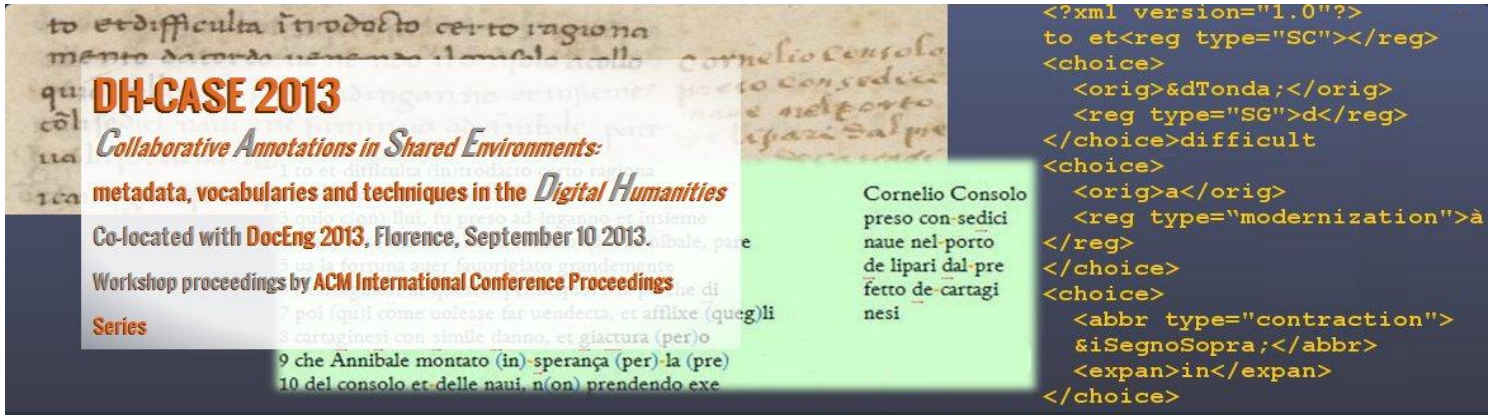
DH-CASE 2013

Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities



1st International Workshop **DH-Case '13**
Florence, September 10th

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

Ilaria Bartolini
Dept. of Computer Science and Engineering
University of Bologna

Federico Condello
Dept. of Classical Philology and Italian Studies
University of Bologna

Mirko Degli Esposti
Dept. of Mathematics
University of Bologna

Valentina Garulli
Dept. of Classical Philology and Italian Studies
University of Bologna

Matteo Viale
Dept. of Classical Philology and Italian Studies
University of Bologna

Goal and content

- Explore **quantitative** and **qualitative** practices generally exploited in different scientific fields (philology, mathematics, quantitative linguistics, computer science) in order to **reveal forgery**.
- Study 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 for developing a **data model** to be potentially used in all forgery situations.

A question of points of view

- 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**.

Approaches and methods

- **Philologists** usually adopt **qualitative** and **comparative** methods.
- **Computational methods** are instead essentially **statistical**.
- **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 a **computer scientist**, text could be represented also, for example, by means of the **image of the text itself** (e.g. a manuscript page).

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.
- 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…
- It 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.
- This is why the traditional methods of A.A., based on **mere statistical data**, are not sufficient to evaluate how „Montalian" is our text.

Philological approach

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.

It is significant 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.**

Mathematical approach

- **similarity distances** based on n-grams;
- **compression algorithms**;
- **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.

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.

Quantitative linguistic approach

- The typical approach of quantitative linguistics to the issues of A.A. and identification of forgery is the “bag-of-words”.
- ***lexical connection index****.* 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.
- ***intertexual distance index*****.** This index is not based on the simple number of shared occurrences but on a calculation which compares the frequency of each occurrence in the wordlists of the two texts.

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

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**.

Image analysis approach

- 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.
- 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.

The emerging levels

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**.

- The **macro-levels/categories of annotation** (highest concepts of a classification scheme), emerged from our first analysis and considered here as points of view on the source (philological, mathematical, linguistics, image analysis), are:
  - characters;
  - words and segments;
  - linguistic features;
  - literary phenomena;
  - lexical data;
  - image pattern.

Level - categories	Features – subcategories	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	`<span>, <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>`

The annotation elements

The annotation elements are 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 data model

- 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** for forgery.

Thank you!

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Francesca Tomasi
francesca.tomasi@unibo.it
Dept. of Classical Philology and Italian Studies,
University of Bologna

Via Zamboni, 32 40126 Bologna (BO)
Tel. +39 051,2098539 Fax. +39 051.228172