An n-gram based approach to the automatic classification of schoolchildren's writing

Jordi Cicres
Universitat de Girona, Spain
jordi.cicres@udg.edu

Sheila Queralt
Laboratorio SQ-Lingüistas Forenses, Barcelona, Spain
sheila.queralt@cllicenciats.cat

Abstract

This article focuses on the analysis of schoolchildren's writing (throughout the whole primary school period) using sets of morphological labels (n-grams). We analyzed the sets of bigrams and trigrams from a group of literary texts written by Catalan schoolchildren in order to identify which bigrams and trigrams can help discriminate between texts from the three cycles into which the Spanish primary education system is divided: lower cycle (6- and 7-year-olds), middle cycle (8- and 9-year-olds) and upper cycle (10- and 11-year-olds). The results obtained are close to 70% of correct classifications (77.5% bigrams and 68.6% trigrams), making this technique useful for automatic document classification by age.

Keywords: writing, n-grams, primary school, morphological categories, automatic classification

Resumen

Este artículo trata del análisis de la escritura de los escolares (a lo largo de la educación primaria) utilizando un conjunto de etiquetas morfológicas (n-gramas). Se han analizado un conjunto de bigramas y trigramas de un conjunto de textos literarios escritos por escolares catalanes con el objetivo de identificar qué bigramas y trigramas pueden discriminar los textos según los ciclos en los que se divide la educación primaria en España: el ciclo inicial (6 y 7 años), medio (8 y 9 años) y superior (10 y 11 años). Los resultados muestran cerca del 70% de clasificaciones correctas (el 77,5% en bigramas y el 68,6% en trigramas), lo que permite afirmar que la técnica es útil para la clasificación automática de los documentos según la edad.

Palabras clave: escritura, n-gramas, educación primaria, categorías morfológicas, clasificación automática

1. Introduction

The development of writing competence in the age corresponding to the primary school period (from 6 to 12 years old) is a key factor for both the expression of ideas (Graham, 2006) and cognitive development (Björk & Blomstrand, 2000). Writing is thus an essential tool for learning (Graham and Herbert, 2011). In view of this, the study of texts produced by schoolchildren is highly relevant and justified. Various approaches have been proposed for analyzing texts produced during the school period, including error analysis (Sofkova Hashemi, 2003); the analysis of the main textual properties —i.e. cohesion, coherence and adequacy (Sotomayor, Lucchini, Bedwell, Biedma, Hernández & Molina, 2013)—; the writing and revision processes (Flower & Hayes, 1981; Fitzgerald and Markham, 1987; Camps, 1990; Graham, 2006); and the analysis of the literary formal aspects of texts.

The present study focuses on the analysis of stylometric aspects. From this perspective, the center of attention of stylometry has concentrated, on the one hand, on analyzing the writing style of specific authors (ranging from the most famous controversies over the authorship of Shakespeare plays (Efron & Thisted, 1976; Lowe and Matthews, 1995; Merriam, 1996) to studies on the Federalist Papers (Mosteller & Wallace, 1964; Holmes & Forsyth, 1995; Tweedie, Singh & Holmes, 1996), or on constructing profiles that can help identify the author's gender, dialectal origin, educational level, etc., on the other.

For instance, in order to determine the author's gender in digital texts (specifically in tweets), it has been discovered that, in English, men use more determiners and prepositions, whereas women use more personal pronouns, auxiliary verbs and conjunctions. It has also been observed that women use more emoticons, ellipses (...), words with multiplied vowels (*nooo waaay*), repeated exclamation marks, combined punctuation marks (especially *?* and *!*), the *omg* abbreviation (from *Oh my God*) and onomatopoeic words (*ah, hmm, ugh, grr*), whereas the only common thing among male authors is the frequent use of *yeah* and *yea* (Bamman, Eisenstein & Schnoebelen, 2012).

As a result, several linguistic and computational approaches have been proposed, whose aim is to define a group of variables that can be used to discriminate the authors of texts according to sociolinguistic variables (gender, ethnicity, age, educational status...) or to identify the style of a specific author. For example, Cheng, Chandramouli

& Subbalakshmi (2011) used up to 545 parameters related to psycholinguistic and linguistic preferences according to gender, together with stylometric parameters, including character-based features (such as the ratio of letters, numbers, uppercase characters, spaces or the ratio of special characters in relation to the total number of characters), word-based features (including measures such as mean word-length, lexical richness, long- and short-word ratio, the ratio of hapax legomena and hapax dislegomena, etc.), syntactic features (with variables related to punctuation marks) and, lastly, structural features (with variables such as the number of sentences and paragraphs, mean number of sentences and of words per paragraph, the ratio of sentences starting with upper or lower case, etc.). This set of variables succeeded in correctly classifying texts according to the author's gender in 85.1% of the cases in extensive corpora (the Enron Corpus and the first volume of the Reuters Corpus).

Other techniques to describe the author's style consist in the analysis of n-grams, which are sets of *n* elements appearing together. In different areas of linguistics, in particular in studies on information theory and psycholinguistics (Jurafsky, 2003), learning theories (Anderson, 1982; Newell, 1990) and more recently computational linguistics, these categories have been employed not only for descriptive purposes, but also as classifiers (to classify genres or authors).

Although n-grams are usually based on lexical categories, in this article we concentrate on the use of sets of morphological labels, since several studies have shown their efficiency in describing the style of specific authors or literary genres.

Different studies have focused on extracting the syntactic information of texts. Baayen, van Halteren & Tweedie (1996) were the first to implement n-grams based on syntactic information on authorship attribution analysis using an annotated English corpus. Their study proved that authorship attribution analysis based on syntactic n-grams was more successful than the one based on lexical measures such as vocabulary richness. Stamatatos, Fakotakis & Kokkinakis (2000) later implemented sentence and chunk boundaries in order to discriminate between authors of Modern Greek texts. Their approach therefore used simpler information than the one by Baayen et al. (1996). Hirst and Feiguina (2007) used bigrams of syntactic labels and were able to obtain optimal results in authorship attribution with very short texts (about 200 words long). Other researchers, like Nazar & Sánchez Pol (2007), Spassova & Turell (2007) and Queralt & Turell (2013), obtained very successful results in authorship attribution through the application of a Part-of-Speech (POS) tagger to Spanish texts using bigrams and trigrams.

The use of n-grams has yielded very successful results in the determination of the authorship of written texts within the field of forensic linguistics, since this method focuses on syntactic structure. Although syntactic variables are more complex and therefore present more obstacles for automatic analysis (compared to more superficial variables like sentence length or the use of punctuation marks, for instance), it is also more difficult for writers to modify them at their will. For this reason, they represent the concept of idiolectal style better than other variables (Turell, 2010; Queralt & Turell, 2013).

Nevertheless, these techniques have not yet been applied to the automatic analysis and classification of school texts. In this paper, the n-gram technique is used to analyse texts produced by primary school students (ranging between 7 and 12 years old) and to classify them according to their authors' ages. It is worth noting that children make hugely significant progress in their acquisition of reading and writing skills during the primary school years. In Spain, most schools initiate the teaching and learning of these skills when students are between 3 and 4 years old (Teberosky, 2001), so that most children have reached the alphabetic phase by the time primary education begins (at age 6). From then on, more intensive writing practices are introduced and the teaching of orthographic rules is initiated.

2. Objectives and hypotheses

The present article deals with the analysis of schoolchildren's writing (throughout the whole primary school period) using sets of morphological labels (n-grams). Our goal is to identify which bigrams and trigrams can help discriminate between texts written by children in each of the 3 cycles into which the Spanish primary education system is divided: lower cycle (6- and 7-year-olds), middle cycle (8- and 9-year-olds) and upper cycle (10- and 11-year-olds). An additional aim is to establish a means of automatically classifying new texts as belonging to one of the 3 cycles of primary school.

Therefore, the following hypotheses are considered:

1. It will be possible to find a combination of bi- and trigrams that characterize the writing style of each of the three cycles of primary education.

2. The combination of bi- and trigrams will allow us to correctly classify new texts, i.e., assign them to their corresponding age group.

3. Methodology

3.1. Corpus

The texts are written in Catalan by children attending school in the town of Balaguer (Catalonia). The native languages of all the participants are Catalan and Spanish.

The corpus used in this study comprises 169 fragments of literary texts in Catalan (a specific version of *Little Red Riding Hood)* written by 7- to 11-year-old children as an activity in their regular classrooms. We have not considered 6- and 7-year-old children because their command of the written language is still insufficient to write a long text, as required by the proposed exercise.

The children did not receive specific instructions on how to perform the writing task other than that they had to explain the *Little Red Riding Hood* story "in their own way". Table 1 shows the distribution of the corpus by the total number of samples in each class. Classes are grouped into cycles and the number of samples is divided by gender. Other measures shown are the mean number of words per text and the standard deviation.

Table 1. Dist,ribution of the corpus.

Cycle	Age	N	Boys	Girls	Average Length of Words	SD
Lower	7-8	42	24	18	157.90	60.779
Middle	8-9	39	12	27	202.69	74.836
	9-10	30	10	20	214.80	73.091
Upper	10-11	42	23	19	194.02	56.759
	11-12	16	7	9	289.50	124.125
Total		169	76	93	199.78	80.671

3.2. Analysis and labeling of the morphological categories

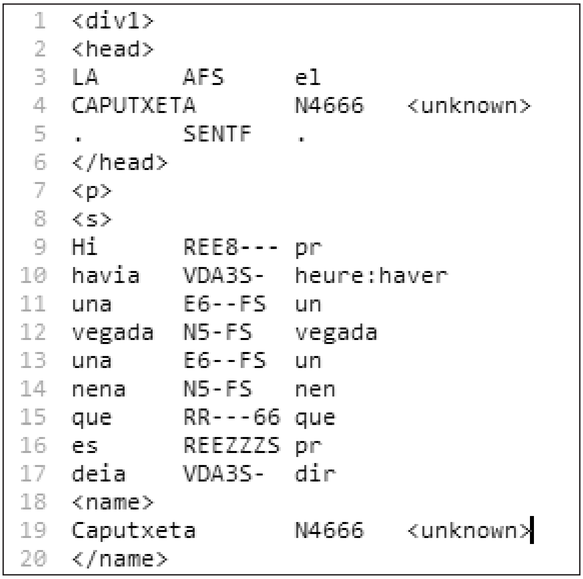
The text analysis process followed 4 steps. First, the texts were pre-processed. This step includes the digitalization of the texts (all the texts were originally written by hand). During the second step the researchers corrected spelling mistakes without altering the syntactic structures in the texts (except for specific cases in which the syntactic mistakes would have prevented correct morphological analysis and labeling).

Next, the morphological labeling process was performed, which is initially an automatic process. In this case, however, we used HectorWWW (available on http://eines.iula.upf.edu/cgi-bin/hectorwww/hectormain.pl), a morphological analyzer and disambiguator developed by the University Institute of Applied Linguistics (IULA) at the Pompeu Fabra University in Barcelona. This tool works with three languages: Catalan, Spanish and English. The output provided is a text file with a list of all entries and their corresponding morphosyntactic label, with a standard format (Morel, Torner, Vivaldi, De Yzaguirre & Cabré, 1998). After automatic labeling, a manual revision was conducted in order to correct any errors in the labels. An extract of the linguistic characterization of the labels used by HectorWWW is shown in Table 2. An example of the output is illustrated in Figure 1 in which the Catalan sentence *La Caputxeta. Hi havia una vegada una nena que es deia Caputxeta /Once upon a time there was a girl called Little Red Riding Hood/* is tagged by HectorWWW.

Table 2. Linguistic characterization of HectorWWW's tag set.

Hectorwww's tag set	**Key**
AFP	Article-Feminine-Plural
EP12MS	Specifier-Possessive-1possessor-2nd Person-Masculine – Singular
JQ-FP	Adjective-Description- Feminine-Plural
N5-MS	Noun- Common- Masculine-Singular
REO-2MP	Pronoun-Personal-Strong-2nd Person-Masculine-Plural
VDP2S-	Verb Indicative Perfect 2nd Person-Singular

Figure 1. Output of the sentence *La Caputxeta. Hi havia una vegada una nena que es deia Caputxeta* tagged by HectorWWW.



```
<div1>
<head>
LA      AFS     el
CAPUTXETA       N4666   <unknown>
.       SENTF   .
</head>
<p>
<s>
Hi      REE8--- pr
havia   VDA3S-  heure:haver
una     E6--FS  un
vegada  N5-FS   vegada
una     E6--FS  un
nena    N5-FS   nen
que     RR---66 que
es      REEZZZS pr
deia    VDA3S-  dir
<name>
Caputxeta       N4666   <unknown>
</name>
```

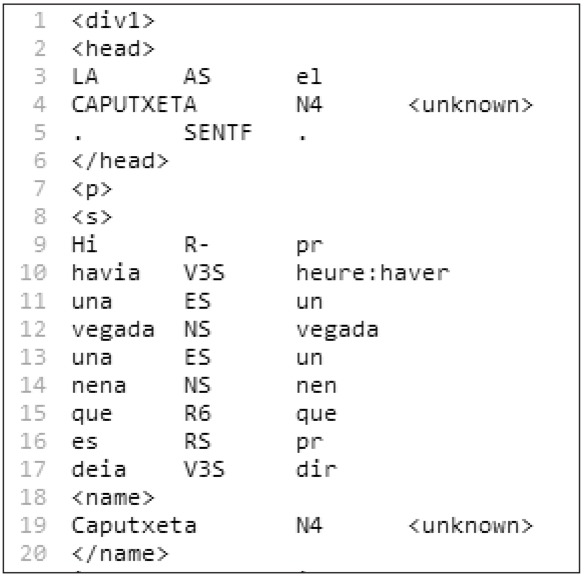
Continuing with the third step, the labels were simplified following the model suggested by Bel, Queralt, Spassova and Turell (2012). In the case of conjugated verbs, only the number and the person were kept. For impersonal verb forms (V), only the type of form (infinitive or gerund) is kept, except for participles, where the number is also retained. As for nouns, they are classified into proper (N4) and common (N) nouns, with the latter also including information on their number (singular or plural). With respect to the other categories, such as articles (A), adjectives (J) and pronouns (R), only the numerical information is maintained. Categories which do not require any additional information are adverbs (D), conjunctions (C) and punctuation (DLD). Table 3 below shows the simplified tag set and its key meaning, while Figure 2 shows the previous example sentence with the new tag set.

Table 3. Linguistic characterization of tag set number 13. Source: Bel, N., S. Queralt, M. S. Spassova, and M. T. Turell. (2012: 196).

Hectorwww's tag set	Tag set No. 13	Key
AFP	AP →	Article-Plural
EP12MS	ES →	Specifier-Singular
JQ-FP	JP →	Adjective-Plural

N5-MS	NS →	Noun-Singular
REO-2MP	RP →	Pronoun-Plural
VDP2S-	V2S →	Verb 2nd Person-Singular

Figure 2. Output of the sentence *La Caputxeta. Hi havia una vegada una nena que es deia Caputxeta* with Tag set No. 13.

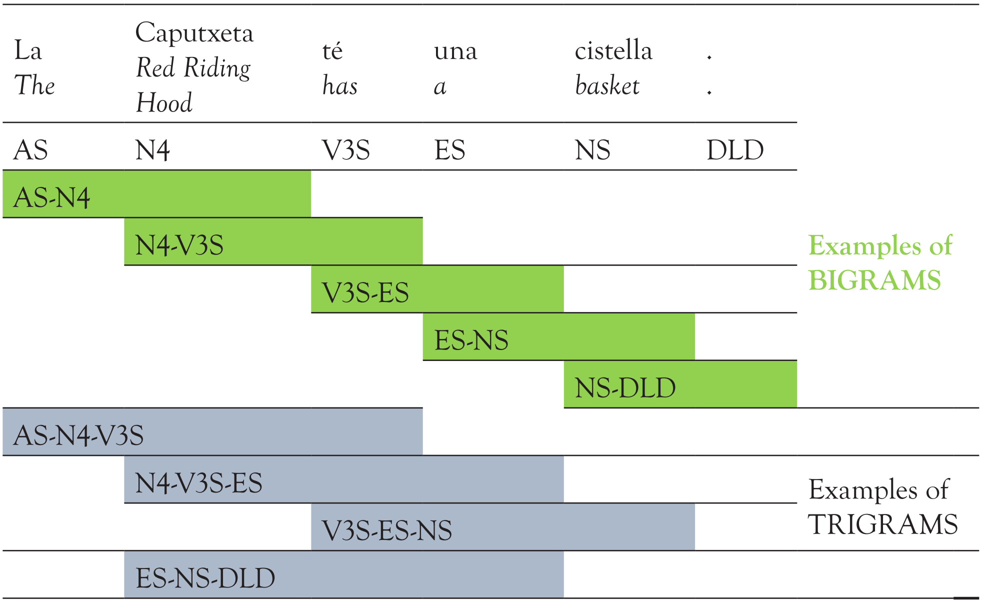


```
<div1>
<head>
LA        AS        el
CAPUTXETA           N4        <unknown>
.         SENTF     .
</head>
<p>
<s>
Hi        R-        pr
havia     V3S       heure:haver
una       ES        un
vegada    NS        vegada
una       ES        un
nena      NS        nen
que       R6        que
es        RS        pr
deia      V3S       dir
<name>
Caputxeta           N4        <unknown>
</name>
```

Finally, the last step consists in extracting n-grams (bigrams and trigrams) using the ForensicLab's private tool, known as Legolas software, specifically developed in the University Institute of Applied Linguistics (IULA) to work with output files produced by HectorWWW. At the end of this process, the n-gram results are obtained in a format which is adequate for their statistical treatment.

It must be stressed that only sets of two and three morphological categories (bi- and trigrams) have been examined since most researchers that have used n-grams have concluded that bigrams and trigrams are the combinations offering the highest performance (e.g. Baayen et al., 1996; Stamatatos, Fakotakis and Kokkinakis, 2000; Hirst and Feiguina, 2007; Nazar & Sánchez Pol, 2007; Spassova, 2007; Spassova & Turell, 2007; Grant, 2007; or Bel et al., 2012). Bel et al. (2012) also recommend the restriction of the number of bigrams and trigrams to the 40 most used (out of the hundreds possible) to facilitate analysis. An example of bigrams and trigrams is shown in Figure 3 below.

Figure 3. Examples of bigram and trigram.



3.3. Statistical analysis

Given the nature of the data and the goals pursued, two related statistical techniques were used. On the one hand, an ANOVA test was carried out to determine which variables showed significant-enough differences to classify the texts into the three age groups analysed. In addition, the post-hoc Dunnett's T3 test was applied to the data to reveal which groups differed from which. The Dunnett's T3 test was chosen due to the nature of the sample: unbalanced groups (containing a different number of individuals in each group) and unequal variances (the Levene test showed that the variances are not similar in all the groups).

On the other hand, a linear discriminant analysis (LDA) was carried out. This multivariable statistical technique has a twofold goal. Firstly, it identifies the features which can be used to differentiate two or more groups of cases and it constructs discriminant functions based on them. Secondly, it can classify new cases as belonging to one group or another (Pardo and Ruiz, 2002: 499).

The LDA technique consists in determining the characteristics that differentiate the distinct groups and, from those, finding the optimal plane where the projection of the observations best separates the groups. Subsequently, this optimal plane allows us to classify new cases, i.e., to assign a new observation to one of the existing groups based on the values taken by their original variables.

In this study, conducted through SPSS software (19th version), we used the stepwise inclusion method and Lambda de Wilks, with the criteria of values being F: 3.84 as input and 2.71 as output (the standard values in SPSS). In order to check the discriminative power, we used the cross-validation method (leaving one out). With this method, one observation of the analysis is removed, and the discriminant functions are generated. After that, this observation is classified into one of the groups. Since the group to which the excluded observation of the analysis belongs is known beforehand, it is possible to check if the subsequent classification is correct. This process repeats itself for each observation.

4. Results

4.1. Bigrams

Table 4 below shows the combination of 2 grammatical categories which present significant differences between one or more than one group. In total, 25 bigrams were selected, 16 of which present differences between the lower and the middle cycles and 17 between the lower and upper cycles. 7 bigrams show differences between the middle and upper cycle. Only 2 of the variables (DLD-D and DLD_V3S) distinguish between the 3 cycles.

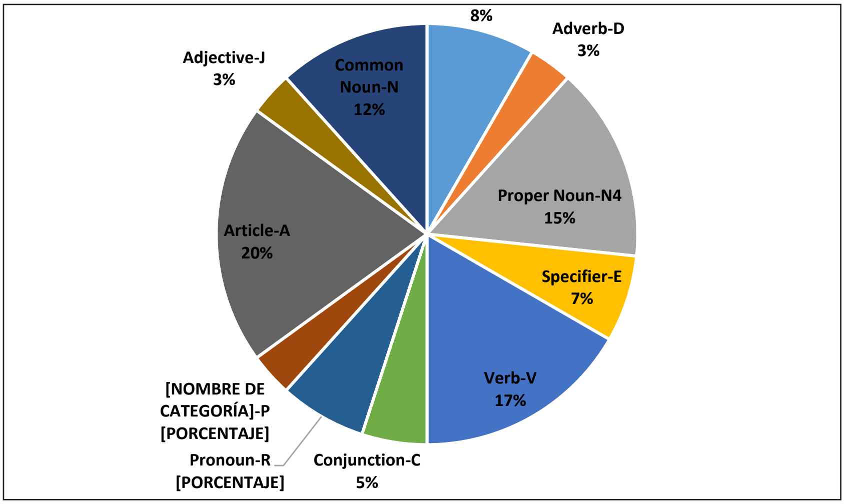
When observing the mean frequency of each bigram for each group, ascending and descending tendencies can be established, as well as cases in which the middle cycle presents small fluctuations. The bigrams which are most frequent in the lower cycle tend to be the most basic and common syntactic structures, while trigrams with ascending tendencies in their use are related to the use of punctuation (DLD) in 72% of the cases.

Table 4. Bigrams that present significant differences between groups.

Bigram	Differences lower-middle cycle	Differences lower-upper cycle	Differences middle-upper cycle
AS_N4	√	√	
P_AS	√		√
AS_NS			√
NS_P		√	
NS_C	√	√	
C_V3S	√		
NS_DLD	√	√	
C_AS		√	
P_VI			√
VI_AS	√	√	
N4_DLD	√	√	
C_RS		√	
DLD_AS	√	√	
DLD_D	√	√	√
N4_N4	√	√	
AS_ES	√	√	
DLD_C	√	√	
D_JS		√	
V3S_ES			√
VI_DLD	√		√
N4_RS	√		
NS_D		√	
D_DLD	√	√	
DLD_V3S	√	√	√
Total	**16**	**17**	**7**

From this table, we can also learn which categories are most significant when establishing differences between the cycles. As has already been explained, the most discriminant category is punctuation, followed by the use of articles, as well as proper nouns, verbs and conjunctions. The categories which show the smallest differences are adjectives, specifiers and pronouns. These results can be visualized in Figure 4:

Figure 4. Categories within the most significant bigrams for the detection of differences between cycles.



As concerns the results of the linear discriminant analysis, this was used to conduct a multivariate analysis of variance to test the hypothesis that the lower, middle and upper cycles would differ significantly on a linear combination of bigram variables. The overall Chi-square test turned out to be significant (Wilks □ = .290, Chi-square = 53.625, df = 8, Canonical correlation = .772, p <. 001). Regarding the variance explained in bigrams, 79% was explained by the first function and 21% by the second.

The stepwise discriminant analysis selected 9 of the 40 variables (Table 5). These 9 bigrams are variables which discriminate between the groups. As in the case of the bigrams, this satisfies the minimum sample size criterion (N=169) of having 10 cases per variable and the requirement that the number of cases in each group be equal to or exceed the number of variables recommended by Brown and Tinsley (1983), as well as Huberty's (1975) criterion to include at least 3 cases for every variable in each group. As shown, these results match the ones obtained for the ANOVA tests, since the bigrams are made up of the categories which presented a higher degree of significance in the ANOVA tests. The bigrams formed by punctuation marks (DLD), articles (A) and verbs (V) stand out.

Table 5. Variables included in the analysis.

P_AS	Preposition - Article Singular
NS_DLD	Noun Singular - Punctuation mark
P_VI	Preposition - Verb Infinitive
VI_AS	Verb Infinitive - Article Singular
N4_DLD	Proper Noun – Punctuation mark
DLD_D	Punctuation mark – Adverb
AS_ES	Article Singular – Specifier Singular
DLD_C	Punctuation mark – Conjunction
DLD_V3S	Punctuation mark – Verb Third Person Singular

Next, Table 6 presents the standardized discriminant function coefficients, and Table 7 shows the two functions at the group centroids.

Table 6. Standardized Canonical Discriminant Function Coefficient.

	Function	
	1	**2**
P_AS	-.230	.362
NS_DLD	.450	-.124
P_VI	.221	-.286
VI_AS	-.328	-.233
N4_DLD	.467	-.198
DLD_D	.094	.764
AS_ES	.546	.397
DLD_C	.322	-.047
DLD_V3S	.314	-.343

Table 7. Functions of Group Centroids.

Course	**Function**	
	1	2
lower cycle	-2.069	-.165
middle cycle	.879	-.594
upper cycle	.453	.826

Unstandardized canonical discriminant functions evaluated at group means

The classification results are shown in Table 8. Classification of cases based on the canonical variables was highly successful: 77.5% of the cases were correctly reclassified into their original categories. Cross-validation results were also successful, with 74% of the cases correctly classified. Based on the results, it can be observed that students in the lower and middle cycle are correctly classified in a very high percentage of cases (90.5% and 81.2%, respectively), and that those proving to be the most difficult to classify are students in the upper cycle, with a low 53.4% success rate in the classification. Therefore, the groups which are confused most frequently are the middle and upper cycles, while students in the lower cycle are clearly distinguished.

Table 8. Classification results.

		Course	Predicted Group Membership			Total
			lower cycle	middle cycle	upper cycle	
Original	Count	lower cycle	39	3	0	42
		middle cycle	1	58	10	69
		upper cycle	7	17	34	58
	%	lower cycle	92.9	7.1	.0	100.0
		middle cycle	1.4	84.1	14.5	100.0
		upper cycle	12.1	29.3	58.6	100.0
Cross-validated [b]	Count	lower cycle	38	3	1	42
		middle cycle	1	56	12	69
		upper cycle	8	19	31	58
	%	lower cycle	90.5	7.1	2.4	100.0
		middle cycle	1.4	81.2	17.4	100.0
		upper cycle	13.8	32.8	53.4	100.0

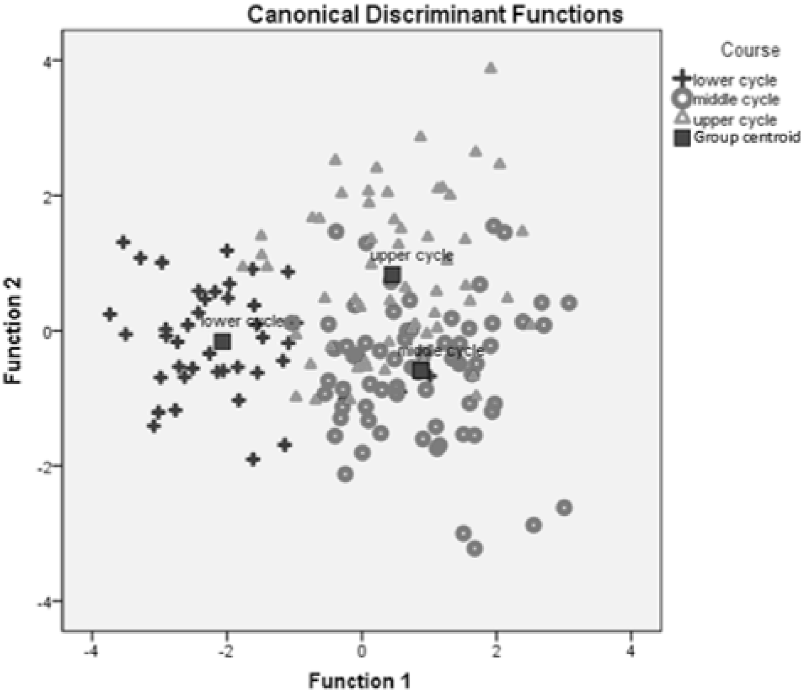
a. 77.5% of original grouped cases correctly classified.

b. Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.

c. 74.0% of cross-validated grouped cases correctly classified.

Next, the results of the original classification are presented graphically. In Figure 5, each of the lower cycle samples are represented by a cross, the middle cycle samples are displayed as a circle, and the upper cycle ones as a triangle. It can be clearly observed that the centroid for the lower cycle is located far from the rest of the centroids and that there is only one case in which the sample finds itself nearer to another centroid. As regards the centroids for the upper and middle cycles, despite being closer to each other, both cycles can also be graphically distinguished, although there are more overlapping cases. Function 1 explained most of the differences (namely 79%), hence the importance of very dissimilar values in that function, as in the case of the lower cycle, which is located in negative values whereas the higher cycles show positive values which are close to each other. Nevertheless, Function 2 allows us to differentiate more clearly between the middle and the upper cycles, locating them in negative and positive values, respectively.

Figure 5. Canonical Discriminant Functions.



4.2. Trigrams

Table 9 contains the trigrams which present differences between the groups and shows between which groups the differences are found. There are a total of 21 sequences of 3 grammatical categories which show differences between the groups. Specifically, 13 of the variables present differences between the lower and the middle cycles, 14 variables show differences between the lower and upper cycles, and only 2 variables distinguish between the middle and upper cycle. Therefore, the most easily distinguished group is the lower cycle, both from the middle and upper cycles. The combination of middle and upper cycle presents very few differences. None of the variables show differences between the 3 cycles.

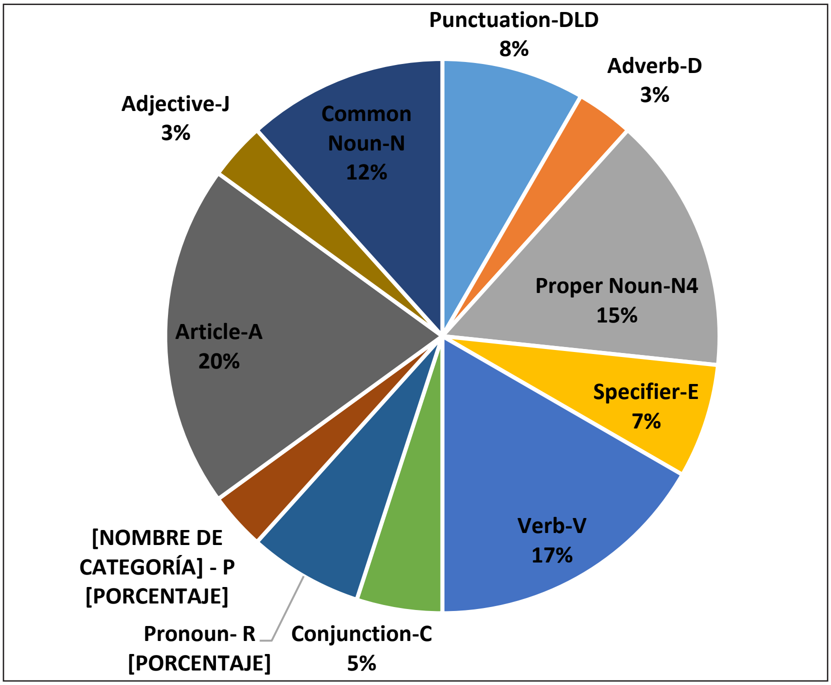
In the case of the mean frequency of each of the trigrams, in 62% of cases the students in the lower cycle repeat concrete trigrams which do not contain a punctuation label with a higher frequency and that their use diminishes as we ascend in the cycles: in other words, the higher the cycle, the higher the cycle, the less frequently are they used. This could be attributable to the fact that students in the lower cycle possess less syntactic richness and therefore tend to repeat a given structure more often. Another interesting fact is that in 50% of the trigrams whose frequency is inverse (that is, the higher the level, the higher the frequency of use), we notice that the trigrams are related to the punctuation label (DLD).

Table 9. Trigrams that present significant differences between groups.

Trigram	Differences lower-middle cycle	Differences lower-upper cycle	Differences middle-upper cycle
VI_P_AS	✓		
C_AS_N4		✓	
V3S_VI_AS		✓	
AS_ES_NS	✓	✓	
AS_N4_C	✓	✓	
VI_AS_N4	✓	✓	
DLD_AS_N4	✓	✓	
AS_N4_DLD	✓	✓	
V3S_ES_NS			✓
AS_NS_C	✓		
P_VI_RS			✓
AS_N4_RS	✓		
N4_RS_V3S	✓		
V3S_VI_DLD	✓	✓	
AS_N4_N4	✓	✓	
ES_NS_DLD	✓	✓	
AS_NS_DLD	✓	✓	
NS_D_JS		✓	
R_V3S_ES		✓	
P_AS_ES		✓	
NP_D_JP		✓	
Total	13	15	2

As regards the categories which form the trigrams with significant differences between the groups, it can be observed that they follow a pattern similar to that of bigrams, since the categories presenting more differences are verbs, articles and punctuation.

Figure 6. Categories within the most significant trigrams for the detection of differences between cycles.



Discriminant analysis was used to conduct a multivariate analysis of variance test of the hypothesis that the cycles would differ significantly on a linear combination of trigram variables. The overall Chi-square test proved to be significant (Wilks = .375, Chi-square = 29.354, df = 8, Canonical correlation = .742, p <. 001). Of the variance explained in trigrams, 86% was explained by the first function and 14% by the second.

The stepwise discriminant analysis discarded 9 of the 40 variables (Table 10). These 9 trigrams are discriminant variables between the groups. As in the case of bigrams, the results satisfy the minimum sample size criterion, the requirement that the number of cases in each group be equal to or exceed the number of variables, and also the criterion of at least 3 cases for every variable in each group. Again, the most frequent categories are verbs, punctuation and articles, although proper nouns and specifiers also prove to be interesting in the case of the most discriminant trigrams.

Table 10. Variables included in the analysis.

Trigrams	Key
ES_NS_DLD	Specifier Singular – Noun Singular – Punctuation mark
AS_NS_ DLD	Article Singular – Noun Singular – Punctuation mark
V3S_VI_ DLD	Verb Third Person Singular – Verb Infinitive – Punctuation mark
AS_N4_DLD	Article Singular – Proper Noun – Punctuation mark
AS_ES_NS	Article Singular – Specifier Singular – Noun Singular
VI_AS_N4	Verb Infinitive – Article Singular – Proper Noun
R_V3S_ES	Pronoun – Verb Third Person Singular – Specifier Singular
V3S_ES_NS	Verb Third Person Singular – Specifier Singular – Noun Singular
V3S_VI_C	Verb Third Person Singular - Verb Infinitive - Conjunction

Table 11 presents the standardized discriminant function coefficients. Table 12 shows the two functions at the group centroids.

Table 11. Standardized Canonical Discriminant Function Coefficient.

	Function	
	1	*2*
AS_ES_NS	*0.348*	*-0.428*
V3S_VI_C	*0.265*	*0.364*
VI_AS_N4	*-0.358*	*0.471*
AS_N4_DLD	*0.494*	*0.099*
V3S_ES_NS	*0.272*	*1.042*
V3S_VI_DLD	*0.466*	*0.271*
ES_NS_DLD	*0.349*	*-0.225*
AS_NS_DLD	*0.332*	*0.294*
R_V3S_ES	*-0.584*	*-0.667*

Table 12. Functions of Group Centroids.

Functions of Group Centroids		
Course	Function	
	1	2
lower cycle	-1.903	0.043
middle cycle	0.695	0.452
upper cycle	0.552	-0.569
Unstandardized canonical discriminant functions evaluated at group means		

The classification results are shown in Table 13. Classification of cases based on the canonical variables was successful in 68.6% of the cases, which were correctly reclassified into their original categories. Cross-validation results are also successful in 65.1% of the cases. From the results, it can be observed that the students in the lower cycle are correctly classified in a very high percentage of cases (88.1%). The classification presents issues with the middle and upper cycles, which are only correctly classified in 55.1% and 60.3% of the cases, respectively. Again, these two groups are confused with each other, while the samples by students in the lower cycle are clearly differentiated.

Table 13. Classification results.

		Course	Predicted Group Membership			Total
			lower cycle	middle cycle	upper cycle	
Original	Count	lower cycle	39	1	2	42
		middle cycle	6	39	24	69
		upper cycle	7	13	38	58
	%	lower cycle	92.9	2.4	4.8	100
		middle cycle	8.7	56.5	34.8	100
		upper cycle	12.1	22.4	65.5	100
Cross-validated [b]	Count	lower cycle	37	2	3	42
		middle cycle	6	38	25	69
		upper cycle	7	16	35	58
	%	lower cycle	88.1	4.8	7.1	100
		middle cycle	8.7	55.1	36.2	100
		upper cycle	12.1	27.6	60.3	100

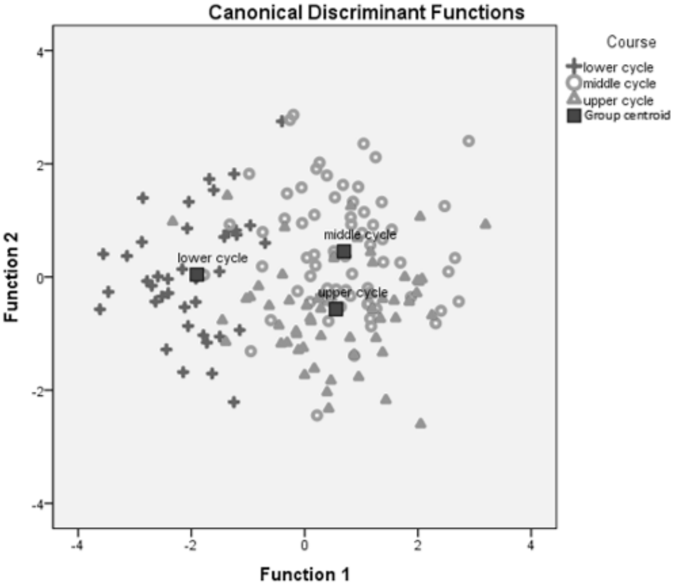
a. 68.6 % of original grouped cases correctly classified.

b. Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.

c. 65.1% of cross-validated grouped cases correctly classified.

Next, the results of the original classification are displayed graphically. In Figure 7, as in the previous illustration, each of the samples of the lower cycle is shown as a cross, those of the middle cycle as a circle and the upper cycle ones as triangles. This graph shows that the centroid for the lower cycle is clearly distanced from the other cycles, while the middle and upper cycle centroids are located closer to one another. Therefore, the overlap between the samples of these two groups is also larger. It should be noted that Function 1 explained 86% of the variance, which means that the most notable differences must be found in the x-axis. Thus, we can see that the centroid for the lower cycle is located around the value -2, that is, very far from the other cycles. Function 2, with a low 14% of variance, presents weaker differences between the groups. However, it again allows us to distinguish between the higher cycles, since the middle cycle centroid is found in positive values whereas the upper cycle shows negative values.

Figure 7. Canonical Discriminant Functions.



5. Discussion

The analytical methods used in this study have shown once again that discriminant analysis is a very useful tool to analyse and describe the differences between groups of samples, as well as to classify new samples based on the differences and similarities presented by the use of their variables.

From a global perspective, it can be observed that bigrams which include a punctuation mark (DLD) present significant differences between the cycles on 18 occasions, followed by 12 bigrams which include a singular article (AS). The rest of the bigrams with other morphological categories present significant differences less frequently: ranging from a single instance of the singular adjective (JS) category, to seven instances of proper nouns (N4), adverbs (D) and conjunctions (C).

The discriminant analysis performed does not include all of the bigrams, but only those which can be used to better discriminate between the groups. The results of the discriminant analysis enable the classification of texts belonging to the three school cycles according to the frequency of bigram use. Thus, we can see that texts corresponding to the first cycle are correctly classified in 92.9% of instances (mainly due to the first discriminant function, as can be seen in Tables 6 and 7, since it clearly distinguishes this cycle from the middle and upper cycles). As to the frequency of bigrams, the main features include the low use of the combinations P_

AS (preposition-singular article), VI_AS (infinitive verb-singular article) and —with a lower frequency— DLD_D (punctuation mark-adverb). At the same time, the AS_ES (singular article- singular specifier), NS_DLD (singular noun-punctuation mark), N4_ DLD (proper noun-punctuation mark), DLD_C (punctuation mark-conjunction) and DLD_V3S (punctuation mark-third person singular verb) bigrams are found with a highest ratio in the middle cycle (although they are also present in the upper cycle), but with little presence in the first cycle. Lastly, for the texts of middle and upper cycles, misclassifications represent around 30% of instances, where a text written by a student in the middle cycle was falsely attributed to the upper cycle. This process of misclassification occurred in the opposite direction in 14.5% of cases. The main variables which can be used to distinguish between the middle and upper cycles include DLD_D (punctuation mark-adverb), AS_ES (singular article-singular specifier) and P_AS (preposition-singular article), which characterize the texts in the upper cycle (and which are much less frequent in those by students of the middle cycle). These variables are related to a greater complexity in the texts. Thus, the students of the initial cycle tend to write shorter sentences, so that more bigrams appear in which a noun and a punctuation mark (mainly full stops) are combined, while the students of the middle and upper cycles make longer and more complex phrases. On the other hand, the bigrams that characterize the middle and upper cycles best are, on the one hand, the combination of a punctuation mark and an adverb, and on the other hand, a preposition and an article. It is shown, thus, that older children are more likely to use adjuncts (introduced by the adverb) at the beginning of the phrases, as well as introduce more complements of the name or adjuncts (introduced by the combination preposition and article).

These data suggest that students experience an important qualitative leap in their essay writing when they move from the first to the middle cycle. Furthermore, the change in the use of punctuation marks is significant (Hall, 1999; Sing & Hall, 2009).

As for trigrams, the morphological category found in those which present significant differences between the groups is the singular article (AS), with a total of 20 instances, followed by the proper noun (N4), with 13, and punctuation marks (DLD), with 10. Again, the first cycle is the easiest to distinguish (the discriminant analysis was successful in 92.9% of the cases). The texts in this group are characterized by the low use of the trigrams VI_AS_N4 and R_V3S_ES. Regarding the first of these two trigrams (VI_AS_N4), first-cycle students use it very repetitively (in most cases the AS_N4 is the object of the verb). Older children choose to introduce in their texts more clitic pronouns, so that the structures are more varied and richer. It is also interesting to emphasize the use of pronouns (R) in the second trigram: even though Catalan is a pro-drop language, younger children do not have enough variety of linguistic resources to mark the subject, so they use significantly more personal pronouns with this syntactic function than the more competent writers.

The texts in the middle cycle are correctly classified in 56.5% of cases (most of the misclassifications occur with the upper cycle). The most frequent trigrams in this cycle are V3S_ES_NS, VI_AS_N4 and V3S_VI_C, as well as those containing punctuation marks. From this group of trigrams, the last one stands out because it includes a conjunction (C), which indicates that students at that age already use subordinate sentences frequently. This trigram is also frequent in the students of the upper cycle. Lastly, the texts by students in the upper cycle are correctly classified in 65.5% of cases and they are characterized by a lower frequency of the trigrams AS_ES_NS, R_V3S_ES and ES_NS_DLD. In this group, the low frequency of trigrams that include pronouns (R) stands out —which correspond mainly to personal pronouns with a subject function in the texts—, since children in the upper cycle are already able to adopt other strategies to mark the subject or choose to omit it. In contrast, trigrams including punctuation are frequent, as they are increasingly more competent with the use of punctuation, especially commas. Once more, we can see that the trigram analysis reinforces the divide between the first and the other two cycles.

6. Conclusions

The results have shown that the analysis of bigrams and trigrams of morphological labels is useful for classifying texts according to the age of the children. The overall percentage of correct classifications is around 70% (77.5% in bigrams and 68.6% in trigrams). With regard to bigrams, it has been observed that those that include a punctuation mark are relevant for discriminating between groups. The differences between age groups with regard to some bigrams that include prepositions (and which usually include complements) and adverbs (which work as adjuncts) are also significant. Regarding the trigrams, once again those that include punctuation marks allow to discriminate between age groups. Also relevant are those that include conjunctions (which introduce mostly subordinate clauses) and personal pronouns (which mostly serve as the subject, and which the older children use in a greater proportion, since they do not have enough syntactic resources to avoid repetition of explicit subjects).

The majority of the bigrams and trigrams studied allow discrimination between the initial cycles (6-7 years), the middle cycles (8-9) and upper cycles (10-11). On the other hand, there are few differences between the middle and upper cycles. Thus, globally, the study can confirm the idea that when students turn 9 they experience a significant change in their writing competence, since they begin to use more complex syntactic structures, they use the punctuation marks more efficiently and show more ability to avoid repetitions (mainly because they introduce the use of clitic pronouns).

References

Anderson, J. R. (1982) Acquisition of cognitive skill. *Psychological Review* 89: 369-406.

Baayen, R.H., van Halteren, H. & Tweedie, F.J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3): 121-131.

Bamman, D., Eisenstein, J. & Schnoebelen, T. (2012). *Gender in Twitter: Styles, Stances, and Social Networks*. Unpublished manuscript.

Bel, N., S. Queralt, Spassova, M.S. & Turell, M.T. (2012) The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tombilin, N. MacLeod, R. Sousa-Silva, & M. Coulthard (eds.) *Proceedings of the International Association of Forensic Linguists' Tenth Biennal Conference*. Birmingham: Center of forensic linguistics.

Björk, L. & Blomstrand, I. (2000) *La escritura en la enseñanza secundaria. Los procesos del pensar y del escribir*. Barcelona: Graó.

Brown, M. T. & Tinsley. H.E.A. (1983) Discriminant analysis. *Journal of Leisure Research* 15(4): 290-310.

Camps, A. (1990) Modelos del proceso de redacción: algunas implicaciones para la enseñanza. *Infancia y Aprendizaje* 49: 3-19.

Cheng, N., Chandramouli, R. & Subbalakshmi K.P. (2011) Author Gender Identification from Text. *Digital Investigation* 8: 78-88.

Efron, R. & Thisted, B. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. *Biometrika* 63(3): 435-447.

Enron email dataset (2005, April). [Online]. Available: http://www-2.cs.cmu.edu/~enron/.

Fitzgerald, J. & Markham, L. (1987) Teaching children about revision in writing. *Cognition and Instruction* 4(1): 3-24.

Flower, L. & Hayes, J. (1981) Plans that guide the composing process. In C. Fredericksen and J. Dominic (eds.) *The nature, development and teaching of written communication, 2, Writing: process, development and communication*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (eds.) *Handbook of educational psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Graham, S. & Hebert, M. (2011) Writing to Read: A Meta-Analysis of the Impact of Writing and Writing Instruction on Reading. *Harvard Educational Review* 81(4): 710-744.

Grant, T. (2007) Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law* 14(1): 1-25.

Hall, N. (1999) Young Children's Use of Graphic Punctuation. *Language and Education* 13(3): 178-193.

Hirst, G. & Feiguina, O. (2007) Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22(4): 405-417.

Holmes, D. I. & Forsyth, R. (1995) The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* 10(2): 111-127.

Huberty, C. J. (1975) Discriminant analysis. *Review of Educational Research* 45: 543-598.

Jurafsky, D. (2003) Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In R. Bod, J. Hay, and S. Jannedy (eds) *Probabilistic Linguistics*. Cambridge, Massachusetts/London, England: The MIT Press.

Lowe, D. & Matthews, R. (1995) Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities* 29: 449-461.

Merriam, T. (1996) Marlowe's hand in Edward III revisited. *Literary and Linguistic Computing* 11(1): 19-22.

Morel, J.,Torner, S, Vivaldi, J. De Yzaguirre, L & Cabré, M.T. (1998). *El corpus de l'IULA: etiquetaris. Papers de l'IULA, sèrie informes, 18*. Barcelona: IULA, Universitat Pompeu Fabra.

Mosteller, F. & Wallace, D.L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Nazar, R. & Sánchez Pol, M. (2007) An extremely simple authorship attribution system. In M. T. Turell, M. S. Spassova, & J. Cicres (eds) *Proceedings of the Second European IAFL Conference on Forensic Linguistics/Language and the Law.* Barcelona: Publicacions de l'IULA. 197-203.

Newell, A. (1990) *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

OECD (2004) *Reviews of National Policies for Education. Denmark: Lessons from PISA 2000*. Paris, OECD.

Pardo, A. & Ruiz, M.A. (2002) *SPSS 11. Guía para el análisis de datos*. Madrid: McGraw–Hill.

Reuters corpora (2000). [Online]. Available: http://trec.nist.gov/data/reuters/reuters.html.

Queralt, S. & Turell, M.T. (2013). A semi-automatic authorship attribution technique applied to real forensic cases involving Judgments in Spanish. In R. Sousa-Silva, R. Faria, N. Gavaldà, & B. Maia. (eds.) *Bridging the Gap(s) between Language and the Law: Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*. Porto: Faculdade de Letras da Universidade do Porto.

Shaoul, C. & Westbury, C. (2011). Sequences: Do they exist and do they matter?. *The mental lexicon* 6(1): 171-196.

Sing, S. & Hall, N. (2009). Listening to children think about punctuation. In A. Carter, T. Lillis, and S. Parkin (eds.) *Why Writing Matters. Issues of access and identity in writing research and pedagogy.* Studies in Written Language and Literacy, 12. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Sofkova Hashemi, S. (2003). *Automatic Detection of Grammar Errors in Primary School Children's Texts. A Finite State Approach.* PhD Thesis. Department of Linguistics. Göteborg University.

Sotomayor, C., G. Lucchini, P. Bedwell, M. Biedma, C. Hernández & D. Molina. (2013) Producción escrita en la Educación Básica: análisis de narraciones de alumnos de escuelas municipales de Chile. *ONOMÁZEIN* 27: 53-77.

Spassova, M. S. (2007) The relevance of inter and intra authorial variation in authorship attribution. Some findings on syntactic identification markers. *8th Biennial Conference on Forensic Linguistics/Language and the Law*. University of Washington, Seattle.

Spassova, M. S & Turell, M.T. (2007) The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S. Spassova, and J. Cicres (eds.) *Proceedings of the Second European IAFL Conference on Forensic Linguistics/Language and the Law*. Barcelona: Publicacions de l'IULA. 229-237.

Stamatatos, E., N. Fakotakis & G. Kokkinakis. (2000) Automatic text categorization in terms of genre and author. *Computational Linguistics* 26(4): 471-495.

Teberosky, A. (2001) *Proposta constructivista per aprendre a llegir i a escriure.* Barcelona: Editorial Vicens Vives.

Turell, M. T. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech Language and the Law* 17(2): 211-250.

Tweedie, F. J., S. Singh & D. I. Holmes. (1996) Neural network applications in stylometry: The federalist papers. *Computers and the Humanities* 30(1): 1-10.