Archeologia e Calcolatori
Supplemento 6, 2014, 191-202

THE AUTHOR'S FINGERPRINT. A COMPUTERISED ATTRIBUTION METHOD

1. Stylometry and computerised attribution methods

This research originates from the consideration, initiated less than a decade ago, of the possible interaction between Philology and Information Theory (Canettieri *et al.* 2005, 2006, 2008; Canettieri 2012). Its purpose is to propose systematically a Unified Theory of the Text (UTT), whose application consists of the possibility of measuring the distance existing between two or more texts on different levels: graphic, phonetic, morphologic, and semantic. On a theoretic level, for each one of these scales the range of distance fluctuates from 0 to 1, where 0 indicates that two texts are identical and 1 that they are completely different from each other.

Based on the he UTT it is thus possible to evaluate, with a single automated operation:

1) the distance existing between manuscripts or copies of a same work (taking both the errors and the variables into consideration and providing them with differentiated weights);
2) the distance existing between texts of a same author, extrapolated from the same or from different works (authorship intertextuality or intratextuality);
3) the distance existing between texts belonging to the same genre, school or poetic movement;
4) the distance existing between different works that cannot be associated with each other in an intertextual manner but present a certain degree of intertextual relationship.

The values that can be obtained are various and articulated, but each homogeneous group of results can be visually represented by means of a graphic tree structure where the textual objects that present a certain similarity with each other are gathered in clusters and where the similarities gradually decrease while proceeding from the tree leaves up to its roots.

An example of representation is offered by the Italian Duecento Poetry tree, which has been variously reproduced on various occasions in the works indicated above and for which it is possible to visualise the integral version online. Moreover, it is not a coincidence that, within the UTT frame, the computerised procedures based on the basics of Information Theory -which have been adapted by Roman Jakobson for the linguistic and literary subject, heuristically appear to be the most efficient (Jakobson 1960; Van de Walle 2008). Such procedures provide, in effect, by definition discriminating elements in the gathering of texts based of their distance.

The UTT finds practical confirmation and application in the field of authorial attribution, where the interaction between scientists in text, mind, physics and mathematics may offer important contributions to the community. I take the liberty to refer to the work that I consider the most relevant of my "philological" works, the debate held during the trial for the homicide of Massimo D'Antona, which can now be listened on Radio Radicale's website and to which I took part as an expert witness for the defence: in such case, the study of the attribution problems allowed the full acquittal before of the Court, after nine month of imprisonment, of one on the accused subjects, who had been implicated in this trial by an attributive appraisal based on an erroneous methodology. In fact, dealing with anonymous documents or with works of uncertain attribution represents a great, important and useful challenge for philologists. The attribution process has been described as «the operation, which culminates in a critical judgement and which aims at, in the absence of or to completion of or to check the historical records, the identification of the author of an artefact, that is the assignment of the authorship of an anonymous work to a specific producer or the identification of the historic-cultural environment in which this artifact was conceived and produced» (Besomi, Caruso 1992).

Such procedure entails the confrontation with a series of problems, as for example:

1) a text is attributed to different authors by different witnesses (contrasting attributions);
2) a text has been transmitted in an anonymous form and it is necessary to ascertain its author;
3) a text has been attributed to an author, but there are strong doubts that it might have been written by someone else and it is thus necessary to ascertain whether it is authentic or apocryphal;
4) a text attributed to an author might contain interpolations of one or more different authors and it is thus necessary to understand which sections of the text were interpolated and to whom they ought to be attributed.

In an attempt to resolve such problems, philologists have resorted to traditional tools of textual analysis, first of all, in regard to antique texts, the *stemma codicum*, used to identify the most reliable witnesses or the point in the tradition where the attributive disturbance occurred: it is clear that the correct attribution, in terms of stemmatics, is related to the selection of variables and cannot be founded on the amount of the conveyed manuscripts but rather on their critical quality. Furthermore, when the *stemma codicum* consists of two branches, the choice would generally go to the less famous author, since the condition of minor notoriety has the same value as the concept of *lectio difficilior* in textual reconstruction. The aetiology of the attributive error has also been analytically described in some *ad hoc* essays (Pulsoni 2001).

In addition to the stemma, in the field of traditional attributive science, other criteria are used, which are usually classified in "external" and "internal" criteria (Erdman, Fogel 1966; Contini 1984): «Attribution studies distinguish conventionally between internal and external evidence. Broadly, internal evidence is that from the work itself and external evidence that from social world within which the work is created, promulgated and read; […] External evidence […] covers the following kinds: (1) Contemporary attributions contained in incipits, explicits, titles, and from documents purporting to impart information about the circumstances of composition […]; (2) Biographical evidence, which would include information about a putative author's allegiances, whereabouts, dates, personal ties, and politic and religious affiliations; (3) The history of earlier attributions of the work and the circumstances under which they were made. Internal evidence […] covers (1) Stylistic evidence; (2) Self-reference and self-presentation within the work; (3) Evidence from the themes, ideas, beliefs and conceptions of genre manifested in the work» (Love 2002, 51).

The first ones, which are by nature substantially related to the content, assist in evaluating whether the information contained in the text correspond to historical data: they include evidences from other authors, historic-cultural or biographical references within the text, and analysis of the sources. The internal evidences, on the other hand, include formal aspects such as rhetorical, metric, stylistic and intertextual analysis. Well-known is the method for the attribution of artistic works implemented by Giovanni Morelli and based on the internal criteria of stylistic particulars (ears, hands, folds of the clothes, etc.), that, according to Morelli, would have brought to the individualisation of the style of a specific artist, distinguishing him from his imitators (Ginzburg 1979, 57-106).

Among the internal criteria is also included the stylometry, i.e. the quantitative and statistical study of the literary style, in the present research intended as «all the formal features characterising (in sum or in a particular moment) the expressive manner of an individual or the writing manner of an author», therefore meaning the expressive and creative features that are typical of each individual rather than «all the formal features characterising a group of works, constituted on a typological or historical basis» (Segre 1985). Stylometry is based on two premises: in the first place, it should be possible to quantify, and thus to measure, the stylistic features of a text; in the second place, the texts of the same author should present similar features with each other. For example, if the suspect exists that an author did actually not write a text that is traditionally attributed to him, both such text and the other texts written by the same author could be analysed and compared through different attribution methods. In the event that an important statistical difference between the text under study and other texts certainly written by this author exists, this could represent the evidence that the uncertain document has not been written by the hand that wrote the other documents.

Stylometry is a procedure of analysis that, at this point, is extensively applied to literature: since it aims at identifying, measuring and confronting the features of what we call style, it proposes to decompose the text in order to understand which stylistic features distinguish a work and its author from other works and other authors. In the stylometric procedure the data interpretation follows an analytic stage that includes, along with the usual tools, also the statistical representation and the comparison of the numerically elaborated results. The numerical comparability allows a direct and precise confrontation of texts, authors and passages of a same work or pertaining to the same author. The proposals that followed each other over time suggested the analysis of different graphic/linguistic elements, ranging from the average sentence length, expressed in number of words, characters, or syllables, to the study of the vocabulary, by counting the average length of words, expressed in characters, the average number of syllables per word, the frequency of monosyllabic words, the frequency of empty words (that thus do not depend on the content, as in English the words *a, all, an, and, any, as, but, by, in, it, no, not, of, that, the, to, up, upon, with, without*).

Some other stylistic features were considered to be significant, such as the percentage of each different part of the discourse (nouns, verbs, adjectives, etc.). Recently, the measurement method of the "intertextual distance", introduced by Labbé and Labbé, has raised a passionate debate in France. After realising a complete lemmatisation of the texts under study, the two scholars have analysed the distance occurring between the obtained dictionaries: the shorter the distance, the greater the possibility that the two texts are attributable to the same author, or belong to the same literary genre or to the same period, or relate to the same subject (Labbé, Labbé 2001). Some statistical models have been developed to evaluate the "lexical richness" of an author, the average distance in which new words are generated in a text, the frequency of *hapax legomena* and *dislegomena*, the so-called "Weighted Precision/Recall" (WPR) lexical units reports, such as, for example, the relationship between the Rate of Occurrence of an English article at the beginning of a sentence and the number of sentences, the relationship between the Rate of Occurrence of a conjunction followed by an adjective and the overall occurrences of the same conjunction, or the preference given to a synonym rather than to another (for example the relationship between the number of occurrences of *any* and *all* and the occurrences of *any*), etc. (Mendenhall 1887; Yule 1938; Williams 1940; Fucks 1952; Wake 1957; Ellegård 1962; Brinegar 1963; Mosteller, Wallace; Fuks, Lauter 1965; Morton 1965; Somers 1966; Antosch 1969; Brainerd 1973; 1974; Bruno 1974; Sichel 1974; Morton 1978; Kjetsaa 1979; Marriot 1979; Larsen *et al.* 1980; Burrows 1987; Hilton 1988, 1990).

The comparative application of the different methods to texts of *Known authors* has pointed out that the phrastic analysis does often produce unreliable

results, also because of the practical definition of the concept of "sentence" as a unit bounded by two gradually different elements of punctuation, like two full stops, dot and exclamation (or question) mark, dot and semicolon, two commas, etc. (Ellegård 1962; Mosteller, Wallace 1964; Marriot 1979). The lexical richness turned out as well to be insufficiently probative, since many of the experiments conducted on texts of the same known author have highlighted the possible reliance of this element on the practiced genre: it is different to write a letter, a newspaper article, a poem or a novel. Nevertheless, the interference between the author's typical stylistic features and the features related to the content or to the genre of the text (either literary or not) represents a very sensitive issue (Clement, Sharp 2003).

On the other hand, how the stylistic elements are stable over time and how they are influenced by each individual's spiritual evolution, as well as the way in which they consciously or unconsciously change over time still need to be investigated. Certain is that, in order to apply efficiently this type of analysis, the compared texts need to be, as well as comparable in terms of genre, language and content, long enough, and the analysed stylistic features need to be structural, frequent, easily measurable and sufficiently independent from the author's conscious control (Bailey 1979).

In this sense the WPR, the frequency of empty words, the computation of elements outside of the conscious control, both the frequency of the use of certain letters rather than others, and the investigations conducted on the relative frequency of morphemes and minimal linguistic segments have shown good differentiation and assimilative skills. We know that it is now possible to automate analysis procedures aiming at the authorship attribution through an increasingly efficient technology that has multiplied exponentially the calculation rapidity, providing opportunities that were still unthinkable few years ago; so, even though many of the currently employed strategies remain essentially measurements of the words (in terms of length, rate of occurrence, frequency ratio) and of the sentences (number of words and average length in terms of characters), stylometry has led to the awareness that a certain number of textual structures can be described in quantitative terms and, consequently, various tools of mathematics and statistics have been introduced in attribution science: so, gradually, the interest has shifted from indicators based on discrete linguistic components to methods in which the text is analysed through models rather not dissimilar to those in use to compare other chains of symbols, as for example the DNA.

In this view, the systems applied since the beginning of the XXI century have been using series of units, also linguistically non discrete ones, analysed through methods that range from Markov's chains to the compression algorithms, the Bayesian classifiers and, finally, to the numerous methods of "Machine Learning", a field of Artificial Intelligence that realises algorithms based on the learning of data coming from different types of samples and on

the following statistical evaluation of the relationships between the observed variables, in order to achieve the data summary and then new knowledge. In all these approaches the text is considered as a sequence of symbols and the lexical elements have no more meaning than other symbols' aggregates, while the statistics of the sequences of *n* consecutive characters (the so-called n-grams) naturally appear as fundamental subjects of the research (Khmelev, Tweedie 2001; Benedetto *et al.* 2002; Keselj *et al.* 2005; Basile *et al.* 2008).

In a competition organised in 2003 by Patrick Juola, different attribution methods were compared by applying them to the same composite *corpus* of texts "Ad Hoc Authorship Attribution Competition" (AAAC): the best results were obtained by scholars who had applied to the traditional stylometric parameters (unstable words, empty words, most frequent words) a machine learning method called Support Vector Machine. Juola himself provided a valuable guide for the questions of authorship attribution, which can also be seen as the theory underlying the JGAAP (Java Graphical Authorship Program), a downloadable program for analysis, text categorisation and attribution, written in the Java programming language (Juola, Baayen 2003; Juola 2006).

JGAAP uses a modular architecture, whose base levels are the graphical standardisation/regularisation of the text Canonicalisation, the stylometric element that is meant to be processed (Event Set Generation), the modality of selection of the element (Event Culling) and the statistical analysis of the acquired data (Analysis). Each one of these levels is handled by one single generic Java class: the Canonicalisation module is so handled by the canonicaliser class, the Event Set Generation by the Event Drivers class, the Event Culling by the Event Cullers class and the Analysis module by the Analysis Methods class. Among the Event Drivers we may select single characters or contiguous n characters gathered from a sliding window (Characters and Character Grams), as well as single words or contiguous n words (Words and Word Grams), the first word of each sentence (First Word in Sentence), the words with a variable number of letters or vowels (M-N Letter Words and Vowel M-N Letter Words, where *M* and *N* are variable parameters), the empty words or the function-words used in Mosteller and Wallace study on the Federalist Papers (MW Function Words), the rare words, such as those employed once or twice in each document (Rare Words), the sentence length measured in words (Sentence Length), the suffixes, understood as the last three letters of each words (Suffixes), the syllables per word, with a very simplified system where each vowel is counted as a syllable (Syllables per Word), etc.

In order to select the analysis modality it is necessary to activate the function Event Culling, which allows to sound out, in all the documents, the least common n phenomena or the most common n phenomena or even the phenomena present in all the samples (Least Common Events, Most Common Events, Xtreme Culler), etc. It is then possible to select, among the countless

types of statistical analysis, among which we bear in mind Burrow's Delta, Support Vector Machine (SVM), in its Gaussian version (Gaussian SVM) and in its linear version (Linear SVM), Linear Discriminant Analysis (LDA), Markov Chain Analysis, Naïve Bayes Classifier, PCA, SPCA, WEKA. Some methods require the selection of the Distance Functions such as the Cross Entropy, the Lempel-Ziv-Welch Nearest Neighbor Classifier and many others.

2. Application to Romanic texts

Although the good performance of many computer methods of attributive analysis has now been demonstrated and although such methods are employed extensively and with good results in other areas of literary studies, especially in Anglo-Saxon ones, their application to the texts of the Romanic literatures (whether medieval or not) has not received sufficient impulse yet. In several works in collaboration with physicist Vittorio Loreto of La Sapienza University of Rome, we have been using methods based on the Information Theory, applying to the texts of Italian poetry a zipping program and different programs for the elaboration of phylogenetic trees (Canettieri *et al.* 2005; 2006; Canettieri 2011; 2012).

The results have been overall very satisfying, with percentages over 90% for the known author on known author attribution. As has been demonstrated, in this type of approach the outcome of the process is constituted by a tree in which the analysed elements are grouped on a scale in clusters, from the closest to the farthest. Reckoning the advancement of researches in the already boundless field of authorship attribution, I have been able to verify in first person the possibility of extending and integrating such researches. So, I have been using the program developed by Juola to test the different "known author on known author" methods, also in order to achieve results more solid than those obtained previously.

I have thus developed a series of experiments in which I have tried to verify how a specific approach was able to identify the author of a text string, taking into account the crucial variables of this type of research: text length, number of examples, genre variety. I have also recently extended the research, in collaboration with two scholars of authorship philology, to twentieth-century authors, in order to validate or falsify the method in the event of ascertainable authorship diachrony. Although I have not crossed all the possibilities offered by the system yet, I believe that, for early Romanic texts, considered the huge graphic oscillation, which makes each exclusively lexicon-based analysis much less accurate, the most productive approach is the one that computes pairs of letters (Characters Bigrams), using Linear SVM as a method of statistical analysis and selecting the items with Xtreme Culler: the application of these parameters to different *corpora* has in fact always produced excellent results, which I will now briefly describe.

The first, extremely simple experiment only concerned the *Roman de la Rose*. The text was divided into parts of about 1000 lines each, up to v. 13058. The first documents, named *Roselorris (1-4)*, contained the part of the *Roman de la Rose* attributed to Guillaume de Lorris; the remaining nine documents, named *RoseMeung (1-9)*, the part attributed to Jean de Meung. Then two files were constituted in the *Known Authors*' box: all the *Roselorris* documents were included in the first file, named *Guillaume*, all the *RoseMeung* documents in the second one, named *Jean*. The experiment consisted in rotating in turns all the documents, eliminating them one by one from the *Known Authors*' file and entering the corresponding removed document in the file of *Unknown authors*. In the first test, for example, *Roselorris1* was listed among *Unknown Authors* and *Roselorris2*, *Roselorris3* and *Roselorris4* in the *Known Authors*' file; in the second test, the same procedure was followed by placing *Roselorris2* among the *Unknown Authors* and *Roselorris1*, *Roselorris3* and *Roselorris4* among the *Known Authors*, and so on, by rotating the texts to be verified. The results gave 100% of correct attributions: all the documents named *Roselorris* were attributed to *Guillaume* and all the documents named *RoseMeung* were attributed to *Jean*. In addition to the functionality of this method for issues such as the one here proposed, this simple experiment demonstrates, if it is still necessary, that the double authorship of the *Roman de la Rose* cannot be called into question.

The second experiment involved troubadours' lyric poetry, with much smaller portions of text. Each one of the first ten documents, named *Guglielmo (1-10)*, contained a *vers* of Guglielmo IX; each one of the next six documents, named *Jaufre (1-6)*, the *vers* of Jaufre Rudel. The two series were included in the *Known Authors*' box, in files named respectively *Guglielmo* and *Jaufre*. Also in this case, the texts were rotated one by one in the *Unknown Authors*' file, with the result of a full recognition: all the texts attributed to Guglielmo, including the discussed *chansoneta nueva*, and all the texts attributed to Jaufre were recognised. The recognition of 100% of the texts also occurred with the extension of this *corpus* first to nine poems of Raimbaut d'Aurenga, of certain attribution, and then to three authors of his "manner" (Elias de Barjols, Gaucem Faidit, Peire Vidal), with a limited number of texts (four per author). In this last experiment we also added to the texts of *Known authors* a composition attributed to Elias de Barjols (BdT 132,8), for which the attributive *varia lectio* also mentions the name of Gaucelm Faidit; this attribution, in the light of the presented results, should probably be rejected.

Full recognition known on known results were also obtained for the Trouvères poetry, which was stressed with lyrics of Adam de la Halle, Andrieu Contredit, Thibaut de Champagne: in this case I nominated each text with the name of the Trouvère followed by numbers 1, 2, 3 and 4 (therefore adam1, adam2, adam3, adam4, andrieu1, andrieu2, andrieu3, andrieu4, thibaut1, thibaut2, thibaut3, thibaut4). Then I rotated all the documents both in the

Known authors' file, with three compositions per author at a time, and in the *Unknown authors'* file, with one document per Trouvère, obviously the ones not included among the *Known authors*.

In an additional experiment, I tested the method on some authors of the Italian Duecento poetry, in an attempt to verify or falsify the results obtained with the compression method: Amico di Dante (30 sonnets), Bonagiunta Orbicciani (20 sonnets and 5 madrigals), Brunetto Latini (Tesoretto), Guido Cavalcanti (30 sonnets), Cecco Angiolieri (30 sonnets), Chiaro Davanzati (30 sonnets), Cino da Pistoia (30 sonnets), Dante Alighieri (30 sonnets), Dante da Maiano (30 sonnets), Folgore da San Gimignano (30 sonnets), Guittone d'Arezzo (30 sonnets), Monte Andrea (30 sonnets), Rustico Filippi (30 sonnets), *Fiore* (30 sonnets), *Intelligenza* (4500 vv.) and *Mare Amoroso*. In the *Known Authors'* file were placed two documents, nominated with the abbreviated name of the author or of the anonymous work (thus also *Fiore*, *Intelligenza* and *Mare Amoroso*), followed by numbers 1 and 2 (amico1, amico2, bonag1, bonag2, brunetto1, brunetto2, etc.), each containing 10 sonnets or, in their absence (see for example *Tesoretto* of Brunetto Latini or *Intelligenza*), the evaluated syllabic-based equivalence of text amount (approximately 1500 syllables per document).

In the *Unknown Authors'* file were included documents containing another 10 sonnets (or textual equivalence) of each one of these authors or anonymous works, and named like the documents above, but followed by number 3 (amico3, brunetto3, standing for Bonagiunta, and in the absence of sonnets the madrigals were included, thus bonagiuntacan, etc.). The result was 100% "known author on known author" recognition. The three anonymous works were attributed to the authors of the textual portion of the same work included among the *Known Authors* (thus fiore3 was attributed to Fiore; intelligenza3 to *Intelligenza* and mare3 to *Mare Amoroso*).

In the next experiment, *Fiore*, *Intelligenza* and Mare Amoroso were not included in the *Known Authors'* file; they are thus attributed to the authors to be considered as the closest. The result is astonishing since, in front of a "known author on known author" attributive precision of 100%, in the case of the anonymous poems the result is a complete diffraction: in seven cases *Fiore* was attributed to Rustico Filippi, in four cases to Dante Alighieri, in two cases to Cecco Angiolieri and in two cases to Cino da Pistoia; *Intelligenza* was attributed to Rustico Filippi in two cases, in one case to Folgore da San Gimignano, in one case to Guido Cavalcanti and in one case to Dante Alighieri; Mare Amoroso was attributed in one case to Dante Alighieri and in two cases to Rustico Filippi. In my opinion, this result provides the evidence of the fact that the authors of the involved works should be sought outside the group of those included in the *corpus*. I will comment elsewhere the (apparent) discrepancy existing between this result and the result obtained for *Fiore* in the experiments communicated above.

As I already outlined, the method here proposed, in addition to the works of early Romanic literatures, is also being applied to texts of contemporary authors: in collaboration with two scholars of authorship philology, Simone Celani and Paola Italia, I tested it on two interesting and complementary cases, on the one hand some texts of Fernando Pessoa for which there is no clear attribution yet, on the other hand Montale's *Diario postumo* for which serious doubts of apocryphia subsist (at least for a number of poems) (Celani 2005; Italia 2013, 173-196; Celani, *infra*). In both cases, and despite the quite limited size of the analysed textual portions, the analysis of bigrams using *Linear SVM* in *Xtreme Culler* proved itself useful to discriminate authorship, with always quite substantial percentages of "known author on known author" correct attributions; furthermore, interesting results were also obtained in extreme cases such as that of Pessoan heteronyms or that of the probably apocryphal texts of Montale's *Diario Postumo*, generally in confirmation or assistance of the results obtained through downright philological analysis (Canettieri, Italia forthcoming; Canettieri, Celani forthcoming).

Altogether the method of analysis here developed appears to be particularly useful in cases where no vast portions of text are available, for example single poems, and where it is necessary to attribute those to not particularly large groups of authors. Paradoxically, the expansion of textual amount that is available for the analyser, precisely because of the used analyser, might reduce the system's efficiency. Naturally the upper and lower limits beyond which the system loses its efficiency still need to be verified: we know that these limits exist, but their precise identification in relation to the different *corpora* under exam, is another step in the process that will have to bring to systematic exploration of computerised authorship attribution, which still represents, for Romanists, a *terra incognita*.

Paolo Canettieri
Dipartimento di Studi Europei, Americani e Interculturali
Sapienza Università di Roma

REFERENCES

Antosch F. 1969, *The Diagnosis of Literary Style with the Verb-Adjective Ratio*, in R.W. Bayle (ed.), *Statistics and Style*, New York, American Elsevier.

Bailey R.W. 1979, *Authorship Attribution in Forensic Setting*, in *Advances in Computeraided Literary and Linguistic Research*, Birmingham, AMLC University of Aston.

Basile C., Benedetto D., Caglioti E., Degli Esposti M. 2008, *An example of mathematical attribution*, «Journal of Mathematical Phisics», 49, 125-212.

Benedetto D., Caglioti E, Loreto V. 2002, *Language Trees and Zipping*, «Physical Review Letters», 88/4, 1-4.

Besomi O., Caruso C. 1994, *L'attribuzione: teoria e pratica. Storia dell'arte, musicologia e letteratura*, *Atti del Seminario (Ascona 1992)*, Basel, Birkhäuser, 3-4.

Brainerd B. 1973, *On the Distinction Between a Novel and a Romance: A Discriminant Analysis*, «Computers and Humanities», 7, 259-270.

Brainerd B. 1974, *Weighing Evidence in Language and Literature: A Statistical Approach*, Toronto, University of Toronto Press.

Brinegar C.S. 1963, *Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship*, «Journal of American Statistical Association», 58, 85-96.

Bruno A.M. 1974, *Toward a Quantitative Methodology for Stilistic Analyses*, Berkeley, University of California Press.

Burrows J.F. 1987, *Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style,* «Journal of the Association for Literary and Linguistic Computing», 2, 61-70.

Canettieri P. 2011, *Il Fiore e il Detto d'Amore*, «Critica del Testo», 14/1, 519-30.

Canettieri P. 2012, *Unified Theory of the Text (UTT) and the Question of Authorship Attribution*, «Memoria di Shakespeare», n.s. 8, 65-67.

Canettieri P., Celani S. (in press), *Pessoa Heteronymy and New Attribution Methods*.

Canettieri P., Italia P. (in preparation), *Un caso di attribuzionismo novecentesco: il "Diario Postumo" di Eugenio Montale*.

Canettieri P., Loreto V., Rovetta M., Santini G. 2005, *Higher Criticism and Information Theory*, «Cognitive Philology», 3.

Canettieri P., Loreto V., Rovetta M., Santini G. 2006, *Philology and Information Theory: Towards an Integrated Approach*, in P. Baret, A. Bozzi, C. Macé (eds.), *Textual Criticism and Genetics*, «Linguistica Computazionale», 104-126.

Canettieri P., Loreto V., Rovetta M., Santini G. 2008, *Philology and Information Theory*, «Cognitive Philology», 1.

Celani S. 2005, *Il Fondo Pessoa: problemi metodologici e criteri di edizione*, Viterbo, Sette Città.

Clement R., Sharp D. 2003, *N-gram and Bayesian Classification of Documents for Topic and Authorship*, «Literary and Linguistic Computing», 18/4, 423-447.

Contini G. 1984, *Il Fiore e il Detto d'amore attribuibili a Dante Alighieri*, Milano, Mondadori.

Ellegård A. 1962, *A Statistical Method for Determining Authorship: The Junius Letters 1769-1772,* «Gothenburg Studies in English», 13, 1-115.

Erdman D.V., Fogel E.G. 1966, *Evidence for Authorship: Essay on Problems of Attribution*, Ithaca, Cornell University Press.

Fucks W. 1952, *On the Mathematical Analysis of Style*, «Biometrika», 39, 122-129.

Fucks W., Lauter J. 1965, *Mathematische Analyse des Literarischen Stils*, in H. Kreuzer, R. Gunzenhauser (eds.), *Mathematik und Dichtung: Versuche zur Frage einer exakten Literaturwissenschaft*, München, Nymphenburger Verlagshandl, 107-122.

Ginzburg C. 1979, *Spie. Radici di un paradigma indiziario*, in A. Gargani, *Crisi della ragione*, Torino, Einaudi.

Hilton John L. 1988, *Some Book of Mormon Word Print Measurements Using Wrap-around Block Counting*, Provo, Utah, FARMS.

Hilton John L. 1990, *On Verifying Wordprint Studies: Book of Mormon Authorship*, «BYU Studies», 30/3, 89-108.

Keselj V., Peng F., Cercone N., Thomas C. 2005, *N-gram Based Author Profiles for Authorship Attribution*, in *Proceedings of the Conference Pacific Association for Computational Linguistics*, PACLING'03, Dalhousie University, Halifax, 255-264.

Khmelev D.V., Tweedie F.J. 2001, *Using Markov Chains for Identification of Writers*, «Literary and Linguistic Computing», 16/4, 299-307.

Kjetsaa G. 1979, *And Quiet Flows the Don Through the Computer*, «Association for Literary and Linguistic Computing Bulletin», 7, 248-256.

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

Jakobson R. 1960, *Closing Statement: Linguistics and Poetics,* in Sebeok 1960, 350-377.

Juola P., Baayen H. 2003, *A Controlled-Corpus Experiment in Authorship Attribution by Crossentropy*, in *Proceedings of ACH/ALLC*, Athens, GA, 59-67.

Juola P. 2006, *Authorship Attribution*, «Foundations and Trends in Information Retrieval», 1/3, 233-334.

Labbé C., Labbé D. 2001, *Inter-Textual Distance and Authorship Attribution. Corneille and Molière*, «Journal of Quantitative Linguistics», 8/3, 213-231.

Langlois E. (ed.) 1914-1924, *Le Roman de la Rose* par Guillaume de Lorris et Jean de Meun, Paris, Firmin-Didot.

Larsen W.A., Rencher A.C., Layton T. 1980, *Who Wrote the Book of Mormon? An Analysis of Wordprints*, «BYU Studies», 20/3, 225-251.

Marriot I. 1979, *The Authorship of the Historia Augusta: Two Computer Studies*, «Journal of Roman Studies», 69, 65-77.

Mendenhall T.C. 1887, *The Characteristic Curves of Composition*, «Science», 9, 237-249.

Morton A.Q. 1965, *The Authorship of Greek Prose*, «Journal of the Royal Statistical Society A», 128, 169-233.

Morton A.Q. 1978, *Literary Detection*, New York, Scribners.

Mosteller F., Wallace D. 1964, *Inference and Disputed Authorship: The Federalist*, Reading Ma., Addison-Wesley.

Pulsoni C. 2001, *Repertorio delle attribuzioni discordanti nella lirica trobadorica*, Modena, Mucchi.

Sebeok T. 1960 (ed.), *Style in Language*, Cambridge Ma., The MIT Press.

Segre C. 1985, *Avviamento all'analisi del testo letterario*, Torino, Einaudi.

Sichel H.S. 1974, *On a Distribution Representing Sentence Length in Written Prose*, «Journal of the Royal Statistical Society A», 137, 25-34.

Somers H.H. 1966, *Analyse Statistique du Style*, Paris, Louvain.

Yule G.U. 1938, *On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authority*, «Biometrika», 30, 363-90.

Van de Walle J. 2008, *Roman Jakobson, Cybernetics and Information Theory: A Critical Assessment*, «Folia Linguistica Historica», 29, 87-123.

Wake W.C. 1957, *Sentence Length Distributions of Greek Authors*, «Journal of the Royal Statistical Society A», 120, 331-346.

Williams C.B. 1940, *A Note on the Statistical Analysis of Sentence Length as a Criterion of Literary Style*, «Biometrika», 31, 356-361.

ABSTRACT

Methods borrowed from Information Theory are applied to the traditional text criticism. A critique of the raw cladistic methods and an interpretation of the dichotomy-phenomenon are offered. The same methods are applied to 13th century Italian poetry to determine authorship attributions and to verify commonly accepted literary taxonomy. Philology is a human science primarily applied to literary texts and traditionally divided into lower and higher criticism. Lower criticism tries to reconstruct the author's original text and higher criticism is the study of the authorship, style, and provenance of texts. The use of methods borrowed from information theory makes it possible to bring together methodologically some of the sectors of the two fields. The outcome of the experiments in both text criticism and text attribution has been encouraging. In the former, the tests performed on three different traditions have provided results very similar to those obtained by traditional methods requiring a great amount of time. The experiments carried out both on 13th century Italian poets and schools have shown that it is possible to draw texts closer to one another. Furthermore, the method we have used makes it possible to attribute anonymous writings.