Delta Inside Valle-Inclán

Stylometric Classification of Periods and Groups of His Novels

José Calvo Tello (Würzburg)

ABSTRACT: In this article, I analyze the internal periodization and group of works in a corpus of novels by Valle-Inclán. The technique used is supervised Machine Learning (classification, K-Nearest Neighbors), using the stylometric Delta distances (cosine Delta, 5000 MFW) as features. The evaluation shows that this technique classifies correctly the undisputed texts by literary scholars. The prediction about the disputed texts establish that the year of greatest stylistic change in Valle was 1920 and not 1905. This technique also allows us to classify the works by Valle in only three groups: the *sonatas*, a *social-bélico* group, and the later esperpentic works.

ZUSAMMENFASSUNG: Im vorliegenden Beitrag analysiere ich die interne Periodisierung und Werkgruppen in einem Valle-Inclán Roman-Korpus. Die angewandte Methode ist überwachtes Machine Learning (classification, K-Nearest Neighbors). Dabei werden stilometrische Delta Distanzen (cosine Delta, 5000 MFW) als Parameter verwendet. Die Überprüfung zeigt, dass mithilfe der angewendeten Methode die nicht-umstrittenen Romane Valle-Incláns richtig klassifiziert werden. Die These bezüglich der umstrittenen Texte wird dahingehend bestätigt, dass der bedeutende stilistische Wechsel innerhalb des Romanwerks von Valle-Inclán 1920 stattfand und nicht wie bisher behauptet 1905. Die Methode erlaubt es uns außerdem, Valles Werk in drei Gruppen einzuteilen: die *sonatas*, eine *social-bélico*-Gruppe und später dann die *esperpentischen* Werke.

KEYWORDS: stylometry; Valle-Inclán; novel; Machine Learning; classification

SCHLAGWÖRTER: Stilometrie; Valle-Inclán; Roman; überwachtes Machine Learning; Klassifizierung

1. The Novels by Valle-Inclán

Valle-Inclán is one of the most influential Spanish writers of the 20th Century. Among other genres, he published 13 novels during his life: four *Sonatas*, three historical novels about the *Carlista* War, three avant-garde-historical novels[1] (part of *El ruedo ibérico*) and three other novels (*Flor de Santidad, La media noche* and *Tirano Banderas*).

[1] The last one, *Baza de Espadas*, was only published partly in a magazine before Valle's death.

The traditional studies of Valle-Inclán's works show some dispute about the classification of his novels as far as periodization and grouping his works is concerned. While it is clear that the early novels (*Sonatas*) and the late novels (*El ruedo ibérico*) are part of two different periods of Valle's writing, it is unclear which group the rest of the novels should be assigned to.

In a similar way, while some novels belong clearly to a certain group of works (the four *sonatas*, the three novels about the *Carlista* war, the three avant-garde-historical novels), the classification of the other three is unclear.

2. Stylometric Clustering and Classification

Since Burrows proposed Delta (2002), it has become one of the most widely used methods of Digital Humanities for authorial recognition. Many researchers have already pointed out that other aspects (signals) can affect its results, like genre,[2] gender,[3] narrative perspective, or period.[4] For that reason corpora tend to be created trying to neutralize these phenomena. Hoover[5] and Jannidis and Lauer[6] have used Delta with corpora of a single author in order to analyze the internal structure of the works clustering the Delta values.

In my proposal for the Hispanic Digital Association Conference in 2017[7] I explored to what extent the homogeneity of the clusters based on Delta values correlate with the metadata in single author corpora. The results of this

[2] John Burrows,"'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3 (2002): 267–87, here 271; Christof Schöch, Ulrike Henny, José Calvo Tello and Stefanie Popp, *The CLiGS Textbox*, (Würzburg: University of Würzburg, 2015) https://github.com/cligs/textbox.

[3] Shlomo Argamon, Moshe Koppel, Jonathan Fine and Shimoni Anat Rachel, "Gender, Genre, and Writing Style in Formal Written Texts," *Text and Talk*, 23 (2003): 321–46. https://doi.org/0165-4888/0023-0321.

[4] David L. Hoover, "Testing Burrows's Delta," *Literary and Linguistic Computing* 19, 4 (2004): 453–75, here 455.

[5] David L. Hoover, "A Conversation Among Himselves: Change and the Styles of Henry James" in *Digital Literary Studies*, edited by David L. Hoover, Jonathan Culpeper and Kieran O'Halloran, (New York & London: Routledge, 2014) 90–119.

[6] Fotis Jannidis and Gerhard Lauer, "Burrows's Delta and Its Use in German Literary History" in *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, (Rochester: Camden House, 2014) 29–54, gerhardlauer.de/index.php/download_file/view/335/1/.

[7] José Calvo Tello, "What Does Delta See inside the Author?: Evaluating Stylometric Clusters with Literary Metadata," in *III Congreso de La Sociedad Internacional Humanidades Digitales Hispánicas: Sociedades, Políticas, Saberes* (Málaga, 2017), 153–61. http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.

research show that form, period, and the group of works tend to be differentiated by the clusters of Delta.

Although more present in the stylometric practice, clustering, an unsupervised machine learning method, has some disadvantages in compare to supervised methods: first, the known and unknown cases are not clearly divided. Second, the evaluation of the algorithm is more complicated than in classification tasks. Third, the amount of information in a dendrogram is strongly reduced: we see the closeness between the most similar texts but the rest of the distances to the other texts are only represented indirectly by the distances between the subclusters. And finally, the dendrogram used normally in stylometry does not show clear groups, making the evaluation even harder.

However, the results for the clustering of the corpus of the novels of Valle (stylo version 0.6.5, cosine Delta, 5000 MFW) based on the Delta matrix are:

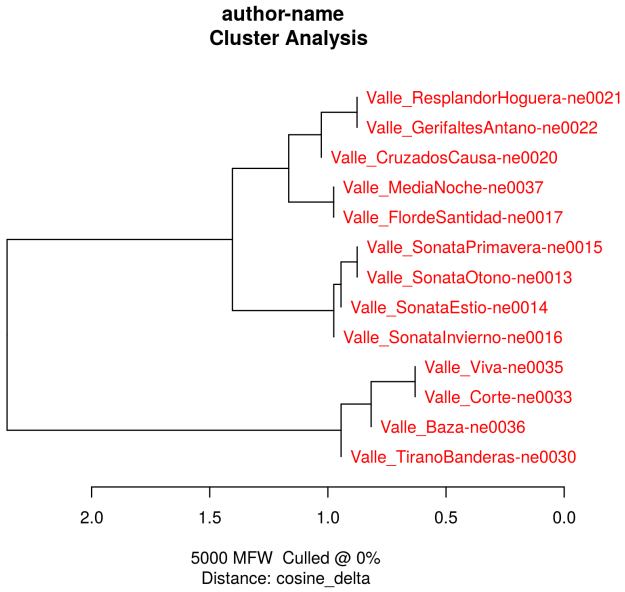

Figure 1: Dendrogram of Valle's novels

Although the dendrogram contains a clear *sonatas* subcluster in the middle, the rest is rather imprecise: do the three novels on the top constitute their own subcluster or should they be considered in conjunction with the

following two? What is the correct number of subclusters to be considered in this cluster? Two? Seven? Something in between?

In general, the kind of tasks for which stylometry is used can be fitted in a supervised task, such as classification, which helps to better answer the questions mentioned in the last paragraph. That is why in this article I use the Delta matrix as features for the classification task. The idea is to conserve the features that we know are the best for the authorial research (Delta values), but instead of deleting information to cluster them, we will use all the distances to the rest of the texts to classify them.

The corpus is composed of 13 works by Valle-Inclán, all of them part of the *Corpus of Novels of Spanish Silver Age* (CoNSSA). This collection is encoded in XML-TEI, validated and enhanced with metadata. The texts were already digitized by projects such as Gutenberg Project, Archive.org and ePubLibre. The CoNSSA will be available at the end of the CLiGS project and some of Valle's works are already part of the Textbox.[8] In any case, all the texts, metadata and code (as Jupyter Notebook) for this article are available in a GitHub repository[9]. In order to clarify the specific steps, the analysis is reproducible and gives the opportunity to reuse the data for other purposes.

3. Periods of Valle-Inclán's writing

I will first analyze the various periods in which Valle-Inclán's novels are inscribed. Some scholars have pointed out that the year of 1920 entails an important change in Valle's style.[10] After this year, he produces the *esperpentic* works in two different genres: novels (*Tirano Banderas, Corte de los milagros*....) and theater (*Cara de Plata, Luces de bohemia*). This hypothesis would group nine novels in the first period, and four in the second one.

Nevertheless, this date is not completely unproblematic. A contemporary intellectual to Valle, Maeztu, assigns the change of period not to 1920 but to 1905.[11] Although less accepted, some information about the subgenres and

[8] Schöch et al., *The CLiGS Textbox*.

[9] https://github.com/morethanbooks/publications/tree/master/Delta_Valle.

[10] José-Carlos Mainer, "Ramón Del Valle-Inclán," in *Historia y crítica de la literatura española 6*, (Barcelona: Crítica, 1980) 289–97, here 292; José-Carlos Mainer, *La Edad de Plata (1902–1939). Ensayo de interpretación de un proceso cultural* (Madrid: Cátedra, 2009), here 166; Felipe B. Pedraza Jiménez and Milagros Rodríguez Cáceres, *Manual de literatura española. 8: Generación de fin de siglo: introducción, líricos y dramaturgos* (Pamplona: Cénlit Ed., 1986, 2. ed.), here 613–16.

[11] Adolfo Sotelo Vázquez, "Valle-Inclán y Ramiro de Maeztu (Dos semblanzas de Valle por Maetzu: 1899 y 1936)" *Cuadernos Hispanoamericanos* 438 (1986): 83–114, here 90.

even characters of the works would support this theory: after 1905 Valle wrote two different series of novels that can be considered historical: *Las guerras carlistas* (1908–1909) and *El ruedo ibérico* (1927–1932). Valle even started a collection of plays called *Las comedias bárbaras* in 1907, which was finished only in 1923, a collection that would surpass the frontier of 1920. The protagonist of the theater play of 1923, Cara de Plata, is also the main character in a novel of 1909. The hypothesis of change in 1905 would then group five novels in the first period (the *sonatas* and *Flor de santidad*) and eight in the second one.

Both hypotheses can be condensed in the following table:

Title	Year of publ.	Hypothesis 1905	Hypothesis 1920	Certain first period	Certain second period
Sonata de otoño	1902	1st period	1st period	yes	no
Sonata de estío	1903	1st period	1st period	yes	no
Sonata de primavera	1904	1st period	1st period	yes	no
Flor de santidad	1904	1st period	1st period	yes	no
Sonata de invierno	1905	1st period	1st period	yes	no
Los cruzados de la Causa	1908	2nd period	1st period	no	no
El resplandor de la hoguera	1909	2nd period	1st period	no	no
Gerifaltes de antaño	1909	2nd period	1st period	no	no
La media noche	1917	2nd period	1st period	no	no
Tirano Banderas	1926	2nd period	2nd period	no	yes
La Corte de los Milagros	1927	2nd period	2nd period	no	yes
Viva mi dueño	1928	2nd period	2nd period	no	yes
Baza de espadas	1932	2nd period	2nd period	no	yes

Table 1: Works of Valle arranged in two competing chronological groups

As we see, both hypotheses diverge on the four novels published after 1905 and before 1920, but they do concur regarding the other nine. That means that we can approach this problem from a supervised classification task in Machine Learning: we take the novels whose period is certain, put them in a train corpus and evaluate the model. If acceptable results are achieved, the model can be applied to the novels for which the hypothesis differ and let the model predict their periods.

3.1. Methodology

For this work the Delta matrix is created with *stylo*[12] (5000 MFW, cosine Delta) using the entire corpus of the novels. After that, the Delta Matrix is split in two matrices: one with the distances of the texts whose periodization is clear (the train corpus) and another with the disputed texts (the test corpus). That is, the distances of each text to the rest of Valle's corpus are treated as features and the metadata about period as labels.

Different classifiers were tested: Decision Trees, Logistic Regression, Random Forest, Supported Vector Machines and K-Nearest Neighbors (KNN). The latter two showed the best results and finally KNN was chosen since it does not require the different classes to be equally represented by data points (that means, the corpus does not have to contain equal amount of texts for, in this case, each period).

3.2. Evaluation and Results

To evaluate the model, I used cross validation, a technique in which "the data is instead split repeatedly and multiple models are trained".[13] In this case the cross validation was repeated four times since the second period only had four different undisputed texts. All of the four times, the accuracy scores were 1.0: the algorithm achieved every time a perfect classification in the train corpus. The evaluation is also correct if we add to the corpus Valle's other prose texts:[14] three collections of short stories (*Femeninas, Jardín umbrío* and *Corte de Amor*) and an essay (*Lámpara maravillosa*).

If the algorithm classified the texts with an unchallenged period correctly, what is its prediction for the four disputed texts published after 1905 and before 1920? The algorithm provides the following results:

Title	Year	Predicted Period
Los cruzados de la Causa	1908	first
El resplandor de la hoguera	1909	first
Gerifaltes de antaño	1909	first
La media noche	1917	first

Table 2: Predicted period of disputed text

[12] Maciej Eder, Mike Kestemont and Jan Rybicki, "Stylometry with R: A Package for Computational Text Analysis," *The R Journal* 16,1 (2016): 1–15.

[13] Andreas C. Müller and Sarah Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists* (Beijing: O'Reilly, 2016), here 254.

[14] See texts and code in the subfolder of GitHub "/data/prose/".

As we observe, each disputed text is classified by the algorithm as part of the first period. That means, the algorithm agrees with the hypothesis of 1920 as the year of change for Valle over the one that sets the change in 1905. Still, we don't know to what extend the hypothesis of 1905 is correct or wrong. To quantify the agreement of each hypothesis with the algorithm, we can evaluate both hypotheses with cross validation (four times) and show the accuracy scores as box plots:

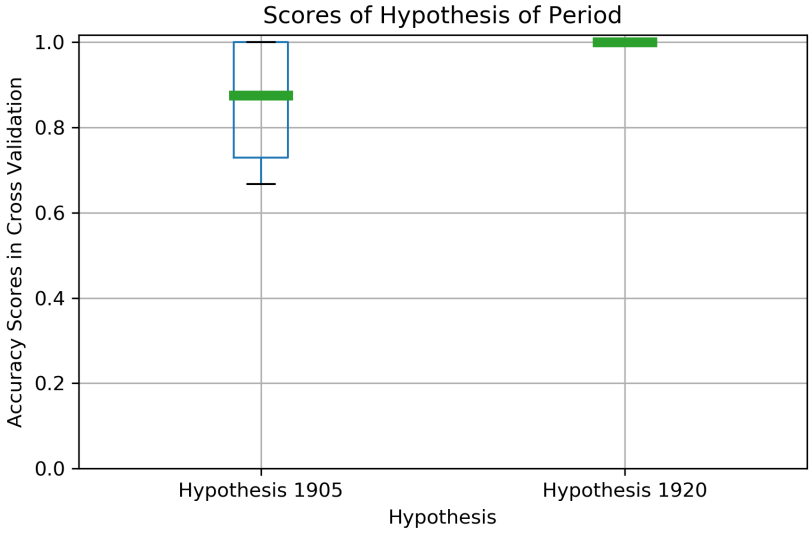

Figure 2: Accuracy scores for hypothesis of periodization in Valle's novels

Here we can conclude that while the median of accuracy scores of the hypothesis of 1905 is 0.88, all the scores of 1920 achieve perfect results of 1.0. That means, 1905 does represent a change of Valle's style, but 1920 defines clearer periods. The shift of Valle's style towards the *esperpento* is stronger than the similarity between the two groups of historical novels.

4. Groups of Valle-Inclán's Works

A more complicated classification of Valle's works regards how to group his texts in relative smaller collections of texts (in Spanish *series*), which will be referred here as *group of works*. All the *sonatas* and the *carlista* novels share subtitles ("Memorias del Marqués de Bradomín" in the first case, "La guerra carlista" in the second), while it is undisputed that the last three novels belong

to "El ruedo ibérico", a group that should have contained nine works but only the first three were published during Valle's life.[15]

Different researchers have proposed alternative ways to group the other three texts (*Flor de santidad*, *Tirano Banderas* and *La media noche*). In the following table three different classifications of groups of works, proposed by literary scholars such as García de Nora (1963), Pedraza Jiménez and Rodríguez Cáceres (1986) and de Juan Bolufer (2000), are summarized:

Title	Year	Subtitle	Pedraza-Rodriguez	Nora	Bolufer
Sonata de otoño	1902	Memorias del Marqués de Bradomín	sonatas	modernista	sonatas
Sonata de estío	1903	Memorias del Marqués de Bradomín	sonatas	modernista	sonatas
Sonata de primavera	1904	Memorias del Marqués de Bradomín	sonatas	modernista	sonatas
Flor de santidad	1904	Historia milenaria	flor	modernista	flor
Sonata de invierno	1905	Memorias del Marqués de Bradomín	sonatas	modernista	sonatas
Los cruzados de la Causa	1908	La guerra carlista	guerra carlista	carlista	bélica
El resplandor de la hoguera	1909	La guerra carlista	guerra carlista	carlista	bélica
Gerifaltes de antaño	1909	La guerra carlista	guerra carlista	carlista	bélica
La media noche	1917	Visión estelar de un momento de guerra	otras	not analyzed	bélica
Tirano Banderas	1926	Novela de tierra caliente	tirano	esperpéntica	esperpéntica
La Corte de los Milagros	1927		ruedo	esperpéntica	esperpéntica
Viva mi dueño	1928		ruedo	esperpéntica	esperpéntica
Baza de espadas	1932		ruedo	esperpéntica	esperpéntica

Tab. 3: Works of Valle arranged in several competing groups of works

Even if the different scholars use different specific labels (for example "guerra carlitas", "carlista", or "bélica" for the *carlista* novels), they all agree that eleven novels constitute three clear groups: four *sonatas*, three *carlistas*

[15] The first two as book, the last one only in magazines.

and three *El ruedo ibérico*. The remaining three are classified in different ways:

- *Flor de santidad* is classified by García Nora together with the *sonatas*, while the other two form an ad hoc class[16]
- *La media noche* is treated by de Juan Bolufer together with the *carlista* novels;[17] García Nora does not analyze this text and Pedraza-Rodriguez[18] use another ad hoc class
- Pedraza-Rodriguez assign an ad hoc class for *Tirano Banderas*, while the other two group it within the esperpentic group.

4.1. Evaluation and Results

Like in the previous section, this problem can be analyzed as a classification task in which the undisputed texts are placed in a train corpus and the disputed texts are placed in a test set. For this, I used the exact same methodology that I explained for the previous case. Now the cross validation could only be run three times (because the *carlista* novels only had three novels). Again, the accuracy scores (based on cosine Delta matrix, 5000 MFW, KNN), for the evaluation in the train corpus achieve a perfect value of 1.0 every time.

The prediction of the corresponding group for the three disputed texts is as follow:

Title	Year	KNN predicted Group of Works
Flor de santidad	1904	carlista
La media noche	1917	carlista
Tirano Banderas	1926	esperpéntica

Table 4: Predicted groups of disputed text

The results of the prediction concur with Nora and Bolufer to classify *Tirano Banderas* as an esperpentic novel and also with Bolufer in considering *La media noche* along with the *carlista* novels. The new nuance that the prediction provides is the classification of *Flor de santidad* in this same group of the *carlista* novels as well, an unexpected association.

[16] Eugenio García de Nora, *La novela española contemporánea* (Madrid: Ed. Gredos,1963).

[17] Amparo de Juan Bolufer, *La técnica narrativa en Valle-Inclán* (Santiago de Compostela: Universidade de Santiago de Compostela, 2000).

[18] Pedraza /Rodríguez, *Manual de literatura española*.

Let us reflect on this new group of novels by Valle that Delta and the classification method have revealed: a group of works published between 1904 and 1917 and therefore with a start date before the *sonatas* are finished (published between 1902 and 1905). This group of works has a thematic and chronological core with the three novels of the *carlista* war, another war novel (about World War I) and a sui generis novel (*Flor de santidad*) in a timeless Galicia about a female character who suffers hallucinations and has sexual relationships with a man who claims to be Jesus Christ. The relation between the three *carlista* novels is very clear, and also the closeness to the other war-novel is predictable. But how do we fit in *Flor de santidad* with the rest of these novels? Although very different, *Flor de santidad* does share some literary phenomena with these novels: its story shares settings with the *carlista Los cruzados de la Causa*. It can be argued that the action of *Sonata de invierno* also takes place in a rural Galicia, one of the reason that makes Nora group the disputed novel together with the *sonatas*. But other aspects do differentiate *Flor de santidad* from the *sonatas*: they have an autodiegetic narrator and share an upper-class protagonist: Marqués de Bradomín. In contrast, all the novels of this new group have a heterodiegetic narrator and the conditions of lower-class characters are portrayed in detail. That is why I am proposing to use the following label (in Spanish) for this group: *social-bélico*.

In addition to the already existing three hypotheses from literary scholars, I am proposing a new way of classifying Valle's work based on the results of the Delta matrix and the prediction of the algorithm. Now, the different hypothesis (the already existing ones and the new one) can be evaluated and compared to the results of the accuracy scores of a cross validation (4 times).

This shows that Pedraza/Rodríguez, Bolufer and Calvo Tello have a median of accuracy of 1.0. Bolufer and Pedraza/Rodríguez mismatch some texts in the cross validation evaluation, while the new hypothesis with the *social-bélico* group is correctly classified every time. From the already existing way of classifying the groups, Bolufer achieves the best results, an outcome that was easily predictable, since she is the only expert on the specific field of Valle (the other two address bigger collections of texts).

Given the possible existence of other ways of classification proposed by other literary scholars, which have not been considered, here I would like to encourage scholars to report any missing information through the issues of GitHub and I will update the results in the newest versions of the Jupyter notebook cited before. Likewise, if new digital editions of the texts with bet-

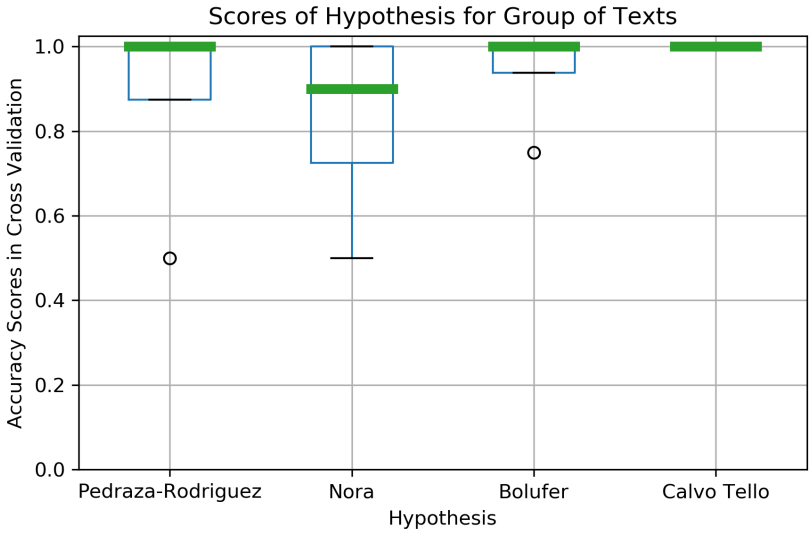

Figure 3: Accuracy scores in cross validation (KNN) of different hypotheses

ter quality or a more precise philological control are published, for example with the publication of the archive GIVIUS[19] the analysis can be run again.

If we go back to the cluster, now we can represent the color of the subclusters with the labels from the classification, as it is shown in the figure 4.

5. Conclusion

In this article, I have analyzed two specific questions regarding the corpus of Valle-Inclán's novels: first, what is the internal periodization of Valle's novels? And second, what is the internal classification within groups of works? I have applied and evaluated different hypotheses proposed by literary scholars. To answer these questions, I have used Delta distances from the novels as features in a classification task, evaluating whether this methodology sorts the undisputed texts correctly. For both cases the algorithm achieved perfect results.

The results show that 1920 is the frontier between the first and the second period of Valle's novels.

[19] Margarita Santos Zas and Carmen Elena Vílchez Ruiz, "Valle-Inclán en red: el Archivo Digital del GIVIUS," in *Sociedad, políticas, saberes*, (Málaga: HDH, 2017) 85–87, http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.

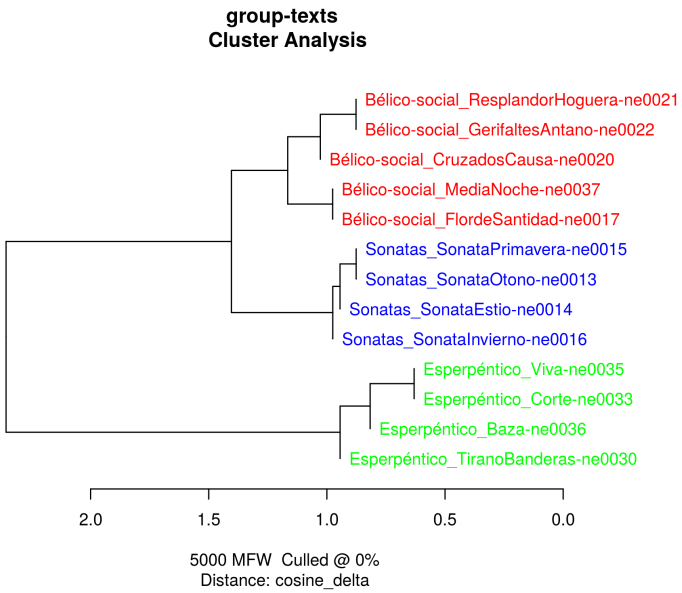

Figure 4: Dendrogram of Valle's Novels with groups from the classification

The predictions of the algorithm for the three disputed cases of group of works match in two cases with hypotheses from literary scholars and also establish an unexpected and new twist: the fact that *Flor de santidad* belongs to the war novels group. With these results, I am proposing a new classification for Valle's novels with only three groups: *sonatas* (four novels), *bélico-sociales* (five novels) and *esperpénticas* (four novels). These groups not only share lexical features (captured in the form of Delta distances), but also literary phenomena and chronological closeness.

This article has only analyzed Valle's novels, although his oeuvre is more extensive and contains different genres. Other scholars are currently working on the digitization of Valle's works,[20] so I hope that in the near future we will be able to reanalyze these questions with more data that will make our work even more solid. In any case, in this article I am proposing a way

[20] Santos Zas, "Valle-Inclán en red"; Concepción María Jiménez, Elena Martínez Carro, María Teresa Santa María, José Calvo Tello, María Simón Parra, Roxana Beatriz Martínez Nieto and María García Sánchez, "BETTE: Biblioteca Electrónica Textual Del Teatro En Español de La Edad de Plata" in *Sociedad, políticas, saberes*, (Málaga: HDH, 2017) 88–91, http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.

of mutual evaluation between traditional literary studies and digital humanities: evaluation of digital methods through solid literary knowledge, on the one hand, but also evaluations of diverging literary hypotheses through solid digital methods, on the other.