Archeologia e Calcolatori
Supplemento 6, 2014, 203-215

ARTIFICIAL ADAPTIVE SYSTEMS FOR PHILOLOGICAL ANALYSIS: THE PESSOA CASE

1. The Pessoa Archive

Just as Fernando Pessoa offers an extreme example of an open and complex literature, resulting from a heteronymical imagination and constant inclination for expressing the contradictions of human logic and feelings, the same can be said of the Archive that preserves his original manuscripts and which without doubt constitutes an extreme case within the philology of modern and contemporary authors. The features of his writing are in fact well represented in a huge labyrinthine mass consisting of 28,000 documents, which are thematically various and rich in sketches and incomplete works, often written on low quality paper in a hand-writing that is almost always difficult to read. In addition, to crown it all, after his death the papers were organised according to criteria that were at times incongruous and for the most part nullified the arrangements left by the author (Castro 1990; Celani 2005, 2007, 2013).

From poetry to prose, from the essays to the translations, from philosophy to politics, occultism, economics – just to mention some of the subjects that interested him – it seems like there is almost no field of human knowledge that Pessoa did not concern himself with. Hundreds of works have come out of this incredible Archive over the past 70 years, but few of them have been published in a philologically correct form. Even though awareness has grown over the past twenty years about the need to produce critically reliable editions, resulting in a series of volumes that were more conscientious in their attention to the actual material contained in the original manuscripts (I am referring here in particular to the editions published within the project for the national publication of Pessoa's work, being carried out by the "Equipa Pessoa" coordinated by Ivo Castro and published by the Imprensa Nacional-Casa da Moeda publishers of Lisbon), most of Pessoa's works nevertheless continue to circulate in non-critical editions, based either on questionable or not clearly or completely defined criteria. In addition, the challenge faced by the philologist in editing the texts is daunting and it is in part understandable that the series of national editions is proceeding slowly and with extreme difficulty and that the volumes are often doomed to limited circulation among specialists, while the general public prefers slimmer volumes unencumbered with critical notes and commentary (in this sense, the series dedicated to Pessoa co-ordinated

by Teresa Rita Lopes and published by Assírio & Alvim publishers of Lisbon have enjoyed great success).

One of the main difficulties presented by Pessoa's works lies in the author's method of composing and conserving his writing. He in fact tended to work contemporaneously on different projects, some of which were written over a span of ten, if not twenty or more, years. Moreover, Pessoa did not often date his sheets. His works, for the most part unpublished during the author's lifetime, tend to be unfinished and present few structural indications; therefore, the problem of identifying objective criteria for their arrangement becomes central. A solution to this problem could be the reconstruction of the documents' chronology, which can in turn be useful in reconstructing the writing process. This was what was done for example with one of the most complex – from the editing point of view as well – of Pessoa's works: the *Livro do Desassossego*, a prose work written in a lyrical style, halfway between an intimate diary, a notebook of reflections, and the narration of the inner thoughts and feelings of a fictional character who would over time take on the definitive name of Bernardo Soares. Given the great disparity of themes and the complete absence of any narrative thread, each editor of the work opted for his own textual reconstruction, following different and often subjective criteria.

The first critical edition focusing on an objective criterion based on the reconstruction of the chronology of the individual passages did not appear until 2010 (Pessoa 2010). The chronology, reconstructed by means of traditional methods of synoptic comparison of the material characteristics of the original documents (Pessoa 2010, II, 530), however, did not produce optimal results: of the 445 passages included in the edition, 316 maintain a very hypothetical dating.

Undoubtedly, the basic criterion is correct, inasmuch as the material features of the originals certainly make it possible to trace the precise stages of the writing, which can be anchored to direct or indirect dates for the purpose of constructing a definite chronology of the literary works. But there are many variables to take into consideration and the process is rather complex. The only way to obtain complete and reliable results is to resort to computerised procedure, which can automatise the crosschecking of the data and obtain wide-ranging results which, if correctly interpreted, can permit a consistent overall view of the different stages and intersections in Pessoa's writing. A first attempt in applying this procedure, based on a sample of 128 original documents from the Archive, has been carried out in the last two years by the ARCHEOSEMA research group (Ramazzotti 2013). The remaining sections of this paper include a description of this work, the results obtained and those which may be obtained in future by extending the dataset to include all the documents in the Pessoa Archive.

2. Basic idea and creation of the dataset

Those who have in some way dealt with the Pessoa Archive for long periods of time will have noticed that in the apparent chaos and random nature of the writing paper used, there are in fact many recurrent elements, ranging from specific formats (size, colour, division into squares, etc.) to headings and watermarks. These elements can provide links among the different parts of the Archive, which make it possible to identify related sections, useful in reconstructing the original stratification of the material. The entropy resulting from the numerous manipulations of the material and the attempts to organise it after the author's death, while eliminating every trace of the original arrangement, nevertheless do not prevent further attempts to reconstruct it. The first step in this direction is to seek to identify all the variables which might be useful in reconstructing the manner and time of the original writing process. For my doctoral thesis, completed in 2004, I worked on an edition of a Pessoa text – an unpublished (and uncompleted) detective story entitled *O Caso Vargas* (Pessoa 2006) – of which I described the material features. The work, consisting of a collection of 128 sheets, was used as a sample *corpus* for the ARCHEOSEMA experimentation.

Each side of every page (for a total of 184 – not all the sheets were written on both recto and verso), identified by its press mark within the Archive, constituted a record in the dataset and was described on the basis of a large number of variables relating to its material characteristics.

The supports on which the texts were written were identified according to parameters such as size (length and width), colour, presence (and typology) of watermarks, headings or other printed text on recto or verso, presence of printed lines or squares, and presence of any cut, folded or perforated parts. The instruments used for writing were on the other hand classified according to type (hand-written, type-written or a combination), use of pen or pencil, and colours used. Finally, all the indications of explicit connections between one document and another have been included, particularly for fragments occupying more than one sheet. In all, as many as 95 variables were indicated for each record (41 for the writing support, 13 for the writing instrument and 41 for the connections among the papers). Starting with the support variables, we can observe that some of the 41 types occur with much greater frequency than others, starting with type 11 (a white sheet cut along one side and measuring 22.1×16.2cm), which occurs in as many as 25 sheets out of 128; type 10 (very similar to the preceding type, being a white sheet cut along one side and measuring 22×16.3cm), which accounts for 16 sheets; type 27 (white sheet measuring 27.4×21.3cm), which includes 12 sheets; and type 19 (which is the back of a form entitled "Anúncios – Tabela de preços", measur-

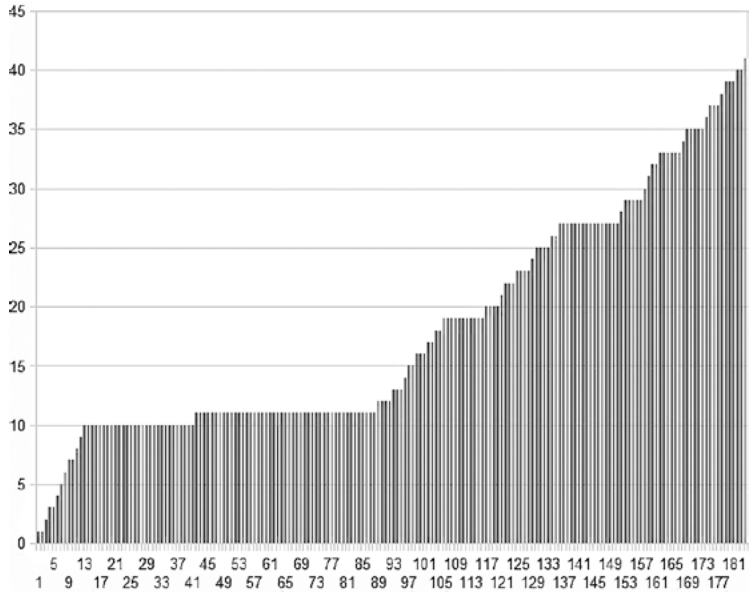

Fig. 1 – Distribution of support types.

ing 26.3×19.6cm) which includes 11 sheets. Thus, half of this collection of papers can be referred to only four types of support (Fig. 1).

With regard to the writing instruments, it is observed that type 5 (pencil) is entirely predominant, and is used in all the papers belonging to support types 10 and 11, in 7 of the type 19 sheets and in 5 of the type 27 sheets (Fig. 2). The matrix obtained from the dataset was then inserted into Intelligent Data Mining developed by Massimiliano Capriotti at the Semeion Research Centre. The dataset was developed according to three different metrics: Linear Correlation (LC), Prior Probability (PP) and Auto-Contractive Maps (Auto-CM). For each of these a Minimum Spanning Tree (MST) was calculated, and represented in the form of a graph using GEPHI v. 0.8.1 software, which made it possible to obtain a concise visual representation of the results. At this point, the data were analysed.

3. Analysis of the data

The graphs in Figs. 3, 4, and 5, based respectively on LC, PP and Auto-CM, were obtained through the application of two main filters: betweenness centrality (which indicates the centrality of a node in the network, obtained on the basis of the number of shortest paths that pass through that

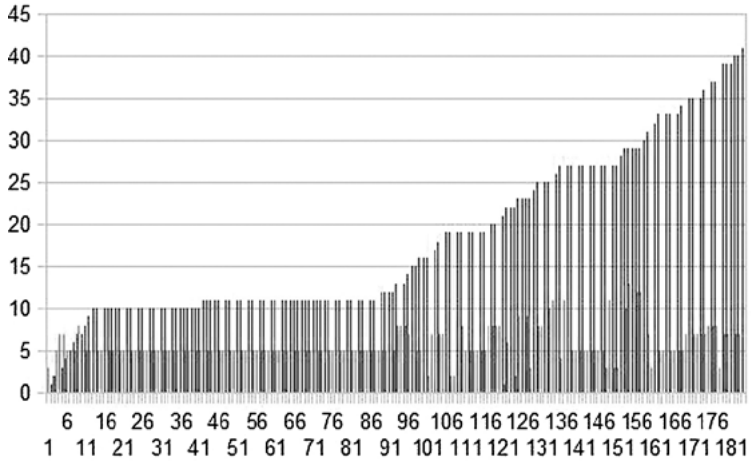

Fig. 2 – Distribution of writing instrument types compared to distribution of support type.

node from all vertices to all others) and modularity (which sub-divides the nodes into a pre-determined number of clusters on the basis of their affinity). Each point (or node) indicates one of the documents which make up the text; the different classes, designated by different colours, identify the different sections of the document set; the lines that connect the nodes indicate the material contiguities among the different records – contiguities which can indicate closeness in conception or drafting of the text. By comparing the graphs produced with the three different algorithms, we can observe that Auto-CM produces a more intelligible result which is in line with the data already known in relation to the content of the individual fragments. Both LC (Fig. 3) and PP (Fig. 4) on the other hand produce graphs which tend to be polycentric and less efficient and which show values in the weights of the connections that are on average low. On the other hand, by analysing the weighted graph obtained with Auto-CM (Fig. 5), it is possible to immediately identify five clusters, of which the first is clustered around the central node while the others are designated by four clearly distinct branches.

The four peripheral clusters correspond, in whole or in part, to the four most frequent types of support. Cluster 2, identified in the upper branch on the left, corresponds completely to support type 11; cluster 3, identified in the lower branch on the left, corresponds almost completely to support type 10; cluster 4, designated by the upper right branch, consists for one third of support type 19; and cluster 5, indicated by the lower right branch, consists for 30% of support type 27. Finally, 35% of cluster 1, which corresponds to the records clustered on the central node, consists of support type 33 (white

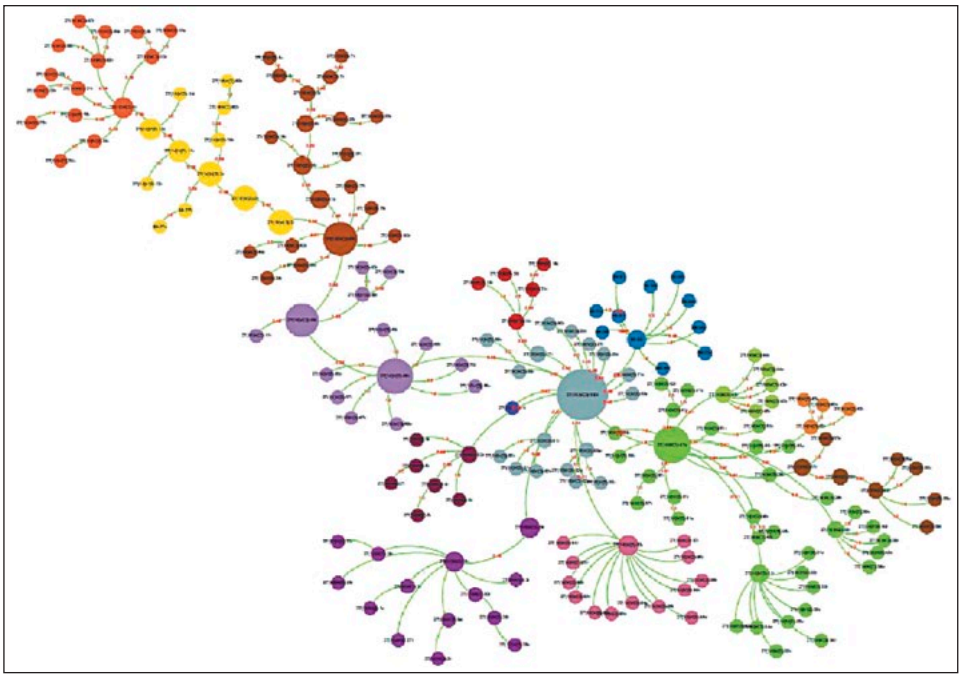

Fig. 3 – LC graph.

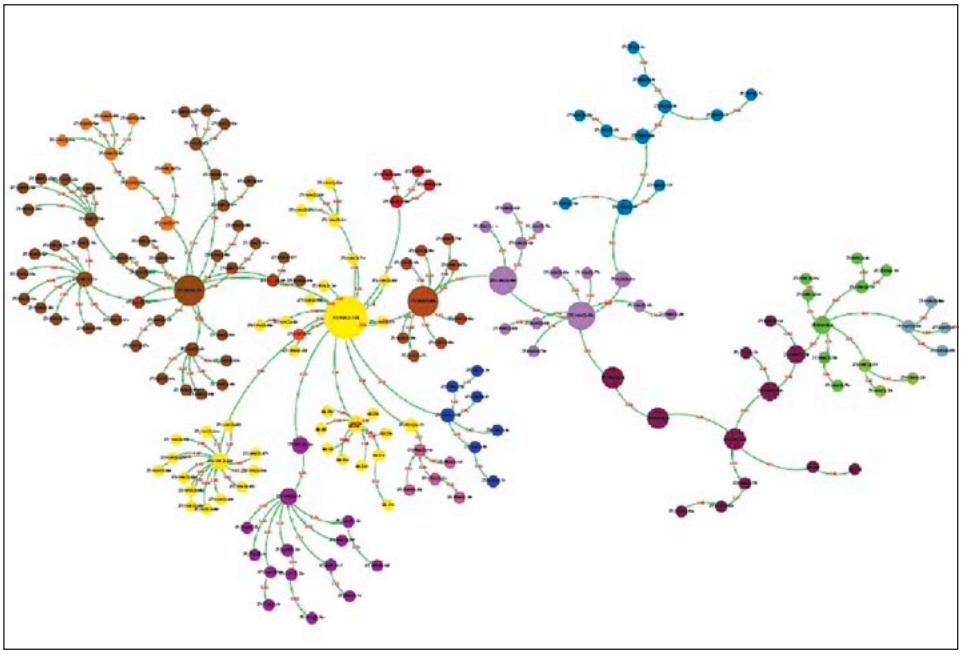

Fig. 4 – PP graph.

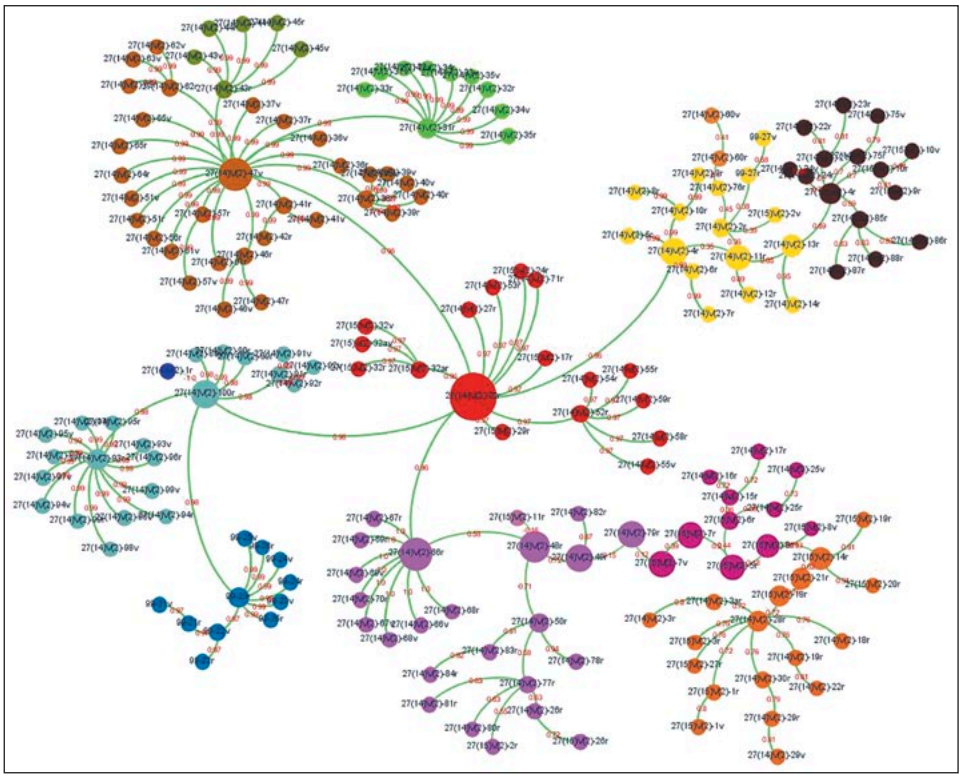

Fig. 5 – Auto-CM graph.

sheet measuring 27.5×21.6 cm) while 23% is made up of support type 12 (folded white sheet headed "F. Caetano Dias – Perito-contabilista" measuring 22.2×14 cm. This was Fernando Pessoa's brother-in-law, managing editor of the magazine «Revista de Comércio e Contabilidade», in which the Portuguese poet was to publish a great number of articles regarding the economic sphere in 1926).

The complete identification between support type and writing type in cluster 2, makes it possible to identify a highly homogeneous block, which is also very cohesive due to the constant force of the connections among its elements (which is always 0.99) and which undoubtedly indicates a clearly distinct phase in the drafting of *Caso Vargas*. The documents are all handwritten in pencil. At this stage, Pessoa would have worked, during very close periods of time, on sections of Chapters 2, 3, 4, 10, 12 and 15.

As already indicated above, another very homogeneous block is cluster 3, consisting entirely of support type 10. The documents are all hand-written

with the exception of one. This stage likely corresponds to the writing of parts of Chapters 9 and 12. Cluster 4 on the other hand is less homogeneous; it consists of a central block, which can be identified with support type 19, with a couple of more peripheral segments. The weights of the connections are medium-high, but definitely lower than the average in clusters 2 and 3. The documents are for the most part hand-written. This section, which is more dispersed, consists of writing which is less uniform in content and corresponds to parts of Chapters 2, 4, 6, 8, 10, 12 and 13. Cluster 5 is the most complex and least homogeneous and cohesive of all; the documents are almost equally hand and type-written. A central section consists prevalently of support type 27, while the more peripheral branch presents greater complexity and has much lower average values for the weights of the connections. This cluster includes the writing of sections of Chapters 1, 2, 6, 7, 8, 10, 12, 13 and 15. Finally, cluster 1 appears quite cohesive (with weights of the connections being constant at around 0.97), but less homogeneous than clusters 2 and 3. The documents are all hand-written. The writing of parts of Chapters 2, 3, 7, 12, and 15 belongs to this stage. This description of the five clusters is confirmed in a version of the same graph filtered through the degree parameter (which classifies the nodes according to the number of connections they possess; Fig. 6).

Here it is possible to identify the more cohesive and homogeneous clusters with greater clarity, starting with cluster 2, followed by cluster 3, then 1 and finally by 4 and 5. Let us now try to connect this structure to a likely chronology. *O Caso Vargas* was written during a period of time that goes from the early 1920s (probably after 1923) to 1935, the year of Pessoa's death. The work is mentioned in the famous letter on the genesis of the heteronyms, sent to Adolfo Casais Monteiro on 13 January 1935; even though the title is not mentioned explicitly, it is very likely that the text he is referring to is to be identified as *O Caso Vargas* (Pessoa 1998, 252). These dates were postulated on the basis of information taken directly from the originals, thanks in particular to the presence of headings, watermarks or other texts printed on the sheets used. Such elements make it possible to observe in particular that some of the texts in cluster 5, especially its "peripheral" branch, containing among the material also the fragments which make up Chapter 1 of the work, can be dated to a period after 1923 (the fragments are written on back of a pamphlet by Pessoa titled *Sobre um manifesto de estudantes*, published in April 1923), but probably not much later than that date. Within cluster 2 are found on the other hand passages written during and later than the period 1924-1925 (since they were written on forms for the publication of the magazine «Athena», of which Pessoa was co-editor and which published its five numbers precisely between 1924 and 1925) and others dated to the years between 1926 and 1928, since they were written on sheets with the above-mentioned heading "F. Caetano

Dias – Perito-contabilista", of which there are other samples in the Archive, some of which have direct dates going back to 1926-1928 (Pessoa 2010, II, 351). Finally, some fragments belonging to cluster 4, found once again in a peripheral branch, can be dated in the year 1931. These are texts written on watermarked sheets headed "Graham Bond Registered"; almost all the texts on the same support present in the Archive, when dated, in fact date back to 1931 (Pessoa 2010, II, 351). There are no traces of dates for fragments belonging to clusters 2 and 3, even though the close connections they have with cluster 1 may suggest that they fall within the central stage of the writing of the text, which is around the second half of the 1920s.

Many of the fragments belonging to cluster 5 in effect include texts that have no direct references to *Caso Vargas*, but contain instead reflections on criminal psycho-pathology of the type that are more characteristic of an essay style. The reflections are completely lacking in any direct references to the events or characters of the work, and were incorporated into it only at a later date. A clear example is sheet $27^{14}V^{2}$-79^{r} in which a hand-written note referring to the character of the murderer is added – most certainly at a later date – to a type-written text tending towards the schematic and regarding a general classification of pathologies which can lead to murder.

The same Auto-CM graph can be visualised in a partially linear manner, reconstructing a possible sequence of the writing stages of the work (Fig. 7). Here cluster 1 is still at the centre, cluster 5 is to the left, cluster 4 to the right, cluster 3 is above and cluster 2 below cluster 1. It thus becomes possible to connect the various clusters (or stages) shown in the graph to an absolute chronology, a procedure that can be extended by increasing the sample subjected to analysis. The evident limitations of the results obtained so far are in fact due to the small size of the *corpus* under study.

4. Conclusions

The Pessoa Archive can be seen as an archaeological site damaged for years by those who delved into it. The huge accumulation of papers still holds traces of the original stratification, which makes it possible to identify many cohesive strata and establish their chronology. Pessoa was in the habit of contemporaneously writing different works, some of which have a time span that is very wide, reaching at times even ten, if not twenty, years. A case in point is *Livro do Desassossego*, whose *trechos* were produced, albeit not uninterruptedly, over a period that goes from 1913 to 1934. But many other works in fact have similar chronologies: we might mention, for example, *Fausto* (1908-1934), *O Caso Vargas* (1923 ca.-1935), *Mensagem* (whose first poems were written as early as 1913, whereas the work itself, as is well-known, was not published until 1934. A date after which in any case Pessoa's work

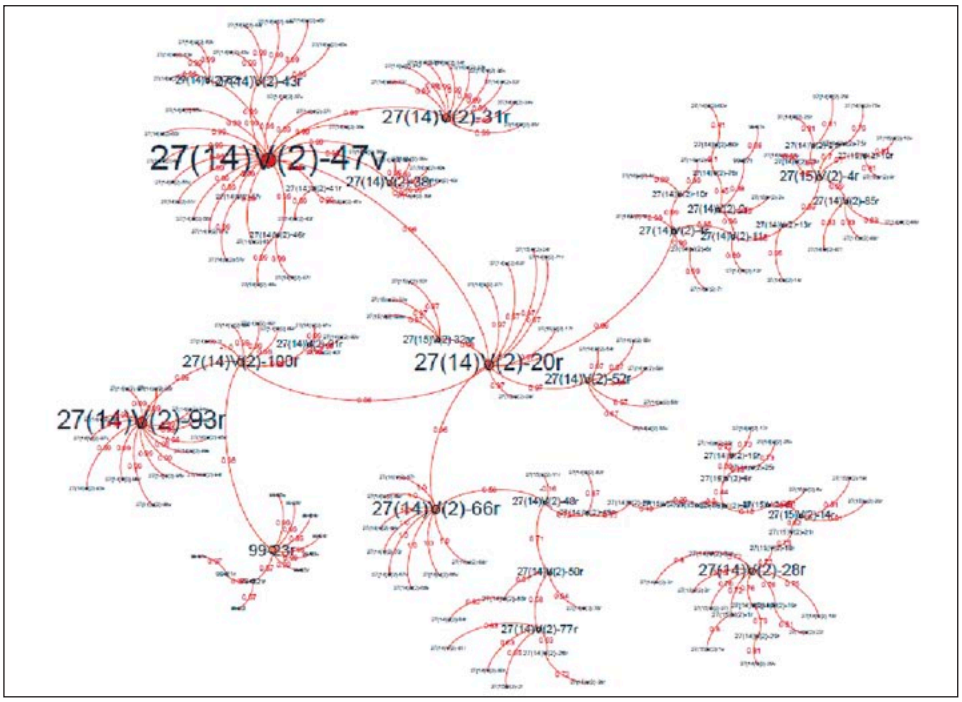

Fig. 6 – Auto-CM graph filtered through the degree parameter.

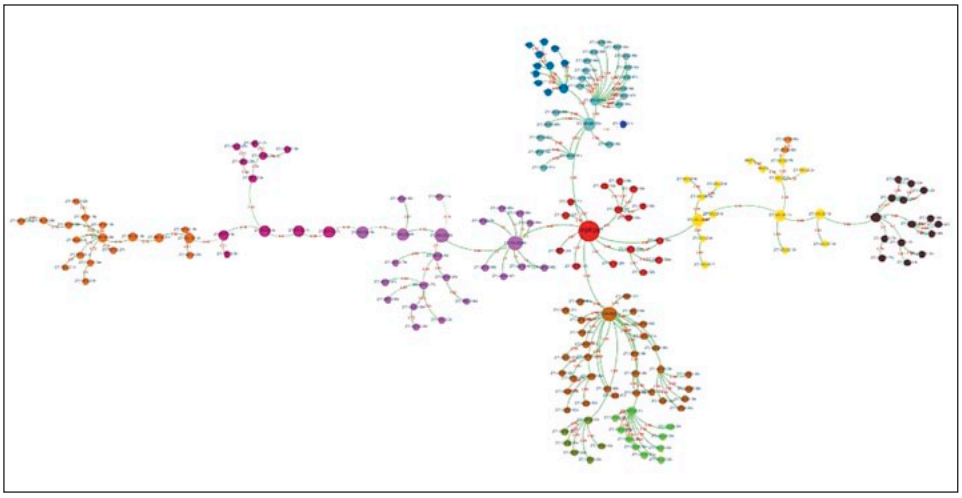

Fig. 7 – A possible chronological visualisation of Auto-CM graph.

on the texts was not completed, as indicated by a printed copy of the work with numerous hand-written variations).

Dealing with this will necessarily require comprehensive study – before any edition is prepared for publication – aimed at gathering up the threads of all the intersections, including those related to content and to the material aspect of the documents in the Archive. Obviously, an endeavour of this scope, carried out on such an extensive *corpus*, will involve a complex process that cannot be achieved with the procedures hitherto adopted: in this sense, a synoptic comparison carried out "by sight", like the one used by Pizarro for *Livro do Desasocego*, besides requiring a long period of time is also doomed to furnish only partial views in which essential elements can escape attention – both in the details or the general aspect.

By subjecting a dataset including the data contained in the entire collection of papers in the Archive to the above-described procedure, it would finally be possible to obtain a comprehensive map of the stratification of Pessoa's writing, in a reconstruction of all the diachronic and synchronic aspects of his creative process, thus illuminating the different stages in the writing of the individual works and placing them within the wider context of his entire literary production. This would be an extremely useful tool for publishing purposes which could permit the construction of a comprehensive scheme for the publication of Pessoa's works, beginning not with the individual works, but with the whole of his literary output. Within such a scheme, each work would have its own specific place and the relations (as well as the not infrequent over-laps) among the different works would be much clearer and more evident. Underlying the overall – albeit incomplete – project of Pessoa's works there is an organised structure, clearly visible despite the apparently fragmentary nature of a great part of his literary output. Pessoa devoted almost as much time to planning as he did to writing his works.

The Archive is full of schemes and lists as well as introductions and prefaces seeking to explain the *ratio* hidden behind his works. Every volume, every collection, every essay is placed within a larger scheme which connects them to other works within a comprehensive and unifying vision. The works of the different heteronyms, like the one of the orthonym, can be read independently of one another, but they take on a further level of meaning only when they are placed alongside one another, in a web where the individual parts are closely interwoven.

The proliferation of heteronyms and literary personalities does not in itself imply a fragmentation of the literary output, but rather a clear attempt at organising and cataloguing the numerous parts of the works, for which each alternative name functions as an explicative label, a space in which to contain elements which in some way are homogeneous.

By carefully reconstructing the connections and homogeneous strata that link the documents present in the Archive, it will finally be possible to read the scheme as arranged through the process of its creation and constitution. But in order to be able to embrace it in its entirety, we need instruments which can concisely represent its complexity without simplifying it. The response to this need may perhaps be found in the utilisation of adaptive artificial networks.

Simone Celani
Dipartimento di Studi Europei, Americani e Interculturali
LAA&SAA
Sapienza Università di Roma

REFERENCES

Canettieri P., Loreto V., Rovetta M., Santini G. 2005a, *Ecdotics and Information Theory*, «Filologia Cognitiva», 3 (http://w3.uniroma1.it/cogfil/ecdotica.html).

Canettieri P., Loreto V., Rovetta M., Santini G. 2005b, *Higher Criticism and Information Theory*, «Filologia Cognitiva», 3 (http://w3.uniroma1.it/cogfil/attribuzioni.html).

Carile P., Mandich A.M. 1995 (eds.), *Discorrere il metodo. Il contributo della francesistica agli studi metodologici*, Ferrara, Centro Stampa Università.

Castro I. 1990, *Editar Pessoa*, Lisboa, Imprensa Nacional-Casa da Moeda.

Celani S. 2005, *Il Fondo Pessoa*, Viterbo, Sette Città.

Celani S. 2007, *Il Fondo Pessoa: una sciarada filologica*, «Quaderno del Premio Letterario Giuseppe Acerbi», 8, 87-92.

Celani S. 2012, *Fernando Pessoa*, Roma, Ediesse.

Celani S. 2013, *Quale Pessoa? Ultime edizioni e nuove prospettive*, «Critica del Testo», 16/2, 335-353.

Cherchi P. 2001, *Filologie del 2000*, «Rassegna europea di letteratura italiana», 17, 135-153.

Ciotti F., Crupi G. 2012, *Dall'informatica umanistica alle culture digitali*, Roma, La Sapienza-Digilab.

Ferrari P. 2008, *Fernando Pessoa as a Writing-reader: Some Justifications for a Complete Digital Edition of his Marginalia*, «Portuguese Studies», 24/2, 69-114.

Fiormonte D. 2003a, *Scrittura e filologia nell'era digitale*, Torino, Bollati Boringhieri.

Fiormonte D. 2003b, *Scrittura, filologia e varianti digitali*, «Filologia Cognitiva», 1 (http://w3.uniroma1.it/cogfil/varianti.html).

Giaveri M.T. 1995, *L'edizione genetica: tradizioni filologiche e orizzonti informatici*, in Carile, Mandich 1995, 149-155.

Italia P. 2013, *Editing Novecento*, Salerno, Salerno Editrice.

Italia P., Raboni G. 2010, *Che cos'è la filologia d'autore*, Roma, Carocci.

Lebrave J.-L. 1999, *L'édition critique au XXI^e siècle*, in *I nuovi orizzonti della filologia. Ecdotica, critica testuale, editoria scientifica e mezzi informatici elettronici*, Roma, Accademia Nazionale dei Lincei, 127-132.

Mordenti R. 2001, *Informatica e critica dei testi*, Roma, Bulzoni.

Mordenti R. 2012, *Filologia digitale (a partire dal lavoro per l'edizione informatica dello Zibaldone Laurenziano di Boccaccio)*, «Humanist Studies & the Digital Age», 2.1 (http://oregondigital.org/hsda/article/view/2991).

Orlandi T. 1990, *Informatica umanistica*, Roma, La Nuova Italia.

Orlandi T. 2010, *Informatica testuale. Teoria e prassi*, Roma, Laterza.

Pessoa F. 1998, *Cartas entre Fernando Pessoa e os directores da Presença*, E. Martinez (ed.), Lisboa, Imprensa Nacional-Casa da Moeda.

Pessoa F. 2006, *Il caso Vargas*, S. Celani (ed.), Viterbo, Il Filo.

Pessoa F. 2010, *O Livro do Desasocego*, 2 vols., Lisboa, Imprensa Nacional-Casa da Moeda.

Ramazzotti M. 2013, *ARCHEOSEMA. Sistemi Artificiali Adattivi per un'acheologia analitica e cognitiva dei fenomeni complessi*, «Archeologia e Calcolatori», 24, 283-303.

Rossi L.C. 2007, *La filologia della letteratura italiana sul confine tra cartaceo ed elettronico*, «Studi di Filologia Italiana», 65, 401-405.

ABSTRACT

Fernando Pessoa represents an extreme case in the context of contemporary author's philology. The breadth of his legacy, the large number of unpublished works at his death, the disorganisation and incompleteness of his materials and the entropy caused by the early processes of inventory produced an archive, now largely in the possession of the Portuguese National Library, partially refractory to the application of traditional text-criticism methods. This paper will demonstrate, through some application examples, that a careful study of material aspects concerning the originals of the Pessoa archive, made through the use of Artificial Adaptive Systems, will shed new light on the complex and multi-layered writing system created by Pessoa and identify new genetic relationships among his works, useful for the construction of an overall mapping of his literary output.