# Author Identification Using Imbalanced and Limited Training Texts

Efstathios Stamatatos
*University of the Aegean*
*stamatatos@aegean.gr*

## Abstract

*This paper deals with the problem of author identification. The Common N-Grams (CNG) method [6] is a language-independent profile-based approach with good results in many author identification experiments so far. A variation of this approach is presented based on new distance measures that are quite stable for large profile length values. Special emphasis is given to the degree upon which the effectiveness of the method is affected by the available training text samples per author. Experiments based on text samples on the same topic from the Reuters Corpus Volume 1 are presented using both balanced and imbalanced training corpora. The results show that CNG with the proposed distance measures is more accurate when only limited training text samples are available, at least for some of the candidate authors, a realistic condition in author identification problems.*

## 1. Introduction

Authorship analysis has a long history mainly due to research on literary works of disputed or unknown authorship [9]. In recent years, researchers have paid increasing attention to authorship analysis in the framework of practical applications, such as verifying the authorship of emails and electronic messages [2], plagiarism detection in student essays [13], and forensic cases [3].

Author identification is the task of predicting the most likely author of a text given a predefined set of candidate authors and a number of text samples per author of undisputed authorship [11]. This task can be seen as a single-label multi-class text categorization problem [10]. In many cases, at least for some of the candidate authors, there are extremely limited text samples available to be used for training. This constitutes the class imbalance problem and any author identification approach should be able to deal with it.

One great challenge is the definition of an appropriate text representation so that to reveal the stylistic choices of the author. Many stylometric features have been proposed, including function word frequencies, high-frequency word frequencies, vocabulary richness measures, word-class frequencies, syntactic analysis measures, grammatical errors etc.

A promising text representation technique for stylistic purposes, that has recently been proposed [6, 11], is a 'bag of character *n*-grams'. Character *n*-grams (contiguous characters of fixed length) are able to capture complicated stylistic information on the lexical, syntactic, or structural level. For example, the most frequent character 3-grams of an English corpus indicate lexical ('the', ' to', 'tha'), word-class ('ing', 'ed '), or punctuation usage ('. T', ' "T') information.

The Common N-Grams (CNG) approach to author identification [6] is based on profiles consisting of the most-frequent character *n*-grams found in a text. Although quite simple, the CNG method has achieved remarkable performance in many author identification experiments [5]. However, the distance measure used to compare the text profiles is not stable for large values of the profile length when extremely limited training text samples are available for at least one author [4]. In this paper, new distance measures, especially designed for author identification tasks, are presented. The proposed distance measures are not affected by high values of the profile length and provide more accurate results when only limited text samples (at least for some of the authors) are available, a realistic condition in author identification tasks. A series of experiments based on texts on the same topic from the Reuters Corpus Volume 1 (RCV1) are presented and the effectiveness of the proposed approach is evaluated when using both balanced and imbalanced training text corpora.

## 2. The common n-grams approach

The CNG method represents each text sample as a bag of character *n*-grams taking into account case-sensitive information [6]. CNG is a profile-based approach, that is, the available training texts per author are concatenated and, then, an author profile is extracted based on the resulting large training text. A profile *P* is a set of *L* pairs $\{(g_1, f_1), (g_2, f_2), \ldots, (g_L, f_L)\}$, where $g_1, g_2, \ldots, g_L$ are the *L* most frequent *n*-grams of

1529-4188/07 $25.00 © 2007 IEEE
DOI 10.1109/DEXA.2007.5

the text (in decreasing order) and $f_1, f_2,\ldots, f_L$ their normalized (wrt text-length) frequencies of occurrence, respectively. A test text sample is assigned to an author using a dissimilarity function to compare the test text profile with the profiles of all the candidate authors. Let **A** be the set of the candidate authors and $T_a$ be the training text (the concatenation of all text samples) of the author $a$ ($a \in \mathbf{A}$). For a given $n$ and $L$, consider $P(x)$ and $P(T_a)$ as the profile of the test text and the author $a$, respectively. If $f_x(g)$ and $f_{Ta}(g)$ are the frequencies of the $n$-gram $g$ in the test text and the author $a$'s training text, respectively, then the distance (or dissimilarity) measure $d_0$ between $P(x)$ and $P(T_a)$ is defined as follows:

$$d_0(P(x),P(T_a)) = \sum_{g\in P(x)\cup P(T_a)} \left( \frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)} \right)^2$$

where $f(g)$=0 if $g \notin P$. A $k$-nearest neighbor approach with $k$=1 is then followed in order to guess the most likely author of a text $x$:

$$author(x) = \arg\min_{a\in\mathbf{A}} d_0(P(x),P(T_a))$$

## 2.1. Pros and cons

The CNG method has a number of important advantages. First, it is language-independent. The bag of $n$-grams representation is able to capture stylistic properties on the lexical, the syntactical level, and the use of upper case, punctuation etc. The CNG method is easy to follow. No stemming or other 'deep' NLP techniques are involved for preprocessing the text.

Moreover, it has been proved to be quite effective for author identification problems. Keselj *et al.* [6] tested this approach in various test collections of English, Greek, and Chinese text, improving previously reported results. Moreover, the CNG method achieved the best results in the ad-hoc authorship attribution contest [5], a competition based on a collection of 13 text corpora in various languages (English, French, Latin, Dutch, and Serbian-Slavonic). The performance of CNG was remarkable especially in cases with multiple candidate authors (>5). In addition, the basic idea of this method has been applied to other problems of similar characteristics, like source code author identification [4], detection of malicious code [1], and lexical analysis of spontaneous speech [12].

As a profile-based method, CNG has the advantage of using one big training file per author. So, there is no need for having (or segmenting a big training text into) multiple text samples per author to be used as training set for a machine learning algorithm, such as support vector machines.

On the other hand, there are some drawbacks to this method. In more detail, it has two basic parameters that have to be tuned in order to obtain the best results for a given corpus: $n$-gram length ($n$) and profile size ($L$). Experimental results in a variety of corpora have shown that $3\leq n\leq 5$, and $1{,}000\leq L\leq 5{,}000$ give the best results in most of the cases. However, the exact values of $n$ and $L$ for getting the best results have to be found using a validation corpus. Moreover, larger values of $L$ have not been tested thoroughly. Especially when a text is short, it is not convenient to deal with a predefined profile length. For example, the predefined profile length may be $L$=5,000 and all the ordered $n$-grams from 4,000 to 6,000 may have the same normalized frequency. In such cases, it is convenient to have a predefined profile length as high as possible.

Another problem has to do with the particular distance function ($d_0$). In case one author profile is shorter than $L$ (the predefined profile length), that author has an important advantage over the others. In more detail, if a $n$-gram $g$ of $P(T_a)$ is not included in the $P(x)$ (i.e., $f_x(g)$=0), then its contribution to $d_0$ is $2^2$=4. If an author profile $P(T_{a1})$ is shorter than another author profile $P(T_{a2})$, then $P(T_{a1})$ will have a higher proportion of $n$-grams in common with $P(x)$ than will $P(T_{a2})$. Hence, $d_0(P(x), P(T_{a1}))$ is more likely to be smaller than $d_0(P(x), P(T_{a2}))$. Unfortunately, this case is not unrealistic in author identification problems. Very often, many training texts are available for one candidate author, while only a few training texts for another. In such a case, the longest profile length (i.e., the total number of distinct $n$-grams) for that author will be much shorter than the others.

The latter problem of CNG was first indicated in [4] where an alternative distance measure is proposed. Given the simplified profile $SP=\{g_1, g_2,\ldots, g_L\}$ of a text (the set of the $L$ most frequent $n$-grams, for predefined values of $n$ and $L$), the similarity of a test text to the training text of the author $a$ is the Simplified Profile Intersection (SPI):

$$SPI(SP(x),SP(T_a)) = |SP(x) \cap SP(T_a)|$$

where |.| denotes the cardinality. In other words, *SPI* merely counts the common $n$-grams in the test text and the author profiles. Note that *SPI* does not make use of the frequency information for each $n$-gram. Moreover, it is a similarity (rather than dissimilarity) measure, that is, the higher the $SPI(SP(x), SP(T_a))$, the more likely for the test text $x$ to be assigned to author $a$:

$$author(x) = \arg\max_{a\in\mathbf{A}} SPI(SP(x),SP(T_a))$$

**Table 1. Details for the text corpora used for evaluation in this study.**

| | C50ir | C50ig | C50b50 | C50b10 |
|---|---|---|---|---|
| Training corpus | Imbalanced | Imbalanced | Balanced | Balanced |
| Test corpus | Imbalanced | Balanced | Balanced | Balanced |
| Training corpus (text samples) | 7,962 | 1,234 | 2,500 | 500 |
| Test corpus (text samples) | 883 | 2,500 | 2,500 | 2,500 |
| Longest training text (KB) | 812 | 170 | 179 | 43 |
| Shortest training text (KB) | 288 | 6 | 100 | 18 |
| Longest training profile (3-grams) | 11,817 | 7,326 | 7,955 | 4,504 |
| Shortest training profile (3-grams) | 8,244 | 1,807 | 5,956 | 2,890 |

This measure has been successfully applied to source code author identification tasks. However, it suffers when all the authors but one have very short profiles (i.e., when there are many training texts available for one author and only a couple for all the others). In such a case, (for large values of *L*) the author with the long profile will be the most likely author. In such cases, the training texts for that author can be reduced so that the training text size per author to be more balanced. Note also that the normalization of the *SPI* measure (e.g., by the size of the author profile) did not work well in preliminary experiments.

### 2.2. New dissimilarity measures

In order to improve the CNG method a number of new dissimilarity (or distance) functions, especially designed for the author identification problem, are introduced in this paper. First, a simple variation of $d_0$ is to account only for the *n*-grams that belong to the test text profile (i.e., $g \in P(x)$):

$$d_1(P(x),P(T_a)) = \sum_{g\in P(x)} \left(\frac{2(f_x(g)-f_{T_a}(g))}{f_x(g)+f_{T_a}(g)}\right)^2$$

Note that although $d_0$ is a symmetrical function (i.e., $d_0(x,y)=d_0(y,x)$), $d_1$ is not. The first argument of $d_1$ should be the test text profile and the second argument should be an author's training profile. That way, it is ensured that all the distances of the test text profile from the training profiles will be calculated based on the same number of terms, which is equal to the predefined profile length *L* (or the size of *P*(*x*) in case it is shorter than *L*). Hence, a short author profile will not affect the overall accuracy of the method for high values of *L*. However, in case the test text is much longer than the training text of a candidate author, that author is less likely to be selected (i.e., more *n*-grams of the test text profile will not be included in the profile of that author, hence, the distance measure increases too much). On the other hand, this is not a realistic scenario in author identification. Usually, the test text is shorter than the concatenation of all the training texts per author. In any case, the test text can be easily divided into smaller segments of fixed length.

The second proposed distance measure is based on the 'training corpus norm', that is, the concatenation of all available training texts of all candidate authors. The corpus norm profile indicates what the average profile for all the training texts should look like. The new distance measure $d_2$ is an extension of $d_1$ and incorporates the distance of the test text profile *P*(*x*) from the corpus norm profile *P*(*N*):

$$d_2(P(x),P(T_a),P(N)) = \sum_{g\in P(x)} \left(\frac{2(f_x(g)-f_{T_a}(g))}{f_x(g)+f_{T_a}(g)}\right)^2 \cdot \left(\frac{2(f_x(g)-f_N(g))}{f_x(g)+f_N(g)}\right)^2$$

where $f_N(g)$ is the normalized frequency of the *n*-gram *g* in the corpus norm profile ($f_N(g)=0$ if $g \notin P(N)$). The second term of $d_2$ can be viewed as a weight to each *n*-gram of the *P*(*x*). The more a *n*-gram *g* of *P*(*x*) deviates from its 'normal' frequency $f_N(g)$, the more it contributes to the distance function. If $f_x(g)= f_N(g)$, *g* is not taken into account (its weight is zero). In addition, $d_2$ (like $d_1$) should also be more stable for high values of *L* in comparison to $d_0$.

## 3. Evaluation corpora

The corpus used in this study is taken from the RCV1. Parts of the RCV1 corpus has already been used in author identification experiments. In [7] the top 50 authors (with respect to total size of articles) were selected. Moreover, in the framework of the *AuthorID* project, the top 114 authors of RCV1 with at least 200 available text samples were selected [8]. In contrast to these approaches, in this study, the criterion for selecting the authors was the topic of the available text samples. First, a list of 27,754 duplicates (exactly the same or plagiarized texts) were removed [7]. Then, the top 50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) were selected. That way, it is attempted to minimize the topic factor in distinguishing among the texts.

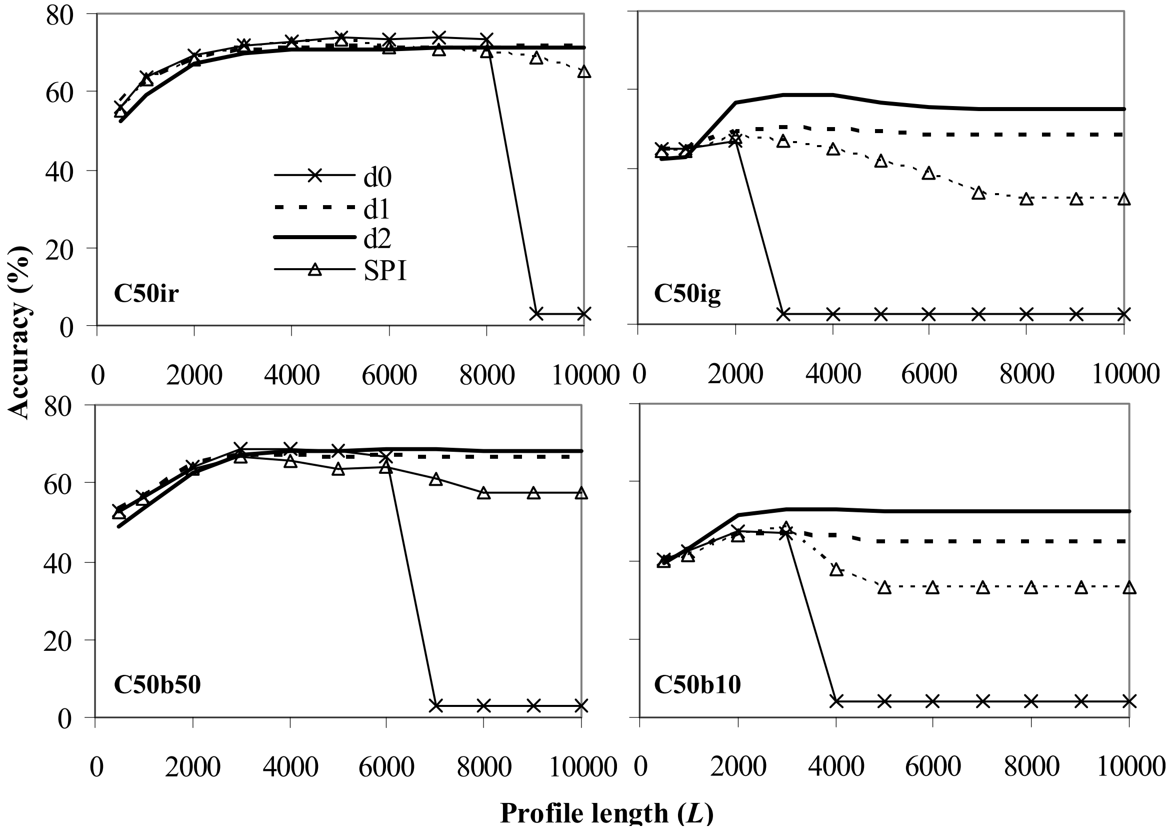

**Figure 1. Performance of CNG with $d_0$, $d_1$, $d_2$, and *SPI* for $n$=3 and various *L* values.**

Therefore, since steps to reduce the impact of genre have been taken, it is to be hoped that authorship differences will be a more significant discriminating factor. Hereafter, this corpus will be called C50. The C50 was used to produce both balanced and imbalanced training/test corpora. In this study, four cases were tested:

**C50ir:** All the available texts per author were divided into 10 equal parts (measured in texts per author), and, then, 9 out of 10 formed the training corpus while the remaining part formed the test corpus. We call this corpus 'imbalanced representative' because its distribution over the authors reflects the initial distribution of C50.

**C50b50:** 50 texts per author were used for training and another (not overlapping) set of 50 texts per author were used for test.

**C50b10:** 10 texts per author were used for training. The test corpus remains the same with that of C50b50.

**C50ig:** A Gaussian distribution was applied to the training texts of C50b50 resulting in an imbalanced training corpus ranging from 2 training texts for some authors to 50 training texts for other authors. That better resembles a realistic case. We call this 'imbalanced Gaussian' corpus.

More details for these corpora are given in Table 1. The 'longest/shortest training text' line refers to the texts produced after the concatenation of all training texts per author. Similarly, 'longest/shortest training profile' refers to the profiles extracted from the author training texts when all the *n*-grams are considered.

## 4. Results

The CNG method has been applied to the corpora C50ir, C50ig, C50b50, and C50b10. In each case, the training texts of the authors produce the author profiles. Then, the distance between each test text and the author profiles based on $d_0$, $d_1$, $d_2$, and SPI provides the most likely author. The microaverage accuracy results are shown in Figure 1. In all cases, 3-grams were used and various values of *L* were examined (from 500 until 10,000).

As can be seen, until a certain point, all the distance measures have a similar performance curve. This point corresponds to shortest training profile (see Table 1). When *L* exceeds the shortest training profile, the performance of $d_0$ falls rapidly to very low accuracy rates in all cases. This verifies what has been underlined in the previous section, that is, the performance of CNG with $d_0$ is dramatically affected when the predefined profile length ($L$) is longer than the longest possible profile of at least one author (all the texts are assigned to that author). *SPI* is also affected but not so drastically. On the other hand, the performance of $d_1$ and $d_2$ remains at the same high level indicating that these new measures are more robust for high values of profile length. Actually, in the cases of C50ig and C50b10 the performance of CNG with $d_2$ continues to increase beyond the shortest training profile limit.

In most cases $d_2$ outperforms $d_1$. This is an indication that the training corpus norm weighting factor of $d_2$ considerably assists the classification procedure, especially when limited texts per author are available for training. On the other hand, the performance of $d_1$ and $d_2$ for C50ir is a strong indication that the training corpus norm does not significantly contribute to the classification model when the training text size per author is quite long.

## 5. Conclusions

This study presented new distance measures for the CNG method for author identification. The proposed approach provides a more stable solution than traditional CNG for high values of profile length. This is particularly important, especially in cases where there are only limited training texts for at least one of the candidate authors. On the other hand, when adequate training text samples are available, the traditional CNG outperforms CNG with $d_2$. However, especially in forensics or criminal investigation applications, the latter is not a representative case.

All the experiments performed in this study were based on character 3-grams. Similar results can be achieved when using longer *n*-grams (i.e., 4-grams and 5-grams). In that case, the shortest (and longest) profile length is considerably increased and the classification results are slightly improved. However, the selection of the best *n* value depends on the particular corpus used and the language of the texts.

A promising future work direction is to use variable-length *n*-grams to achieve even better classification results. Moreover, open-set author identification (i.e., when the true author of an unseen text is not necessarily among the candidate authors), another realistic scenario, should be examined thoroughly.

## References

[1] T. Abou-Assaleh, N. Cercone, V. Keselj, and R. Sweidan, "Detection of New Malicious Code Using N-gram Signatures". *Proc. of the 2nd Annual Conference on Privacy, Security, and Trust*, 2004.

[2] S. Argamon, M. Saric, and S. Stein, "Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results". *Proc. of the 9th ACM SIGKDD*, 2003, 475-480.

[3] C. Chaski, "Empirical Evaluations of Language-based Author Identification Techniques". *Forensic Linguistics*, 8(1), 2001, pp. 1-65.

[4] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. Katsikas, "Effective Identification of Source Code Authors Using Byte-level Information". *Proc. of the 28th Int. Conf on Software Engineering*, 2006.

[5] P. Juola, "Ad-hoc Authorship Attribution Competition". *Proc. of ALLC/ACH Joint Conf.*, 2004, pp. 175-176.

[6] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based Author Profiles for Authorship Attribution". *Proc. of the Conf. of Pacific Association for Computational Linguistics*, 2003.

[7] D. Khmelev, and W. Teahan, "A Repetition Based Measure for Verification of Text Collections and for Text Categorization". *Proc. of the 26th ACM SIGIR*, 2003, pp. 104–110.

[8] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author Identifcation on the Large Scale". *Proc. of CSNA*, 2005.

[9] F. Mosteller, and D. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, New York, 1984.

[10] F. Sebastiani, "Machine Learning in Automated Text Categorization". *ACM Computing Surveys*, 34(1), 2002, pp. 1-47.

[11] E. Stamatatos, "Ensemble-based Author Identification Using Character N-grams". *Proc. of 3rd Int. Workshop on Text-based Information Retrieval*, pp. 41-46.

[12] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic Detection and Rating of Dementia of Alzheimer Type Through Lexical Analysis of Spontaneous Speech". *Proc. of IEEE ICMA*, 2005.

[13] H. van Halteren, "Linguistic Profiling for Author Recognition and Verification". *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 199-206.