Towards a taxonomy of suspected forgery in authorship attribution field. A case: Montale's *Diario Postumo*

Francesca Tomasi
Dept. of Classical Philology and Italian Studies
University of Bologna
via Zamboni 32
40126 Bologna (Italy)
+390512098539
francesca.tomasi@unibo.it

Ilaria Bartolini
Dept. of Computer Science and Engineering
University of Bologna
Viale Risorgimento, 2
40126 Bologna (Italy)
+390512093550
i.bartolini@unibo.it

Federico Condello
Dept. of Classical Philology and Italian Studies
University of Bologna
via Zamboni 32
40126 Bologna (Italy)
+390512098539
federico.condello@unibo.it

Mirko Degli Esposti
Dept. of Mathematics
University of Bologna
Piazza di Porta S. Donato 5
40126 Bologna (Italy)
+390512094409
mirko.degliesposti@unibo.it

Valentina Garulli
Dept. of Classical Philology and Italian Studies
University of Bologna
via Zamboni 32
40126 Bologna (Italy)
+390512098529
valentina.garulli@unibo.it

Matteo Viale
Dept. of Classical Philology and Italian Studies
University of Bologna
via Zamboni 32
40126 Bologna (Italy)
+390512098585
matteo.viale@unibo.it

ABSTRACT

This paper wants to explore quantitative and qualitative practices generally exploited in different scientific fields (philology, mathematics, quantitative linguistics, computer science) in order to reveal forgery. Our study will be conducted on Montale's *Diario postumo* that shows all the typical features of a suspected forgery. The final aim is to merge all these methods in order to define a taxonomy of annotation elements useful, in this particular context of authorship attribution, for developing a data model to be potentially used in all forgery situations.

Categories and Subject Descriptors

H.3 [**Information Storage And Retrieval**]: H.3.1 Content Analysis and Indexing; I.4 [**Image Processing And Computer Vision**]: I.4.7 Feature Measurement; I.5 [**Pattern Recognition**]: I.5.m Miscellaneous; I.7 [**Document And Text Processing**]: I.7.2 Document Preparation; J.5 [**Arts And Humanities**]: *Linguistics*, *Literature.*

Keywords

forgery, quantitative linguistics, image analysis, mathematics, philology, annotation, data model, TEI.

1. INTRODUCTION

The question "what a text is" is not a new topic. The variance of this concept implies different methods that could be exploited for managing an informational resource. A charming value of the text, in the domain of authorship attribution (A.A.), concerns how to reveal forgery.

Mathematicians, computer scientists, philologists, quantitative linguists and digital humanists have different points of view on what a text is; this entails different strategies in order to reveal forgery. We argue that only constructive interactions between different approaches might help with complex problems such as forgery.

Philologists usually adopt qualitative and comparative methods on the basis of phenomena like anachronisms concerning events and language, inconsistencies in style, patchwork effect, anomalies in the material medium or in handwriting style.

Computational methods are instead essentially statistical. Authenticity, dubious attribution, plagiarism, interpolation are typical subjects of stylometry and quantitative linguistics.

Quantitative linguists, but also mathematicians, use two different approaches: 1) texts as character strings, regardless of their meaning (algorithmic approach); 2) texts as word sequences that have to be studied statistically ("bag of words" approach);

From the point of view of computer scientists, text could be represented also, for example, by means of the image of the text itself (e.g. a manuscript page). In this context, the application of pattern analysis and (dis)similarity search techniques (characterizing the handwriting of the page in term of "low-level features") could help in solving the problem of authorship attribution.

However, similarity is not a satisfactory criterion in order to attribute authorship in the case of suspected forgery: it is not surprising that a forgery is *similar* to the author's work: the problem is how to verify if a text is *too similar* to the author's work, and if such types of similarity cannot be found elsewhere in the extant *corpus*.

Our first case study will be Montale's *Diario postumo*, that shows all the typical features of a suspected forgery: first of all, *an excess* – rather than a lack – of textual similarities (single words, word groups, sentence patterns, etc.) with Montale's authentic works (these similarities are mixed, of course, to many inconsistencies at the level of both style and meaning). This is

DH-case '13, September 10 2013, Florence, Italy

http://dx.doi.org/10.1145/2517978.2517989

why the traditional methods of authorship attribution, based on mere statistical data, are not sufficient to evaluate how 'Montalian' is our text.

The final aim of the project described in this paper is to explore these methods in order to define an annotation model able to formally represent all the different points of view on the text. The results derived from these different approaches will be used for the annotation of the whole poetical *corpus*, in order to acquire information useful to define an extensible taxonomy of all the phenomena that could reveal forgery. The taxonomy will be used for exploring the possibility to establish a data model with classes and predicates regarding forgery.

2. RELATED WORKS

The approach that we want to describe here has no former similar works. The methods we intend to analyze here have been studied independently from each other.

In philology there are some well tested approaches used to reveal forgery: i.e. the study of the medium, the analysis of handwriting, the presence of anachronisms, the exam of contents and obviously of style. Secondary literature on this topic is really rich (for a general discussion see [24] and [25]; a 'classical' history of the question is [26]).

Computational methods used to face these problems could be mostly classify under the quantitative approach of the A.A. domain. A.A. has thus a quite long history and the interested reader is referred to [61] and [32]. But forgery has not exactly the same problems of the A.A.

The idea of applying quantitative, although not always mathematically founded, ideas to the problem of recognizing the author of an anonymous or apocryphal text is not new: it dates back to the end of the 19th century at the latest, when two studies by the mathematician de Morgan [23] and the geophysicist Mendenhall [41] (wrongly) suggested that the average length of words in the works of different writers be calculated and compared them in order to establish authorship.

From the first works by de Morgan and Mendenhall up to recent times there has been a shift in the methods used for attribution; the interest has moved from indicators based on words, which are very natural because words are the basic components of language (see section 4.3.), to methods where no syntactical structure of the text is taken into account (see section 4.2.).

This approach, where the text is considered merely as a sequence of symbols, is far from being new. Already Markov [39], [40] and Shannon [59] regarded the texts as a sequence of symbols, where the words as basic components of the text have no more meaning than other aggregates of symbols. In this approach it is the statistics of sequences of n consecutive characters (the so called *n–grams*) that appears naturally as the fundamental object of investigation.

Focusing now on authorship attribution, even if approaches based on words or other syntactical units are still frequently used, several works of the past decade adopted the "n-gram approach": Clement and Sharp, for example, proposed in 2003 [17] a method based on n-gram frequencies, whereas in 2001 Khmelev and Tweedie [33] published some results obtained by considering texts as first order Markov chains, i.e. by calculating the empirical one-step transition matrix between characters from the reference texts of an author and then using it to establish the probability for a given anonymous text to have been written by that author. Around the same time, novel and efficient mathematical methods for A.A. based on compression algorithms have been introduced [7]. Recently the combinations of these two approaches, namely the one based on n-gram frequencies and the one based on compression algorithms have been successfully applied to two concrete problems in A.A. [6], [8] and [9]. These methods, roughly speaking, allow to introduce similarity distances between texts and can be used for designing methods of attribution with an approach which is in some way distinct from "usual" computer science approach to classification, where a text is basically represented as a vector of frequencies of words (bag-of-words model) and then, eventually after some dimension reduction (feature extraction) and in the spirit of machine learning, suitable algorithm/machine (neural nets, supported vector machine [SVM], Bayesian tools, etc...) are trained and used to discriminate and classify texts into known classes [61], [32].

Recently, the bag-of-words approach made further progress through the so-called *intertexual distance index* created by the French scholar E. Brunet [13] and then adjusted by D. Labbé, who applied it in studies on the attribution of literary works of uncertain authors [36], [37]. New calibration and applications of this index are having particularly fruitful results (see e.g. [20], [54]).

As regards the specific field of image analysis, OCR is a traditional approach based on pattern-recognition techniques that enable a computer to read texts (i.e. scanned images of a texts). However, if this is a feasible solution on printed text, its use for manuscripts is rather problematic. In general, OCR applied to handwritten texts is far from being perfect because of the issue of "variations". For example, the same letter drawn by the same person is slightly different each time, as well as letters drawn by different hands. These variations make it hard for the computer to read the writing correctly and to make a successful match in the context of authorship attribution [14]. System for Paleographic Inspections (SPI) is the first tool for the study of handwritten manuscripts. SPI can solve the problem of variations by training and working on prototypes of letters, i.e. collecting abstracted models of a single person's handwriting. The prototype comes with a pre-defined set of limits between which the letter belonging to the unidentified document may deviate from the prototype. This decides whether a letter is written by the same hand, or someone else drew it [16]. The main limit of SPI is that the segmentation process focuses on the shape of individual letters only. Thus, the overall appearance of the manuscript page and its immediate context is completely ignored [44].

Finally, the Digital Humanities community has a long tradition in A.A. studies [28] and in quantitative analysis of texts [29]. Literary text analysis is a practice that reveals the complexity of the concept of text and the potential stratification of interpretative levels [51].

3. THE CASE STUDY

Diario postumo is a collection of 84 poems written by Montale between 1969 and 1979 (according to the official version), and given by the poet to his young friend Annalisa Cima, with the precise order to publish the texts only after his death. Cima testifies that the poet himself arranged his poetic legacy as follows: ten sealed envelopes (numbered from I to X), each containing 6 poems, and a bigger packet, not numbered, containing another envelope with 6 poems (numbered XI) and 18 additional compositions [15, 89][1]. In 1986, five years after the

[1] It is important to observe that the only witness of the real existence of these envelopes is Cima herself, because Bettarini

poet's death, the Foundation Schlesinger (Lugano) published the first 6 texts in a plaquette titled *Poesie inedite di Eugenio Montale*[2]. In 1991, the first section of *Diario postumo* (30 texts) was published by A. Mondadori. In 1996, all the remains of the legacy were edited by Annalisa Cima, with the substantial ecdotic contribution of one of the most important scholars in the field, Rosanna Bettarini [46]. It was then that the philological controversy flared up. In the previous years a poet like Giovanni Raboni, among others, had raised a doubt about the authenticity of the texts (although on the basis of an aesthetic judgment only [49]), but immediately after the publication of [46] another of the most authorative Montalian philologists, Dante Isella, launched his formidable attack against the authenticity [30]. In addition to philological arguments (based especially on the patchwork nature of many texts, which look like a mosaic of Montalian quotations: see section 4.1), Isella underlined – through an impressive expertise by Armando Petrucci, perhaps the most famous Italian paleographer – the highly implausible appearance of the supposed autographs published, in the meantime, by Annalisa Cima [30, 20]). The reply by the supporters of the authenticity of *Diario* was immediate: those attending a Conference in Lugano, in autumn 1997 [1], where the first exhibition of the discussed autographs took place, affirmed the authenticity of the collection unanimously. A key witness was provided by Maria Corti, who testified Montale's perfidious intention – revealed to her in 1971 by the poet himself – of providing some kind of posthumous provocation directed against critics and philologists [21]. On the basis of such a witness, nowadays *Diario* is regarded by many scholars as an authentic, albeit ironic and self-ironic, Montale's work. However, some doubts remain, and the communis opinion was perhaps too hastily accepted: the illustrious philologist Pier Vincenzo Mengaldo has strongly supported Isella's position [42]. A skeptical view on the authenticity of *Diario postumo* is now provided by N. Scaffai [55] and P. Italia [31] (the best recent surveys on the question).

4. APPROACHES

4.1 Philology

The traditional and well-tested approach of philologists, in order to reveal forgery, is based on some typical clues such as e.g.:

1. implausible or impossible features of the material medium (concerning the material itself, but also the techniques used and, of course, its age);
2. implausible or impossible features of the visual aspects of the object (e.g. *mise en page* of a text);
3. in the case of a written document, implausible or impossible features of the handwriting (discordant either from a single author's hand, if known, or from the use of his/her age);
4. anachronisms both factual (mention of events, persons, customs, etc. which are chronologically incompatible with the age of the supposed author) and linguistic (words, forms, expressions belonging to a later stage of language);
5. recognition of the sources from which the text seems to derive, if these sources are not compatible with either history or nature of the text;
6. contradictions at the content level (themes, ideas, data) with the other works of the supposed author.

Long experience shows that very rarely a forger does not make at least one (usually more than one) of these typical missteps. But if none of this clues supports the inquiry of the philologists, they resign themself to a less certain and conclusive criterion: the analysis of style. But it is difficult to draw compelling conclusions from a supposed "constant" or "typical" style (*usus scribendi*) of an author: not only because style can considerably change during an author's life and can depend on the literary genre practiced by the author, but mainly because a deliberate forgery (if not too ingenuous or rough) is characterized precisely by a scrupulous imitation of the individual author style. Therefore, if a stylistic analysis can help both in case of unknown or uncertain authorship, and in case of plagiarism, the same method is likely to be useless or even counterproductive when applied to a case of "pseudoepigraphy" (intentional false attribution).

The *Diario postumo* attributed to Eugenio Montale is a truly representative example of the impasses that are thrown up once we try to apply stylistic analysis to a text which is suspected to be a forgery. It is significant, in particular, that the high frequency of quotations from Montale's works (from almost all the previous poetical works) could represent an argument both *pro* and *contra* the authenticity. According to Isella, this is actually the chief proof of a forgery, i.e. of a patchwork text composed through an artificial and awkward bricolage of authentic Montalian expressions, mixed with an incredible amount of oddities and literary banalities (many of which find some interesting parallels in Annalisa Cima's own poetry) ([30, 7-15], and [42]). On the opposite, according to Bettarini and many others, the presence of such impressive similarities (sometimes word for word) between *Diario postumo* and Montale's previous production demonstrates Montale's authorship, especially because ironical self-quotations were a typical device of the works of his last period. And for those who believe in the theory of a "posthumous mockery", arranged by the poet to the detriment of his critics (see section 3), such a high degree of similarity becomes a further argument to assert the authentic, although ironical, nature of the work (see e.g. [12], [56], [48], [50], [10]).

But does the "self-quotation" practiced, e.g., by the author of *Satura* or *Quaderno di quattro anni* really agree with the quotation technique used by the author of *Diario postumo*? This is the question implicitly asked by Grignani [27] in an important contribution confined in an appendix of the Lugano Proceedings [1]. In order to answer this question, we need a preliminary taxonomy of phenomena such as repetition of words or words-series, quotation, self-quotation, etc., at least in the last decade of Montale's production. Such a taxonomy should provide us with distinctions both on the quantitative and the qualitative level, depending e.g. on phenomena such as:

1. number of words involved in a quotation or self-quotation;
2. linguistic nature of the words involved;
3. stylistic level of the words involved;
4. number of words replaced by synonyms;
5. stylistic level of these synonyms;
6. stylistic level of the quotation context;

declares that she was not personally present when the envelopes were progressively opened [31, 182].

[2] The gradual publication of the poems went on, year after year, in groups of six poems at a time.

7. age or period of the work which is the source of the quotation;

Etc.

The intersection of these and other criteria can help to draw a real "map" of Montalian intertextuality (with a special regard to forms and methods of the "internal intertextuality" in Montale's works dating from his last decade). Such a map is perhaps the best starting point for an evaluation of what is typical and what is atypical (therefore suspect) in *Diario postumo* (for other pertinent levels of literary analysis see table 1).

4.2 Mathematics

Why and how should A.A. (or *stylometry*, as it is sometimes called) be the object of a mathematical study?

An important point is that whatever is the approach used, described in the related work section (either similarity distances based on n-grams and compression algorithms, or feature extractions combined with machine learning), all these methods have been so far almost always applied to very typical and scholastic scenarios in A.A.: one or more unknown texts must be attributed to one (and only one) author selected from a finite number of known authors.

It is worth mentioning that already the concrete and frequent case when one has to decide if a given text have been written by a given author or not (the so called *Authorship Verification* problem) presents enormous difficulties and, to the best of our knowledge, no quantitative systematic approach exists in literature (see [34] for an interesting attempt).

Of course these limitations restrict the range of applicability, excluding several interesting and important real cases met in modern philology. This is exactly why we believe that a genuine multidisciplinary approach (quantitative, linguistic, philological, computational, image analysis) can be quite fruitful and we are convinced that the problem of the attribution of the "*Diario Postumo*" by Montale represents an excellent challenge. Here we are very far from any traditional scheme, far from classical A.A. problems and even far from an Authorship Verification problem (already very difficult to approach with mathematical models). First of all we face poetry, and any statistical method based on the presence of recurring patterns and long correlations cannot be applied in a straightforward way. For example the methods and the results must be invariant by the shuffling of the poems inside a given opera. But – most important – when we try to follow all the animated dispute, we can restrict the number of possible and realistic solutions to the problem essentially to three:

1. Annalisa Cima, using some raw material from Montale'writings (authentic and published works, but also voice recordings, sketches on paper, etc…), "created" *Diario Postumo*, often emphasizing and hence trivializing Montale's style.
2. *Diario Postumo* is a genuine creation by Montale, perhaps affected by a long time process of mental and physical degradation.
3. It was just a "joke" by Montale to make fun of people after his death.

But one could also argue that the *Diario* is a genuine small *corpus* stuffed with forgery (2 + 1).

Of course, if we have any evidence or suspect that (3) is the solution to the debate, then there is nothing to do for a mathematician (or others), but there are no hints supporting (3), as far as we know. Assuming now that this is not the case, (1) and (2) raise interesting and challenging problems. For example, assuming (2) and knowing the overall poetic opera of Montale along few decades[3], detecting, quantifying and measuring a degradation in his writing style poses an incredible challenge for quantitative methods. At this moment, we do not have any experimental results able to support the claim, but we are convinced that a mathematical approach to the problem[4], possibly combined with modern image analysis of the handwriting style and its evolution could hopefully shade some light on hypothesis (2).

Finally, if we want to investigate hypothesis (1), we clearly have to abandon the safety and prudent realm of typical A.A. problems (attribution among a finite number of authors, each one with a proper and *distinguishable style*) and move towards quantitative methods for forgery or plagiarism, two sides of the same medal: an author trying either to mimic or camouflage another author creation (or fragments of it). Also in this case, quantitative approaches have been recently introduced and tested [47], often inspired by the related A.A. algorithms here briefly discussed, and we believe that they might represent a solid starting point for our investigation.

4.3 Quantitative linguistic

The typical approach of quantitative linguistics to the issues of A.A. and identification of forgery is the already mentioned "bag-of-words" which stems from the traditional statistical analysis of words – considered as mere graphical forms (type or token) or alternatively as lemmas – through different quantitative methods varying on a case-to-case basis.

One of the first tools developed to test in more global terms the lexical similarity between two corpora is the *lexical connection index* which is the ratio between the shared part of vocabulary and the two corpora joined together, which measures the portion of vocabulary shared by two texts [19, 52-55].

Afterwards, many statistical indexes have focused on particular aspects such as the lexical richness (e.g. type-token ratio), the words' length, the repeated segments of words, the position and recursion of specific keywords (see [38], [32], [64], [35], [61]) for a recent literature review on the topic; see [54] for a presentation and discussion of the different methods as well as the related references.).

Recently, the so-called *intertexual distance index* (see related works) is proving particularly fruitful. This index is not based on the simple number of shared occurrences – as in the case of the lexical connection index – but on a calculation which compares the frequency of each occurrence in the wordlists of the two texts. The main advantage of adopting this statistical tool is that it allows to compare texts of different sizes, relativizing through a

[3] At least concerning publishing, Montale's poetic works are quite evenly distributed in time: *Ossi di seppia* (1925); *Le occasioni* (1939); *Finisterre* (1943); *Quaderno di traduzioni* (1948); *La bufera e altro* (1956); *Xenia* (1966); *Auto da fè* (1966); *Satura* (1971); *Diario del '71 e del '72* (1973); *Quaderno di quattro anni* (1977); *Altri versi* (1980). Now we can add the posthumous *La casa di Olgiate e altre poesie* (2006).

[4] Here we think that also some recent approaches to textual data based on network analysis could turns out to be quite useful [60],[2].

mathematical proportion the different number of occurrences [[19, 55-58].

Labbé [36], [37] used some threshold values which allow to identify, in the case of more than one author, not only the different authors or the authorship attribution with various levels of possibility, but also the textual genres as well as the topics.

A further improvement of the index offers a new calculation procedure based on repeated observations of intertextual distance between pairs of equal-sized text chunks [20]. The above-mentioned study provides some tests on the thresholds proposed by Labbé aimed at considering the specificities of the Italian language.

However, this type of approach does not suit the present case study, the *Diario Postumo* of Eugenio Montale. As a matter of fact, in order to test the different hypotheses on such work it is necessary to determine if the author is Montale himself and consider that the alleged plagiarist attempts to imitate the style of the author by means of wise devices that can thwart the usefulness of these quantitative tools.

Nevertheless, the bag-of-words techniques can contribute significantly to the authorship attribution. From a methodological point of view, a preliminary important issue is the challenge of determining the existence of a threshold able to identify the author of a work by taking into account the changes related to the author's stylistic evolution with time. This is strictly connected to the study of the stylistic features of an author which remain invariable with time and can defeat the attempts of the alleged plagiarist.

The analysis of linguistic and formal features of literary texts are dealt with by stylistics through traditional non-quantitative approaches such the study of lexical register, key-words, syntactic peculiarities, position of POS, punctuation, etc. [43], [18].

From an operational point of view, the issue – which has not been analyzed comprehensively in the literature on the topic, particularly applied to the Italian poetry – can be framed by resorting to useful studies: on the one hand, those analyzing Montale's poetic production certainly attributed[5] to the poet, and on the other hand those dealing with the poetic production of authors subject to stylistic evolution and aiming at determining the distinctive features beyond the stylistic changes occurred in time.

In broader terms, a quantitative analysis of the main stylistic elements of literary stylistics may contribute significantly, although not decisively, to the identification of plagiarism.

4.4 Image analysis

Text should be represented also by means of the image of the text itself through, for example, the digital representation of a manuscript page. The application of pattern analysis and (dis)similarity search techniques, able to characterize the handwriting of a page in term of "low-level features", could help in solving the problem of authorship attribution.

The problem therefore is translated in term of "analysis and comparison of handwritten pages" with the aim of establishing whether a manuscript corpus is authentic or not with respect to a specific author. Of course, this implies the need of a pre-processing phase where the analysis of some handwritten pages of authentic writings is executed in order to build a "ground truth" reference information for comparing suspicious writings to the authentic ones.

Our data and comparison models are inspired by the Windsurf ones [5], where images are composed by elements (i.e. relevant parts of the handwritten text). Each element is described by means of automatically extracted *low level features* that represent, in an appropriate way, the content of the element itself (e.g. the handwritten style of an author with respect to a specific set of graphic aspects). We regard such features as a first step of automatic *low-level annotations* of our digital corpora. As for the comparison model, given an input (*query*) manuscript page, composed of *m* relevant elements, and an element distance function that measures the dissimilarity of a given pair of elements using their features, we want to determine automatically if the query manuscript page could be considered authentic with respect to a specific author.

Dissimilarity between manuscript pages is numerically assessed by way of a page distance function that somehow "combines" the single element distances into an overall value. The efficient resolution of comparisons over features is ensured by an index structure built on top of elements (e.g. syllables).

In particular, manuscript pages are first segmented in parts (e.g. syllables, words, sentences, etc.). From each element, visual salient characteristics, able to define specific graphic aspects (such as shape, module, *ductus*, writing angle, hatching, and ligatures) and thus differentiate the handwriting of an author, are automatically extracted. Image elements are compared according to their visual features using an ad hoc distance metric. Elements scores are then opportunely matched to aggregate distance values of matched elements.

Finally, in order to further enrich data representation, pages are also annotated with high level (semantic) descriptors which are text labels representing the meaning of syllables, words, sentences, etc. Such descriptors are in the form of keywords (or tags) and are semi-automatically assigned at both the whole image and elements levels by means of a semi-automatic annotator.

Note that our rich data characterization ensures to capture the appearance and the layout of the whole manuscript page, thus guaranteeing an effective and efficient representation of our corpora.

5. THE ANNOTATION MODEL

A.A. generally does not consider annotations as a method useful for deducing attribution because stylometry, that is a statistical/numerical approach, is the most attested practice. At the same time annotation has been used as a tool for the lemmatization or for the creation of semantic or syntactic treebank [4]. The POS (Part Of Speech) annotation could help in some NLP (Natural Language Processing) problems.

Given that the quantitative approach help us in counting n-grams (see section 4.2), the above mentioned phenomena (see all section 4) could be formally described through semantic tags. The aim of the annotation model we want to define here is to take into account all the approaches, trying to define a possible taxonomy starting from the vocabulary of the TEI schema [62]. Some experiment in "forensic philology" will be useful in order to verify whether TEI is sufficient in order to describe all the features [57].

[5] The groundbreaking statistical-linguistic analysis of the first collections of Montale carried out by Luigi Rosiello [52] provides valuable insights into the characterisation of Montale's poetic production and the analysis of the stylistic features of each collection of the Italian author. See [19] for an accurate discussion of the study conducted by Rosiello.

The macro-levels of annotation, emerged from our first analysis and considered here as points of view on the source (philological, mathematical, linguistics, image analysis), will be:

- characters;
- words and segments;
- linguistic features;
- literary phenomena;
- lexical data;
- image pattern.

These levels are the highest concepts of a classification scheme intended to define categories and subcategories of phenomena related to forgery. The taxonomy could be described at a more granular level as shown in table 1. Levels are the categories, features are the subcategories of the classification scheme and TEI elements/attributes are the formal way used in order to declare properties.

Table 1. A taxonomy of forgery

Level	Features	TEI elements/attributes
Characters		<c>
Punctuation		<pc>
Words	nouns, adjectives, verbs, adverbs, function words/lexical words	<w>, @lemma
Segmentation	sentences, phrases, clause and syntactic aspects; verse (rhyme and metrical patterns)	<s>, <phr>, <cl>, @function; <l>, @met, @rhyme
Linguistic features (at the level of word, sentence, phrase, clause and verse)	fine-grained grammatical categories and morphological aspects (POS: NN, PP, NP, VP, etc.)	@type/@ana
Literary phenomena	rhetorical aspects (sound/meaning); quotation, self-quotation	, <interp>; <q>, <cit> @when, @type
Lexical data	archaisms, neologisms, foreignisms, keywords, borrowings, hapax legomena	<foreign>, <distinct>, <term>, @type, @ref="URI"
Image pattern (at the level of single character, syllabi, word, segment)	character/glyph, ligatures, dimensions, shape, module	<g>, <gliph>, <char>, <desc>

As a result, such a collection of elements/attributes will be attached to the source (both text and image) in order to describe each phenomenon and associate features to non interpreted strings/portion of image.

The annotation elements will be used in order to 1) count all the single phenomena both in *Diario* and in the rest of the *corpus* for detecting forgery (also on the basis of an excess of similarity); 2) compare the suspected (annotated) forgery text with the rest of the poetical (annotated) *corpus* in order to understand what is a typical, or atypical, Montalian way of writing with special regard to the *Diario*; 3) compare the suspected (annotated) forgery text with other texts, first of all Annalisa Cima's poems.

The process, that will be firstly managed manually, will surely reveal that embedded markup is complex (because of the overlapping of descriptive elements) and a stand-off approach will be adopted [for a recent discussion see 58].

The final aim of the process is to deduce a data model from the annotation, in order to specify classes and predicates of forgery elements. The taxonomy will be used as the first tool for defining concepts and establishing relationships between the defined concepts. Macro-levels (categories) will be the classes of the data model and the relationships between classes and subclasses will be managed as predicates in order to create a domain ontology (through OWL/RDF).

If this process will produce, on this specific case, some worth results it will be obviously used in order to verify the degree of reliability on other different kind (genre) of texts (other poems, prose and drama).

6. ACKNOWLEDGMENTS

The research aimed to lead to these results has received funding from the University of Bologna in 2013 for the project "*Authorship*, varianti, stile: frontiere dell'analisi testuale tra filologia, linguistica, matematica e *computer science*" ("Authorship, variants, style: textual analysis between philology, linguistics, mathematics and computer science") under grant agreement FARB ("Finanziamento di Ateneo alla Ricerca di Base" – Local grant for research).

7. REFERENCES

[1] AA.VV. 1998. *Atti del seminario sul Diario postumo di Eugenio Montale*. Lugano (24-26 ottobre 1997). Milano, Scheiwiller.

[2] Amancio, D.R., Oliveira, O.N., Jr and da F. Costa, L. 2012. Identification of Literary Movements Using Complex Networks to Represent Texts. *New Journal of Physics* 14 (April 2012 – 043029). arXiv:1302.4099.

[3] Argamon, S. and Olsen, M. 2009. Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis. *Digital Humanities Quarterly* 3, 2 (Spring 2009).

[4] Bamman, D. and Crane G. 2011. The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage*, C. Sporleder, A. van den Bosch and K. Zervanou, Eds. Springer, Berlin.

[5] Bartolini, I., Patella, M., and Stromei, G. 2011. The Windsurf library for the efficient retrieval of multimedia hierarchical data. In *Proceeding of the International*

Conference on Signal Processing and Multimedia Applications (Seville, Spain, July, 2011).

[6] Basile, C., Benedetto, D., Caglioti, E., Degli Esposti, M. 2008. An example of mathematical authorship attribution. *Journal of Mathematical Physics* 49, 125211. DOI=10.1063/1.2996507

[7] Benedetto, D., Caglioti, E., and Loreto, V. 2002. Language Trees and Zipping. *Physsical Review Letters* 88, 048702. DOI=10.1103/PhysRevLett.88.048702.

[8] Benedetto, D., Degli Esposti. M., Maspero, G. 2013. Authorship Attribution and Small Scales Analysis Applied to a Real Philological Problem in Greek Patristics, *QuaLiCo 2012 proceedings* (Belgrade, Serbia, April 26-29, 2012) (in press).

[9] Benedetto, D., Degli Esposti. M., Maspero, G. 2013. The puzzle of Basil's Epistula 38: a mathematical approach to a philological problem. *Journal of Quantitative Linguistics* (in press).

[10] Benevento, A. 2000. Il parapiglia del Montale postumo. *Rivista di Letteratura Italiana*, 1, 113-124.

[11] Bettarini, R. 1996. Apparato critico. In [46], 87-117.

[12] Bettarini, R. 1998. Il bello viene dopo. In [1], 7-12.

[13] Brunet, E. 1988. Une mesure de la distance intertextuelle: la connexion lexicale. *Revue informatique et statistique dans les sciences humaines*, Centre Informatique de Philosophie et Lettres, Université de Liège.

[14] Bunke, H. and Wang, M.S.P. 1997. *Handbook of character recognition and document image analysis*. World Scientific Publishing Company, Singapore.

[15] Cima, A. 1998. Montale postumo e l'accademico spregiudicato. In [1], 13-18.

[16] Ciula, A. 2005. Digital paleography: using the digital representation of medieval script to support paleographic analysis. *Digital Medievalist* 1 (Spring 2005).

[17] Clement, R. and Sharp, D. 2003. Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing* 18, 4, 423-447.

[18] Colella, G. 2010, *Che cos'è la stilistica*. Carocci, Roma.

[19] Cortelazzo, M. and Tuzzi, A. 2008, *Metodi statistici applicati all'italiano*. Zanichelli, Bologna.

[20] Cortelazzo, M.A., Nadalutti, P. and Tuzzi, A. 2013. Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics* 20, 2, 125-152.

[21] Corti, M. 1998. La storia lontana del Diario postumo. In [1], 42-45.

[22] Craig, H. 2004. Stylistic Analysis and Authorship Studies. In *A companion to Digital Humanities*, S. Schreibman, R. Siemens and J. Unsworth, Eds. Blackwell Publishing Ltd, Malden, MA, USA. DOI=10.1002/9780470999875.ch20.

[23] De Morgan, A. 1851. *Memoirs of Augustus de Morgan by His Wife Sophia Elizabeth de Morgan with Selections from His Letters*. Longmans, Green, London.

[24] Eco, U. 1990. Falsi e contraffazioni. In *I limiti dell'interpretazione*, Id. Bompiani, Milano, 162-192.

[25] Eco, U. 2012. La falsificazione nel Medioevo. In *Scritti sul pensiero medievale*, Id. Bompiani, Milano, 731-774.

[26] Grafton, A. 1990. *Forgers and critics. Creativity and duplicity in western scholarship*. Princeton UP, Princeton.

[27] Grignani, M.A. 1998. Intervento dal pubblico. In [1], 163-166.

[28] Holmes, D. 1994. Authorship Attribution. *Computers and the Humanities* 28, 87-106.

[29] Hoover, D.L. 2008. Quantitative Analysis and Literary Studies. In *A companion to Digital Literary Studies*, S. Schreibman, R. Siemens and J. Unsworth, Eds. Blackwell Publishing Ltd, Malden, MA, USA.

[30] Isella, D. 1997. *Dovuto a Montale*. Archinto, Milano.

[31] Italia, P. 2013. *Editing Novecento*. Salerno Editrice, Roma.

[32] Juola, P. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1, 3, 233–334.

[33] Khmelev, D.V. and Tweedie, F.J. 2001. Using Markov chains for identification of writers. *Literary and Linguistic Computing* 16, 3, 299-307. DOI=10.1093/llc/16.3.299.

[34] Koppel, M. and Schler, J. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning* (Banff, Canada, July, 2004), 489-495.

[35] Koppel, M., Schler, J., and Argamon, S. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science & Technology* 60, 1, 9-26.

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

[37] Labbé, D. 2007. Experiments on authorship attribution by intertextual distance in english. *Journal of Quantitative Linguistics* 14, 1, 33-80.

[38] Love, H. 2002. *Attributing Authorship: An Introduction*. Cambridge University Press, Cambridge.

[39] Markov, A.A. 1913. Primer statisticheskogo issledovanija nad tekstom `Evgenija Onegina' illjustrirujuschij svjaz' ispytanij v tsep (An example of statistical study on the text of `Eugene Onegin' illustrating the linking of events to a chain). *Izvestija Imp. Akademii nauk* 6, 3, 153-162.

[40] Markov, A.A. 1916. Ob odnom primenenii statisticheskogo metoda (On some application of statistical method). *Izvestija Imp. Akademii nauk* 6, 4, 239-242.

[41] Mendenhall, T.C. 1887. The Characteristic Curves of Composition. *Science* ns-9, 237-246. DOI= 10.1126/science.ns-9.214S.237.

[42] Mengaldo, P.V. 1998. Ma com'è goffo il Montale postumo. *Corriere della Sera* (12 marzo 1998).

[43] Mengaldo, P.V. 2007. *Prima lezione di stilistica*. Laterza, Roma-Bari.

[44] Meredith, M., Ainsworth, P. 2010. Digging into image data to answer authorship-related questions. *UK All Hands eScience meeting* (Cardiff, Wales, September 2010).

[45] Merriam, T. 2002. Intertextual Distances Between Shakespeare Plays, with Special Reference to Henry V (verse). *Journal of Quantitative Linguistics* 9, 3, 261-273.

[46] Montale, E. 1996. *Diario postumo. 66 poesie e altre*. Mondadori, Milano.

[47] PAN Conferences*: international workshop on uncovering plagiarism, authorship, and social software misuse* (2010, 2011, 2012).

[48] Parronchi, A. 1998. “Diario postumo”: echi montaliani da sé e da altri. In [1], 118-126.

[49] Raboni, G. 1997. Quegli apocrifi? L’avevo detto dieci anni fa, ma la querela non arrivò. *Corriere della Sera* (25 luglio 1997).

[50] Ramat, S. 1998. Su un frammento del Diario postumo (“Tornerà la musica”). In [1], 130-139.

[51] Rockwell, G. 2003. What Is Text Analysis Really? *Literary and Linguistic Computing* 18, 2, 209-19. DOI=10.1093/llc/18.2.209.

[52] Rosiello, L. 1965. Analisi statistica della funzione poetica nella poesia montaliana. In *Struttura, uso e funzioni della lingua*, L. Rosiello, Ed. Vallecchi Editore, Firenze.

[53] Savoca, G. 1997. *Concordanza del «Diario postumo» di Eugenio Montale. Facsimile dei manoscritti. Testo, Concordanza*. Olschki, Firenze.

[54] Savoy, J. 2012. Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics* 19, 2, 132-161.

[55] Scaffai, N. 2007. Un “apocrifo d’autore”: argomenti per il “Diario postumo” di Montale. In *La regola e l’eccezione. Saggi sulla letteratura italiana del Novecento*, Id. Ed. Le Monnier, Firenze, 120-139.

[56] Scheiwiller, V. 1998. Eugenio Montale, una burla riuscita. In [1], 19-21.

[57] Schlitz, S.A. 2009. The TEI as luminol: Forensic philology in a digital age. *Literary and Linguist Computing* 24, 2, 173-185. DOI=10.1093/llc/fqp001.

[58] Schmidt, D. 2010. The inadequacy of embedded markup for cultural heritage texts. *Literary and Linguist Computing* 25, 3, 337-356. DOI=10.1093/llc/fqq007.

[59] Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423 and 623-656.

[60] Solé, R.V., Corominas Murtra, B., Valverde, S. and Steels, L. 2010. Language networks: Their structure, function, and evolution. *Complexity* 15, 20–26.

[61] Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, 3, 538-556.

[62] Text Encoding Initiative. 2008. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/`.

[63] Weiss, S.M., Indurkhya, N. and Zhang, T. 2010. *Fundamentals of Predictive Text Mining*. Springer-Verlag, London-New York.

[64] Zheng, R., Li, J., Chen, H. and Huang, Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science & Technology* 57, 3, 378–393.