Stylometric Investigation of Dante's Divina Commedia by Means of Multivariate Data Analysis Techniques

Edoardo Saccenti[1], Leonardo Tenori[2]
[1]Biosystems Data Analysis Group
Swammerdam Institute for Life Science
University of Amsterdam
Science Park 904, 1098 XH, Amsterdam
The Netherlands
[2]FiorGen Foundation
via Luigi Sacconi 6
50019, Sesto Fiorentino
Florence, Italy
e.saccenti@uva.nl, tenori@cerm.unifi.it

ABSTRACT: *We analyzed the three parts of Dante's Divina Commedia using word frequencies as style markers for statistical analysis. Partial least square discriminant analysis was used to provide separation among the cantos of the three parts of the Comedy, Inferno, Purgatorio and Paradiso; the statistical models were able to successfully discriminate the poetic tones used by Dante to characterize and diversify the three parts, demonstrating the existence of a particular stylistic substratum in each cantica of the Commedia.*

Keywords: Dante, Divina Commedia, Multivariate Analysis, Stylometry, Partial Least Square Discriminant Analysis

Received: 11 January 2012, Revised 1 March 2012, Accepted 8 March 2012

1. Introduction

Style is an integral part of natural language in written, spoken or machine generated forms. Humans have been dealing with style in language since the beginnings of language itself, but computers and machine processes have only recently begun to process natural language styles [29].

Stylometry is the application of the study of linguistic style (stylistics), and it has legal as well as academic and literary applications, ranging from the question of the authorship of anonymous or disputed documents to the plagiarism discovery in forensic linguistics. An embryo of stylometric application was the Lorenzo Valla's 1439 proof that the Donation of Constantine was a falsification: using solid philological argument based on the analysis of the Latin used in authentic 4th Century documents he demonstrated that the vernacular style of the Donation dated conclusively to a later era (8th Century).

Advanced statistical techniques have been fruitfully applied to stylometry investigations for the discovery of author identity [44], gender [5], native language [31], and even whether an author has dementia [41].

In the present paper we aimed to demonstrate the efficacy of the word frequency based statistical learning approach to enlighten the stylistic differences that characterize the three parts of the Dante's *Divina Comedia*.

The paper is organized as it follows. The Background section 2.1 offers an introduction to Dante's *Comedy*: the overall structure of the poem is illustrated together with a summary of relevant facts and figures. A short discussion about the use of the vernacular language and Dante's style is presented to help the reader to frame the present computational approach to the Commedia in the broader context of Dante's studies. Background section 2.2 is dedicated to a brief overview of the literature dedicate to computational linguistics to the Comedy and the use of multivariate analysis in this field. Background sections 2.3 and 2.4 set a rationale for the choice of word frequency as stylistic marker for the analysis of stylistic substrata of the three parts of the *Comedy* and the use of multivariate analysis. The Material and methods section gives details about text mining strategy, statistical tools applied, with a focus on word frequency and with a concise illustration of partial least square discrimination.

Section 4 present the results of the statistical analysis while Section 5 offers a detailed discussion of the latter with a specific focus on the Purgatory. Finally, criticalities and perspectives are outlined in the Discussion section.

2. Background

2.1 Dante's *Divina Commedia*

Dante Alighieri (Florence, 1265 - Ravenna, 1321) attended the composition of the *Commedia* from 1306 until his death. The poem was lately renamed *Divina Commedia* (Divine Comedy) by the Tuscan poet Giovanni Boccaccio (1313-1375). The poem is divided in three parts (*cantiche*): *Inferno* (Hell), *Purgatorio* (Purgatory) and *Paradiso* (Paradise). The three parts consist of 34 (*Inferno*) and 33 (*Purgatorio* and *Paradiso*) single poems (*canti*). See Table 1 for a glossary of Italian terms.

Italian	English
Infernou	Hell
Purgatorio	Purgatory
Paradiso	Paradise, Heaven
Cantica / Cantiche	Cantica / Canticas
Canto / Canti	Canto / Cantos
Terzina	Tercets

Table 1. Glossary of Italian-English terms

In the *Commedia* Dante tells his journey in the afterlife realms, from the moment he gets lost in the dark wood on the threshold of the Hell to the glorification of the sight of God in last *canto* of the Paradise. Dante's journey starts on the night of Good Friday of April 7 of 1300 and ends on Easter Wednesday April 13, 1300. Dante is guided through the Hell and the Purgatory by the Latin poet Virgil and through the Paradise by Saint Bernard and Beatrice. Figure 1 presents a graphical illustration of the structure of Dante's afterlife.

Figure 1 Scheme of the three realms of the afterlife according to Dante's narration (from Michelangelo Cactani, "*La Materia della Divina Commedia*", 1855). Dante's Hell is a hollow cone dig by Satan at the moment of its fall from the Heaven and it is situated below the city of Jerusalem. The Purgatory is mountain, located at Jerusalem antipodes, formed by the very soil displaced by Satan's fall. The Paradise is outer of the space and organized in ten different heavens [33].

The poem, which counts more than fourteen thousand verses, is written in hendecasyllables, organized in rhymed tercets, with a *terza rima* rhyme scheme: ABA BCB CDC ... A summary of information is given in Table 2.

The *Commedia* is a poetic summary of Dante's human, intellectual, religious, philosophical, scientific and political life. Scholarly study and analysis of the text began straight after the completion of the poem (one of the first was the exegetical comment *Expositio seu Comentum* by Filippo Villani (1325 – 1407)) and never stopped since then. The *Commedia* is regarded as one of the biggest cultural achievement of human civilization.

In the discussed letter to Cangrande della Scala [21], Dante explains why he termed his poem *comedy* [2]:

Nam si ad materiam respiciamus, a principio horribilis et foetida est, quia Infernus, in fine prospera, desiderabilis et grata,

quia Paradisus; ad modem loquendi, remissus est modus et humilis, quia locutio vulgaris in qua et mulierculae communicant.

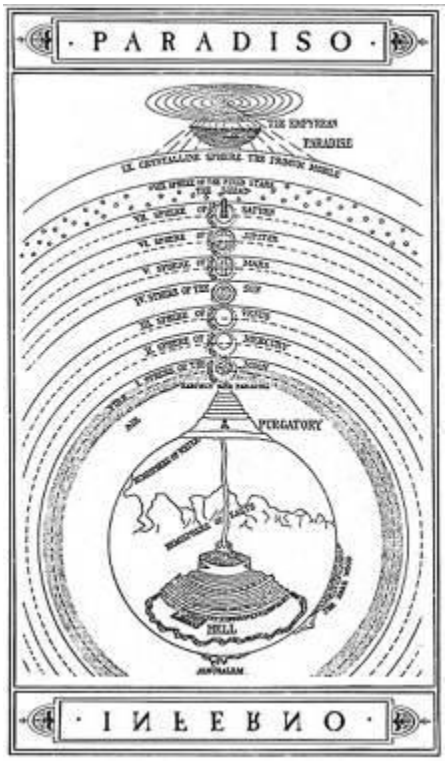

Figure 1.Scheme of the three realms

Cantica	# Cantos	Year	# Verses	Cum# verses
Hell	34	1306-09	4720	4720
Purgatory	33	1308-12	4755	9475
Paradise	33	1315-21	4758	14233

Table 2. *Divina commedia* relevant data summary [33]

The *Commedia* begins in a "*horrible and foul* "manner and ends "*happily*" (the journey from the Hell to the Paradise) in contrast with the tragedy that begins "*admirably and quietly*" and ends in a "*foul and horrible*" manner. The comedy is written in vernacular, using a "*lower and humble*" style, (which Dante, with a medieval lack of political correctness, deems intelligible even by women) while the style of the tragedy, usually written in Latin, is "*elevated and sublime*" [23].

As a matter of fact, Dante does not keep his styles sharply delineated in the *Commedia:* Auerbach, was probably the first to cast the idea of a poem written in a mixed style [23]. Dante's mingling of styles results in a complex plurilingualism [6]. The language of the Commedia encompasses neologisms, technical jargon (from mathematics, physics, theology, philosophy, medicine…), Latin words and invented words (*Pape Satàn, pape Satàn aleppe!*). Nevertheless, Dante succeed in preserving the poetic and structural unity of the poem when there are at least three major contents blocks (Hell, Purgatory and Paradise): these blocks are set off stylistically through the creation of three distinct poetic tones [23]. Each *cantica* is dominated by a particular style which depends on the content and that defines the peculiar color and sound of the three afterlife realms. Using a musical metaphor, each *cantica* is written in a different tonality but Dante allows modulation within each *cantica* [23].

2.2 Computational linguistic approaches to Dante's *Comedy*

Computational linguistic studies have addressed a variety of problems and many tools have been developed or applied to the analysis of different authors and works. The works of William Shakespeare [37, 40, 39], the Gospel [17, 34, 38],

the Bible [4, 27], The Book of Mormon [22, 47], Columbus' diary [25], the Federalist papers [27] have been subjects of investigation. At our surprise it came that such a fundamental text of western literature like Dante's Comedy has received relatively little attention: it has been used in a variety of studies [11, 12, 28, 35] but, at the best of our knowledge, it has never been object of a systematic and dedicated investigation. Statistics of word frequency and rhymes occurrence have been compiled during the time as a natural complement and aid for scholarly studies [9, 18, 56], but an investigation using modern computational approaches focused solely on the *Divina Commedia* seems to be lacking; this paper may be the first attempt initiate a new line of research on Dante and his work.

2.3 A rationale for text analysis

Analyzing the frequency of words in a text is a well-established method for text analysis which dates back to the seminal paper of Luhn [36]. Word frequency analysis has been widely used to look for similarities/dissimilarities in texts of groups of text [30], to detect text genres [49] or for automatic authorship attribution [50]. In this paper we investigate Dante's style in the *Divina Commedia* by means of a multivariate analysis of word frequencies. The analysis of word frequencies to investigate Dante's writing style is somehow justified by Dante himself. In *De vulgari eloquentia (On Eloquence in the vernacular)*, his theoretical treatise on the use poetical language [3] written in 1302-1305, Dante addresses the problem of selecting words appropriate to the style one has choose: the poetic style translates into the use of certain words or categories of words in respect to others [54]. This approach fully justifies the use of word frequencies for the analysis of Dante's style. In this paper we aim to disentangles three distinct poetic tones with which the Divina Commedia is written by means of multivariate statistical analysis of the text to highlight the existence of a distinctive stylistic substratum in each *cantica* of the *Commedia* [23]. Some scholars have already incidentally noted that the frequency with which different words or category of words appear in the *Commedia* seems to correlate with the dominant style of each *cantica* [7]. At the best of our knowledge this aspect has never been treated with a rigorous and state of the art statistical approach. In this paper we focus on the use of word frequencies in combination with multivariate analysis.

2.4 A multivariate stylometric approach to Dante's Comedy

Multivariate analysis is the art of exploring, reducing and analyzing complex data sets to extract significant features. To do this, the text needs to be reduced to low level feature, *i.e.* to entities that can be mathematically manipulated for statistical analysis. In stylometric analysis, data mining is routinely used to extract measurable text attributes such as syllables counts, word types, word length or word frequencies which are referred as style markers [20, 43].

Using multivariate analysis techniques, we aim to address two main research questions: *i*) does each of the three *cantiche* of Dante's *Commedia* possess and individual stylistic signature? *ii*) Does this signature allow a statistical discrimination among the poems of *Inferno*, *Purgatorio* and *Paradiso*?

3. Materials and Methods

3.1 Text sources

We used the standard Italian version of the *Commedia* in the *vulgata* by Giorgio Petrocchi [14] retrieved on line at etcweb.princeton.edu/dante/index.html.

3.2 Word frequencies

Weighted word frequencies were calculated by using the *tf–idf* (term frequency–inverse document frequency) approach [26]. The importance of a word in the *Divina Commedia* corpus is measured using the *tf–idf* (term frequency–inverse document frequency) weighting scheme. The *tf–idf* value increases proportionally to the number of times a word appears in the document, but is weighted by the frequency of that word in the analyzed corpus. This helps to control for the fact that some words are generally more common than others. This weight is a statistical measure directly proportional to the number of times a word appears in a given *canto* and inversely proportional to the frequency of the word in the whole *Commedia*. A high weight in *tf–idf* is reached by a high term frequency in the given *canto* and a low document frequency of the term in the whole *Commedia*; the weights hence tend to filter out common terms. The term frequency for a term t in a document d, $tf(t,d)$, is the ratio of the number of times the term occurs in the document (nt) to the total number of words in that document (nw).

$$tf(t,d) = \frac{nt}{nw}$$

The inverse document frequency, *idf* (*t,C*), is a measure of whether the term *t* is common or rare across all documents *d* in a corpus *C*. It is obtained by dividing the total number of documents in the corpus (*nc*) by the number of documents containing the term (*nd*), and then taking the logarithm of that quotient.

$$idf(t,C) = \log \frac{nc}{nw}$$

Finally, the *tf–idf* (*t,d,C*), is calculated as:

$$tf - idf(t,d,C) = tf(t,d) \times idf(t,C)$$

A high weight in *tf– idf* is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out not relevant terms. Since the ratio inside the *idf* 's log function is always greater than 1, the value of *idf* (and *tf– idf*) is greater than 0. As a term appears in more documents then ratio inside the log approaches 1 and making *idf* and *tf– idf* approaching 0.

3.3 Text mining

For any statistical calculation the *R* software was used [24]. *R* is a free open source statistical environment endowed with a full featured scripting language and a graphical system. It is highly customizable tanks to the huge amount of external package, freely available on the net (http://cran.r-project.org/). The R package "*TM* " [19] was used for text mining: the plain text files (.txt notepad) with UTF-8 encoding were scanned to retrieve the *tf– idf* of each word in all the 100 Comedy poems. We excluded words shorter than 4 letters to exclude articles, propositions and other empty non informative complete text. The text mining resulted in a data matrix of size 100 × 371 containing the *tf– idf* of 371 different relevant words for the whole corpus of the 100 Comedy poems. This matrix is the classical representation of the vector space model as proposed by Salton [45] were columns represent the space dimensions and the rows are objects laying in that space.

3.4 Univariate Analysis

To assess which words frequencies were significantly different between the *Inferno* and *Paradiso* group of poems, the nonparametric Kruskal-Wallis test [32] was applied together with a Bonferroni correction for multiple testing [10] on a nominal *p*-value of 0.05.

3.5 Partial Least Square Discriminant Analysis

Partial least-squares (PLS) regression is a technique used with data that contain correlated predictor variables. It is routinely used in chemometrics and *omics* disciplines to analyze and extract relevant features from complex data sets [53, 55, 54].

This technique constructs new predictor variables, known as components, as linear combinations of the original predictor variables. PLS constructs these components while considering the observed response values, leading to a parsimonious model with reliable predictive power. PLS finds combinations of the predictors that have a large covariance with the response values. The mathematical formulation is briefly illustrated is as it follows in a simplified fashion.

Let be **X** the independent variables data matrix of size $n \times p$ (*n* observations and *p* predictors) and **Y** the matrix of dependent variables of size $n \times q$ (*n* observations and *q* independent variables) . The goal of PLS regression is to predict **Y** from **X** and to describe their common structure. PLS regression decompose **X** and **Y** the product of a set of common orthogonal factors and a set of specific loadings.

The scores are a linear combination of the original variables. This can be expressed in matrix notation as $\mathbf{T} = \mathbf{XW}$ where **T** is the matrix of the scores ($n \times r$) and **W** is the matrix of the weights ($p \times r$) for the linear combination.

The score are "*few*" and orthogonal, that is $\mathbf{T}^T\mathbf{T} = \mathbf{I}$). The scores are such that the original data matrix **X** can be written as $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$ where **P** is the matrix of the loadings, summaries of **X** such that the residuals **E** are small. The **Y** matrix can be decomposed as $\mathbf{Y} = \mathbf{TC}^T$ where **C** is a matrix of weight ($r \times q$); it follows that **Y** can be approximated as $\mathbf{XWC}^T = \mathbf{XB}$ where **B** are the PLS-regression coefficients. In PLS regression the weights **W** and **C** are found such that the linear combinations of the columns of **X** and **Y** show maximum covariance. We refer the reader to [1, 58, 57] for more details on PLS and associated algorithms for the estimation of the **T**, **P**, **W**, **C**, **B** matrices [57].

Partial least square discriminant analysis [8] (PLS-DA) consists in a standard partial least square regression in which the continuous response vector **Y** is substituted by a vector of discrete class memberships. The aim of PLS-DA is the prediction of group membership from the levels of continuous predictor.

PLS discriminant analysis was performed by using the classical SIMPLS algorithm (de Jong 1993) as implemented in the R-library "*plsgenomics*".

Accuracy for classification was assessed by means of a double cross-validation scheme. The original data set was split in to a training set (90%) and a test set (10%) prior to any step of statistical analysis. Parameter selection (*i.e.* best number of PLS components) was carried out by 7-fold cross validation on the 90% training set [52, 55]. Briefly, part of the cantos was left out during the building of the model. In this way the model has no *prior* knowledge whether the canto that is predicted belongs to a given cantica. This approach allows unbiased prediction results. The whole procedure was repeated 100 times inside a Monte Carlo cross validation scheme. Accuracy and specificity were calculated as described in [52].

4. Results

4.1 Words frequency in the *Divina Commedia*

We used the text mining tools *R* package TM to extract word frequencies in the Comedy. The frequency of a word was calculated according the *tf– idf* method as detailed in the Material and Methods Section. This choice was dictated by the need of filtering words that are frequently appearing but are not informative in term of information retrieval; this approach provided better results in term of discrimination accuracy in respect to the standard word frequency (see section: Stylistic discrimination of the three *cantiche* of the *Commedia*). A summary of the 20 most frequent words, according to *tf– idf*, for the three *cantiche* is given in Table 3. Figure 2 shows a heat map of the *tf– idf* word frequencies data set.

Rank	Hell	Purgatory	Paradise
1	duca	gent(e)	amor
2	maestro	occhi	luce
3	diss	quando	questo
4	loco	sanza	ciel(o)
5	ancor	esser	mondo
6	elli	miei	lume
7	vidi	Perch(é)	grazia
8	altri	dove	tanto
9	poco	quanto	quest
10	tutto	altro	esser
11	altra	altra	veder
12	avea	diss	Cristo
13	terra	fuor	quanto
14	quando	questa	occhi
15	gran	eran(o)	quella
16	questi	tosto	prima
17	quella	terra	donna
18	fuor	tutti	qual
19	altro	qual	perch(é)
20	quei	poco	questa

Table 3. The twenty most relevant words in the three *cantiche* of the Divine Comedy ranked according to their *tf–idf* frequency

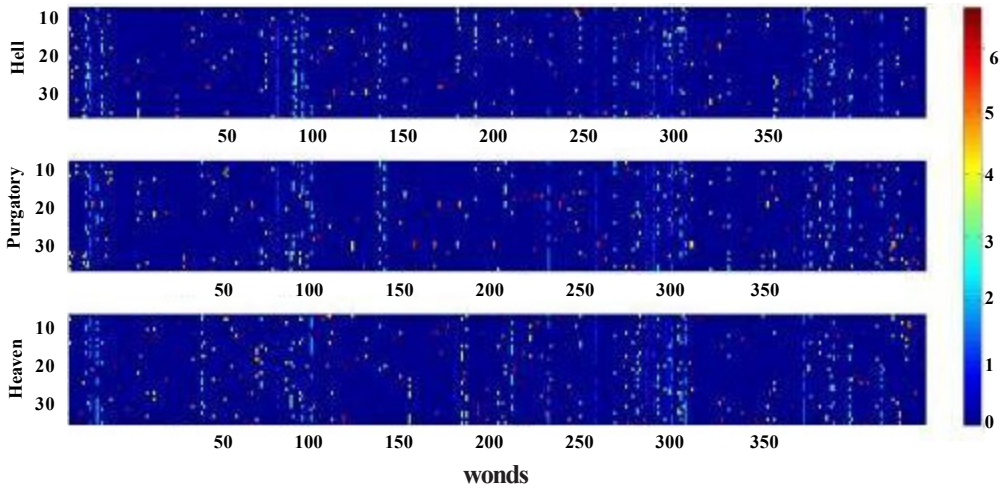

Figure 2. Heat map of the *tf-idf* word frequencies data set for the *Divina Commedia data set*

Many words patterns appear constantly across the three *cantiche*: for examples the words around word # 10 (*altro*, *altri*, others), #100 (*dove*, where) or #250 (*qual*, *quella*, that *quando*, when). Other frequency patterns are typical of one of the three parts. For instance, word #59 (*ciel*, sky, heaven) appears mostly exclusively in the *Paradiso.* The word #150 (*grazia*, grace) appears only in the last ten cantos of the *Paradiso.* On the contrary, the words #39 (*bestia*, beast) appears almost only in the Hell and in the Purgatory.

4.2 Univariate and multivariate analysis

When subjected to univariate analysis only the *tf– idf* frequencies of the word *duca*, *amor* and *dissi* appear to be significantly different (p-value < 0.01) between the *Inferno* and the *Paradiso.*

We used partial least square discriminant analysis with the purpose of investigating if a discriminative stylistic signature could be captured by the word count frequency of the three parts of the *Commedia.* We first applied PLS modeling to the *Inferno* and *Paradiso* data set (67 × 376). Figure 3 shows a plot of the first two PLS components: a net separation (clustering) between the two groups of cantos of the *Inferno* and *Paradiso* along these first two components is evident.

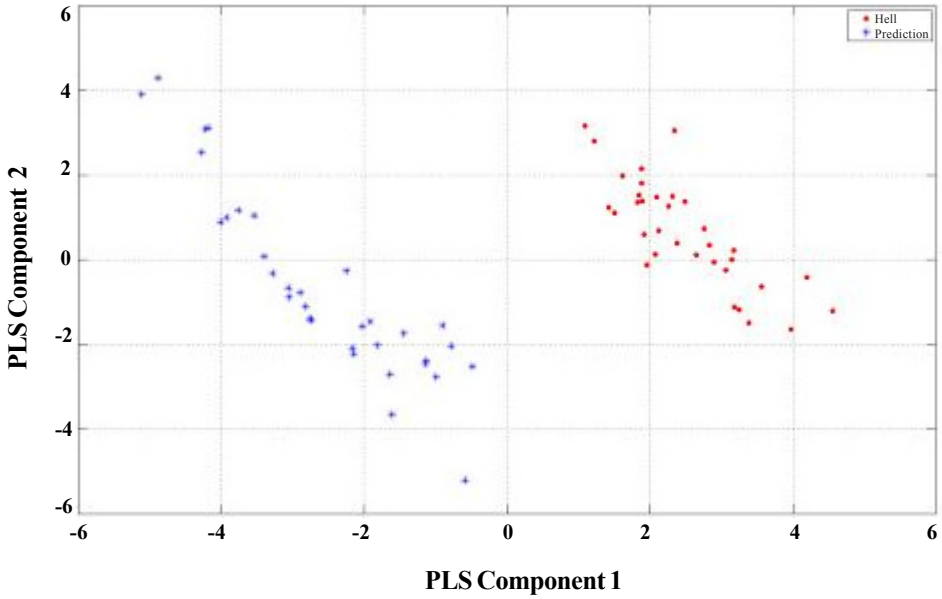

Figure 3. Descriptive PLS clustering of *Hell* and *Paradise* cantos

Although indicative, a simple PLS plot gives no information on the predictive power of the statistical models, *i.e.* if the word frequencies profiles contain information apt to discriminate among the styles of *Inferno* and *Paradiso.* We aimed indeed to assess if a given *canto* can be classified as belonging to the *Inferno* or to the *Paradiso* just on the basis of its word frequencies profile. To do this we applied PLS-DA in a double cross validation scheme as detailed in the Material and Methods section. The PLS model proved to possess a high predictive power: when asked to classify an unknown *canto*, the accuracy and the specificity resulted to be 0.94 and 0.96 respectively. When the standard word frequency was used these results dropped to 0.84 and 0.85 respectively. The confusion matrix is given in Table 4.

	Hell prediction	**Paradise prediction**
Hell	0.97	0.03
Paradise	0.09	0.91

Table 4. Confusion matrix for the predictive discrimination of the cantos of the Hell and the Paradise

When applied to the complete data set (100 × 376) comprising all the one hundred *cantos*, accuracy and specificity decrease significantly as shown in Table 5. Table 6 shows the classification results for the thirty three cantos of the Purgatory.

	Hell prediction	**Purgatory prediction**	**Paradise prediction**
Hell	0.64	0.27	0.09
Purgatory	0.25	0.43	0.32
Paradise	0.04	0.27	0.69

Table 5. Confusion matrix for the predictive discrimination of the cantos of the Hell, Purgatory and the Paradise

Figure 4 shows a predictive clustering of the cantos of the Purgatory obtained by projecting the Purgatory word frequency data set onto the PLS space defined by the model built on the Hell and Paradise data set shown in Figure 3.

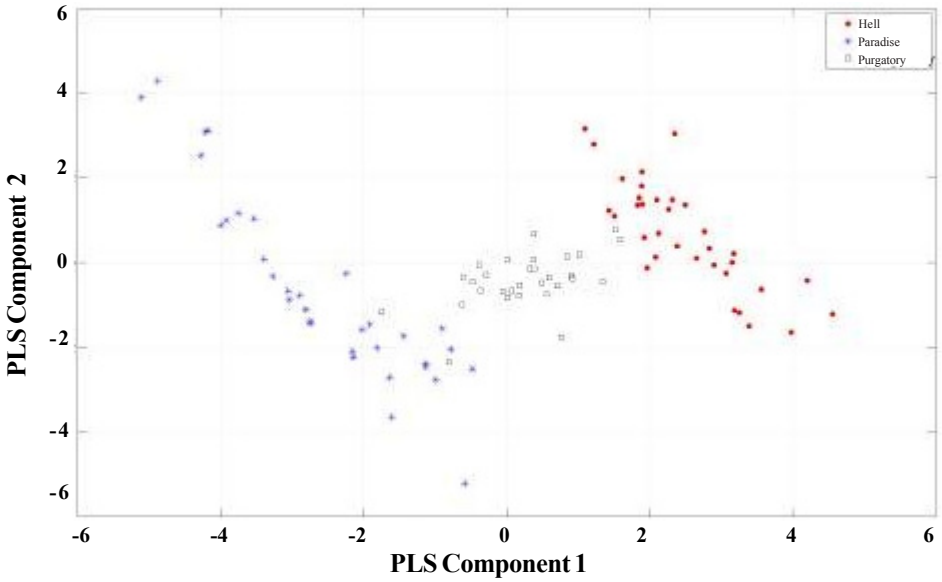

Figure 4. Predictive clustering of the cantos of the *Purgatory*

Purgatory Canto	Theme content	Predicted as
I	Ante-Purgatory: Arrival of the souls in the Purgatory	Paradise
II	Ante-Purgatory: The excommunicates	Hell
III	Ante-Purgatory: The excommunicates	Hell
IV	Ante-Purgatory: The late repentant	Hell
V	Ante-Purgatory: The late repentant	Hell
VI	Ante-Purgatory: The late repentant	Paradise
VII	Ante-Purgatory: The late repentant	Hell
VIII	Ante-Purgatory: The late repentant	Hell
IX	Ante-Purgatory: The late repentant	Hell
X	The proud	Hell
XI	The proud	Paradise
XII	The proud	Hell
XII	The envious	Hell
XIV	The envious	Hell
XV	The wrathful	Paradise
XVI	The wrathful	Paradise
XVII	The wrathful	Paradise
XVIII	The slothful	Hell
XIX	The covetous	Hell
XX	The covetous	Hell
XXI	The covetous	Hell
XXII	The covetous	Hell
XXIII	The gluttonous	Hell
XXIV	The gluttonous	Hell
XXV	The lustful	Hell
XXVI	The lustful	Hell
XXVII	The lustful	Hell
XXVIII	The Earthly Paradise	Paradise
XXIX	The Earthly Paradise	Paradise
XXX	The Earthly Paradise	Paradise
XXXI	The Earthly Paradise	Paradise
XXXII	The Earthly Paradise	Hell
XXXIII	The Earthly Paradise	Paradise

Table 6. Classification results for the thirty three cantos of the Purgatory

Figure 4 Predictive clustering of the cantos of the *Purgatory*. The clustering is obtained by projecting the Purgatory cantons onto the space defined by the PLS model build on the *Hell* and *Paradise* cantos.

Strikingly, the Purgatory *cantos* fall half way between the two clustering defined by the *cantos* of the *Inferno* and the *Paradiso*. We asked ourselves weather the *cantos* falling close to the Hell cluster correspond to the early cantos of the Purgatory and

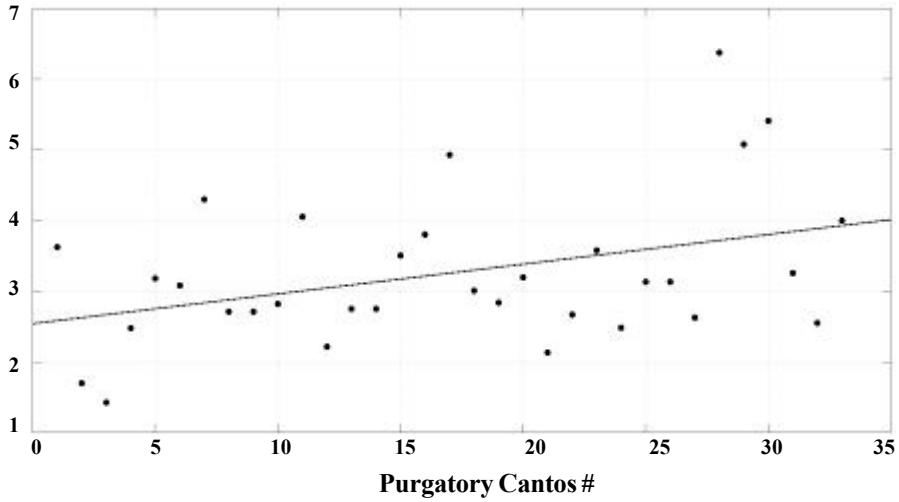

Figure 5. Distance from Hell *dH* for the thirty three *cantos* of the Purgatory

those falling close to the Paradise corresponds to the last ones. As the two dimensional plot provide only a limited view of a multidimensional space, we calculated the distance (in the PLS component space) from the centroid of the cluster of the Hell (d_H distance from Hell) for each of the Purgatory cantos. A plot of d_H *versus* the cantos number I-XXXIII is presented in Figure 5. A positive trend in the distances does exist, showing that the style of the Purgatory evolves, moving from the low style of the Hell to the high style of the Paradise.

Figure 5 Distance from Hell *dH* for the thirty three *cantos* of the Purgatory. A trend of the distance of the Purgatory *cantos* from those of the Hell does exist in the multidimensional space individuated by the PLS model.

5. Discussion of the results

5.1 Words frequency in the *Divina Commedia*

Most occurring words found by the analysis of word frequencies are of course basic grammar elements like indefinite (*altro*, *altri*) or demonstrative (*qual*, *quella*) adjectives. On the other side, other words reflect a precise poetical choice: for instance, Dante reserves the use of the word *grazia* (intended as Divine grace), to the last cantos of the *Paradise.*

5.2 Univariate and multivariate analysis

The results of the univariate analysis are not surprising, as they reflect Dante's exclusive use of certain words in association with certain characters or situation which are peculiar of the Hell or of the Paradise. For instance, in the Paradise Dante is guided by Beatrice and not by Virgil who, in the Hell is usually addressed by Dante as *duca* or *maestro.* It is also interesting to note how the uses of the word *dissi* (*I said*) decreases moving from the Hell to the Paradise, mirroring Dante's difficulties in telling his heavenly experience.

Multivariate analysis shows a clear discrimination and prediction (Table 4 and Figure 3) which points to the existence of a distinctive stylistic substratum in the *Inferno* and *Paradiso cantos* of the *Commedia* which is captured by the word frequencies profiles.

When all the three cantica's are included in the model, it can be noted that the models somehow fails to correctly classify the cantos of the Purgatory (see confusion matrix in Table 5). This means that word frequencies do not carry a strong distinctive stylistic substratum for this part of the *Commedia* as opposite to what found in the case of the *Inferno* and the *Paradiso*. This is certainly to be ascribed to the fact that Dante uses a mixed style, a sort of hybrid of the low style proper of the Hell and the high style of the Paradise; where the Hell is a gloomy and dark place and the Paradise is imbibed by a heavenly bright light, the Purgatory is immersed in a suffuse Spring light. The style and the linguistics of the Purgatory bridges between the anti-music and the cacophony of the Hell and the euphony of the music of the spheres of the Paradise [51]. Dante's climbing of the mountain of the Purgatory is dominated by the dizzing effect of styles changes which reflect and anticipate the pyrotechnics of the language of the Paradise [42].

O navicella mia , com mal se ***carca*** (Purg. XXXII, 129)
Sembiava ***carca*** *ne la sua magrezza* (Inf. I, 50)
E sol quandio fui dentro parve ***carca*** (Inf. VIII, 27)

Seder sovresso una ***puttana*** *sciolta* (Purg. XXXII, 149)
A la ***puttana*** *e a la nova belva* (Purg. XXXII, 160)
Taïde è, la ***puttana*** *che rispuose (*Inf. XVIII, 133*)*

A me rivolse, quel feroce ***drudo*** (Purg. XXXII, 149)
Al ***drudo*** *suo quando disse… (*Inf. XVIII, 134*)*

The first *canto*, describing the arrival of the souls in the Purgatory, is classified as belonging to the Paradise groups, a fact that indicates a net change in the style respect to last cantos of the Hell. The last six cantos (XXVIII-XXXIII) describe the Earthly Paradise: they are all classified as belonging to the Paradise with the exception of the canto XXXII. This abnormality can be explained by considering the content of this *canto* which is the longest in the Comedy and the more dense in term of symbolisms (Sermonti et al. 1994). Dante witnesses the metamorphosis of the mystic chariot, symbolizing the Church. After being attacked by a dragon, the chariot sprout three hornet and monstrous heads before turning into a naked whore (*puttana sciolta*) guarded by a jealous giant who beats her after she turned a seductive glance to Dante. The giant drags her and disappears in the forest. To depict this repulsive scene, Dante uses a low language, characterized by harsh rhymes and tortured rhythms typical of the Hell *cantos* [23]. Dante uses an array of "*infernal*" words [7]:

Interestingly the words *puttana* (whore), *carca* (freighted) and *drudo* (paramour) appears only in the XXXII canto of the Purgatory and in several cantos of the Hell.

These findings show the existence of a stylistic substratum in each *cantica* of the *Commedia* and it can be hypothesized that the way Dante uses and chooses words reflects a well-planned and organized writing scheme conceived by the poet well ahead of the completion of the poem.

6. Conclusions

In this paper we showed how standard multivariate data analysis techniques such as partial least square discriminant analysis can be proficiently applied to investigate writing style and stylistic differences when paired with standard information retrieval and text mining methods. The analysis focused on the analysis of *tf- idf*. The use of PLS-DA combined with word frequencies showed to be a powerful and flexible framework for stylometric investigation of literature corpora.

The superiority of the *tf-idf* over the standard word frequency was assessed by comparing the results of the multivariate analysis performed on both *tf- idf* and the standard word frequency.

By applying them to Dante's *Divina Commedia* we were able to quantify and disentangle the poetic tones used by Dante to characterize and diversify the three parts of the *Commedia*. The scope of the present study was to investigate the existence of a stylistic substratum (which collapses form and content together) in each *cantica* of the *Commedia* and not characterizing Dante's style *per se* as usually done when author attribution is pursued. For the latter task, an analysis of writing invariant (such as word lengths, average length of word and sentences frequency of noun, verb and adjective usage frequency and frequency of function words) would be more appropriate. On the other side, vocabulary richness could have been used but this was not attempted in the present investigation and it will be the subject of a follow up study in which we plan to also apply the proposed approach to the analysis of the effects of translation on a text. As a final remark, we want to stress that the word frequencies apparently do not have any relationship with the underlying semantics. This could be a possible limitation of our approach, however, as reported in (Charkic 2008) style and semantics are linked in a causal relationship: style is the cause, semantics the effect.

E quindi uscimmo a riveder le stelle (Inf. XXXIV, 139).

References

[1] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics,* 2 (1) 97-106.

[2] Alighieri, D., Cecchini, E. (1995). *Epistola a Cangrande* (Giunti).

[3] Alighieri, D., Botterill, S. (1996). *De vulgari eloquentia* (Cambridge University Press).

[4] Alviar, J. J. (2008). Recent advances in computational linguistics and their application to biblical studies, *New Testament Studies,* 54 (1) 139.

[5] Argamon, S., et al. (2003). Gender, genre, and writing style in formal written texts, *Text,* 23, 3.

[6] Baranski, Z. G. (1986). Significar per verba: Notes on Dante and plurilingualism, *The Italianist,* 6 (1) 5-18.

[7] Bargin, T. (1964). *Perspectives on the Divine Comedy* (Bloomington: Indiana University Press).

[8] Barker, M. Rayens, W. (2003). Partial least squares for discrimination, *Journal of Chemometrics,* 17 (3) 166-73.

[9] Bernardo, A. S. (1970). A Dante Milestone:" The Commedia on Computer", *Dante Studies,* 88, 169-74.

[10] Bonferroni, Carlo Emilio. (1935). Il calcolo delle assicurazioni su gruppi di teste., *Studi in Onore del Professore Salvatore Ortu Carboni* (Rome). p.13-60.

[11] Cantone, D., Faro, S. (2003). On the frequency of characters in natural language texts, *Twenty-first Twente workshop on language technology*, 69.

[12] Carpena, P., et al. (2009). Level statistics of words: Finding keywords in literary texts and symbolic sequences, *Physical Review E,* 79 (3) 035102.

[13] Charkic, M. Z. (2008). Semantics and style, *Stylistycka,* 17, 5-16.

[14] Collins, J., et al. (2004). Detecting Collaborations in Text Comparing the Authors Rhetorical Language Choices in The Federalist Papers, *Computers and the Humanities,* 38 (1) 15-36.

[15] Dante, Petrocchi, G. (1966). *La Commedia secondo lantica vulgata* (Firenze: Le Lettere).

[16] de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems,* 18 (3) 251-63.

[17] Denaux, A. (2006). Style and stylistics, with special reference to Luke 1, *Filología Neotestamentaria,* 31-51.

[18] Fay, E. A. (1888). *Concordance of the Divina commedia* (The Dante Society).

[19] Feinerer, I., Hornik, K., Meyer, D. (2008). Text mining infrastructure in R, *Journal of Statistical Software,* 25 (5) 1-54.

[20] Forsyth, R. S., Holmes, D. I. (1996). Feature-finding for test classification, *Literary and Linguistic Computing,* 11 (4) 163.

[21] Hollander, R. (1993). *Dantes Epistle to Cangrande* (University of Michigan Press).

[22] Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts, *Journal of the Royal Statistical Society. Series A (Statistics in Society).* 91-120.

[23] Iannucci, A. A. (1973). Dantes Theory of Genres and the Divina Commedia, *Dante Studies, with the Annual Report of the Dante Society*, 1-25. I

[24] Haka, R., Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of computational and graphical statistics*, 299-314.

[25] Irizarry, E. (1993). The two authors of Columbus Diary, *Computers and the Humanities,* 27 (2) 85-92.

[26] Jones, K. S. (1993). A statistical interpretation of term specificity and its application in retrieval, *Journal of documentation,* 28 (1) 11-21.

[27] Kenny, A. (1986). *A stylometric study of the New Testament* (Clarendon Press).

[28] Khalaf, W. M. (2012). On the Use of Supervised Learning Method for Authorship Attribution, *Engineering and Technology Journal,* 30 (2) 282 - 92.

[29] Khosmood, F., Levinson, R. A. (2008). Automatic natural language style classification and transformation.

[30] Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora, *Sixt ACL/ SIGDAT Workshop on Very Large Corpora pages*, 231- 45.

[31] Koppel, M., Schler, J., Zigdon, K. (2005). Determining an authors native language by mining a text for errors, (ACM). 624-28.

[32] Kruskal, W. H., Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association,* 47 (260) 583-621.

[33] Lansing, R. H., Barolini, T. (2000). *The dante encyclopedia* (Garland Publishing).

[34] Lee, J. (2007). A computational model of text reuse in ancient literary texts, (45) 472.

[35] Lü, L., Zhang, Z. K., Zhou, T. (2010). Zipfs law leads to heaps law: analyzing their relation in finite-size systems, *PloS one,* 5 (12) 661-703.

[36] Luhn, H. P. (1958). The automatic creation of literature abstracts, *IBM Journal of research and development,* 2 (2)159-65.

[37] Matthews, R.A.J. and Merriam, T.V.N. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher, *Literary and Linguistic Computing,* 8 (4) 203-09.

[38] Mealand, D. L. (1989). Positional Stylometry Reassessed: Testing a Seven Epistle Theory of Pauline Authorship, *New Testament Studies,* 35 (02) 266-86.

[39] Merriam, T. (2009). Untangling the derivatives: points for clarification in the findings of the Shakespeare Clinic, *Literary and Linguistic Computing,* 24 (4) 403-16.

[40] Merriam, T. V. N., Matthews, R. A. J. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe, *Literary and Linguistic Computing,* 9 (1) 1-6.

[41] Pang, Bo., Lee, L.,Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques, *Empirical Methods in Natural Language Processing* (Philhadelphia). 79-86.

[42] Paolucci, A. (1965). Art and Nature in the Purgatorio, *Italica,* 42 (1) 42-60.

[43] Patton, J. M., Can, F. (2004). A Stylometric Analysis of Ya ar Kemals nce Memed Tetralogy, *Computers and the Humanities,* 38 (4) 457-67.

[44] Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions, *Computers and the Humanities,* 31 (4) 351-65.

[45] Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval, *Information processing & management,* 24 (5) 513-23.

[46] Salton, G., Wong, A., Yang, C. S. (1975). A vector space model for automatic indexing, *Communications of the ACM,* 18 (11) 613-20.

[47] Schaalje, G. B., et al. (2011). Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes, *Literary and Linguistic Computing,* 26 (1) 71.

[48] Sermonti, V., Alighieri, D., Contini, G. (1994). *Il purgatorio di Dante* (Rizzoli).

[49] Stamatatos, E. (2000). Text genre detection using common word frequencies, *Association for Computational Linguistics*, 808-14.

[50] Stamatatos, E., Fakotakis, N., Kokkinakis, G. (1999). Automatic authorship attribution, 158-64.

[51] Stillman, M. (2005). The Music of Dantes Purgatorio, *The Online Graduate Journal of Medieval Studies*.

[52] Szymanska, E., et al. (2011). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies , *Metabolomics*.

[53] van Velzen, E. J. J., et al. (2008). Multilevel data analysis of a crossover designed human nutritional intervention study, *Journal of Proteome Research,* 7 (10) 4483-91.

[54] Vincent, E. R. (1955). Dantes Choice of Words, *Italian Studies,* 10 (1) 1-18.

[54] Westerhuis, J. A., et al. (2009). Multivariate paired data analysis: multilevel PLSDA versus OPLSDA, *Metabolomics,* 6 (1) 119-28.

[55] Westerhuis, J. A., et al. (2008). Assessment of PLSDA cross validation, *Metabolomics,* 4 (1) 81-89.

[56] Wilkins, E. H., Bergin, T. G. (1965). *A concordance to the Divine comedy of Dante Alighieri* (Belknap Press).

[57] Wold, S., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems,* 58 (2) 109-30.

[58] Wold, S., Sjöström, M., Eriksson, L. (1998). Partial least squares projections to latent structures (PLS) in chemistry, *Encyclopedia of computational chemistry*.