Umanistica Digitale – ISSN: 2532-8816 – n. 10, 2021
A. S. Lippolis, G. Totaro – *Variants Mining. Computational investigations on authorial variants: a comparison between Leopardi and Manzoni*
DOI: http://doi.org/10.6092/issn.2532-8816/12627

Variants Mining. Computational investigations on authorial variants: a comparison between Leopardi and Manzoni

[1]Anna Sofia Lippolis, [2]Gabriella Totaro

[1,2]University of Bologna, Italy
[1]annasofia.lippolis@studio.unibo.it
[2]gabriella.totaro@studio.unibo.it

Abstract

What is the meaning of the authors' interventions on their own texts? Variants Mining aims to answer this question by gaining empirical insight into creative thinking. The research focuses on pre-established categories of authorial variants to provide scholars with statistics of corrections, automatically computed starting from either the PDF document or the TEI encoding of the referring critical edition of the text. Finally, the output is made easily understandable thanks to a specifically tailored data visualization tool. This innovative comparative analysis on the correction modalities would allow to provide new philological readings that shed light on the creative process of different authors and assess if these interventions can be read as recurring patterns. Therefore, this methodology promotes a more conscious insight on the historical evolution of the authorial thought through a straightforward representation. A pilot analysis will focus on the case studies of Giacomo Leopardi's *Canti* and Alessandro Manzoni's *I promessi sposi* to understand the difference not only between the two key authors of Italian Romanticism, but also between poetry and prose.

Qual è il significato delle correzioni d'autore? Variants Mining cerca di rispondere a questa domanda tramite l'acquisizione di dati empirici sul pensiero creativo. Tramite una previa categorizzazione delle modalità di correzione autoriale, lo studio mira a fornire agli studiosi statistiche sulle correzioni, calcolate in maniera automatica a partire dal PDF dell'opera o dalla codifica TEI della rispettiva edizione critica. L'utilizzo di grafiche di Data Visualization permette infine di rendere il risultato immediatamente comprensibile. L'analisi comparativa delle modalità correttorie che ne risulta potrebbe favorire la nascita di nuove letture critiche volte a svelare i processi creativi degli autori e valutare la possibilità di considerare le correzioni come pattern ricorrenti. Questa metodologia innovativa si propone inoltre di stimolare gli utenti verso un approccio più consapevole all'evoluzione storica del pensiero creativo servendosi di una rappresentazione immediata dei risultati. L'analisi dei due case-study scelti per lo studio pilota, i *Canti* di Giacomo Leopardi e *I promessi sposi* di Alessandro Manzoni, permetteranno di

comprendere le differenze esistenti tra i due autori chiave del Romanticismo italiano e tra poesia e prosa intese come dispositivi narrativi.

Introduction

Can authorial manuscripts, also known in Italian as "scartafacci", reveal something about the creative thinking of an author? Discarding the teleologism that characterized traditional philological hypothesis, the revolutionary approach to literary texts furthered by Gianfranco Contini ([12]) made it possible to shed light on the hermeneutic, linguistic and philosophical value of authorial variants and to consider them as clues of the subjective approach to the act of writing.

While boasting a century-old tradition, it is only due to the foundational contribution of Dante Isella that Authorial Philology has been recognized as a subject in its own right compared to the more traditional "philology of the copy", which differs in terms of purposes and methodologies. The first essential distinction between the two disciplines concerns the conception of the *critical apparatus*, which also reflects their different goals: while in Textual Criticism it is aimed at representing in a synthetic and transparent way the variants of the witnesses in the process of reconstruction/amendment of the archetype, to allow the reader to understand the choices of the publisher in the total absence of evidences of the author's ones, the method of Authorial Philology shows all the genetic phases of the text, from the very first creative idea of the author until his or her last will.

Other than providing the discipline with an official definition, in his 1987 essay *Le Carte Mescolate. Esperienze di Filologia d'autore* ([18]), Dante Isella developed a formalization method to represent authorial corrections based on their critical interpretation that has served as a model for all the following critical editions.

Since manuscripts are made up of many heterogeneous elements that must be compared to be efficiently analyzed, Isella grouped variants into different categories based on the function they performed in the manuscript and provided an individual representation for each of them. Genetic and evolutive variants, which are those affecting the final reading, are distinguished from, on the one hand, alternative or intra-textual variants, that co-exist with the text as the author had not yet reached an undeniable decision about them; on the other hand, meta-textual notes, which constitute a separate category of linguistic references and critical or compositional notes written by the author ([20]).

The streamlined system developed by Isella made it possible to represent all the detected correcting phenomena: exclusive abbreviations are associated with single variants (such as deletions, insertions, displacements, variants taken from others, overwritten and underwritten variants), whereas numeric and alphabetic exponents identify phases and sub-phases that allow tracing the evolution of syntactic units.

The resulting *apparatus* allows scholars to represent all the correction phenomena in a diachronic and systemic way, granting the reader the possibility of identifying the stratification of the editorial process. Thanks to this innovative framework, critical apparatuses do not present corrections transcribed in a "diplomatic" form but are rather critically interpreted, significantly diverging from the model used by similar philological traditions such as *Critique génétique* ([23]).

Given the variety of available documentary material and case studies in the Italian literary tradition, and the tendency to study them in an isolated way, a comparative analysis of the process lying behind the composition of these works from the perspective of Authorial Philology has never been carried out before. Nevertheless, this approach would prove useful to enter the author's workshop and deeply understand their poetics and the creative process starting from material evidence.

Recent studies have shown the possibility to find patterns in correction on such grounds. Starting from the consideration of manuscripts as "geographical maps", used to keep track and categorize the elaborative processes of the authors, Paola Italia ([22]) proposed a first typology based on some famous cases of Italian literature and highlighted the possible ecdotic implications of such approach. A first distinction to be outlined is that among the types of authors who correct "by map" or "by compass": the essential difference, therefore, lies in the planning of the writing that translates into the manuscript page, transforming it into a conceptual map or into the battlefield of ideas that compete for the physical space of the page.

Since authorial corrections can be considered clues to investigate the creative thinking of an author, performing a comparative analysis of the different correction modalities would thus allow to assess whether authorial interventions depend on personal attitudes or are influenced by external events.

Authorial Philology and Digital Approaches

The digital turn has led to a real methodological revolution in philological research emphasizing the dynamic character of literary texts and their social value. Nevertheless, because of the apparatuses' complexity and their relatively recent affirmation, so far, digital approaches to Authorial Philology have been mainly focused on investigating the shift that the brand-new editorial practices have produced in the centuries-old tradition of Textual Criticism, either by enhancing it or by emphasizing its criticalities, through the aid of scientific disciplines and new technologies. In particular, almost all the attempts in this field have been aimed at representing the history of the textual tradition, rather than at reconstructing the process behind an author's rewriting interventions on their texts.

Instead of trying to render the textual movement as a whole, Variants Mining focuses on its minimum underlying entities, namely authorial interventions, to emphasize their importance in defining the evolution of the author's self-representation.

This work thus deviates from this widespread approach for it results reductive when compared to the whole process of interpretation of a text, which provides a complete understanding of a literary work beginning from the representation level. Accordingly, Variants Mining aims at taking advantage of the current coexistence between analog and digital philology by exploiting the transition process using traditional formulas of philological apparatuses to provide a comprehensive evidence-based analysis of pre-established categories of authorial corrections, starting either from the PDF version of the critical apparatus of literary works or from their TEI-encoded versions, when available.

It sets an even more ambitious goal compared to current developments in the field, that is, to prove that the digital medium could enhance the representation of authorial variants, and also favor quantitative analyses, allowing scholars to understand more intuitively how an author works, and gain empirical insight on their philological hypotheses in the awareness that this could contribute not only to extend traditional philology to new perspectives of analysis but also facilitate the social democratization of philological knowledge.

The rationale behind the apparatuses of Authorial Philology, together with the high standard of technical development reached in the XX century, makes it worth the attempt to analyze them using computational methods. By exploiting a highly multidisciplinary approach that integrates philology, cognitive philology, and computer science, and uses new technologies as methodological framework of analysis and as means of information visual representation, the project not only aims at proposing innovative theoretical insights, but also at achieving scientific verifiable results through an evidence-based analysis and empirical insight into the material premises of creativity.

Related work

Automatic processing of a critical *apparatus* is a brand-new field of research. However, similar work concerning the automatic representation and processing of corrections, derived by a custom TEI model of the text, has already been carried out starting from different premises and a different purpose. For instance, in the *Samuel Beckett Digital Manuscript Project* (SBDMP, [36]), the number of added, deleted and modified words is displayed through a pie chart for each genetic edition of Beckett's works, along with the possibility to compare one another, first drafts versions and a broader view of the full genetic dossier. This choice of visualization corresponds to complex editorial principles which mirror Beckett's poetic anxiety and attitude towards thematic despair and failure topic-wise. Both additions and deletions are indeed encoded in the project with attributes referring to the author of the intervention (Beckett, anyone or unknown): its place (up to nine possibilities), the writing tool used and a particular "type" value for the case of alternative variants. Specific tags are also employed as children of the main ones to cover the cases of restorations of deletions (the "restore" tag), or to add annotations to signal further information about the text ("metamark").

Another proposal of automatic processing of corrections concerns the automatic identification of types of alterations in historical manuscripts for the case study of the digital scholarly edition *Berlin intellectuals 1800-1830* ([26]), a collection of letters and texts representing communication conditions and developments for intellectuals in that specific historical and cultural time frame. The input for this analysis was a custom diplomatic TEI transcription of the manuscripts. Like the SBDMP, in fact, the digital edition gives value to interoperability, so that a reader can both see additions and deletions of the text as well as material elements of the manuscript relevant to its contextual comprehension ([2]). The analysis thus exploits recent machine learning techniques to help categorize content-related alterations into mistakes, stylistic and moral changes from such a labelled dataset.

Both projects show the multiple perspectives of analysis that can be derived from a manual TEI encoding of the authorial corpus, as the transcription and its automatic analysis show the attempt to find a balance between objectivity and applicability to the specific resource and, at the same time, between a general impartial transcription of the object and the availability of critical elements to enable its understanding. Apart from the difficulty to achieve an efficient solution in a short time period, problems of this operation concern its being time consuming and requiring dedicated work.

Instead, by exploiting the categories of Authorial Philology, Variants Mining allows individual scholars to extract immediately granular information, which can be then grouped into aggregate categories of analysis, and using open-source tools only, hence the use of digitized critical editions—whose choice would fill the gap between born-digital texts and printed ones— will be taken as the main reference for the analysis and the proposal of a TEI model will be outlined in the view of scalability.

So far this task requires a completely different approach to critical apparatuses, which is based on the ways corrections are made by the author rather than on their content thus designating a "style" of corrections.

Moreover, the lack of related work in the field requires addressing a theoretical problem that is implicitly the elephant in the room in the field of computational studies applied to literary sources. Text criticism carried out on the results of a computational analysis claims an arbitrary relationship between formal patterns and interpretation that is difficult to justify. In fact, formal patterns are products of *a posteriori* interpretations, where the term has every time its own meaning, thus a different position towards objectivity. It is true this sense of computational approach is different from the traditional sciences', and it has to be treated as such.

Methodology

The first step of the methodology was to choose the categories of correction suitable to the analysis. Thanks to a first qualitative analysis based on close reading and the principles of Authorial Philology, categories of correction have been identified, which have then been grouped into macro categories (immediate variants, late variants, additions). Whereas immediate variants

are distinguished by late ones due to the indication of an intervention at the time of the composition[1], additions are signs of a thematic or linguistic innovation.

PDF preprocessing and automatic analysis of the critical apparatus

Because one of the main purposes of Variants Mining was to bridge a gap between paper and digital editions, there were two main inputs for the analysis—the PDF version of the considered work and its encoding in TEI. While the latter has been already exposed in De Biase, Lippolis and Totaro ([14]), the development of the project on the basis of the digitized document requires further explanation steps.

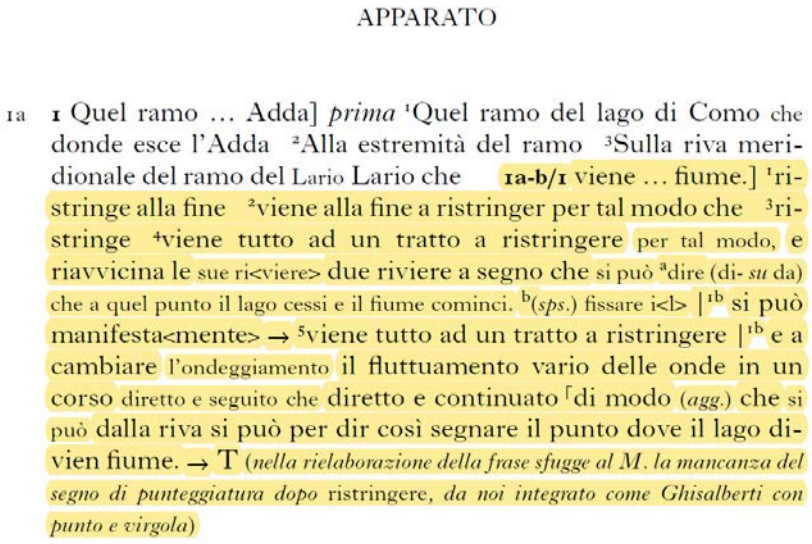
Before delving into the analysis, it is important to mention that PDF, which stands for Portable Document Format, is not meant to be processed. It was created by Adobe in 1993 as a layout-based format that preserved content and configuration of documents in view of the print while being able to be exchanged platform-independently. As a result, it has been defined as a "collection of objects which describe how one or more pages must be displayed" ([4]: 4): the semantic building blocks of the text are non-existent, since the data related to them is conceived as characters independent from one another, each with a specific position on the page. As such, while the human eye naturally recognizes words, paragraph boundaries and semantic roles within the page, text understanding for the machine stems from instructions grouping letter sizes and positions into words and paragraphs. In this way, content ordering and presentation ordering are two separate levels of document identification.

Furthermore, as PDFs can include nonstandard fonts and encoding, as well as extra spaces between letters or, vice versa, not enough space, it was necessary to convert the file in XML format so that every information related to each character could be preserved precisely as it was. In the case of the apparatus of the critical edition, this meant to maintain superscripts of sequences and the smaller corpus of the text for immediate corrections within sequences, while with a simple PDF-to-text conversion the former would turn into common numbers that could be confused with others in the text and the latter would be outputted in the same size as the text.

Among the different Python libraries aimed at PDF text extraction, PDFminer has been considered the most efficient one not only for its very accurate text rendering, but also for the possibility to directly convert the format into an XML file and the constant maintenance of the "six" version. Despite the incorrect handling of ligatures, characters with diacritics, hyphenated words and tabular content ([3]: 6), the library has the great advantage of being highly customizable: to properly deliver the right order of characters as corresponding to the document layout, the function variable LAParams() collects specific parameters that can be changed in accordance with the structure of the text taken into consideration. In this way, it was possible to get an XML file that preserved the characters' position and essential information related to them (size and font), necessary to identify the typographical features that are at the core of the formalization method of Authorial Philology. A first look at the converted document made it

[1] These considerations have been based on the critical editions of the exhamined authors. Dubious variants have not been taken into account for the analysis.

clear that what were superscripts and words with minor corpus in the text, in the XML file had different sizes that have been uniformed so that no category of correction could have been missed. Nevertheless, the major obstacle to precision concerned the beginning of a new line. In fact, as the "textline tag" correctly identified every line of the text, it forced readings multiple lines long to be separated every time a new line started.



APPARATO

1a **1** Quel ramo … Adda] *prima* [1]Quel ramo del lago di Como che donde esce l'Adda [2]Alla estremità del ramo [3]Sulla riva meridionale del ramo del Lario Lario che **1a-b/1** viene … fiume.] [1]ristringe alla fine [2]viene alla fine a ristringer per tal modo che [3]ristringe [4]viene tutto ad un tratto a ristringere per tal modo, e riavvicina le sue ri<viere> due riviere a segno che si può [a]dire (di- *su* da) che a quel punto il lago cessi e il fiume cominci. [b](*sps.*) fissare i<l> |[1b] si può manifesta<mente> → [5]viene tutto ad un tratto a ristringere |[1b] e a cambiare l'ondeggiamento il fluttuamento vario delle onde in un corso diretto e seguito che diretto e continuato ⌜di modo (*agg.*) che si può dalla riva si può per dir così segnare il punto dove il lago divien fiume. → T (*nella rielaborazione della frase sfugge al M. la mancanza del segno di punteggiatura dopo* ristringere, *da noi integrato come Ghisalberti con punto e virgola*)

Figure 1: Sample of the apparatus showing a reading multiple lines long

As a matter of fact, the first big difficulty of *apparatus* processing concerned the gap between each formalized typographical representation of the text, which resulted very clear to the human eye, as opposed to the impossibility to give a solid basis for the algorithm to understand it. Indeed, because the readings were separated by a bigger space than the one between words — although it did not correspond every time to the same number—, they were able to be isolated. However, there was no key point besides an occasional hyphen to rely on when trying to recognize when a reading was longer than one line (Figure 1). The same issue emerged when a new page started, since even deleting the page marks would have maintained a new line. This prevented long sequences to be recognized as one occurrence instead of multiple. The only partial solution was to identify the hyphen between a line or another and specify in the algorithm to delete it and consider both lines as a single reading. As for the other cases, no fix has been found.

After grouping every character into words, and the words into readings as they were on the critical *apparatus* through the calculation of the gap between "bounding box" attributes, punctuation and round brackets have been deleted as they could invalidate the regular expression patterns that were to identify the smallest units corresponding to categories of corrections. As such, two dictionaries have been created to improve precision in the distinguishing of single variants and complex ones, that is to say, one with all the text contained by the tag "textline" that had at least

one occurrence of the font dimension of the superscript, which indicated the existence of sequences, and the other with the rest.

It was thus straightforward, after deleting punctuation and round brackets that could invalidate our regular expression patterns with the algorithm functions *clean* and *clean_sequences*, to identify a first pattern that matched at once the bracket, the whitespace and the word of the category of correction. In these cases, symbols and numbers were identified as they were, while whitespace has been represented by "\s", with an eventual asterisk indicating when more than one could be present, and multiple words with "\w".

Identifying patterns of sequences was a way more complex task than that of the other categories of correction. For instance, to isolate a reading group of three elaborations with text reuse the pattern had to take into account all the possible combinations:

] + whitespace + 1 + words (numbers excluded) + 2 + words (numbers excluded) + → + T or 3T;

] + whitespace + 1 + words (numbers excluded) + → + 2 + words (numbers excluded) + T or 3T;

] + whitespace + 1 + words (numbers excluded) + → + 2 + words (numbers excluded) + → + T or 3T.

Which, translated into regular expressions became:

Find_sequence_1 = re.compile(r"\]\s*1\s*([\w\s]+)\s*2([\w\s][^3]+)\s*→\s*T");

Find_sequence_2 = re.compile(r"\]\s*1\s*([\w\s]+)\s*→\s*2([\w\s]+)\s*(T|3\sT)");

Find_sequence_3 = re.compile(r"\]\s*1\s*([\w\s]+)\s*→\s*2([\w\s]+)\s*→\s*(T|3\sT)[^→]").

This process became very complex if we consider the presence of sequences with up to ten elaboration stages.

After retrieving all the matches, the Python library Pandas was used to store all the numbers of the frequencies in a dataframe, a two-dimensional data structure, which can then be converted into different formats of spreadsheets to be saved locally. In this case, the two dataframes that resulted from the algorithm —one for the results of the sequences, one for the rest of the categories of corrections —have been saved in the Comma Separated Value (CSV) format, with the names of the categories as columns and frequencies as values. Data has been grouped to create a chart with the macro categories of correction and one referring to the use of text derivation in both authors. To make the results comparable, relative frequencies have been computed for each opera, making an average calculation of the works of reference.

Data visualization process

Data, information and knowledge are three terms commonly used interchangeably, since they all serve as both the input and the output of a visualization process ([11]). The huge quantity of data available on the internet and the fact that in the last thirty years psychology has unveiled many brain mechanisms involved in visual studies ([29]; [39]; [33]) have made visualization of various quantities of information a dominating practice nowadays for conveying knowledge.

Although the purpose of data visualization is indeed to display, explore and discover information where people can draw insight on, Yi et al. ([42]) highlight the discrepancies in literature in the definition of the term "insight" and provide an explanation of such a process in four steps: provide overview about the data, adjust to filter and group relevant information, detect patterns and trends and match a mental model, that is, to reduce the distance between the data and the user's mental model of it. As data visualization is defined as "functional art" ([8]), in that design is tied to the purposes of the presentation, it can be thus seen with the same traits of a technology: a mean for an audience to help accomplish specific aims. In view of these studies, after the algorithm of Variants Mining collected and ordered relevant data on the corrections made in the first and second draft of *I promessi sposi* and Leopardi's *Canti*, it was necessary to display it in a way that made it easy to be interpreted under different points of view and help the reader understand its patterns (see Figure 2 and Figure 3).

It was thus possible to gain a general overview on data, thanks to the various options offered by Flourish, an open-source visualization website. Although the data distribution could have been displayed directly from Python, the programming framework offered a narrower range of choices than Flourish and the results lacked interactivity. For the same reasons, although RawGraphs could have been a good representation alternative, it has not been chosen.

The first attempt at visualization concerned the representation of the categories extracted from the algorithm, without any grouping other than the division of immediate and late sequence into single and multiple to avoid excessive confusion in the chart. Because the categories were still too many to gain insight on the data, an aggregate system of representation has been tried out. As A. Cairo states, in fact, "graphics should not simplify messages. They should clarify them, highlight trends, uncover patterns, and reveal realities not visible before" ([8]: 79). Following this principle and Shneiderman's Information Seeking Mantra, according to which "overview first, zoom and filter, then details-on-demand" ([38]), another exploratory representation has been carried out, by grouping the data in immediate and late corrections, both single and in sequence; along with additions. Because the categories were much less, it was also possible to include relative frequency values for each one in the visualization. Thanks to the Pyramid chart, which allows intuitive comparison, the result is much clearer, considering that Flourish makes it possible to filter out selected categories.

Provided that it is highly suggested ([37]) to use multiple coordinated views to represent all the necessary information, another chart concerning the difference in categories with reuse and those without has been developed for each work of the author. As a result, the following charts have been developed:

- Macro categories of correction (immediate variants, late variants, ascriptions, additions);
- Comparison of categories with text derivation and categories without it.

In order for the analysis of Variants Mining to be publicly accessible, a Github repository with open license has been created.

Case studies

Giacomo Leopardi's *Canti* and Alessandro Manzoni's *I promessi sposi* have been selected as case studies to be compared. While the two works belonging to different literary genres may make it seem that the influence of the narrative devices are decisive on the results of the work, recent studies ([14]) show that the division between immediate and late variants do not depend upon prose and poetry. Apart from the two authors being among the most important ones in Italian Romanticism, what lead to their choice was also the availability of their works' comparable critical editions.

Giacomo Leopardi

The illustration of the development of Criticism of Variants explored so far has been useful to highlight the essential value of the first critical editions as they would later become the theoretical base of Authorial Philology and, vice-versa, to show how much they served as a stimulus for the evolution of the discipline itself.

With his almost secular history, the critical edition of Leopardi's *Canti* constitutes the pivotal work that the most recent Authorial Philology and its critical counterpart interfaced with. The four critical editions published in Italy from the late 20's up to 2006 witness, on the one hand, the importance of such poetry collection in the Italian literature and, on the other hand, the different layouts and editorial practices that were implemented to deal with the textual problems presented in it.

The edition of Leopardi's *Canti* provided by Moroncini in 1927 represents a turning point for both the philological interest in Leopardi's works and the new philology, because it constitutes the first attempt at scientifically representing authorial corrections retracing the twenty-years-long history of the text. As Moroncini himself underlined in the Proemial Discourse ([27]: LVII), a conspicuous tradition of studies preceded its edition, showing, therefore, how the critical interest in Leopardi's paper-foul originated between the late XIX and the early XX century, long before the affirmation of the criticism of variants.

Because of the few available autographs, although less remarkable than the pioneering edition by Moroncini about the achieved results, the previous works bear evidence of the first attempts at publishing both the handwritten and printed variants (through the graphic representation of the autographs), and to tackle them using a critical approach.

Relying in part on these embryonic attempts, Moroncini created an indispensable tool to study variants in Leopardi's works. As De Robertis wrote in his commentary on the Leopardi's Canti "With this edition, Authorial Philology accomplishes a decisive step forward […] and Leopardian studies enter a new dimension opened by the complete revelation of the poet's work around the text" ([28]: IX).

Leopardi's manuscripts reveal his attention to detail in their appearance, which is the main distinctive feature of his writing. While the nature of Leopardi's works, considered as transcriptions rather than first drafts, is well known, the many corrections and variants added by

the author while preparing the copy for typography, witness that the book grew over time, progressively incorporating the poet's ongoing meditations and stylistic experiments.

In this regard, Gavazzeni ([17]: 409-420) provided a detailed description of Leopardi's method, which proved fundamental to identify a steady pattern in self-emendations of the author's creative thinking. Furthermore, his insights helped scholars to shed light on the many corrections and re-elaborations processes Leopardi's poems have continuously been subjected to over the years.

The distinctive features of each manuscript mirror the poet's ongoing meditations, presenting issues related to the specific poetics of which they are witnesses. This is particularly evident when it comes to comparing the manuscripts of the Canzoni, published in 1824 in the "*Neapolitan notebook of the Idylls*". On the one hand, a series of heterogeneous elements (namely corrections, variants, annotations, bibliographical references), comprehensively known as varia lectio, appears on the margin of the sheet and testifies the substantial philological work that later led the author to outline a new poetic language; on the other hand, the significant decrease of such components in the manuscript of the Idylls allows scholars to infer the homogeneous character of these compositions considering the whole system.

The series of Leopardi's poems entitled *Idylls* appeared in its earliest extant forms in the autograph manuscript known as "*Neapolitan notebook of the Idylls*". The work consists of a booklet with a lined page on which the author's handwriting is clearly and sharply visible. Written in separate phases, the poems were then corrected by the author using identifiably different pens before being copied into Visso's manuscript, where they were again revised.

Gavazzeni took advantage of the digital reproduction of the "*Neapolitan notebook of the Idylls*". In fact, he used it to specify all the different corrective campaigns relating to each of the three compositional phases of the texts and did so by identifying the stratigraphy of the ink, deepening the intuitions originally furthered by Moroncini, and then shared by Peruzzi e De Robertis. Accordingly, four different pens have been identified, namely A, B, C, and D, which were used to write the new texts and correct the previous ones.

The highly dynamic structure of this poetry collection is also reflected in the publishing history of the book. Although a detailed illustration of the various reprintings is out of the scope of this work, both a synchronic and diachronic analysis of print witnesses is fundamental because it allows for an understanding of the overall design of a strongly unitary work. As noted above, the textual history of Leopardi's *Canti* is fascinating because much of the Italian Authorial Philology has been based upon it. In this regard, the latest edition provided by Franco Gavazzeni constitutes a thorough guidebook for a complete understanding of this kind of approach, for it provides users with the critical edition of both manuscripts and printed versions, plus detailed textual notes recording all handwritten changes, as the abovementioned notation on the different pens.

Alessandro Manzoni

Like Leopardi's case, the huge attention received by Alessandro Manzoni's editions of *I promessi sposi* was at the same time drive and symptom of a new phase of literary studies, when it became

necessary to rethink about the national literary heritage and the works that conveyed a sense of identity both politically and linguistically.

As a matter of fact, this novel was already a subject of research when it came out, starting, for instance, from the project of linguistic comparison between the two editions of the work carried out by Giovanni Battista De Capitani straight after the *Quarantana* edition got published ([35]: 21) or Gilberto Boraschi's index of corrections relating *Ventisettana* and *Quarantana* editions ([6]). Considered the foundational text of the Italian literary tradition, *I promessi sposi* was indeed the result of a deep reflection, proven at first by the long period of time in which the development of the work had been carried out.

Manzoni's relationship with the papers he wrote seems to mirror his own personality both in an ethical and stylistic sense. In fact, the fast, rough and continuously renovated work of writing carried out right after thorough reflections was counterbalanced by long periods of meditation and a reserved character that made him reticent to public interventions. Although the author did not leave any autobiography, he made his works talk for himself, in the context of an antiheroic attitude that privileged the collective dimension and the pedagogical purpose not only in his literary pieces, but even in the management of his own beloved library ([35]: 43).

Ever since the author's early writings, literary activity has been considered ethical and centred on the interlacement of truth, utility and beauty. However, the means to reach this theoretical purpose were to be experimental, through confrontation in structure and language, as an inherent dialectics between the reflexive moment and the creative one constituting a "genetic principle" ([7]: 19): not by chance all Manzoni's most important theoretical essays follow a creative circumstance.

In this sense, the dialogic dimension is of primary importance in the whole work of Manzoni, not only because it is possible to witness the priority he attributed to the written word—he composed a lot and drew very little ([35]: 17) —, but most of all, for the attention he paid to the debates of the time and the discussions with his friends, for the interplay of mutual genesis and influence between his works and even for the connections of his creations to the several documents he browsed, in a dialectical perspective of knowledge acquisition and transmission that was considered superior to the mere enunciation of facts. Accordingly, the second Parisian stay of 1819-1820 was decisive for the maturation of the reflections that pervaded the composition of this second tragedy and the start of *I promessi sposi*: Manzoni had been trying to reach the French capital for some time, hoping to escape his political concerns that caused him also nervous breakdowns. In such a cultural turmoil, Walter Scott's Ivanhoe was a ground-breaking work for the liberal intellectual circle near Claude Fauriel. At the time, the novel was becoming widespread in France, in an environment of reevaluation of history as a mean, along with philosophy, to fight the French absolutism. According to Augustin Thierry's review of the novel, for the first time the events were focused on the reactions of individuals to the major social and national circumstances, no more on the private passions of single characters.

It is not by chance, then, that Manzoni chose to follow an analogous path when he realized the problems of *Adelchi* and decided to begin *Fermo e Lucia* (a title derived from a reference of Ermes Visconti in a letter to Gaetaneo Cattaneo, but most likely it was *Gli sposi promessi*, as shown by

[24]), at a time of rise of Italian riots for the independency from the Austrian domain, two years after the cancellation of the review *Conciliatore* and the incarceration in 1821 of his director, Silvio Pellico.

Although it is a first draft and was never published during the life of Manzoni, *Fermo e Lucia* is a complete and definite work that must not be considered as subordinate to its best known revision. Organized in thirty-seven chapters, written between April 24, 1821 and September 17, 1823 according to the two dates that can be found at the beginning and the end of the text, the autograph allows to understand the way Manzoni worked and conceived his novel. It is, first of all, a "pure modern invention" ([30]: 586) that represents real life in the shape of both historical events and the everyday world of private matters doomed to be left behind, thus worth to "live within it" ([15]: 352-353) and to be narrated to a potentially wide public. It is not only a first step towards the democratization of literature, but also the first attempt at opening the path to the realist frontier of the novel, with a focus on how the major events of history have impacted on the people of smaller and forgotten realities ([30]: 5).

Although the difference between *Fermo e Lucia* and the *Ventisettana* was well-known, making the first draft to be considered as an autonomous text, even if unrefined under many points of view, the need to discern and untangle the interventions related to the second draft from those successive to the writing of the *Fermo e Lucia* were equally as recognized, but never carried out in a complete way before the 2012 edition ([31]).

Despite its huge number of interventions, the second draft of *I promessi sposi* had in fact always been considered as subordinate to its printed counterpart, which shares the same narrative structure with, unlike the first and second draft. Instead, it is a necessary document to understand the development of Manzoni's reflection on the relationship between language, truth of the story and truth of the art, especially because of the reuse of the pages of the first draft's first tome, which underwent an intense re-elaboration.

In this perspective, the first eight chapters for the first and the second draft has been chosen as a sample for the Variants Mining analysis because they were comparable on the one hand, but on the other they witnessed a substantial development in the author's creative process.

Leopardi and Manzoni: a comparison between two working methods

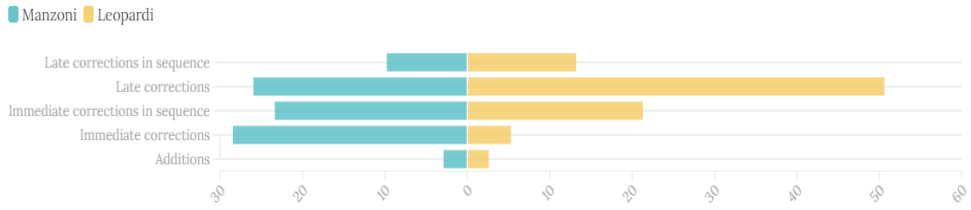


Figure 2: Comparison between macro categories of correction in Leopardi and Manzoni's works (average of the relative frequencies of the selected works)

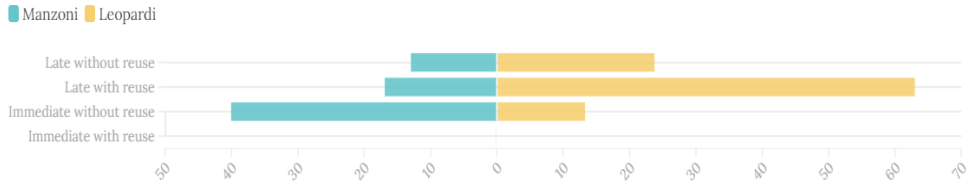


Figure 3: Comparison between categories with or without reuse in Leopardi and Manzoni's works (average of the relative frequencies of the selected works)

The analysis carried out on the chosen examples marks a clear difference between the two authors. As shown in Figure 2, while Leopardi's average frequency of corrections is quite unevenly distributed, Manzoni's drafts seem to have more balance in distribution. The high prevalence of late corrections in Leopardi's selected works clearly shows how his working method differs from Manzoni's. In fact, following the already mentioned pioneering study of Italia ([22]), which considers Leopardi a writer "by map", as he had a specific planning for his compositions, the many late corrections, notes, glosses, annotations, and variants visible in his papers are the most truthful testimony of the effort to achieve what the author calls "minute perfection" ("minutissima perfezione"): it arises from a profound faith in the word as truth, understood not as a mere sign of communication but as an ontological entity, characterized by its physicality.

On the other hand, the elevate number of immediate variants shown in Manzoni's papers induces to think about a method of correction "by compass", as from the confusion of the papers it is clear the author did not have a preconceived plan for his composition, but rather imposed the draft a definitive character that did not belong to it, where his reflections had to gradually focus on the details through prompt corrections despite ignoring what the final text would be. As a result, in the manuscripts of the author it is possible to witness a spontaneous dialogic proceeding of the thought rather than an established intention both in content and language ([35]: 17).

A common aspect between both two writers is the low incidence of additions. As it is evident from Figure 3, Leopardi's works are characterized by the prevalence of categories that entail reuse of the previous text, rather than being an innovation. In the case of Manzoni, the data can be motivated by Giovanni Nencioni's statement, according to whom "the changes made by Manzoni are not innovations that cut a weave and a tone essentially aulic, but touch-ups of adjustment and levelling on a weave and tone fundamentally common already at the beginning" ([32]: 241). A quantitative content analysis of the *apparatus* could shed light on this category's lexical diversity with respect to the remaining text and confirm or disprove these theses.

For what concerns Figure 3, reuse of previous text constitutes the key point to really understand the peculiarities of the two different approaches to the act of writing. In fact, Leopardi's interventions turn out to be strictly influenced by formal constraints that characterize poetry and in particular by the search for a pleasant musicality of the verse and by compensatory techniques due to the introduction of innovative metric solutions adopted by the author, such as the *canzone libera*.

On the contrary, in Manzoni's case immediate corrections without any text derivation are in clear prevalence, proving once again that the author's thought process acted immediately and without a precise direction in mind. It is thus evident from the chart that he tried to concretize and shape his thoughts in the very moment of textual composition, resulting in a high majority of immediate corrections without text reuse that outnumber all the others. Therefore, the obtained data allow to hypothesize an influence of the different narrative devices, namely prose and poetry, on the two authors' elaboration processes, that are on the one hand the mechanisms of the description and both the meter and musicality on the other.

Furthermore, the data also shows that due to the prevalence of immediate variants without reuse, the second draft of *I promessi sposi*, which is included in one of the works that make up the average output of the results, can be considered as a manuscript written in one go like the first one. As a matter of fact, both first and second draft are tendentially written right away, but despite knowing it needed further adjustments, Manzoni considered *Fermo e Lucia* as a text that had to be autonomous. Despite the complete status of the draft and the reuse of some of its parts in the second draft, what had to change was created *ex novo*.

Future developments

As already underlined, one of the primary purposes that were set at the beginning of this work was to extract the information contained in the apparatus directly from the PDF of the reference editions due to the almost total absence, in the panorama of critical digital editions, of experiments in the field of Authorial Philology.

However, the potential of the implications resulting from the analysis of an author's correction categories can expand well beyond the essential, but traditional literary criticism. In fact, this study was born in the wake of the conviction that corrections should be considered the author's fingerprints, i.e., responding to his general attitude towards artistic creation.

As such, the Variants Mining algorithm can also be helpful for the study of authorship attribution, as it is based on the consistent occurrence of specific features in a corpus of texts.

More so, when considering the direct extraction of frequencies from the critical *apparatus* of Authorial Philology, as it accounts for elaboration phases that can also involve multiple lines of texts instead of comparing single terms' rate of variation. In fact, recent studies highlight the importance in authorship attribution detected through machine learning techniques not to consider only word-level features, which can be misleading when trying to detect forgery, but also syntactic elements, as they indicate the writers' way to construct their texts and are more difficult to fake, especially in multi-author documents ([43]). In this regard, Chaski ([10]) suggests that syntactic features are less common in authorship attribution because of the difficulty to extract such information with precision. However, the critical *apparatus* already provides information at this level so that it is not necessary to involve complex but less precise computational studies.

At the same time, researchers have recently been trying new ways to capture this kind of information, such as Kim et al. with a k-embedded-edge subtree ([25]); Tshuggnall and Specht ([40]) with syntax trees or the more traditional dependency grammars ([13]), making it possible to even combine information of the critical *apparatus* with a more and more granular identification of syntactic structures. Another advantage of Variants Mining in this field of research would thus concern the fact there is no need to necessarily encode the text and its variants to obtain a complete analysis for authorship attribution, but rather to perform the same algorithm on both the reference texts and their corresponding critical *apparatus* and consequently draw conclusions on their comparison.

Among the various innovative elements that characterize the method proposed in this work, the study of the critical *apparatus* based on the way corrections are made, rather than on their content, certainly represents the richest implications in terms of theoretical potential. In this context, the study on late style theory can be a fruitful research field to understand whether it can be supported with empirical evidence. If late style is "a distinctive signature that characterizes the end of a career, as the contours of the aging body are mapped onto the weave of writing" ([34]: 147), this change should also reverberate on the way an author corrects their texts. In it being a peculiar approach to authorship, not as something static but rather as a process of change, "late style" seems rather fitting to the study of authorial variants, as Contini, when facing the problem of the meaning of a study of variants, defined the text itself as a process whose dynamism has to be taken into consideration when dealing with it in a critical way.

Apart from widening the sample corpus to the entire work of the author, a comparative study of the correction methodologies, implemented by expanding the study carried out in this research project, could allow to investigate the presence of similar correction patterns even among very different authors, both in terms of age and field of action.

As it has been already underlined, despite the ecdotic development of Authorial Philology has reached a very high level, especially in the Italian context, it has remained confined to the separate study of individual authors, effectively preventing a comparative study of the authors' compositional and correcting methods (see [5]: 263-264). Accordingly, the extraction of data

from digital critical editions annotated from the PDF of the text or its TEI encoded version allows to extend the analysis to authors outside the Italian context to verify how the dynamics of correction change in different eras and in the texts of authors who reference to areas of knowledge that are also external to the strictly literary one such as the scientific, historical or philosophical one.

The taxonomic categorization of the phenomenology of corrections and their analytical analysis made possible through the proposed methodology could, therefore, contribute to delineate a general theory of corrections: this would allow to consider the concept of creativity from a different perspective, providing empirical evidence that emphasizes its inevitable material and time-dependent premises, discarding the notion of the modern manuscripts as products of the authors' genius ever since the Enlightenment.

Moreover, it must be emphasized that this study aims to configure itself as state of the art in a moment of fundamental transition towards a born-digital literature and the development of archives of digital native texts, such as the very recent ALDiNa, a project born under the aegis of Associazione per l'Informatica Umanistica e la Cultura Digitale, which aims to guarantee adequate documentation and enhancement of the authorial digital native.

The attention to the philological dimension therefore becomes fundamental to ensure respect for the author's last will in a context, that of the digital, which accentuates the fluidity of the text and its ephemeral character.

Therefore, starting from these premises, the approach of the proposed methodology emerges evidently in an anthropological sense, testifying on corrective methodologies of the past while the study of author variants is inevitably moving towards new practices.

Conclusion

A panoramic analysis of the most recent projects developed in the field of digital philology clearly showed how the digital turn has led to a real methodological revolution in philological research. However, the scholars' attempts were mainly aimed at exploiting digital infrastructures to simplify the use of critical apparatuses thanks to the possibility of publishing all the witnesses, thus focusing on the content of the corrections rather than on the way in which they are made.

This paper aims to fill this gap by providing an innovative technique able to deepen the conventional philological readings relying on the one hand on the solid theoretical infrastructure provided by the authorial philology and its critical counterpart, namely the Criticism of Variants, and on the other on the potential of computational and data visualization tools that allow to intuitively understand the obtained data in a more immediate and effective way.

Although pursued thanks to the aid of automatic and quantitative methods, this study's scope is ultimately rooted in literary criticism because it aims to provide evidence-based analysis that participates in the long-lasting debate on creativity by emphasizing the active role played by

authorial corrections in better defining the mechanisms underlying the author's creative thinking.

The analysis carried out on the chosen case-study, namely Manzoni and Leopardi's samples of writing, has revealed important feature of their working method: in particular, the focus on the correction categories allowed to highlight that differences in the elaboration processes are likely to mirror differences in the authors' intellectual disposition such as Manzoni's tendency to proceed by gradual focusing on the object and Leopardi's aptitude to return to the texts at a later time to perfect them.

Moreover, it seems possible to hypothesize that the prevalence of authorial interventions characterized by some reuse of the previous text are determined by some constraints of the poetic texts that also are peculiar of Leopardi's style, such as the attention to the musicality of the verse and the not always linear relationship between meter and syntax.

These results made it clear that the automatic detection of correction patterns must be developed even further. By expanding this study to as many authors as possible, Variants Mining would make it possible to write a grammar of correction which all the different approaches to the act of writing could be referred to.

References

[1] Baillot, Anne and Sabine Seifert. 2013. "The Project "Berlin Intellectuals 1800–1830" between Research and Teaching". Journal of the Text Encoding Initiative [Online], 4.

[2] Baillot, Anne, and Anna Busch. 2015. "Editing for man and machine: the digital edition 'Letters and texts. Intellectual Berlin around 1800' as an example". Users of Scholarly Editions: Editorial Anticipations of Reading, Studying and Consulting.

[3] Bast, Hannah, and Claudius Korzen. 2017. "A Benchmark and Evaluation for Text Extraction from PDF". 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

[4] Blonce, Alexandre, Eric Filiol and Laurent Frayssignes. 2008. Portable document format (PDF) security analysis and malware threats. Amsterdam: Black Hat Europe Conference.

[5] Bonsi, Claudia. and Italia Paola. 2021. Riscrittura, revisione e editing in Storia dell'italiano scritto VI Supporti, forme, pratiche di scrittura. Roma: Carocci.

[6] Boraschi, Gilberto. 1912. I promessi sposi nelle due edizioni del 1840 e del 1825 di Alessandro Manzoni. Milano: Trevisini.

[7] Brogi, Daniela. 2005. Il genere proscritto. Manzoni e la scelta del romanzo. Pisa: Giardini Editori.

[8] Cairo, Alberto. 2012. The functional art: an introduction to information graphics and visualization. Indianapolis: New Riders Publishing.

[9] Cairo, Alberto. 2016. The truthful art: data, charts and maps for communication. Indianapolis: New Riders Publishing.

[10] Chaski, Carole E. 2012. "Author Identification in the Forensic Setting", in Oxford Handbook of Language and Law (Oxford: Oxford University Press): 489–503.

[11] Chen, Min, David Ebert, Hans Hagen, Robert S. Laramee, Robert van Liere, Kwan-Liu Ma, William Ribarsky, Gerik Scheuermann, Deborah Silver. 2009. "Data, Information, and Knowledge in Visualization" in IEEE Computer Graphics and Applications, vol. 29, no. 1 (Jan-Feb): 12-19. doi: 10.1109/MCG.2009.6-.

[12] Contini, Gianfranco. 1970. Varianti e altra linguistica: una raccolta di saggi (1938-1968), Torino: Einaudi.

[13] Covington, Michael A. 2001. "A Fundamental Algorithm for Dependency Parsing", in Proceedings of the 39th Annual ACM Southeast Conference eds. John A. Miller and Jeffrey W. Smith: 95–102.

[14] De Biase, Margherita, Lippolis Anna Sofia and Totaro Gabriella. 2021. "Political Variants Mining: computational investigations on authorial variants", AIUCD 2021 Book of Abstracts (2021), URL: https://aiucd2021.labcd.unipi.it/wp-content/uploads/2021/01/a041.pdf.

[15] Fauriel, Claude and Alessandro Manzoni. 2000. Carteggio. Edited by Irene Botta. Milano: Centro internazionale studi manzoniani.

[16] Fish, Stanley. 2019. "Computational Literary Studies: Participant forum responses, Day 3". In Computational Literary Studies: a Critical Inquiry Online Forum. https://critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-5/.

[17] Gavazzeni, Franco. 2006. "Come copiava e correggeva Leopardi", in Id., Studi di critica e filologia sull'Ottocento e il Novecento. Verona: Valdonega.

[18] Isella, Dante. 1987. Le carte mescolate. Esperienze di filologia d'autore. Padova: Liviana.

[19] Isella, Dante. 2009. Le carte mescolate vecchie e nuove. Edited by Silvia Brusamolino Isella. Torino: Einaudi.

[20] Italia, Paola. and Raboni Giulia. 2010. Che cos'è la filologia d'autore. Roma: Carocci.

[21] Italia, Paola Maria Carmela, "Leopardi e Manzoni: due metodi a confronto", in Di mano propria. Gli autografi dei letterati italiani, Atti del Convegno internazionale (Forlì, 24-27 novembre 2008), eds. Guido Baldassarri, Matteo Motolese, Paolo Procaccioli, Emilio Russo (Roma: Salerno Editore, 2010), 493-519.

[22] Italia, Paola Maria Carmela. 2017. "Carte geo-grafiche: prosatori al lavoro". Autografo 57: 23-37.

[23] Italia, Paola Maria Carmela. 2019a. "Aux origines de la 'Critique des brouillons'". Genesis: 47 – 59.

[24] Italia, Paola Maria Carmela. 2019b. "Un nuovo testimone della Lettera sul Romanticismo". Annali Manzoniani, no. 2 (December): 175-202.

[25] Kim, Sangkyum, Hyungsul Kim, Tim Weninger, Jiawei Han and Hyun Kim. 2011. "Authorship Classification: A Discriminative Syntactic Tree Mining Approach," in Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval: 455-464. DOI: 10.1145/2009916.2009979

[26] Lassner, David, Anne Baillot, Sergej Dogadov, Klaus-Robert Muller, Shinichi Nakajma. 2020. Automatic Identification of Types of Alterations in Historical Manuscripts.

[27] Leopardi, Giacomo. 1927. Canti. Edited by Francesco Moroncini, Bologna: Cappelli. [anast. 1978 con introduzione di Gianfranco Folena].

[28] Leopardi, Giacomo. 1984. Canti, edizione critica e autografi, edited by Domenico De Robertis, Milano: Il Polifilo.

[29] Lohse, Gerald L. 1997. "The role of working memory on graphical information processing". Behaviour and Information Technology, 16(6):297-308. DOI: 10.1080/014492997119707.

[30] Manzoni, Alessandro. 2006. Fermo e Lucia: prima minuta (1821-1821). Edited by Barbara Colli, Paola Italia and Giulia Raboni. Milano: Casa del Manzoni.

[31] Manzoni, Alessandro. 2012. Gli Sposi promessi: seconda minuta (1823-1827). Edited by Barbara Colli and Giulia Raboni. Milano: Casa del Manzoni.

[32] Nencioni, Giovanni. 1993. La lingua di Manzoni. Avviamento alle prose manzoniane. Bologna: Il Mulino.

[33] Nielsen, Cydney B. 2016. "Visualization: A Mind-Machine Interface for Discovery". Trends Genet 32:73-75.

[34] Piper, Andrew. 2018. Enumerations: Data and Literary Study. Chicago: The University of Chicago Press.

[35] Raboni, Giulia. 2017. Come lavorava Manzoni. Roma: Carocci.

[36] Samuel Beckett Digital Manuscript Project, 2019. Accessed January 9, 2021. https://www.beckettarchive.org/.

[37] Shimabukuro M.H., E. F. Flores, M. C. F. de Oliveira and H. Levkowitz. 2004. "Coordinated views to assist exploration of spatio-temporal data: a case study". In Proceedings. Second International Conference on Coordinated and Multiple Views in Exploratory Visualization, 107-117. DOI: 10.1109/CMV.2004.1319531.

[38] Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". In Proceedings of the IEEE Symposium on Visual Languages, 336-343. Washington: IEEE Computer Society Press.

[39] Tory, Melanie, and Torsten Moller. 2004. "Human factors in Visualization Research". IEEE Transactions on visualization and computer graphics, vol. 10, no. 1: 72-84.

[40] Tschuggnall, Michael and Gunther Specht. 2014. "Enhancing Authorship Attribution By Utilizing Syntax Tree Profiles", in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics:195-199.

[41] Van Hulle, Dirk. 2008. Manuscript Genetics, Joyce's Know-How, Beckett's Nohow. Gainesville. University Press of Florida.

[42] Yi, Ji Soo, Youn-ah Kang, John T. Stasko. Julie A. Jacko. 2008. "Understanding and characterizing insights: how do people gain insights using information visualization?". Proceedings of the 2008 Workshop on Beyond time and errors: novel evaluation methods for Information Visualization, no. 4: 1-6. https://doi.org/10.1145/1377966.1377971

[43] Yu, Brian. 2019. Stylometric Features for Multiple Authorship Attribution. Bachelor's thesis, Harvard College.

Last access: 27th March 2021.