been produced did not exceed 18 months. It is unknown if idiolect changes over time – a concern also expressed by Coulthard (2004) and de Vel (2001). Therefore, to replicate the context of forensic investigation more authentically, the time frame has to be relatively narrow.

Chaski's (2001) study analyses texts produced by authors matched very closely socio-linguistically: same sex, and of similar age, dialect and educational level. However, due to the data for my study having been taken from the public domain, it was possible to control all variables and achieve such close sociolinguistic matches. Moreover, I am aware that if texts are taken from one socio-linguistically close group, the results might only be representative of that group. For this reason, I have chosen texts written by people from a variety of backgrounds united obviously by one factor: university education.

Due to the data being in the public domain, there were no ethical issues. However, to ensure protection of identity for all the writers, I have removed all names (including place names) from the corpora.

Concerns – texts may have been tampered with

The biggest concern with academic texts is that they may have been edited by a third party. It is unknown who may have corrected the author's mistakes or changed the text's wording etc. Rudman (1998) also expresses his concerns about the possibility of texts being edited by others and thus losing some of their idiolectal properties. Two prominent features of undergraduate academic writing that can distort the analysis of the author's style are paraphrasing and direct quotes. To avoid idiolect confusion, sentences containing long direct quotes were removed from the texts, with a few exceptions where they did not exceed seven words and functioned as clauses.

Paraphrasing, on the other hand, was not considered a hindrance. Keck's (2006) research shows that although students (particularly first years) tend to copy substantial pieces of the original text, the rate of what Keck labels as 'Near Copies' among English L1 students is below 5% in 250-300 word texts. Minimal and moderate revisions constitute about 15% each, and substantial revisions amount to around 11%. Moreover, it has been found that, students at more advanced level tend to leave the same content words but change the sentence's structure when paraphrasing. Therefore, it could be argued that the function words that they use in syntactic structures are typical of their idiolect.

Taking into account the length of the texts and the percentages, it can be presumed that paraphrasing and self-editing do not make a significant impact on the author's idiolect in the text. Moreover, the sentences that Keck (2006) provides as examples of paraphrasing clearly indicate that students make conscious lexical choices and retain the features of their idiolects while citing other authors.

Data Analysis

The data analysis was carried out in five steps:

1. Every extract was imported into TextStat concordance software (http://neon.niederlandistik.fu-berlin.de/en/textstat/), and word frequencies were exported into Microsoft Excel tables. After this, certain function words that occurred more than 3 times in every extract were highlighted and their relative frequencies (Word/Token ratio in text) calculated. Words that did not occur in all extracts or occurred less than three times in any text were disregarded.

2. Relative frequencies of potential authorship markers (ratios of function words as well as Lexical Richness (LR), which is the Hapax/Token ratio) were exported into separate tables for each author (see Appendix 2 for the tables). The ratios for the tables were analysed and Mean and Standard Deviation (StDev) were calculated for each marker.

3. For the ANOVA analyses, for each author, the Marker/Token ratios were grouped into four combinations; extracts 1, 2, and 3 became combination 123, etc. (see Table 1 below for an example). Tables like that were created for each author.

Table 2: Authorship markers and extract combinations

	123		124		134		234	
Marker	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
TO	3.8830%	0.2789%	4.2490%	0.3556%	4.0920%	0.5499%	4.0840%	0.5509%
OF	3.7400%	0.1775%	3.8070%	0.0946%	3.7470%	0.1781%	3.6890%	0.1113%
IN	2.9000%	0.2547%	2.6220%	0.2762%	2.7170%	0.4208%	2.7910%	0.4243%
THAT	2.3500%	0.1614%	2.1620%	0.4790%	2.0560%	0.3964%	2.0950%	0.4466%
AS	1.2840%	0.4440%	1.6320%	0.7500%	1.4850%	0.7652%	1.7810%	0.5241%
THIS	1.4070%	0.5265%	1.4030%	0.5305%	1.0940%	0.0226%	1.3930%	0.5390%
TTR	41.3230%	4.0559%	43.4770%	5.2262%	44.5860%	3.3466%	41.8950%	4.9769%
LR	27.1150%	2.5943%	28.8160%	4.2848%	29.9540%	2.5997%	28.2570%	4.2854%

Means and Standard Deviations of values in those groups were calculated and the extract combinations were compared in pairs (see Appendix 2 for tables), applying the ANOVA statistical technique to carry out Intra-Author Analysis, and find out the likelihood of these extract combinations belonging to the same author. An ANOVA calculator available online was used for analysis (http://www.danielsoper.com/statcalc/calc43.aspx).

4. After ANOVA, all authors underwent T-Test analysis, where I took three extracts as K (known) texts and compared them to a fourth one, pretending that it was a Q (Query) Text. Thus, the analysis included four ‘make believe’ Q Texts for each author for each marker, resulting in 20 to 28 tests overall. One sample T-test calculator, available online was used (http://www.graphpad.com/quickcalcs/OneSampleT1.cfm)

5. After intra-author analysis, every authorship marker was subject to inter-author analysis using T-Test. This was done in two steps.

i) Using a T-test function in Excel, one by one, authors were compared to one another using all four extracts. The results can be found in Appendix 3.
ii) One sample T-Test was carried out for differentiation for each marker using four extract by one author as K texts and one extract by another author as a Q Text. The results in tables and graphs can be found in Appendix 3. Success and error rates were calculated for each test and expressed as a ratio (number of successful tests/number of all tests) and as a percentage.

Analysis and Discussion

Data analysis was carried out in several steps. First of all, as part of intra-author analysis, each individual marker was reviewed and evaluated for consistency. Secondly, ANOVA and T-Test were used to cluster and differentiate between authors. The results are presented in Appendices 2 and 3, and discussed below.

Since it is difficult to express what 'beyond reasonable doubt' means in numbers, several criteria have been established. These are listed in a Table 3.

Table 3: Criteria for probability of same authorship (PSA)

Criterion	**Meaning/explanation**
A ratio:[1] Number of test results where Probability of Same Authorship is above 50% All test results	50% certainty is a civil court benchmark
A ratio: Number of test results where Probability of Same Authorship is above 80% All test results	80% is my subjective 'success mark' (Very likely that the text have been written by the same author)
A ratio: Number of test results where Probability of Same Authorship is below 40% All test results	40% is my subjective 'failure mark' (**Unlikely** that the texts have been written by the same author)
A ratio:[2] Number of test results where Probability of Same Authorship is below 20% All test results	**Very unlikely** that the texts have been written by the same author
A ratio: Number of test results where Probability of Same Authorship is below 10% All test results	**Extremely unlikely** that the texts have been written by the same author
A ratio: Number of test results where Probability of Same Authorship is below 5% All test results	'Beyond Reasonable Doubt'

[1] All ratios are expressed as percentages.

[2] This and the following criteria were used only for discriminating between authors, not clustering.

Intra-Author Analysis

Reliability of Authorship Markers based on their frequencies

Tables in Appendix 1 show that it is very difficult to call any of the identified authorship markers very reliable – the standard deviations for the majority of the markers are far too large. However, it must be noted that, in comparison with the traditional computational authorship analysis studies (e.g. Burrows, 2003, or Wallace and Mosteller, cited in Holmes and Forsyth, 1995), the texts are very short and few. This might account for within-author variation in my data. Moreover, it can be seen that StDev for some markers is very low (e.g. LR in authors C and D; OF in author A) and for 14/26 (53.85%) markers across all authors StDev does not exceed 20% of the mean.

Some markers that occurred in all extracts by one author but not in all authors (e.g. IN for author A; IT for authors B and D) showed a varying degree of consistency. StDev for marker IT in author B was almost 50%, whereas the StDev for the other two was relatively small.

One problem repeated itself several times throughout the analysis. Alongside the three relatively even percentages, there was one which was significantly higher or lower (e.g. Author A – THAT, THIS; Author B – TO, OF, THAT; Author C – AS, FOR; Author D – THAT). A hypothesis was formed that this ratio belonged either to an introduction or a conclusion of the essay. Introductions and conclusions are often structured differently from the main body, and it would be worth testing whether the data becomes more uniform without them. This hypothesis has been tested and its results are described below under the heading Post-experimental Test and in Appendix 3.

Markers that do not occur in all authors

The absence of certain markers is just as important as their presence. For example, authors A and B use THIS fairly consistently, which cannot be said about authors C and D. Authors B and D use IT; Author A uses IN; Although in some cases the ratios are not very uniform throughout all extracts, there is still a degree of consistency. It could be presumed that different authors prefer to use different words, and this feature (presence or absence of certain words) could be considered idiolectal.

Lexical Richness (LR)

Lexical richness is a particularly significant marker. Not only its StDev is comparatively small for all authors, but it is also very distinctive for each individual author. Unlike Type/Token Ratio (TTR), which can vary greatly depending on text length (Grant and Baker, 2001) and 'loses its utility... within the confines of forensic investigation with its typically short documents' (Chaski, 2001: 3), LR has proven to be very consistent regardless of the text's length. The chart below shows LR for all extracts by all authors.

As it can be seen, although some extracts have very similar LR, the lines never cross each other. The graph indicates relative consistency within one author and disparity between different authors.

Analysis of Variance (ANOVA)

The results of ANOVA analysis are quite convincing, but only for the civil court. The highest number of successful attribution tests, where the Probability of Same Authorship (PSA) rises above 80% (my subjective success mark) is 33.33% (author C), which is hardly 'beyond reasonable doubt'. On the other hand, if we consider this experiment a civil case, then even the lowest percentage of successful attribution tests that rises above the 50% mark (63.89% for author D) is enough.

However, such fairly optimistic results could be explained by the fact that there was no query text involved in ANOVA analysis. Moreover, I used groups of three extracts for each test, and, as Table 1 shows (see page 12), these groups shared two texts out of three.

T-Test

The results of the T-Test analysis differ from ANOVA significantly. First of all, the percentage of successful tests is notably lower (the highest percentage of attributions above the success line is 25%, while the lowest is 5%). Moreover, only between 20% and 41.67% of test results rise above the civil court benchmark (50%). These results would not suffice even for the civil court.

Another problem with T-Test is that it is very inconsistent with expressing PSA. For the same marker, in text of the same author, the 'One Extract Against Three' test returns two very low probabilities (often below 10%) and two very high ones (70% and higher). (e.g. Author A – TO, OF, LR; Author B – OF, Author C – AS).
On the other hand, the results become more optimistic in the light of the Inter-Author Analysis, the results of which are provided in Appendix 3.

Inter-Author Analysis Using T-Test

All authors against each other in pairs

When all authors were compared against each other using the T-Test, it turned out that the differentiation between authors was much more successful than clustering the same authors. Combined misattribution for all markers for all authors (i.e. PSA rising above 50%) is only 20% of all tests, and definite failure to differentiate only occurs in 13% of cases (See Appendix 3). Compared with the percentage of tests that failed to cluster the authors (over 50% of unsuccessful tests for authors A and B, and 75% for both authors C and D), it could be called good performance.

Once again, Inter-Author analysis has proven that LR is a good authorship marker, with a very low PSA, as it can be seen from Tables 5 and 6 in Appendix 3. Marker TO also worked well; THAT, on the other hand, is particularly problematic as it fails to differentiate between authors A-B, A-C and B-C to such extent that almost clusters them, Probability of Same Authorship shooting above 90%. The other markers' performance was average and it is not certain which factors influenced it.

One author at a time against extracts by different authors

Overall, differentiation could be called very successful. The ratio of unsuccessful tests (PSA over 50%) is relatively low for all authors, not exceeding 22.33% (author C). The ratio of absolute failures (PSA over 80%) is even lower, 6.67% being the highest.

Analysis of reliability for each individual marker is provided in the table below:

Table 4: Marker analysis

Marker	Comments
TO	Very reliable only when PSA ≥ 40%. As soon as PSA bar drops, error rate shoots up; received 100% error rate under the strict conditions (PSA ≥ 10%) once.
OF	Moderately reliable. Performs well until the point when PSA reaches ≥ 20% mark. After that, error rates rise, reaching 91.67% once. However, for Authors A and B it performed very well, error rates not rising above 25%.
AS	Not very reliable. Error rates reach 25% when PSA ≥ 40% and rise to 33%-50% at PSA ≥ 20 stage. However, performed very well in tests for Author D.
THAT	Very unreliable across all authors – enormous error rates; PSA shooting over 50% most of the time. On the basis of this analysis, probably should be disregarded altogether.
LR	Excellent marker; very consistent – error rate rose to 25% once, under the strictest conditions (PSA ≥ 5%). Most of the time error rate was 8.33% (1/12).

However, as it was mentioned above, the problem may be not with the markers or the method, but with the data (e.g. the length of the text and the number of texts). Grant (2010) writes that possibly ‘a substantial and varied body of text would be required before manifest idiolectal features became noticeable or measureable’. Chaski (2001) also expresses concerns about the optimal amount of data. Rudman adds: ‘Authorship studies must not fall into the trap of discarding style markers ... because they didn’t work in some other study.’ (1998, p. 361).

For example, in this study, Author C has the smallest word count in the extracts, and the smallest number of markers. This could be the reason behind T-Test for this author’s texts returning the largest number of failed tests for both inter and intra-author analyses. This presumption is backed up by Grant and Baker (2001), who say that a small sample size may lead to incorrect results.

In Grant's (2007) study, when the data sample is reduced to five texts per author, the analysis started breaking down. I examined four authors with four texts per author, and got comparatively significant results (albeit only for the civil court). Therefore, this study should be treated as a pilot and more research should be done on the subject in order to determine whether ratios of function words and hapax legomena within the text can serve as authorship markers.

A suitable number of authorship markers should also be examined in greater detail. Holmes and Forsyth (1995) suggest that perhaps the number of potential markers should be ≥50. They also mention that Mosteller and Wallace used 30 markers. Burrows (2003) uses 50 function words for his analysis. I have used five that occurred in all authors for inter-author analysis and five to six that occurred in all texts by the same author for intra-author analysis. There are two reasons for such a small number. First of all, my texts are shorter and the number of times a particular word occurs will be lower than in the aforementioned studies. Due to this, it is hard to find many words with significant ratios that occur in all texts by all authors. Secondly, Mosteller and Wallace (cited in Holmes and Forsyth, 1995), Holmes and Forsyth (1995) and Burrows (2003) all used different techniques from the ones I used. Therefore, it is natural that the data set will be different as well. Moreover, one of the purposes of this study is to find salient markers and test their applicability. Due to the time constraints, it was not possible to test all function words in the text, and it is also unfeasible if they occur too few times or do not occur in all texts.

Another issue with the texts might be not with their length but with the socio-linguistic variables of their authors. Grant and Baker (2001) write that 'where two authors, A and B, are socio-linguistically close, distinguishing between them will be a difficult test for any marker'. From what the results show, it can be said that the authors might have shared some socio-linguistic variables, or might have been exposed to similar language as some point in time.

Post-experiment test

In order to test the hypothesis that introductions and conclusions in essays are structured differently from the main body, I took two essays by different authors and split them, first into two, and then into three extracts. The lengths of the extracts are provided in the table below.

Table 5: Token Count for Post-Experiment Test Corpora

Author E	**ES1**	**ES2**	**ES3**		**EL1**	**EL2**
Word Count	590	611	599		870	911
Author F	**FS1**	**FS2**	**FS3**		**FL1**	**FL2**
Word Count	518	554	596		921	747

Abbreviation meanings: E, F – author code; S – short texts, L – long texts.

The analysis was carried out only to see how consistent the markers will be within each of the authors. The results are provided in Appendix 4. As it can be seen from the tables, the markers are much more consistent than tables in Appendix 2 show. The largest standard deviation for the seemingly unreliable marker THAT is only 13.42%, in comparison with nearly 34% in the main analysis. Other markers, like TO, OF and IT were also very consistent. The test has also shown that the length of the text does affect the intra-author consistency slightly (particularly in the case of LR) but the effect does not seem significant enough to cause idiolect confusion.

The data was subject to inter-author analysis using T-Test. The table below provides the results of the analysis, outlining the Probability of Same authorship for each marker that the authors shared.

Table 6: T-test results for authors E and F (inter-author analysis)

Marker	PSA (paired T-Test)	PSA (unequal variance)
TO	10.7677%	5.7237%
OF	1.8409%	0.6160%
IN	**46.8041%**	**39.7300%**
IT	10.3521%	6.1325%
THAT	16.0625%	9.9243%
THEIR	**20.9488%**	11.7665%
LR	0.0385%	0.1177%

NB: Values in bold are not reliable and were calculated from markers with high StDev

As it can be seen, the PSA is low in most cases – T-Test works well for differentiation. The results shown in bold are not reliable, but the markers they represent had very high StDev in both authors. Therefore, it could be presumed that within-author consistency is an important factor that may influence inter-author differences.

Overall, intra-author variation decreased, whilst inter-author differences remained rather significant (i.e. differences between means for the same marker used by authors E and F). However, there are certain limitations: due to the time constraints I was not able to carry out a full-scale T-Test analysis and my results may not be as conclusive as I would like them to be.

Conclusion

This paper has examined authorship markers in undergraduate students' assignments. The analysis has demonstrated that the same marker may have different consistency within each author. This, however, could be attributed to intra-author variation. Moreover, as the post-experiment tests have shown, introductions and conclusion can indeed make the data imperfect, as these parts of the essay are written in a different style from the main body. Marker frequencies were more consistent when only main bodies of the essays were used.

Therefore, somewhat misleading results could stem not from the markers or the methods, but from the noisy data.

Lexical Richness was found an excellent marker with a very low StDev, but during the Post-experiment test it also seemed to change slightly depending on the text's length. However, these changes were not significant, as StDev remained relatively low. Other markers performed with a varying degree of success. Importantly, the Post-experiment test has shown that introductions and conclusions do indeed make the data more confusing, and the markers that performed badly in the main analysis improved their within-author consistency. It can be presumed that it happens because these parts of an essay are structured differently from the main body.

Moreover, T-Test (both normal and one-sample) was found to be particularly useful for discriminating between different authors, but returned very poor results for clustering. ANOVA, on the other hand, has proven to be good at clustering, but its results would only suffice in civil court cases. On the other hand, in defence of T-Test as a method, it must be said that although it failed to cluster quite often, failure to discriminate was much lower. Taking into account the overall performance of the tests, it could be said that they worked well and could be used, if not to say 's/he wrote it', then at least to say 's/he didn't write it'.

The post-experimental test has shown that introductions and conclusions do indeed make the data more confusing. It can be presumed that it happens because these parts of an essay are structured differently from the main body. All authorship markers performed surprisingly better in the post-experiment test, and their within-author consistency may have led to better discrimination rates.

References

BURROWS, J. (2003) Questions of Authorship: Attribution and Beyond. *Computers and Humanities*, 37, 5-23.

CHASKI, C. E. (1997) Who Wrote It? Steps Towards a Science of Authorship Identification. *National Institute of Justice Journal*. (September Issue). [Online]. Available from: http://www.ncjrs.gov/pdffiles/jr000233.pdf [Accessed 31 January 2010].

CHASKI, C. E. (2001) Empirical evaluations of language-based author identification techniques. *International Journal of Speech, Language and the Law*, 8 (1), 1-65.

CHASKI, C. E. (2005) Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence* [Online] 4 (1), pp. 1-14. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.3852&rep=rep1&type=pdf [Accessed 31 January 2010].

COULTHARD, M. (1998) Identifying the Author. *Cahiers de Linguistique Française* [Online] 20, pp. 139-161. Available at: http://clf.unige.ch/display.php?idFichier=168 [Accessed 28 January 2010].

COULTHARD, M. (2004) Author Identification, Idiolect and Linguistic Uniqueness. *Applied Linguistics*, 25 (4), 431-447.

COULTHARD, M. and A. JOHNSON (2007). *An Introduction to Forensic Linguistics: Language in Evidence.* Abingdon: Routledge.

DE VEL, O. (2001) Multi-Topic E-mail Authorship Attribution Forensics. In*: ACM Conference on Computer Security – Workshop on data mining for security applications.* November 8, 2001.Phildelphia, PA [Online]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.9951&rep=rep1&type=pdf [Accessed 31 August 2010].

GARCIA, A. M. and J. C. MARTIN (2007) Function Words in Authorship Attribution Studies. *Literary and Linguistic Computing*, 22 (1), 49-66.

GRANT, T. (2007) Quantifying Evidence in Forensic Authorship Analysis. *International Journal of Speech, Language and the Law*, 14 (1), 1-25.

GRANT, T. D. (2010). Txt 4n6: idiolect free authorship analysis? In: *Routledge Handbook of Forensic Linguistics*. Abingdon: Routledge.

GRANT, T. and K. BAKER 2001. Identifying reliable, valid markers of authorship: a response to Chaski. *International Journal of Speech, Language and the Law*, 8 (1), 66-79.

HOLMES, D. I. and R. S. FORSYTH (1995) The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10 (2), 111-127.

KECK, C. (2006) The use of paraphrase in summary writing: A comparison of L1 and L2 writers*. Journal of Second Language Writing*, 15 (4), 261-278.

RUDMAN, J. (1998) The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, 31, 351–365.

Websites

Textstat	http://neon.niederlandistik.fu-berlin.de/textstat/
ANOVA Calculator	http://www.danielsoper.com/statcalc/calc43.aspx
T-test Calculator	http://www.graphpad.com/quickcalcs/OneSampleT1.cfm
Grade Guru	www.gradeguru.com

Appendix 1: Salient potential authorship markers and their consistency within each author.

Table 1: reliability of markers for Author A

	Extract 1	Extract 2	Extract 3	Extract 4	Mean	StDev	StDev/Mean
Marker							
TO	4.0559%	4.0307%	3.5608%	4.6595%	4.0767%	0.4503%	**11.0461%**
OF	3.9161%	3.7428%	3.5608%	3.7634%	3.7458%	0.1455%	**3.8854%**
IN	2.6573%	2.8791%	3.1652%	2.3297%	2.7578%	0.3531%	**12.8039%**
THAT	2.3776%	2.4952%	2.1761%	1.6129%	2.1655%	0.3912%	**18.0666%**
AS	0.8392%	1.7274%	1.2859%	2.3297%	1.5456%	0.6362%	41.1643%
THIS	1.1189%	2.0154%	1.0880%	1.0753%	1.3244%	0.4610%	34.8105%
LR	29.3706%	24.2802%	27.6954%	32.7957%	28.5355%	3.5431%	**12.4164%**

Table 2: reliability of markers for Author B

	Extract 1	Extract 2	Extract 3	Extract 4	Mean	StDev	StDev/Mean
Marker							
TO	2.3278%	3.8168%	3.4714%	3.1746%	3.1977%	0.6365%	**19.9057%**
OF	2.2308%	2.9686%	1.5677%	2.1164%	2.2209%	0.5764%	**25.9532%**
AS	1.3579%	0.3393%	0.8000%	0.6349%	0.7830%	0.4280%	54.6635%
THAT	3.2008%	1.4419%	1.9037%	2.2222%	2.1922%	0.7448%	33.9776%
IT	1.8429%	1.2723%	0.4479%	1.0582%	1.1553%	0.5763%	49.8798%
THIS	1.0669%	1.1026%	1.2318%	1.4815%	1.2207%	0.1877%	**15.3799%**
LR	19.3986%	21.5437%	26.9877%	20.2116%	22.0354%	3.4179%	**15.5109%**

Table 3: reliability of markers for Author C

	Extract 1	Extract 2	Extract 3	Extract 4	Mean	St Dev	StDev/Mean
Marker							
OF	3.6364%	3.6913%	3.9855%	2.7950%	3.5271%	0.5115%	**14.5032%**
TO	2.7273%	2.0134%	4.3478%	4.3478%	3.3591%	1.1783%	**35.0780%**
AS	2.2727%	3.3557%	0.7246%	1.8634%	2.0541%	1.0872%	52.9273%
THAT	1.8182%	2.6846%	2.1739%	1.8634%	2.1350%	0.3990%	**18.6902%**
FOR	0.9091%	1.0067%	0.3623%	1.2422%	0.8801%	0.3724%	42.3174%
LR	35.0000%	36.2416%	39.8551%	38.8199%	37.4792%	2.2450%	**5.9901%**

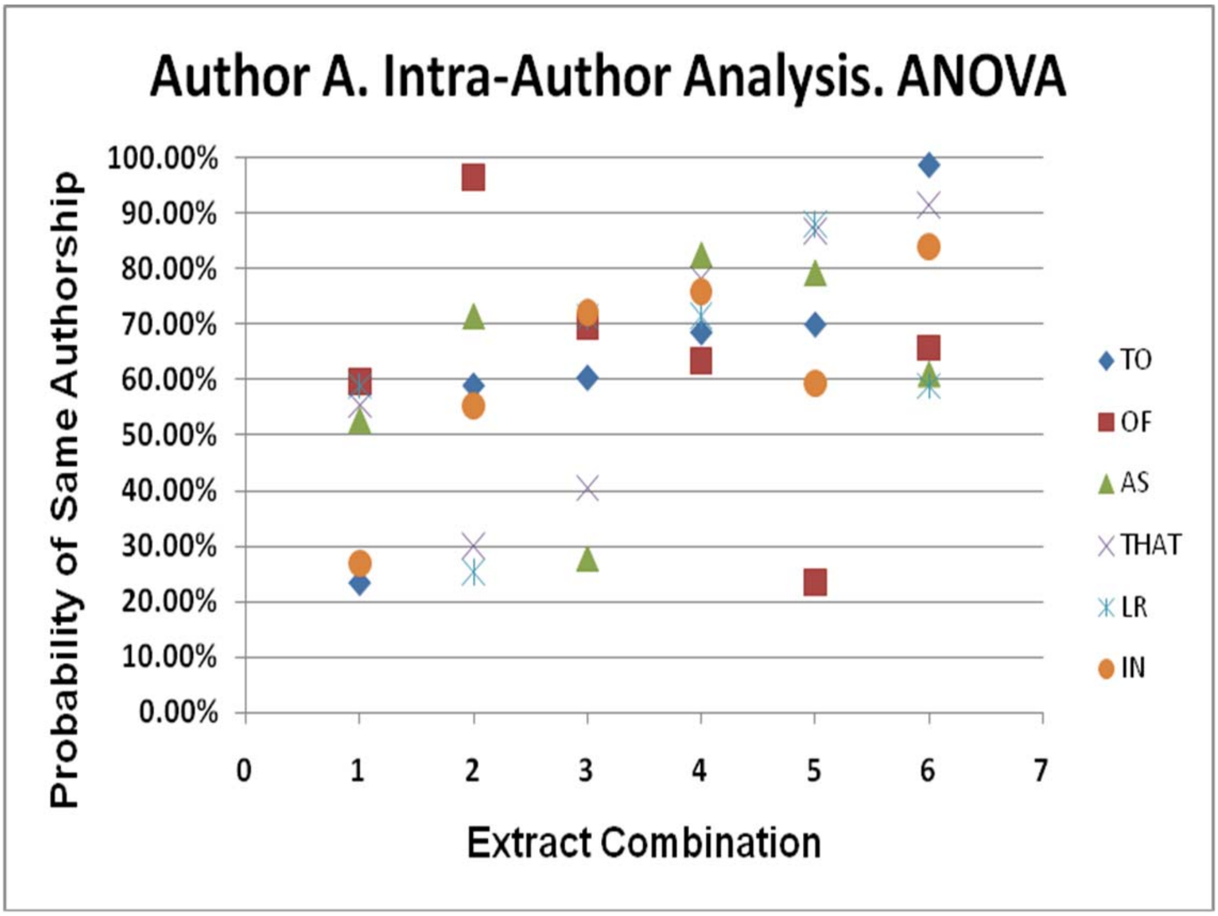
Table 4: reliability of markers for Author D

	Extract 1	Extract 2	Extract 3	Extract 4	Mean	StDev	StDev/Mean
Marker							
OF	3.5225%	2.8302%	4.1916%	2.4590%	3.2508%	0.7665%	**23.5801%**

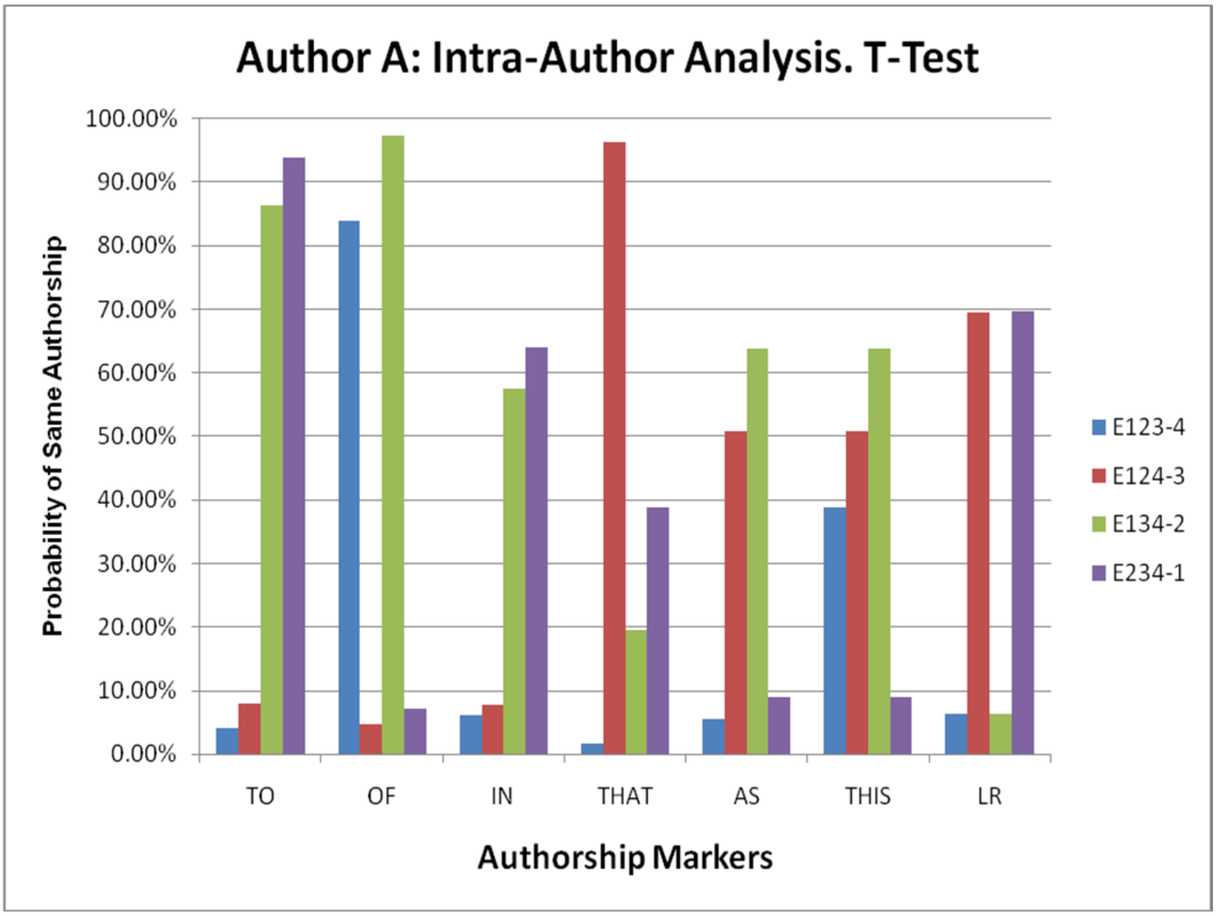
TO	3.9139%	3.7736%	3.5928%	3.0055%	3.5714%	0.3996%	**11.1874%**
THAT	2.3483%	3.7736%	1.7964%	3.5519%	2.8676%	0.9498%	33.1211%
AS	1.1742%	0.9434%	1.7964%	1.6393%	1.3883%	0.3972%	28.6118%
IT	1.1742%	0.9434%	0.8982%	0.8197%	0.9589%	0.1524%	**15.8907%**
LR	38.5519%	41.8239%	40.7186%	39.6175%	40.1780%	1.4095%	**3.5080%**

NB: During intra-author analysis, ALL authorship markers that occurred in ALL texts by the same author were analysed (e.g. *in* for Author A, or *it* for Authors B and D). However, during inter-author analysis, only markers that occurred in *all* authors' texts were analysed (*TO*, *OF*, *AS*, *THAT* and *LR*). Also, some markers that did not occur in all authors' texts and had large standard deviation were excluded even from the intra-author analysis.

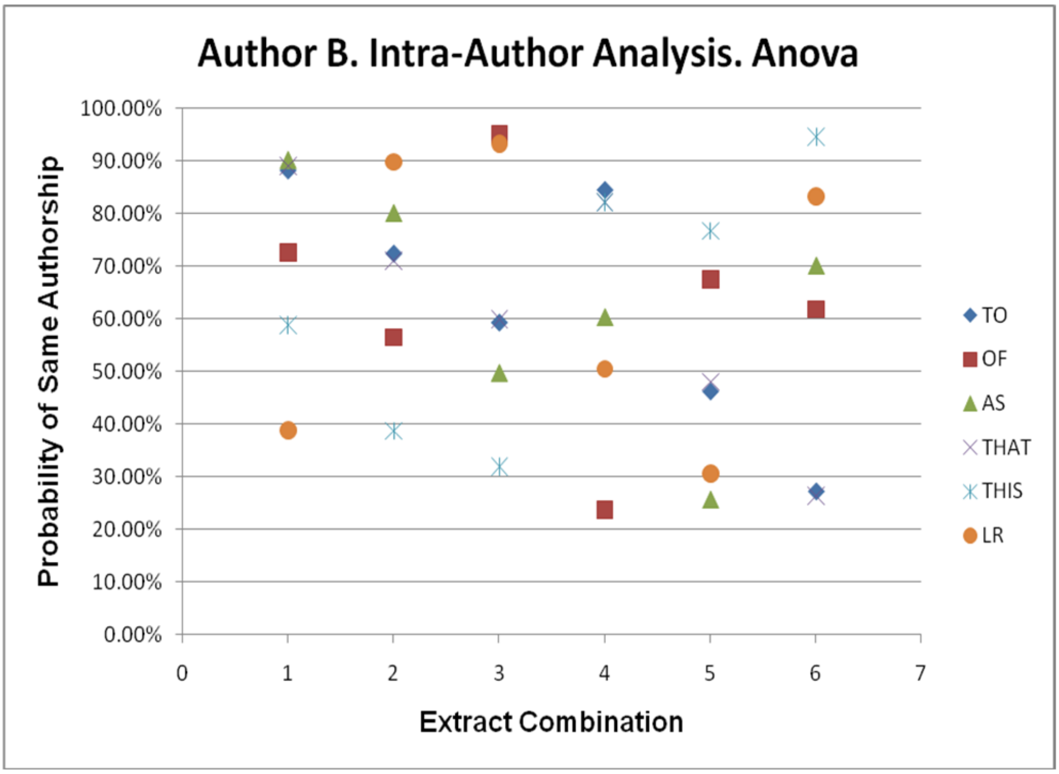
1. ANOVA and T-test Analysis



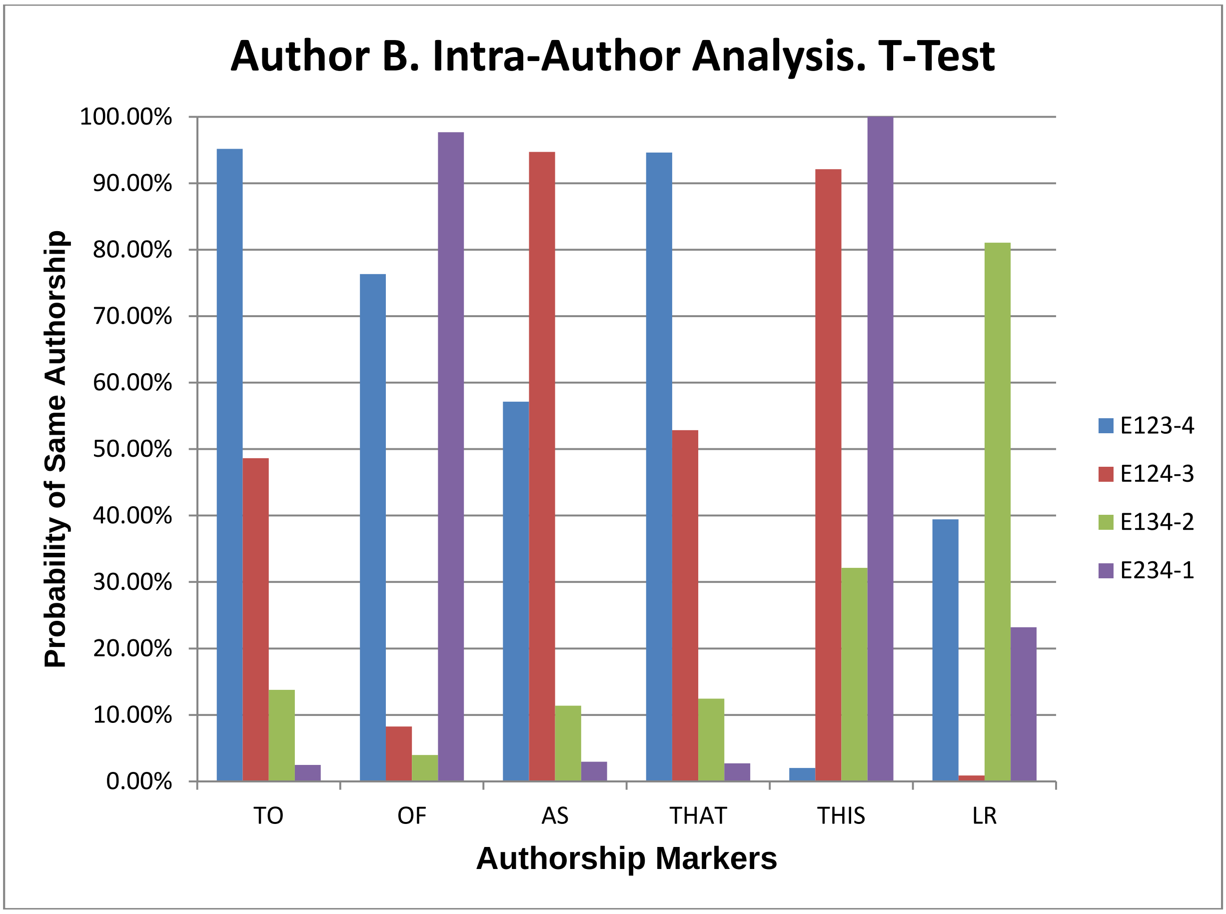
	Ratio	Percentage
Below Fail Mark (40%)	**7/36**	**19.44%**
Above Success Mark (80%)	**7/36**	**19.44%**
Above Civil Court Benchmark (50%)	**26/36**	**72.22%**



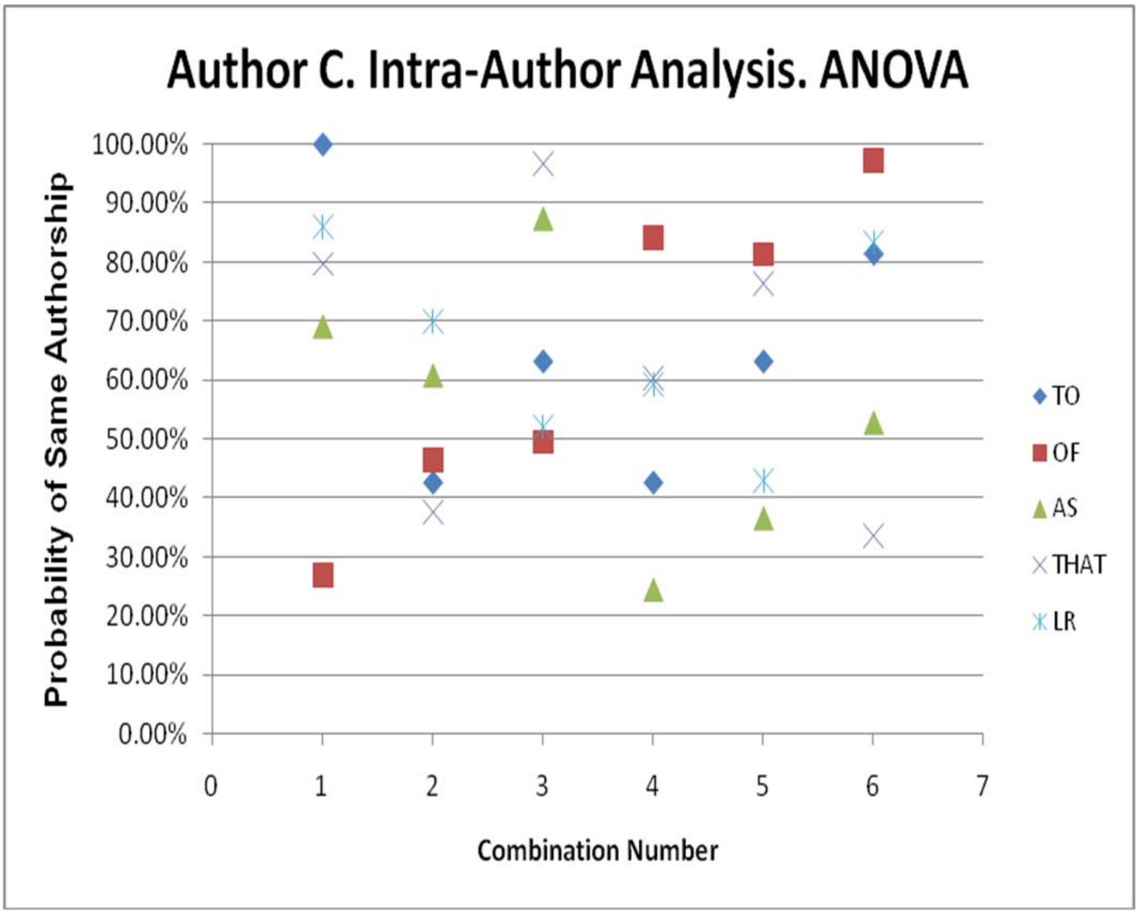
	Ratio	Percentage
Below Fail Mark (40%)	**15/28**	**53.57%**
Above Success Mark (80%)	**5/28**	**17.86%**
Above Civil Court Benchmark (50%)	**11/28**	**39.29%**



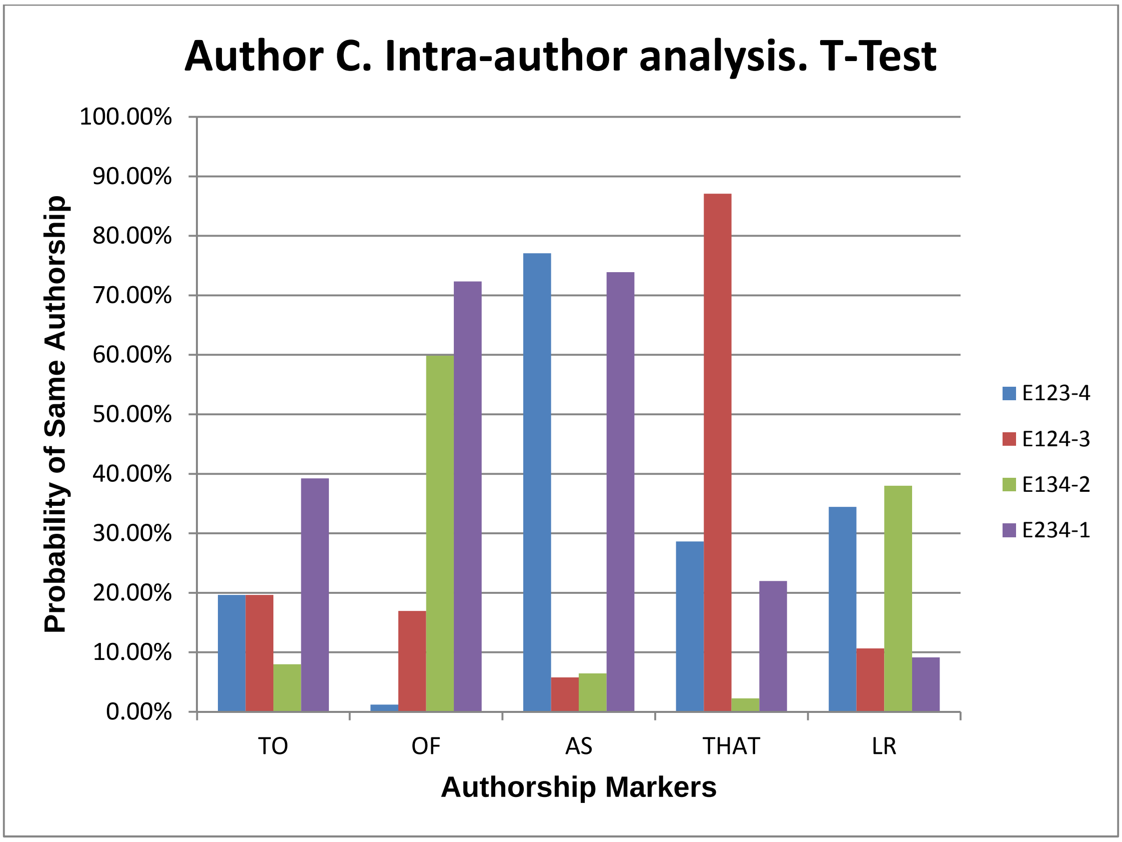
	Ratio	Percentage
Below Fail Mark (40%)	**8/36**	**22.22%**
Above Success Mark (80%)	**11/36**	**30.56%**
Above Civil Court Benchmark (50%)	**24/36**	**66.67%**



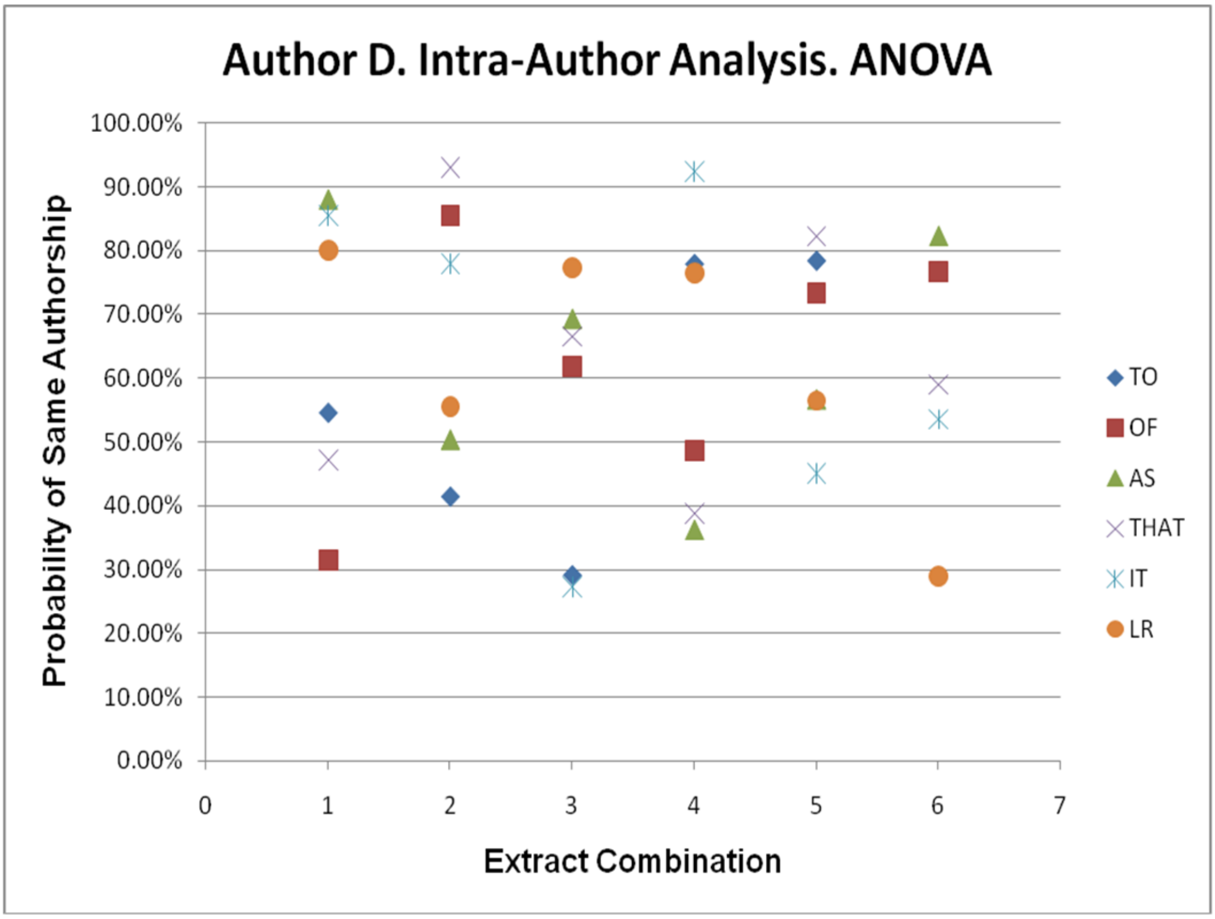
	Ratio	Percentage
Below Fail Mark (40%)	**13/24**	**54.17%**
Above Success Mark (80%)	**6/24**	**25%**
Above Civil Court Benchmark (50%)	**10/24**	**41.67%**



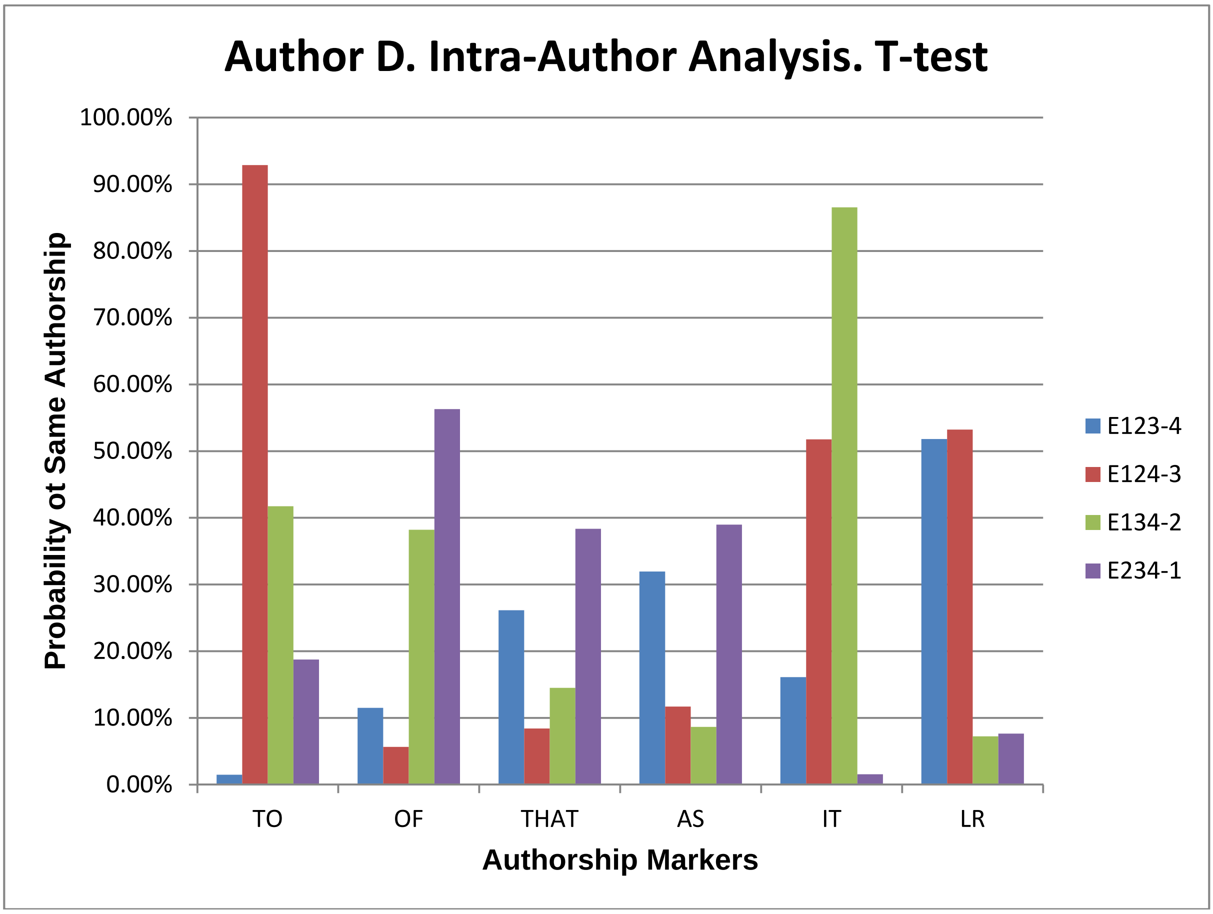
	Ratio	Percentage
Below Fail Mark (40%)	**5/30**	**16.67%**
Above Success Mark (80%)	**10/30**	**33.33%**
Above Civil Court Benchmark (50%)	**20/30**	**66.67%**



	Ratio	**Percentage**
Below Fail Mark (40%)	**15/20**	**75%**
Above Success Mark (80%)	**1/20**	**5%**
Above Civil Court Benchmark (50%)	**4/20**	**20%**

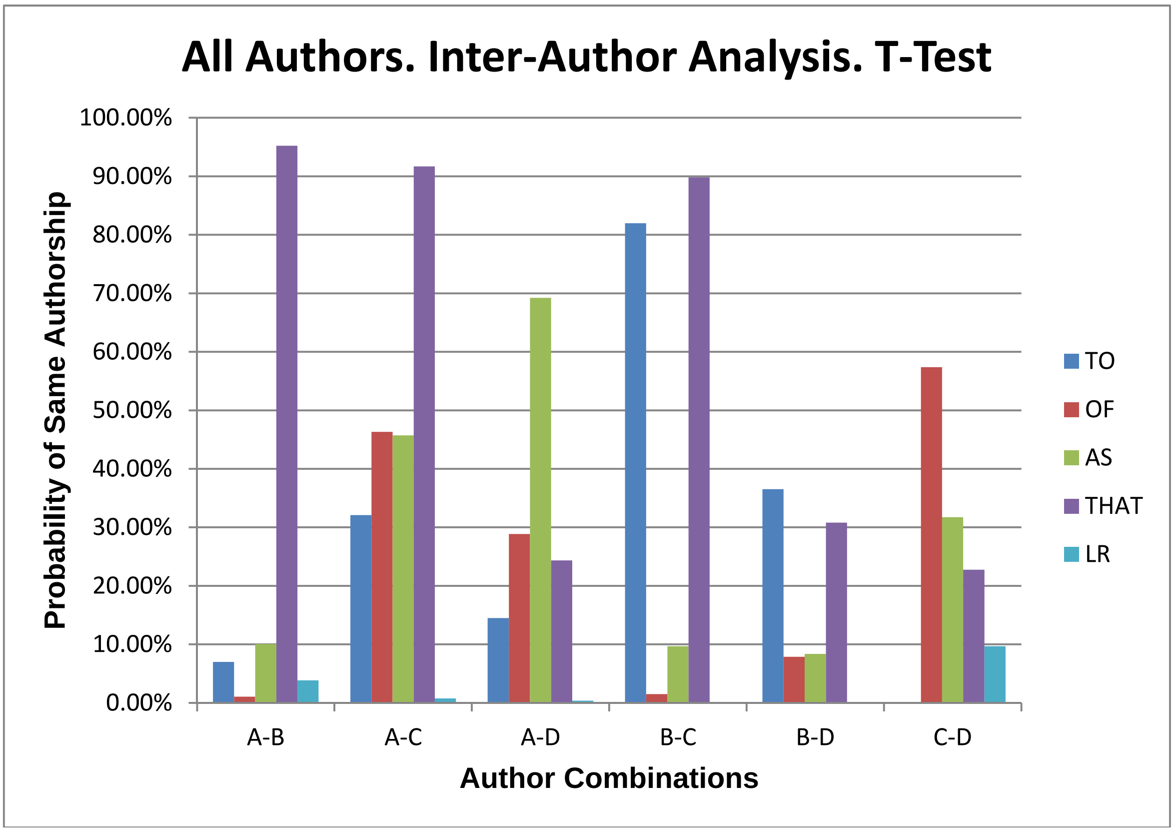


	Ratio	Percentage
Below Fail Mark (40%)	**6/36**	**16.67%**
Above Success Mark (80%)	**7/36**	**19.44%**
Above Civil Court Benchmark (50%)	**23/36**	**63.89%**



	Ratio	Percentage
Below Fail Mark (40%)	**18/24**	**75%**
Above Success Mark (80%)	**2/24**	**8.33%**
Above Civil Court Benchmark (50%)	**5/24**	**20.83%**

Appendix 2: Inter-Author Analysis.



	Ratio	Percentage
Overall successful differentiation (below 40%)	**22/30**	**73.33%**
Successful Differentiation below 30%	**13/30**	**43.33%**
Failed to differentiate (above 50%)	**6/30**	**20%**
Definite Failure (over 80%)	**4/30**	**13.33%**

Table 5: Differentiation between authors. The numbers express the likelihood of SAME authorship.

	TO	OF	AS	THAT	LR
A-B	6.9878%	1.0634%	10.0638%	95.2091%	**3.8550%**
A-C	32.0873%	46.3343%	45.7247%	91.6843%	**0.7721%**
A-D	14.4933%	28.8591%	69.2335%	24.3582%	**0.3866%**
B-C	81.9842%	1.5000%	9.6799%	89.8173%	**0.0550%**
B-D	36.5119%	7.8856%	8.3777%	30.8185%	**0.0611%**
C-D	0.0000%	57.3882%	31.7404%	22.7606%	**9.6809%**

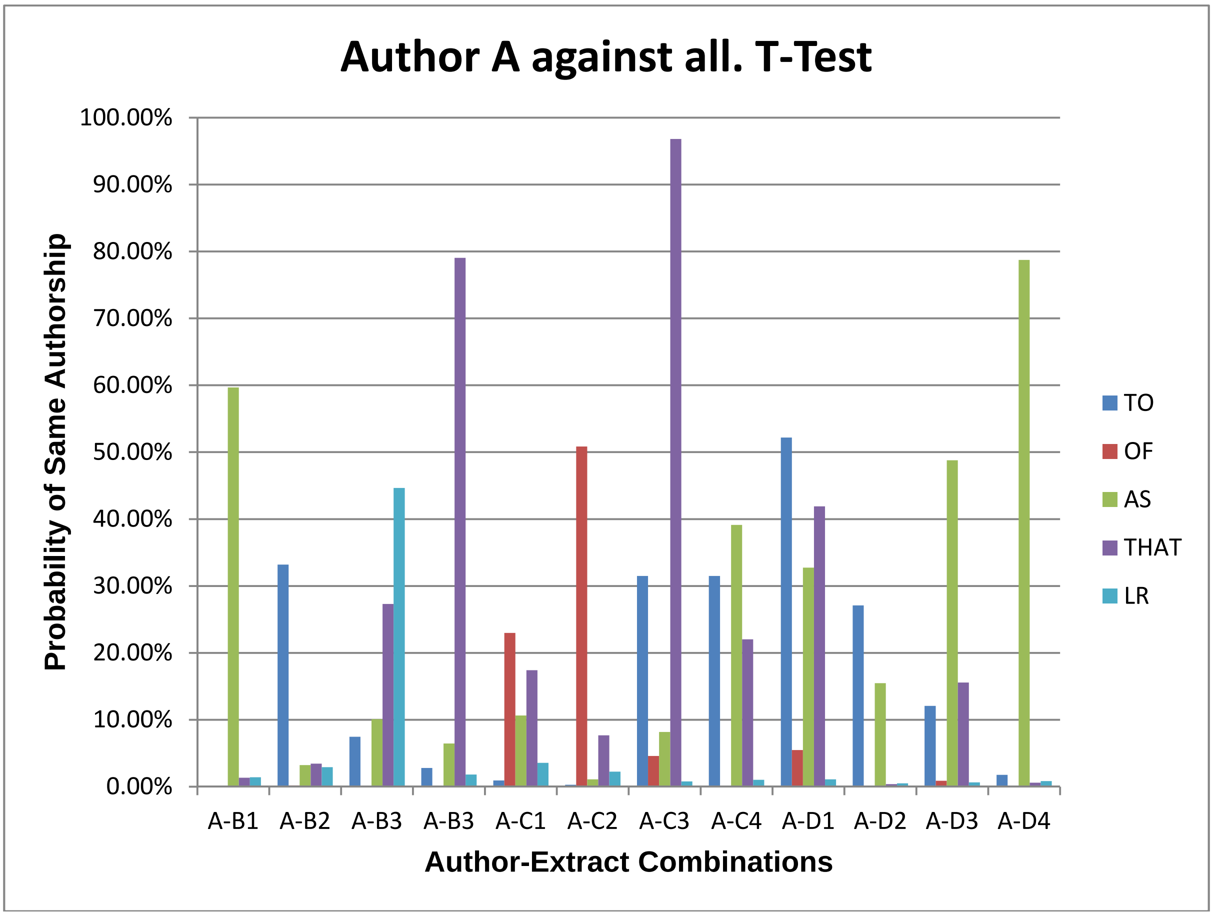
Table 6: Error rates for each marker:

TO	OF	AS	THAT	LR
16.67%	33.33%	33.33%	50%*	0%

Breakdown of Analysis:

T-Test results for inter-author analysis (see previous page) reveal several interesting trends. As the Tables show, some markers have very large standard deviations (e.g. *as* or *that*). Marker *that* is particularly problematic, as it fails to differentiate between authors A-B, A-C and B-C to such extent that almost clusters them, probability of same authorship shooting above 90%. In one case out of six, marker *to* also does that, but the overall average differentiation rate is quite low (in one case it is equal to zero).

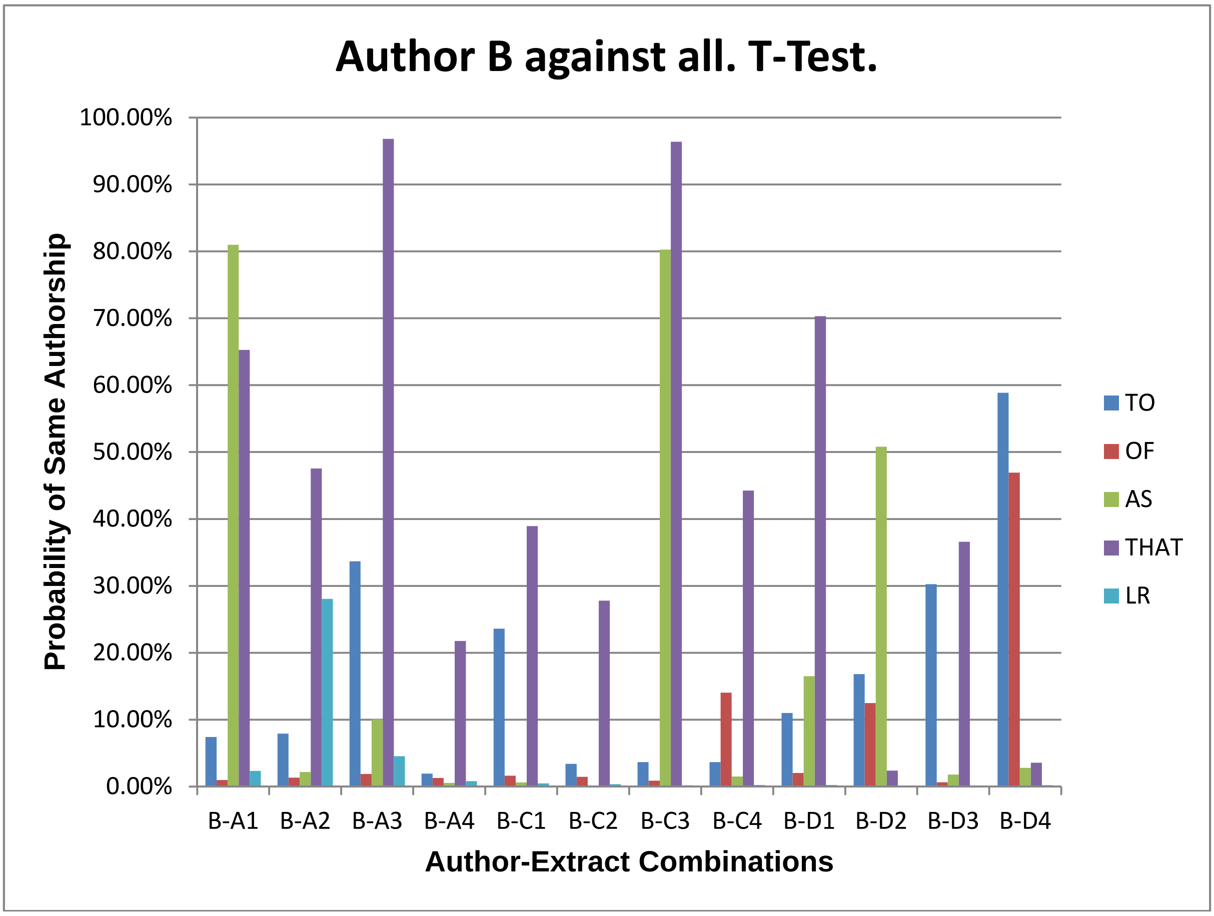
The biggest problem seems to be with the marker *That*. Out of 6 tests, it failed three, with all results shooting over 90%. This allows one to presume that, in the case of this study, *that* is not a significant authorship marker.



Overall Successful Differentiation (below 40%)	**51/60**	**85%**
Successful Differentiation below 20%	**42/60**	**70%**
Definite Failure (over 80%)	**1/60**	**1.67%**
Failed to Differentiate (over 50%)	**6/60**	**10%**

Author A against all: Table of probabilities

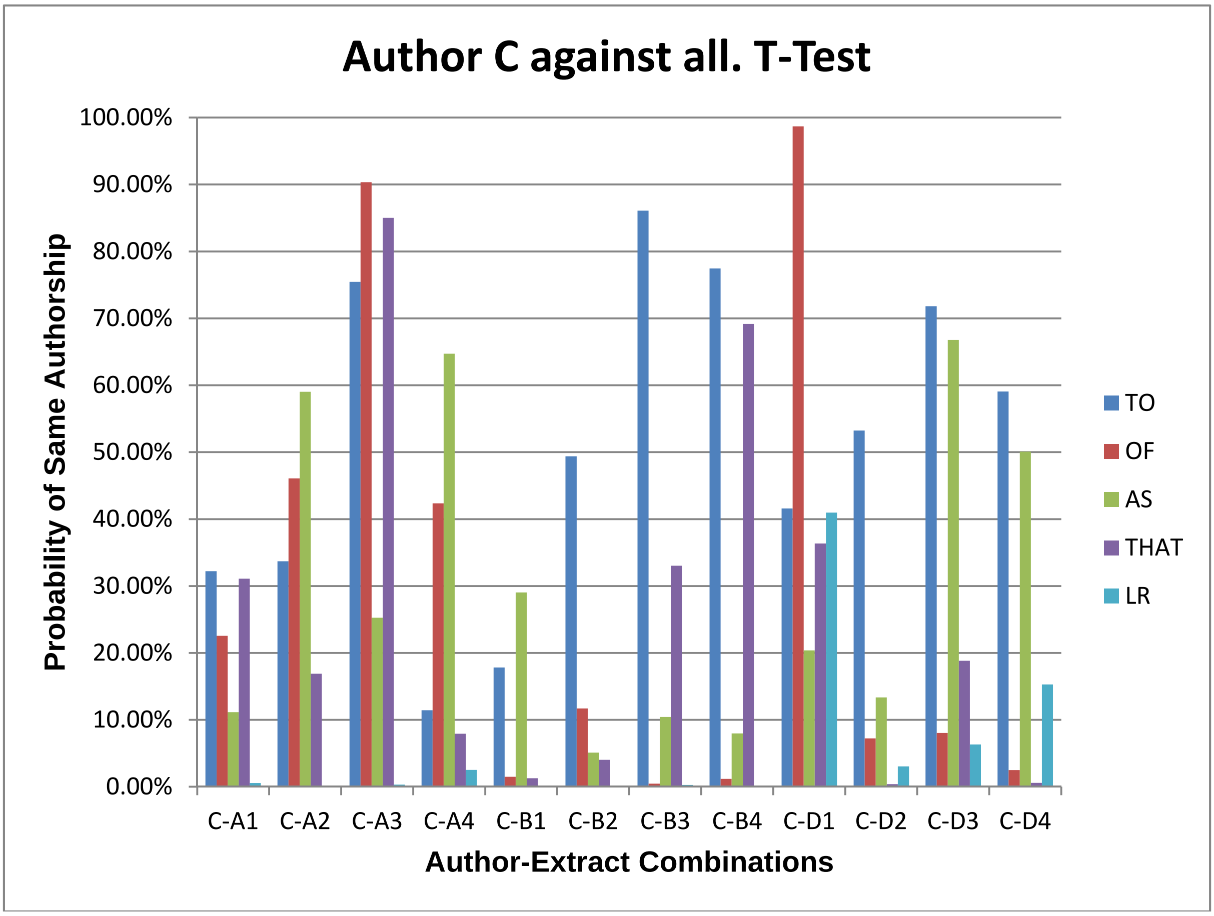
	TO	**P<=0.2**	**P<=0.4**	**OF**	**P<=0.2**	**P<=0.4**	**AS**	**P<=0.2**	**P<=0.4**	**THAT**	**P<=0.2**	**P<=0.4**	**LR**	**P<=0.2**	**P<=0.4**
A-B1	0.0004	1	1	0.0002	1	1	**0.5967**	0	0	0.0132	1	1	0.0141	1	1
A-B2	**0.3319**	0	1	0.0018	1	1	0.0322	1	1	0.0343	1	1	0.029	1	1
A-B3	0.0745	1	1	0.0001	1	1	0.1009	1	1	**0.2732**	0	1	**0.4466**	0	0
A-B3	0.0279	1	1	0.0002	1	1	0.0644	1	1	**0.7906**	0	0	0.0182	1	1
A-C1	0.0093	1	1	**0.2299**	0	1	0.1064	1	1	0.174	1	1	0.0355	1	1
A-C2	0.0027	1	1	**0.5084**	0	0	0.0108	1	1	0.0767	1	1	0.0224	1	1
A-C3	**0.3149**	0	1	0.0459	1	1	0.0817	1	1	**0.9683**	0	0	0.0078	1	1
A-C4	**0.3149**	0	1	0.001	1	1	**0.3913**	0	1	**0.2203**	0	1	0.0102	1	1
A-D1	**0.5219**	0	0	0.0546	1	1	**0.3274**	0	1	**0.4189**	0	0	0.011	1	1
A-D2	**0.2709**	0	1	0.0011	1	1	0.1547	1	1	0.0038	1	1	0.0049	1	1
A-D3	0.1208	1	1	0.0087	1	1	**0.4879**	0	0	0.1557	1	1	0.0063	1	1
A-D4	0.0176	1	1	0.0004	1	1	**0.7874**	0	0	0.0058	1	1	0.0082	1	1
Counts		7	11		10	11		7	9		7	9		11	11
Success Rate		**58.33%**	91.67%		83.33%	91.67%		**58.33%**	**75.00%**		**58.33%**	**75.00%**		91.67%	91.67%
Error Rate		**41.67%**	8.33%		16.67%	8.33%		**41.67%**	**25.00%**		**41.67%**	**25.00%**		8.33%	8.33%
	TO	**P<=0.05**	**P<=0.1**	**OF**	**P<=0.05**	**P<=0.1**	**AS**	**P<=0.05**	**P<=0.1**	**THAT**	**P<=0.05**	**P<=0.1**	**LR**	**P<=0.05**	**P<=0.1**
Counts		5	6		9	10		2	4		4	5		11	11
Success Rate		**41.67%**	50.00%		75.00%	83.33%		**16.67%**	**33.33%**		**33.33%**	**41.67%**		91.67%	91.67%
Error Rate		**58.33%**	50.00%		25.00%	16.67%		**83.33%**	**66.67%**		**66.67%**	**58.33%**		8.33%	8.33%



Overall Successful Differentiation (below 40%)	**49/60**	**81.67%**
Successful Differentiation below 20%	**41/60**	**68.33%**
Definite Failure (over 80%)	**4/60**	**6.67%**
Failed to Differentiate (over 50%)	**8/60**	**13.33%**

Author B against all. Table of Probabilities

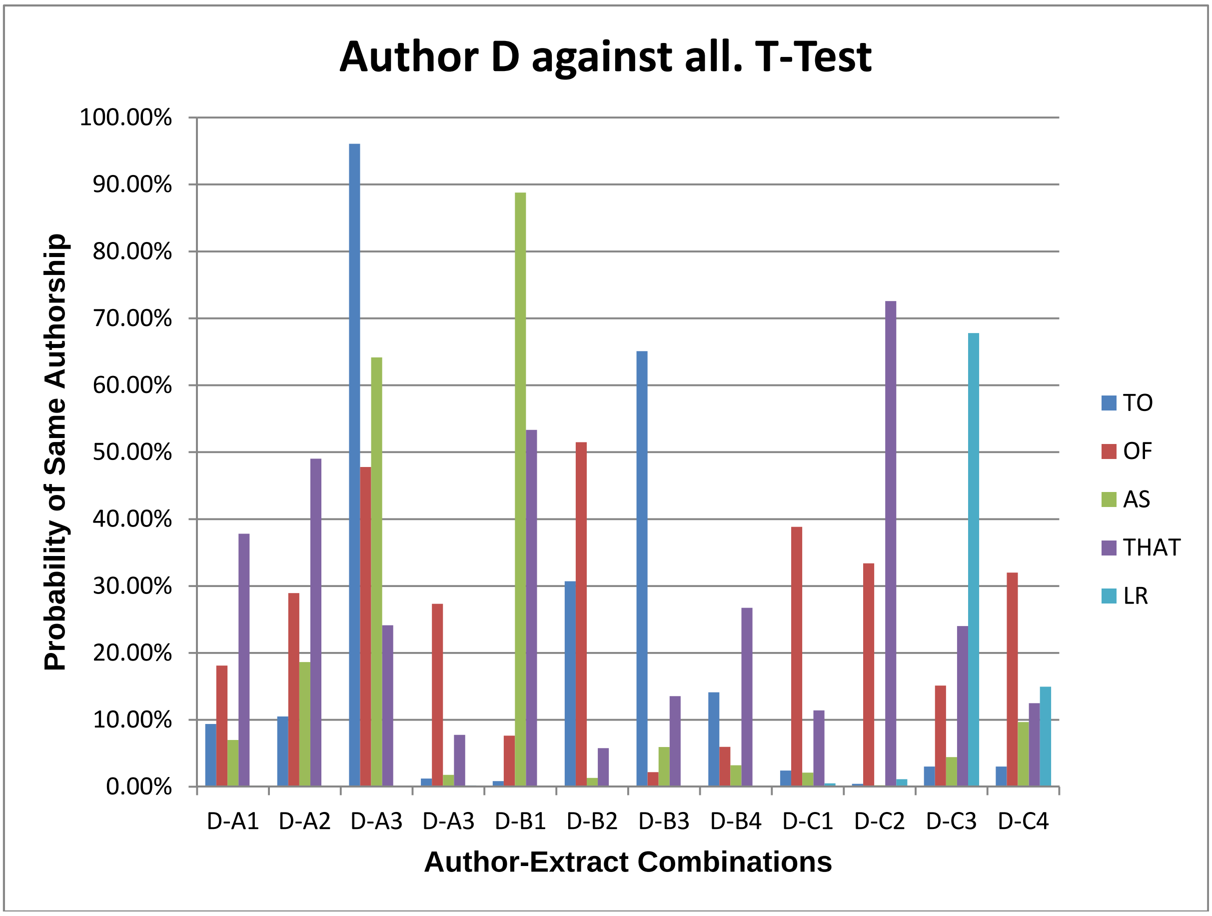
	TO	**P<=0.2**	**P<=0.4**	**OF**	**P<=0.2**	**P<=0.4**	**AS**	**P<=0.2**	**P<=0.4**	**THAT**	**P<=0.2**	**P<=0.4**	**LR**	**P<=0.2**	**P<=0.4**
B-A1	0.074	1	1	0.0098	1	1	**0.8099**	0	0	**0.6527**	0	0	0.0233	1	
B-A2	0.0792	1	1	0.0132	1	1	0.0216	1	1	**0.4754**	0	0	**0.2804**	0	
B-A3	**0.3367**	0	1	0.0188	1	1	0.1003	1	1	**0.9683**	0	0	0.0453	1	
B-A4	0.0194	1	1	0.0128	1	1	0.0055	1	1	**0.2177**	0	1	0.0081	1	
B-C1	**0.236**	0	1	0.0162	1	1	0.0061	1	1	**0.3893**	0	1	0.0048	1	
B-C2	0.0338	1	1	0.0146	1	1	0.0012	1	1	**0.2778**	0	1	0.0036	1	
B-C3	0.0364	1	1	0.0088	1	1	**0.8026**	0	0	**0.964**	0	0	0.0019	1	
B-C4	0.0364	1	1	0.1404	1	1	0.015	1	1	**0.4424**	0	0	0.0022	1	
B-D1	0.1099	1	1	0.0203	1	1	0.165	1	1	**0.7032**	0	0	0.0024	1	
B-D2	0.168	1	1	0.1249	1	1	**0.508**	0	0	0.0239	1	1	0.0014	1	
B-D3	**0.3026**	0	1	0.0064	1	1	0.0179	1	1	**0.3659**	0	1	0.0016	1	
B-D4	**0.5886**	0	0	**0.4692**	0	0	0.028	1	1	0.0355	1	1	0.002	1	
Counts		8	11		11	11		9	9		2	6		11	1
Success Rate		**66.67%**	91.67%		91.67%	91.67%		**75.00%**	**75.00%**		**16.67%**	**50.00%**		91.67%	100.00
Error Rate		**33.33%**	8.33%		8.33%	8.33%		**25.00%**	**25.00%**		**83.33%**	**50.00%**		8.33%	0.00
	TO	**P<=0.05**	**P<=0.1**	**OF**	**P<=0.05**	**P<=0.1**	**AS**	**P<=0.05**	**P<=0.1**	**THAT**	**P<=0.05**	**P<=0.1**	**LR**	**P<=0.05**	**P<=0.**
Counts		4	6		9	9		7	7		2	2		11	1
Success Rate		**33.33%**	**50.00%**		75.00%	75.00%		**58.33%**	58.33%		**16.67%**	**16.67%**		91.67%	91.67
Error Rate		**66.67%**	**50.00%**		25.00%	25.00%		**41.67%**	41.67%		**83.33%**	**83.33%**		8.33%	8.33



Overall Successful Differentiation (below 40%)	**41/60**	**68.33%**
Successful Differentiation below 20%	**33/60**	**55%**
Definite Failure (over 80%)	**4/60**	**6.67%**
Failed to Differentiate (over 50%)	**14/60**	**22.33%**

Author C against all. Table of Probabilities

	TO	**P<=0.2**	**P<=0.4**	**OF**	**P<=0.2**	**P<=0.4**	**AS**	**P<=0.2**	**P<=0.4**	**THAT**	**P<=0.2**	**P<=0.4**	**LR**	**P<=0.2**	**P<=0**
C-A1	**0.3221**	0	1	**0.2256**	0	1	0.1115	1	1	**0.311**	0	1	0.0055	1	
C-A2	**0.3371**	0	1	**0.4609**	0	0	**0.5902**	0	0	0.1688	1	1	0.0013	1	
C-A3	**0.7546**	0	0	**0.9034**	0	0	**0.2525**	0	1	**0.8501**	0	0	0.0032	1	
C-A4	0.1144	1	1	**0.4236**	0	0	**0.6471**	0	0	0.0792	1	1	0.0251	1	
C-B1	0.1783	1	1	0.0148	1	1	**0.2903**	0	1	0.0128	1	1	0.0005	1	
C-B2	**0.4938**	0	0	0.117	1	1	0.0511	1	1	0.0402	1	1	0.0008	1	
C-B3	**0.861**	0	0	0.0046	1	1	0.1043	1	1	0.3302	0	1	0.0026	1	
C-B4	**0.7747**	0	0	0.0117	1	1	0.0796	1	1	**0.6917**	0	0	0.0006	1	
C-D1	**0.4158**	0	0	**0.9869**	0	0	**0.2039**	0	1	**0.3635**	0	1	**0.4097**	0	
C-D2	**0.5324**	0	0	0.0723	1	1	0.1336	1	1	0.0038	1	1	0.0305	1	
C-D3	**0.7181**	0	0	0.0805	1	1	**0.6678**	0	0	0.1882	1	1	0.0632	1	
C-D4	**0.5907**	0	0	0.025	1	1	**0.5009**	0	0	0.0057	1	1	0.1529	1	
Counts		2	4		7	8		5	8		7	10		11	
Success Rate		**16.67%**	**33.33%**		**58.33%**	**66.67%**		**41.67%**	**66.67%**		**58.33%**	83.33%		91.67%	91.67
Error Rate		**83.33%**	**66.67%**		**41.67%**	**33.33%**		**58.33%**	**33.33%**		**41.67%**	16.67%		8.33%	8.33
	TO	**P<=0.05**	**P<=0.1**	**OF**	**P<=0.05**	**P<=0.1**	**AS**	**P<=0.05**	**P<=0.1**	**THAT**	**P<=0.05**	**P<=0.1**	**LR**	**P<=0.05**	**P<=0**
Counts		**0**	**0**		**4**	**6**		**0**	**2**		**4**	5		9	
Success Rate		**0.00%**	**0.00%**		**33.33%**	**50.00%**		**0.00%**	**16.67%**		**33.33%**	**41.67%**		75.00%	83.33
Error Rate		**100.00%**	**100.00%**		**66.67%**	**50.00%**		**100.00%**	**83.33%**		**66.67%**	**58.33%**		25.00%	16.67



Overall Successful Differentiation (below 40%)	**50/60**	**83.33%**
Successful Differentiation below 20%	**40/60**	**66.67%**
Definite Failure (over 80%)	**2/60**	**3.33%**
Failed to Differentiate (over 50%)	**8/60**	**13.33%**

Author D against all. Table of Probabilities.

	TO	**P<=0.2**	**P<=0.4**	**OF**	**P<=0.2**	**P<=0.4**	**AS**	**P<=0.2**	**P<=0.4**	**THAT**	**P<=0.2**	**P<=0.4**	**LR**	**P<=0.2**	**P<=0.4**
D-A1	0.0937	1	1	0.181	1	1	0.0699	1	1	**0.3781**	0	1	0.0006	1	1
D-A2	0.1051	1	1	**0.2894**	0	1	0.1863	1	1	**0.4902**	0	0	0.0002	1	1
D-A3	**0.9608**	0	0	**0.4778**	0	0	**0.6416**	0	0	**0.2414**	0	1	0.0004	1	1
D-A3	0.0122	1	1	**0.2735**	0	1	0.0178	1	1	0.0775	1	1	0.0019	1	1
D-B1	0.0084	1	1	0.0763	1	1	**0.888**	0	0	**0.5334**	0	0	0.0001	1	1
D-B2	**0.307**	0	1	**0.5149**	0	0	0.0132	1	1	0.0576	1	1	0.0001	1	1
D-B3	**0.6509**	0	0	0.0219	1	1	0.0594	1	1	0.1354	1	1	0.0003	1	1
D-B4	0.1412	1	1	0.0596	1	1	0.0321	1	1	**0.2673**	0	1	0.0001	1	1
D-C1	0.0242	1	1	**0.3885**	0	1	0.0211	1	1	0.1141	1	1	0.0052	1	1
D-C2	0.0044	1	1	**0.3338**	0	1	0.0022	1	1	**0.7257**	0	0	0.0113	1	1
D-C3	0.0302	1	1	0.1511	1	1	0.0443	1	1	**0.2402**	0	1	**0.678**	0	0
D-C4	0.0302	1	1	**0.3199**	0	1	0.0966	1	1	0.1248	1	1	0.1496	1	1
Counts		9	10		5	10		10	10		5	9		11	11
Success Rate		**75.00%**	83.33%		**41.67%**	83.33%		83.33%	83.33%		**41.67%**	75.00%		91.67%	91.67%
Error Rate		**25.00%**	16.67%		**58.33%**	16.67%		16.67%	16.67%		**58.33%**	25.00%		8.33%	8.33%
	TO	**P<=0.05**	**P<=0.1**	**OF**	**P<=0.05**	**P<=0.1**	**AS**	**P<=0.05**	**P<=0.1**	**THAT**	**P<=0.05**	**P<=0.1**	**LR**	**P<=0.05**	**P<=0.1**
Counts		6	7		1	3		6	9		0	2		10	10
Success Rate		**50.00%**	58.33%		**8.33%**	25.00%		**50.00%**	75.00%		**0.00%**	**16.67%**		83.33%	83.33%
Error Rate		**50.00%**	41.67%		**91.67%**	75.00%		**50.00%**	25.00%		**100.00%**	**83.33%**		16.67%	16.67%

Appendix 3: Post-Experimental Analysis

Two essays, one essay per author, introductions and conclusions excluded from the analysis.

Tables for Author E:

	ES1	ES2	ES3	Mean	StDev	StDev/Mean
TO	4.2373%	4.0917%	3.8397%	4.0562%	0.2012%	4.9593%
IN	**3.0508%**	**0.8183%**	**1.8364%**	**1.9018%**	**1.1177%**	**58.7689%**
IS	2.5424%	2.1277%	2.6711%	2.4471%	0.2840%	11.6044%
THAT	1.8644%	2.1277%	1.6694%	1.8872%	0.2300%	12.1874%
OF	1.6949%	1.6367%	2.1703%	1.8340%	0.2927%	15.9612%
IF	1.1864%	0.8183%	1.3356%	1.1134%	0.2663%	23.9132%
IT	1.6949%	1.4730%	1.3356%	1.5012%	0.1813%	12.0772%
ON	0.8475%	1.3093%	1.3356%	1.1641%	0.2745%	23.5821%
THEIR	**0.8475%**	**1.4730%**	**1.3356%**	**1.2187%**	**0.3287%**	**26.9736%**
THIS	**0.8475%**	**1.6367%**	**1.5025%**	**1.3289%**	**0.4223%**	**31.7759%**
LR	26.1017%	24.3863%	25.7095%	25.3992%	0.8988%	3.5388%

	EL1	EL2	Mean	StDev	StDev/Mean
TO	3.9517%	4.2529%	4.1023%	0.2130%	5.1917%
IS	2.5247%	2.4138%	2.4693%	0.0784%	3.1758%
IN	**2.3052%**	**1.4943%**	**1.8998%**	**0.5734%**	**30.1825%**
THAT	2.0856%	1.7241%	1.9049%	0.2556%	13.4194%
OF	1.7563%	1.9540%	1.8552%	0.1398%	7.5355%
IF	**1.0977%**	**0.5747%**	**0.8362%**	**0.3698%**	**44.2259%**
ON	0.9879%	0.8046%	0.8963%	0.1296%	14.4617%
THEIR	**0.9879%**	**1.4943%**	**1.2411%**	**0.3581%**	**28.8517%**
THIS	**0.9879%**	**1.7241%**	**1.3560%**	**0.5206%**	**38.3903%**
FROM	**0.7684%**	**0.3448%**	**0.5566%**	**0.2995%**	**53.8143%**
LR	20.7464%	21.8391%	21.2928%	0.7727%	3.6287%

	ES1	ES2	ES3	EL1	EL2	Mean	StDev	StDev/Mean
TO	4.2373%	4.0917%	3.8397%	3.9517%	4.2529%	4.0747%	0.1795%	4.4045%
IN	**3.0508%**	**0.8183%**	**1.8364%**	**2.3052%**	**1.4943%**	**1.9010%**	**0.8407%**	**44.2251%**
IS	2.5424%	2.1277%	2.6711%	2.5247%	2.4138%	2.4559%	0.2049%	8.3450%
THAT	1.8644%	2.1277%	1.6694%	2.0856%	1.7241%	1.8942%	0.2071%	10.9316%
OF	1.6949%	1.6367%	2.1703%	1.7563%	1.9540%	1.8424%	0.2188%	11.8744%
IF	**1.1864%**	**0.8183%**	**1.3356%**	**1.0977%**	**0.5747%**	**1.0025%**	**0.3045%**	**30.3687%**
IT	1.6949%	1.4730%	1.3356%	1.6465%	1.4943%	1.5289%	0.1441%	9.4261%
ON	0.8475%	1.3093%	1.3356%	0.9879%	0.8046%	1.0570%	0.2518%	23.8240%
THEIR	0.8475%	1.4730%	1.3356%	0.9879%	1.4943%	1.2277%	0.2937%	23.9204%
THIS	**0.8475%**	**1.6367%**	**1.5025%**	**0.9879%**	**1.7241%**	**1.3397%**	**0.3964%**	**29.5871%**
LR	26.1017%	24.3863%	25.7095%	20.7464%	21.8391%	23.7566%	2.3690%	9.9718%

Tables for Author F:

	FS1	FS2	FS3	Mean	StDev	StDEv/Mean
OF	3.2819%	3.7906%	2.3490%	3.1405%	0.7311%	23.2806%
TO	2.7027%	3.9711%	3.6913%	3.4550%	0.6664%	19.2875%
IN	**2.5097%**	**2.8881%**	**1.5101%**	**2.3026%**	**0.7120%**	**30.9191%**
THAT	2.3166%	1.9856%	2.0134%	2.1052%	0.1836%	8.7215%
IT	**1.7375%**	**1.6245%**	**2.5168%**	**1.9596%**	**0.4858%**	**24.7931%**
THEIR	1.1583%	0.9025%	0.8389%	0.9666%	0.1691%	17.4911%
FOR	**0.9653%**	**0.3610%**	**0.6711%**	**0.6658%**	**0.3022%**	**45.3867%**
BY	0.7722%	0.7220%	0.6711%	0.7218%	0.0506%	7.0037%
LR	31.2741%	30.5054%	34.3960%	32.0585%	2.0605%	6.4273%

	FL1	FL2	Mean	StDev	StDev/Mean
OF	2.7144%	3.6145%	3.1645%	0.6365%	20.1130%
TO	3.8002%	3.0790%	3.4396%	0.5100%	14.8263%
IS	**3.3659%**	**2.4096%**	**2.8878%**	**0.6762%**	**23.4164%**
THAT	2.0630%	2.1419%	2.1025%	0.0558%	2.6536%
IT	2.1716%	1.7403%	1.9560%	0.3050%	15.5922%
IN	2.0630%	2.5435%	2.3033%	0.3398%	14.7515%
AS	**1.6287%**	**0.9371%**	**1.2829%**	**0.4890%**	**38.1195%**
FOR	**0.4343%**	**0.9371%**	**0.6857%**	**0.3555%**	**51.8497%**
THEIR	0.9772%	0.9371%	0.9572%	0.0284%	2.9624%
THERE	**0.5429%**	**0.8032%**	**0.6731%**	**0.1841%**	**27.3471%**
BY	0.7600%	0.6693%	0.7147%	0.0641%	8.9743%
LR	28.4473%	29.8527%	29.1500%	0.9938%	3.4092%

	FS1	FS2	FS3	FL1	FL2	Mean	StDev	StDev/Mean
OF	3.2819%	3.7906%	2.3490%	2.7144%	3.6145%	3.1501%	0.6072%	19.2764%
TO	2.7027%	3.9711%	3.6913%	3.8002%	3.0790%	3.4489%	0.5358%	15.5368%
IN	**2.5097%**	**2.8881%**	**1.5101%**	**2.0630%**	**2.5435%**	**2.3029%**	**0.5313%**	**23.0719%**
THAT	2.3166%	1.9856%	2.0134%	2.0630%	2.1419%	2.1041%	0.1328%	6.3115%
IT	1.7375%	1.6245%	2.5168%	2.1716%	1.7403%	1.9581%	0.3759%	19.1953%
THEIR	1.1583%	0.9025%	0.8389%	0.9772%	0.9371%	0.9628%	0.1205%	12.5150%
FOR	**0.9653%**	**0.3610%**	**0.6711%**	**0.4343%**	**0.9371%**	**0.6738%**	**0.2782%**	**41.2860%**
BY	0.7722%	0.7220%	0.6711%	0.7600%	0.6693%	0.7189%	0.0482%	6.7015%
LR	31.2741%	30.5054%	34.3960%	28.4473%	29.8527%	30.8951%	2.2153%	7.1704%

Notes:
ES1, FS1 etc. – short texts by authors E and F
EL1, FL1 etc. – long texts by authors E and F
Rows typed in Bold font contain unreliable data that was excluded from the analysis.