G. Sidorov
Instituto Politécnico Nacional, Av. Juan de Dios Batiz, s/n, 07320,
Ciudad de México, Mexico

Automatic Authorship Attribution Using Syllables as Classification Features

For the Authorship Attribution (AA) task, some categories of character n-grams are more predictive than others, both under single- and cross-topic AA conditions. Taking into account the good performance of character n-grams, in this paper, we examine different features: various types of syllable n-grams as features (for single- and cross-topic AA in the English and Spanish languages). We show that the the use of syllable n-grams is usually better than bag-of-words baseline and it overcomes the results obtained using character n-grams in one dataset of three (by 6%), on the other two corpora the difference in results is not favorable, but it is relatively small (2%).

Keywords: computational linguistics, machine learning, authorship attribution, stylometry, feature selection, syllables.

Г. Сидоров
Национальный Политехнический Институт, Ав. Хуан де Диос Батис, б/н,
07320, город Мехико, Мексика

Автоматическое определение авторства с использованием слогов как классификационных признаков

Для задачи автоматического определения авторства, некоторые типы буквенных n-грамм дают лучшие результаты чем другие типы, как для текстов написанных на одну тему, так и для текстов на разные темы. Базируясь на хороших результах буквенных n-грамм для определения авторства, в данной статье мы применяем другие признаки (похожие на буквенные n-граммы): разные типы слоговых n-грамм (для опредения авторства текстов как на одну тему, так и для текстов на разные темы, для английского и испанского языков). Мы показываем, что полученные результаты лучше чем

результаты со словами, и в некоторых случаях (на одном корпусе из трех) превосходят результаты полученные с буквенными n-граммами (на 6%), в двух же других корпусах разница в результатах не в нашу пользу, но она достаточно мала (2%).

Ключевые слова: компьютерная лингвистика, машинное обучение, определение авторства, стилометрия, выбор признаков, слоги.

1. Introduction

Mainstream of the modern computational linguistics is based on application of machine learning methods. We represent our task as a classification task, represent our objects formally using features and their values (constructing vector space model), and then apply well-known classification algorithms. In this pipeline, the crucial question is how to select the features. For example, we can use as features words, or n-grams of words (sequences of words) or sequences of characters (character n-grams), etc. An interesting question arises: Can we use syllables as features? It is very rarely done in computational linguistics, but there is certain linguistic reality behind syllables. This paper explores this possibility for the Authorship Attribution task.

The Authorship Attribution (AA) task aims at determining who authored an anonymous text given text samples of a number of candidate authors [Juola 2008]. The history of AA research goes back to the late nineteenth century, when [Mendenhall 1887] studied the disputed authorship among Bacon, Marlowe, and Shakespeare.

In twentieth century, there were several notorious cases related with the authorship. One of them is *The Federalist* papers published in 1787-1788. They contain 85 essays arguing in favor of the U.S. Constitution, which are considered a remarkable example of political prose. There were three authors of these essays. In 63 cases, the authorship is known, while in other 12 cases, it is disputed and it was the object of early stage authorship attribution studies [Mosteller and Wallace 1964].

Another case is if the author of The Gospel of Luke also wrote the Book of Acts. There are opposite points of view. Many arguments are based on theological ideas or circumstantial facts, like was Luke the doctor or not.

Another notable case is the dispute about the authorship of the novel “And Quiet Flows the Don” by Mikhail Sholokhov. Some critics claimed that he was too young at the moment of its writing having just 23 years, while the novel is very mature. It was suggested that the other possible author is Fyodor Kryukov, who also wrote about the life of Cossacks. It was supposed that after the death of F. Kryukov in 1920, M. Sholokhov might have access to his manuscripts and used them. A study in 1977 by Scandinavian scientists using existing at that moment computing tools and techniques (sentence lengths, frequencies of several frequent POS n-grams, etc.) confirms the authorship of M. Sholokhov. Maybe, it is worth trying to compare

previous results in these cases with modern authorship attribution approaches based on machine learning.

In recent years, the AA task has experienced an increase in interest due to the growing amount of textual data available on the Internet and a wide range of AA applications: it contributes to marketing, electronic commerce, and security purposes, as well as to terrorism prevention and forensics applications, for example, by limiting the number of candidate authors of a text under investigation [Abbasi and Chen 2000]. The interest in this task is partly driven by the annual organization of the PAN evaluation campaign[1], which is held as part of the CLEF conference and is considered as the main *fora* for AA and relates tasks on digital text forensics.

There are two common approaches to tackle the AA task: statistical and machine learning. Statistical methods were widely used in earlier studies. They include histograms of word-length distribution of various authors [Mendenhall 1887], principle component analysis of function words [Burrows 1987], etc. Machine learning approach consists in representing the objects (text samples) as a set of features and their values, that is, building a vector space model (VSM). This model defines a space of N dimensions; each dimension in this space corresponds to a feature. Then a machine learning algorithm is applied to training data to learn to assign class labels (authors' names) to objects (text samples) using previously selected features. Thus, the AA task from a machine learning perspective is targeted as a multi-class, single-label classification problem, in which the set of class labels (authors' names) is defined beforehand.

Authors' writing style is the most important information required for solving AA problems. Character n-grams are considered among the most discriminating stylometric features for both single- and cross-topic AA task [Stamatatos 2013; Sapkota et al. 2015; Markov et al. 2017; Markov et al. 2017c], as well as for similar tasks, such as Author Profiling [Markov et al. 2017b], and others. In single-topic AA task, all the authors write about the same topic, while in cross-topic AA the thematic areas of training and test corpora are disjoint [Stamatatos 2013]. Cross-topic AA is a more realistic scenario to the development of practical applications of this task, since the thematic area of the target anonymous text is often different from the documents available for training a machine learning classifier. One possible explanation of the good performance of character n-gram features is their capacity to capture various aspects of authors' stylistic choices, including lexical and syntactic nuances, as well as punctuation and capitalization information [Stamatatos 2009; Kestemont 2014]. [Sapkota et al. 2015] showed that not all categories of character n-grams are equally indicative. They claim that character n-grams that capture information concerning affixes and punctuation marks (morpho-syntactic and stylistic information) are more effective than when using the whole set of character n-grams, that is, when including those n-grams that capture thematic content (word-like n-grams).

[1] http://pan.webis.de [last access: 27.12.2016]. All other URLs in this document were also verified on this date.

One of the challenges when using character n-grams is to determine the optimal size of these features. The size can depend on the language and corpus, as well as on the category of character n-grams [Markov et al. 2017]. Syllables are able to capture the same information as typed character n-grams, being linguistically adjusted to the appropriate size. It makes it interesting to examine the predictiveness of syllables as features for the task of AA. Moreover, to the best of our knowledge, no work has been done on using syllables as features for this task.

In this paper, we conduct experiments adopting the categories of character n-grams proposed by [Sapkota et al. 2015] to syllables. We examine the predictiveness of these features under single- and cross-topic AA conditions using English corpora, as well as under cross-topic AA conditions using a Spanish dataset. We compare the obtained results with the typed character n-grams approach proposed in [Sapkota et al. 2015], with the bag-of-words approach, and with random baseline.

The research questions addressed in this paper are the following:

1. Can syllables be used as predictive features for single- and cross-topic AA task?
2. Which categories of syllables are more effective for the English and Spanish languages?
3. Is the conclusion reported in [Sapkota et al. 2015], that for the English language, the best performing model is based solely on affix and punctuation n-grams, valid also for syllables? Is this conclusion valid for the Spanish language?

The remainder of this paper is organized as follows. Section 2 describes related work in single- and cross-topic AA. Section 3 presents the procedure of adopting character n-gram categories to syllables. Section 4 describes the datasets used in this work. Section 5 and 6 present the experimental settings and the obtained results. Finally, Section 7 draws the conclusions from this work and points to the directions of future work.

2. Related Work

Over the last years, a large number of methods have been applied to Authorship Attribution (AA) problems. Most of the prior work focused on single-topic AA conditions, when texts for training and evaluation are written on the same topic. A great variety of feature types aiming at capturing stylistic properties of author's style, feature representations, and machine learning algorithms were examined for single-topic AA (see [Stamatatos et al. 2014] for a more detailed overview of single-topic AA studies). Function words [Kestemont 2014; Holmes 1994], part-of-speech n-grams [Diederich et al. 2003], functional lexical features [Argamon et al. 2007], and functions of vocabulary richness [Stamatatos et al. 2000] are considered to be reliable markers of author's style.

[Qian et al. 2014] showed that the combination of different feature types improves the performance of the AA models.

[Van Halteren 2004] introduced a linguistic profiling technique, when counts of linguistic features are considered to compare separate authors to average profiles.

[Luyckx and Daelemans 2008] analyzed the effect of using a larger set of authors (145 authors) on feature selection and learning, and the effect of using limited training data in the AA task.

[Koppel and Winter 2014] introduced the "impostors" method based on repeated feature sub-sampling methods.

[Gómez-Adorno et al. 2015] achieved promising results extracting textual patterns based on features obtained from shortest path walks over integrated syntactic graphs. However, a single-topic condition considerably simplifies the realistic scenario of AA problems, since in many realistic situations in which stylometry is applied (e.g., forensics), it is very unlikely to obtain examples of the writing on the same topic and genre.

Recently, the focus of the AA community has shifted towards cross-topic and cross-genre AA conditions [Stamatatos et al. 2015], which is a more challenging but yet a more realistic scenario to the development of practical applications of this task, since style is affected by both genre and topic. [Posadas-Durán et al. 2016] demonstrated that feature representation plays an important role in this task. The authors achieved high performance using doc2vec-based feature representation.

[Sapkota et al. 2014] showed that out-of-topic data, that is, training texts from multiple topics instead of a single topic, allows achieving higher results under cross-topic AA conditions.

Various independent studies report a substantial accuracy drop under cross-topic AA conditions when compared to single-topic and single-genre conditions, suggesting that obtaining high cross-topic and cross-genre results remains challenging in AA [Stamatatos et al. 2015. Detailed overview of cross-topic and cross-genre AA studies is given in [Stamatatos et al. 2015].

Several prior studies have demonstrated the predictiveness of character n-gram features both under single- and cross-topic AA conditions [Stamatatos 2013; Sapkota et al. 2015; Markov et al. 2017; Luyckx and Daelemans 2008]. These language-independent stylometric features have proved to provide good results in this task due to their sensitivity to both the content and form of a text, among other reasons [Kestemont 2014; Daelemans 2013].

More recently, the work by [Sapkota et al. 2015] showed that some categories of character n-grams perform better than others for the English language. The authors concluded that there is no need to consider the whole set of character n-gram features, claiming that excluding word-like n-grams enhances AA accuracy.

2.1. Syllables and their Use in Authorship Attribution

Most of the studies in AA [Fucks 1952; Grieve 2005] and other natural language processing (NLP) tasks, such as Author Profiling [Pentel 2015], text cohesion and text difficulty estimation [McNamara et al. 2010], automatic readability assessment [Feng et al. 2010], among others, that explore syllables, use the average number of syllables per word as features and not syllables as such. Reviewing the AA-related literature, we did not encounter any mentions of using syllables as features for this task nor for related tasks. This makes it important to study the impact of syllables and different categories of syllables in single- and cross-topic AA. The assessment of their performance in the English and Spanish languages is the basis of the main research question addressed in this work.

The absence of application of syllables in the tasks of the computational linguistics is surprising. Especially if we compare it with the wide use of character n-grams. In fact, syllables is a linguistic reality, especially it is a psycholinguistic reality. Say, when a person starts learning to read, he usually is composing letters into syllables. On the other hand, the first phonetic writing systems were based on syllables, which later turned into letters corresponding to single sounds.

We will not discuss in this paper the definition of syllables. Note that we are dealing with the written texts, so for our purposes it is sufficient to have some syllabification rules. Further we consider various types of syllables depending on the word structure (position of a syllable in a word), similar to typed/untyped character n-grams.

3. Untyped and Typed Syllables

In this work, the same categories of character n-grams as introduces in [Sapkota et al. 2015] are applied to syllables. All character n-grams and syllables are grouped into three super categories (affix-, word-, and punctuation-related n-grams/syllables). The definitions of the categories of character n-grams and syllables are provided below. We slightly refine some of the original definitions of character n-gram categories provided in [Sapkota et al. 2015], with the aim of making them more accurate. Thus, the categories of character n-grams are the following:

Affix character n-grams

prefix: An n-gram that covers the first n characters of a word that is at least n+1 characters long.

suffix: An n-gram that covers the last n characters of a word that is at least n+1 characters long.

space-prefix: An n-gram that begins with a space and that does not contain any punctuation mark.

space-suffix: An n-gram that ends with a space, that does not contain any punctuation mark, and whose first character is not a space.

Word character n-grams

whole-word: An n-gram that encompasses all the characters of a word, and that is exactly n characters long.

mid-word: An n-gram that contains n characters of a word that is at least n+2 characters long, and that does not include neither the first nor the last character of the word.

multi-word: An n-gram that spans multiple words, identified by the presence of a space in the middle of the n-gram.

Punctuation character n-grams (abbreviated as **punct**)

beg-punct: An n-gram whose first character is a punctuation mark, but the middle characters are not.

mid-punct: An n-gram whose middle character is a punctuation mark.

end-punct: An n-gram whose last character is punctuation, but the first and the middle characters are not.

The similar categories of syllables are the following:

Affix syllables

prefix: First syllable of a word that has at least two syllables.

suffix: Last syllable of a word that has at least two syllables[2].

Word syllables

whole-word: One-syllable word.

mid-word: Middle syllables of a word that has at least three syllables.

multi-word: Last syllable of a word and first syllable of the next word.

Punctuation syllables (abbreviated as **punct**)

beg-punct: A punctuation mark followed by first syllable of the next word.

mid-punct: Last syllable of a word followed by a punctuation mark followed by first syllable of the next word.

end-punct: Last syllable of a word followed by a punctuation mark.

Tables 1 and 2 show the extracted character n-grams and syllables, respectively, when applying the proposed categories to the following sample sentence:

(1) *She said, "My mother will arrive tomorrow night."*

[2] For syllables, space-prefix and space-suffix categories are omitted since they correspond to prefix and suffix categories.

Table 1. Character n-grams (n=3) per category for the sample sentence (1), where SC stands for Super Category.

SC	Category	Character trigrams
Affix	***prefix***	*sai mot wil arr tom nig*
	suffix	*aid her ill ive row ght*
	space-prefix	*_sa _mo _wi _ar _to _ni*
	space-suffix	*he_ My_ er_ ll_ ve_ ow_*
Word	***whole-word***	*She*
	mid-word	*oth the rri riv omo mor orr rro igh*
	multi-word	*e_s y_m r_w l_a e_t w_n*
Punct	***beg-punct***	*,_“ “My*
	mid-punct	*d,_ “M_ t.”*
	end-punct	*id, ht.*

Table 2. Syllables per category for the sample sentence (1), where SC stands for Super Category.

SC	Category	Syllables
Affix	***prefix***	*moth ar to*
	suffix	*er rive row*
Word	***whole-word***	*She said My will night*
	mid-word	*mor*
	multi-word	*She_said My_moth er_will will_ar rive_to row_night*
Punct	***beg-punct***	*“My*
	mid-punct	*–*
	end-punct	*said, night.*

We examine three models of syllables, using the models applied by Sapkota et al. (2015) to character n-grams. Note that we use different feature sets (syllables) in each case, i.e., we obtained these feature sets in different manner.

1. ***All-untyped***: when the categories of syllables are ignored; any distinct syllable is a different feature.
2. ***All-typed***: when syllables of all available categories (**affix+word+punct**) are considered.
3. ***Affix+Punct***: when the syllables of the **word** category are excluded.

The conclusion of [Sapkota et al. 2015] was that in the English language, models based on *affix+punct* features were more efficient than models trained using all the features. In this paper, these three models were applied in order to examine whether this conclusion is also valid for syllables; moreover, we examine whether this conclusion is also valid for the Spanish language.

4. Datasets

We conduct two sets of experiments using (i) English single- and cross-topic corpora and (ii) using Spanish cross-topic corpus. In case of English, we use the same datasets as in [Sapkota et al. 2015]. That is, for single-topic experiments, we use a subset of the Reuters Corpus Volume 1 (Lewis et al. 2004), which consists of corporate news written by 10 different authors with 100 newswire stories per author on the same thematic area. Following [Sapkota et al. 2015], balanced training and test settings of this corpus were considered. We refer to this single-topic English corpus as CCAT 10. For cross-topic experiments in English, we used *The Guardian* corpus. This corpus is composed of opinion articles published in *The Guardian* newspaper in four thematic areas (Politics, Society, World, and U.K.). The texts were written by 13 different authors. Following previous studies [Stamatatos 2013; Sapkota et al. 2015], ten documents per author were considered for each of the four thematic areas.

A new cross-topic Spanish corpus was built automatically using crawler developed in the Python programming language. Given a set of URL seeds, the crawler extracted the names of the authors and the corresponding articles from the news website Cultura Collectiva[3]. The developed Spanish corpus consists of articles written in six thematic areas (Cinema, Food, Photography, Art, Design, and Lifestyle) by 6 different authors. We did not construct similar corpus for single-topic condition for Spanish, since we did not have enough resources for this. We preferred cross-topic condition, which is the modern trend.

The corpus is unbalanced in terms of documents written by authors on six topics (Table 3), since the use of a balanced subset of the corpus was not feasible due to a very short number of authors with a relevant number of texts in all six considered topics. Therefore, the developed cross-topic Spanish corpus addresses more challenging but at the same time more realistic AA conditions, when the same number of text samples written by different authors is not available.

Table 4 shows some of the statistics of the CCAT 10, *The Guardian*, and Spanish corpora.

Table 3. Number of documents written by authors on six topics for the Spanish corpus.

Author #	Cinema	Food	Photo.	Art	Design	Lifestyle	Total per author
Author 1	52	16	16	22	9	34	149
Author 2	51	16	17	31	12	54	181
Author 3	26	5	21	18	44	33	147
Author 4	6	4	9	13	5	14	51

[3] http://CulturaCollectiva.com

Author 5	14	7	4	13	8	12	58
Author 6	9	2	7	12	12	15	57
Total per topic	158	50	74	109	90	162	–

Table 4. Statistics of the CCAT 10, *The Guardian*, and Spanish corpora.

Corpus	#authors	#docs/author/topic	#sentences/doc	#words/doc
CCAT 10	10	100	19	425
The Guardian	13	13	53	1034
Spanish Corpus	6	see Table 3	36	962

To perform tokenization, we used Natural Language Toolkit[4] tokenizer.

5. Automatic Syllabification

After analyzing existing modules for syllabic division for both English and Spanish (Pyphen[5], PyHyphen[6], and Hyphenate[7]), we noticed that a large number of words encountered in the corpora are not present in the dictionaries of these modules and/or are divided incorrectly into syllables. Therefore, we decided to use existing lexical resources in order to perform syllabic division. Thus, for the English language we used Moby Hyphenator[8], which contains 186,097 hyphenated entries. If a word encountered in the corpus is not present in the Moby Hyphenator, we used alternative lexical resources that allow syllabic division[9]. For the Spanish language, we used the dictionary of syllabic division OSLIN-Es[10], which contains 110,527 hyphenated entries. Table 5 presents some statistics on how many words (%) were encountered in the dictionaries for each of the three considered corpora.

Table 5. Statistics about the dictionaries used for syllabic division.

Corpus	N of words in corpus	N of words encountered in the dictionary	% of words encountered in the dictionary
CCAT 10	20,073	13,747	68.49%
The Guardian	25,518	21,680	84.96%

[4] http://www.nltk.org
[5] https://pypi.python.org/pypi/Pyphen
[6] https://pypi.python.org/pypi/PyHyphen
[7] https://pypi.python.org/pypi/hyphenate
[8] http://icon.shef.ac.uk/Moby/mhyph.html
[9] https://www.howmanysyllables.com; https://ahdictionary.com; http://www.dictionary.com
[10] http://es.oslin.org/syllables.php

Spanish corpus	36,995	23,049	62.30%

It is clear that the syllabification techniques with dictionaries do not cover the whole set of words (see Table 5). Thus, we applied a heuristic for extension of the coverage based on the morphological structure of the words. We considered the sets of prefixes and suffixes for English and Spanish languages, and used them for division into syllables. This method of division is justified by the idea that we are interested in typed syllables related to affixes. This heuristic allowed to improve the coverage up to 90%.

Another consideration with respect to the coverage is that the words that are not in the mentioned dictionaries should be relatively rare words or named entities. So we expect that their influence on authorship attribution is minimal.

Probably, simpler heuristic based on phonetical rules would give better results, for example, division into syllables using just positions of vowels or combinations of vowels. We leave this option for future work.

6. Experimental Methodology

In this paper, we apply standard methodology based on application of machine learning methods: (1) we represent our task as classification problem, (2) then we select the features and their values (i.e., we construct Vector Space Model, VSM), (3) further we prepare the data (corpus), which is marked with necessary information (in our case, it is just the author of each text), and (4) finally, we apply traditional machine learning algorithms (like Naive Bayes or Support Vector Machines, which allow for evaluation of results: typically, calculation of accuracy or F1-measure) over the marked corpus using features from the VSM.

One of the evaluation schemes, which we applied in this paper, consists in dividing the data into training and test sets. Then the machine learning algorithm learns from the training set and make decisions for further evaluation over the test set. Note that these two sets should always be different, otherwise overfitting occurs. Other possible approach to evaluation is k-fold cross-validation.

Another important part of research is comparison with other methods from state-of-the-art or baseline methods. As the baseline, usually a very simple (“silly”) method is applied, for example, random selection or selection of the majoritarian class.

One of the most traditional approaches using machine learning in text related tasks is to use words as features in the VSM. This is called bag-of-words model (BoW). It is “bag” in the sense that there are no relations between words, they all are independent. This supposition of independency, which is too strong, is usually overcome in other models, for example, by using n-grams of various types. Note that bag-of-words approach is commonly considered as a strong

method for AA and related tasks [Jarvis et al. 2014; Markov et al. 2017a], though it is relatively simple and computationally inexpensive method.

Now let us describe the application of this general scheme to our task.

AA as a classification problem. As we already mentioned, authorship attribution task from a machine learning perspective is targeted as a multi-class, single-label classification problem, in which the set of class labels (authors' names) is defined beforehand. Our objects are texts, their labels are the authors. Note that there can be many authors, i.e., it is multi-class problem. We learn from the texts with the known authorship (training data), and then we should decide the author of the new texts (test data).

Feature selection. This is central point of the machine learning approaches. [Sapkota et al. 2015] used for the same task case-sensitive typed character n-grams of length 3 and considered only those features that occur at least 5 times in the training corpus. Note that this is the full description of the model that they used. We used the same approach with the same threshold for the frequency, but instead of typed character n-grams we use various types of character n-grams obtained from syllables as described in Section 3.

Data preparation. We use three corpora marked with authors of the texts, one for single-topic and two for cross-topic authorship attribution, as described in Section 4.

Application of machine learning algorithms. As in [Sapkota et al. 2015], we used WEKA's [Hall et al. 2009] implementation of Support Vector Machines (SVM) for classification. SVM classifier has been shown to be effective for the AA task and was the classifier of choice under cross-topic conditions at PAN 2015 [Stamatatos et al. 2015].

Evaluation of experiments. Each model is evaluated in term of accuracy on the test corpus. For publicly available single-topic corpus CCAT 10, the division into test and training data is already established. So, we just performed the evaluation over the existing data. For cross-topic experiments, testing is performed on one topic, while training on the rest of topics. This procedure is repeated for each topic and the results are averaged. We carried out experiments applying three models described in Section 3 to syllables. These models correspond to the categories of features: *all-untyped*, *all-typed*, and *affix+punctuation*.

Comparison. We used the variation of the bag-of-words approach, when punctuation marks are excluded, that is, we considered only the frequency of the words. Next, we conducted experiments using character 3-grams. We applied the algorithm of [Sapkota et al. 2015] to the three corpora described above. Though we implemented the algorithm following as exactly as possible the description, the obtained results on the CCAT 10 and *The Guardian* corpora are slightly different (less than 1%). Correspondingly, we compare the results obtained using syllables with our own implementation of the algorithm of [Sapkota et al. 2015]. We also compare with random baselines (see next section).

7. Experimental Results

The results in terms of classification accuracy using the bag-of-words baseline, three models of character 3-grams and syllables on the CCAT 10, *The Guardian*, and the Spanish corpora are shown in Table 6[11]. Note that all the results are much superior as compared to random baseline (and they are good in this sense. i.e., all methods have reasonable performance). Say, if we have 10 authors (CCAT 10 corpus), the random baseline, i.e., assigning the author by chance, is 10%. In *The Guardian* corpus, we have 13 authors, so the random baseline is 7.7%. In Spanish corpus, there are 6 authors, so the baseline is 16.7%. In Table 6, we show the results using bag-of-words approach and three models (*untyped*, *typed*, and *affix+punctuation*) of character n-grams and syllables. The best accuracy for each dataset is highlighted in bold typeface and underlined, the result of the second best method is highlighted.

Table 6. Accuracy (%) of the results on the CCAT 10, *The Guardian* and Spanish corpora.

Approach	***All-untyped***	***All-typed***	***Affix+punct***
	CCAT 10		
Random baseline	10.0	-	-
Bag-of-words	76.2	-	-
Char. n-grams	78.2	78.0	**78.8**
Syllables (our)	**76.6**	77.0	72.8
	The Guardian		
Random baseline	7.7	-	-
Bag-of-words	46.0	-	-
Char. n-grams	**52.5**	50.0	52.3
Syllables (our)	50.1	45.0	**58.1**
	Spanish corpus		
Random baseline	16.7	-	-
Bag-of-words	**55.1**	-	-
Char. n-grams	56.0	56.3	**56.5**
Syllables (our)	54.2	54.8	52.9

In Table 7 we present the number of features used in each experiment. Number of features is the number of dimensions in the corresponding Vector Space Model: for bag-of-words, it is the number of words used as features; for character n-grams or syllables, it is the number of n-grams or syllables of the corresponding type. It is important that though these numbers are large (from 2,000 to 17,000), they are still tractable, i.e., they are not too large, like hundreds of thousands of features, which would be intractable. Let us remind that for character n-grams we used the threshold for frequency (more than 5 appearances). Note that the information of the

[11] Programming of the method was performed by H. J. Hernández and E. López.

number of features is useful for comparison of machine learning methods, but it is mainly supplementary information. It is worth noting that the size of the feature set is larger when using syllables, except for the case of *affix+punctuation,* when the size is 15-30% less.

Table 7. Number of features used in the CCAT 10, *The Guardian* and Spanish corpora.

Approach	***All-untyped***	***All-typed***	***Affix+punct***
CCAT 10			
Bag-of-words	5,166	-	-
Char. n-grams	9,258	10,859	6,296
Syllables (our)	16,991	18,826	5,497
The Guardian			
Bag-of-words	2,595	-	-
Char. n-grams	5,728	6,903	3,779
Syllables (our)	7,947	2,691	2,201
Spanish corpus			
Bag-of-words	2,087	-	-
Char. n-grams	4,914	5,735	3,005
Syllables (our)	5,802	6,420	1,821

It is well-known that the results in cross-topic scenario are lower than in single-topic conditions, because it is much more difficult to make authorship attribution using cross-topic data. For example, in modern competitions on authorship attribution (like PAN) only cross-topic scenario is considered as being more challenging. The reason of this major difficulty is the radical changes in the vocabulary: for a new topic, the authors do not use the majority of words corresponding to a previous topic. This tendency is also observed in our results: there is about 20% of difference in the results for CCAT 10 (single-topic) and for the other two corpora (cross-topic).

For the English language, our approach outperforms the bag-of-words method. On one corpus (CCAT 10) we obtained lower results than those obtained with character 3-grams (difference 2.2%), while on the other one, *The Guardian*, we won by 5.8%. This result is obtained using *affix+punctuation* features. The other feature types are slightly worse for syllables.

For Spanish, the results obtained using syllables are lower than both the BoW method and the character 3-grams approach, but the difference is very small (0.3% for BoW and 1.8% for character n-grams).

Speaking about character 3-grams, *affix+punctuation* is the best model for this type of features on two of the three corpora, but the difference is less than 1%. Thus, the supposition of [Sapkota et al. 2015] is valid for the Spanish language, but it is not conclusive since the difference is rather small. Thus, we cannot confirm that it will always hold.

It is worth noting that our syllabification techniques do not cover the whole set of words (see Table 5), and though we further applied heuristic based on the the morphological structure, which augmented the coverage to 90% of words, better syllabification methods should be applied in future.

8. Conclusions and Future Work

In this paper, we tackled the authorship attribution (AA) task using different categories of syllables as features. Syllables have not been previously used as features for this task, while they are capable of capturing all the information that is captured by typed character n-grams, which are considered among the best predictive features for this task [Sapkota et al. 2015].

Syllables obtained better results on one of the three corpora (5.8% better) and worse results (difference of 2.2% and 1.8%) on other two corpora.

One of the possible explanations is that a complete dictionary of syllabification is required for the proper performance of this approach, while existing libraries and developed dictionaries still do not cover all the words encountered in the corpora. The proposed heuristic based on the morphological word structure should be reconsidered as well.

One of directions for future work would be to improve our dictionaries for both the English and Spanish languages, as well as to build dictionaries for other languages is not addressed in the current work. We will examine other representations of the proposed features, including doc2vec-bsed feature representation, which is known to provide good results for AA [Posadas-Durán et al. 2016]. Moreover, we will examine the effect of applying latent semantic analysis, which allows significantly reducing the number of dimensions in the vector space model [Sidorov et al. 2016]. We will also conduct experiments under cross-genre and cross-corpus AA conditions, as well as consider languages other than English and Spanish.

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT project 240844, SNI, COFAA-IPN, and SIP-IPN 20161947, 20171813). I thank I. Markov, J.-P. Posadas and G. Posadas for their invaluable help in preparation of this paper. I also thank H. J. Hernández and E. López for the programming of the described method.

References

Abbasi and Chen 2005 – Abbasi A. and Chen H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems.* 2005. Vol. 20. No. 5. Pp. 67–75.

Argamon et al. 2007 – Argamon S., Whitelaw C., Chase P., Hota S.R., Garg N., Levitan S. Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science and Technology*. 2007. Vol. 58. No. 6. Pp. 802–822.

Burrows 1987 – Burrows J. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*. 1987. Vol. 2. No. 2. Pp. 61–70.

Daelemans 2013 – Daelemans W. Explanation in computational stylometry. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*. 2013. Pp. 451–462.

Diederich et al. 2003 – Diederich J., Kindermann J., Leopold E., Paass G. Authorship attribution with support vector machines. Applied Intelligence. 2003. Vol. 19. No. 1–2. Pp. 109–123.

Feng et al. 2010 – Feng L., Jansche M., Huenerfauth M., Elhadad N. A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010. Pp. 276–284.

Fucks 1952 – Fucks W. On the mathematical analysis of style. *Biometrica*. 1952. Vol. 39. No. 1–2. Pp. 122–129.

Gómez-Adorno et al. 2015 – Gómez-Adorno H., Sidorov G., Pinto D., Markov I. A graph based authorship identification approach. *Working Notes Papers of the CLEF 2015 Evaluation Labs*. 2015. Vol. 1391.

Grieve 2005 – Grieve J. *Quantitative authorship attribution: A history and an evaluation of techniques*. MSc dis. Simon Fraser University. 2005.

Hall et al. 2009 – Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The WEKA data mining software: An update. *SIGKDD Explorations.* 2009. Vol. 11. No. 1. Pp. 10–18.

Holmes 1994 – Holmes D. Authorship attribution. *Computers and the Humanities*. 1994. Vol. 28. No. 2. Pp. 87–106.

Jarvis et al. 2014 – Jarvis S., Bestgen Y., Pepper S. Maximizing classification accuracy in native language identification. *Proceeding of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*. 2013. Pp. 111–118.

Juola 2008 – Juola P. Authorship attribution. *Foundations and Trends in Information Retrieval.* 2008. Vol. 1. No. 3. Pp. 233–334.

Kestemont 2014 – Kestemont M. Function words in authorship attribution. From black magic to theory? *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*. 2014. Pp. 59–66.

Koppel and Winter 2014 – Koppel M., Winter Y. Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology*. 2014. Vol. 65. No. 1. Pp. 178–187.

Lewis et al. 2004 – Lewis D.D., Yang Y., Rose T.G., Li F. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*. 2004. Vol.

5. Pp. 361–397.

Luyckx and Daelemans 2008 – Luyckx K., Daelemans W. Authorship attribution and verification with many authors and limited data. *Proceedings of the 22nd International Conference on Computational Linguistics*. 2008. Pp. 513–520.

Markov et al. 2017 – Markov I., Baptista J., Pichardo-Lagunas O. Authorship attribution in Portuguese using character n-grams. *Acta Polytechnica Hungarica*. 2017. Vol. 14. No. 3. Pp. 59–78.

Markov et al. 2017a – Markov I., Gómez-Adorno H., Posadas-Durán J.-P., Sidorov G., Gelbukh A.: Author profiling with doc2vec neural network-based document embeddings. *Proceedings of the 15th Mexican International Conference on Artificial Intelligence*. 2017. Vol. 10062. Pp. 117–131.

Markov et al. 2017b – Markov I., Gómez-Adorno H., Sidorov G. Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. *Working Notes Papers of the CLEF 2017 Evaluation Labs*. 2017. Vol. 1866.

Markov et al. 2017c – Markov I., Stamatatos E., Sidorov G. Improving cross-topic authorship attribution: The role of pre-processing. *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*. 2017.

McNamara et al. 2010 – McNamara D., Louwerse M., McCarthy P., Graesser A. Cohmetrix: Capturing linguistic features of cohesion. *Discourse Processes*. 2010. Vol. 47. No. 4. Pp. 292–330.

Mendenhall 1887 – Mendenhall T. The characteristic curves of composition. *Science*. 1887. Vol. 9. No. 214. Pp. 237–249.

Mosteller and Wallace 1964 – Mosteller F., Wallace, D. L. Inference and Disputed Authorship: The Federalist. *Reading*, MA: Addison-Wesley Publishing Company. 1964. (Reprinted: Stanford: Center for the Study of Language and Information. 2008.).

Pentel 2015 – Pentel A. Effect of different feature types on age based classification of short texts. *Proceedings of the 6th International Conference on Information, Intelligence, Systems and Applications*. 2015. Pp. 1–7.

Posadas-Durán et al. 2016 – Posadas-Durán, J.-P., Gómez-Adorno H., Sidorov G., Batyrshin I., Pinto D., Chanona- Hernandez, L. Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing. 2016. Vol. 21. No. 3. Pp. 627–639.

Qian et al. 2014 – Qian T., Liu B., Chen L., Peng Z. Tritraining for authorship attribution with limited training data. *Proceeding of the 52nd Annual Meeting of the Association for Computational Linguistics*. 2014. Pp. 345–351.

Sapkota et al. 2014 – Sapkota U., Solorio T., Montes-y-Gómez M., Bethard S., Rosso P. Cross-topic authorship attribution: Will out-of-topic data help? *Proceedings of the 25th*

International Conference on Computational Linguistics. 2014. Pp. 1228–1237.

Sapkota et al. 2015 – Sapkota U., Bethard, S., Montes-y-Gómez, M., Solorio, T. Not all character n-grams are created equal: A study in authorship attribution. *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies.* 2015. Pp. 93–102.

Sidorov et al. 2016 – Sidorov G., Ibarra Romero M., Markov I., Guzman-Cabrera R., Chanona-Hernández, L., Velásquez, F. Detección automática de similitud entre programas del lenguaje de programación Karel basada en técnicas de procesamiento de lenguaje natural. *Computación y Sistemas*. 2016. Vol. 20. No. 2. Pp. 279–288.

Stamatatos 2009 – Stamatatos E. A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*. 2009. Vol. 60. No. 3. Pp. 538–556.

Stamatatos 2013 – Stamatatos E. On the robustness of authorship attribution based on character n-gram features. Journal of Law & Policy. 2013. Vol. 21. Pp. 427–439.

Stamatatos et al. 2014 – Stamatatos E., Daelemans W., Verhoeven B., Stein B., Potthast M., Juola P., Sánchez-Pérez M.A., Barrón-Cedeño A. Overview of the author identification task at PAN 2014. *Working Notes of CLEF 2014 - Conference and Labs of the Evaluation forum*. 2014. Pp. 877–897.

Stamatatos et al. 2015 – Stamatatos E., Daelemans W., Verhoeven B., Juola P., López-López A., Potthast M., Stein B. Overview of the author identification task at PAN 2015. *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*. 2015.

Stamatatos et al. 2000 – Stamatatos E., Kokkinakis G., Fakotakis N. Automatic text categorization in terms of genre and author. *Computational Linguistics*. 2000. Vol. 26. No. 4. Pp. 471–495.

Van Halteren 2004 – Van Halteren H. Linguistic profiling for author recognition and verification. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 2004.

Григорий Олегович Сидоров — к. филол. н.; профессор лаборатории естественного языка и обработки текста Центра Компьютерных Исследований Национального Политехнического института (Мексика).

Grigori Sidorov — Ph. D. in Computational Linguistics, full professor; research professor of Natural Language and Text Processing Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Mexico City, Mexico.

Email: sidorov@cic.ipn.mx