https://doi.org/10.7203/Celestinesca.11.19663

CELESTINA, AUTHORSHIP, AND THE COMPUTER

James L. Wyatt
Florida State University

[Professor Wyatt presented a slightly longer version of this paper at the Seventh Congress of Applied Linguistics (Brussels) to an audience primarily of people interested in computational linguistics. It was not, therefore, intended for literary scholars although, when the project is completed, the results will undoubtedly interest *Celestina* scholars around the world. One final caution: Professor Wyatt is not attempting to assign authorship to anyone; rather, his goal is to identify, based on cluster analysis of linguistic features, the probable *number* of authors/collaborators in the *Comedia* and *Tragicomedia* versions. Ed.]

For nearly 500 years readers of the Renaissance Spanish classic, *Celestina*, have reflected upon the actual authorship of the work. The work first appeared anonymously in Burgos in 1499 with the title *Comedia de Calisto y Melibea*: the natural question was, "who wrote it?" With the appearance of an enhanced edition (Seville 1501), an introductory letter "El Autor a Vn su Amigo" said that the (still anonymous) author had found a manuscript which interested him enormously, and he had taken that work as the first act and added fifteen acts of his own authorship. However, following the letter appeared eleven eight-line stanzas, and the first-letter acrostic, reading down the lines of verse, revealed that "el bachjller fernando de royas *acabo* la comedia de calysto y melybea y fve nascjdo en la pvebla de montalvan." All this raised new questions. Were there really two authors, one for the first act, and another for the remaining fifteen?

When an expanded edition was published in Seville, later than the one just mentioned, it bore a new title, *Tragicomedia de Calisto y Melibea*. This new version contained 21 acts; there was a prologue which made reference to the addition of new material, an altered letter, and some changes in the acrostic verses. With this edition the questions

change: Was there a single author? were there two authors? or could there be three (or even more) authors?

From time to time these questions of authorship have been considered or revisited. The editor of the edition I have used for this research, Julio Cejador y Frauca, stated in 1913 that he strongly felt that there were two authors. He believed that Fernando de Rojas certainly wrote the original 16 acts, and he believed that Alonzo de Proaza, editor of the Seville *Tragicomedia* was most probably the author of the added parts.[1]

Spain's renowned critic and literary historian Marcelino Menéndez y Pelayo stated in his work *Orígenes de la novela* (1910), that "the *Celestina* with 16 acts and the *Celestina* with 21 belong to the same author, and for the reasons given the author could be none other than Fernando de Rojas."[2] Menéndez y Pelayo's study of the *Celestina* has been considered important enough in Spanish criticism that it has been republished four times, apart from the other material contained in the *Orígenes*.

Countering the judgment of Menéndez y Pelayo, Manuel Criado de Val concluded after analyzing more than 10,000 verb forms that act one of the *Celestina* was not written by Fernando de Rojas but all the other original acts and interpolations were.[3] Two other researchers maintained in a 1971 study that Juan del Encina actually wrote *Celestina* and that Fernando de Rojas altered it.[4]

Since the computer has already been utilized as an aide to infer authorship of Biblical writings, of Shakespeare's works, and of *The Federalist Papers*, it seemed to me altogether appropriate to use the computer to make various analyses of *Celestina*. What follows will describe my research-in-progress; it does not attempt to arrive at final results or conclusions. It describes, rather, my preparation of *Celestina* in machine-readable form, methods of error checking, the process of normalization of spellings, the creation of files for processing with a view to producing frequency lists and statistical studies, and the nature of certain syntactical studies.

My question at the very beginning was whether the additions to the original *Celestina* would be copious enough to constitute a valid sample for comparison with its other constituent parts. The authors of the computer-aided study of *The Federalist Papers* concluded that a sample of 2000 words was statistically valid in attributing authorship of

those papers. (They compared the occurrence of function words in writings of likely known authors with those in the anonymous writings comprising *The Federalist Papers*).[5]

The early processing of the text of *Celestina* showed that the added parts contained approximately 13,000 words, while the original *Celestina* contained about 48,000 words, the first act accounting for 8,665 of these. These calculations were an indication that the added parts and act one, respectively, represented very generous and adequate sample sizes for analysis.

Since the study of *The Federalist Papers* concerned function words, and that was to be a major focus in this study, I tested the 2000-word sample size by comparing samples from one work of three known authors and samples from three different works of one known author. The purpose was to determine whether the use of function words could distinguish three known contemporary authors, and whether works of a single author would display a consistent use of function words. In all, 24 samples of 2,000 words each were processed to produce frequency lists, which were then compared as to function word use. The results of these preliminary studies of modern American fiction encouraged me to continue with the authorship study of *Celestina.*

The text for this project, except as noted in later remarks, is the edition with introduction and notes by Julio Cejador y Frauca, first published in 1913 and reprinted eleven times since (Madrid, Espasa-Calpe, 11th ed. 1984). Cejador utilized Foulché-Delbosc's 1902 edition of the *Comedia* (Burgos 1499), with its sixteen acts, and the 1514 Valencia edition of the *Tragicomedia*, reproduced in 1899-1900 by E. Krapf in Vigo. Cejador y Frauca replicated the 1499 edition by setting it in roman type, and all the later additions in italics. He made some corrections and restored some mistaken "corrections" to their original form.

The text contains added parts in all acts except act 21. Act one has but 14 added words, act two has 32 added words, and act five contains five added words. However, Acts 15, 16, 17, and 18 are, in their entirety, new. In between these two extremes the added parts amount to the insertion of several words at a time and groups of several sentences at a time. Act 19 is approximately 80 percent new, consisting entirely of added material from the beginning to a point where the original text resumes uninterrupted to the end of the act.

The work was done on a terminal at my home and stored on the Florida State University's CDC Cyber 120 model 760. My use of the CDC Text Editor made error detection and correction a simple task, since errors spotted reading the lines of text on the terminal's screen could be corrected on the spot, making verification instantaneous. But the Text Editor was even more useful in the task of normalizing the spelling of the text, necessary because of various spellings of the same words and the same spelling of different words. Often the same word is spelled variously in the same sentence. Some words are spelled alternately with initial *h* and initial *f*, while others are spelled with or without an initial *h*. A number of words are alternately spelled with *b* or *v* in free variation. Some words spelled with *-ct-*, *-ss-*, *-bd-* and *-ll-* are also spelled with *-t-*, *-s-*, *-ud-*, and *-l-*, respectively. Some high frequency verbs have variant spellings of their stems, and several high frequency words spelled uniquely today to distinguish grammatical function share spellings with other words in the text. Accent marks are used or not used with three very high-frequency function words (the preposition *a* and the conjunctions *o* and *e*), and the conjunction *e* is spelled alternately *e* and *y*. Several verbs have alternate paradigmatic endings (*do*, *doy*; *está*, *estoy*) and some lexical items are written as one or two words. Accents indicating grammatical function (*como*, *cómo*) are used or not used on a number of function words. The text's orthography had to be normalized if frequency and statistical calculations were to be made.

Necessary changes fell into two categories, those which were context free,and hence global, and those which were context sensitive and had to be considered on a one-at-a-time basis by inspection of the environment to determine function.

The Text Editor made quick work of the context free changes. The important thing was to state accurately the several punctuational environments of lexical items to be changed. Each possible punctuational environment required a search-and-change operation using the Text Editor. A list of spelling change rules was prepared and the rules were then applied with all possible left and right patterns of punctuation stated.

Each of the many hundreds of context-free changes required that the Text Editor make a character-by-character search of the entire text for each punctuational environment. Only several seconds were required for each of the commands to make changes. After each search and change operation, the Text Editor stated the number of changes made.

In the case of context-sensitive changes, the Text Editor found the location of each potential change in a matter of several seconds, then human intervention at each point had to determine whether any change was to be initiated. An example was the occurrence of the form *ay* in the text. This could represent an interjection, locative adverb, or verb. Other examples were the occurrence *fijo* which could represent a noun or an adjective, *so* an adverb or a verb, *solo* an adjective or an adverb, *este* a demonstrative pronoun or adjective, *que* an introducer of clauses or an interrogative, *mas* a conjunction or an adverb, and so on. Contractions had to be restored to their separate forms.

The Text Editor was a great aid, but it was not a programming language. For those tasks which required more than simple searches, the SNOBOL4 programming language, or rather its SPITBOL version, was put to use (SNOBOL4 is a programming language developed at the Bell Telephone Laboratories expressly for the manipulation of natural language). SPITBOL was used to create two files used for still more error checking.

The principal error checking process finally over, SPITBOL was used to create separate files of *Comedia* words and words from the interpolations in the *Tragicomedia*, simply as word lists without punctuation, one word per record. These files were then alphabetically sorted by a CDC utility program. Next a SPITBOL program read each file of sorted words, counted like words and created still another file listing each word and its frequency and giving the total number of words read and the total number of different words read. These word frequency lists will now be used in arriving at judgments concerning authorship; but they serve also another purpose. They make it easy to identify high frequency function words to be used as input for running a cluster analysis program in FORTRAN, contained in *BMDP-79, Biomedical Computer Programs, P-Series*.[6] Additionally, SPITBOL is now being utilized to reformat the files of original and added words as required by the FORTRAN statistical package.

While I have already spent many hundreds of hours on this project and will no doubt spend hundreds of additional hours before its completion, the use of the computer with its Text Editor and sort utility packages, the SPITBOL version of the BNOBOL4 Programming Language, and the Biomedical cluster analysis program are greatly facilitating a project which some years back would have required years or even a lifetime to complete.

Up to this point attention has focused on single lexical items only. Some of these items in themselves indicate syntactic features, among them certain noun phrase constituents, conjoined elements, negation, interrogative constructions, adjective or noun clauses or exclamations (undifferentiated), adverbial clauses of time, place, and manner, and subordinate clauses.

The CDC Text Editor has been used to initiate a series of searches for groups of words indicating syntactic features. The groups of words are being joined by arbitrary symbols to cause them to be picked up as single items in future runs to produce frequency counts. Again, the results will also be used as input for the subsequent cluster analysis program.

After consideration of the problem presented by the differences between the 1499 edition of the *Comedia* and the enhanced *Tragicomedia*, I plan to consider the problem of the authorship of act one of *Celestina*, supposedly picked up and added to by the author of acts two through 16.

Omedia intitulada Teſo
rina la materia dela quales vnos amores
de vn penado por vna ſeñora / y otras per-
ſonas adherentes. Hecha nuevamẽte por
Jayme de Huete. Pero ſi / por ſer ſu natu-
ral lengua Aragoneſa / no fuere por muy cẽrados termi-
nos / quãto a eſto merece perdon Los Interlocutores ſon
los infrapueſtos / y es de notar que el Fravle es ſaleador.

Jaime de Huete. COMEDIA TESORINA [obra celestinesca, h. 1530]

NOTES

[1]Fernando de Rojas *La Celestina*, ed. Julio Cejador y Frauca. Clásicos Castellanos, 20 y 23, Madrid: Espasa-Calpe, 1966.

[2]Marcelino Menéndez y Pelayo. *La Celestina* (estudio). Col. Austral 691, Buenos Aires: Espasa-Calpe, 1947. The original study first appeared as part of vol. 3 of the author's *Orígenes de la novela* (1910).

[3]Manuel Criado de Val. *Indice verbal de 'La Celestina'*. Anejos de la Rev. de Filología Española, 64, Madrid: Rev. de Filología Española, 1955.

[4]A. Sánchez Sánchez-Serrano and María R. Prieto de la Yglesia. *Solución razonada para las principales incógnitas de 'La Celestina'*. Madrid: Gráficas Breogán, 1971.

[5]Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Reading, Massachusetts: Addison-Wesley Publishing Co., 1964, on page 249.

[6]*BMDP-79, Biomedical Computer Programs, P-Series*. Ed. by W. J. Dixon and M. B. Brown. Berkeley: Univ. of California Press, 1979.

Comedía llamada Vidriana compueſta poz Jayme de Gueta agoza vnena- mẽte; enla qual ſe recitan los amozes de vn ca- uallero yde vna ſeñozadeAragõ a cuya peticiõ poz ſer les muy ſierno ſe ocupo en la obza pze- ſente: el ſucceſſo y fin de cuyos amozes va meataphozicamẽ- te tocado juſta elpzoceſſo y execucion de aquellos.hay los in terlocutozes ſiguyentes.·.

Jaime de Huete. COMEDIA VIDRIANA, obra celestinsca, h. 1530.