

Début: the undergraduate journal of  
languages, linguistics and area studies  
Volume 2  
Number 1  
2011

ISSN: 2044-7256

## **Contents**

### **Editorial**

The future of Debut: have your say **1**

*John Canning*

### **Articles**

An exploration of two manuscripts written between AD 900 and AD 1299  
as exemplars of Old and Middle English **3**

*Coleen Borrer*

Edited, paraphrased and still unique? Authorship analysis and uniqueness  
of encoding in undergraduate student essays **13**

*Vlad Mackevic*

**Instructions for authors** **49**

## **Editorial. The future of *Début*: have your say**

John Canning

*LLAS: Centre for Languages, Linguistics and Area Studies, School of Humanities,  
University of Southampton, Southampton, UK*

[j.canning@soton.ac.uk](mailto:j.canning@soton.ac.uk)

Most work written by undergraduates (or postgraduates for that matter) is seen by no-one except for the assessors. However a growing number of outlets are publishing the work of undergraduates in a variety of disciplines. In the UK these publications include the multidisciplinary *Reinvention* based at Warwick University, *Geoversity* at Oxford Brookes, *Bioscience Horizons* published by Oxford Journals and *SURJ* from the University of Surrey. In the USA undergraduate journals are even more widespread. *Watcher Junior: the undergraduate journal of Whedon Studies*, now into its sixth year is probably the most intriguing (for the uninitiated Joss Whedon is author of *Buffy, the Vampire Slayer*).

A few years ago I become aware that students are graduating from university without awareness of the peer review process and how knowledge is published in journals. In all honesty I can't remember how or when I learnt about the process myself. Some may regard knowledge of such things as trivial, but it is a key part of how disciplines operate. Reviewers and journals set the boundaries of the discipline, deciding what should and should not be published in contributing to that discipline's scholarship (see Becher and Trowler, 2001). Success in academia depends on publishing work and publishing it in the right outlets. To this end *Début* is presently published under a traditional peer review model. Submissions are anonymised and reviewed by experts in the field (usually academic teaching staff or postgraduate students). I had hoped that this traditional system would enable students, particularly those hoping to have an academic career, to gain experience of how the system works (or doesn't work).

The peer review system has its limitations of course and *Début* is no exception to this. Firstly, we cannot be totally sure that the articles published are better than the articles that were rejected.

A second barrier is the adage that students (in general) don't like doing work which will not count directly towards their degree. Work submitted to *Début* is often derived from essays or dissertations which have already been assessed (and have probably received good marks). Authors do not always appreciate 'negative' feedback on their first class work and there have been instances where authors have not made revisions to their work despite my encouragement.

A third issue is that of standards. What standard of work can reasonably be expected from an undergraduate who has only studied the subject for a couple of years? Furthermore, the papers that students write are not necessarily a specialist area for the teacher who taught the student. About half of papers submitted to *Début* are not published and I suspect that this is often the reason. For example, what seems like a

good attempt on Yoruba linguistics to a linguistics lecturer who does not know the language is not necessarily a good attempt in the estimation of an expert Yoruba linguistics researcher.

### **Reviewing *Début***

The decision of the UK's Higher Education Academy to withdraw funding from its 24 subject centres from August 2011 makes a review of how *Début* operates particularly timely. *Début* is among the activities to be continued by the new Centre for Languages, Linguistics and Area Studies whose staff will include many currently employed at the Subject Centre. Various ideas have been put forward to me including:

1. Having teams of students edit special editions of the journal
2. Putting up draft papers for public consultation prior to publication in the journal with editors using this consultation to reach their decision
3. Publish papers without any review process (expect maybe a check for plagiarism)
4. Publish any undergraduate paper accompanied by a reference/recommendation by a lecturer
5. Accept research presented in alternative formats, e.g. posters, videos.
6. Maintain the current double-blind review system

Walkington and Jenkins (2008) observe that the research process (or cycle) is not completed until research is published, therefore most undergraduate research does not complete the process. Whatever 'model' or combination of models is selected the journal will continue to provide a place to demonstrate the high research capabilities of undergraduate students.

To feedback your opinion on the future of *Début*, please email me (j.canning@soton.ac.uk) or post your comments on the *Début* facebook page [www.facebook.com/group.php?gid=158224702055](http://www.facebook.com/group.php?gid=158224702055)

### **References**

Becher, T and P. Trowler (2001), *Academic Tribes and Territories*, 2nd ed. (Milton Keynes: Open University Press).

Walkington, H. and A. Jenkins (2008) Embedding undergraduate research publication in the student learning experience. *Brookes eJournal of Learning and Teaching* 2(3). Available at: <http://bit.ly/b0Bbm3>

## **An exploration of two manuscripts written between AD 900 and AD 1299 as exemplars of Old and Middle English**

Coleen Borrer

*University of Brighton, Brighton, UK*

coleenkaren@yahoo.co.uk

Coleen Borrer graduated with a first-class honours in English Language and English Literature from the University of Brighton in 2010. She has a keen interest in history and the English language. Her career interests include the preservation and archiving of collections within the cultural and heritage sector.

### **Abstract**

This article examines the change in the English language that occurs between the Old and Middle English periods. It looks at an extract of Cædmon's Story taken from Bede's *Ecclesiastical History of the English People* from the Old English period, and the opening 12 lines of *The Owl and the Nightingale* taken from MS. Cotton Caligula A. ix from the Middle English period. This article first briefly explores some of the background events that influenced the Old and Middle English periods. Secondly, the linguistic features of the texts are explored and compared to one another in order to chart the changes that have occurred between the two periods.

### **Background to the English language**

The aim of this article is to explore the changes that occurred in the vocabulary and grammar of the English language between the Old and Middle English periods. The different periods of the English language are difficult to define, as changes to the language did not occur instantaneously with the events that caused these changes. For the purpose of this article, I define the Old English period as being between AD 450 – 1100, and the Middle English period as being between AD 1100 – 1500. This is based on Graddol (1996), and Brinton and Arnovick's (2006) definitions as shown below:

- (1) Early Old English (450 – c. 850)
- (2) Later Old English (c. 850 – 1100)
- (3) Middle English (c. 1100 – 1450)  
(Graddol, *et al.* (eds.) 1996, p.41).

- (1) Old English (c. 450 – 1100)
- (2) Middle English (1066 – 1500)  
(Brinton and Arnovick 2006, pp.9 – 10).

The Old English extract I use is taken from Cædmon's Story, which was written by Bede in his *Ecclesiastical History of the English People*. The main influence on the development of the language during the Old English period was the invasion of the West Germanic tribes in the fifth century, who brought with them their native dialects. Overall, they had a great effect on Modern English, and it is estimated that 83% of the 1000 most regularly used words today have their origins in Old English (Brinton and Arnovick 2006, p.166). Towards the end of this period, the Vikings also invaded Britain. As the West Germanic Tribes continued to make up the majority of the

population, Old Norse, the language of the Vikings, became assimilated into Old English, although some features did not begin to appear until the early 12th century (Crystal 1995, p.25).

The second extract is taken from *The Owl and the Nightingale*, which was written during the Middle English period. The Norman Conquest of AD 1066 was the major influence on the language during this time. The Normans introduced Norman French to England, although it is estimated that 90 – 98% of the population continued speaking English (Stockwell and Minkova 2001, p.35). Overall, approximately 10,000 words entered the language from Norman French (Bragg 2003, p.37), and it is estimated that 75% of these are still in use (Baugh and Cable 2002, p.178). As will be explored in the following pages, the language changes to such an extent that according to Crystal, Middle English becomes “familiar” to the modern reader (Crystal 2004: 105).

I chose these two texts, as Cædmon’s Story is from later in the Old English period. This ensured that the language used in this text had become more established than a text written in AD 500 and would begin to include some influences from Old Norse. However, *The Owl and the Nightingale* was written in the early/mid Middle English period. This meant that by this time Norman French had started to influence the language. As the boundaries between language periods are difficult to define, this also ensures that few features of Early Modern English, AD 1500 – 1700 (Brinton and Arnovick 2006, pp.9–10), would be present in this text.

## **Background to the texts**

### ***Text one: Cædmon’s Story from Ecclesiastical History of the English People by Bede***

Cædmon’s Story tells the story of a cowherd named Cædmon who lived at Whitby Abbey (North and Allard (eds.) 2007, p.189). Bede tells that Cædmon had never learnt how to compose poetry but one night after leaving a feast before he was asked to sing, he had a dream and thereafter was able to sing and compose verses praising God (Bede 1990, pp.248 – 249). The poem Cædmon created in his dream is now known as Cædmon’s Hymn, and as these events are estimated to have taken place between AD 657 – 680, and because his work has survived, Cædmon is considered as the “first named English poet” (North and Allard (eds.) 2007, pp.189 – 190).

Bede was born in AD 673 and was placed in the monastery at Wearmouth aged seven in order to receive an education. He later moved to the monastery at Jarrow where he stayed until his death in AD 735 (Farmer in Bede 1990, pp.19 – 20; 22). There is little information about Bede’s life, but he wrote over 30 books in Latin, although he also believed that English should be used (Bragg 2003, p.12). *Ecclesiastical History of the English People* was finished in AD 731 and has become his best-known piece of work. It is from this that much of the history of Britain and the coming of Christianity is known (Bede’s World 2003). In the ninth century, King Alfred ordered *Ecclesiastical History of the English People* to be translated into Old English (Crystal 2004, p.17). The extract used is from a translation written between AD 900 – 950 (Mielniczuk 2004).

### ***Text two: The Owl and the Nightingale***

*The Owl and the Nightingale* is a poem of 1794 lines (Burrow and Turville-Petre 1996, p.81) that tells of a debate between two birds: an owl and a nightingale (Stone 1971, p.155). The poem describes the birds' argument and at the end they fly off to receive judgement from Nicholas of Guildford (Stone 1971, p.155), which also leaves the reader to form their own opinion as the debate is not resolved (Treharne 2004, p.370).

The author of the poem is unknown, although scholars have tried to guess (Wells (ed.) 1907, pp.xxv – xxvi). Many scholars believe that it was written by Nicholas of Guildford (Cartlidge, ed. 2001, p.xiv) as the birds fly off to receive his judgement. However, Wells (1907) says the majority disagree (Wells (ed.) 1907, p.xxvi). This is because the poem starts in the first person, as the author is present at the debate. As at the end of the poem they fly off to see Nicholas, it is unlikely that he is the author, as he probably would not begin to write about himself in the third person (Wells (ed.) 1907, p.xxvi).

The composition date of the poem is unknown. As King Henri is mentioned, many scholars believe this to be King Henri II who died in AD 1189 and therefore puts the composition date at between AD 1189 and 1216 (Burrow and Turville-Petre 1996, p.81). However, Cartlidge (2001) believes that this may actually be later with the possibility of it being composed after AD 1272 and the death of King Henri III (Cartlidge 2001, p.xv).

### ***Extract from text one***

- (1) Wæs he se mon in weoruldhade geseted oð þa tide þe he  
*Was he the man in secular life settled until the time that he*
- (2) wæs gelyfdre ylde, ond he næfre nænig leoð geleornode. Ond he  
*was of-advanced age, and he never any poem learned. And he*
- (3) for þon oft in gebeorscipe, þonne þær wæs blisse intinga  
*therefore often at banquet, when there was of-joy occasion*
- (4) gedemed, þæt heo ealle sceolden þurh endebyrdnesse be hearpan  
*decided, that they all should by arrangement with harp*
- (5) singan. Þonne he geseah þa hearpan him nealecan, þonne aras he  
*to sing. When he saw the harp him approach, then arose he*
- (6) for scome from þæm symble ond ham eode to his huse. Þa he  
*for shame from the feast and home went to his house. When he*
- (7) þæt þa sumre tide dyde, þæt he forlet þæt hus þæs  
*that a certain time did, that he left the house of the*
- (8) gebeorscipes, ond ut wæs gongende to neata scipene,  
*banquet, and out was going to of-cattle stall*
- (9) þara heord him wæs þære neahte beboden, þa he ða þær  
*of which keeping him was that night entrusted, when he there*
- (10) in gelimplice tide his leomu on reste gesette ond onslepte,  
*at suitable time his limbs at rest set and fell asleep,*
- (11) þa stod him sum mon æt þurh swefn ond hine halette  
*then stood him a certain man beside in dreams, and him hailed*
- (12) ond grette ond hine be his noman nemnde: 'Cedmon, sing me  
*and greeted, and him by his name called 'Cædmon, sing me*
- (13) hwæthwugu.' Þa ondswarede he, and cwæð: 'Ne con Ic noht  
*something.' Then answered he, and said: 'Not can I nothing*

- (14) singan; ond Ic for þon of þeossum gebeorscipe ut eode, ond hider  
*sing; and I for that from this banquet out went and hither*
- (15) gewat, for þon Ic naht singan ne cuðe.  
*came, because I nothing to sing not knew how.'*

Source: Oxford Bodleian Library, Tanner 10 (extract taken from Treharne 2004, pp.4 – 6, translation taken from Crystal 1995, p.20).

***Extract from text two***

- (1) Ich was in one sumere dale,  
*I was in a summer valley,*
- (2) In one suþe diȝele hale:  
*In a very hidden corner*
- (3) Iherde ich holde grete tale  
*Heard I held great tale*
- (4) An hule and one nigtingale.  
*An owl and a nightingale.*
- (5) Þat plait was stif & starc & strong,  
*That plea was stiff and firm and strong,*
- (6) Sumwile softe & lud among;  
*Formerly soft and loud among;*
- (7) An aiþer aȝen oþer sval,  
*An either against other swell,*
- (8) & let þat vole mod ut al;  
*and let that evil mood out all;*
- (9) & eiþer seide of oþeres custe  
*and either side of other's character*
- (10) Þat alre worste þat hi wuste;  
*That all worst that they knew;*
- (11) & hure & hure of oþeres songe,  
*and at intervals of other's song,*
- (12) Hi holde plaiding suþe stronge.  
*They held pleading very strong.*

Source: British Library, London, MS Cotton Caligula A. ix (extract taken from Cartlidge 2001, p.2, translation by the Author after Mayhew and Skeat 2005; McSparran (ed.) 2001).

**Linguistic analysis of the texts**

***Vocabulary***

Old English words appear very different from Modern English. It is estimated that over 80% of Old English vocabulary has been replaced (Stockwell and Minkova 2001, p.46); however, 76% of the most common 100 Old English words are still in use in Modern English (Brinton and Arnovick 2006, p.165). Although many words appear unfamiliar in extract one, some are recognisable. In lines five and 15, the



word 'singan' is the Old English origin for the Modern English word 'sing' (Harper 2010). 'Ondswarede' in line 13 means 'answered', and although the spelling of the word has changed, there is similarity between the two (Crystal 1995, p.22). However, many of these words are now obsolete. In line nine, the word 'beboden' translates into Modern English as 'entrusted', however 'beboden' is a form of the word 'bebéodan' (Old English Made Easy 2008) and is now obsolete (Crystal 1995, p.22).

In extract two, many words are becoming more familiar to Modern English making the piece easier to read than extract one. In line ten, the word 'worste', which translates as 'worst', appears and is identical to Modern English with the exception of the final 'e'. Also, in line one the word 'sumere' translates as 'summer' and the two appear very similar. However, some words are now obsolete. In line two, the word 'digele' translates as 'hidden', but this word is no longer in use (Freeborn 1998, p.125). This is also seen with 'hale' in the same line, which translates as 'corner' (Freeborn 1998, p.125). In line five there is evidence of alliteration with 'stif & starc & strong', which translates as 'stiff and firm and strong'. Alliteration was widely used in poetry in order for oral poems to be remembered easily. The same is true of rhyming (Crystal 2004, p.93), which is seen through the use of rhyming couplets in extract two such as 'custe' and 'wuste' from lines nine and ten. The use of rhyming couplets in poetry and eight-syllable lines originates from the French (Bragg 2003, p.49). However, in extract two there is only one French word, which is 'plait', meaning 'plea', in line five (Graddol, *et al.* (eds.) 2007, p.68). 'Plea' is a legal word and these entered the language first from French (Bragg 2003, p.37).

The alphabet of Old English was also different to Modern English. The Roman alphabet was introduced by the Irish missionaries and contained 24 letters (Bragg 2003, pp.11 – 12), some of which were influenced by the runic alphabet (Crystal 1995, pp.16 – 17). The word 'wæs' in line one of extract one contains the letter 'æ' known as an 'ash' (Smith 1999, p.47) and this was eventually replaced by 'a' prior to the end of the 13th century (Freeborn 1998, p.25). Other symbols no longer in use are the 'thorn' written as 'þ' and 'eth', written as 'ð' (Smith 1999, p.47). These are seen in line four with the word 'þæt', and line one with 'oð'. The 'thorn' is from the runic alphabet of the West Germanic tribes (Brinton and Arnovick 2006, p.154), whilst the 'eth' is taken from Irish writing (Freeborn 1998, p.24). Both letters represents the 'th' sound (Brinton and Arnovick 2006, p.154) and they were often used interchangeably (Smith 1999, p.47). The 'eth' stopped being used by approximately AD 1300 (Freeborn 1998, p.25), whilst the 'thorn' was eventually replaced with 'th' during the 15th century (Freeborn 1998, p.25).

In extract two, the letters begin to resemble the Modern English alphabet. In this extract the 'eth' (ð) and 'ash' (æ) are not used, although the thorn (þ) is still in use as seen in the word 'oþer' in line seven, which translates as 'other'. Both the 'eth' and 'ash' were obsolete by the end of the 13th century (Freeborn 1998, p.25) and this may explain why they are not used in the text. However, other letters are still in use. The word 'niȝtingale' in line four contains the letter 'ȝ' known as a 'yogh' (Brinton and Arnovick 2006, p.154), which represented the 'g' and 'j' sounds (Crystal 1995, p.16).

Another letter in use is 'ƿ', known as a 'wynn' from the runic alphabet (Brinton and Arnovick 2006, p.154). The 'wynn' represented the 'w' sound, and prior to its introduction either a 'u' or 'uu' was used (Freeborn 1998, p.24). The Normans started replacing the 'wynn' with a 'w' in the 11th century (Crystal 1995, p.264), although it is still used throughout this manuscript, which was written in the 13th century. In

extract two, the 'wynn' has been modified from the original manuscript. An example of this is seen in line one where the word 'was' is originally spelt with a 'wynn' (ƿ). Additionally, in much of the poem a '&' is used, whereas in the original manuscript '7' is used, which means 'and' (Brinton and Arnovick 2006, p.154).

Old English at first may appear very unlike Modern English due to the spelling. However, by changing the obsolete letters into their modern forms, it is easier to perceive some words as Modern English (Graddol, *et al.* (eds.) 1996, p.112). This is seen with the word 'þæt' in line seven of extract one, which when changed becomes 'that'. Although it is not known how Old English sounded, it is likely that the spelling of words was related to their pronunciation (Singh 2005, p.76). The spelling in Old English was not regular and varied according to the preference of the scribe and their dialect (Crystal 1995, p.17).

The influence of the Norman scribes can be seen in extract two. The Normans changed words spelt with a 'u' into 'ou' (Culpeper and Archer 2009, p.251) but this has not occurred by the time this text was written. In line eight, the word 'out' is still spelt as 'ut'. This spelling is unchanged since Old English and is also found in line eight of the extract from Cædmon's Story. However, other spellings have changed from the earlier period. In line four of Cædmon's Story, the word 'ealle', which means 'all', is spelt differently to 'al' in line eight of *The Owl and the Nightingale*; although there still appears to be no set spelling as it is also spelt as 'alre' in line ten of the same extract. Another influence of the Norman's scribes was to drop the initial 'h' prior to consonants (Crystal 1995, p.42). The Old English word for 'loud' was 'hlud' (Old English Made Easy 2008). In extract two, line six the 'h' has been dropped giving 'lud'. The influence of the Normans would later change the 'u' to 'ou' giving the Modern English 'loud'.

As stated above, the dialect of the scribes also influenced their spelling. Altogether, there are four recognised Old English dialects: Northumbrian, Mercian, Kentish and West Saxon (Crystal 1995, p.29). The majority of Old English texts are written in the West Saxon dialect (Baker 2003, p.10) including the above extract from Cædmon's Story (Treharne 2004, p.1). Overall, dialect had the major effect on spelling as words were spelt differently according to the region and pronunciation which led to no uniform system throughout England (Graddol, *et al.* (eds.) 1996, p.72). In line 13 of extract one there is the word 'cwæð', which later developed into 'quoth' (Old English Made Easy 2008) and in Modern English means 'said'. However, in the *Lindisfarne Glosses*, written in the Northumbrian dialect in approximately AD 950 (van Gelderen 2006, p.48) this word is spelt differently as 'cweð' (van Gelderen 2006, p.59). Another example of this spelling variance is found in the *Rushworth Glosses*, written circa AD 970 in the Mercian dialect (van Gelderen 2006, p.48), where 'cweþ' is used (van Gelderen 2006, p.59).

By the Middle English period, dialects are still largely based on the same areas as Old English but there are now five (Brook 1965, p.60), which are Northern, East Midland, West Midland, Southern and Kentish (Crystal 1995, p.50). The dialect used in extract two is Southern (Crystal 1995, p.36), which used the '-ing' suffix to express the continuous while '-and' was used in the Northern dialect (Crystal 1995, p.50). In line 12 the word 'plaiding' translates as 'pleading' and uses the '-ing' participle. Additionally, the Southern dialect used the third person pronoun 'hi', seen at the beginning of line 12) whereas the Northern and Midland dialects used 'they' (van Gelderen 2006, p.135).

## **Grammar**

The word order of Old English also varies from Modern English. Today English is a subject-verb-object (SVO) language, but in Old English, there is evidence of subject-object-verb (SOV) (Graddol *et al.* (eds.) 1996, p.115). An example of this is found in line two of extract one as ‘he næfre nænig leoð geleornode’, which translates directly as ‘he never any poem learned’ (Crystal 1995, p.20), showing a clear SOV structure. This is not the only word order found in Old English. In lines five and six of the same extract, the clause ‘þonne aras he for scome’ translates as ‘then arose he for shame’ (Crystal 1995, p.20). This shows a verb-subject order and is often found when adverbials introduce a main clause (Singh 2005, p.88).

In text two, *The Owl and the Nightingale* shows a subject-verb (SV) word order. An example of this can be seen in line one with ‘ich was in one sumere dale’. This translates directly as ‘I was in a summer valley’ which clearly shows the subject before the verb. However, as this extract progresses, the word order becomes more irregular and constructions such as ‘Iherde ich holde grete tale’ in line three, which translates directly as ‘heard I held great tale’, are used. This is because this extract is taken from a poem, therefore the need to keep the rhyming means that the word order is manipulated (Graddol *et al.*, (eds.) 1996, p.124).

Old English was able to have more freedom with word order due to the inflections in the language, which carried the meaning (Bragg 2003, p.24). Although Modern English uses inflections, Old English inflections were more complex as there were many different forms depending on case, number, person and gender (Brinton and Arnovick 2006, p.183). In Modern English, personal pronouns are still affected by these same things but there are less forms and the main difference concerns number. Modern English only differentiates between singular and plural (Crystal 1995, p.210) whereas Old English used to have dual forms to refer to two people (Crystal 1995, p.21) when speaking in the first or second person (Smith 1999, p.72). Some forms of the third person singular pronouns are similar to Modern English. This can be seen with the personal pronoun ‘his’. In line one of extract one, ‘he’ is used as the subject of the first clause, the same as today. However, in line 12 the form has changed to ‘hine’ as Cædmon has become the direct object. In Modern English, ‘him’ would be used (Graddol *et al.*, (eds.) 1996, p.117).

During the Middle English period, personal pronouns still varied according to case, number, gender and person, but changes had occurred to this system (Singh 2005, p.119). The main changes include the loss of the dual forms, the combining of the dative and accusative known as the objective, and the introduction of Scandinavian third person plural pronouns (Brinton and Arnovick 2006, pp.274 – 275). The different dialects of Middle English resulted in different personal pronoun forms being used throughout the different regions (Singh 2005, p.119). In *The Owl and the Nightingale*, the third person plural pronoun forms are not yet fully developed as the Old English forms are still being used. This can be seen in line 12 of extract two as the word ‘hi’ translates as ‘they’. These Old English personal pronoun forms were used in the Southern dialect throughout the Middle English period. However, by the end of this period the Scandinavian forms of ‘they’, ‘their’ and ‘them’ were being used throughout the country (Crystal 1995, p.50).

Verbs in Old English are usually divided into strong or weak depending on how they formed the past participle or preterites (Singh 2005, pp.85 – 86). Weak verbs would generally add a dental suffix with the past participle also adding a ‘ge-’ prefix

(Brinton and Arnovick 2006, pp.207 – 208). This is seen with the word ‘geleornode’ from extract one line two, meaning ‘learned’, which adds the dental suffix ‘-od’.

By the Middle English period, verbs are largely unchanged as they are still divided into strong and weak forms, which inflect in similar ways with a few exceptions. (Singh 2005, p.121). Weak verbs still add a dental suffix, but they have lost the ‘ge-’ prefix on the past participle. Instead, they use a ‘y-’ or ‘i-’ prefix (Singh 2005, p.122). An example of this is seen in extract two, line three with the word ‘iherde’ meaning ‘heard’.

Adjectives in Old English inflected according to the number, case and gender of the noun they modified. Weak forms are used when determiners precede the adjective, and strong forms are used at all other times (Smith 1999, pp.53 – 54). Strong adjectives in the Old English period used eight different inflections (Brinton and Arnovick 2006, p.189). However, during the Middle English period, many of these inflections were lost (Baugh and Cable 2002, p.160). The change in inflections went through two phases, one in early Middle English, and another in late Middle English (Brinton and Arnovick 2006, p.269). The first change resulted in there only being four inflections on strong adjectives (Brinton and Arnovick 2006, p.269). An example of this can be seen with the plural nominative and accusative cases. In Old English, where the masculine, neuter and feminine of these cases used the inflections ‘e’, ‘u’ and ‘a’ respectively, only the ‘e’ was used for all genders by the early Middle English period. By the end of the Middle English period these inflections changed again as they were no longer marked by case or gender. As a result the only inflection used was the ‘-e’ for both strong and weak plural adjectives (Brinton and Arnovick 2006, p.269).

### **Discussion**

By comparing the two extracts, some of the changes that have occurred between the Old and Middle English periods can be instantly recognised. This is because words are instantly recognisable and there have been many changes over this 400 year period. Towards the end of the Old English period, it is estimated that there were over 50,000 words in the English language (Crystal 2004, p.162). Following the Norman Conquest over 80% of these words were replaced (Stockwell and Minkova 2001, p.46) with 10,000 new words from Norman French. It is estimated that 75% of these are still used today (Baugh and Cable 2002, p.178). The vocabulary of these two extracts has changed greatly in this short time period. In *The Owl and the Nightingale*, the alphabet has dramatically changed from Cædmon’s Story and many words become familiar when changed to the modern alphabet. Spelling is also starting to change, although it is still not regular.

The grammar of these extracts also clearly shows the language becoming modernised. In *The Owl and the Nightingale* the word order has become more regularised and pronouns are now used more often to compensate for the loss of inflections. The loss of many inflections in both verbs and adjectives sees the language becoming more modern and identifiable with Modern English.

Overall, many changes have occurred throughout this period as seen through the exploration of these extracts. Analysis of the Old English reveals that it is not as irregular as it may first appear, and there are rules and there is consistency, particularly in relation to its grammar. However, these rules are not found in Modern English and it is the lack of understanding that makes it appear difficult to modern

readers. The Middle English text begins to use Modern English constructions and as we are already familiar with these, the text becomes recognisable, despite the irregularities that are still occurring.

Following the Norman Conquest the language appeared to change dramatically, although in many cases this did not become immediately apparent. However, while Norman French had an effect on the language, it would be wrong to say that all these changes occurred due to the Normans. The loss of inflectional endings had already started during the late Old English period (Singh 2005, pp.103 – 104) and instead much of what we see above is due to natural language change. If the Normans had not invaded England then many of these changes would still have happened, although English would not have the rich vocabulary it later acquired.

## **Bibliography**

- BAKER, P. S. (2003) *Introduction to Old English*. Malden: Blackwell Publishing.
- BAUGH, A. C. and T. CABLE (2002) *A History of the English Language*. (5th ed.) Abingdon: Routledge.
- BEDE (1990) *Ecclesiastical History of the English People*. London: Penguin Books.
- BEDE'S WORLD (2003) *Academic: The Venerable Bede* [online]. Available: URL: <<http://bedesworld.co.uk/academic-bede.php>> [Access date: 14th April 2010].
- BRAGG, M. (2003) *The Adventure of English*. London: Hodder and Stroughton.
- BRINTON, L. J. and L. K. ARNOVICK (2006) *The English Language: A Linguistic History*. Don Mills: Oxford University Press.
- BROOK, G. L. (1965) *English Dialects*. (2nd edn.) London: Andre Deutsch.
- BURROW, J. A. and T. TURVILLE-PETRE (1996) *A Book of Middle English*. (2nd edn.) Oxford: Blackwell Publishers.
- CARTLIDGE, N. (ed.) (2001) *The Owl and the Nightingale: Text and Translation*. Exeter: University of Exeter Press.
- CRYSTAL, D. (1995) *The Cambridge Encyclopedia of the English Language*. London: BCA.
- CRYSTAL, D. (2004) *The Stories of English*. London: Penguin.
- CULPEPER, J. and D. ARCHER. (2009) The History of English Spelling. IN: CULPEPER, J., F. KATAMBA, P. KERSWILL, R. WODAK and T. MCENERY (eds.), *English Language: Description, Variation and Context*. Basingstoke: Palgrave Macmillan. 244 – 258.
- FREEBORN, D. (1998) *From Old English to Standard English*. (2nd edn.) London: Macmillan Press.
- GRADDOL, D., D. LEITH and J SWANN (eds.) (1996) *English: History, Diversity and Change*. Abingdon: Routledge.
- GRADDOL, D., D. LEITH, J. SWANN, M. RHYS and J. GILLEN (eds.) (2007) *Changing English*. Abingdon: Routledge.
- HARPER, D. (2010) *Online Etymology Dictionary*. Available: URL: <<http://www.etymonline.com/index.php>> [Access date: 7th May 2010].
- MCSPARREN, F. (2001) *Middle English Dictionary* [online]. Available: URL: <<http://quod.lib.umich.edu/m/med/>> [Access date: 27th May 2010].
- MAYHEW, A. L. and W. W. SKEAT (2005) *A Concise Dictionary of Middle English: From A.D. 1150 to 1580*. Clearwater: R. A. Sites Books.
- MIELNICZUK, L. (2004) *Bodleian Library, University of Oxford: MSS. Tanner*. Available:URL:<<http://www.bodley.ox.ac.uk/dept/scwmss/wmss/online/medieval/tanner/tanner.html>> [Access date: 19th March 2010].

- NORTH, R and J. ALLARD (eds.) (2007) *Beowulf and Other Stories*. Harlow: Pearson Education.
- OLD ENGLISH MADE EASY (2008) *Dictionaries*. Available: URL: <<http://home.comcast.net/~modean52/index.htm>> [Access date: 10th May 2010].
- OXFORD BODLEIAN LIBRARY (2004) MSS *Tanner*. Available: URL: <<http://www.bodley.ox.ac.uk/dept/scwmss/wmss/online/medieval/tanner/images/00103782.jpg>> [Access date: 21st May 2010].
- SINGH, I. (2005) *The History of English: A Student's Guide*. London: Hodder Arnold.
- SMITH, J. J. (1999) *Essentials of Early English*. London: Routledge.
- STOCKWELL, R. and D. MINKOVA, (2001) *English Words: History and Structure*. Cambridge: Cambridge University Press.
- STONE, B. (1971) *The Owl and the Nightingale / Cleanness / St Erkenwald*. Harmondsworth: Penguin Books.
- TREHARNE, E. (2004) *Old and Middle English c.890 – c.1400: An Anthology*. (2nd edn.) Malden: Blackwell Publishing.
- VAN GELDEREN, E. (2006) *A History of the English Language*. Amsterdam: John Benjamins Publishing.
- WELLS, J. E. (ed.) (1907) *The Owl and the Nightingale*. Boston: D. C. Heath and Co.

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

Vlad Mackevic

Aston University, UK  
[mackeviv@aston.ac.uk](mailto: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 than 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 is 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 at 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

Marker	123		124		134		234	
	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)

<b>Criterion</b>	<b>Meaning/explanation</b>
A ratio: <sup>1</sup> Number of test results where <u>Probability of Same Authorship is above 50%</u> All test results	50% certainty is a civil court benchmark
A ratio: Number of test results where <u>Probability of Same Authorship is above 80%</u> 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 <u>Probability of Same Authorship is below 40%</u> All test results	40% is my subjective ‘failure mark’ <b>(Unlikely)</b> that the texts have been written by the same author)
A ratio: <sup>2</sup> Number of test results where <u>Probability of Same Authorship is below 20%</u> All test results	<b>Very unlikely</b> that the texts have been written by the same author
A ratio: Number of test results where <u>Probability of Same Authorship is below 10%</u> All test results	<b>Extremely unlikely</b> that the texts have been written by the same author
A ratio: Number of test results where <u>Probability of Same Authorship is below 5%</u> All test results	‘Beyond Reasonable Doubt’

<sup>1</sup> All ratios are expressed as percentages.

<sup>2</sup> 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

<b>Marker</b>	<b>Comments</b>
<b>TO</b>	Very reliable only when PSA $\geq$ 40%. As soon as PSA bar drops, error rate shoots up; received 100% error rate under the strict conditions (PSA $\geq$ 10%) once.
<b>OF</b>	Moderately reliable. Performs well until the point when PSA reaches $\geq$ 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%.
<b>AS</b>	Not very reliable. Error rates reach 25% when PSA $\geq$ 40% and rise to 33%-50% at PSA $\geq$ 20 stage. However, performed very well in tests for Author D.
<b>THAT</b>	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.
<b>LR</b>	Excellent marker; very consistent – error rate rose to 25% once, under the strictest conditions (PSA $\geq$ 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  $\geq 50$ . 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

<b>Author E</b>	<b>ES1</b>	<b>ES2</b>	<b>ES3</b>		<b>EL1</b>	<b>EL2</b>
<b>Word Count</b>	590	611	599		870	911
<b>Author F</b>	<b>FS1</b>	<b>FS2</b>	<b>FS3</b>		<b>FL1</b>	<b>FL2</b>
<b>Word Count</b>	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	<b>46.8041%</b>	<b>39.7300%</b>
IT	10.3521%	6.1325%
THAT	16.0625%	9.9243%
THEIR	<b>20.9488%</b>	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. Philadelphia, 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	<a href="http://neon.niederlandistik.fu-berlin.de/textstat/">http://neon.niederlandistik.fu-berlin.de/textstat/</a>
ANOVA Calculator	<a href="http://www.danielsoper.com/statcalc/calc43.aspx">http://www.danielsoper.com/statcalc/calc43.aspx</a>
T-test Calculator	<a href="http://www.graphpad.com/quickcalcs/OneSampleT1.cfm">http://www.graphpad.com/quickcalcs/OneSampleT1.cfm</a>
Grade Guru	<a href="http://www.gradeguru.com">www.gradeguru.com</a>

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

Table 1: reliability of markers for Author A

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

Table 2: reliability of markers for Author B

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

Table 3: reliability of markers for Author C

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

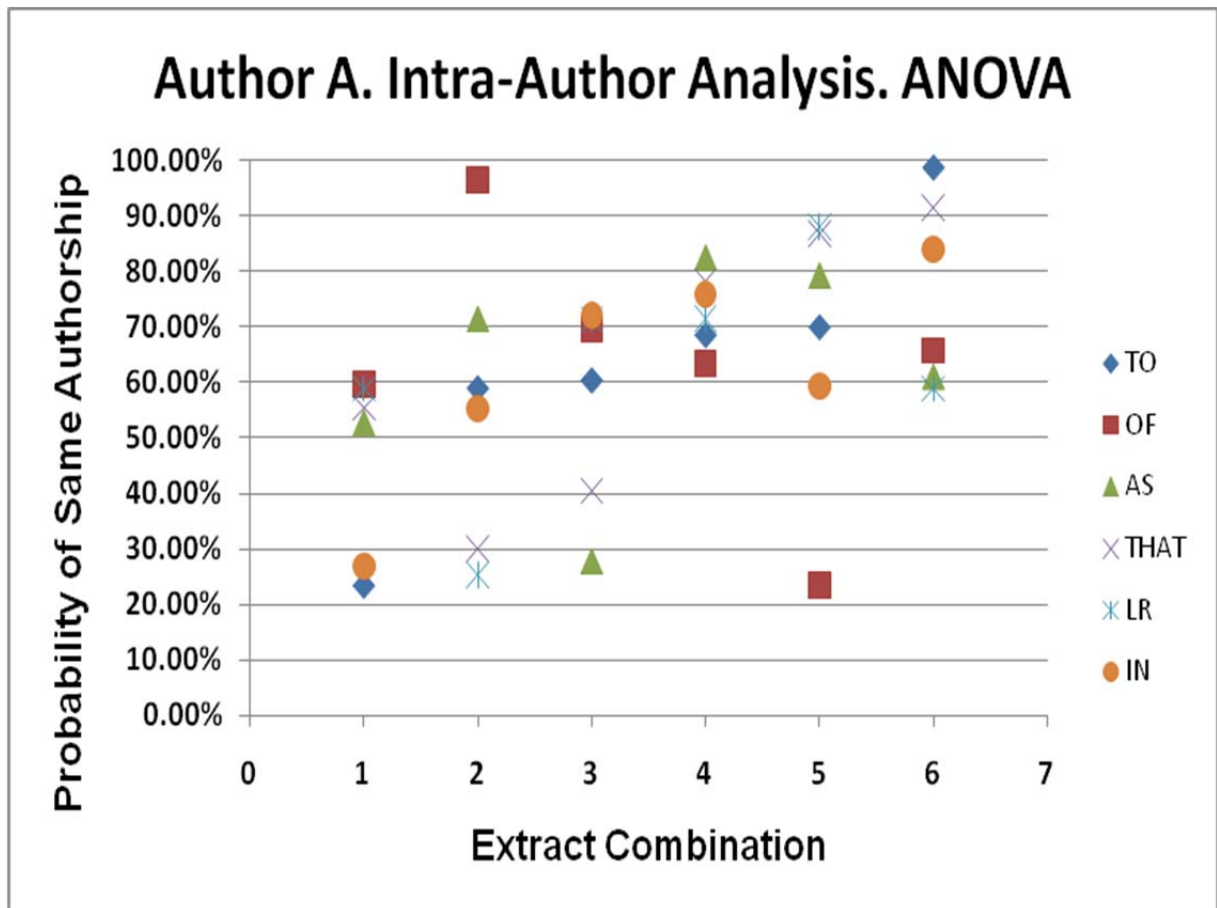
Table 4: reliability of markers for Author D

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

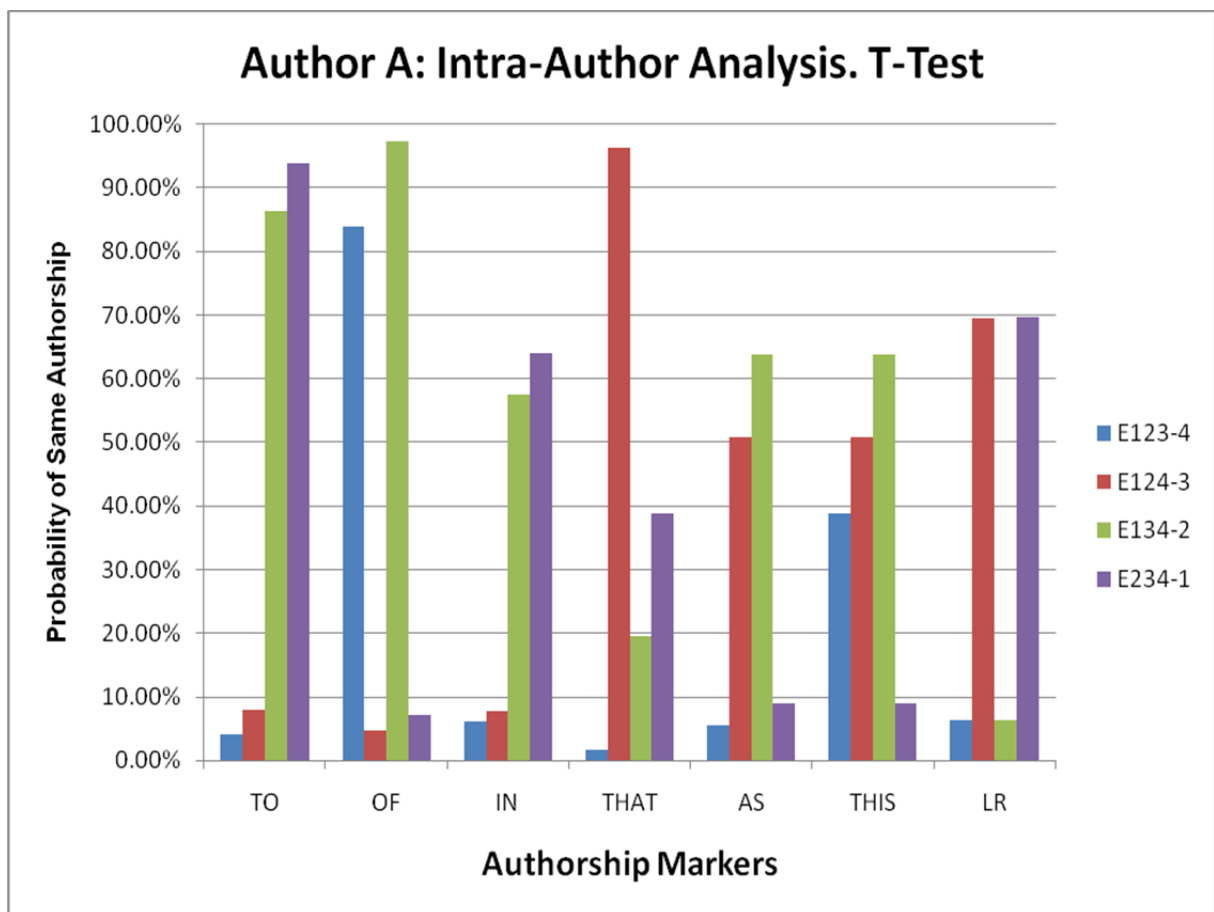
<b>TO</b>	3.9139%	3.7736%	3.5928%	3.0055%	3.5714%	0.3996%	<b>11.1874%</b>
<b>THAT</b>	2.3483%	3.7736%	1.7964%	3.5519%	2.8676%	0.9498%	33.1211%
<b>AS</b>	1.1742%	0.9434%	1.7964%	1.6393%	1.3883%	0.3972%	28.6118%
<b>IT</b>	1.1742%	0.9434%	0.8982%	0.8197%	0.9589%	0.1524%	<b>15.8907%</b>
<b>LR</b>	38.5519%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

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.

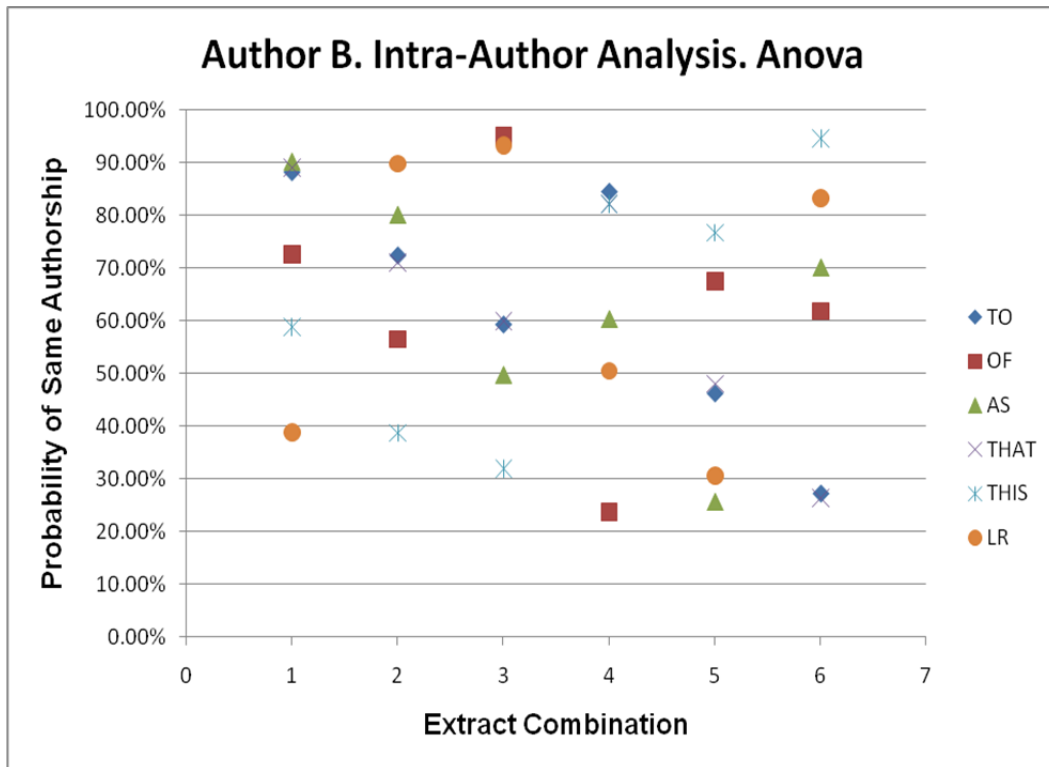
### 1. ANOVA and T-test Analysis



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

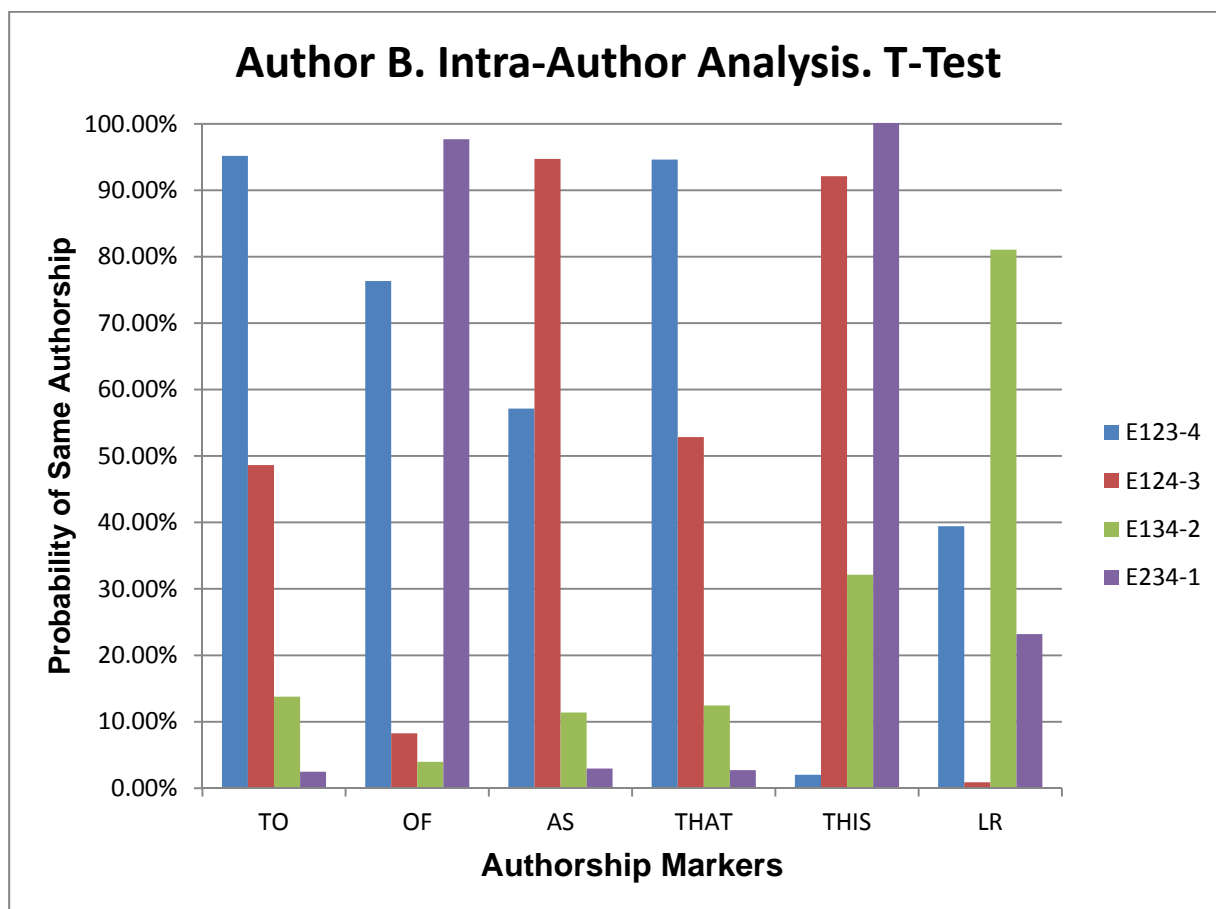


	Ratio	Percentage
<b>Below Fail Mark (40%)</b>	<b>15/28</b>	<b>53.57%</b>
<b>Above Success Mark (80%)</b>	<b>5/28</b>	<b>17.86%</b>
<b>Above Civil Court Benchmark (50%)</b>	<b>11/28</b>	<b>39.29%</b>

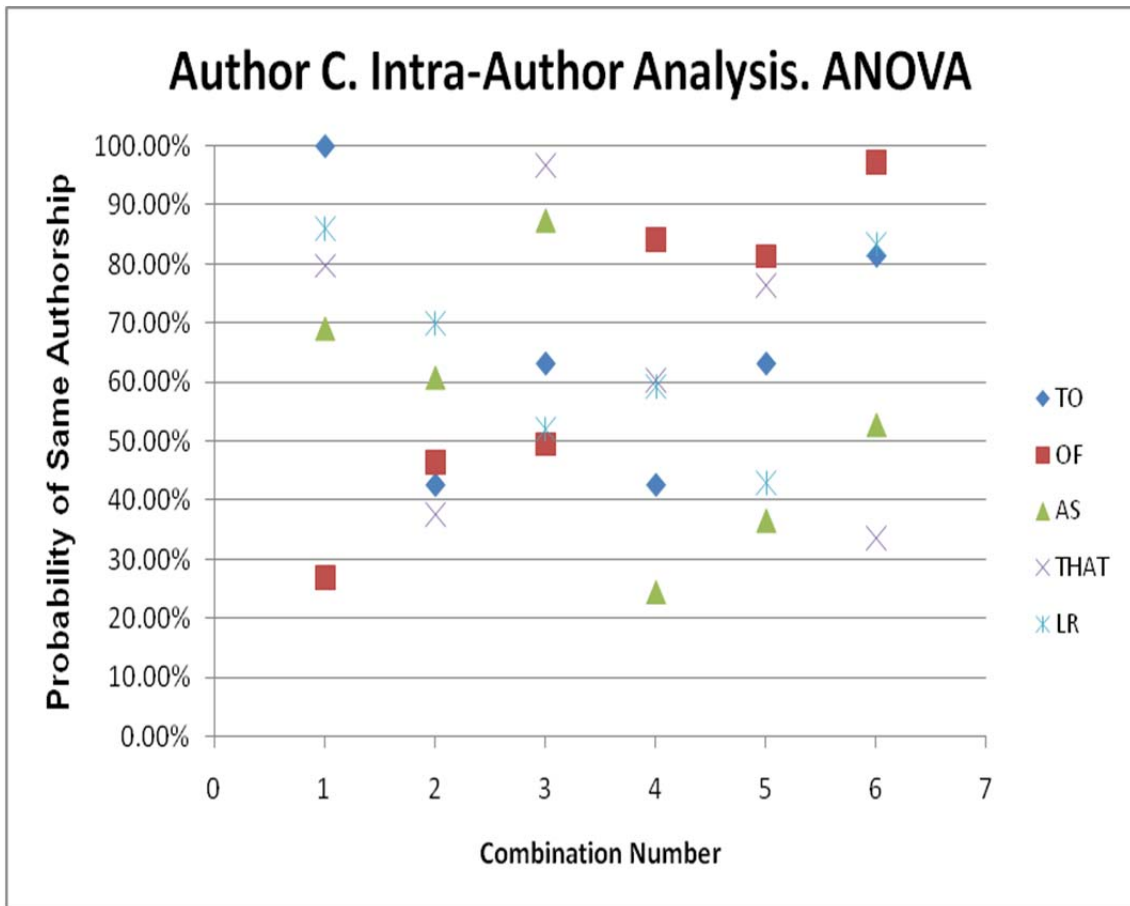


	<b>Ratio</b>	<b>Percentage</b>
<b>Below Fail Mark (40%)</b>	<b>8/36</b>	<b>22.22%</b>
<b>Above Success Mark (80%)</b>	<b>11/36</b>	<b>30.56%</b>
<b>Above Civil Court Benchmark (50%)</b>	<b>24/36</b>	<b>66.67%</b>

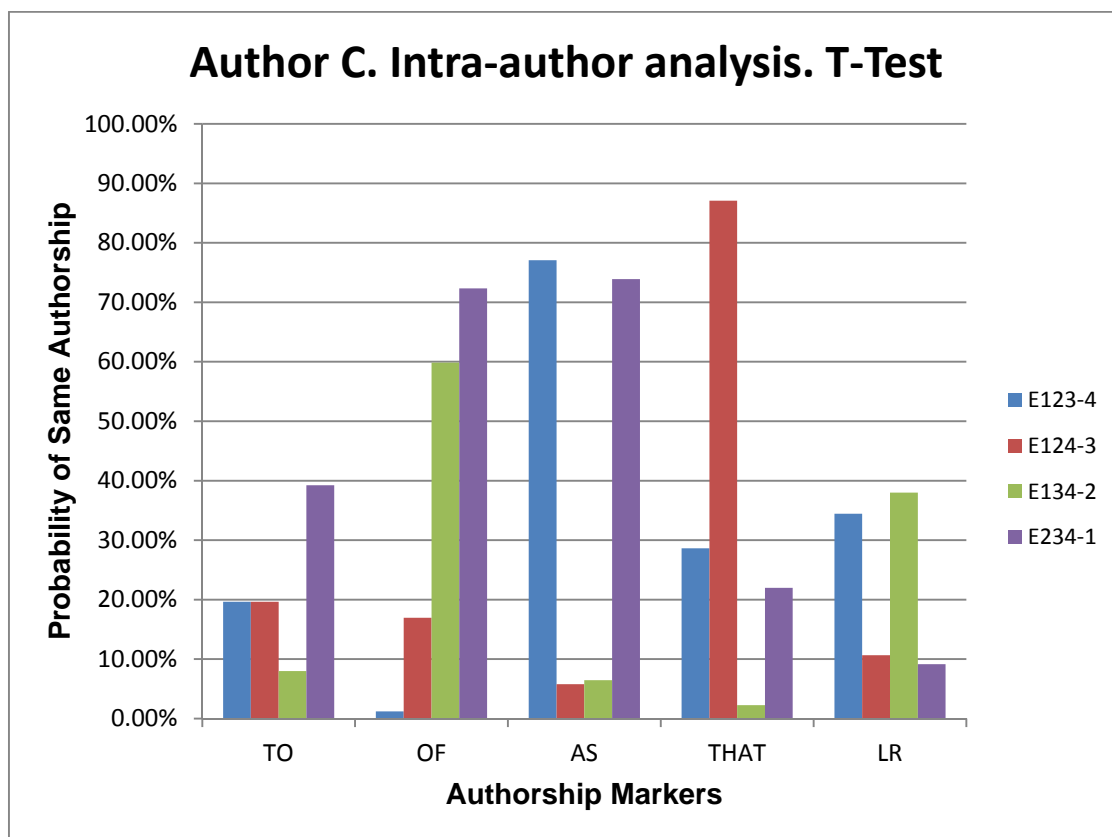




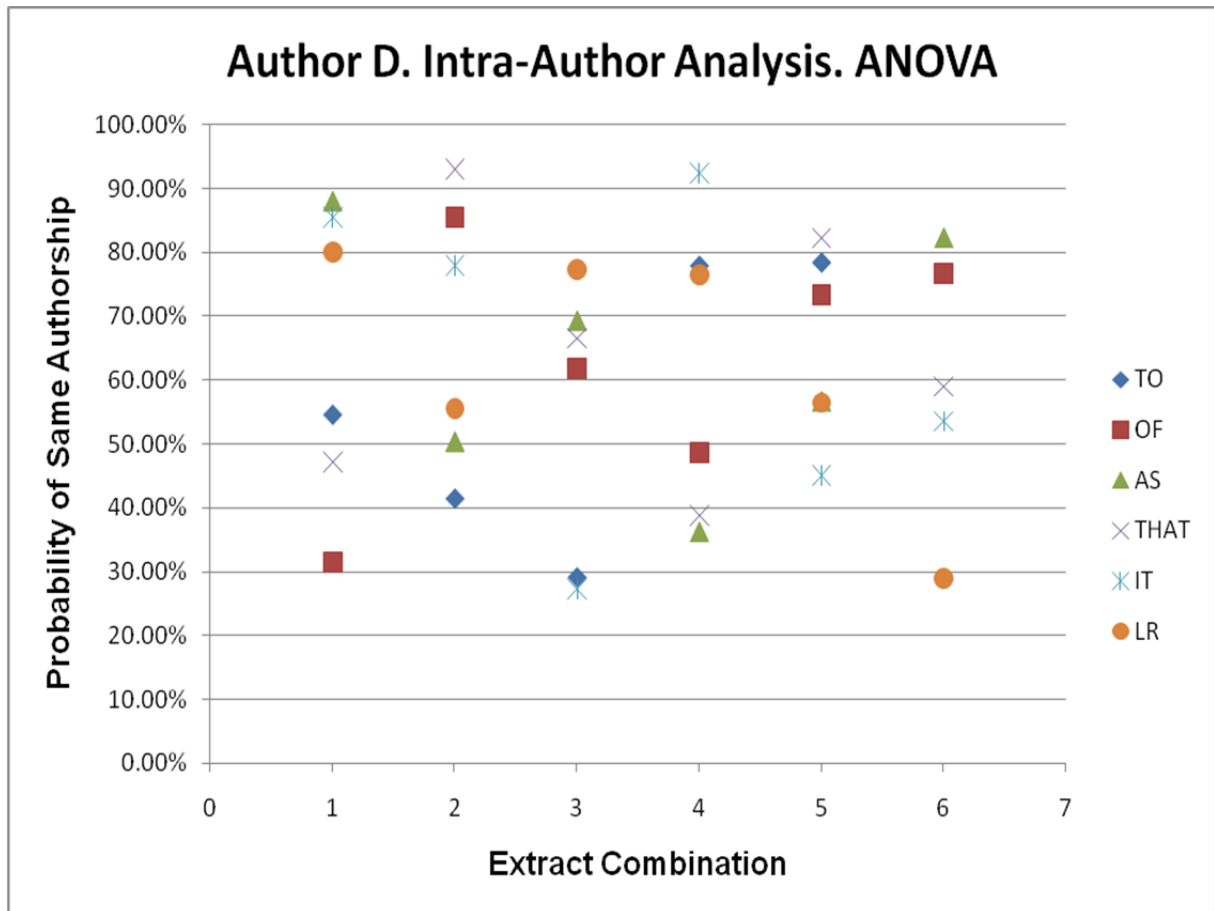
	Ratio	Percentage
<b>Below Fail Mark (40%)</b>	<b>13/24</b>	<b>54.17%</b>
<b>Above Success Mark (80%)</b>	<b>6/24</b>	<b>25%</b>
<b>Above Civil Court Benchmark (50%)</b>	<b>10/24</b>	<b>41.67%</b>



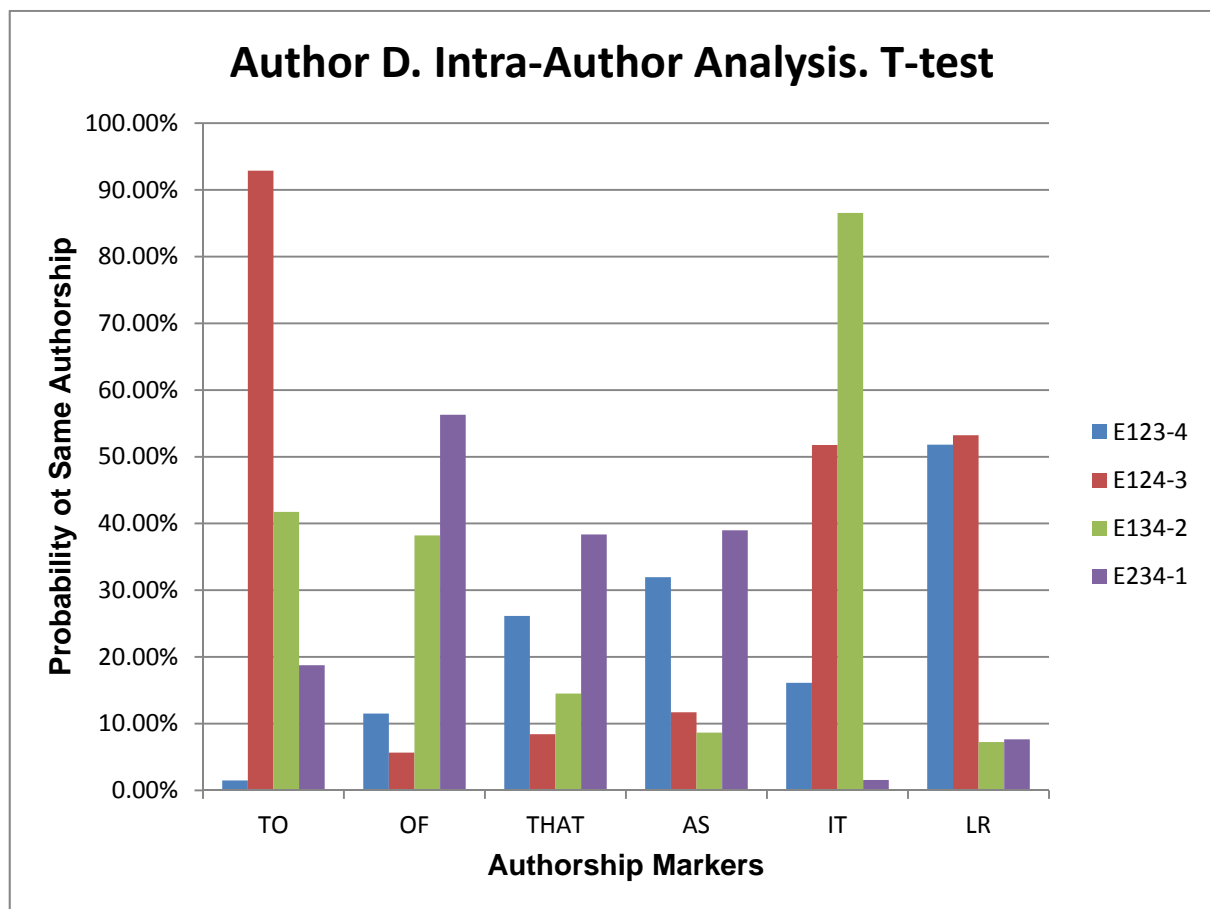
	<b>Ratio</b>	<b>Percentage</b>
<b>Below Fail Mark (40%)</b>	<b>5/30</b>	<b>16.67%</b>
<b>Above Success Mark (80%)</b>	<b>10/30</b>	<b>33.33%</b>
<b>Above Civil Court Benchmark (50%)</b>	<b>20/30</b>	<b>66.67%</b>



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

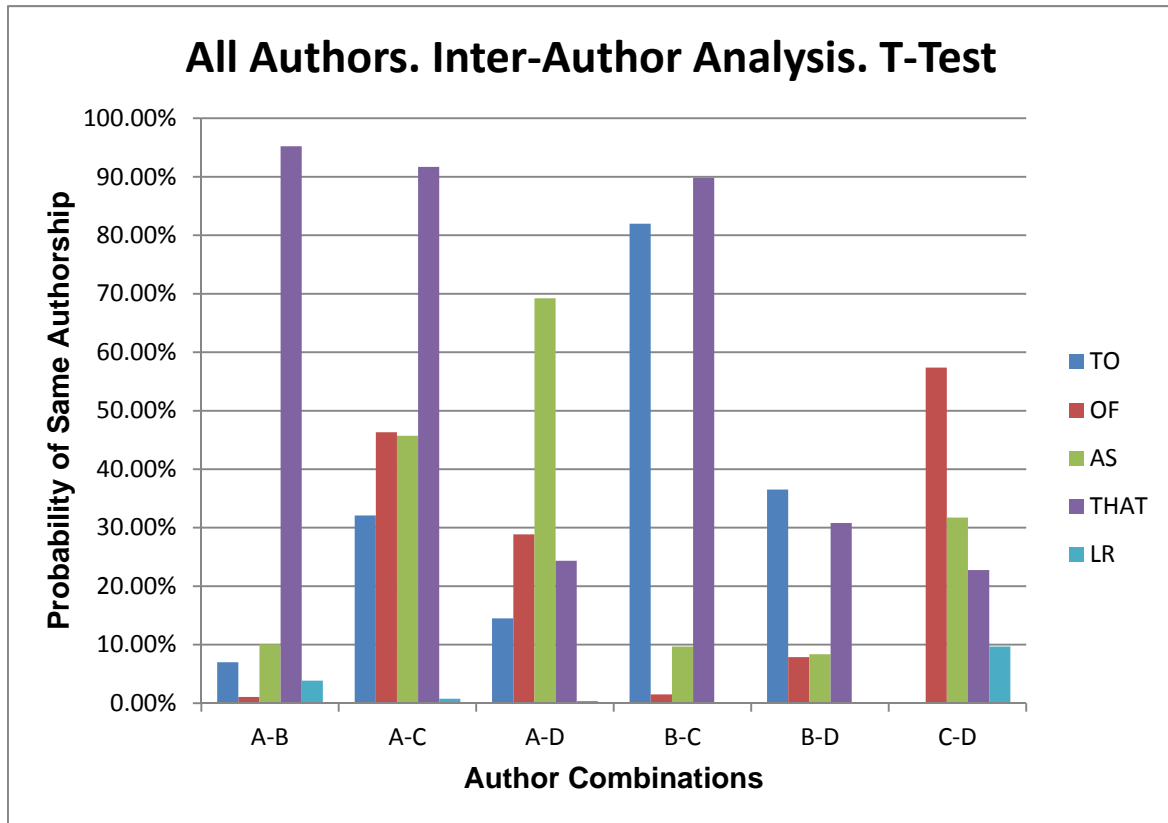


	Ratio	Percentage
<b>Below Fail Mark (40%)</b>	<b>6/36</b>	<b>16.67%</b>
<b>Above Success Mark (80%)</b>	<b>7/36</b>	<b>19.44%</b>
<b>Above Civil Court Benchmark (50%)</b>	<b>23/36</b>	<b>63.89%</b>



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

**Appendix 2: Inter-Author Analysis.**



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

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

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

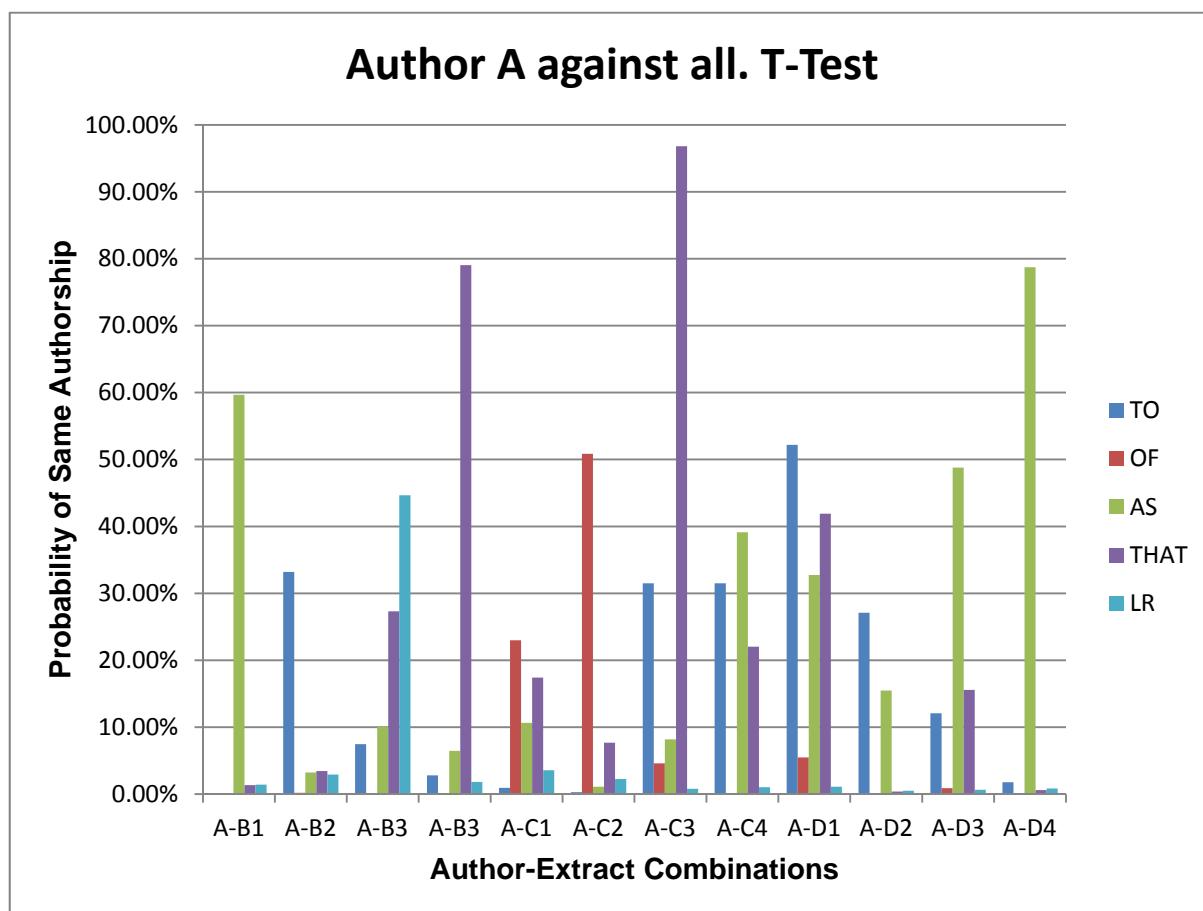
**Table 6: Error rates for each marker:**

<b>TO</b>	<b>OF</b>	<b>AS</b>	<b>THAT</b>	<b>LR</b>
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.

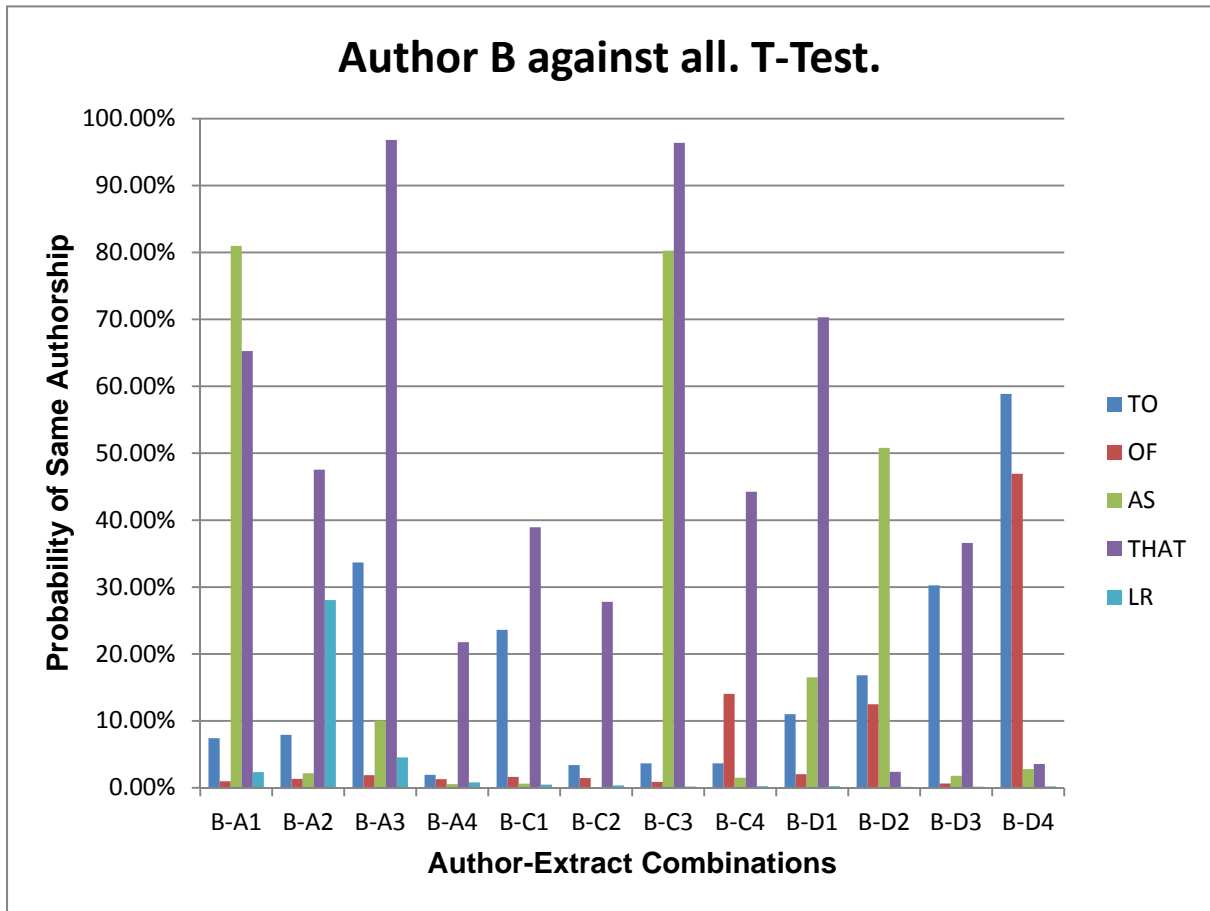


<b>Overall Successful Differentiation (below 40%)</b>	<b>51/60</b>	<b>85%</b>
<b>Successful Differentiation below 20%</b>	<b>42/60</b>	<b>70%</b>
<b>Definite Failure (over 80%)</b>	<b>1/60</b>	<b>1.67%</b>
<b>Failed to Differentiate (over 50%)</b>	<b>6/60</b>	<b>10%</b>



Author A against all: Table of probabilities

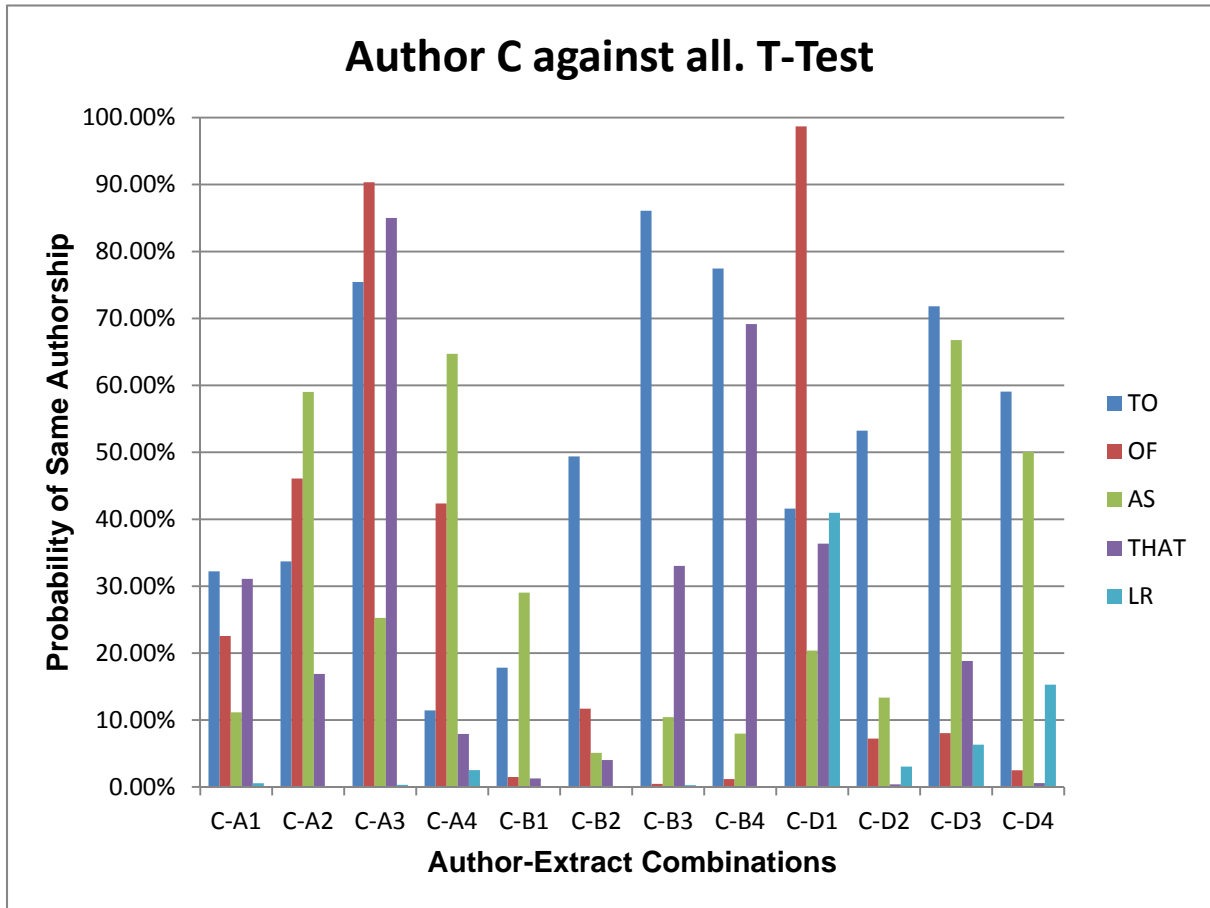
	TO	P<=0.2	P<=0.4	OF	P<=0.2	P<=0.4	AS	P<=0.2	P<=0.4	THAT	P<=0.2	P<=0.4	LR	P<=0.2	P<=0.4
A-B1	0.0004	1	1	0.0002	1	1	<b>0.5967</b>	0	0	0.0132	1	1	0.0141	1	1
A-B2	<b>0.3319</b>	0	1	0.0018	1	1	0.0322	1	1	0.0343	1	1	0.029	1	1
A-B3	0.0745	1	1	0.0001	1	1	0.1009	1	1	<b>0.2732</b>	0	1	<b>0.4466</b>	0	0
A-B3	0.0279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



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

Author B against all. Table of Probabilities

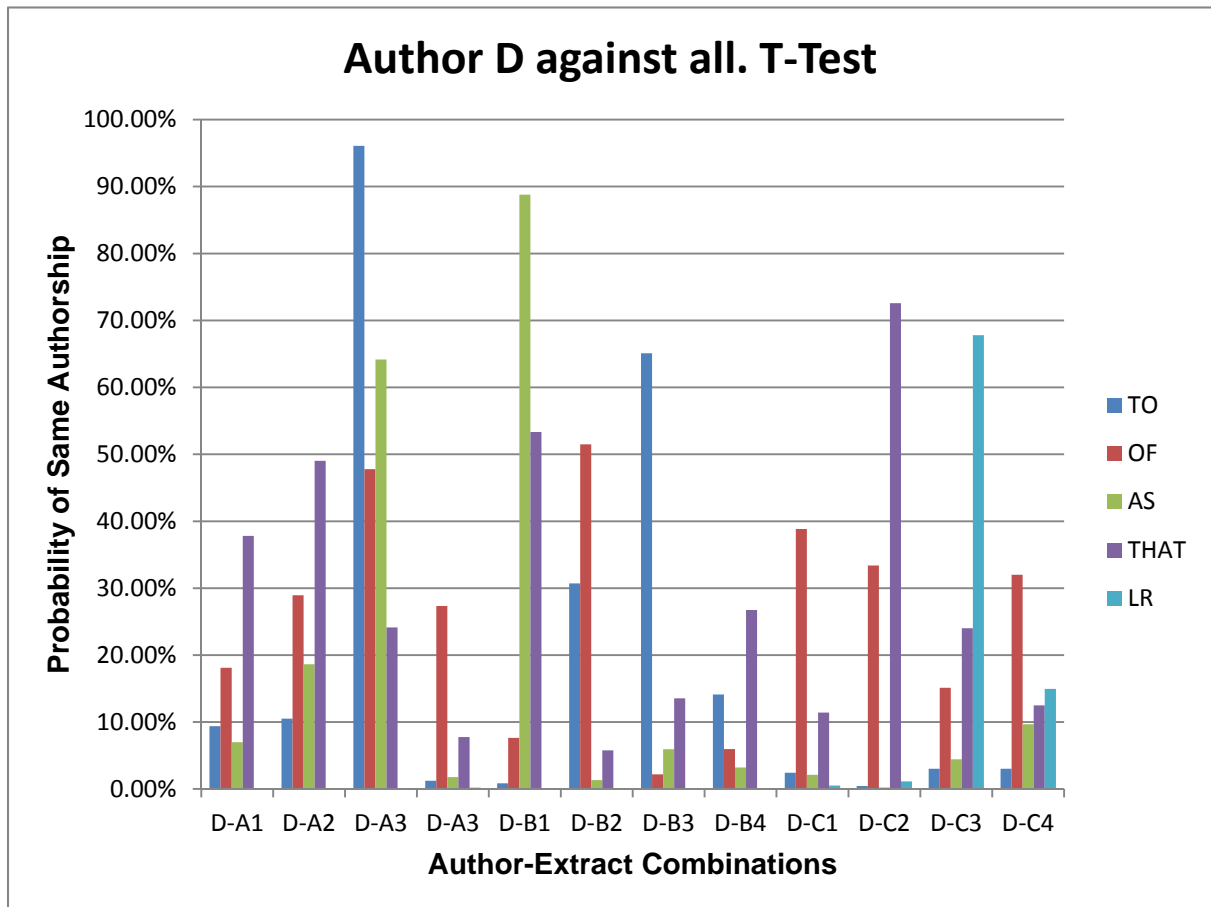
	TO	P<=0.2	P<=0.4	OF	P<=0.2	P<=0.4	AS	P<=0.2	P<=0.4	THAT	P<=0.2	P<=0.4	LR	P<=0.2	P<=0.4
<b>B-A1</b>	0.074	1	1	0.0098	1	1	<b>0.8099</b>	0	0	<b>0.6527</b>	0	0	0.0233	1	
<b>B-A2</b>	0.0792	1	1	0.0132	1	1	0.0216	1	1	<b>0.4754</b>	0	0	<b>0.2804</b>	0	
<b>B-A3</b>	<b>0.3367</b>	0	1	0.0188	1	1	0.1003	1	1	<b>0.9683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



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

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



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

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	<b>0.3781</b>	0	1	0.0006	1	1
D-A2	0.1051	1	1	<b>0.2894</b>	0	1	0.1863	1	1	<b>0.4902</b>	0	0	0.0002	1	1
D-A3	<b>0.9608</b>	0	0	<b>0.4778</b>	0	0	<b>0.6416</b>	0	0	<b>0.2414</b>	0	1	0.0004	1	1
D-A3	0.0122	1	1	<b>0.2735</b>	0	1	0.0178	1	1	0.0775	1	1	0.0019	1	1
D-B1	0.0084	1	1	0.0763	1	1	<b>0.888</b>	0	0	<b>0.5334</b>	0	0	0.0001	1	1
D-B2	<b>0.307</b>	0	1	<b>0.5149</b>	0	0	0.0132	1	1	0.0576	1	1	0.0001	1	1
D-B3	<b>0.6509</b>	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	<b>0.2673</b>	0	1	0.0001	1	1
D-C1	0.0242	1	1	<b>0.3885</b>	0	1	0.0211	1	1	0.1141	1	1	0.0052	1	1
D-C2	0.0044	1	1	<b>0.3338</b>	0	1	0.0022	1	1	<b>0.7257</b>	0	0	0.0113	1	1
D-C3	0.0302	1	1	0.1511	1	1	0.0443	1	1	<b>0.2402</b>	0	1	<b>0.678</b>	0	0
D-C4	0.0302	1	1	<b>0.3199</b>	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		<b>75.00%</b>	83.33%		<b>41.67%</b>	83.33%		83.33%	83.33%		<b>41.67%</b>	75.00%		91.67%	91.67%
Error Rate		<b>25.00%</b>	16.67%		<b>58.33%</b>	16.67%		16.67%	16.67%		<b>58.33%</b>	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		<b>50.00%</b>	58.33%		<b>8.33%</b>	25.00%		<b>50.00%</b>	75.00%		<b>0.00%</b>	<b>16.67%</b>		83.33%	83.33%
Error Rate		<b>50.00%</b>	41.67%		<b>91.67%</b>	75.00%		<b>50.00%</b>	25.00%		<b>100.00%</b>	<b>83.33%</b>		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	<b>3.0508%</b>	<b>0.8183%</b>	<b>1.8364%</b>	<b>1.9018%</b>	<b>1.1177%</b>	<b>58.7689%</b>
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%
<b>THEIR</b>	<b>0.8475%</b>	<b>1.4730%</b>	<b>1.3356%</b>	<b>1.2187%</b>	<b>0.3287%</b>	<b>26.9736%</b>
<b>THIS</b>	<b>0.8475%</b>	<b>1.6367%</b>	<b>1.5025%</b>	<b>1.3289%</b>	<b>0.4223%</b>	<b>31.7759%</b>
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%
<b>IN</b>	<b>2.3052%</b>	<b>1.4943%</b>	<b>1.8998%</b>	<b>0.5734%</b>	<b>30.1825%</b>
THAT	2.0856%	1.7241%	1.9049%	0.2556%	13.4194%
OF	1.7563%	1.9540%	1.8552%	0.1398%	7.5355%
<b>IF</b>	<b>1.0977%</b>	<b>0.5747%</b>	<b>0.8362%</b>	<b>0.3698%</b>	<b>44.2259%</b>
ON	0.9879%	0.8046%	0.8963%	0.1296%	14.4617%
<b>THEIR</b>	<b>0.9879%</b>	<b>1.4943%</b>	<b>1.2411%</b>	<b>0.3581%</b>	<b>28.8517%</b>
<b>THIS</b>	<b>0.9879%</b>	<b>1.7241%</b>	<b>1.3560%</b>	<b>0.5206%</b>	<b>38.3903%</b>
<b>FROM</b>	<b>0.7684%</b>	<b>0.3448%</b>	<b>0.5566%</b>	<b>0.2995%</b>	<b>53.8143%</b>
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	<b>3.0508%</b>	<b>0.8183%</b>	<b>1.8364%</b>	<b>2.3052%</b>	<b>1.4943%</b>	<b>1.9010%</b>	<b>0.8407%</b>	<b>44.2251%</b>
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	<b>1.1864%</b>	<b>0.8183%</b>	<b>1.3356%</b>	<b>1.0977%</b>	<b>0.5747%</b>	<b>1.0025%</b>	<b>0.3045%</b>	<b>30.3687%</b>
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	<b>0.8475%</b>	<b>1.6367%</b>	<b>1.5025%</b>	<b>0.9879%</b>	<b>1.7241%</b>	<b>1.3397%</b>	<b>0.3964%</b>	<b>29.5871%</b>
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%
<b>IN</b>	<b>2.5097%</b>	<b>2.8881%</b>	<b>1.5101%</b>	<b>2.3026%</b>	<b>0.7120%</b>	<b>30.9191%</b>
THAT	2.3166%	1.9856%	2.0134%	2.1052%	0.1836%	8.7215%
<b>IT</b>	<b>1.7375%</b>	<b>1.6245%</b>	<b>2.5168%</b>	<b>1.9596%</b>	<b>0.4858%</b>	<b>24.7931%</b>
THEIR	1.1583%	0.9025%	0.8389%	0.9666%	0.1691%	17.4911%
<b>FOR</b>	<b>0.9653%</b>	<b>0.3610%</b>	<b>0.6711%</b>	<b>0.6658%</b>	<b>0.3022%</b>	<b>45.3867%</b>
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%
<b>IS</b>	<b>3.3659%</b>	<b>2.4096%</b>	<b>2.8878%</b>	<b>0.6762%</b>	<b>23.4164%</b>
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%
<b>AS</b>	<b>1.6287%</b>	<b>0.9371%</b>	<b>1.2829%</b>	<b>0.4890%</b>	<b>38.1195%</b>
<b>FOR</b>	<b>0.4343%</b>	<b>0.9371%</b>	<b>0.6857%</b>	<b>0.3555%</b>	<b>51.8497%</b>
THEIR	0.9772%	0.9371%	0.9572%	0.0284%	2.9624%
<b>THERE</b>	<b>0.5429%</b>	<b>0.8032%</b>	<b>0.6731%</b>	<b>0.1841%</b>	<b>27.3471%</b>
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%
<b>IN</b>	<b>2.5097%</b>	<b>2.8881%</b>	<b>1.5101%</b>	<b>2.0630%</b>	<b>2.5435%</b>	<b>2.3029%</b>	<b>0.5313%</b>	<b>23.0719%</b>
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%
<b>FOR</b>	<b>0.9653%</b>	<b>0.3610%</b>	<b>0.6711%</b>	<b>0.4343%</b>	<b>0.9371%</b>	<b>0.6738%</b>	<b>0.2782%</b>	<b>41.2860%</b>
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.**

## **Instructions for authors**

Début is published by the UK Subject Centre for Languages, Linguistics and Area Studies. We work to support teaching and learning in our subject areas.

Début welcomes scholarly papers written by undergraduate students in languages, linguistics and area studies. Each paper is reviewed by two academics or postgraduate students with expertise in the subject of your paper. Papers could be based on dissertations, (including year abroad projects), class projects or your own independent research. All papers should reference appropriate existing scholarship. There is now a Facebook group for Début where you can discuss preparing your articles and ask questions.

Début welcomes scholarly papers written by undergraduate students in languages, linguistics and area studies. Those who have graduated in the past 12 months may also submit work. Each paper is reviewed by academics or postgraduate students with expertise in the subject of your paper. Papers could be based on dissertations, (including year abroad projects), class projects or your own independent research. All papers should reference appropriate existing scholarship.

Papers (including references) should be 3000-5000 words and use the Harvard referencing style <http://www.llas.ac.uk/resourcedownloads/3088/harvardcitations.pdf> The first page should include your name, postal and email addresses. Please include an abstract of up to 150 words. Your work should be submitted to the editor, John Canning by email at [j.canning@soton.ac.uk](mailto:j.canning@soton.ac.uk) in Microsoft Word (.doc) or Rich Text Format (.rtf). The language of the journal is English. Articles are considered for publication in Début on the understanding that they are not being considered for publication in any other outlet.

If your paper requires the use of phonetic symbols, please consult the UCL Department of Speech, Hearing and Phonetic Sciences webpage <http://www.phon.ucl.ac.uk/resource/phonetics/> This will ensure that your paper can be read by the editor and the reviewers.

Authors submit their work to Début on the understanding that the Subject Centre may publish their work online or in print form for non-commercial purposes. Authors retain the copyright to their work. It is the author's responsibility to ensure that permissions are obtained for objects that appear in the article, including screenshots, photographs, audio/video files and diagrams.

### ***Checklist***

Before submitting your article, please check the following:

- Your article does not exceed 5000 words
- If your article is based on a class essay or dissertation that all references to module number/ lecturer name etc. are removed
- Your article does not include materials copyrighted to other people (this could include pictures and photocopies from other books/ articles)
- That your references are formatted according to the Début style
- That your article does not include any appendices