

Computing, memory and writing: some reflections on an early experiment in digital literary studies

Giorgio Guzzetta¹, Federico Nanni²

¹ Italian Department, University College Cork

² Department of Philosophy and Communication Studies, University of Bologna

guzzettg@gmail.com, federico.nanni8@unibo.it

Abstract

English. In this paper we present the first steps of a research that aims at investigating the possible relationship between the emergence of a discontinuity in the study of poetic influence and some early experiments of humanities computing. The background idea is that the evolution of computing in the 1960s and 1970s might have triggered a transformation of the notion of influence and that today's interactions between the field of natural language processing and digital humanities are still sustaining it.

In order to do so, we studied a specific interdisciplinary project dedicated to the topic and we reproduced those experiments. Then we compared the results with different text mining techniques in order to understand how contemporary methods can deal with the rethinking of the notion of influence.

Italiano. *In questo articolo presentiamo i primi passi di una ricerca che intende investigare il possibile legame tra l'emergere di una discontinuità nello studio dell'influenza poetica e alcuni tra i primi esperimenti di informatica umanistica. La nostra idea è che l'evoluzione dell'informatica tra gli anni Sessanta e Settanta abbia avviato tale trasformazione nel concetto di influenza e che le contemporanee interazioni tra i campi di natural language processing e digital humanities la stiano ancora sostenendo. Di conseguenza abbiamo deciso di studiare uno specifico progetto interdisciplinare dedicato all'argomento e abbiamo riprodotto gli esperimenti descritti nell'articolo. Quindi abbiamo con-*

frontato i risultati ottenuti con quelli ricavati applicando contemporanei approcci di text mining per capire come tali metodi possano permetterci di comprendere più approfonditamente il ripensamento della nozione di influenza.

1 Introduction

In recent years, when the institutional presence of digital humanities grew stronger, the need for a closer look at the history of humanities computing (the name under which this community of researchers was originally gathered) became more urgent, and some answers have been attempted (Nyhan et al., 2012; Sinclair and Rockwell, 2014). Rehearsing the history of humanities computing proved to be a challenging task, because of the hybrid and interdisciplinary nature of it and because of its entanglement up with a field like computing, which is epistemologically unstable and interdisciplinary as well (Castano et al., 2009).

This paper is a contribution to this attempt to develop an history of humanities computing, trying to combine together the histories of computing, of computational linguistics, and of literary studies. We present here the first steps of a research that aims at investigating the possibility that the emergence of a “discontinuity” in the notion of literary influence (and, consequently, of literary source), or at least the critical awareness of it, might be related to some early experiments of humanities computing. The background idea is that the evolution of computing in the 1960s and 1970s might have contributed to this transformation of the notion of influence, and that today's interactions between the field of natural language processing and digital humanities might have further developed it, possibly in new directions. In order to do so, the paper is organised as

follows. First of all, we define the problem of influence from the point of view of literary studies and literary theory. With that background in mind, we study a specific interdisciplinary project dedicated to the topic (Raben, 1965) and we analysed Raben key-role on Goodman and Villani (1965) computational analysis. Moreover, to get a complete understanding of the methods applied in this work, we recreated the same approach with a script in Python.

Then, we decided to update Goodman and Villani's approach by adopting contemporary text mining methods. While conducting this specific task our purpose was not only to compare the different techniques and to highlight possible mistakes in Goodman and Villani's approach, but we aimed especially at understanding whether contemporary methods commonly adopted in natural language processing and digital humanities are able to answer questions and solve problematic issues emerged during the rethinking of the notion of influence in recent years.

2 The problem of literary influence

In the 1970s Harold Bloom “developed a theory of poetry by way of a description of poetic influence, or the story of the intra-poetic relationships” (Bloom, 1997), introducing the notion of “anxiety of influence” as a feature of creative writing. After him, critics used the term influence “to designate the affiliative relations between past and present literary texts and/or their authors” (Renza, 1990). That poets were influenced by previous masters of the craft is not a new idea, in fact it is as old as poetry itself. What is new is Bloom's aim to “de-idealize our accepted accounts of how one poets helps another” (Bloom, 1997). The traditional humanistic notion of intra-poetic relations “performed a conservative cultural function” because critical assessment of it was “focusing on the ways literary works necessarily comprise revision or updating of their textual antecedents”, emphasising “homogeneity” and “continuity” in Western (canonical) literature:

Grounded in nineteenth-century philological notions of historical scholarship [...] this tradition-bound position regards literary influence as a benign, even reverential, endorsement of humanism: the ongoing project to transform

the world into the image and likeness of human beings (Renza, 1990).

The past was not at all a burden in this humanistic tradition. Since the 1970s, though, thanks mainly to *The Anxiety of Influence* (Bloom, 1973), it became one. A few years earlier, in 1970 - but we have to consider that the first draft of Bloom's *Anxiety* was written in 1967 - Walter Jackson Bate, whom Renza considers Bloom's main critical precursor, introduced a discontinuity in this traditional view of influence, discussing the burden of the past in English poets:

[Bate] concedes that the mimetic view of influence pertains to the major portion of Western literary history. Only in the eighteenth century does the poets first suffer “the burden of the past”; only then does he experience a “loss of self-confidence” about what to write and how to write it “as he compares what he feels able to do with the rich heritage of past art and literature” (Renza, 1990).

3 Trasferring of memory and language: Raben's project

In 1964, during one of the first conferences announcing the coming of age of Digital Literary Studies, the *Literary Data Processing Conference*, sponsored by IBM, Joseph Raben, founder and main editor of *Computers and the Humanities* in 1966, described an interdisciplinary project in which computers played a significant role. Raben's project is interesting because it introduced some differences from a tradition that by then was consolidated (even though far from being a mainstream literary trend), involving the use of statistical methods to define and describe the style of an author. The two main elements of this tradition, that dominated the first experiments in literary humanities computing, are concordances on one side and authorship attribution based on stylometry on the other. Both were developed between late nineteenth/early twentieth century, and in both we can notice an essentialist idea of style, sort of a writer's trademark quantifiable from a linguistic point of view. With this experiment they were not trying to define the character of an author, a notion still essentialist and stiff; instead, they were

more interested in the agency of the writer that reuse the material at his disposal, transforming and manipulating it for his own purposes. The identity of the writer is not therefore an essence, but an active process of rewriting and remediation.

3.1 A specific case study

In his paper Raben described a specific case study based on the automatic detection of similarities between Milton’s *Paradise Lost* and Shelley’s *Prometheus Unbound*. The aim of Raben was “to illuminate the relationship of Shelley and Milton”. To do this, he discussed Shelley’s ideas on poetry and imitation, which were both considered “mimetic art”:

It creates, but it creates by combination and representation. Poetical abstractions are beautiful and new, not because the portions of which they are composed had no previous existence in the mind of man or in nature, but because the whole produced by their combination has some intelligible and beautiful analogy with those sources of emotion and thought, and with the contemporary condition of them: one great poet is a masterpiece of nature which another not only ought to study but must study. (Raben, 1965)

Combination of linguistic expressions previously used by great poets was indeed part of the creative writing process. Shelley himself was doing that in his own creation, as Raben tried to demonstrate using computational techniques to confront his work with that of Milton:

Shelley is not merely echoing Milton’s language: he is absorbing the sense of Milton’s lines, transmuting it and re-expressing it in some of the same words in which it was originally invested. (Raben, 1965)

The ambition was that of entering in the (re)writing process as a creative endeavour, giving a very detailed description of the way in which Shelley reused and reshaped images and vocabulary from Milton.

4 Detecting similarities between two texts

The method adopted in order to detect textual similarities between *Promethus Unbound* and

Paradise Lost has been described by Goodman and Villani (1965), the two researchers who helped Raben in his work.

The authors described in detail the different steps of their analysis. First of all, they clarified the specific reason of their focus on detecting sentence-based similarities. In fact, Goodman and Villani remarked that they preferred to compare sentences instead of lines, as the former usually contain “a whole idea”.

For every sentence each token was lemmatised (“reduced to its root”) and stop-words were removed. To detect similarities between sentences, documents were then merged in a combined sentence-term matrix. By using it, results were ordered following the number of common words between two sentences from different books.

In order to re-create Goodman and Villani’s experiment we used a set of tools available in Python NLTK library. We performed tokenization, stop-words removal and lemmatization¹. Then, we detected the sentences that have more words in common and we ranked the results following this value. By doing this, we confirmed Goodman and Villani’s discovery, in fact we retrieved as one of the first results the two sentences presented in their paper (see Fig. 1).

All else had been subdued to me; alone The soul of man, like unextinguished fire, Yet burns towards heaven with fierce reproach, and doubt, And lamentation, and reluctant prayer, Hurling up insurrection, which might make Our antique empire insecure, though built On eldest faith, and hell’s coeval fear; And though my curses through the pendulous air, Like snow on herbless peaks, fall flake by flake, And cling to it; though under my wrath’s night It climb the crags of life, step after step, Which wound it, as ice wounds unsaddled feet, It yet remains supreme o’er misery, Aspiring, unrepressed, yet soon to fall; Even now have I begotten a strange wonder, That fatal child, the terror of the earth, Who waits but till the destined hour arrive, Bearing from Demogorgon’s vacant throne The dreadful might of ever-living limbs Which clothed that awful spirit unhehld, To redescend, and trample out the spark.	The conquered, also, and enslaved by war, Shall, with their freedom lost, all virtue lose, And fear of God—from whom their piety feigned In sharp contest of battle found no aid Against invaders; therefore, cooled in zeal, Thenceforth shall practise how to live secure, Worldly, or dissolute, on what their lords Shall leave them to enjoy; for the Earth shall bear More than enough, that temperance may be tried. So all shall turn degenerate, all depraved, Justice and temperance, truth and faith, forgot; One man except, the only son of light In a dark age, against example good, Against allurements, custom, and a world Offended. Fearless of reproach and scorn, Or violence, he of their wicked ways Shall them admonish, and before them set The paths of righteousness, how much more safe And full of peace, denouncing wrath to come On their impotence, and shall return Of them derided, but of God observed The one just man alive: by his command Shall build a wondrous Ark, as thou beheld’st, To save himself and household from amidst A world devote to universal wrack.
--	---

Figure 1: Two sentences with a high similarity score following Goodman and Villani’s approach. The one on the left is from *Prometheus Unbound* and the other from *Paradise Lost*.

¹These algorithms have been trained on contemporary documents. For this reason we conducted a post-evaluation of the performances and we noticed that, for the purpose here described (tokenization, stop-words removal, lemmatization and later part of speech tagging) their performances could be in general considered solid. For what concerns stop-words removal we excluded four other extremely frequent tokens (“thou”, “thy”, “ye”, “thro”). In the future we intend to improve this step of our work by developing in-domain tools.

4.1 A different approach, pt. 1

In order to get a better understanding of Goodman and Villani's approach, we compare their results with a standard natural language processing procedure for detecting similarities between two sentences. We re-ran their experiment considering each sentence as a tf-idf word vector (Sparck Jones, 1972). We removed extremely short sentences (with less than 4 words) and we use NLTK Part of Speech tagger to avoid homonyms. Then we computed the cosine similarity between each passage and we ranked the results according to that value.

In making these changes our aim was to normalise the word-frequency by the length of the sentence, in order to avoid that extremely long sentences would come out as first results (as in Goodman and Villani's work).

By using the inverse document frequency, sentences which have in common words that are rare in the books were ranked in higher positions.

Prometheus Unbound
Thus I am answered : strange!
Paradise Lost
Whom thus the angelick Virtue answered mild.

Figure 2: Two sentences with a high similarity score with the approach described in 4.1.

4.2 A different approach, pt. 2

We then performed a transformative improvement to this approach by looking at the mutual position of co-occurrences in the two texts. We ranked in higher positions couples that share two or more rare words which appear close in both sentences.

Prometheus Unbound
How canst thou hear? Who knowest not the language of the dead?
Paradise Lost
What callest thou solitude? Is not the Earth With various living creatures, and the air Replenished, and all these at thy command To come and play before thee? Knowest thou not Their language and their ways? They also know, And reason not contemptibly: With these Find pastime, and bear rule; thy realm is large.

Figure 3: Two sentences with a high similarity score with the approach described in 4.2.

By looking at the two sentences with a high

similarity-values which we obtained recreating Goodman and Villani's approach, the major issue in their method is clearly evident. As they didn't normalise the word frequency by the length of the sentences, they ranked long sentences as top results. However it is clear that the similarity between these sentences is not very indicative of influence between authors.

If we instead consider two of the highest sentences detected by using our approaches we could notice that the common use of "rare words" and the close co-occurrence of specific couple of words could be useful to highlight poetic influence.

Starting from these encouraging results, in the near future our intention is to extend this study by considering other computational approaches, which, for example, have been developed in the area of text re-use (Clough et al., 2002), and by selecting a reliable metric for evaluating the different outputs of these ranking systems.

5 Results and conclusion

To conclude, for what concerns more specifically the possible boundary between humanities computing and the study of influence in literary studies, we believe that computational techniques helped to develop a keen sense of the issues involved, foregrounding the role of mechanical reading in de-idealising the problem. The use of machines in turn triggered Bloom's creation of his own "machine for criticism" (Bloom, 1976) built in order to understand how the anxiety of influence was dealt with by writers. If it's true that Bloom's theory "anxiously responds to [...] the subliminally perceived threat of textual anonymity promoted by the 'mechanical reproduction' of available texts" (Renza, 1990), that is to say something that with the digital will soon reach an entirely new, we can consider Raben's work as moved by a similar tension. Compared to Bloom, however, the advantage of the latter was that he was working from within, so to speak, the machine language. For this reason, he was able to begin transforming the traditional approach, that emphasised continuity within literary traditions, with a different one more focused on the way in which creative writing works, and on what the role of linguistic memory and of "creative" reading (or, as Prose (2006) said, reading as a writer).

References

- Harold Bloom. 1973. The anxiety of influence. *Journal of PsychoAnalysis*, 76(1): 19-24.
- Harold Bloom. 1976. *Poetry and Repression: Revisionism from Blake to Stevens*. New Haven, USA.
- Harold Bloom. 1997. *The anxiety of influence: A theory of poetry*. Oxford University Press, Oxford, UK.
- Silvana Castano, Alfio Ferrara and Stefano Montanelli. 2009. *Informazione, conoscenza e Web per le scienze umanistiche*. Pearson, Milano, IT.
- Paul Clough, Robert Gaizauskas, Scott S.L. Piao and Yorick Wilks. 2002. Meter: Measuring text reuse. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.
- Seymour Goodman and Raymond D. Villani. 1965. An algorithm for locating multiple word co-occurrences in two sets of texts. *Literary Data Processing Conference Proceedings*, 275-292.
- Julianne Nyhan, Andrew Flinn, and Anne Welsh. 2012. A short Introduction to the Hidden Histories project and interviews. *DHQ: Digital Humanities Quarterly*, 6(3): 275-292.
- Francine Prose. 2006. *Reading Like a Writer*. Harper-Collins, New York, USA.
- Joseph Raben. 1965. A computer-aided study of literary influence: Milton to Shelley. *Literary Data Processing Conference Proceedings*, 230-274.
- Louis A. Renza. 1990. *Influence*, in Lenticchia and McLaughlin, eds., *Critical Terms for Literary Study*. University Press of Chicago, Chicago, USA.
- Stefan Sinclair and Geoffrey Rockwell. 2014. Towards an archaeology of text analysis tools. Unpublished paper (DH2014, Lausanne, 7-12 July: <http://dharchive.org/paper/DH2014/Paper-778.xml>)
- Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, (28.1): 230-274.