Authorship and Writing Style: An Analysis of Syntactic Variable Frequencies in Select Texts of Alejandro Casona

Lisa Barboun
Coastal Carolina University

Arunava Chatterjee
Hyperdigm, Inc.

Abstract

This research is a partially computer-assisted transformational-generatively based analysis of select dramatic works of the Spanish author Alejandro Casona. Of interest are the frequencies with which certain syntactic variables occur in the works.
The syntactic structures considered are analyzed within a framework somewhat loosely based on a transformational-generative model first proposed for English in *Syntactic Structures* (Chomsky, 1957), later modified in *Aspects of a Theory of Syntax* (Chomsky, 1965), and adapted for Spanish in *A Transformational Grammar of Spanish* (Hadlich, 1971).
By calculating cross-correlation functions for syntactic variables in selected works of Casona, the researchers are able to compare the frequency of these variables between literary subcategories, as well as between plays that Casona has written for adults and those he has written for children. Of interest is examining whether there is an overall consistency in Casona' s syntactic style. Such consistency might suggest that there are identifiable features of Casona' s style, and perhaps, by extension, the style of other authors. Being able to provide evidence to support the notion that one can determine a blueprint of an author' s style is of interest given the research that has arisen concerning texts that have a disputed or unknown author (e.g. *La segunda Celestina*, *The Federalist Papers*, or works attributed to Shakespeare).

Keywords: computational linguistics, stylistics, syntactic variables.

1. Introduction

The procedure for this study involved identifying and selecting works of Casona that lent themselves to the proposed comparisons, drawing samples from these works, and selecting variables which reasonably accurately could be identified in a text file using computer programs designed for this purpose.

The next steps included creating a sample text file from one of Casona' s works that was not included in the final analysis, manually analyzing this work for the variables in questions, and making a trial run of the programs to determine if they handled the variables as planned. Since some of the variables were not represented in this text file, the programs designed to locate these were also run using contrived text files into which these variables purposely had been integrated.

The last steps were inputting the samples from the selected works, running the computer programs on these samples, grouping the results, calculating the correlations, and considering the similarities and differences between the works that were compared.

2. The Works

The works used for the final analyses included two plays that Casona wrote for children, *¡A Belén pastores!* and *Pinocho y Blancaflor*, and four that he wrote for adults: *El caballero de*

las espuelas de oro, La sinfonía inacabada, Los árboles mueren de pie, and *Prohibido suicidarse en primavera*. *Caballero* and *Sinfonía* are historical works, while *Árboles* and *Prohibido* are fantastic works.

Translations, adaptations and co-authored works were eliminated as candidates. This was to assure that the representative works best reflected only Casona' s style.

3. The Samples

Once the population had been defined (i.e. the specific works to be used had been selected), samples were chosen. Each text file contained a series of records (each consisting of one sentence) which comprised a representative sample of running text from one of the works selected for analyses.

It has been found that neither sentence length nor type is randomly distributed (de Haan, 1993). Assuming that both of these factors are directly related to the occurrence of certain syntactic variables, the non-random arrangement of language would have made random sampling problematic. Therefore, the samples for this study are representative and not random.

3.1 Quantitative Considerations

Sentence length distribution has been found to differentiate among genres (Marckworth and Bell, 1967, qtd. in Marckworth and Baker, 1980). Although this study does not deal with genre, per se, it did consider subcategories of Casona' s plays. Therefore, it was assumed that the works to be compared might display various average sentence lengths. In order to avoid undesirably small or large samples, this study used samples containing a certain number of words rather than a certain number of running sentences.

To make the planned comparisons for the research, it was necessary to analyze samples from at least one work representing each of the following three categories: historical plays for adults, fantastic plays for adults, and plays for children. Since an author may write differently at the start of a work (Kenny, 1982), it could not be assumed that Casona' s style remained consistent throughout a given work. Therefore, three or four samples were drawn from each play analyzed. Moreover, samples were drawn from two works in each category so that comparisons could be made within categories as well as between. Additionally, drawing the samples for each category from different works and acts reduced the possibility of thematically caused repetition of certain structures.

There were a total of twenty samples. All ended with the last sentence completed at or just becoming 1,200 words. This number was based on past stylometric research (Mosteller and Wallace, 1964; Marckworth and Baker, 1980; Stoddard, 1985). The first sample from each work started at the beginning of the text. Additional samples were drawn starting with segments beginning at a breakpoint (e.g. the beginning of an act or scene). For each of the six works, three samples were taken. Each sample came from a different act, or in the case of *Caballero*, section. A fourth sample was taken for both *Árboles* and *Sinfonía*. These additional samples were drawn from the second scenes of the already sampled 3rd acts.

3.2 Qualitative Considerations

Many qualitative considerations, such as treatment of stage directions, foreign terms and elided material had to be settled prior to the analyses; however, it is not essential to detail them here.

4. The Variables

The 59 variables, all quantifiable, were located using various computer programs written by the researchers. Three of these variables were average word length, average number of words per sentence, and number of different words. These were the only variables not analyzed within the adapted transformational-generative framework. Also examined were perfect and progressive constructions. These constructions deal with possible expansions of the predicate in Spanish and are generated via PS rules, but do not involve transformations. There were several variables, including negation and passive voice with *ser*, which involved one-string transformations. Additional variables, including subordination, and comparative and superlative constructions, involved two-string transformations.

5. The Computer Programs

The programs were used to help locate and calculate the number of occurrences of the variables in the text files. It would have been desirable to have been able to design programs that by themselves could determine accurately the frequency of each variable. However, some of the programs were designed to find forms that could be associated with several variables. Since these forms did not all have consistent and different contexts that the computer could be further programmed to recognize for a given variable, it would have been fairly difficult to create programs that by themselves could determine the relative frequency of all the variables with which the forms could have been associated.

To exclude false matches, follow-up manual analyses were conducted on the output.

6. An Example: Double Subordination

This section presents a description of a single variable which the present researchers refer to as ' double subordination' . This description includes: 1. A narrative summary of the transformation, noting exceptions, options, and obligatory rules; 2. The transformation proper, including the structural change, structural description, relevant conditions and an example, and 3. An explanation of the corresponding computer program.

6.1 Narrative Description

In the transformation involving double subordination, an S2 that does not express a known or expected fact is attached via the form *si* to an S1 that expresses a potential proposition. The S1 predicate is expressed in the conditional tense and the S2 predicate in the imperfect subjunctive. The clauses may be inverted; however, *si* must precede the S2 clause. Some native speakers use the imperfect tense instead of the imperfect subjunctive and conditional.

6.2 Transformation

Consider a structural description (SD), where S1 = 1 and S2 = 2, S1 expresses a potential proposition, and S2 does not express a known or expected fact. This structural description undergoes the following structural change (SC) when double subordination occurs: 1 2 --> 1 (conditional) *si* 2 (imperfect subjunctive).

The following is an example:

SD: S1 *Yo soy una culpable*.
1
S2 *Me habría matado ayer*.
2

SC: 1 2 --> *Yo sería una culpable si me hubiera matado ayer.*
1 2

Optional inversion of the clauses is possible, as seen in the following example:

Yo sería una culpable si me hubiera matado ayer. -->
...si me hubiera matado ayer, yo sería una culpable...

Some native speakers use the imperfect tense, as seen in the example below:

Si me hubiera matado ayer, yo sería una culpable. -->
Si me había matado ayer, yo era una culpable.

6.3 Computer Program

Two keyword sets were defined in the program designed to locate examples of double clause subordination. Keyword set 1 included the following forms preceded by an arbitrary number of characters and followed by a blank: ía, ías, íamos, íais and ían. Keyword set 2 included the following forms preceded by an arbitrary number of characters and followed by punctuation (it should be noted that the candidates for 'punctuation' included a blank space): ara, aras, áramos, aran, arais, era, eras, éramos, erais and eran. The program looked for either of the following: 1. a member of keyword set 2 preceded by *si* and followed by a member of keyword set 1, or 2. a member of keyword set 1 followed by an arbitrary number of characters followed by *si* followed by a member of keyword set 2

7. Results

Five comparisons were made by calculating cross-correlation functions for the n-tuple variables (v_1 ... v_n). These included: 1. comparisons between the first and second scenes of the last acts of *Árboles* and *Sinfonía*; 2. comparisons between the acts in a given work (e.g. using the correlation <v, v´> to calculate values such as $v_n p_1 a_1$ x $v_n p_1 a_2$ where v represents a given variable, p represents the first play considered, and a_1 and a_2 refer to the first two acts from which samples were taken from p_1; 3. comparisons of the average variable values between the works in a given category (e.g. between *Pinocho* and *Belén*, *Prohibido* and *Árboles*, and *Sinfonía* and *Caballero*; 4. comparisons of the average variable values between the fantastic (*Prohibido* and *Árboles*) and historical (*Sinfonía* and *Caballero*) plays for adults, and 5. comparisons of the average variable values between the plays written for children (*Pinocho* and *Belén*) and those written for adults (*Prohibido*, *Árboles*, *Sinfonía* and *Caballero*).

Since the findings did not indicate significant act/scene differences, the three remaining types of comparisons were based on data averaged across the acts. The average variable values from the plays from which an additional scene was drawn were determined by adding the values from the first and second scenes of the twice sampled act, dividing the result by two, adding this result to the values from acts 1 and 2, and dividing by three.

8. Conclusions

Although no predictions were made prior to the analyses, three outcomes seemed intuitively probable. One was that the language in the plays for adults would be more complex structurally (i.e. contain more transformations) than that in the plays for children. Another was that the frequency of the syntactic structures examined would differ for the fantastic and historical plays. The third was that plays written for children would average fewer different lexical forms than those written for adults. Any of these outcomes might have complicated efforts to show that Casona has a predictable syntactic style, or to actually identify his style.

None of these outcomes, however, occurred. In fact, a striking feature of Casona' s syntactic style is its consistency across works. This consistency, regardless of type of play or intended audience, suggests that there are indeed identifiable features of Casona' s syntactic style, at least across his dramatic works. This finding is of interest since it supports the assumption on which some stylistic studies are based--that an author has an identifiable style. Using distributions based on the results of this study, the present researchers hope to be able to begin to form a blueprint of Casona' s syntactic style that they can use to examine other types of his works.

References

Casona A. (1967). ¡A Belén, pastores!. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. Los Árboles mueren de pie. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. La barca sin pescador. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. El caballero de las espuelas de oro. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. Prohibido suicidarse en primavera. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. La sinfonía inacabada. In Sainz de Robles, F. editor, *Obras completas*. Aguilar.
---------. La tercera palabra. In Sainz de Robles, F. editor, *Obras completas*. Aguilar
---------. (1983) Pinocho y Blancaflor. In Arce, E. editor, *Tres farsas infantiles*. Ediciones Noega.
Chomsky N. (1957). *Aspects of a Theory of Syntax*. MIT Press.
---------. (1965) *Syntactic Structures*. Mouton.
de Haan P. (1993). Sentence Length in Running Text. In Souter, C. and Atwell, E. editors, *Corpus Based Computational Linguistics*. Rodopi.
Hadlich R. (1971). *A Transformational Grammar of Spanish*. Prentice Hall, Inc.
Kenny A (1982). *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. Pergamon Press, Inc.
Marckworth M. and Baker Wm. (1980). A Discriminant Function Analysis of Co-variation of a Number of Syntactic Devices in Five Prose Genres. In Prideaux, G., Derwing, B. and Baker Wm. editors, *Experimental Linguistics*. E. Story-Scientia.
Mosteller F. and Wallace D. (1964). *Inference and Disputed Authorship: The Federalist Papers*. Addison-Wesley Publishing Co.
Stoddard S. (1985). Determining the Relative Cohesiveness of Written Texts. In Johnson, E. editor, *The Proceedings of the 1985 Conference on English Language Literature Applications of SNOBOL and SPITBOL*. Dakota State College.