Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles

[1]George K. Mikros and [2]Kostas A. Perifanos
[1]Department of Italian Language and Literature, [2]Department of Linguistics,
National and Kapodistrian University of Athens, Greece
[1]gmikros@isll.uoa.gr, [2]kperifanos@phil.uoa.gr

Abstract

The aim of this study is to explore authorship attribution methods in Greek tweets. We have developed the first Modern Greek Twitter corpus (GTC) consisted of 12,973 tweets crawled from 10 Greek popular users. We used this corpus in order to study the effectiveness of a specific document representation called Author's Multilevel N-gram Profile (AMNP) and the impact of different methods on training data construction for the task of authorship attribution. In order to address the above research questions we used GTC to create 4 different datasets which contained merged tweets in texts of different sizes (100, 75, 50 and 25 words). Results were evaluated using authorship attribution accuracy both in 10-fold cross-validation and in an external test set compiled from actual tweets. AMNP representation achieved significant better accuracies than single feature groups across all text sizes.

Introduction

Automatic Authorship Identification (AAI) has a long history dated back to the end of the 19th century when Mendenhall (Mendenhall 1887) examined for the first time the word length in the works of Bacon, Shakespeare and Marlowe in order to detect quantitative stylistic differences. Since the late 1990's quantitative stylistic analysis has known a new impetus based on developments in a number of key research areas such as Information Retrieval, Machine Learning and Natural Language Processing (Stamatatos 2009). Furthermore, online text is now massively available and Web 2.0 has added to the now standard internet genres of email, web page and online forum message, new forms of online expression such as blogs, tweets and instant messaging.

AAI research now is concerned not only with problems of authorship in the broad field of the Humanities (Literature, History, Theology), but also with applications in various law-enforcement tasks such as Intelligence, Forensics e.g. (Chaski 2005; Iqbal et al. 2010a; de Vel et al. 2001; Li, Zheng, and Chen 2006).

The major application areas are described briefly below:

1. Authorship attribution: This is the most common authorship identification analysis with the study of the Federalist Papers by Mosteller and Wallace (1984) being a typical example. In this case we are trying to find who is the author of one or more disputed texts among a closed set of 2, 3,..., n known authors. This scenario assumes that we are certain that at least one of the possible authors is actually the author of the disputed texts and that an adequate corpus in size and quality for every possible author is available (Juola 2008).
2. Author verification: In this case we are investigating whether certain text(s) were written by a specific author. We are assuming an open set of authors and each dubious document must be attributed to the specific author without reference to corpora from other authors (Iqbal et al. 2010b; Koppel and Schler 2004; Van Halteren 2007).
3. Author profiling: In some applications related to Information Retrieval or Opinion Mining and Sentiment Analysis we are interested in identifying the author's gender (Argamon et al. 2007; Koppel, Argamon, and Shimoni 2002; Schler et al. 2006), age (Argamon et al. 2003) or psychological type (Argamon et al. 2005; Luyckx and Daelemans 2008a; 2008b).

In this study we are concerned with the authorship attribution in Greek tweets. More specifically we explore the effectiveness of n-gram features of different sizes and levels to the authorship classification accuracy taking into consideration the text size of the training and the test set. Since tweets are extremely short texts we are examining different methods of constructing training sets in order to develop a robust text size threshold for authorship attribution in tweets of Modern Greek language.

Authorship attribution in tweets

Since its launch in 2006, Twitter has expanded rapidly and became one of the most active social networking sites worldwide. Last usage statistics for 2011[1] revealed that approximately 200 million tweets per day are sent from more than 200 million registered accounts.

Twitter, has transformed radically the ways information is spread over the internet and created a new language genre with specific linguistic conventions and usage rules. Users form messages in 140 characters or less producing text that

[1]Source: http://visual.ly/following-twitter

is semantically dense, has many abbreviations and often carries extralinguistic information using specific character sequences (smileys, interjections etc.) (Crystal 2008). Although Twitter resembles many other text-size restricting services like Instant Messaging (IM) in Internet and Short Message Service (SMS) in cell phones, considerable differences have been noticed between them. According to Denby (2010), tweets represent a more standard linguistic norm including a higher percentage of standard punctuation usage, a low number of logograms and pictograms, and a complete lack of logograms in which a character is used to directly represent a phonetic sound, features that have all been observed within text messaging and instant messaging.

Authorship attribution methods have already been applied to tweets, since Twitter is an extremely popular service and cybercrime frequently uses it for illegal activities. Layton, Watters and Dazeley (2010) collected the 200 most recent tweets from 14,000 Twitter users. They applied a methodology named Source Code Authorship Profile (SCAP) which is based in character n-grams and the evaluation of their relative distance from the documents of undisputed authorship. More specifically, for each author they joined all his/her documents and extracted the L most frequent n-grams. The produced list constitutes the main author's profile and each of the test documents is evaluated using the relative distance of its n-gram profile to the verified one. The proposed methodology was tested using different n-gram sizes as well as varying L numbers. SCAP methodology was applied to a subset of 50 authors and produced accuracy rates of 70% with n-gram sizes ranging from 3 to 6. Furthermore, it turned out that @reply was a highly informative feature since its removal led to 27% accuracy decrease denoting that the social network of each author was highly informative for his/her identity. Another important finding was that authorship attribution accuracy was stabilized when the training set had reached approximately 120 tweets per author.

Sousa Silva et al. (2011) compiled a Portuguese Twitter corpus of 200,000 users and over 4 million messages. From this corpus they selected the 120 most active users with at least 2,000 distinct and original messages (i.e. excluding retweets). The basic experimental setup was based on randomly selecting 40 3-author groups and measuring authorship attribution precision, recall and F-value. Datasets of size 75, 250, 1,250 and 2,000 tweets per author were created in order to test the text-size influence in the authorship attribution effectiveness. A number of different non-lexical feature groups were used including quantitative markers (word length spectrum, @replies, #tags, URLs etc.), marks of emotion (smileys, LOLs, interjections), punctuation, and abbreviations. The machine learning algorithm selected was Support Vector Machines (SVM), due to its robustness in text classification tasks and sparse datasets. Results indicated that emotion symbols were highly effective in identifying the author of a tweet and that reliable authorship attribution results can be obtained using as little as 100 tweets per author.

Boutwell (2011) developed a multimodal classifier for tweets in order to match a cell phone with a specific user. Separate Naive Bayes (NB) classifiers were built for authors and phones. Tweets were attributed to specific users training the NB classifier with character n-grams and phones were linked to users by examining the GSM modulation characteristics. At a final stage these two classifiers were combined and the results indicated that the combination of natural-language and network-feature classifiers identifies a user to phone binding better than when the individual classifiers are used independently. The initial sample corpus contained 4,045 tweets from 53 active users and a following feed increased dramatically this number reaching 114,000 tweets. Preprocessing contained the removal of @replies, #tags from the text and tweets smaller than 3 words. Character n-grams were measured using n from 2 to 6 and fed the NB classifier. Additionally, an attempt to model the pattern of life of each user was undertaken using time stamps of the tweets and @replies as separate features. Results indicated that character n-gram feature vectors using separate tweets as basic text units performed poorly (40.3%) authorship attribution accuracy in the 50 authors' dataset). However, if each text unit contains multiple tweets the accuracy of authorship attribution improves radically (99.6%) with text units merging 23 tweets). Very high percentages of authorship attribution accuracy (94%) were also obtained by using @replies feature verifying further previous findings by Layton et al. (2010), while timestamps didn't perform as expected giving only 35% accuracy.

Research Methodology

The Greek Twitter Corpus

None of the available corpora of Modern Greek contain social media communication data. For this reason we had to develop from scratch a new corpus of tweets written in Modern Greek. We utilized *trending.gr* a free service that monitors the activity of Greek users in Twitter and publishes many statistics regarding their activity including the top users in followers, mentions etc. We selected 10 users based on their popularity (number of followers) and their activity (number of tweets in a month).

In order to extract tweets from the specific users we used the twitteR R package[2]. The Twitter API can only return up to 3,200 statuses per account including native retweets in this total. To cope with this API's rate limit restrictions, an incremental approach was adopted, keeping track of the most recent tweet id per author for each API and repeating the whole procedure in frequent time intervals. The descriptive statistics of the Greek Twitter Corpus (GTC) are displayed in table 1.

During the corpus preprocessing we removed all @replies, #tags and manual retweets (RT's). The main reason for this decision was that we wanted to develop an authorship attribution methodology that is based exclusively on linguistic features and study the effects of the training text-size without bias from extralinguistic features such as the social status of the user. Previous studies (Boutwell 2011, Layton et al 2010) have indicated that @replies is a powerful authorship marker since most of the times users

[2]http://cran.r-project.org/web/packages/twitteR/index.html

Authors	No of Tweets	Total size (words)	Average size (words)	Standard deviation
A	500	5,378	10.75	5.42
B	918	10,515	11.45	5.52
C	2,065	32,098	15.54	6.73
D	455	7,451	16.57	5.48
E	1,347	9,822	7.29	5.01
F	535	3,692	6.90	4.93
G	1,277	9,412	7.37	5.63
H	2,306	26,212	11.36	5.86
I	2,986	18,720	6.26	4.28
J	584	7,618	13.06	6.74
Total	**12,973**	**130,918**		

Table 1: GTC descriptive statistics.

share thoughts and information with a restricted circle of other users who belong to their social network. However, this is not a stable effect. Indeed, some users have a solid circle of contacts but there are also many users who use Twitter as one-way communication channel. In these cases we need a robust authorship attribution methodology which can identify authors using information only from the linguistic part of the tweet.

Author's Multilevel N-gram Profile

Character and word n-grams have been used successfully previously in AAI tasks with character bigrams to appear as early as 1976 in the relative literature (Bennett 1976). Modern applications of character and word n-grams used as features include Coyotl-Morales et al. 2006, Peng, Schuurmans and Wang 2004, Gehrke 2008, Luyckx and Daelemans 2011, Koppel, Schler and Argamon 2011, Grieve 2007 among others. They exhibit significant advantages over other stylometric features since their identification can be achieved easily and they are language-independent.

N-gram profiles, i.e. sequences of different n-gram sizes have also been used as feature group in AAI research but they are limited in one level each time, either using characters or words. Character n-gram profiles have been used by Keŝelj et al. 2003 and Peng 2003 in authorship attribution datasets in English, Modern Greek and Chinese language improving considerably the attribution accuracy over previous experiments in the same data. Furthermore, Peng et al. 2004 have compared word vs. character n-gram profiles in authorship attribution using Modern Greek dataset and concluded that word-level n-grams significantly outperform character-level models. They state that word and character-level n-grams function complementary and capture different text regularities and combining these two levels could result in more robust and accurate results (Peng et al. 2004).

Taking into consideration the complementary nature of character and word level information, we propose a combined vector of both character and word n-grams of different size. Following previous successful work in email authorship attribution (Mikros and Perifanos 2011) we extracted the 1,000 most frequent character and word n-grams with n=2 and 3 resulting in a total vector of 4,000 features. We used a tokenization pattern that allowed us to treat as separate tokens different punctuation marks (space included) in order to capture a large number of idiomatic sequences of characters typically met in tweets (e.g. ;..:-) .... ;;;; !;). The resulting vector represents the Author's Multilevel N-gram Profile (AMNP), a document representation that captures in a parallel way both character and word sequences. AMNP representation draws its roots from the Prague School of Linguistics and the notion of double articulation (Nöth 1995 p. 238). The specific notion states that language is organized in two separate layers or articulations. The first layer contains meaningful units (morphemes) and their sequence produce the grammatical pattern of language. The second layer contains only minimal functional units (phonemes) which do not carry any meaning themselves. It is the parallel combination of units from the second and first layer that produces grammatically correct linguistic production. Style could be analyzed in a similar manner. Stylistic information is constructed in blocks of segments of increasing semantic load, from character n-grams, to word n-grams (see figure 1).

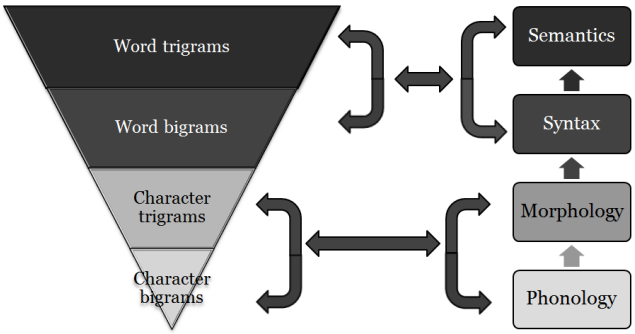

Figure 1: AMNP: An hierarchical representation of n-gram features and related linguistic levels.

Using AMNP we combine information from different linguistic levels and we capture stylistic variation across a wide range of linguistic choices. In all our features we calculated their normalized frequency in order to avoid text length bias in subsequent calculations. Feature normalized frequency (fnf) of a feature i in a document j is defined as follows:

$$fnf_{i,j} = \frac{frf_{i,j} * 100}{\sum_{k} frf_{k,j}}$$

where $frf_{i,j}$ is the raw frequency of the feature i in the document j multiplied by 100 and divided by the sum of number of occurrences of all features (k) in the document j, that is, the size of the document $|j|$.

Classification algorithm

For the purposes of the classification tasks, we used multiclass support vector classification by Crammer and Singer (2001), provided by LIBLINEAR library (Fan et al. 2008). This method is based on the generalization of the notion of

separating hyperplanes and margins for multiclass problems and yields significant time efficiency improvements.

Experimental Setup

The main research questions addressed in this study are:

1. Can we obtain reasonable authorship attribution accuracy in Modern Greek tweets?
2. Does AMNP produces better stylometric representations of each author, or separate n-grams profiles are more efficient?
3. What is the optimal way to construct training sets for this kind of study? More specifically:
   (a) Is it better to use single tweets for testing or do we need to merge tweets producing bigger text units?
   (b) In the case of merging tweets, what is the text size that produces the best attribution results?

In order to address the above research questions we used GTC to create 4 different datasets which contained merged tweets in texts of sizes (100, 75, 50 and 25 words). Furthermore, we tested the authorship attribution accuracy with each feature group separately and compared it with AMNP. Accuracy figures are calculated on two different conditions: a) 10-fold cross-validation (cv) in the merged tweets text units b) External test set which contained 500 single tweets not included in the training set (35-60 per author).

Results

Authorship attribution in Greek tweets can be performed with remarkable accuracy when we use a training set in which the basic text units contain merged tweets. Best results were obtained using 100-word and 75-words text chunks (0.952 and 0.918 respectively). The influence of the text size in the attribution accuracy in 10-fold cv is displayed in figure 2.

10-fold cv results confirms the basic premise of data science that there is nothing better to data than more data. Increasing text size produces better accuracy figures. Increase from 50 to 100-words sizes advances accuracy in a linear way. However, there is visible differentiation of the improvement rate when we are moving from 25 to 50-word chunks meaning that 50 words is a significant threshold for the stability of the AMNP.

A different picture was revealed when we used the external tweets as test set (figure 3). Accuracy rates seem to move to the opposite direction with smaller text-size chunks producing better attribution rates than bigger ones. Best authorship attribution accuracy (0.854) is observed when we use the smallest text-size chunk (25 words) while the worst (0.763 and 0.731) are taken from the training set with the 100 and 75-words text chunks respectively. This opposite effect could be related to the normalization of the measurement and the sparseness of our data. AMNP coupled with multi-class support vector classification exhibits patterns of symmetrical behavior and significant inequalities in the sizes of training and test documents seem to affect it. In general, our methodology displays lower accuracies in the external

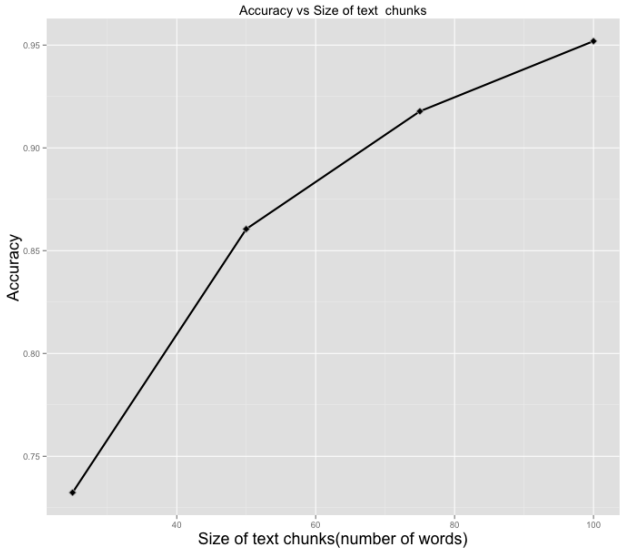

Figure 2: Impact of text size of merged tweets in the authorship attribution accuracy (10-fold cv).

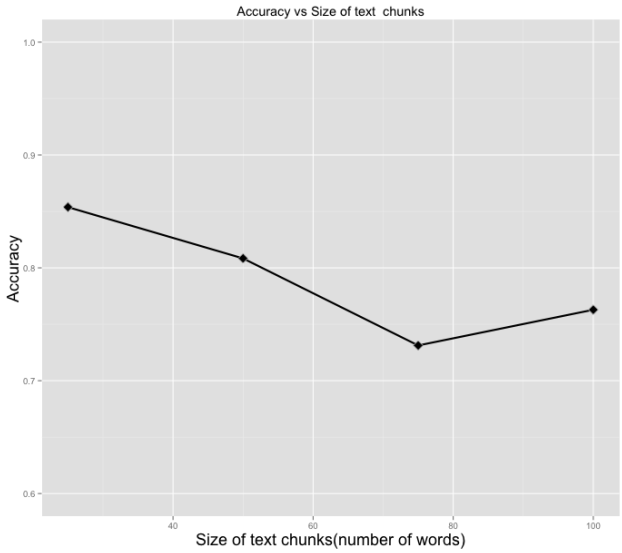

Figure 3: Impact of text size of merged tweets in the authorship attribution accuracy (external set of tweets)..

tweets test set than the 10-fold cv. This is however justifiable, because single tweets authorship attribution is a far more difficult task since tweets exhibit large variation in size and linguistic content.

A second experiment was conducted in order to evaluate whether AMNP representation captures better the stylometric profile of the tweets than using separate n-gram profiles. For this reason we repeated the authorship attribution task in the four datasets of varying text size chunks under both testing conditions (10-fold cv and external tweets). The results

are displayed in figures 4 and 5.

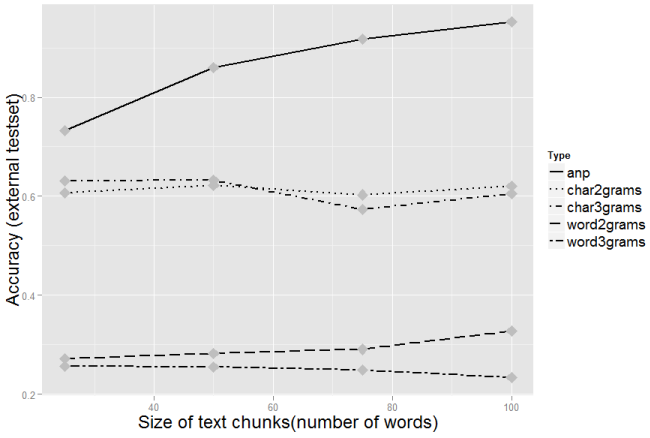

Figure 4: Impact of text size and feature representation method in authorship attribution accuracy using cross-validation

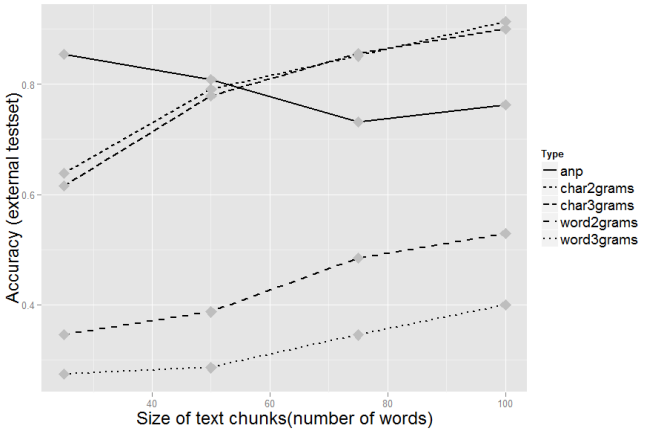

Figure 5: Impact of text size and feature representation method in authorship attribution accuracy using external test set

In the 10-fold cv evaluation AMNP supersedes any other type of feature representation across all text sizes. The observed differences were evaluated using one-sample t-test between percents. AMNP accuracies in each text size were compared with the corresponding accuracies achieved by the single feature groups in the same text size category. Since we had multiple comparisons a Bonferroni correction was applied to the p level of significance ($p = 0.01$) of all the t-test conducted. In all the comparisons employed, AMNP obtained statistical significant higher classification accuracy over each one of the single feature groups providing support to our claim that a combined feature group consisting of features from multiple and different linguistic levels captures more efficiently an author's style. In the external dataset AMNP performs slightly less than character 2-grams and 3-grams. The observed differences however are not statistically significant. One-sample t-test between percents was performed and AMNP accuracy compared to accuracies from each single feature group was found to be not statistically significant ($p > 0.05$) across all text sizes. This result indicates that AMNP is a robust document representation method even in the case of single tweets and should be preferred compared to single feature group representations.

Conclusions

The present study explored different authorship attribution methods in Modern Greek tweets. Since there wasn't any relative Modern Greek corpus available, we compiled one, named Greek Twitter Corpus (GTC). For the needs of our research we selected 10 of the most prolific Greek Twitter users and retrieved 12,973 tweets totaling 130,918 words. In this corpus we extracted the Author's Multilevel N-gram Profile (AMNP), i.e. n-gram features of increasing size and linguistic unit. In total we extracted the 1,000 most frequent character 2-grams, 3-grams and word 2-grams, 3-grams resulting in a vector of 4,000 features. This vector fed the multi-class support vector classification algorithm and its classification performance was evaluated using 10-fold cross-validation and external dataset of 500 actual tweets. Returning to our original research questions we found that:

1. Authorship attribution in tweets of Modern Greek is a feasible task. Our top performance (0.951 accuracy in 10-fold cv using 100-word text chunks) is a good indication that the tweet's linguistic structure is a significant carrier of authorship information. A near 10% accuracy drop was observed when we applied the same methodology in the external dataset of single tweets. However, 0.854 is still an acceptable accuracy rate taken into consideration that we are based only on linguistic cues and do not utilize any other user information (social network, topic, timestamps etc).
2. AMNP representation is based on a solid linguistic semiotic theoretical background. A large number of authorship attribution studies have previously implied that authorship traits span across a wide spectrum of linguistic levels. N-grams still represent one of the best suited features for representing stylometric profiles of small size texts. AMNP proved highly efficient compared to single n-gram feature groups in all text sizes.

3.(a) The obtained results indicated that optimal results are achieved when both training and testing sets for authorship attribution contained merged tweets.

(b) The text-size effect in the authorship attribution accuracy correlates directly with the test set. When the test set consists of text units containing merged tweets, we found that bigger texts imply higher attribution accuracies. However the trend is not linear. We identified that 50-words text chunks represent a significant threshold in our methodology, since authorship attribution accuracy falls sharply in text sizes smaller than 50 words. Using the external test dataset with single tweets we observed a different tension. Accuracy exhibits a negative correlation with text size. As the size of the text

units increases, accuracy drops, meaning that optimal performance is obtained only when training and test instances are roughly of the same size.

In the near future we plan to extend the present study in a number of ways. We will continue to develop GTC in order to enlarge its users and record their full metadata. Furthermore, we will expand AMNP feature representation with bigger n-grams and compare different machine learning algorithms in order to find the best blend. Finally, we will enrich our test sets with tweets from users that are not part of the training data in order to explore more systematically the efficiency of our methodology in cases of authorship verification.

References

Argamon, S.; Koppel, M.; Fine, J.; and Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3):321–346.

Argamon, S.; Dhawle, S.; Koppel, M.; and Pennebaker, J. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification, 8-12 Jun 2005*.

Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2007. Mining the blogosphere: Age, gender and the varieties of selfexpression.

Bennett, W. R. 1976. *Scientific and engineering problem-solving with the computer*. Englewood Cliffs, N.J.: Prentice Hall.

Boutwell, S. R. 2011. *Authorship attribution of short messages using multimodal features*. Ph.D. Dissertation, Naval Postgraduate School, Monterey, California.

Chaski, C. E. 2005. Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1):1–13.

Coyotl-Morales, R.; Villaseor-Pineda, L.; Montes-y Gmez, M.; and Rosso, P. 2006. Authorship attribution using word sequences. In *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4225 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 844–853.

Crammer, K., and Singer, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2:265–292.

Crystal, D. 2008. *Txting: The Gr8 Db8*. Oxford: Oxford University Press.

de Vel, O.; Anderson, A.; Corney, M. W.; and Mohay, G. 2001. Multi - topic e-mail authorship attribution forensics. In *Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*.

Denby, L. 2010. *The Language of Twitter: Linguistic innovation and character limitation in short messaging*. Ph.D. Dissertation, University of Leeds.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classication. *Journal of Machine Learning Research* 9:1871–1874.

Gehrke, G. T. 2008. *Authorship discovery in blogs using Bayesian classification with corrective scaling*. Ph.D. Dissertation, Naval Postgraduate School.

Grieve, J. W. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22(3):251–270.

Iqbal, F.; Binsalleeh, H.; Fung, B. C. M.; and Debbabi, M. 2010a. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation* 7(1-2):56–64.

Iqbal, F.; Khan, L. A.; Fung, B. C. M.; and Debbabi, M. 2010b. E-mail authorship verification for forensic investigation. In *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC '10), March 22-26, 2010, Sierre, Switzerland*, 1591–1598. New York: ACM.

Juola, P. 2008. Authorship attribution. *Foundations and Trends in Information Retrieval* 1(3):233–334.

Keelj, V.; Peng, F.; Cercone, N.; and Thomas, C. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference of Pacific Association for Computational Linguistics (PACLING'03), 22-25 August 2003, Dalhousie University, Halifax, Nova Scotia, Canada*, 255–264.

Koppel, M., and Schler, J. 2004. Authorship verification as a one-class classification problem. In *Proceedings of 21st International Conference on Machine Learning, July 2004*, 489–495.

Koppel, M.; Argamon, S.; and Shimoni, A. R. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4):401–412.

Koppel, M.; Schler, J.; and Argamon, S. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45(1):83–94.

Layton, R.; Watters, P.; and Dazeley, R. 2010. Authorship attribution for twitter in 140 characters or less. In *2nd Workshop on Cybercrime and Trustworthy Computing Workshop (CTC), 19-20 July 2010, Ballarat, Australia*, 1–8.

Li, J.; Zheng, R.; and Chen, H. 2006. From fingerprint to writeprint. *Communications of the ACM* 49(4):76–82.

Luyckx, K., and Daelemans, W. 2008a. Personae: A corpus for author and personality prediction from text. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odjik, J.; Piperidis, S.; and Tapias, D., eds., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28-30 May 2008*.

Luyckx, K., and Daelemans, W. 2008b. Using syntactic features to predict author personality from text. In *Proceedings of Digital Humanities 2008 (DH 2008)*, 146–149.

Luyckx, K., and Daelemans, W. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1):35–55.

Mendenhall, T. C. 1887. The characteristic curves of composition. *Science* 9(214):237–249.

Mikros, G. K., and Perifanos, K. 2011. Authorship identification in large email collections: Experiments using features that belong to different linguistic levels. In *Proceedings*

of PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19-22 September 2011, Amsterdam.

Mosteller, F., and Wallace, D. L. 1984. *Applied bayesian and classical inference. The case of The Federalist Papers*. New York: Springer-Verlag, 2nd edition.

Nöth, W. 1995. *Handbook of semiotics*. Bloomington: Indiana University Press.

Peng, F.; Schuurmans, D.; Keelj, V.; and Wang, S. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL '03), 12-17 April 2003, Budapest, Hungary*, volume 1, 267–274. Stroudsburg, PA, USA: Association of Computational Linguistics.

Peng, F.; Schuurmans, D.; and Wang, S. 2004. Augmenting naive bayes classifiers with statistical language models. *Journal of Information Retrieval* 7(3-4):317–345.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. 2006. Effects of age and gender on blogging. In *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 27-29 March 2006, Stanford, California*, 199–205.

Sousa-Silva, R.; Laboreiro, G.; Sarmento, L.; Grant, T.; Oliveira, E.; and Maia, B. 2011. twazn me!!! ;(automatic authorship analysis of micro-blogging messages. In Muoz, R.; Montoyo, A.; and Mtais, E., eds., *Natural Language Processing and Information Systems*, volume 6716 of *Lecture Notes in Computer Science*. Berlin / Heidelberg: Springer. 161–168.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3):538–556.

Van Halteren, H. 2007. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)* 4(1):1–17.