Report of Dr Nicci MacLeod

Dated: 12th June 2012

Specialist field: Forensic linguistics

Subject matter: Comparison of tweets from tabloidtroll to tweets from dennisricemedia for evidence of linguistic consistency.

Dr Nicci MacLeod
Centre for Forensic Linguistics
School of Languages & Social Sciences
Aston University
Aston Triangle
Birmingham
B4 7ET
n.macleod1@aston.ac.uk

Tel.: 0121 204 3769

1. Introduction

1.1 About me

I am Dr Nicci MacLeod, Research Associate at the Centre for Forensic Linguistics, Aston University. My specialist areas are forensic discourse analysis, authorship attribution, and sociolinguistic profiling. I am a member of the International Association of Forensic Linguists (IAFL). A summary CV is attached.

The opinions expressed herein are entirely my own.

1.2 Background to the case

This case centres on the authorship of tweets originating from a Twitter account, tabloidman (@tabloidtroll). I was tasked with comparing around 2000 tweets from this account with a corpus of around 600 tweets from the account dennisricemedia, whose author is presumed to be Dennis Rice, for the purposes of assessing the level of linguistic consistency between the two sets. Where appropriate I have referred to smaller corpora of tweets from other male tabloid journalists, and to the live Twitter feed itself, as 'control' corpora.

1.3 Summary of conclusions

Based on my linguistic analysis I conclude the following:

- There are multiple points of linguistic consistency between the tabloidtroll and dennisricemedia data sets.
- There are no significant points of linguistic difference between the tabloidtroll and dennisricemedia data sets.

2. Authorship analysis

Authorship analysis procedures often involve a combination of linguistic, descriptive and statistical methods. All authorship analyses additionally require a less statistical, linguistic expertise in understanding the causes of consistency and variation in language. This involves not only an appreciation of what is unusual in a text but also the recognition of possible explanations for the occurrence and use of interesting features.

The analysis carried out here depends upon a close reading of the texts looking for features or idiosyncrasies in written style of the known tweets which might indicate authorship of the queried tweets. An analysis such as this starts with a reading of the documents of known authorship attempting to find style markers which are consistent enough within the texts of known authorship such that they will 'carry over' to the query texts.

Standard stylistic markers such as lexical richness, frequency of function words, or syntactic measures, known to perform well with longer, 'standard' language texts, perform worse with texts as short as tweets. In this context, 'low level' features such as emoticons, initialisms and punctuation choices have proved useful in authorship analysis tasks (see MacLeod & Grant, 2012; Sousa Silva *et al.* 2011).

It should be kept in mind that while none of these features can independently individuate a particular author, the *combination* of features that build up an individual's repertoire provides increasingly strong evidence as more features are added.

The table below summarises these features. It should be noted that only a selection of these are relevant for the current analysis.

Table 1: Features of short form messages (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 pronounciation 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	Replacing entire words with a character	"Meet you @ the bus stop"

Feature	Description	Example
substitution		
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?!?!?!?"

3. Findings

3.1 Initialisms

dennisricemedia makes use of a number of initialisms, including 'btw' for 'by the way', 'lol' for 'laugh out loud', 'NOTW' for 'News of the World' and 'mf' for 'more follows', which is used at the end of a tweet that is to be taken as one of a series. All four of these also appear in the tabloid troll tweets. As well as the mere use of a feature, attention is paid here to *how* it is used – issues of case, spacing, punctuation and location. The preference across both sets is for tweet-final 'lol' *not* to be followed by a full stop, although there are exceptions within the tabloidtroll set. There are also occasions when tabloidtroll makes use of capital letters rather than lower case. There were no such occurrences in dennisricemedia. However, we must bear in mind that with only 3 overall occurrences of this initialism across the whole (substantially shorter) dennisricemedia set, the feature is somewhat sparse. I would therefore not interpret non-occurrence of upper case 'LOL' in dennisricemedia as evidence of inconsistency.

The occurrence of the initialism NOTW in itself is not significant given that the two accounts overlap significantly in terms of content (journalism, media standards, Leveson, etc.). However, there are a number of choices available for this initialism in terms of which of the words are initialised and which letters are capitalised. A cursory glance at Twitter and at retweets contained within the tabloidtroll and dennisricemedia data reveals the options NoW, NotW, NOtW, NoTW and notw. Both tabloidtroll and dennisricemedia consistently use NOTW. Thus, for this feature, tabloidtroll and dennisricemedia are consistent in the choices they make.

A further initialism that appears across both sets of texts is ‘btw’ for ‘by the way’. Both accounts consistently display ‘btw’ in lower case. Furthermore there are grammatical similarities in the way it is used – at the end of a clause – whereas a number of Twitter users display a preference for using it at the beginning of a clause. There are also examples of consistency between the two sets in terms of the use of a dashed aside immediately following the initialism.

‘mf for ‘more follows’ in tweet final position is used once in the dennisricemedia set and 24 times in the tabloidtroll set. While sparse in dennisricemedia, the occurence of the feature supports the idea that the author(s) of both have a shared, or at least overlapping, lexicon of initialisms.

3.2 Substitution

The use of a numeral, letter or other character in place of either a syllable or an entire word is a common feature of short form messages, but authors differ in the range of options they make use of. Both tabloidtroll and dennisricemedia make use of ‘b4’ for ‘before’ (letter-syllable and numeral-syllable substitution) and ‘4’ for ‘for’ (numeral-word substitution).

It should also be noted that both authors make use of the full form ‘before’ (tabloidtroll 20 occurrences, dennisricemedia 5 occurences), as well as the substituted ‘b4’. Thus, while both sets display a preference for the full form, the substituted variant forms part of both their repertoires. Comparisons with control corpora suggest this is not a particularly frequent feature in the style of this type of user. Once more, this suggests consistency between the two sets.

‘4’ for ‘for’ forms part of the repertoires of the author(s) of both sets of texts. As with ‘b4’, the full form is preferred by both authors, although it is difficult to provide exact figures due to the prevalence of the phrase ‘Twitter for iPhone’.

‘u’ for ‘you’ also occurs in both sets of texts.

3.3 Spacing & capitalisation

In representing time, both sets display a preference for lower case am/pm and NO space between the number and the letters, i.e. ‘4am’ as opposed to ‘4 am’, ‘4AM’ or ‘4 AM’. Once again, authors have a range of possible options but the two sets of data are consistent in their use of one particular variant over others.

3.4 Prosodic emphasisers

Both authors make use of repetition of the letter ‘o’ in their spelling of ‘so’, with one occurrence in each set of ‘sooo’ and one in each of ‘soooo’, and one occurrence of ‘soo’ in tabloidtroll.

3.5 Emoticons

The most frequently used emoticon across the two sets is :), with 13 occurrences in dennisricemedia and 51 in tabloidtroll. The most frequent position for the emoticon is at the end of the tweet (‘tweet-final’), with 12 of dennisricemedia’s and 39 of tabloidtroll’s appearing here. 100% of these across both sets appear with no final punctuation mark. While not infrequent – many users rely on the emoticon itself as a closing – there are other options which are not exploited by either tabloidtroll or dennisricemedia. It is also worth mentioning that the choice of *which* emoticon is consistent - there are no occurrences of the equivalent :-), :-D or :D in either set, all of which occur frequently across Twitter as a whole.

The author(s) of both sets of texts produce the emoticon followed by an upper-case letter on the next word when at a sentence boundary, and a lower- case letter on the next word when mid-sentence. Emoticons are not followed by punctuation marks in either final or medial position for both sets, with the exception of one occurrence in tabloidtroll of a full stop after the emoticon in medial position.

3.6 Onomatopoeia

The most frequent onomatopoeic expressions across both sets of texts are ‘err’, a ‘hesitation’ marker used to indicate a problematic proposition, and ‘hmmm’, used to indicate either that the author is giving something thoughtful consideration, or in a sarcastic sense to indicate the opposite. ‘Ahem’ also occurs a number of times across the sets.

‘err’

There are 34 occurrences of ‘err’ in tabloidtroll and 3 in dennisricemedia.

There is a scarcity of feasible alternative expressions in the data, and no alternative spellings in either set. Furthermore, there are consistencies in the way the item is used in both sets. Both show occurrences of ‘err’ immediately following 3 full stops ‘...’, and elsewhere. Perhaps most markedly, both show ‘err’ following the first person plural pronoun ‘we’ and preceding a negative statement. A quick search on Twitter reveals that the majority of the top 50+ hits for “we err” are either in languages other than English, are using ‘err’ as a verb (e.g. we err because we are human), or as a phonetic stylisation of ‘ever’. Where it is used in its onomatopoeic sense, it is clear that there are other options for its use that are inconsistent with the way it is used in the two data sets, e.g. followed by a comma or a single full stop.

Lastly, the dashed aside " – err not" appears in both sets:

dennisricemedia @tabloidwatch Yes what an awful Royal story out of 1000s of positive ones. And so many people hurt by it - err not. #pathetic 12:24 PM Jan 9th from Twitter for iPhone

abloidtroll @JamieSmiff Online news writer? Well that immediately places you at the heart of the industry doesn't it - err not. ROFL 2:44 PM Mar 19th from Twitter for iPhone

Taken together I consider these to be strong markers of consistency.

'Hmmm'

There are 46 occurences of 'hmmm' in tabloidtroll and 12 in dennisricemedia.

Here too, there is no variability in the spelling of 'hmmm' with the exception of 1 occurrence of 'hmmmm' in tabloidtroll. This is easily explainable as a typographic error.

In a contrary pattern to 'err', the preferred position for 'hmmm' is before rather than after 3 full stops '...', and both data sets contain occurrences of this pattern. The data sets show consistency in that both capitalise the first 'H' on 'Hmmm' when it appears at the start of a tweet ('tweet-initial') or at a sentence boundary. Both make use of both capital and small case 'h' if the 'hmmm' follows an account name (e.g. '@JournoSharon hmmm').

4 Conclusions

There are a significant number of examples of consistency between the two sets of texts, and these are significant because as authors we are faced with a variety of options, and within each range of options the author(s) of both these sets of texts has selected the same option. Put another way, the analyses presented here have demonstrated that for all the features selected, each of which allows for a limited set of choices to be made, the author of tabloidtroll and the author of dennisricemedia make the same choice.

Based on my analyses, I conclude that there are multiple points of linguistic consistency, and no significant points of linguistic difference. It is my opinion that the use of emoticons and onomatopoeic expressions is the most marked, and therefore provides the strongest evidence of consistency.

5 References

MacLeod, N. & Grant, T. (2012) 'Whose Tweet? Authorship analysis of micro-blogs and other short form messages'. *Electronic Proceedings of the International Association of Forensic Linguists' 10th Biennial Conference, Aston University, Birmingham, UK, July 2011*

Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E. & Maia, B. (2011) ''twazn' me!!! ;(' Automatic authorship analysis of micro-blogging messages'. *NDLB 161-168*

Summary C.V.

Dr Nicola J MacLeod
Centre for Forensic Linguistics
Aston University
Birmingham
B4 7ET

e-mail: n.macleod1@aston.ac.uk
Tel.: 01212043769

Education:

2010: PhD in Forensic Linguistics, Aston University, Birmingham, U.K.
2004: MA in Forensic Linguistics, Cardiff University, U.K.
2003: BA in English Language, Bangor University, U.K.

Employment:

05/2011 – present Aston University, Birmingham
Research Associate, Centre for Forensic Linguistics

01/2011 – 08/2011 University of Birmingham, Birmingham
Sessional Lecturer (Forensic Linguistics), Centre for English Language Studies

01/2011 – 05/11 Aston University, Birmingham
Contract Research Associate, 'Authorship Analysis of Short Form Messages', Centre for Forensic Linguistics

05/2010 – 02/2011 University of Aberdeen, Aberdeen
Research Fellow, 'Language & Linguistic Evidence in the 1641 Depositions', School of Language & Literature

Selected Publications:

MacLeod, N. & Grant, T. (2012) 'Whose Tweet? Authorship analysis of micro-blogs and other short form messages'. Electronic Proceedings of the International Association of Forensic Linguists' 10th Biennial Conference, Aston University, Birmingham, UK, July 2011.

MacLeod, N., & Fennell, B. (forthcoming September 2012) 'Lexico-grammatical portraits of vulnerable women in war: The 1641 Depositions' to appear in The Journal of Historical Pragmatics 13(2).

MacLeod, N. (forthcoming 2012) 'Rogues, villaines & base trulls: Constructing the other in the 1641 Depositions' to appear in E. Darcy, A. Margey & E. Murphy (eds.) The 1641 Depositions and the Irish Rebellion London: Pickering & Chatto.

MacLeod, N. (forthcoming 2013) 'Forensic Linguistics' The Blackwell Encyclopedia of Applied Linguistics C.A. Chapelle (Ed.) Oxford: Wiley-Blackwell.

MacLeod, N. (2011). 'Risks and benefits of selective (re)presentation of interviewees' talk: Some insights from discourse analysis'. British Journal of Forensic Practice,13 (2), 95 – 102.

MacLeod,N. (2009) ''Well did you feel jealous?' Control & ideology in police interviews with rape complainants' Critical Approaches to Discourse Analysis Across Disciplines 3 (1), 46 – 57.