Whose Tweet? Authorship analysis of micro-blogs and other short-form messages

Nicci MacLeod[1] *and Tim Grant*[2]
Centre for Forensic Linguistics, Aston University, UK
[1]*n.macleod1@aston.ac.uk*
[2]*t.d.grant@aston.ac.uk*

Abstract

Approaches to authorship attribution have traditionally been constrained by the size of the message to which they can be successfully applied, making them unsuitable for analysing shorter messages such as SMS Text Messages, micro-blogs (e.g. Twitter) or Instant Messaging. Having many potential authors of a number of texts (as in, for example, an online context) has also proved problematic for traditional descriptive methods, which have tended to be successfully applied in cases where there is a small and closed set of possible authors.

This paper reports the findings of a project which aimed to develop and automate techniques from forensic linguistics that have been successfully applied to the analysis of short message content in criminal cases. Using data drawn from UK-focused online groups within Twitter, the research extends the applicability of Grant's (2007; 2010) stylistic and statistical techniques for the analysis of authorship of short texts into the online environment. Initial identification of distinctive textual features commonly found within short messages allows for the development of a taxonomy which can then be used when calculating the 'distance' between messages containing instances of these feature types. The end result is an automated process with a high level of success in assigning tweets to the correct author. The research has the potential to extend the scope of reliable and valid authorship analysis into hitherto unexplored contexts. Given the relative anonymity of the internet and the availability of cloaking technology, linguistic research of this nature represents a crucial contribution to the investigative toolkit.

Keywords: AUTHORSHIP ANALYSIS; STYLISTIC METHODS; STATISTICAL METHODS; ONLINE MESSAGING

1. Introduction

It has been widely noted that there is increasing use of online communications for the organisation and dissemination of a wide range of criminal activities and material. The fundamental anonymity offered by the internet and the ease with which multiple identities can be created enables individuals to share such information in relative security. State-of-the-art work in authorship analysis has had considerable success for cases where there is a small and known set of authors, and sufficient quantity of text of known authorship. These methods do not easily translate into computer-mediated communication where there may be a large and unknown number of authors all contributing an unknown number of short messages. This paper reports on a project that extended existing work in forensic linguistics that had been successfully applied (at evidential standard) to criminal investigations involving SMS text messages, by developing an automated process that can be applied to online environments by non-specialist users.

Recent attempts to develop methods for attributing authorship have emerged from two broad disciplines—linguistics (e.g. Chaski, 2001; Grant & Baker, 2001; McMenamin 1993; 2002) and computing (e.g. Argamon, 2008; Hoover, 2003; Koppel *et al.* 2006; 2011). Traditionally concerned with literary, biblical and political texts, interest has shifted in recent times to the identification of authors of shorter texts such as blogs (Koppel *et al.* 2011) and SMS texts (Grant, 2010). As Zheng *et al.* (2005) point out, the misuse of online messages for inappropriate and/or illegal purposes has become a serious concern in recent times. Aside from the ease of anonymity for online authors and the brevity of the texts, difficulties in establishing robust methods have been compounded by the large and open ended nature of the set of potential authors in this context. Online texts are 'shorter, noisier and they have a greater number of candidate authors' (Abbasi & Chen, 2005: 67).

Features such as relative frequencies of function words and word frequency distributions have traditionally been brought together in multivariate models for attributing authorship, and indeed the individual's variation in their use of function words remains a popular method to this day (Grant & Baker, 2001). Other researchers in the area (e.g. Miranda-García & Calle-Martín, 2005; Smith & Kelly, 2002) have had some success with lexical richness (the frequency of rare words, e.g. *hapax legomena* and *hapax dislegomena*) and repetition (the frequency of common words). Furthermore, average word, sentence, clause and paragraph lengths, word type frequencies and distributions, collocation and content analysis have all been utilised for the task, although it has been noted that these are often used in combination for maximal discriminatory power—identification is achieved through an aggregate of markers (Grant & Baker, 2001; McMenamin, 2001). Chaski's (2001) approach, although not without its critics on account of some significant methodological weaknesses (e.g. Grant & Baker, 2001; McMenamin, 2001) tested a number of features for authorship analysis, including syntactic analysis, syntactically classified punctuation, sentential complexity, vocabulary richness, readability, content analysis, spelling errors, punctuation errors, word form errors, and grammatical errors, and found that only syntactic analysis and syntactically classified punctuation successfully discriminated and clustered documents.

Koppel *et al.* (2011) note that almost all existing research in the field of authorship attribution 'considers only the simplest version of the problem' (p. 84), that is to say, those instances where a relatively long anonymous text is attributed to one of a small, closed set of candidates. As they point out, this version of the authorship attribution is rare in the real world—conversely, we are often faced with the potential of thousands of candidate authors; the possibility that none of the known candidates authored the text; and the likelihood that either the known texts and/or anonymous text may be limited. Addressing these limitations, Koppel *et al.* (2006) report on their own technique for solving authorship attribution even when the candidate set numbers in the many thousands. With a test candidate set of 10,000 bloggers, they aim to determine which individual authored a given 500 word snippet. Their approach involves determining whether a given snippet includes a set of linguistic features unique to a given author. Their results showed that this rather crude approach worked to a certain extent, but that only when a response of *Don't Know* was permissible was the method able to achieve reasonably reliable attribution of snippets in the case of thousands of authors.

Koppel *et al.* (2011) describe existing methods for automated authorship attribution as falling into two paradigms—the *similarity based* paradigm, where the distance between two documents and an anonymous document is measured, and attribution is based on the author whose known writing has more in common with the questioned text; and the *machine learning* paradigm, where the known writings of each candidate author are used to construct a

classifier, which is then used to classify anonymous documents. The authors point out that *similarity based* methods are more appropriate when considering a large volume of candidate authors (Koppel *et al.*, 2006), and that using these methods allows for a document to be verified as having been written by a given author 'if the similarity between the document and the author's known writing exceeds some threshold' (2011: 85). They take 4-grams (strings of characters of length four that include no spaces, or strings of four or fewer characters surrounded by spaces) as the basis for their analysis. Character n-grams have been shown to be effective for authorship attribution, and Koppel *et al.* point out that one advantage is their measurability in any language without the need for specialist background knowledge. However, from a linguistic perspective they lack salience, much like the features focussed on by the early stylometrists: 'in forensic analysis there are obvious dangers in computationally pursuing an algorithm which distinguishes authors and yet has no linguistic explanation or validity' (Smith *et al*, 2009). Koppel *et al*'s method was shown to be successful in 46% of cases, which rose to 93.2% precision after the introduction of a '*Don't Know'* option. The authors conclude that their method represents an effective means of handling large candidate sets for which traditional categorization methods were ineffective, but acknowledge that the case of small open candidate sets and limited anonymous text has, as yet, no satisfactory solution.

Burrows (2002), noting that existing methods in computational stylistics are 'better fitted for 'closed' games than open ones' (p.267), offers a method for authorship attribution which is suited to those cases where there is little or no outside evidence to identify the most likely candidate. Burrows points out that most methods currently employed in the area rely on multivariate statistical comparison between certain features of a given example, and an appropriate set of norms. These comprise the frequencies of relatively simple phenomena, and can include alphabetical characters, whole words, or common grammatical forms. As Burrows points out, the advantage of working with whole words lies in their 'accessibility and meaningfulness' (2002: 268), while it has become customary to allow particular variables to "declare themselves', thus obviating...the danger of a pre-determined outcome' (2002: 268). He goes on to explain that a large set of variables that are weak discriminators is likely to offer better results than a small set of strong ones, given that strong discriminators are susceptible to being recognised and manipulated by users. As he succinctly puts it, 'a distinctive 'stylistic signature' is usually made up of many tiny strokes' (2002: 268). The procedure he develops is, he claims, successful in distinguishing the most likely author of texts exceeding 1500 words—but, more relevant to our own purposes, of even greater value in reducing the pool of likely candidates for texts as short as 100 words.

Moving on to authorship attribution methods more obviously rooted in linguistic theory, McMenamin (2010) outlines his approach to forensic texts, which is grounded firmly in stylistics—'the scientific interpretation of style-markers as observed, described and analysed in the language of groups and individuals' (McMenamin, 2010: 488). Conceptualising style markers as 'the observable result of the habitual and usually unconscious choices an author makes in the process of writing' (2010: 488), he goes on to distinguish between a) the choice between optional forms and b) deviations from the norm. Deviations from the norm may often be associated with particular classes of people, as in the case of mixing up homonyms such as 'your' and 'you're' or 'their' and 'there'—deviations that could be 'common to careless or undereducated writers' (2010: 489), or the use of 'then' for 'than', which could be indicative of a particular linguistic variety in which these forms are homonymous. These features, then, are unlikely to be individuating, although their co-selection could be.

McMenamin distinguishes between the *consistency* model, used to determine if particular texts were written by the same author, and the *population* model which must be used when the pool of candidates is large, i.e. not limited to one or two suspect writers: 'in this instance, the resemblance model is used repeatedly on one author after another until all are excluded' (2010: 490). McMenamin's approach is largely qualitative, as reflected in his assertion that 'linguistic assessments of style precede their expression as numerical values and are often a more realistic representation of the facts' (2010: 491), and this focus has been maintained by others, such as Coulthard (reported in Grant, 2010), although other research such as Grant & Baker (2001) and Grant (2010) has sought to quantify the selection and significance of style markers.

Grant (2010), in discussing authorship attribution of SMS text messages, explains how linguistic distinctiveness and linguistic consistency are matters of degree, and that questions of both can be explored using statistical methods. He calls for descriptive methods to be developed further, particularly in terms of enhancing them to enable the quantifiable comparison of consistency and distinctiveness. To this end, Grant utilises Jaccard's co-efficient, a statistical tool for establishing degrees of similarity between cases. The presence or absence of each stylistic feature identified above a certain frequency within the corpus is coded, as 1 or 0 respectively. These codings then allow for statistical comparison for similarity or dissimilarity. Jaccard's coefficient can be used to compare pairs of messages each of which is coded as a series of zeros and ones relating to the absence or presence of specific linguistic features. Jaccard is essentially a correlation coefficient applied to these binary strings and results in a (dis)similarity metric which resolves to a decimal figure between zero and one where one indicates the two text messages contain identical linguistic features and zero indicates no linguistic features in common. An important advantage of Jaccard is that a match of two absence scores across two texts has no effect on the overall similarity score (Smith *et al.,* 2009). As short-form messages are indeed short the absence of a given feature from a text carries no meaning and does not affect the calculation of similarity in either direction.

Building on the work of Grant (2010) the current project uses an extension of Jaccard called Delta-S (Δs). Delta-S was developed in marine biology and forensic psychology (Woodhams, Grant and Price, 2007) to allow the weighting of variables within a Jaccard calculation as being related to one another. In short-form messaging this requires a taxonomic description to be developed which declares, for example, substitution of different digits in a text to represent more similar stylistic choices than an accent stylisation. The taxonomy developed in this project is described and discussed below. The power of Δs is that it allows the recognition of similar but not identical stylistic choices to be represented in the final similarity metric.

2. Methodological approach

2.1. The data

The dataset analysed in the current study is a corpus of microblogs sourced from the social networking site Twitter (see Figure 1). Microblogging is a form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web. Twitter is a relatively new method of mass communication, operating in real-time and designed for mobility (Chang, 2010).

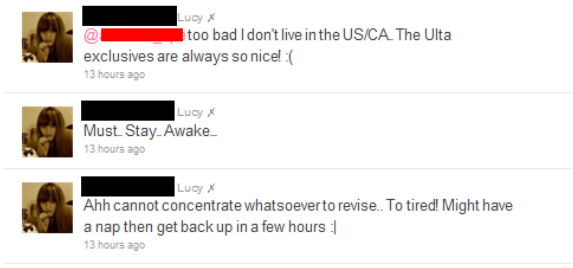

Figure 1: Twitter Screenshot

Because users do not require knowledge of any standardised interaction technique, they are able to customise Twitter to suit their needs, resulting in 'a diverse user base using the service for heterogeneous ends' (Efron & Winget, 2010). There are a number of terms that have sprung up from the Twitter community to aid in organisation and readability. The prefacing of a tweet with 'RT' ('Retweet') indicates that it is a reposting of another user's tweet, while the use of the hashtag—prefacing a word with the symbol '#'—is a convention allowing the filter of tweets by topic (Crystal, 2011; Eliot, 2009), and thus serves as a 'bottom-up user-proposed tagging convention' (Chang, 2010: 1). Users' guides such as twittonary.com offer definitions for words purported to be specific to the Twitter context, but the extent to which these are actually drawn on by users remains unclear.

2.2. Feature selection

While there has been a dramatic increase in the use of microblogging services over the last four years, research into the linguistic features of the texts and the habits and motivations of its users remains minimal (Efron & Winget, 2010). One contribution comes from Crystal (2011), who notes that tweets display a two-part structure, the first being the user's name and the message itself, and the second containing metadata, including its temporal source and Internet origin. Narrowing the focus to the internal grammatical structure of the message, he notes that the use of nonstandard punctuation often makes it difficult to assign tweets unambiguously to a particular syntactic category. Many tweets take a rather fragmented form, and words are sometimes 'juxtaposed in a way which makes an immediate interpretation impossible' (2011: 45). Crystal notes that the average number of words per tweet in his corpus was 14.7, observing that this is higher than is the case for Instant Messages (IM). He also shows ellipsis of the subject and auxiliary verb to be a frequent occurrence in tweets. Based on his corpus, he argues that within tweets there is not the same range of texting abbreviations as in SMS. This brief discussion is concluded with the observation that Twitter is a 'variety in evolution', the norms of which are still in the relatively early stages of development.

The table below demonstrates the features extracted in previous work by Grant and colleagues (Smith *et al.*, 2009) in the area of authorship attribution of SMS texts, for calculation of the Delta-S metric—a more robust version of Jaccard's co-efficient (Smith *et al.*, 2009).

Table 1: Original SMS features list (from Smith et al, 2009)

Feature	Description	Example
Mispellings	Any word not found in an English dictionary	"I saw it on the news this mroing"
Lower case 'I'	Non-captialisation of the word "I"	"i don't think so"
Acronyms	Use of acronyms	"Who are you, the CIA?"
'G' clipping	Dropping the final 'g' of words	"I'm only askin"
Accent stylisation	Using phonetic spelling to convey a specific accent	"Dey don't fink dat it could happen to dem "
Exclamatory onomatopoeia	Using onomatopoeia to convey an exclamation	"Boom, you're dead"
Prosodic emphasisers	Conveying specific pronunciation through spelling	"Boooooooring"
Whole word letter homophone substitution	Replacing entire words with a single letter	"R U still coming out tonight?"
Syllable homophone substitution	Replacing syllables within words with a single letter	"It doesn't matter ne way"
Whole word number homophone substitution	Replacing entire words with a number	"What are you waiting 4?"
Syllable number homophone substitution	Replacing syllables within words with a number	"wait until 2moro"
Whole word typographic homophone substitution	Replacing entire words with a character	"Meet you @ the bus stop"
Syllable typographic homophone substitution	Replacing syllables within words with a character	"I don't know anything about th@"
Shortenings	Common words shortened to a few initial letters	"I need to do this by Sep 10th"
Emoticons	Series of characters used to represent faces	":-)"
Initialisms	Commonly used phrases reduced to their initial letters	"ASAP"
Singular typographic exclamation	Use of a single exclamation mark	"No way!"
Multiple typographic exclamation	Use of a multiple exclamation mark	"No way!!!!!!!!!!"
Mixed typographic exclamation	Use of a mixed characters to convey an exclamation	"What the hell?!?!?!?"

Further to this list, a detailed reading of existing computer-mediated communication (CMC) literature (Crystal, 2008, 2011; Ling & Baron, 2007; Thurlow & Brown, 2003) contributed to the initial set of the type of stylistic features we could expect from our data. Since the linguistic analysis of micro-blogging communication is a relatively new field, the initial list included features of a number of other CMC genres including SMS and Instant Messaging.

The second step in the feature extraction was wholly data driven. Drawing on a development set of around 18500 tweets, a qualitative analysis was performed with the assistance of Wordsmith Tools (Scott, 2008) to identify occurrences of some of the features initially provided by the literature review. Lexicons containing every example of a given feature as it appeared in the corpus were then created by manually extracting items from the

Wordlist tool and creating plaintext files for the programmers. Figure 2 shows a section of the *exclamatory onomatopoeia* lexicon.

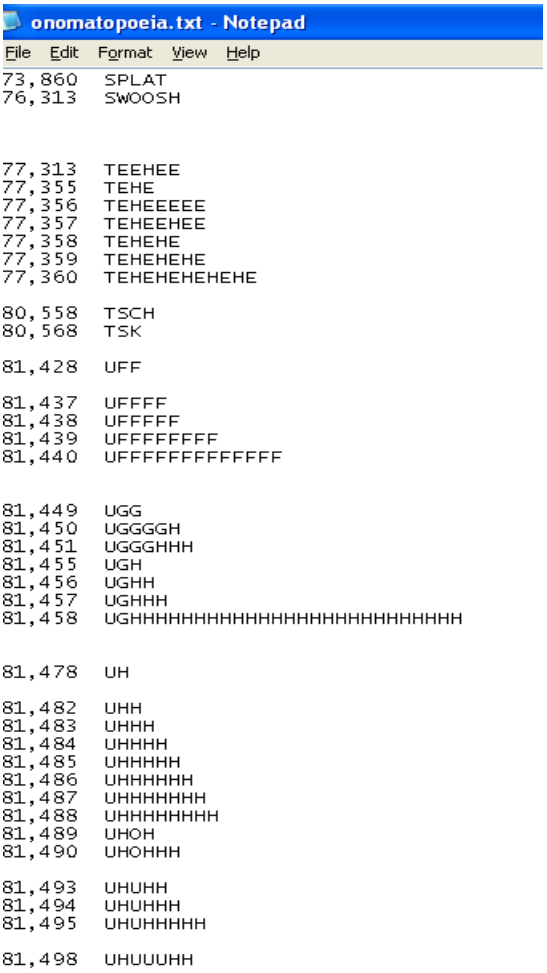

Figure 2: Exclamatory Onomatopoeia Lexicon (Extract)

Lexicons were developed in this way for features such as *exclamatory onomatopoeia*, *initialisms* and *acronyms*, for example, but were less suitable for features such as *whole word numeral substitution*. For these, a rule needed to be developed to allow the computer to distinguish between *2* the number and *2* representing the words *to* (preposition), *to* (infinitive marker) and *too* (adverb). Further to this, it was considered entirely feasible that some authors might regularly substitute the infinitive marker with the numeral (for example), but rarely do so for the prepositional use, and vice versa. Thus, a series of grammatical rules were developed by the first author, which attempted to distinguish the different uses on the basis of the context in which they occurred. There were further distinctions to be made on the basis of spacing—evidence was found in the development corpus of some users inserting a space before and after the *2*, while others inserted one before but not after, others vice versa, and some none at all, relying on the height difference of the character to visually break up the individual words. Some examples appear below.

a) *New article on darkpolitricks: GORDON DUFF: ROGUE AMERICA2* (number, no pre-space)
b) *@Skepta 2girls 1cup :|* (number, pre-space, no trailing space)

c) *@chilemad having a one 2 one* (preposition, pre-space, trailing space)
d) RT @thekatvond: Honking ur horn lk a crazyperson in stopped traffic is a gd example of not bein able 2 accept the uncontrollable-it's al ... (infinitive, pre-space, trailing space).

What resulted from this process was the refining of the original feature category into thirty-two separate features based on all the possible combinations of a) the numeral used, b) which item was replaced, and c) the use of spacing. These distinctions were made at the *bottom* of the feature categorisation system—the *top* four layers are illustrated in Figure 3 overleaf. As the illustration shows, the top-most level at which features were classified was on the basis of *lexis*, *grammar* and *punctuation*, as well as by features peculiar to the *mode of production*, including hashtags and re-tweets. The features classified under the *grammar* heading related mainly to omission of particular classes of word such as verbs and determiners—patterns fairly typical of 'telegraphic' speech, which, as expected, are not particularly useful in assessing matters of authorship. Thus, the analysis focussed in the main on the features classified under *lexis* and *punctuation.* The Delta-S (Δs) distance metric was used to determine and measure distance between two documents by using the presence, and position in the hierarchy, of the stylistic features.

3. Evaluation

After developing the feature set the next step was to test the method's effectiveness at a number of tasks, the end task being the identification of the likely author for a single or small number of short messages, an 'open' problem, where:

· There are many unknown potential authors in the candidate set;
and
· The author of the unknown message may not be present in the candidate set.

The aim was for the system to be able to provide one of the following responses:

· *Author Identified*: the results exceed a specific confidence level;
· *Potential Author*: the results approach the confidence level;
· *Undetermined*: the message contains too few stylistic features to make a judgement above a determined confidence threshold;
· *Not Present*: the author is unlikely to be present in the candidate set.

Where only a small set of messages are available from an unknown author, the decision was made to aggregate these messages in an attempt to improve the chances of attribution. However, it must be borne in mind that, particularly in an online context, there is no guarantee that all the texts in a suspect set were authored by one individual, since a number of authors may have access to a particular micro-blogging account.

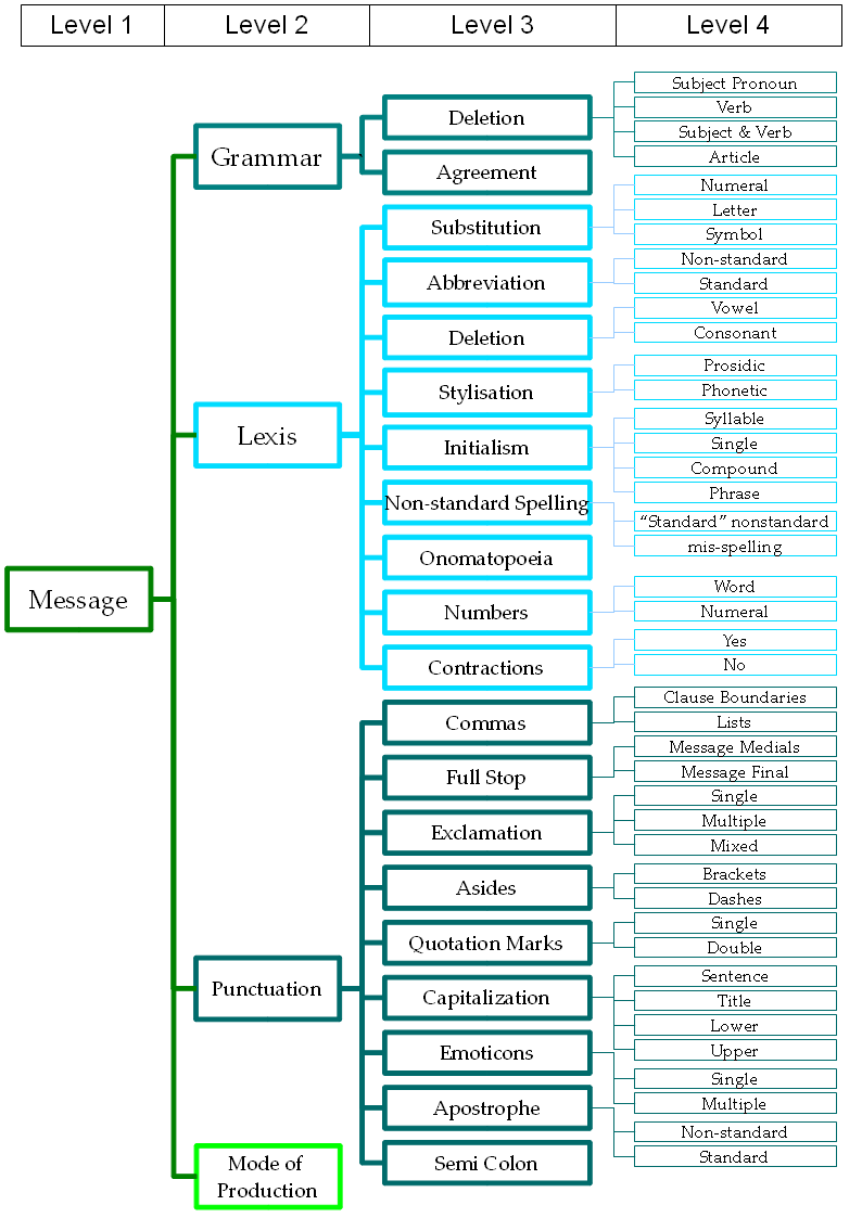

Figure 3: Top four levels of the feature categorisation system.

3.1. Attribution confidence

We reflect here only on the ability of the methodology to discriminate authors of messages: issues concerning scalability to very large datasets are to be the focus of a future project.

The Delta-S (Δs) metric is calculated between a single test message of Author **X** and a set of comparison messages from authors (**A1 , A2 , A3 ,... An**) ; (**B1 , B2 , B3 ,... Bn**), (**C1 , C2 , C3 ,... Cn**), etc. This gives a series of samples:

A{ Δs(**X→A1**), Δs(**X→A2**), Δs(**X→A3**) … Δs(**X→AN**) **};**
B{ Δs(**X→B1**), Δs(**X→B2**), Δs(**X→B3**) … Δs(**X→BN**) **};**
C{ Δs(**X→C1**), Δs(**X→C2**), Δs(**X→C3**) … Δs(**X→CN**) **}** etc.

These can then be compared with a non-parametric statistical significance test (in this case Mann-Whitney U-test) to determine the concordance probabilities:

P(**A** > **B**), P(**A** > **C**) , P(**B**> **C**) etc.

These represent the probability that both samples can be drawn from the same set, with a low probability indicating more significant differences.

3.2. Performance for single messages

The initial test took 10 single, random tweets from a known Author **A** (**Xa**), and generated the Δs distance measures to 100 other tweets from Author **A**, and 100 tweets from author **B**. Three experiments were carried out each with different authors and test messages. The results are presented in Table 2 , which shows the number of messages identified correctly (out of 10 for each trial), the number of these identified correctly with high statistical significance, and the number of messages that could not be assigned. In no case were messages incorrectly linked.

Table 2: Performance for Single Messages

	# Correct	# P ≤ 0.01	# Indeterminate	# P ≤ 0.01
Experiment 1	9	6	1	0
Experiment 2	5	2	5	0
Experiment 3	8	5	2	0

The results show reasonable accuracy and discrimination for a single message, with the correct author identified in the majority of cases, and many of these assigned with a significant level of certainty. Furthermore, no messages were incorrectly assigned to an author. One of the reasons a number of messages could not be assigned is the frequent sparsity of features within such short messages. In this dataset a tweet is typically 12 words long and on average contains fewer than 3 stylistic features. Thus, an approach needed to be developed which could allow for the fact that some will contain many identifying features, whilst others will contain few or none.

3.3. Performance for aggregated messages

The effect of feature sparsity can be reduced by aggregating messages before the Δs calculation. This second test aggregated random tweets from a known author **A** into 10 batches of 1, 2, 5 or 10 messages each. The Δs distance measures were then calculated for each of these aggregations to 100 batches of other messages from author **A** and 100 batches of messages from author **B**. The authors used in this trial were that same as in Experiment 2, the worst performing from the single text test.

Table 3: Performance for Aggregated Messages

Aggregation	# Correct	# P ≤ 0.01	# Indeterminate	# P ≤ 0.01
1	5	2	5	0
2	7	5	3	0
5	8	7	2	0
10	10	10	0	0

The results show an improvement in performance after aggregation. The number of words in the aggregated messages averages around 90 for the 10 message case; still well below the lower limits of stylometric techniques. This author is fairly typical, using an average of 9 words per message, with each message containing an average 2.5 stylistic features. The increase in performance is striking even for modest levels of aggregation. Again, no messages were incorrectly assigned to an author.

3.4. Performance for multiple authors

The next scenario that was considered was one in which multiple authors are present in the candidate set. The test took 10 single, random tweets from a known Author A (Xa), and generated the Δs distance measures between these and 100 other tweets from Author A , 100 tweets from author B, 100 from author C, D etc. Candidate sets of 2, 5, 10 and 20 authors were considered. The results are shown in

Table 4. For each message, the rank order (with 1 being the most similar and 20 being the greatest distance) shows the ranking of the 'correct' author as the likely author of the message in question. The table also shows the level of significance with which messages were incorrectly assigned when they were *not* ranked first.

Table 4: Performance for Multiple Authors

	Rank Order	Correct ?	Incorrect # P ≤ 0.01
Message 1	1	Yes	N/A
Message 2	3	-	0
Message 3	2	-	0
Message 4	1	Yes	N/A
Message 5	5	-	0
Message 6	3	-	0
Message 7	3	-	0
Message 8	1	Yes	N/A
Message 9	1	Yes	N/A
Message 10	1	Yes	N/A

These results show that the methodology has reasonable success identifying authors from a set of 20 authors, a relatively large candidate author set. Of particular interest is that in those cases where the correct author is not top ranked, the significance of the result is not definitive. In practical terms this minimises the risk of false positive results.

4. Concluding remarks

This paper has demonstrated that positive results are possible for typical short message content (SMS text and Twitter), and the approach reported on here advances the state of the art in terms of the size of message to which authorship analysis can be applied. The implementation of feature identification has proved effective in terms of the accuracy and coverage of the feature instances identified and annotated per message. However, some improvements could be made to increase overall performance. Time did not allow, for example, for rigorous part-of-speech tagging of the corpus, which would have allowed for greater use of the rule-based feature categorisation as reported on in section 2.2. Furthermore, although a number of detailed lexicons were developed for feature categorisation tasks (such as onomatopoeic expressions and various sub-categories of initialisms), there are a number of other feature types for which this remains to be completed.

Further improvement to the identification process may be possible by weighting particular features according to how common or rare they are. This would mean that the presence of a very common phrase initialism such as ‘LOL’ (‘Laugh Out Loud’) in both the questioned text and a candidate set of texts would receive a lower weighting than the presence of a rarer one, such as BBIAB (Be Back In A Bit). All these improvements would be likely to contribute to a more refined system with even higher success rates. A further issue for future research is the scalability of the process. In light of the practical reality of online messaging, any operationally useful system would need to generate valid results on the very large data sets typical of the context.

Acknowledgements

Special thanks to Darrell Smith & colleagues at Lexegesys Ltd, with whom we collaborated on this project.

References

Abbasi, A, & Chen, H. (2005) Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20 (5), 67—75.

Argamon, S. (2008) Interpreting Burrows's Delta: geometric and probabilistic foundations *Literary and Linguistic Computing* 23 (2), 131—147.

Burrows, J. (2002) 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17 (3), 267—287.

Chang, H.-C. (2010) A new perspective on Twitter hashtag use: diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology* 47: 1—4.

Chaski, C. (2001) Empirical evaluations of language-based authorship identification techniques. *International Journal of Speech, Language & the Law* 8 (1), 1—65.

Crystal, D. (2008) *txtng: the gr8 db8.* Oxford: OUP.

Crystal, D. (2011) *Internet Linguistics: A Student Guide.* Abingdon: Routledge.

Efron, M. and Winget, M. (2010) Questions are content: a taxonomy of questions in a microblogging environment. *Proceedings of the American Society for Information Science and Technology* 47: 1—10.

Eliot, G. (2009) Common Twitter terms: using and understanding the language of Twitter. http://www.suite101.com/content/common-twitter-termsa101868#ixzz1Ega53YeR

Grant, T. (2007) Calculating TXTual distance in forensic authorship analysis. Paper presented at the International Association of Forensic Linguists 8th Biennial Conference, University of Washington, Seattle, USA, July 12—15, 2007.

Grant, T. (2010) Text messaging forensics: txt 4n6: idiolect free authorship analysis? in M.Coulthard & A. Johnson (eds) *The Routledge Handbook of Forensic Linguistics*, 508—522.

Grant, T. & Baker, K. (2001) Identifying reliable, valid markers of authorship: a response to Chaski. *International Journal of Speech, Language & the Law* 8 (1), 66—79.

Grant, T., MacLeod, N., Exell, A., Smith, D., Spencer, S. & Webb, A. (2011) *Authorship analysis for short form messages.* Unpublished Research Report, Aston University/Lexegesys.

Hoover, D.L. (2003) Multivariate analysis and the study of style variation. *Literary and Linguistic Computing,* 18 (4), 341—359.

Java, A., Song, X., Finin, T. & Tseng, B. (2007) Why we twitter: understanding microblogging usage and communities *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, http://portal.acm.org/citation.cfm?id=1348556

Koppel, M., Schler, J. & Argamon, S. (2011) Authorship attribution in the Wild. *Language Resources & Evaluation* 45, 83—94.

Koppel, M., Schler, J., Argamon, S. & Messeri, E. (2006) Authorship attribution with thousands of candidate authors. *Proceedings of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval* Seattle, Washington.

Ling, R. & Baron, N.S. (2007) Text messaging and IM: Linguistic comparison of American college data. *Journal of Language & Social Psychology* 26 (3), 291—298.

Mcmenamin, G.R. (2001) Style markers in authorship studies. *International Journal of Speech, Language & the Law* 8(2), 93—97.

McMenamin, G.R. (2010) Theory and practice of forensic stylistics. in M. Coulthard & A. Johnson (eds) *The Routledge Handbook of Forensic Linguistics* 487—507, London: Routledge.

Miranda-García, A. & Calle-Martín, J. (2005) The validity of lemma-based lexical richness in authorship attribution: a proposal for the Old English Gospels. *ICAME 29*, 115—130

Scott, M. (2008) WordSmith Tools Version 5, Liverpool: Lexical Analysis Software.

Smith, J. A. & Kelly, C. (2002) Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities* 36: 411—430.

Smith, D.J., Spencer, S. & Grant, T. (2009) *Authorship analysis for counter terrorism* Unpublished Research Report, QinetiQ/Aston University.

Stamatos, E. (2008) A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60 (3), 538—556.

Thurlow, C. & Brown, A. (2003) Generation Txt? The sociolinguistics of young people's text messaging. *Discourse Analysis Online* 1 http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html

Woodhams, J., Grant, T. D., & Price, A. R. G. (2007) From marine ecology to crime analysis: improving the detection of serial sexual offences using a taxonomic similarity measure. *The Journal of Investigative Psychology and Offender Profiling* 4, 17—27.

Zheng, R., Li, J., Chen, H. & Huang, Z. (2005) A framework for authorship identification of online messages: writing style features and classification techniques. *Journal of the American Society for Information Science & Technology* 57(3), 378—393.