Songs of the Self: Using Stylometry to Explore the ‘Pessoa-Whitman Connection’

Digital Humanities: Tools and Methods

Group Project

Henry Hornung (S4156145)

Gheorghe Septelici (S3461912)

Aidan Grefte (S6071597)

Introduction

Portuguese poet Fernando Pessoa (1888 - 1935) is known to be deeply influenced by the works of Walt Whitman. Pessoa himself made no secret of his infatuation with the American's work, naming him as an influence on multiple occasions, and even going so far as to write a salutation to him under the heteronym Alvaro de Campos (Zenith, 2022). It is widely accepted that Pessoa's discovery of Whitman's poetry closely coincided with the beginning of his heteronymous system, somewhere between 1913 and 1916, leading scholars to come to the conclusion that Pessoa was probably inspired to do so by Whitman's work (Brown, 1991; Silva, n.d.). Brown in particular claims that "what in the American poet appealed to Pessoa was his ability to not only affirm the dual nature of the self but to actually generate from that duality promise of the selfs infinitude" (1991, p.4). Thus, in the vein of Whitman's famous statement "I am large – I contain multitudes" (2009, p.103), Pessoa's three most distinct and prolific heteronyms were born: the modernist naval engineer Alvaro de Campos, the Horatian classicist and doctor Ricardo Reis and the non-literate villager Alberto Caeiro – whom Pessoa and the other heteronyms considered their master.

Notably, in this expression of multitudes Pessoa also 'unlocked' the ability of manipulating style. In a recent paper, Skorinkin and Orekhov (2023) demonstrate that the stylistic signature of each heteronym was strong enough to confuse Delta – the most prominent stylometric feature used in computational author-attribution tasks. Their research manages to capture the efficacy of Pessoa's heteronymous project advancing the idea that scholars have to be wary of certain authors' capacity for obfuscating stylistic 'fingerprints' when performing stylometric experiments of authorship attribution. The present analysis builds upon Skorinkin and Orekhov's experiment, albeit altering some of its conditions, in order to investigate if the introduction of a test author that had a known influence on Pessoa – such as Walt Whitman – presents different results in a stylometric analysis. Having established an

eminent connection between Whitman and Pessoa, it can be hypothesized that the former's poetic style will have had an influence on the latter. This study aims to answer to what extent computational methods, namely stylometry, can be used to describe the Pessoa-Whitman connection described in detail by scholars like Brown, thus reinvigorating a humanistic inquiry with the help of contemporary digital humanities tools. This question will be answered in two components. The first component will address the style of Pessoa and his heteronyms in comparison to Whitman to determine whether there are differences in stylistic closeness between Whitman and the heteronyms. The second component will focus on the stylistic comparison between Pessoa's and Whitman's work, as well as the works of two authors selected for a distractor corpus; H.P. Lovecraft and Wilfred Owens. In doing so, a discussion will be opened on the effectiveness of using stylometry to elucidate relations of influence between authors in general. Similarly to Skorinkin and Orekhov, the Stylo package in R will be used in the current inquiry (Eder et. al., 2016). The differences in our corpora and methodology will be expanded upon in the sections that will follow.

Corpus

The corpora used in this study were collected from various sources. The main corpus of Pessoa and his three heteronyms, Alberto Caeiro, Ricardo Reis and Alvaro de Campos, came from the book "A Little Larger Than the Universe", a collection of Pessoa's poems, curated and translated by Richard Zenith (2006). This corpus was initially gathered via scraping from a PDF version of the book using the Python module PyPDF, and various other python modules to clean and format it. The remaining poems by Pessoa; 27 English works and 35 sonnets, were sourced from the 'English Poems' collection on Project Gutenberg and cleaned using Python regular expressions (Septelici, 2025). For the corpus of Whitman, a premade one was found on GitHub, made by JohnmBerger (Berger, 2018), which consisted of a cleaned

.txt file from his book 'Leaves of Grass' which contains 389 poems. Finally, a distractor corpus was selected, including 25 poems by British World War I poet, Wilfred Owen and the American author, H.P. Lovecraft. These were manually copied from "Poems by Wilfred Owen" (Light et al., 2013) via Project Gutenberg (Grefte, 2025), and scraped from the H.P. Lovecraft Archive (2021) using the ScraPy and BeautifulSoup modules in Python respectively (Hornung, 2025).

These corpora were chosen for a variety of reasons. The book on Pessoa and his heteronyms provided a robust corpus of poems for all four personalities, and was used as the basis for most tests in this study. A separate set of Pessoa's poems, originally written in English, was used to test for any translator bias in the stylistic differences between his works. Whitman was included on account of his known influence on Pessoa and his heteronyms. Skorinkin and Orekhov used some contemporaries of Pessoa, Portuguese and Brazilian, in their analysis (Skorinkin and Orekhov, 2023, p.1252). While a reasonable choice, this study aimed to test if a known personal influence of Pessoa such as Whitman, whom he read and annotated fervently (Brown, 1991, p.2), could challenge Skorinkin and Orekhov's findings. Finally, the distractor corpus was selected primarily out of convenience, as it had already been collected by the researchers for previous projects. The purpose of including a distractor corpus was to enable an analysis of Whitman's style against the style of authors with whom he had no immediate connections, and compare this with an analysis of Whitman's style against Pessoa's. Its additional purpose was to test if the distance between Whitman, Pessoa or his heteronyms increases or decreases with the introduction of these unrelated poets.

In this dataset, there was a relatively large difference in corpus size for the works of Pessoa and his heteronyms, and especially large when compared to Whitman's corpus of roughly 120,000 words. To address this, the larger corpora were split into smaller pieces. Pessoa's work has been split between his orthonymous translated works and his self-published English works, yielding two corpora of approximately 10,000 words each. Campos' corpus was split in about 8,000 words each, and the Whitman

corpus into a set of six chunks of roughly 20,000 words each. After the splitting, our corpora were similar in size to those on which Skorinkin and Orekhov performed their analysis and more in line with Lopez-Escobedo et al. previously-mentioned recommendation.

Methodology

The most common use of stylometry is to resolve questions of authorship attribution (Koppel et al., 2009). As there are no questions of authorship with Pessoa's heteronyms, the purpose of this study is rather to explore and evaluate stylometry's uses in other settings. As mentioned previously, this paper uses the 'Stylo' package in R. 'Stylo' integrates methods that allow one to measure the style of an author by the identification of its features of style, created using markers from textual measurements, using statistical methods (Lopez-Escobedo et al, 2013, p.605). This package was used to perform several different analyses and map out the outcomes in various visualisations.

First, the **stylo()** function was used to create consensus trees and cluster analyses using Burrows' Delta, in order to explore whether Walt Whitman, as a known influence of Pessoa and especially of two of his heteronyms - Alvaro de Campos and Alberto Caeiro (Brown, 1991, p.5) - will be clustered together with or nearby any of them.

Next, the **oppose()** function was used to make a contrastive analysis with Craig's Zeta. The basic use of Craig's Zeta analysis with Stylo is to visualize the preferred words by an author in a primary_set and the avoided words when compared with another author in a secondary_set. Since this experiment tests stylometric similarity, the basic use of Zeta renders itself counterproductive. A workaround was designed

where Whitman's corpus was introduced in the test_set, which helps to identify stylistic similarity by way of overlapping the author in the test category with its stylistically closest author in one of the two main sets. This serves as an addition to the analysis done by Skorinkin and Orekhov, who did not use this feature in their investigation.

Given that the aim of the current experiment was to test potential similarity between Whitman and Pessoa or Whitman and one of the two heteronyms known to be influenced by him, Alvaro de Campos and Alberto Caeiro, Ricardo Reis' corpus was excluded from this part of the analysis. Furthermore, the small corpus size of Reis' might have interfered with the stylometric analysis since in stylometry, Lopez-Escobedo et al. claim, shorter texts are less able to form clear clusters for the authors in comparison to longer texts, which might yield misleading results (2013, p.609).

The consensus tree outputs a statistically decided compromise between a number of analysed cluster results, for this paper, using the most frequent words (MFW). In comparison, the cluster analysis produces a dendrogram with the analysed text in a hierarchical format while also using MFW (Eder et al., 2018, p.15). The final visualisation that was used was the contrastive analysis, which generates an analysis of words into a graph format. This graph plots the frequencies of "both word categories, preferred and avoided, for each sample into which texts have been sliced" (p.27). The primary set is presented as circles, having high words preferred frequencies and low words avoided frequencies, and the secondary set as triangles, with low words preferred frequencies, and high words avoided frequencies. The test set's markers are displayed as crosses. The overlap with either the primary or secondary sets suggests stylistic similarity. All three of these methods are essential for this experiment as they measure ranges of stylistic similarity, with Craig's Zeta being especially useful as it depicts exact data points that overlap between different works.

Analysis

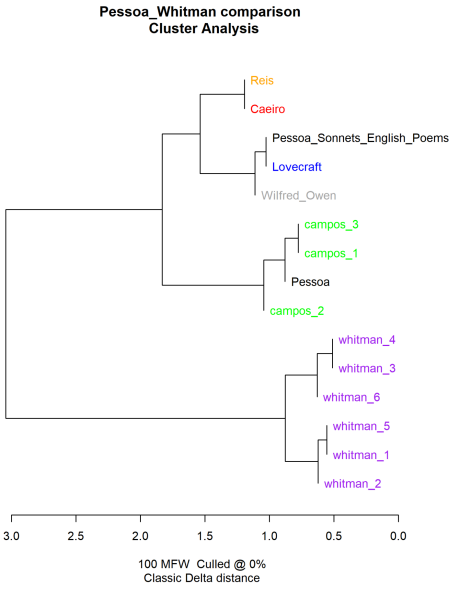

Figure 1

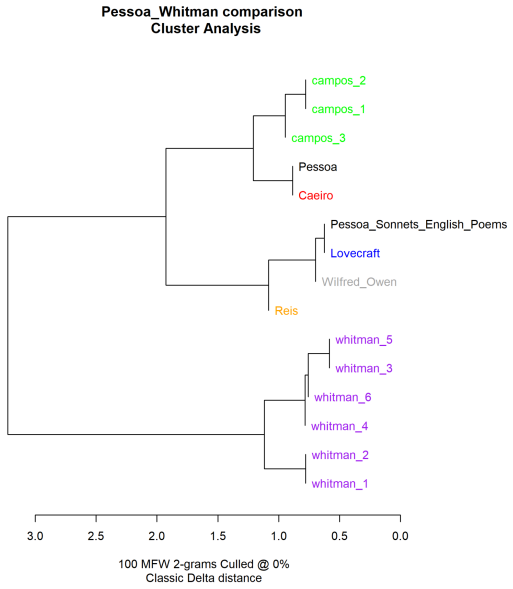

Figure 2

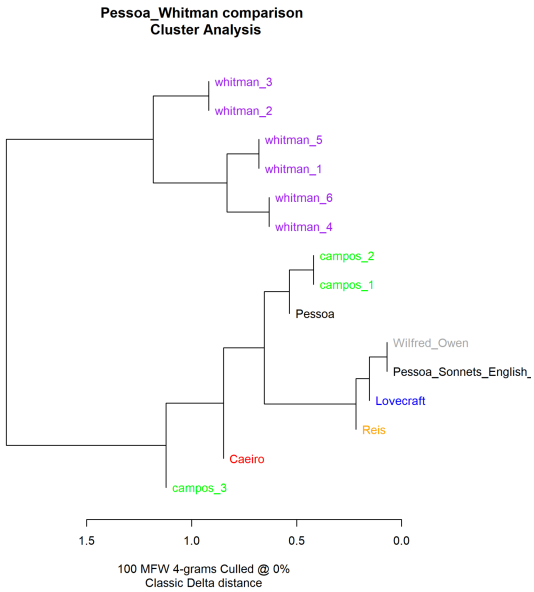

Figure 3

As stated, a cluster analysis produces a dendrogram in a hierarchical format for the 100 MFW. Above, there are 3 Figures. Figure 1 uses a single n-gram, meaning each word is counted separately,

Figure 2 uses a bi-gram, meaning pairs of words are counted, and Figure 3 uses a 4-gram for visualisation, meaning that words are counted in sets of four.

Figure 1 represents a visualization based on the most frequent individual words. This can offer a broad idea of the stylistic similarities found in the text, clear evidence of this fact being the grouping of all of the Whitman texts together. Interestingly, in the grouping of Pessoa with Campos, which is to be expected, it also contains the works of Lovecraft and Owen within the same branch. Although these are more closely linked with the English poems that Pessoa wrote. One explanation for this could be that they are grouped by the period that they published in, Pessoa, Lovecraft and Owen were all contemporaries of each other as early 20th-century poets. Thus, because Whitman was a 19th-century poet, his style is noticeably different and so there is time-related stylistic variation (Reeve, 2018; Rebora and Salgaro, 2018; Stamou, 2008). On the other hand, these results could raise questions about the translator or editor's 'fingerprint'. Since the other works of Pessoa and his heteronyms in our corpus are translated into English by notorious Pessoa translator Richard Zenith, it could be that some of his own stylistic preferences have slipped into the translations. Moreover, Zenith claims that certain editorial decisions that impacted the original work had to be made when collecting and translating the poems, due to them being largely unpublished and in the format of manuscripts with multiple notes, annotations and alternative lines on the side (Zenith, 2006, p.xli). This editorial work might have influenced the distances between the translated work and the other poems in our corpora.

In the bi-gram cluster analysis, Figure 2 presents broadly the same results, although Caeiro is now placed more closely to his fellow heteronyms. The bi-gram analysis works by analysing words in pairs. This allows for more word pairing that may be indicative of shared expressions and context. Single n-grams and bi-grams are considered to be the best source of authorial markers, as these tend to be made up of the most commonly recurring language (Antonia et al, 2014). Furthermore, within these cluster analyses, Lovecraft and Owen are repeatedly placed with Pesso_Sonnets_English_Poems, the works that

Pessoa wrote himself originally in English. This may further suggest an influence on the style of the translated works on the part of Zenith, thus skewing the results as argued by Rybicki (2010; 2013) and Heydel (2013).

Finally, the last cluster analysis was created using a 4-gram system. This may allow for an even broader analysis of patterns and themes, however, it too has limitations. In a corpus, due to the size of the unit, 4-grams and higher tend to make up a comparatively small part of the corpus and so tend to not be useful (Hoover, 2002, p. 162). This limitation is demonstrated within the visualisation as shown by the dynamically changed placement of Campos_3 and Caeiro; these are smaller-sized corpora which may be affected as the n-gram units get expanded. As a result these corpora provide less data for the statistical visualisation and are represented less accurately.

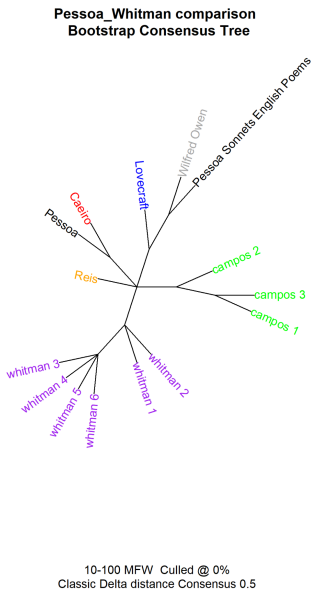

Figure 4

Figure 5

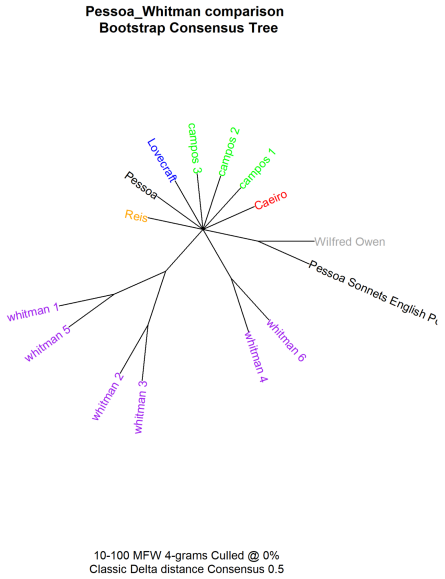

Figure 6

The consensus tree analyses broadly align with the cluster analyses, this is unsurprising as they both are graphically representing similar statistical calculations. When generating the visualisations, the consensus trees were made using the increment setting to allow for an analysis based on the most frequent words as they increased in sets of 10 until it reached the maximum value. This method involves multiple analyses running in a row (Eder et al., 2018, p.13). This creates a statistical compromise between the virtual cluster analyses (p.15), which was then replicated three times using a single n-gram, a bi-gram and a 4-gram.

In Figure 4, the single n-gram table shows that the Whitman papers all statistically align, this is evident from the cluster analyses. This is the same in both Figures 5 and 6. Again, in the single n-gram and the bi-gram the works of Lovecraft and Owen are placed with the English poems of Pessoa. In contrast to this, in Figure 6 the 4-gram consensus tree is obfuscated, likely due to the larger n-gram units applied for this iteration of the experiment. The consistency between the consensus trees and the cluster analyses serve to reinforce the conclusions of stylistic (dis)similarities within the corpus, helping to

increase the validity of this experiment. In line with the experiment carried out by Skorinkin and Orekhov, our results using the stylo() method confirm that Pessoa has indeed developed very strong authorial voices for each of his most prominent heteronyms.

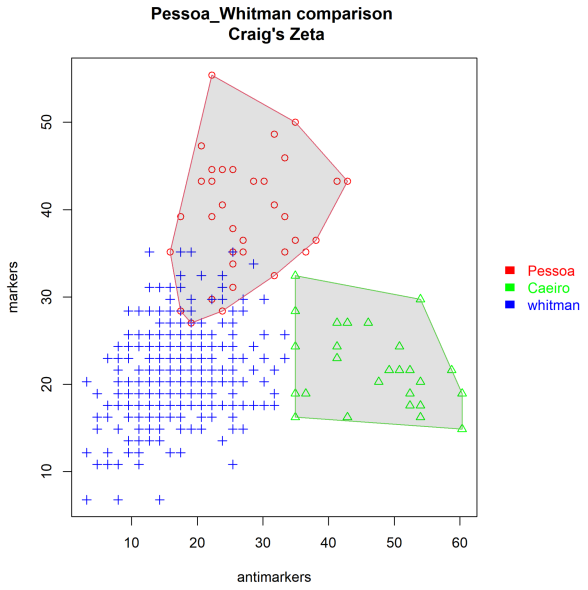

Figure 7

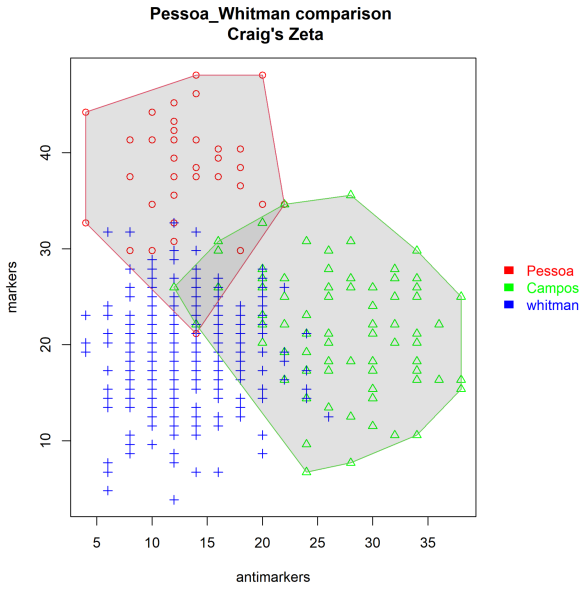

Figure 8

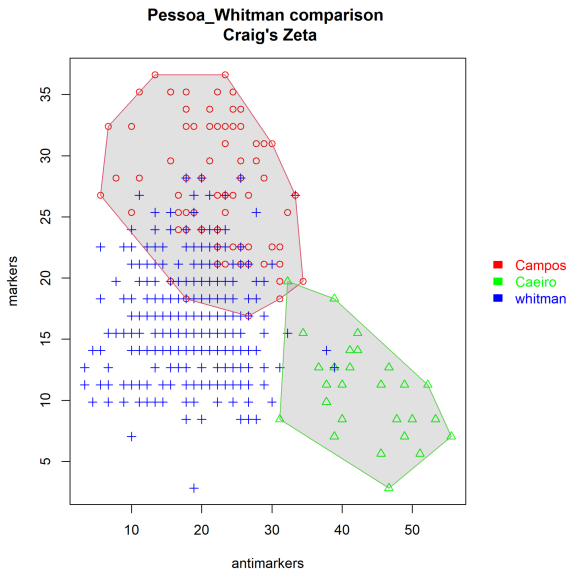

Figure 9

Finally, for this experiment a Craig's Zeta analysis was performed. Given that this method aggregates lists of preferred words by an author in a primary set which are avoided by an author in a

secondary set, and vice versa, the initial obstacle consisted of finding a way to use it in order to test for similarity rather than difference. The 'Stylo' documentation informs that by using a third author as a test set, one can explore overlaps between the test author and the other two authors indicating stylistic similarity (Eder et. al., 2018, p.27). As previously stated, the idea in this part was to test Campos and Caeiro's potential similarity to Whitman, considering the latter's documented influence on the former two heteronyms. Pessoa was also introduced in this part of the experiment, since he is the original authorial voice of the two. Potential stylistic similarities between his work and Whitman's were also tested, and contrasted to the results to the similarity distance between Whitman and the heteronyms. Splitting the corpus was not required for this part of the investigation, so each author or heteronyms' corpus was used in their entirety. Moreover, only Pessoa's translated work was used in order to preserve consistency with the translated works of the other heteronyms.

The first iteration of the experiment used a slice length of 300 and an occurrence of 10 on the corpora of Pessoa, Caeiro and Whitman. This means that the texts have been split in slices of 300 words and only the words with a frequency of 10 and higher were considered. In stylometry, it is claimed that function words such as 'the', 'and', 'as', pronouns etc. are the most accurate markers for stylistic similarity since they are topic and genre independent (Madigan et. al., 2005). A word-frequency of 10 and higher has been chosen in order to exclude unique words and take into account as many of these function words as possible. The results show a mild overlap between Whitman and Pessoa and no overlap with Caeiro. These results indicate similarity between Pessoa and Whitman, albeit on a very low scale. Brown also tells us that Caeiro, who was viewed as Pessoa and the other heteronyms' master, might have appeared as an incarnation of Whitman (Brown, 1991, p.9). It seems that, if one were to entertain the idea of that being true, the incarnation did not manifest itself in stylistic features that can be tested with stylometry. Although we would have to consider the limitation in corpus sizes as a potential factor that would affect these results.

The second iteration of this experiment uses the same slice length and occurrence as previously, this time on the corpora of Pessoa, Campos and Whitman. Although in this case, there is overlap between all three, indicating a level of stylistic similarity between each author, the higher similarity is between Whitman and Campos. One explanation for this may be that Campos' long-form poetry was much more similar to Whitman's own style of writing, which also made the size of their corpora larger. The other heteronyms and Pessoa used medium and short-form poems more commonly in their works.

Lastly, the same settings were used for the third iteration of the analysis, this time on Campos, Caeiro and Whitman's corpora. Here the results are more extreme, with a high overlap between Campos and Whitman and almost no overlap between Caeiro and Whitman. This could consolidate the idea that it was in the long-form structure of their poems that Campos and Whitman were most similar. Given that the cluster analyses separated each author or heteronym perfectly, it is curious that an overlap finally happens in this part of the experiment. Further analyses and more iterations of this experiment with different settings could elucidate better the traces that Whitman's art might have left in Pessoa's own work. Moreover, even though this does not contest the fact that Pessoa managed to create very distinctive authorial voices through his heteronyms, as Skorinkin and Orekhov demonstrate, it does show that methods like oppose() and different test corpora, like Whitman's, can open up the discussion again and invite further investigation on the factors that yielded those results. Moreover, it generates further discussion on the efficacy of stylometry as a method of testing authorial similarity.

Reflection on Stylometry

Stylometry excels at one discipline: highlighting textual similarity or dissimilarity. In order to properly evaluate the role of stylometry in describing the Pessoa-Whitman connection, or other relations of textual influence, two more contrastive analysis figures must be considered, comparing Pessoa, Whitman, and the distractor authors. Figure 10 shows a contrastive chart between the heteronym Alvaro de Campos, who was chosen because his are the most stylistically similar works to Whitman from Pessoa's oeuvre, and Wilfred Owen, with the test set consisting of the Whitman corpus. Figure 11 substitutes Owen for Lovecraft. As with the previous uses of the oppose() function, the slice length is set to 300, and the occurrence threshold is at 10.

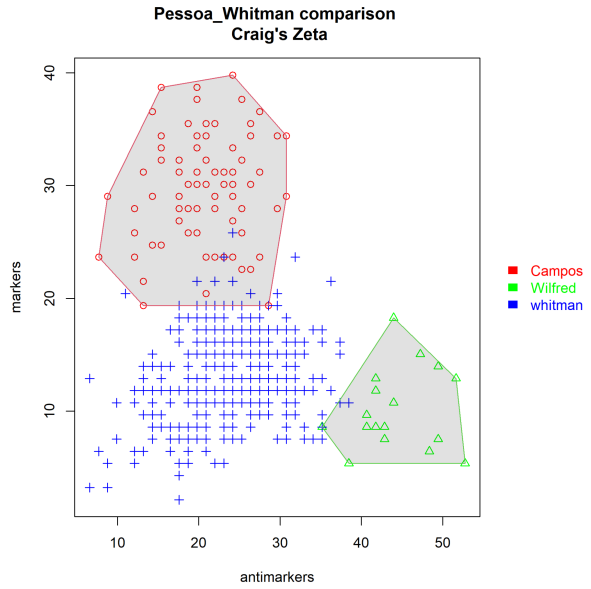

Figure 10

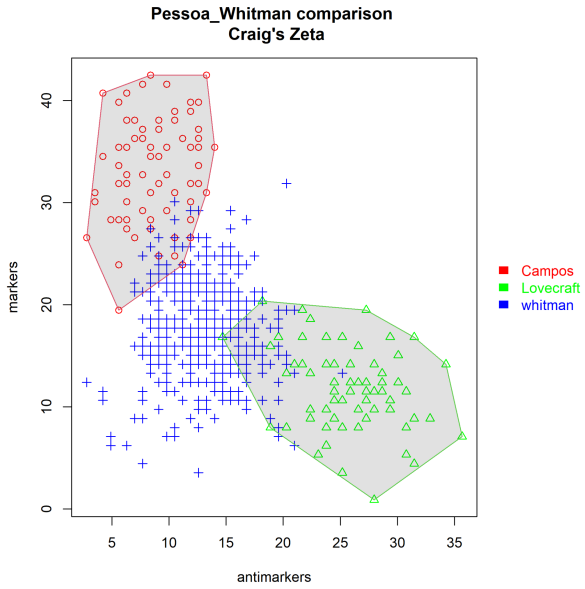

Figure 11

When looking at the points on Figure 10 where Whitman's corpus overlaps with the other two stylistically, there is only slightly more overlap points between Whitman and Campos than there are between Whitman and Owen, meaning that Whitman's style is only tentatively more similar to Campos

than Owen, which is surprising given the strong connection between Pessoa and Whitman. Another interesting result can be seen in the contrastive analysis comparing Whitman to Campos and Lovecraft (Figure 11). In this graph, it appears that Whitman's markers overlap more with those of Lovecraft than Campos, suggesting that the former's style resembles Whitman's more closely than the latter's. This is especially compelling when considering that Lovecraft was likely not particularly fond of Whitman's work, once describing it as "bathetic" (Pérez-Campos, 2010).

As shown earlier, stylometry can be used to an extent to delineate style as a facet of the Whitman-Pessoa connection when considering differences between the various heteronyms. However, the two Figures discussed in this section indicate that there is more to be discussed. There is little to suggest that Whitman is stylistically closer to Campos than to the distractor authors. While it could be the case that Pessoa deliberately altered his language to be more modern than Whitman's, or that Whitman's style resembles Lovecraft's due to them both being American, there is no evidence to substantiate these hypotheses and more research is needed. What is clear, is that at least in this case, literary influence transcends the purely quantitative value of stylometry. Without the known context of Pessoa being influenced by Whitman, the methods applied to the corpora would not clearly show that one had an influence on the other based on the provided graphs, thus, indicating that stylometry does not convey literary influence very well. This study only investigates a single case, however, and further examples of literary influence should be investigated to corroborate this study's findings. In summary, though stylometry can measure style quantitatively, influence is a more qualitative variable, meaning that stylometry can only partially capture it, at best. Thus, stylometry on its own may not be a suitable method for researching literary influence without other methods to complement it.

Conclusion

This study aimed to investigate to what extent computational methods could be used to describe the Pessoa-Whitman connection. Through the conducting of various stylometric tests, it was found that Whitman's writing style was generally distinct from the style of both Pessoa and his heteronyms, and the style of the distractor authors. Of the heteronyms included in the corpus however, Whitman was discovered to be stylistically closest to Campos, likely due to his long-form poetry being similar to that of Whitman. Moreover, by analyzing where Whitman's style was placed on contrastive graphs between Caeiro and the distractor authors, it was determined that, without prior knowledge of the Pessoa-Whitman connection, there wouldn't have been any way to recognize that this connection exists. This called into question the efficacy of stylometry as a method studying literary influence.

Having already discussed several limitations with this study's design and the methods used, there are two more to address for the sake of completeness. Firstly, the majority of the poems by Pessoa used in this study are translated. Each text was translated by Richard Zenith, who is considered an expert on Pessoa's work, ensuring internal consistency within this corpus. All of Caeiro's, Campos' and Reis' works are originally in Portuguese, meaning that it is impossible to avoid using translations, unless one uses the Portuguese language model in Stylo and corpora of Portuguese-speaking authors, as Skorinkin and Orekhov did. Since this project uses Whitman's work written originally in English, the experiments had to be performed on the translated works. Thus, the possibility of the translator's input skewing the results exists, and must therefore be mentioned. Another limitation to consider is that the distractor corpus consisted of two authors who were both Pessoa's contemporaries. This means that it is possible that they use similar language to Pessoa as a result of them being active around the same time, while Whitman, having lived earlier, may employ an older language style, thereby influencing the results of a stylometric analysis. Finally, stylometry is sensitive to genres, and for different genres, the parameters of stylometry

may be varied. As of the writing of this paper, there is still a lack of consensus on which features are most suitable for a stylometric analysis of poetry, thus stifling this study's results' general comparability to other research.

References

Antonia, A., Craig, H., & Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. Literary and Linguistic Computing, 29(2), 147–163.

Berger, J. (2018). Walt-whitbot: A markov bot that tweets in the style of Walt Whitman. GitHub. https://github.com/johnmberger/walt-whitbot

Brown, S. M. (1991). The Whitman-Pessoa Connection. Walt Whitman Quarterly Review, 9(1).

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Eder, M., Rybicki, J., & Kestemont, M. (2017). 'Stylo': a package for stylometric analyses. Computational Stylistic Group, 20.

Grefte, A., (2025). aidan_assignment_2_resubmission. GitHub. https://github.com/Aiclan/aidan_assignment_2_resubmission.git

Hoover, D. L. (2012, July). The Rarer They Are, the More There Are, the Less They Matter. In DH

Hornung, H., (2025). Collecting-Data-Assignment-3. GitHub. https://github.com/HenryAHornung/Collecting-Data-Assignment-3.git

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1), 9–26. https://doi.org/10.1002/asi.20961

Light, A., Johnson, M., Widger, D., 2013, August 19). Poems by Wilfred Owen. Project Gutenberg. https://www.gutenberg.org/ebooks/1034

López-Escobedo, F., Méndez-Cruz, C. F., Sierra, G., & Solórzano-Soto, J. (2013). Analysis of stylometric variables in long and short texts. Procedia-Social and Behavioral Sciences, 95, 604-611.

Lovecraft, H. P. (2021, January 1). Electronic texts of H.P. Lovecraft's works. The H.P. Lovecraft Archive. https://www.hplovecraft.com/writings/texts/

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*.

Pérez-Campos, M. (2010). Lovecraft's "The Bride of the Sea" and the Uses of Bathos. Lovecraft Annual, 4, 3–30.

Septelici, G., (2025). Resubmitted_Pessoa_Corpus_Analysis_SpaCy. GitHub. https://github.com/gheorgheseptelici/Resubmitted_Pessoa_Corpus_Analysis_SpaCy

Silva, R. (n.d.). The Machine and the Garden: Walt Whitman and Fernando Pessoa's Álvaro de Campos. The Mickle Street Review. http://msr-archives.rutgers.edu/archives/Issue%2015/essays/Silva.htm#_ednref1

Skorinkin, D., & Orekhov, B. (2023). Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta. Digital Scholarship in the Humanities, 38(3), 1247–1266. https://doi.org/10.1093/llc/fqad012

Whitman, W. (2009). Leaves of grass, 1860 : the 150th anniversary facsimile edition. http://ci.nii.ac.jp/ncid/BB14147846

Zenith, R. (2006). A Little Larger than the Entire Universe. Penguin Press.

Zenith, R. (2022). Pessoa: An Experimental Life. Penguin Press.