Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature

Jean-Baptiste Camps, Florian Cafiero

Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature*

Jean–Baptiste Camps**,

Florian Cafiero***

** Laboratoire *Études et édition de textes médiévaux* (EA 4349)
Université Paris – Sorbonne
1, rue Victor Cousin
75005 Paris
jbcamps@hotmail.com

*** École centrale de Paris
Grande voie des Vignes
92295 Châtenay-Malabry CEDEX
florian.cafiero@polytechnique.edu

Abstract

The authors present here an exploratory and unspecific method that does not necessitate any *a priori* on the data –or any heavy transformation such as lemmatisation– that would have to be understood as a first step in the apprehension of a corpus. After a first phase of calibration, based on a control sample, the authors will introduce a method which heuristic value is to bring out, at different levels, internal divisions of different kinds (diachronic, diatopic, related to authorship or scribes,...), that can then be analysed specifically. The authors illustrate this method by applying it to a corpus of Occitan medieval texts, the *vidas*. The corpus's authors and origins are in good part unknown.

Les auteurs présentent ici une méthode exploratoire, non spécifique et ne nécessitant pas d'*a priori* sur les données — ni de transformations lourdes comme la lemmatisation —, qui se définit comme une première étape dans l'appréhension d'un corpus. Après une phase de calibration, fondée sur l'utilisation d'un échantillon témoin, sera présentée une méthode dont la valeur heuristique est de pouvoir rendre apparentes, à différents niveaux, des divisions internes de différentes natures (diachroniques, diatopiques, autoriales, scribales,...), pouvant ensuite faire l'objet d'analyses spécifiques. Ils illustrent cette méthode par une application à un corpus à première vue assez homogène de courts textes occitans médiévaux, les *vidas*, dont les auteurs et les origines sont en bonne part inconnues.

*The database used by the authors, along with the *R* scripts, and additional relevant material can be found at http://graal.hypotheses.org/617. Readers are welcome to write to the authors for further information. This article was presented during the 2011 MASHS Conference, and we would like here to thank its organisers and attendees; also, we wish to thank Prof. Fabrice Rossi (SAMM, Univ. Paris 1 Panthéon–Sorbonne) for his precious help with the R *Yasomi* package, as well as David Brossard for proofreading our English.

1 A peculiar corpus: the *vidas*

1.1 Elements of context

Amongst the relatively small corpus of Occitan prose literature from the XIII$^{\text{th}}$ century that has survived to this day, the *vidas*, along with the *razos*, constitute a corpus of roughly 225 texts known as the "biographies" of the troubadours since Raynouard's edition in 1820. Several debates surround these texts' origin: were they performed at first or were they created specifically to be integrated into manuscripts? Did they appear in Northern Italy or do they come from an older Languedocian tradition? Despite the lack of medieval unambiguous vocabulary, there is an apparent distinction between commentaries, explanation of a specific poem (a *razo*), and more general biographical narratives detached from a specific work (*vidas* as we call them).

Written in Northern Italy during the XIII$^{\text{th}}$ century (perhaps, for some of them, from some preexistent Languedocian material[1]), *vidas* seem to answer a specific need for contextualisation, consistent with the needs of a public separated from the poems' authors both by time (up to over a century and a half), space, and native language. These *vidas* also participate in a larger movement linked with the transformation from an oral tradition based on performance to a written one: organisation of the poems, classification in various genres, attribution to the right authors, and even at times philological care for the text and its establishment. In this context, the *vidas* are used both to determine an adequate and unique explanation as well as an exegesis of the troubadour's poems based on the author's life. It also helps organising the troubadours in a hierarchical way, each author receiving a proper place, the esteem and rank he or she is entitled to by talent or birth. In this transformation, the oral living tradition of troubadour's lyric becomes a honourable literature, and the troubadours themselves are dignified and looked upon as respectable authors[2].

The sample this paper focuses on is composed of 52 *vidas*, taken from the same Venetian manuscript of the final quarter of the XIII$^{\text{th}}$ century (Bibl. ap. vat., Reg. lat. 5232, known to philologists as the *chansonnier A*)[3]. Much like other luxurious Venetian manuscripts of the same period (in particular *I* and *K*), *A* uses the *vida*, along with a miniature — often described as a "portrait" — to mark the beginning of the authors' sections and to introduce the poems[4]. The most recent datable *vida* in our corpus is from c. 1274, while the date of the earliest is uncertain. Their authorship, exact location and composition date are in question. Only a few *vidas* are signed, and none are in *A*. Still, in some manuscripts (but not in *A*, the *vida* of Bernart de Ventadorn is "signed" by a known author, Uc de Saint–Circ († c. 1257), one of the troubadours having lived in Italy. He is also considered as a possible author for several other *vidas*, although to a much debated extent[5]. The main purpose of this paper will be to address questions

[1] See the brief summary of this question in Burgwinkle (1999, p. 252–523).

[2] For more information about *vidas*, see, among numerous studies Wilson Poe (1984) and Meneghetti (1992).

[3] We used the edition provided by Françoise Vielliard in Lemaître and Vielliard (2008).

[4] Both the *vida* and the miniature are part of a metatextual apparatus which Meneghetti calls "un doppio filtro metatestuale" (Meneghetti, 1992, p. 348).

[5] Panvini (1952) thinks that Uc can be considered author of 36 *vidas* , while Favati (1961) attributes him a majority of the *vidas* and nearly all *razos*. Meneghetti considers as certain the paternity of Uc on the *vidas* of Raimbaut d'Aurenga, Guillaume IX, Sordello (in the *IK* version) and Guilhem Figueira, and has even stated that it was "ragionevole, quantomeno a livello di ipotesi di lavoro, considerare come composte, o almeno rivedute, da Uc tutte le biografie contenenti allusioni ad avvenimenti anteriori al 1257" (Meneghetti, 1992, p. 243-245), while she recognised that historiography might have fallen, at some point, to a "*furor* attribuzionistico a senso unico", and has also proven the existence of other biographers

of date, authorship and manuscript tradition, without having to develop an automated syntactic analysis, but rather by working from a graphical and lexicometrical point of view.

1.2 Lexicometric characteristics of our corpus

The *vidas* are brief narratives, telling us in a very standardised style the life and deeds of a troubadour, as an introduction to his poems. The fifty–two *vidas* of *A* are usually short (a few lines), but there are some variations in size (fig. 1, p. 4): the shortest one (the *vida* of Richartz de Tarascon) has only 22 words, and the longest one (Raimon Jordan, the viscount of Saint Antonin) has 676 words, the mean being 164 words (and the median 147). Only two especially long *vidas* have more than 400 words and they correspond to very specific stories — Guillem de Cabestaing and the story of the eaten heart[6], Raimon Jordan and his impossible love. While the unusual length of these latter *vidas* might be an indication as to their somewhat later composition, the length of the *vidas* in itself is not necessarily always a clue as to its possible author or date. In many cases, the length of the text is determined from the available sources (mostly the poems attributed to the said troubadour), and also to the fame of the troubadour, the esteem he was given, and his rank in troubadour lyric. The fifty-two *vidas* amount for 8519 words and 1479 word-forms, of which 833 are *hapax* (respectively 9,78% and 56,32% of the total).

Vidas are globally composed according to the same patterns. The genre in itself is clearly recognisable, and they constitute, as Boutière and Schutz (1964, p. viii) stated, "un véritable genre littéraire, dont le moule, le squelette, la langue et la phraséologie ont été si bien respectés au cours du temps qu'il serait aisé d'en faire aujourd'hui des pastiches". In fact, they are built around the same structure and use a limited amount of words and *formulae* to describe the troubadours. Most of them start with the phrase "[Troubadour's name] si fo de [place] e fo [first designation of the troubadour]", and the rest of the text is, for the most part, constructed on this relatively standard and nearly always present social designation, along with a limited set of qualities, and narrative patterns attributed according to this designation [7].

In terms of repeated segments, this leads to two antagonist factors: the lack of graphical standardisation — from modern standards — characteristic of medieval vernacular languages such as Old Occitan, causing the number of forms to increase; the limited amount of vocabulary and the stereotyped *formulae*, leading to a restriction in terms of *lemmata*. This is illustrated by the study of the Repeated Segments[8]. There is a total number of 1228 RS of length (in words) $l \geq 2$ and frequency $freq. \geq 2$ distributed according to Tab. 1.

The first observation to be made is both the elevated number of RS and the relatively low average frequency, which confirms the existence of these two antagonist fac-

(Meneghetti, 2002, p. 148), contesting the stylistic comparative approach of Guida (1996), potentially biased by the very standard and formulaic style of the *vidas*.

[6] For more elements on this very famous story, see Gaunt (2003), or, for a more general synthesis, in English, Gaunt (2006, part. p. 77–79).

[7] In *A*, there is only one *vida* with no such designation, the *vida* of Marcabru, and it is justified by the fact that he was, as a child, "gitaz a la porta d'un ric home, ni anc non saup hom qui·l fo ni don" (was thrown at the door of a powerful man, and nobody ever knew who he was nor whence he came). In the other texts, this designation can be present either in its simpler form "he was a…" or under the filiation form, e.g. "he was the son of…".

[8] For a presentation of the concept of RS and their use, see Salem and Lafon (1983); the repeated segments were computed using the Lexico3 software, developed by the SYLED–CLA2T of University Paris 3 Sorbonne–Nouvelle. The shorter segments systematically comprised in a longer one (that is, preceded and followed always by the same two word-forms) were eliminated by the software.

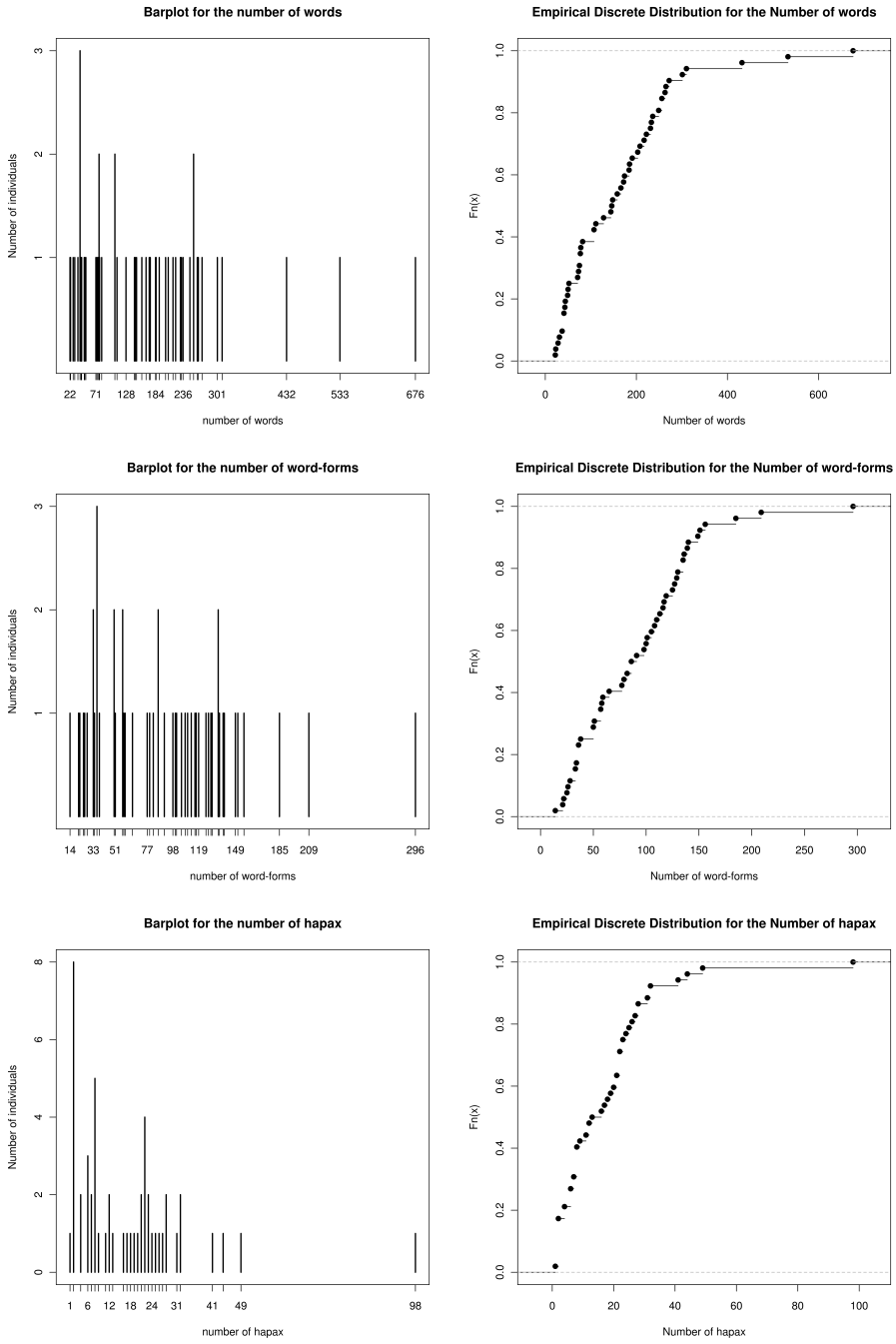


Figure 1: Bar plots and Functions of the Empirical Discrete Distribution for the Number of words, of different word-forms, and of *hapax*, for the *vidas* of A.

l	RS	Max freq.	Average freq.	Total freq.
9	1	2	2,00	2
8	7	3	2,29	16
7	10	6	2,40	24
6	23	4	2,39	55
5	52	7	2,40	125
4	131	10	2,66	348
3	313	23	3,17	992
2	691	50	4,59	3174

Table 1: Length, number of different RS, maximum, average (rounded) and total frequency, for the vidas of A.

tors. If we take a closer look at the RS, we see that in many cases, if simple graphical or flexional changes are ignored, many different RS can be attributed to the same model. For instance, the *formula* meaning "and here are written of his songs" corresponds to different RS and to segments non identified as repeated due to these variations (see Appendix, p. 23, Tab. 5), for a total frequency of 10 and a form that could be summarised as " Et (aqui—aissi) son escriutas (granren—moutas—∅) de las soas (cansos—canssos—chanssos)". The variations can be inherent to the formula itself, as for the *formula* meaning "he was from ... , from a castle called ... ".

It is very likely that lemmatisation and standardisation of the texts would reduce the number of different RS and greatly increase their frequency[9]. Finally, it is to be noted that the RS amount to roughly 60% of the texts. If the *vidas* were to be stripped from all RS, the total amount of word-forms would go from 1479 to 1433, and more importantly the total amount of words from 8519 to 3448 (a decrease of approx. 59,526%). Had the texts been lemmatised, the RS count could have still been higher[10].

In order to analyse these 52 texts, we had to choose between standardising and/or lemmatising, or working directly from the word forms. Classifying texts depending on the vocabulary they use, and thus, working on lemmatised word forms, is a common approach, useful for content or theme–based clustering, but is not always the most relevant (El-Bèze et al., 2005), and especially not here, given the lexical characteristics of the *vidas*. On the contrary, our goal was to identify all the elements that constitute clusters in the corpus: groupings due to branches of the tradition, scribes, diatopic or diachronic variations, or possible authors, and not to determine *a priori* which variations were relevant or not. This led us to work on both unlemmatised and unstandardised word forms, as any graphical variation could be an indication of the origin of a text and of its place in the manuscript tradition. For the same reasons, we chose not to *ex ante* rule out the most frequent words, including "stop words" or "function words", as variations through their usage could also be of interest regarding matters of style or morphosyntactic variations, and be helpful for author identification. In fact, studies tend to demonstrate that variation in the use of the "function words" is more relevant to authorship attribution than the study of the rarer, content–related, words (Stamatatos, 2009; Argamon and Levitan,

[9]For the most common RS, see Appendix, p. 23, Tab. 6; and for the complete set, see the online reference given in the first footnote. For a more detailed analysis of the structure of the *vidas*, the use of *formulae* and narrative patterns, see Camps (2012).

[10]It is worthy to be noted that the use of these *formulae* itself for authorship attribution of the *vidas*, attempted, manually, by Guida (1996), has been severely contradicted by Meneghetti (2002, part. p. 149).

2005), notably because their usage is less conscious and less subject to variations due to change of theme or genre. This explains why, in a second time, we have chosen to restrict our database to the one hundred most frequent word–forms (see below, from 2.2 onwards).

This is why the *vidas* have simply been transformed in a raw word frequency table[11]. Consequently, we had a total of 1480 different word forms across all 52 texts. Each text contained between 14 and 296 word forms, the mean being of 92 (fig. 1). The number of different word forms is obviously affected by the total number of words, and we can observe that they are linked by a non–linear relationship (FIG. 2)[12]. This relationship seems to fit the intuition that, as though the number of words a text may contain is virtually infinite, the vocabulary a language possesses is finite. Consequently, the curvature of this relationship is a fair indication of the amount of vocabulary that an author had at his disposal. This is particularly true in the case of the *vidas* and reveals the standardisation of their vocabulary. The abnormal individual (the same abnormal individual as in the scatter plot, the *vida* of the viscount of Saint Antonin) possesses properties that substantially differ from the other texts[13]. Had we done the non–linear regression again while ignoring the abnormal individual, we would have had an even more efficient model (dashed line on fig. 2).

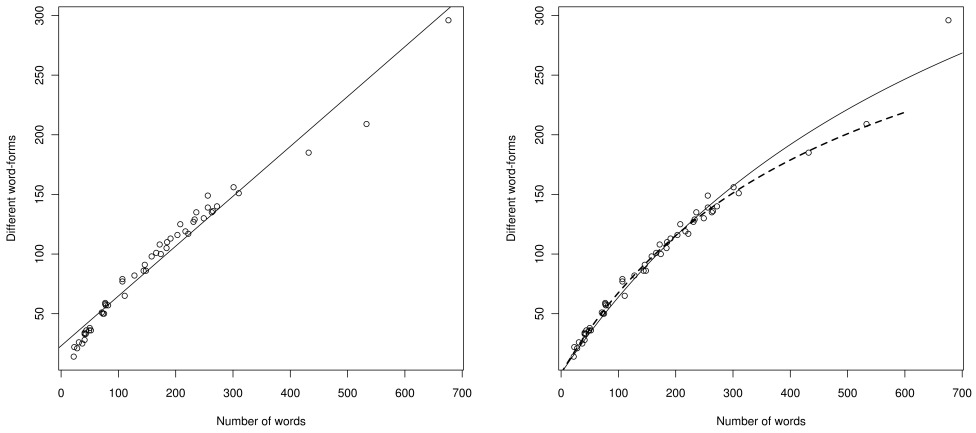


Figure 2: Linear and nonlinear regression of the number of different word forms by the total number of words

1.3 Term weighting and calibration

Two important characteristics of our corpus are the limited amount of vocabulary, a direct function of the total number of words, and the relatively important variation

[11]Informations relative to sentences' length and use of punctuation were not relevant, since both of these are conditioned by the choices of the modern editor, not the author of the text.

[12]The non–linear regression was done by minimising the non–linear least–squares using the Gauss–Newton algorithm.

[13]The wider vocabulary used by this text seems to sustain its classification in a somewhat different group than the other *vidas*, being less of a short biographical narrative and more of a developed narrative, announcing Italian XIVth century *novellas* (such as those contained in Boccaccio's *Decameron*).

in sizes between texts, a known difficulty in term of text classification and authorship attribution methods (Sanderson and Guenter, 2006). Thus, we had to seek a way to avoid the risk of seeing our clusters determined solely by the size of the texts. To solve this problem, we considered using some form of differential term weighting (Dumais, 1991), but the most popular ones (including Dumais') are built to be useful in a theme–based approach different from ours. To compensate specifically for the different text lengths, we adopted a very basic weighting (let F_{ij} be the frequency of term i in document j and T_j the total of terms in document j, $\frac{F_{ij}}{T_j}$)[14] and to compare its results with a simple centring and scaling of the variables. Results were roughly equivalent when used with Hierarchical Clustering (see FIG. 5), but not with the same errors, but our weighting was superior when used with Kohonen's Self organising maps.

Another problem came from the difficulty to judge, *ex ante* as well as *ex post*, which method would be the most helpful to divide our corpus into relevant subgroups. This is why we constituted a control sample (see table 7 in appendix), composed of texts as close as possible to the *vidas* of A: texts from the same manuscript, likely copied by the same scribe in the same period and area. It is to be noted, though, that these texts are from a different literary genre, verse in lieu of prose, and have somewhat different properties as to their lexical richness. Notwithstanding, lacking a corpus of clearly attributed *vidas*, they were our least biased alternative[15]. For this control sample to be effective we chose texts from four different troubadours. We picked them in order to have two strongly different groups (an earlier Northern Occitan group and a later Northern Italian group) so that the two troubadours in each group, while close one to the other, would still be separated from him in terms of chronology and geography. The first group is composed of the *auvergnat* Peire d'Alvernha (fl. c. 1150–1170) and the *limousin* Guiraut de Bornelh (†1215); the second group is composed of Uc de Saint Circ (†1257), a troubadour who was born in Quercy but composed most of his work in Northern Italy, and the Venetian Bertolome Zorzi (†c. 1273).

2 A first approach: Factor Analysis

2.1 Homogeneity or dimensionality?

Our first approach was to use Factor Analysis. In addition to the more usual Principal Components Analysis, we computed a Correspondence Analysis. While partly similar in its principles to Pearson's PCA, it applies the same weighting to lines and columns (the χ^2) and is particularly relevant to the study of homogeneous data sets, contingency tables in general, and frequency tables in particular.

It seemed likely that we could experience here what has been known since Bellman as the "curse of multi-dimensionality". Our corpus is composed of a small number of individuals (*i.e.* objects described by a set of data). For each individual, we study several characteristics, since each word of the frequency table is a characteristic of its text. Geometrically, this results in a distribution of very few points in a high-dimensional space, very often giving results that prove unhelpful with respect to relevance or perceptible contrast (E. Chavez and Marroquín, 2001). Yet, it is possible a technique such as Factor

[14] This was allowed by the relationship between the number of word forms and total number of words: length of the texts would still be rendered by their lexical richness, while not being the most important element of all subsequent analysis.

[15] Problems due to differences in terms of subject matter are easier to handle with authorship attribution methods, for instance by working with the most frequent word–forms ("function words").

Analysis would work. It has been shown that increasing dimensionality can even help as long as the added dimensions are relevant (Houle et al., 2010). Are the number of occurrences of each word relevant dimensions? The only way to find out was to use them in the computation and check.

Figure 3: CA and PCA on weighted word frequencies of the control sample

Unfortunately, both PCA and CA gave out unusable results. Calibration of the CA (fig. 3) has not been effective at all to illustrate which were the relevant groups in our control sample. In PCA, calibration was not effective either and the graph associated to our corpus displayed an indistinct mass of individuals centered on the first plane with two abnormal individuals, corresponding to the two longer texts, and an eigenvalues bar plot showing mostly one meaningful axis.

In order to refine this analysis, we applied different forms of *Differential Term Weighting* on the frequency table (Dumais, 1991). Yet, this method failed to increase contrast and readability[16].

2.2 Reducing dimensionality

How can one explain this failure? First, this could come from our data: the individuals could simply be "too close to call" if they form distinct and coherent subgroups. This would be in favour of the hypothesis of one single author for most of the *vidas*. But it could also come from the dimensionality problem aforementioned. In order to avoid this problem, we implemented two methods.

First, we tried to recompute our two Factor Analyses, taking into account only the top one hundred most frequent words, as is often done in authorship attribution methods. Yet, this method failed to correctly sort our control sample. To avoid dimensionality problems, we then implemented a non–metric Multidimensional Scaling (FIG. 4).

This procedure behaves quite well on our test corpus, and actually separates some groups from a core of texts in the *vidas*, which seems to show that there are indeed distinct parts in our corpus. Yet, the visualisation hardly helps us to determine what the more relevant groupings of these texts would be.

[16] A common solution, bypassing the difficulty rather than confronting it, would be to run the analysis anew ignoring these individuals. But such a method would inevitably result in a loss of information and would deprive us from a complete apprehension of the corpus.

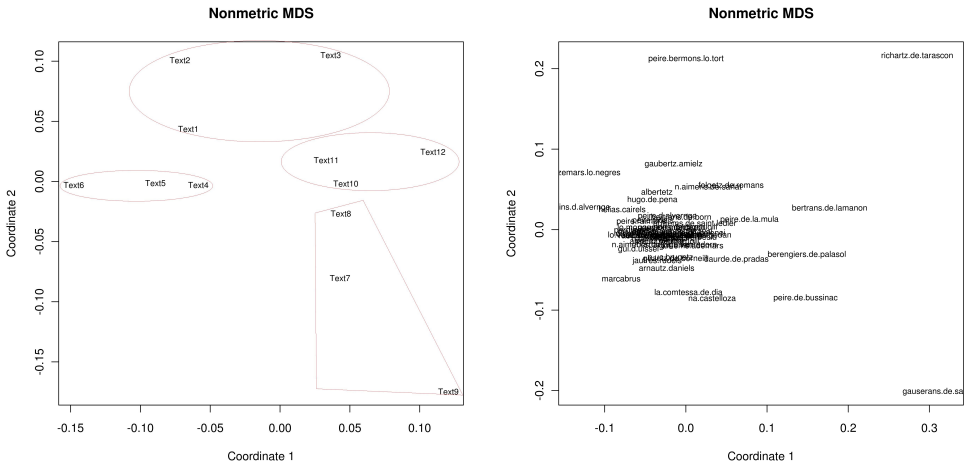


Figure 4: Non-Metric multidimensional scaling with word weighting on our control sample and on the vidas

3 Hierarchical Clustering and Kohonen's Self Organising Map

3.1 Determining an agenda

In order to obtain results on questions relative to authorship, scribes, and manuscript tradition, we needed to operate a clear–cut classification of the individuals, to establish proximities and distances between texts. This is why we sought a method that could allow us to represent these proximities and distances in an human–readable fashion, while still being able to render the hierarchy of the distances, and the abnormal individuals.

We focused however on unsupervised learning, allowing us to gather the more similar individuals in relevant subsets or "clusters", without having to be given any characteristics of what these clusters should be. The method we have chosen to use is a classic type of non–supervised learning called *hierarchical Clustering*.

In the agglomerative version of this method, each individual is aggregated to another single individual considered the "closest", according to a criterion the user has to choose. Then, the clusters formed in this first step are once again aggregated to the closest cluster, the operation being repeated until the clusters are gathered in a unique group containing all the individuals. As we do not define *a priori* constraints on what the common features of individuals belonging to a cluster should be, similarity is here understood as a mere matter of distance between individuals: the closer individuals are on a geometrical viewpoint, the more similar they are.

We then used our test corpus to understand which precise technique of Hierarchical Clustering to use. Several decisions had to be taken: which distance measure? Which linkage criterion? Do we use our term weighting? Do we use all the word forms? Should we restrain to a hundred of them instead, as has often proved effective in authorship attribution methods (Argamon and Levitan, 2005; Stamatatos, 2009)? We ran many possible

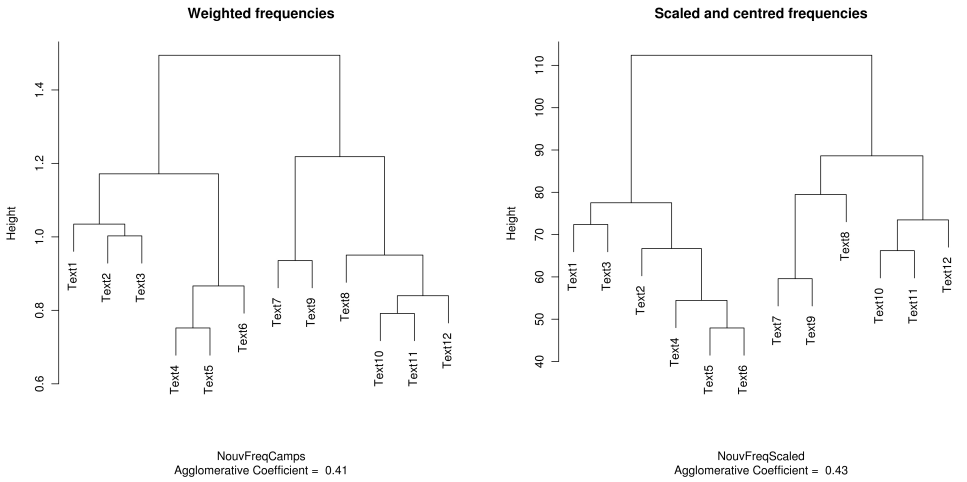


Figure 5: HC of control corpus, computed for the one hundred most frequent words, with Manhattan distances and Ward's method; left with weighted frequencies, right with scaled and centred frequencies.

combinations[17] of these decisions to see which would best divide our test corpus, the best division being to gather texts this way: $\{\{1,2,3\},\{4,5,6\}\},\{\{7,8,9\},\{10,11,12\}\}$. Our best fit was obtained when using only the one hundred most common forms, with Manhattan distance[18] , Ward's method[19] and weighted terms.

The results were then compared to those of a Self Organising Map (Kohonen, 1982)[20] and to a HC done on the prototypes corresponding to observations. We tested several grid-sizes (see.TAB. 2) and settings on our test corpus, using weighted frequencies and the one hundred most frequent word–forms. We were able to notice that the quantisation error decreased with the size of the map, ranging from 8.99e-46 for an 7×7 map, to close to 0 starting from a 16×16 map, while Kaski & Lagus' error (Kaski and Lagus, 1996) seemed to attain a minimum to then stabilise. Nonetheless, the map whose results were the most relevant were obtained with a 8×8 (quantisation error 6.06e-51) and a 10×10 map (quantisation error 1.69e-102), corresponding to a maximisation of the HC agglomerative coefficient at 0.803519 and 0.8053272 (yet, the results obtained with a 9×9 map, agglomerative coefficient of 0.7982797, showed one slight misattribution). The 8×8 map was obtained using Principal Component based initialization, standard Best Matching Unit calculation approach, a radius of 3.968489, exponential–like annealing and a gaussian kernel (see FIG. 6). The results obtained by using our weighting were better than with scaled frequencies.

[17]We tested combinations of: single-linkage, complete-linkage, weighted-average linkage, UPGMA, McQuitty and Ward's method as linkage criterion; euclidean, maximum and Manhattan as distance; with all the words, the one hundred most common and the two hundred most common; with and without our weighting.

[18]Manhattan distance is a concept coming from "taxicab geometry", where the distance between two points is the sum of the absolute differences of their coordinates: $\sum_{i=1}^{n} |x_i - y_i|$

[19]In Ward's method, an individual is not necessarily linked with the closest neighbour; pairs are formed so that they make the least increase to the error sum of squares at each step (Ward, 1963).

[20]To compute the Self Organising Map and obtain the distances between prototypes, we used the R *Yasomi* package (Rossi, 2012).

Map size	Quantisation error	Kaski & Lagus' error	HC agglomerative coefficient
7×7	8.99e-46	0.004965397	0.7460396
8×8	**6.06e-51**	**8.02318e-16**	**0.803519**
9×9	1.68e-94	2.798099e-23	0.7982797
10×10	**1.61e-102**	**3.526616e-18**	**0.8053272**
11×11	5.83e-103	4.344372e-18	0.7855996
12×12	1.05e-187	3.526616e-18	0.797328
	...		
15×15	2.64e-286	1.011429e-17	0.7979713

Table 2: Variation in sizes of the SOM's; maps whose HC delivered the expected results on the control corpus are highlighted.

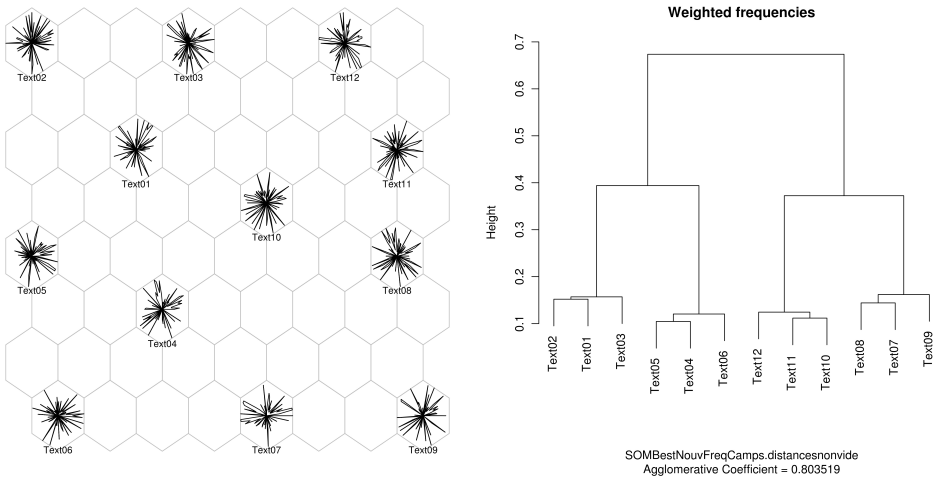


Figure 6: Kohonen's SOM based on the weighted frequencies and the one hundred most frequent word–forms and HC done with Ward's method on the prototypes corresponding to observations, labelled according to the number of the text to which they corresponded.

Map size	Quantisation error	Kaski & Lagus' error	HC agglomerative coefficient
18×18	*3.05e-43*	*0.005577446*	*0.9349739*
19×19	9.93e-51	0.001515318	0.941602
20×20	5.05e-51	0.0006961962	0.9328851
21×21	*1.24e-50*	*0.001980793*	*0.9236621*
22×22	7.59e-51	0.0006961962	0.9371617
23×23	3.41e-56	4.451419e-16	0.9341154
24×24	1.75e-93	3.764552e-18	0.9361364
25×25	2.52e-92	2.985055e-19	0.9362692
26×26	4.09e-98	2.489757e-18	0.9379868
27×27	7.10e-98	7.182769e-18	0.9403673
28×28	*1.35e-95*	*2.867811e-18*	*0.9371938*
29×29	1.45e-117	2.029302e-18	0.9408255
30×30	1.45e-156	2.072804e-18	0.939084

Table 3: Variation in sizes of the SOM's for the *vidas*. Maps that were not retained for analysis are in italics.

We then applied the same method to the text of the *vidas*. Similarly to what happened with our test corpus, quantisation error tended to decrease with the size of the map, and Kaski & Lagus's error decreased to a minimum, to then stabilise (TAB. 3). Based on the experience obtained from the control corpus, we discarded the 18×18 map where errors were too high, as well as the 21×21 and 28×28 maps that showed increased errors when compared to their neighbouring maps. The analysis was founded on the 20×20 (FIG. 8), 23×23, 24×24, 27×27, 29×29 and 30×30 maps, which main structure was robust, while the 19×19, 22×22, 25×25 and 26×26 maps, that showed changes in structure were kept for comparison purposes, along with the HC done directly on the data (FIG. 7).

3.2 Results

The main division of the corpus in two large groups, the former more important than the latter, is common to all HC's, even if the exact individuals composing these groups may vary. Much like the control sample, this division of the corpus is considerably higher than the others (2.44 height against 1.68 for the second, see FIG. 9). The possibility exists that here, as previously, this division is to be interpreted in terms of diachronic and/or diatopic difference. For these two main clusters, if we exclude the "moving" individuals (*i.e.* individuals that "move" between the different clusters in some of the maps listed above)[21], and take only into account the stable ones (*i.e.* individuals that are always in the same cluster)[22], and try to categorise the clusters according to all the word–forms of their texts (using a value test), the categorising variables with the highest probability values are, for the second cluster, the most frequent words of the core structure of the *vidas* (*si*, $p = 0.0228$, *fo*, $p = 0.0083$, *de*, $p \simeq 0$) and words related to the poetic activity which mostly occur as an introduction to the songs following the *vidas* and are also often part of the core structure of the *vidas* (*sirventes*, $p = 0.0083$, *canssos*, $p = 0.0406$, *coblas*, $p = 0.0422$). On the other hand, words characterising the first cluster denote a more complex structure (pronoun *el*, $p = 1e - 04$, determiner *lo*, $p = 3e - 04$, subordinating words *que*, $p = 0.0029$, *don*, $p = 0.0282$, preposition *per*, $p = 0.0034$) and a more developped narrative (verbs of movement *venc*, $p = 0.0075$, *anet*, $p = 0.008$, words related to time, *lonc*, $p = 0.0396$, *temps*, $p = 0.0423$). If we use a database constituted of informations relating to the dates, and the mentions of geographical origin of the troubadours, as well as some external informations concerning the tradition of the texts (notably the other manuscripts in which the texts are attested) and the supposed authors when they are know, none of these criteria are to be linked with a significant (> 0.05) probability value to one of these clusters. It is still interesting to acknowledge that 100% of the *vidas* relating to the older troubadours (first half of the XII$^{\text{th}}$ century) are in the first cluster, as well as 100% of the mention of Limousin as the geographical origin, and that 81.25% of the *vidas* of the first cluster are also to

[21] The *vidas* of Aimeric de belenoi, Aimeric de Sarlat, Arnaut Daniel, Bernart de Ventedorn, Bertolome Zorzi, Bertran de Born, Giraut de Bornelh, la Comtessa de Dia, Peire Bermon lo Tort, Pons de Capduoill, Raimons de Miraval, Sordel, Uc Brunet.

[22] For the first cluster, Aimeric de Peguillan, Albertet, Arnaut de Maruoill, Azemar lo Negre, Cadenet, Dalfin d'Alvernge, Folqet de Marseilla, Gaubert Amiel, Gaucelm Faidit, Gui d'Uissel, Guillem de Cabestain, Helias Cairel, Hugo de Pena, Jaufre Rudel, le Monges de Montaudo, le Monges Gaubertz, le Vescoms de Saint Antoni, Marcabrus, Peire d'Alvernge, Peire Raimon, Peire Rotgiers, Peire Vidals, Peirol, Perdigons, Raembautz de Vacheiras, Uc de Sain Circ; for the second, Berengier de Palasol, Bertran de Lamanon, Daurde de Pradas, Folqetz de Romans, Gauseran de Saint Ledier, Guillem Ademars, Guillem de Bergedan, Guillem de Saint Leidier, Na Castelloza, Peire de Bussinac, Peire de la Mula, Ricart de Berbesiu, Richart de Tarascon.

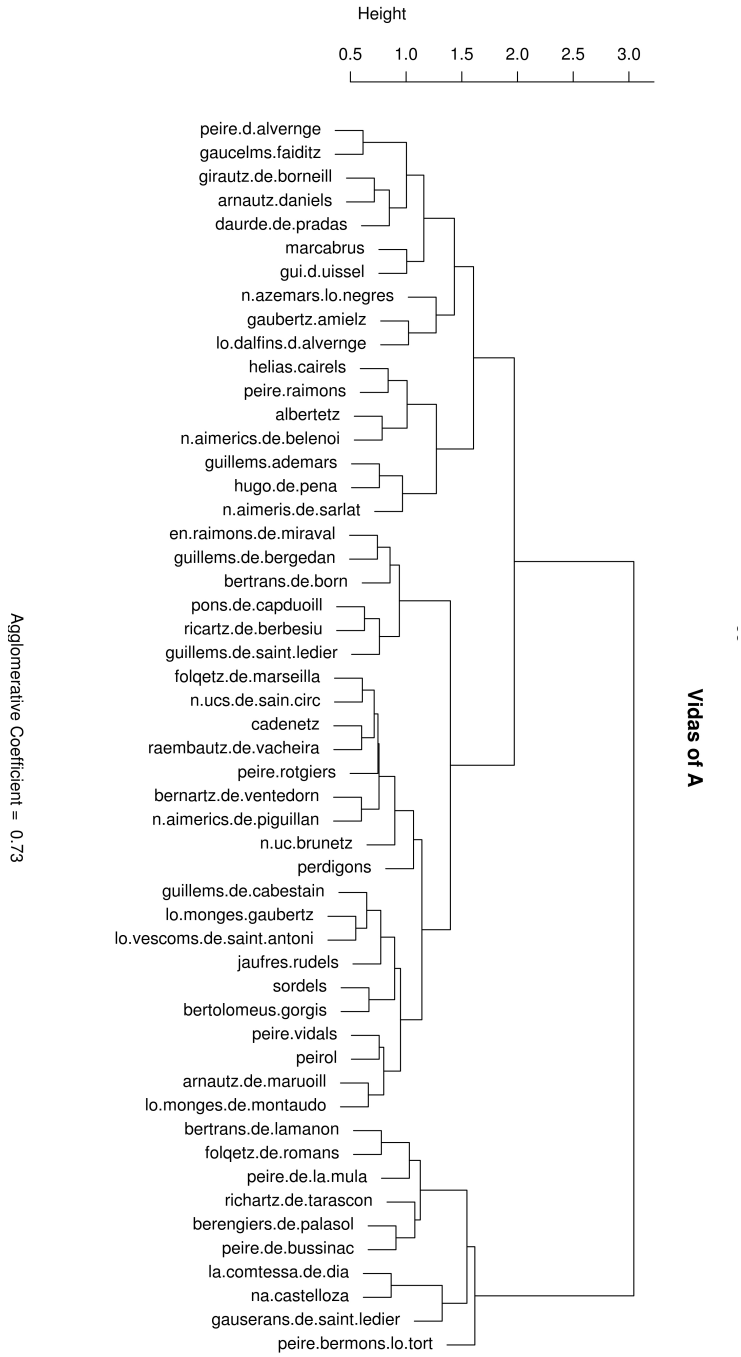


Figure 7: HC done directly on the data (one hundred most frequent word-forms), with Manhattan distances, Ward's method and weighted terms.

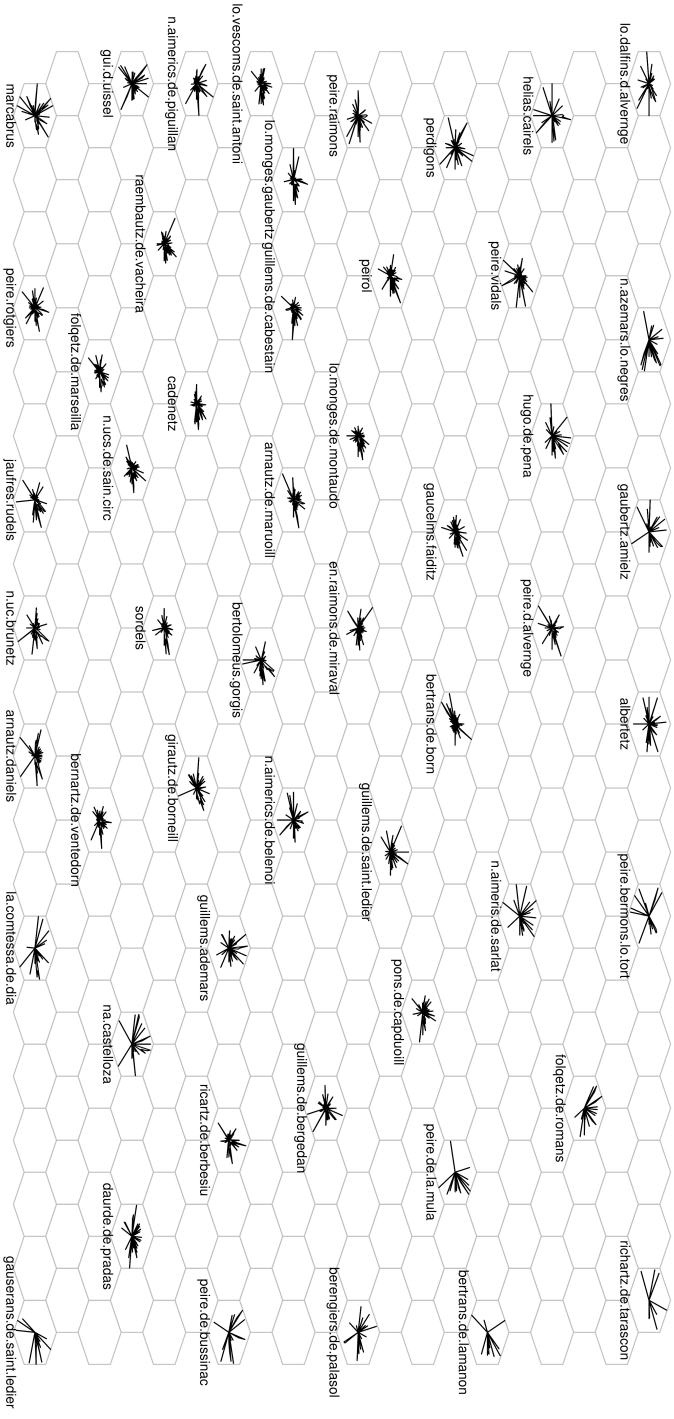
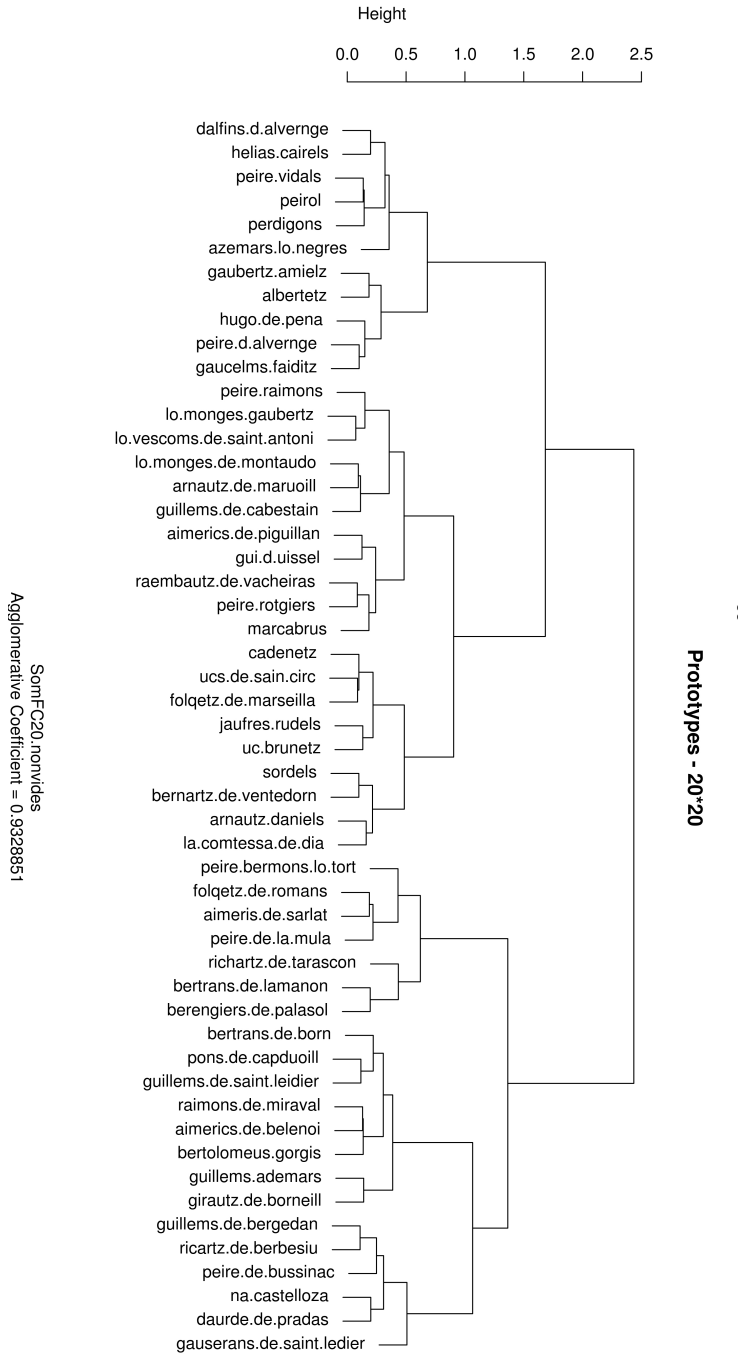


Figure 8: SOM (20×20 grid) based on weighted frequencies for the one hundred most frequent word-forms (left) and HC done on the prototypes, with Ward's method (right).


Height
0.0
0.5
1.0
1.5
2.0
2.5
dalfins.d.alvernge
helias.cairels
peire.vidals
peirol
perdigons
azemars.lo.negres
gaubertz.amielz
albertetz
hugo.de.pena
peire.d.alvernge
gaucelms.faiditz
peire.raimons
lo.monges.gaubertz
lo.vescoms.de.saint.antoni
lo.monges.de.montaudo
arnautz.de.maruoill
guillems.de.cabestain
aimerics.de.piguillan
gui.d.uissel
raembautz.de.vacheiras
peire.rotgiers
marcabrus
cadenetz
ucs.de.sain.circ
folqetz.de.marseilla
jaufres.rudels
uc.brunetz
sordels
bernartz.de.ventedorn
arnautz.daniels
la.comtessa.de.dia
peire.bermons.lo.tort
folqetz.de.romans
aimeris.de.sarlat
peire.de.la.mula
richartz.de.tarascon
bertrans.de.lamanon
berengiers.de.palasol
bertrans.de.born
pons.de.capduoill
guillems.de.saint.leidier
raimons.de.miraval
aimerics.de.belenoi
bertolomeus.gorgis
guillems.ademars
girautz.de.borneill
guillems.de.bergedan
ricartz.de.berbesiu
peire.de.bussinac
na.castelloza
daurde.de.pradas
gauserans.de.saint.ledier
SomFC20.nonvides
Agglomerative Coefficient = 0.9328851
Prototypes - 20*20

be found in the XIV$^{\text{th}}$ century Languedocian manuscripts *ER*, which also points to more important texts, with a broader circulation. Finally, the more categorising variables are actually, in favour of the first cluster, the number of different word–forms ($p = 0.0032$), the length ($p = 0.006$) and number of hapax ($p = 0.0032$). Of this can be deduced that this main division opposes, in a first cluster, longer texts with a stronger authorial value, and in a second cluster, shorter texts (sometimes quite short, 22 words for the *vida* of Richart de Tarascon), that the method fails to attribute, texts lacking clear authorship, and probably some very short *vidas* that were composed by anonymous authors only to be integrated in the manuscripts. This main division seems to tell us little about the authors, and it will be necessary to consider the lower divisions of the tree.

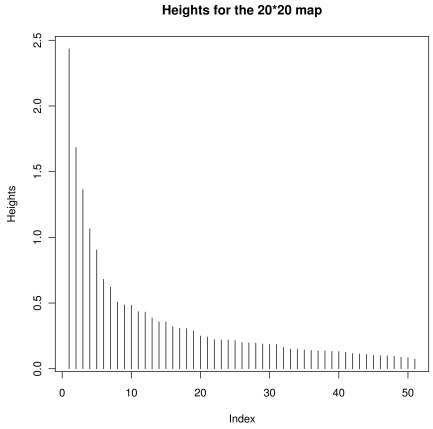


Figure 9: Heights for the CAH done on the prototypes of the 20×20 map.

The second and third divisions are a bit lower (1.68 and 1.36, FIG. 9) but still relatively high. They are found in the same fashion in all dendrograms[23] (except for one small change in the 24×24 map), resulting in four subgroups. The first subgroup is build around nine stable individuals[24] and is opposed to the second subgroup, build around twelve stable individuals[25] on several external criteria: for all the stable individuals, the second subgroup contains all the *vidas* making mentions of locations in Catalonia or Aragon, and the only *vida* in the stable subpopulation to be also present in the Catalonian manuscript *Sg*[26] and all the *vidas* for which there exists a strong and acknowledged suspicion for Uc de Saint Circ to be the author[27]. Moreover, this second subgroup

[23]That is, the tree-like diagrams illustrating the groupings done by the method at each step.

[24]Albertet, Azemar lo Negres, Dalfin d'Alvernge, Gaubert Amiel, Gaucelm Faiditz, Helias Cairel, Hugo de Pena, Peire d'Alvernge, Peire Vidal; Peirol is also almost always included in this subgroup, while Perdigon, Arnaut de Maruoill, lo Monges de Montaudo and Peire Raimon tend to move from this subgroup to the second, and Bertran de Born from this subgroup to the fourth.

[25]Aimeric de Peguilhan, Cadenet, Folqet de Marseilla, Gui d'Uissel, Guillem de Cabestain, Jaufre Rudel, lo Monges Gaubert, lo Vescoms de Saint Antoni, Marcabrus, Peire Rotgier, Raembaut de Vacheiras and Uc de Sain Circ; Uc Brunet is also almost always included in this subgroup (once in the fourth), as is Sordel; many individuals tends to move between this subgroup and the fourth (Arnaut Daniel, Bernart de Ventadorn, Bertolome Zorzi, la comtessa de Dia, Sordel, Uc Brunet).

[26]The *vida* of Raimbaut de Vaqueiras. It is to be noted that two out of the other four *vidas* of *A* also found in *Sg* are part of the moving individuals of this subgroup: nominally, Bernart de Ventedorn, Guiraut de Bornelh. Pons de Capdueil tends to move between the third and fourth subgroup as does Guilllem de Saint Leidier.

[27]The *vida* of Uc himself, among the stable individuals(Meneghetti, 2002), and the one of Bernart

contains longer and more complex *vidas* with more vocabulary than the ones of the first subgroup. The third subgroup contains only three stable individuals[28], all very short texts; the fourth subgroup also contains three[29], in both cases *vidas* of a very limited diffusion, found only in the Venetian manuscripts[30]. As was previously stated, some individuals from the third subgroup tend to move towards the first, and individuals from the fourth towards the second. It is then likely we need to discard these third and fourth unclear subgroups in order to increase the ability to interpret the results. We then need to center our analysis on the first two groups, integrating some of the moving individuals from the third and fourth with a firm anchorage in one of the two first subgroups.

If we restrain our analysis on these first two subgroups, we can define them as containing stable (S) and moving (M) individuals and being themselves constituted, in most analyses, of two different subgroups ($\{\{1a, 1b\}, \{2a, 2b\}\}$); when the belonging of an individual to a subgroup was robust, it as been indicated, otherwise not, as shown in Tab. 4 (individuals moving between those two groups themselves, and individuals only rarely included in the subgroup have been excluded – they are shown in italics and are in the table of the group to which they belong more often and in both in case of equality).

Group	Status	Name	Group	Status	Name
1a	S	Azemar lo Negre	2a	S	Guillem de Cabestain
1a	S	Dalfin d'Alvernge	2b	S	Aimeric de Pegulhan
1a	S	Helias.Cairel	2b	S	Folqet de Marseilla
1b	S	Albertet	2b	S	Jaufre Rudel
1b	S	Gaubert Amiel	2b	S	Peire Rogier
1b	S	Gaucelm Faidit	2b	S	Raimbaut de Vacheiras
1b	S	Hugo de Pena	2b	S	Uc de Sain Circ
1b	S	Peire d'Alvernge	2b	M	Benart de Ventadorn
1	S	Peire Vidal	2b	M	Sordel
1	M	Peirol	2b	M	Uc Brunet
1	*M*	*Aimeric de Belenoi*	2	S	Cadenet
1	*M*	*Aimeric de Sarlat*	2	S	Gui d'Uissel
1	*M*	*Bertran de Born*	2	S	lo Monge Gaubert
1	*M*	*Perdigon*	2	S	lo vescoms de Saint Antoni
1	*M*	*Peire Bremon lo Tort*	2	S	Marcabru
1	*M*	*Pons de Capduoill*	2	M	Bertolome Zorzi
			2	*M*	*Arnaut Daniel*
			2	*M*	*Arnaut de Marueil*
			2	*M*	*Bertran de Born*
			2	*M*	*Guiraut de Bornelh*
			2	*M*	*la Comtessa de Dia*
			2	*M*	*lo Monge de Montaudo*
			2	*M*	*Peire Raimon*
			2	*M*	*Pons de Capduoill*
			2	*M*	*Raimon de Miraval*
			2	*M*	*Rigaut de Barbezieux*

Table 4: The two first subgroups.

It is in $2b$ that we find the texts attributable to Uc with the most certainty (*vidas* of Uc, and of Bernart de Ventedorn, the last one signed in the manuscripts *IKERSg*, but not

de Ventedorn among the moving ones. The *vida* of Rigaut de Barbezieux, sometimes attributed to Uc (Guida, 1996) is not part of the second subgroup on the HC based on the SOM, but is on the HC directly based on the data, fig. 7.

[28]Bertran de Lamanon, Folqet de Roman, Richart de Tarascon.

[29]Daude de Pradas, Gauseran de Saint Ledier, Na Castelloza.

[30]The *vida* of Bertran de Lamanon is attested only in *A*, those of Richart de Tarascon, Daude de Pradas, Gauseran de Saint Ledier in *ABIK*, the one of Folqet de Roman in *AHIK*, and for Na Castelloza in *AIK*.

in *A*, by Uc de Saint Circ), and part of the question is to know if this presumption of authorship can be extended to 2*a* and to undecided individuals. Several *vidas* of this cluster share undeniable similarities both in stylistic and factual terms: the *vidas* of Aimerics de Peguilhan, Bernart de Ventadorn, Cadenetz, Folquet De Marseilla, Peire Rogier, Raimbaut de Vacqueyras, Sordel, and Uc de Saint Circ. In stylistic terms, these *vidas* are distinguished by the use of the first person (singular), elsewhere seldomly present, along with the mention of direct testimony[31]; also, the *senhal* of the troubadours and their ladies is quite often quoted[32]. On a more factual level, theses texts are characterised by a good level of information about the concerned troubadours, in terms of geography[33] or persons[34]. Moreover, these *vidas* seem to be mostly aware of the trips of the troubadours to Spain or Italy and to their participation in courts such as the court of Savaric de Mauleon, of the king Alphonse VIII of Castilla, or of the count Raimon V of Toulouse, even when this is not mentioned in the extant works of the troubadour[35]. These texts

[31] *Vidas* of Folquet("q'ieu vos ditz"), Cadenetz ("E tot lo sieu faich eu saubi per vertat per auzir et per vezer"), Bernart de Ventadorn ("E tot so qu'ieu vos ai dich de lui si me comtet e·m dis lo vescoms n'Ebles de Ventedorn, que fo fills de la vescomtessa q'en Bernartz amet tant"). It is also present in the *vida* of Raimbaut d'Aurenga, for which a space was kept in ms. *A*, but was never copied and is kept only by N^2, and in the *vida* of Gausbert Amiel, placed by the clustering in the first group, as will be discussed later. It is to be noted that the first person is used by the author, very likely Uc de Saint Circ, of the *vida* of Savaric de Mauleon, attested in mss *IK*.

[32] *Vidas* of Cadenetz, "e fasia se apellar Bagas"; Raimbaut de Vaqueiras, "Et apellava la en sas chanssos *mos bels cavalliers*"; Sordel, "et apellava la en los sieus chantars que el fazia per lieis *doussa enemia*". It is also present in the *vida* of Raimon de Miraval, "si fon mout honratz e tengutz en car per lo bon comte Raimon de Tolosa, que·l clamava *n'Audiartz* et el lui" and Rigaut de Berbeziu, whose *vida* is sometimes attributed to Uc (Guida, 1996), but was placed by some of our analyses in the fourth cluster, "et apellava la *Mieils de dompna* en totz sos chantars". One could also add the *vida* of Peire Rogier, that contains the mention "E la clamava *Tort-n-avez*" in other mss but not in *A*.

[33] The *vida* of Folquet de Marseilla mentions a "rica abadia que es en Proenssa, que a nom lo Terondet", informations that are "précises et exactes" (Stroński, 1910); the *vida* of Uc de Saint Circ, a burg named "Tegra", the castle of "Sain Circ", "qu'es a pe de Sainta Maria de Rocamaor, qe fo destruiz e derrochatz per gerra"; the *vida* of Cadenet, a castle "que a nom Cadenet, qu·is en la riba de Durenssa (. . .) destruitz e raubatz per la gen del comte de Tolosa" – "Les ruines d'un vieux château dominent aujourd'hui encore la localité de Cadenet (ch.-l. de cant. de l'arrond. d'Apt, Vaucluse), sise non pas "sur la rive", mais à quelque distance au N. de cette rivière" (Boutière and Schutz, 1964, p. 502) and the event in question can be identified as taking place between 1166 et 1176, during the fights between Raimon V de Toulouse and Alfonse II d'Aragon. The *vida* of Gausbert de Puycibot also mentions a monastery of "Sain Leonart", vraisemblablement "Saint-Léonard-des-Chaumes (près de La Rochelle) dont Savaric de Mauléon (. . .) fut précisément un des bienfaiteurs"; precise informations can also be found here and there in this group, notably in the *vida* of Sordel.

[34] The name of the father of many troubadour is given (*vidas* of Folquet de Marseilla, Uc de Saint Circ, Raimbaut de Vacheiras); the "adoptive father" of Cadenet is also mentioned as "un cavallier que avia nom Guillem del Lantar", likely Guillaume Hunaud de Lantar (. . .) mort en 1222" (Boutière and Schutz, 1964, p. 502). The name of the lady and her husband is also mentioned here and there, as is in the *vida* of Raimbaut de Vacheiras "ma dompna Biatriz, qe fo moiller d'en Haenric del Carret".

[35] Trips to "Peitieus", "Gascoingna", Catalonia, Aragon and Espaigna" are mentioned for Uc de Saint Circ, as well as a stay "en Proenssa" and a trip to "Lombardia et en la Marca" and "en Tervisana" (along with other precisions); for Raimbaut de Vacheiras, his trips to Montferrat and "Romania", as the trips of Americ de Peguilhan to Catalonia and Lombardia, or the trip of Gausbert de Puycibot to Spain and more modestly the trip of Cadenet to Provence; are also mentioned, not exhaustively, stays in the courts of Savaric de Mauleon (Uc de Saint Circ, Gausbert de Puycibot), Guillem de Berguedan (Aimeric de Peguilhan – also mentioned in the *vida* of Guilhem de Berguedan himself, in cluster 4 but perhaps to be included in this one), Alphonse VIII of Castilla (Folquet de Marseilla, Uc de Saint Circ, Peire Rogier, Aimeric de Peguilhan), Alphonse IX de Léon (Uc de Saint Circ), Raimon V of Toulouse (Folquet de Marseilla – not mentioned in his extant poems –, Peire Rogier), Richard I (Folquet de Marseilla). Other minor courts are mentioned here and there. For older courts, the court of Alienor, for instance, is mentioned for Bernart de Ventadorn and the court of the princes of Aurenga is mentioned for Peire Rogier or Raimbaut de Vaqueiras. Detailed informations of this nature are also present in the *vida* of Sordel.

can be dated according to internal criteria, in the state reflected by ms. *A*, from after 1238 (Cadenet entrance in the Order of the Hospital, which is mentioned) and perhaps before 1257 (death of Uc de Saint Circ, not mentioned), a period covering in good parts Uc's supposed period of literary activity in Italy (c. 1220-1253). The attribution to Uc de Saint Circ of these *vidas* could be proposed with a reasonable doubt.

This attribution might be a bit less obvious for the *vidas* of Gui d'Uissel, Marcabru and Uc Brunet, even though they share some similarities with the previous ones[36]. A special consideration must be given to the *vidas* of Jaufre Rudel, Guilhem de Cabestanh, Raimon Jordan *lo vescoms de Saint Antonin* (and perhaps even Gausbert de Puycibot). These quite long and developed *vidas* seem to focus on a narration, whose historical accuracy is questionable, but which develops courteous patterns: *amor de lonh* of Jaufre and the countess of Tripoli, "martyrdom" of Guilhem de Cabestanh and his lover, love between Raimon Jordan and the countess of Pena, impeached by exterior factors. More romanced and longer, these *vidas* tend to draw closer to *novellas* than to biography, and could perhaps be a bit later in date, though it is not completely impossible to integrate them in the group build around Uc. Secondly, the presence, in some of the analyses, of the *vida* of Bertolome Zorzi, which cannot possibly be attributed to Uc, in this cluster, is worthy of some explanations. This seems to be caused by similarities between the *vida* of Bertolome and the one of Sordel, with which it is quite often coupled. Moreover, these two troubadours are close geographically and chronologically to Uc. Finally, a few words deserve to be said about some of the *vidas* included in cluster 4 – which individuals, as stated, tend to move towards cluster 2 – and the moving individuals excluded from the analysis in a first time, and for this a look directly at the SOM, as well as on the HC done directly on data might be useful. One might wonder, in fact, if some *vidas* of cluster 4 ought not to be integrated amongst those of this second cluster.

The groups 1*a* can be dated after 1222 (date of the death of Elias Cairel, which is mentioned here) and perhaps before 1235 (unmentioned death of Dalfin d'Alvernhe); the group 1*b* can be dated after 1221 (death of Albertet) and perhaps before the deaths of Gaubert Amiel or Uc de Pena. The texts of these groups also have some similarities in stylistic and historic terms. On what concerns the stylistic aspects, they are slightly shorter texts, with less details, and more notably, they do not integrate developed courteous narrative patterns or love stories (though weddings are mentioned from time to time)[37]. Instead, those texts focus on the aspects relevant to lyric composition, with a predilection for judgement on the quality of *motz* and *sons* and of the mesure[38]. Whether this aspects are to be attributed to a different focus of the author, to a chronological evolution, or simply to the fact that most troubadours of this groups are actually *joglars*

[36]The *vida* of Marcabru mentions Aldric de Vilar as is adopted father, and gives the surname "Panperdut" – informations probably extracted from [BdT 16b,1] and [BdT 293,43]. The *vida* of Uc Brunenc mentions, though not nominally, the king of Aragon and the count of Toulouse, as well as Bernart d'Andusa and Dalfin d'Alvernha. The *vida* of Gui d'Ussel, even though it gives the names of Gui's brothers and of "Margarita d'Albusson, moiller del vescomte d'Albusson" and mentions the papal interdiction made by Pierre de Castelnau ("lo legatz del papa li fetz jurar que mais non fezes chanssos"), fails to mention his death c. 1225; moreover, on the HC done directly on the data, Gui d'Uissel is integrated in cluster 1 — none of this being, of course, absolutely definitive.

[37]A narrative pattern common to two of these *vidas* is the fact to have lost all of its possessions by an excessive inclination towards games of chance: "E fetz se joglars per ochaion que el perdet a joc tot son aver" (Gaucelm Faidit), "mas el fo grans baratiers de jogar en taverna per q'el fo ades paubres e ses arnes" (Uc de Pena).

[38]For instance "ben escrivia motz e sons" (Elias Cairel), "Albertetz si fetz assatz chanssos que agrons bons sons e motz de pauca valenssa" (Albertet), "e si fetz los sieus vers plus amesuratz que hom q'ieu anc trobes mais" (Gaubert Amiel), "fetz mout bos sons e bons motz" (Gaucelm Faidit), "aquel que fetz los meillors sons de vers que anc fosson faichs" (Peire d'Alvernha).

(*jongleurs*) can be argued. It is also to be noted that the first person singular seems to be used twice by the author in this group[39]. On a more historical point of view, an important element is the mention of Dalfin d'Alvernhe in several of the texts[40]. The case of the *vidas* of Peire Vidal and Peirol (and perhaps even of Peire d'Alvernha) seems worthy of some further estimation on their belonging to this group: even if their integration in it seems confirmed by the HC done on the SOM, they were both included in group 2 by the HC done directly on the data, and have in fact some similarities with the second group and some apparent dissimilarities with the texts of the first group.

Different hypotheses can be formulated regarding this cluster. Perhaps are we to look for a different author. Why not then Uc de Pena, whose *vida* is included in this cluster, and is perhaps an author of *vidas* himself. Formulated by Boutière and Schutz (1964, p. 258), the hypothesis that Uc de Pena could be an author of *vidas* rests on a sentence found in his own *vida*: "e saup granren de las autrui chanssos e sabia mout las generatios dels grans homes d'aqellas encontradas", which is to be compared with Uc de Saint Circ *vida*'s "el apres tenssos e cansos e vers e sirventes e coblas, e·ls faitz e·ls ditz del valens homes que eron adoncs, ni que eron estat denan; et ab aqest saber el s'en joglari". Very few is otherwise know of this troubadour, and even his dates are subject to caution. The fact that his song [BdT 456,2] is imitated by the troubadour Palais gives us a *terminus ante quem*, except for the fact that the dates of Palais are also partially uncertain. If we follow Guida (2006), Palais, who is to be identified with Andrian de Palais, would have composed his songs in the end of the XII$^{\text{th}}$ and the first decades of the XIII$^{\text{th}}$ century (before 1235 for BdT 315,2), and was still alive in 1228. In this case, it would not be impossible to link the supposed date of the *vidas* of this group (after 1222) with the supposed literary activity of Uc de Pena. Moreover, some of the troubadours of this cluster are amongst the known inspirators of Uc de Pena (Gaucelm Faidit for instance, see Cura Curà (2007, p. 12 *et passim*)). On the other hand, nothing is known of a connection of Uc de Pena and the Dalfin d'Alvernha (contrarily to the established connection between Dalfin and Uc de Saint Circ), his *vida* seems to vague for an autobiography, and, finally, Uc de Pena has also been dated differently, by Cura Curà (2007, p. 10), who dates him c. 1248-1283, and notes concerning Boutière's hypothesis that "Si tratta di suggestione che, data la fragilità dell'ipotesi, non è stata ripresa in altri interventi sulla paternità delle biografie trobadoriche".

An other hypothesis could emphasise a chronological difference between those texts and the one found in the second cluster: could the opposition between this cluster and the second be caused by chronological and/or geographical differences more than authorial changes? This could explain the differences of datation, as well as some of the stylistical and historical differences. As for the second group, some individuals from the third cluster should perhaps be integrated to this group, as they are on the HC done directly on the data.

[39]The *vida* of Gausbert Amiel has, in *A*, "fetz los sieus vers plus amesuratz que hom qu'ieu anc trobes mais" (there was in his verses more mesure than in any man I could ever find since), but *IK* have there instead (a reading prefered by Boutière) "e fetz los sieus vers plus (a)mesuratz de hom que anc mais trobes" (there was in his verses more mesure than in any man that has since composed poetry), and the double meaning of the verb *trobar* ("find" and also "compose poetry") as well as the absence of difference between the first and third person of the subjonctive imperfect of the verbs in *-ar* does not allow in itself for a choice between the two; the *vida* of Peire d'Alvernha has a first person along with a mention of testimony, "segon qu·m dis lo Dalfins d'Alverngе en cui terra el nasquet".

[40]To the *vida* of Dalfin himself should be added the mention of the testimony of Dalfin in the *vida* of Peire d'Alvernha, and, in the *vida* of Peirol (undecided between 1*a* and 1*b*, and of which the belonging to this cluster will be discussed), the stay of Peirol at Dalfin's court and his love for Dalfin's sister, Sail de Claustra.

Though cutting the dendrogram at an uniform arbitrary height is customary, the tests on the control corpus tend to demonstrate that it is not always the preferable choice — especially when precise distinctions between clusters are needed — and that the observation of the shape of the dendrogram itself, whenever allowed by a limited amount of individuals, may be a more relevant solution. In any case, one should not be satisfied by the analysis of the clusters generated by a single cut of the dendrogram. As different divisions may contain different informations (perhaps even translate different dimensions), one should probably try to analyse both the main divisions of the population up above and the more consistent subgroups in the bottom. As such, in the case of homogeneous corpus with few individuals and possibly a lot of variables, the heuristic value of Hierarchical Clustering can be put forward in comparison to other, more widespread, methods. In particular, in the study of the "genetics" of texts, the way the divisions inside the corpus are showed seems easier to read and better adapted to the nature of the data than most other clustering methods.

References

Argamon, S. and S. Levitan (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*.

Boutière, J. and A. Schutz (1964). *Biographies des Troubadours. Textes provençaux des XIII^e^ et XIV^e^ siècles*. Paris: A.G. Nizet. 2^e^ édition.

Burgwinkle, W. (1999). The *chansonniers* as books. In *The Troubadours: An Introduction. Ed. Simon Gaunt and Sarah Kay*, pp. 246–262. Cambridge University Press.

Camps, J. B. (2012). Peirols si fo uns paubres cavalliers : lire les troubadours dans les chansonniers A, I et K. *Medioevo romanzo 36*(2), [to be published].

Cura Curà, G. (2007). Le canzoni del trovatore Uc de Pena. *Critica del testo 10*(2), 9–45.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers 23*(2), 229–236.

E. Chavez, B. Navarro, R. B. and J. Marroquín (2001). Searching in metric spaces. *ACM Computing Surveys 33*, 273–321.

El-Bèze, M., J. Torres-Moreno, and F. Béchet (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. In *Traitement Automatique des Langues Naturelles (TALN 2005) – Atelier DEFT'05*, Volume 2, pp. 125–134.

Favati, G. (1961). *Le Biografie trovadoriche : testi provenziali dei secc. XIII et XIV*. Bologne: Libreria Antiquaria Palmaverde.

Gaunt, S. (2003). Le cœur a ses raisons": Guillem de Cabestanh et l'évolution du thème du cœur mangé'. In *Scène, évolution, sort de la langue et de la littérature d'oc: actes du Septième Congrès International de l'Association Internationale d'Études Occitanes, Reggio Calabria-Messina, 7-13 juillet 2002*, pp. 363–73.

Gaunt, S. (2006). *Love and death in medieval French and Occitan courtly literature : martyrs to love*. Oxford ; New York: Oxford University Press.

Guida, S. (1996). Ricerche sull'attività biografica di Uc de Saint Circ. In *Primi approcci a Uc de Saint Circ*, pp. 75–144. Rubbetino: Soveria Mannelli.

Guida, S. (2006). (Andrian de) Palais, trovatore lombardo? In *Studi di filologia romanza offerti a Valeria Bertolucci Pizzorusso*, Volume 1, pp. 685–721. Pisa: Pacini Editore.

Houle, M., H. Kriegel, P. Kröger, E. Schubert, and A. Zimek (2010). Can Shared-Neighbor distances defeat the curse of dimensionality? In *Proceedings of the 21st International Conference on Scientific and Statistical Database Management*, Heidelberg, Germany, pp. 482–500. Springer.

Kaski, S. and K. Lagus (1996). Comparing self-organizing maps. In *Artificial Neural Networks—ICANN 96*, pp. 809–814.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics 43*(1), 59–69.

Lemaître, J. and F. Vielliard (2008). *Portraits de troubadours : initiales du chansonnier provençal* A *(Biblioteca Apostolica Vaticana, Vat. Lat. 5232)*. Città del Vaticano: Biblioteca Apostolica Vaticana.

Meneghetti, M. (1992). *Il Pubblico dei Trovatori, ricezione e riuso dei testi lirici cortesi fino al XIV secolo*. Turin: G. Einaudi.

Meneghetti, M. (2002). Uc e gli altri : sulla paternità delle biografie trobadoriche. In *Il Racconto nel Medioevo romanzo*, Bologna, pp. 147–162. Pàtron.

Panvini, B. (1952). *Le Biografie provenzali. Valore e attendibilità*. Firenze: L. S. Olschki.

Rossi, F. (2012). An introduction to YASOMI (Yet another self organising map implementation).

Salem, A. and P. Lafon (1983). L'inventaire des segments répétés d'un texte. *Mots 6*(1), 161–177.

Sanderson, C. and S. Guenter (2006). Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP'06, Stroudsburg, PA, USA, pp. 482–491. Association for Computational Linguistics.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology 60*(3), 538–556.

Stroński, S. (Ed.) (1910). *Le troubadour Folquet de Marseille : édition critique, précédée d'une étude biographique et littéraire et suivie d'une traduction, d'un commentaire historique, de notes, et d'un glossaire*. Cracovie: Académie des Sciences - Ed. du fonds Oslawski.

Ward, Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Wilson Poe, E. (1984). *From Poetry to Prose in Old Provençal: The Emergence of the* Vidas*, the* Razos *and the* Razos de Trobar. Birmingham: Summa publications.

Appendix

length	form.	freq.	Total freq.
9	Et aqui son escriutas granren de las soas chanssos	1	
9	Et aissi son escriutas granren de las soas chanssos	1	
9	Et aqui son escriutas moutas de las soas chanssos	1	
8	Et aqui son escriutas de las soas chanssos	3	
8	Et aqui son escriutas de las soas canssos	3	
8	Et aqui son escriutas de la soas cansos	1	

Table 5: Various forms of the same formula.

length	form.	freq.
9	hom fo de letras e de sen natural E	2
8	Et aqui son escriutas de las soas canssos	3
8	Et aqui son escriutas de las soas chanssos	3
8	de l abadia d Orllac E l abas	2
8	d Alvernge d un chastel que a nom	2
8	d amor e de dompnei e de totz	2
8	hom de letras e de sen natural e	2
8	pois el se rendet a l orden de	2
7	Et aqui son escriutas de las soas	6
7	de grans bens e de grans mals	2
7	de l evescat de Peiregos d un	2
7	en trobar et en chantar et en	2
7	si fo de Tolosa fills d un	2
7	fetz se joglars e trobet bonas chanssos	2
7	son escriutas granren de las soas chanssos	2
7	amor e de dompnei e de totz	2
7	totz temps volia qu il agesson gerra	2
7	Alvernge d un chastel que a nom	2
6	de letras e de sen natural	4
6	d un castel que a nom	4
6	si fo de l evescat de	4
6	d un chastel que a nom	3
6	fo avinens hom de la persona	3
6	se rendet a l orden de	3
6	e fo fills d un borzes	2
6	e lai el definet e moric	2
6	en Espaigna ab lo bon rei	2
6	d en Guillem de Saint Leidier	2
5	Et aqui son escriutas de	7
5	fo de l evescat de	5
5	se rendet a l orden	4
5	un castel que a nom	4
5	avinens hom de la persona	4
5	e fetz se joglars e	3
5	de l evescat de Peiregos	3
5	fetz se joglars et anet	3
5	Et enamoret se d una	3
5	un chastel que a nom	3
4	de l evescat de	10
4	de las soas chanssos	9
4	Et aqui son escriutas	9
4	d en Guillem de	6
4	fetz se joglars e	6
4	d un castel que	5
4	el s en anet	5
4	si fo de l	5
4	e fetz se joglars	4
3	si fo de	23
3	de las soas	14
3	que a nom	13
3	las soas chanssos	12
3	de l evescat	11
3	que avia nom	11
3	fills d un	10
3	e lai el	9
3	el s en	9
3	fetz se joglars	9
2	si fo	50
2	e de	47
2	de la	43
2	d un	41
2	q el	36
2	e l	32
2	s en	32
2	e fetz	31
2	fo de	30
2	d en	26

Table 6: The 10 most common RS, for each length.

Number	**Incipit**	**BdT**	**Attribution in ms.**
1	Bella mes la flors daiguilen	323.5	Peire d'Alvernha
2	En estiu qan cridal iais	323.17	Peire d'Alvernha
3	Abans qeil blanc puoi sion uert.	323.1	Peire d'Alvernha
4	Er auziretz encabalitz chantars	242.17	Guiraut de Bornelh
5	Ben mera bels chantars	242.20	Guiraut de Bornelh
6	A ben chantar couen amars	242.1	Guiraut de Bornelh
7	Nuills hom no sap damic tro la perdut.	457.26	Uc de Saint-Circ
8	Anc enemics qieu agues.	457.3	Uc de Saint-Circ
9	Seruit aurai longamen.	457.34	Uc de Saint-Circ
10	Pron si deu mais penssar almieu semblan.	74.14	Bertolome Zorzi
11	Entre totz mos cossiriers.	74.5	Bertolome Zorzi
12	Mout fai sobreira foillia.	74.9	Bertolome Zorzi

Table 7: The control sample