Textual Features Indicative of Writing Proficiency in Elementary School Spanish Documents

Diana Yazmín Dueñas Chávez **Gemma Bel-Enguix**
Universidad Nacional Autónoma de México
Grupo de Ingeniería Lingüística
{DDuenasC,GBelE}@iingen.unam.mx

Arturo Curiel
CONACYT-Universidad Veracruzana
me@arturocuriel.com

Abstract

Childhood acquisition of written language is not straightforward. Writing skills evolve differently depending on external factors, such as the conditions in which children practice their productions and the quality of their instructors' guidance. This can be challenging in low-income areas, where schools may struggle to ensure ideal acquisition conditions. Developing computational tools to support the learning process may counterweight negative environmental influences; however, few work exists on the use of information technologies to improve childhood literacy.

This work centers around the computational study of Spanish word and syllable structure in documents written by 2nd and 3rd year elementary school students. The studied texts were compared against a corpus of short stories aimed at the same age group, so as to observe whether the children tend to produce similar written patterns as the ones they are expected to interpret at their literacy level. The obtained results show some significant differences between the two kinds of texts, pointing towards possible strategies for the implementation of new education software in support of written language acquisition.

1 Introduction

Acquiring literacy is not an easy process. Educators have to consider many different variables that may affect student performance, such as their psychological and linguistic development (Flower and Hayes, 1981; McDonald Connor et al., 2011; Defior and Tudela, 1994). The latter is specially relevant when considering that writing isn't the mere transcription of vocal sounds, but an abstract endeavor of language representation. Thus, teachers have to assume that an important cognitive effort is required from the students to understand the nuances of a symbolic encoding, which may be influenced by a myriad of environmental factors (Bissex, 1980; Menn and Bernstein Ratner, 1999).

In this sense, finding an optimal strategy to ensure that a group of students will acquire literacy at the same pace is not straightforward (Bradley, 1988; Anthony and Lonigan, 2004): the learning conditions of each individual are likely different, which may prove challenging for the design of generalized pedagogic approaches (Piaget, 1971; Rogoff, 1984). This situation can complicate critical tasks for the teaching process, such as evaluating the acquisition progress of a group of students. In this regard, data-driven analyses may provide new automatic evaluation tools for teachers, making it possible to dynamically adapt their teaching strategies based on data to improve the learning conditions of specific groups or individuals.

This work presents an exploratory approach to the computational study of written language, oriented towards improving literacy acquisition in school-age children. The idea is to explore whether written productions made by children contain patterns that may be indicative of proficiency, in an effort to pursue novel research on the automatic monitoring of the students' writing skills. To this end, some seminal quantitative analyses were performed over two independent Spanish corpora of child productions. The obtained results were compared against a control corpus, representative of the level of literacy expected from children in the same age group. Early results show

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 113–118
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

that some regularities exist in the texts produced by the children, which contrast with the expected outcome inferred from the control corpus. Identifying these and other possible proficiency indicators may the first step towards the training of robust written acquisition evaluation models.

2 Related work

Research on the written acquisition of Spanish by Zamudio Mesa (2008), Flores Hernández (2012) and Ferreiro & Teberosky (1991) has shown that, starting the acquisition process, children systematically try to codify the words they hear into a simple interleaving of consonants (C) and vowels (V). This translates into a disproportionate use of simple syllabic patters such as CV, VC or CVC, which tends to decrease as the student progresses.

In an ideal learning environment, as the children gain proficiency they should start using more complex patterns such as VCC, CVCC or CCVC (Bowey, 2002; Ferroni et al., 2016). However, some authors claim that, without the proper conditions, children aren't able to perform this transition, which affects their overall academic performance in the future (Ardila and Rosselli, 2014).

Nonetheless, even though some data exists on the evolution of the complexity of children writing in Spanish, as of the authors' knowledge no previous work has explored how it can be assessed automatically by way of a computational method.

Some data on the evolution of reading ability – Bradley (1988), Ferroni and Diuk (2016), Anthony and Lonigan (2004) Bowey (2002) – showed how teachers can prevent future reading and writing children's failures. However, they focused only on speech and not on the complexity of children's writing (Casillas and Goikoetxea, 2007; Levy and Ransdell, 2013).

This paper presents some results obtained with experiments performed over well-known corpora of children writing in Spanish. These results directly contradict the theories of researchers who have previously approached the problem. We show how this contradiction between our data and the language of children as it has been described in the literature is caused by the way the complexity of the texts was measured. In general, the perspicuity tests used to classify the texts assume that writers have a regular proficiency in the use of written language. However, children's writing display phenomena such as lack of punctuation marks and other conventions that have had an impact in the results, as it will be discussed below.

3 Methodology

To identify candidate characteristics that may be indicative of written proficiency, two children-produced corpora were analyzed:

- CEELE[1]: Corpus of 300 documents in Spanish written by children from 7 to 8 years old. The corpus was elicited by asking the subjects to describe their school after showing them an example through a story. Roughly, this prompted the children to write about their daily commute and their usual activities in a normal school day.
- EXCALE[2]: Corpus of 286 documents in Spanish written by children from 7 to 13 years old. It was elicited by showing the students a series of related images and asking them to turn them into a short story (Zamudio Mesa, 2016). Originally, the corpus contains only document scans with no transcriptions, which had to be created for the experiments. In that regard, all documents that were unreadable, incomplete or that didn't hold a story structure (*e.g.* introduction, plot and conclusion) were discarded.

A third corpus of short stories was collected to serve as a control. It served to compare how the children productions fared against adult-written texts for elementary school literacy level:

- Short Stories: 70 texts of between 200 and 250 words written in Spanish, collected from public websites oriented to literacy acquisition in grade school children.

The documents in the three corpora were classified into seven *readability* levels as given by the Sigriszt-Pazos (1993) readability index ($\mathcal{P}$): an adaptation of the Flesch-Kincaid (1948) readability tests for the Spanish language. Equation 1 shows how $\mathcal{P}$ is calculated:

$$\mathcal{P} = 206.835 - 62.3 \cdot \frac{S}{P} - \frac{P}{F} \qquad (1)$$

where:

[1] http://www.corpus.unam.mx:8080/unificado/index.jsp?c=ceele

[2] http://www.inee.edu.mx/index.php/proyectos/excale/corpus-excale

- P corresponds to the total number of words in the document;
- S denotes the total number of syllables; and,
- F is the total number of sentences.

Table 1 shows how documents are classified into seven readability levels according their $\mathcal{P}$ value. An interpretation of each level is provided as well.

$\mathcal{P}$	Level	Interpretation
86-100	1	very easy to read
76-85	2	easy to read
66-75	3	fairly easy to read
51-65	4	plain
36-50	5	fairly difficult to read
16-35	6	difficult to read
0-15	7	very difficult to read

Table 1: Readability level as given by the Sigriszt-Pazos readability index ($\mathcal{P}$).

Once every document in the three corpora was assigned to a level in Table 1, the following measures were calculated for every individual level:

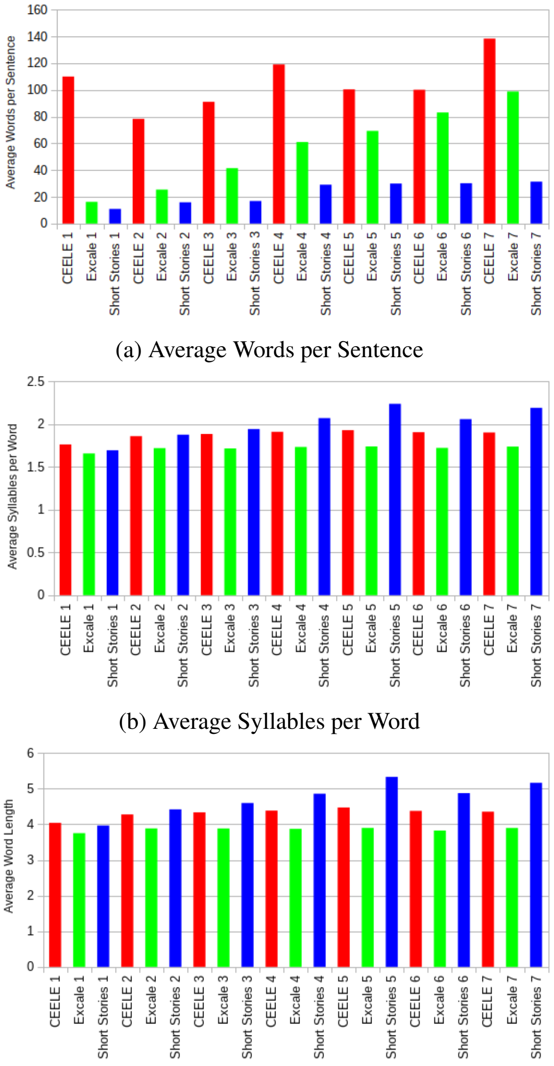
- The average number of words per sentence.
- The average number of syllables per word.
- The average word length.
- The frequency of the syllables per level.
- The frequency of the syllabic patterns appearing in the level.

4 Results

Figure 1 shows three graphs depicting the values calculated for the average number of words per sentence (1a); the average number of syllables per word (1b) and the average word length (1c), for all seven levels in each corpora.

Each graph in Figure 1 shows groups of side-by-side bars for the three corpora, in each of the seven readability levels.

Figure 1a shows that both children produced corpora—CEELE and EXCALE—tend to hold more words per sentence in average than the Short Stories control. Furthermore, the averages per level in both EXCALE and CEELE always surpass the ones from the Short Stories corpus. In general, a strict order between the averages per level is respected: Short Stories < EXCALE < CEELE.



(a) Average Words per Sentence

(b) Average Syllables per Word

(c) Average Word Length

Figure 1: Word statistics for the three corpora.

The generally high number of words per sentence is explained by the lack of punctuation marks in the children corpora. In general, almost no instances of full stops nor semi-colons are to be found in the children's texts; they tended to write the entire document into a single phrase. In itself, this also affected how the documents themselves were classified by the Sigriszt-Pazos formula, as it takes into account the number of words per sentence to calculate the difficulty level. The latter would also help to explain why the correlation between this value and the readability level seems so strong.

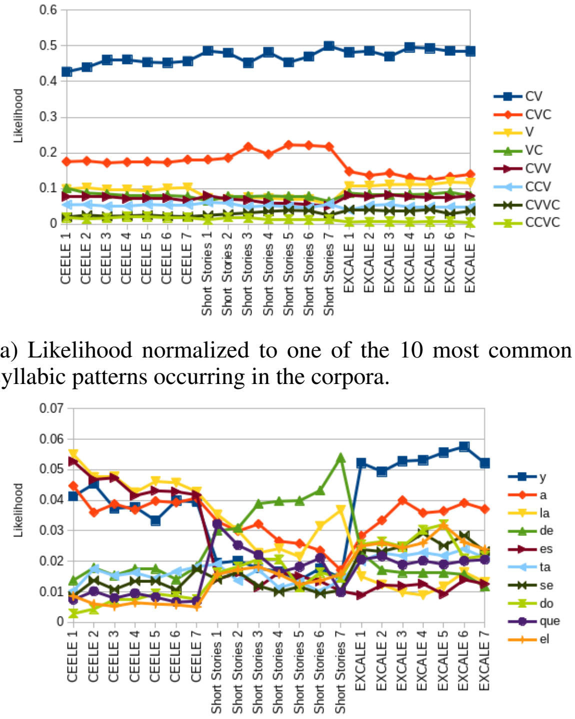
Table 2 shows the Pearson (ρ) correlation values between the average number of words per sentence and the readability level for each corpus.

CORPUS	PEARSON CORRELATION (ρ)
CEELE	0.5456
EXCALE	0.9965
Short Stories	0.8958

Table 2: Correlation between the average number of words per sentence and readability level.

From Table 2 it can be observed that both ρ(Short Stories) and ρ(EXCALE) denote stronger correlation values between the aforementioned variables than ρ(CEELE).

Figures 1b and 1c show that there are no remarkable differences across the three corpora in terms of the average word length (between four and five characters) or the number of syllables per word (around two).



(a) Likelihood normalized to one of the 10 most common syllabic patterns occurring in the corpora.

(b) Likelihood normalized to one of the 10 more common syllables in the corpora.

Figure 2: Likelihood of occurrence of syllabic patterns and syllables.

Figure 2 shows the likelihood of occurrence of the 10 most common syllabic patterns (2a) and syllables (2b) in each readability level of the three corpora. In particular, Figure 2a shows that syllabic patterns tend to occur with similar probability across every readability level and corpus. Simple patterns such as CV and CVC are the most likely to appear with surprisingly regular frequency across corpora. In contrast, Figure 2b shows that the specific syllable realizations of the patterns display a higher level of variability: overall, the relative probabilities for even the 10 most common realizations fall below ten percent. This would indicate that proficient writing skills don't necessarily entail the use of complex syllabic patterns; rather, proficiency would lie on the specific vocabulary used by the speaker, maybe because it contains more words or because it is perceived to be more specialized.

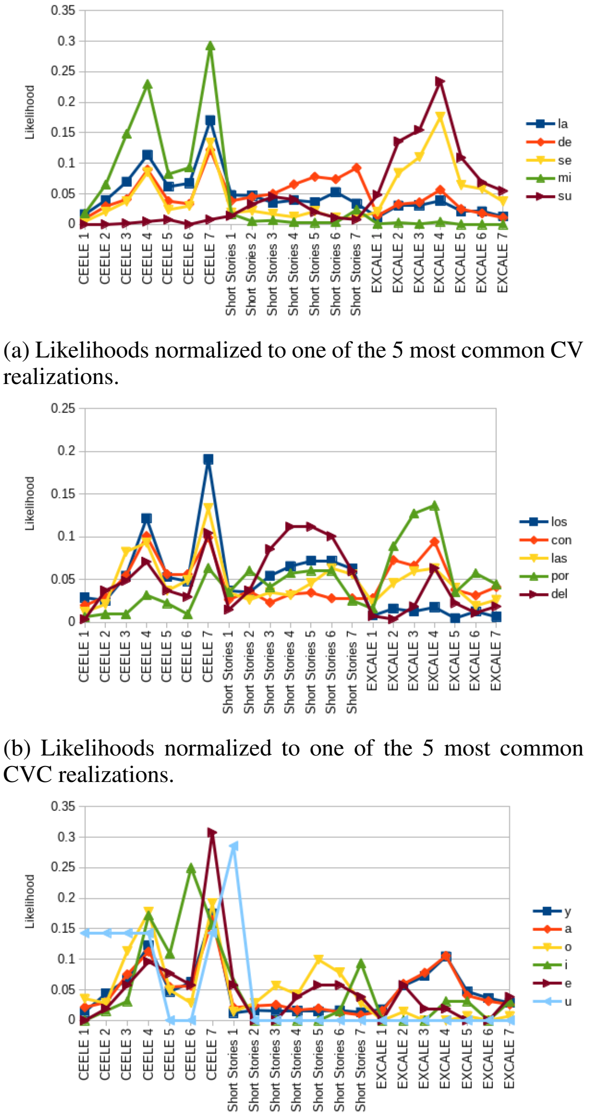
Figure 3 shows the likelihoods of occurrence of the specific realizations for the most frequent patterns: CV (3a), CVC (3b) and V (3c).

Globally, the figure shows that the children in CEELE tend to favor specific syllables in some readability levels for the CV and CVC patterns, such as "mi" and "los" in levels 4 and 7. The EXCALE documents show a similar behavior with syllables "su" and "por". Also, Figure 3c shows that CEELE documents tend to disproportionately favor the use of "e", "u" and "i" as one-character syllables, contrasting with the lower variability shown by both the EXCALE and the control corpus. The results are discussed in the next section.

5 Discussion and Conclusions

The data shows that there might be several characteristics that could help to automatically measure written proficiency. According to the ideas of (Zamudio Mesa, 2008; Flores Hernández and Ramírez Hernández, 2012; Ferreiro and Teberosky, 1991), we expected that children between 7 and 12 years old would already have know how to use punctuation marks and blank spaces between words—particularly, full stops. Clearly, these capabilities had not been acquired by the children whose writing was reported in the corpora, causing very unexpected results.

The more notable deviation corresponds to the average number of words per sentence, which seems to be an strong indicator of literacy; in the CEELE documents, which are expected to show a lower literacy level than the remaining two, the number of average words per sentence explodes. As previously mentioned, the explanation for this is that the children did not use punctuation marks throughout the writings, causing the algorithm to perceive documents as containing only one



(a) Likelihoods normalized to one of the 5 most common CV realizations.

(b) Likelihoods normalized to one of the 5 most common CVC realizations.

(c) Likelihoods normalized to one of the 6 most common V realizations.

Figure 3: Likelihoods of CV, CVC and V pattern realizations.

or two sentences. This happens even with the occurrence of unnaturally long words such as "**CuantosañostienescomoSellamaendondebibes**", product of the erroneous use of whitespace; intuitively this should shorten sentences, however the overall average remained high. More analyses are needed to observe how this variable correlates with others, such as the use of punctuation marks, which might be what is pulling the averages up.

Regarding the use of syllables, the corpora presented instances of invalid Spanish syllabic patterns like strings of consonants without vowels. These irregularities could credibly be indicators of a lack of proficiency; however, the observed probabilities are so low (near zero) that few conclusions can be obtained, as they could correspond to transcription mistakes or else.

For the rest of the patterns, their likelihoods of occurrence remain consistent across all levels on every corpora, meaning that their realizations might give more meaningful information, as explained by the hypothesis of a specialized or more diverse vocabulary. In this regard, Figure 3 provides some evidence that the overuse of simple words and common syllables might be indicative of lack of writing skills. Thus, further exploration is needed on larger corpora, covering written productions by persons with different literacy levels and even learners of Spanish as L2.

Results show that Sigriszt-Pazos readability formula tests productions for expert Spanish writers. Although it measures the complexity of texts written especifically for children, such texts are carefully composed for adapting to the capablities of the readers. However, a child does not have an idea of the parameters that should be used in order to make the text easier. Is it clear, then, that student's productions need different parameters witch calculate their writing proficiency.

In general, more experiments are needed to reach stronger conclusions. Future work will explore how syllabic patterns and syllables combine inside of words, and how this correlates with writing proficiency. This might provide more useful information about the literacy level of the students, rather than just looking at single syllables as it has been done until now.

Finally, it is expected that these studies will lead to the creation of *writability* formulas, which will measure not how readable a text can be, but how difficult it is to write. Moreover, we suggest the creation of a method to measure students' writing skills based on these formulas.

References

Jason Anthony and Christopher J Lonigan. 2004. The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of educational psychology*, 96(1):43.

Alfredo Ardila and Mónica Rosselli. 2014. Spanish and the characteristics of acquired disorders in reading and writing. *Estudios de Psicología*, 35(3):502–518.

Glenda L Bissex. 1980. *GNYS AT WRK: A child learns to write and read*. Harvard University Press.

Judith A Bowey. 2002. Reflections on onset-rime and phoneme sensitivity as predictors of beginning word reading. *Journal of Experimental Child Psychology*, 82(1):29–40.

Lynette Bradley. 1988. Rhyme recognition and reading and spelling in young children.

Angela Casillas and Edurne Goikoetxea. 2007. Syllable, onset-rhyme, and phoneme as predictors of early reading and spelling. *Infancia y Aprendizaje*, 30(2):245–259.

Sylvia Defior and Pio Tudela. 1994. Effect of phonological training on reading and writing acquisition. *Reading and Writing*, 6(3):299–320.

Emilia Ferreiro and Ana Teberosky. 1991. *Los sistemas de escritura en el desarrollo del niño*. siglo XXI.

Marina Ferroni, Beatriz Diuk, and Milagros Mena. 2016. Acquisition of orthographic knowledge: orthographic representations and context sensitive rules. *Psicología desde el Caribe*, 33(3):237–249.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Ana Abigahil Flores Hernández and Esthela Ramírez Hernández. 2012. Jakobsons universalist theory and order of acquisition of consonants in mexican spanish: A case study.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

C Michael Levy and Sarah Ransdell. 2013. *The science of writing: Theories, methods, individual differences and applications*. Routledge.

Fredrick J McDonald Connor, Carol, Barry Fishman, Sarah Giuliani, Melissa Luck, Phyllis S Underwood, Aysegul Bayraktar, Elizabeth C Crowe, and Christopher Schatschneider. 2011. Testing the impact of child characteristics instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, 46(3):189–221.

Lise Menn and Nan Bernstein Ratner. 1999. *Methods for studying language production*. Psychology Press.

Francisco Szigriszt Pazos. 1993. *Sistemas predictivos de legibilidad del mensaje escrito: fórmula de perspicuidad*. Universidad Complutense de Madrid, Servicio de Reprografía.

Jean Piaget. 1971. The theory of stages in cognitive development.

Barbara Rogoff. 1984. *Children's learning in the" zone of proximal development"*. 23. Jossey-Bass Inc Pub.

Celia Zamudio Mesa. 2008. Influencia de la escritura alfabética en la segmentación de sonidos vocálicos y consonánticos. *Lectura y vida*, pages 10–21.

Celia Zamudio Mesa. 2016. Evaluación del corpus excale de escritura.