PhD Abstract

Authorship profiling in a forensic context

Andrea Nini

University of Manchester

Awarding Institution:
Aston University, UK

Date of award: 2015

Keywords: Authorship analysis, authorship profiling, sociolinguistics, gender, age, education, social class, style, threat, register analysis.

Introduction

A task commonly assigned to forensic linguists is the analysis of a piece of dangerous communication for the purpose of profiling the general characteristics of the anonymous author. The most famous case of this kind is probably Shuy's "devil's strip" letter, in which a kidnapper has been profiled as an educated person from Akron, Ohio from the use of the phrase 'devil's strip' (Leonard, 2005). Although linguists are often successful in using the sociolinguistic origin of lexical items for the purposes of profiling, such as in the mentioned case, the research on the variation in writing style and its correlation to social factors, such as gender or age, is still under-developed. Some advances have been carried out within computer science on non-forensic texts, such as blogs or novels (e.g. Argamon *et al.*, 2009; Newman *et al.*, 2008). Because of the pervasiveness of register variation, whether these findings do apply to malicious texts commonly analysed within forensic linguistics is still an open question.

Research questions

The thesis examines whether already established findings regarding the correlations between general patterns of language variation and the social factors of gender, age, level of education, and social class also apply to **malicious texts**. Such texts are defined in the present work as those *texts that are a piece of evidence in a forensic case that involves*

threat, abuse, spread of malicious information or a combination of the above. The goals of the thesis are thus to: (1) elaborate a comprehensive account of the most pervasive patterns of variation of language use correlated with gender, age, level of education, and social class; (2) verify whether these same patterns are found in malicious texts and whether they can help in the profiling of the demographics of the author of a malicious forensic text.

Methodology

To answer the research questions above, a first step consists in gathering a battery of style markers by performing a literature review across all the disciplines that have shown an interest in studying language variation across social factors. After a review of the major studies in different fields, a list of 132 grammatical and lexical linguistic features was compiled.

In order to test whether these same markers could be used to profile authors of malicious texts a corpus of malicious text including the texts' authors demographic details should be compiled. However, due to lack of malicious texts with known author demographics, a set of fabricated texts resembling malicious texts was gathered instead. This corpus, the Fabricated Malicious Texts (FMT) corpus, was compiled in an experimental setting using writings produced by 96 participants from several social backgrounds. These participants were asked to write three texts of about 300 words in length: a letter of complaint to a holiday agency, a letter of complaint to the Prime Minister of the UK, and a letter to their abusive employer threatening damage to their car.

Given the limitation of fabricated data, to control for the possibility that the fabricated data is different from natural occurring data, a corpus of authentic malicious texts was also gathered. This corpus, the Authentic Malicious Texts (AMT) corpus, consisted of 104 texts with an average length of 354 words. This corpus was used to test whether any significant difference in terms of the 132 linguistic features examined existed with the FMT data set and whether, generally speaking, the register adopted in the AMTs was consistent with the register adopted by the participants in the FMTs. Statistical tests of difference indicate, firstly, that no significant register differences are present between the two data sets. Secondly, tests of difference also indicate that no significant differences due to the experimental conditions are found for the 132 linguistic features between the AMT and FMT data sets with the exception of two features. The two findings altogether suggest that findings relative to the FMT data set can also be expanded to real malicious texts.

The analysis of the FMT data set consisted firstly in the use of tests of statistical difference for the 132 linguistic features. Secondly, using the best performing features, logistic regression models were fitted in order to test whether the features presenting statistical differences were also able to predict the gender, age, level of education, or social class of an author.

Results

A first result of the thesis is the literature review that describes the stylistic markers and the most pervasive patterns of variations in language use for the social factors gender, age, level of education, and social class. The review demonstrates that there exists patterns of language variation that are consistently found across several studies using different data sets:

- The genders are typically distinguished by the *report vs rapport* discourse orientations, with men typically adopting a report style realised by more nominal structures while women typically adopting a rapport style realised by more verbal and clausal structures.
- Different age groups are distinguished by the same opposition between nominal and clausal structures, with the incidence of nominal style tending to increase with age. Additionally, grammatical complexity was found to decrease with age.
- For level of education and social class, the literature presents a picture that consists in a general decrease of lexical and grammatical complexity as level of education and socioeconomic metrics decrease.

The analysis of the FMT data set shows that all of the major patterns of variation found in other data sets are also valid for malicious forensic texts. Since register variation was experimentally controlled in this study, stronger effects are found compared to previous studies for all of the social factors. Social class was the social factor for which the strongest effects were observed. Level of education, on the other hand, was dropped from further analysis as the effects observed for this social factor largely overlapped with social class.

The use of logistic regression to train models that can profile the authors of FMT texts was largely successful, although performance varied depending on the register of the text. A logistic regression model trained to distinguish upper from lower social class achieved a classification rate of, on average, 78%. Social class was the easiest social factor to profile, which confirms that social class is the social factor with the most pervasive effect on language production. The logistic regression model for age was the second most successful model, with an average prediction rate of 70% between subjects above and below the age of 40. Finally, gender could be predicted with a rate of 60-70% depending on the register of the letter, with the abusive letter to the boss being the easiest to profile.

In conclusion, the present study demonstrates that the established patterns of variation in language use attested in various registers for the social factors gender, age, level of education, and social class are also present in malicious forensic texts. However, the profiling of the authors' social backgrounds are greatly affected by register variation. Although these linguistic patterns can be used to predict the demographics of an unknown author with a reliability that ranges between 60% to 80%, performance depends on the register of the text, the analysis of which is thus a precursory and necessary step for authorship profiling.

References

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, 52(2), 119.

Leonard, R. A. (2005). Forensic Linguistics: Applying the Scientific Principles of Language Analysis to Issues of the Law. *The International Journal of the Humanities*, 3.

Newman, L. M., Groom, C. J., Handelman, L. D. and Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3), 211–236.