Automatic Detection of Reuses and Citations in Literary Texts

J.-G. Ganascia[1,3], P. Glaudes[2,3] and A. DelLungo[4]

[1]ACASA team, LIP6 - University Pierre and Marie Curie, Paris, France
[2] Littérature française, XIX^{e}-XXI^{e} siècles, Université Paris-Sorbonne, Paris, France
[3] Labex OBVIL, PRES Sorbonne Universités, Paris, France
[4] ALITHILA, Université Charles de Gaulle, Lille 3, Lille, France

1 Introduction

For more than forty years now, modern theories of literature insist on the role of paraphrases, rewritings, citations, reciprocal borrowings and mutual contributions of any kinds. The notion of *intertextuality* was introduced in the sixties to approach these phenomena. *Phoebus* is collaborative project that makes computer scientists from the University Pierre and Marie Curie (LIP6-UPMC) collaborate with the literary teams of Paris-Sorbonne University with the aim to develop efficient tools for literary studies that take advantage of modern computer science techniques.

In this context, we have developed a piece of software that automatically detects and explores networks of textual reuses in classical literature. Written in PROLOG this program has been extensively tested on Isidore Ducasse texts (Lautréamont, 2009) that are known to contain many reuses and on “La comédie humaine” (Balzac, 1976-1981) from Honoré de Balzac, which, according to (Duclos, 2012), reuses some texts of his friend Théophile Gautier (Gautier, 2002). We claim that our approach is more efficient than comparable ones, e.g. (Roe, 2012) (Büchler, Crane, Mueller, Burns, & Heyer, 2011). This abstract describes the principles on which is based this program, the significant results that have already been obtained and the perspectives for the near future.

2 Distinctions between Plagiarism, Pastiches, Citations and Textual Reuses

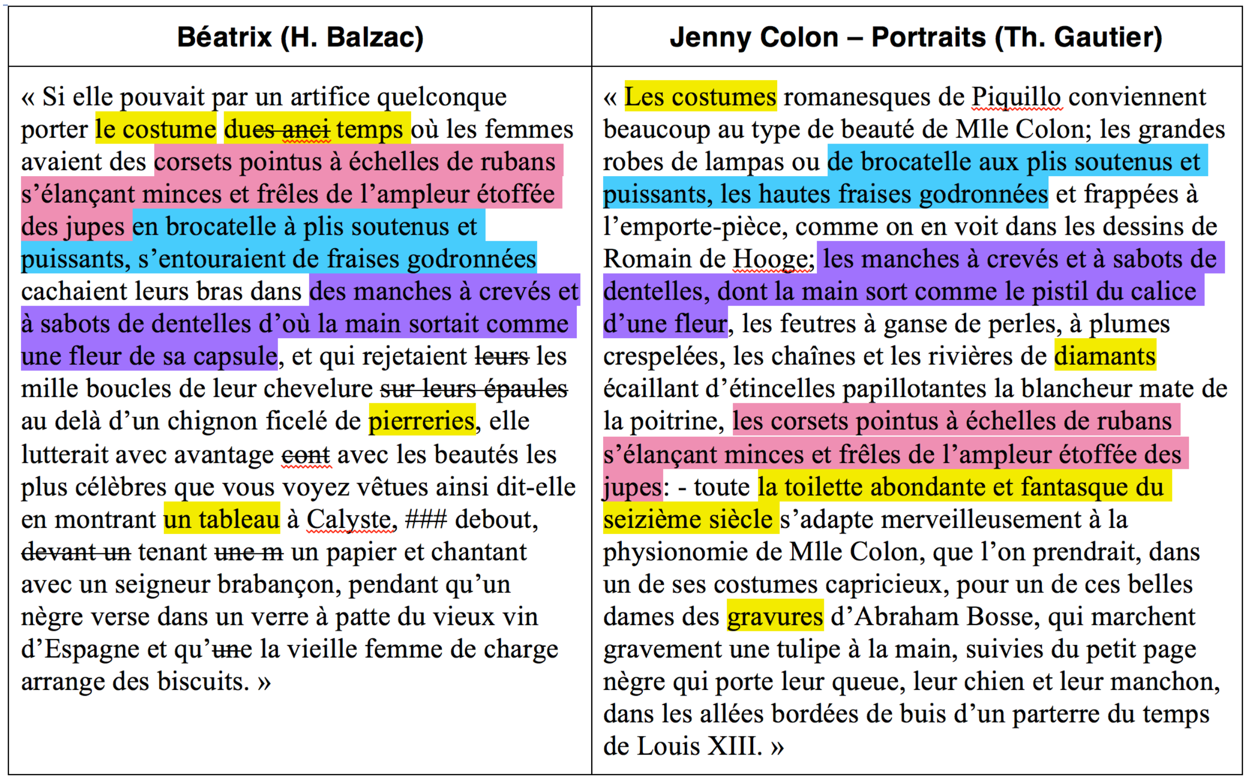
Before going into the detail of the description of the techniques that are used, let us note that the notions of literary textual reuse and citation, which we aim to automatically detect, have to be distinguished from two similar notions: the *plagiarism* and the *pastiche*.

The plagiarism consists in robbing the work of another, i.e. in fraudulently appropriating his/her texts, without mentioning explicitly their origin. As such, the plagiarism is considered as an unethical practice that has to be tracked and prosecuted. Many techniques have been developed to detect plagiarism that is considered as some plague, because intellectual work is stolen. By contrast, the pastiche is an artistic practice that imitates an artist, a style or a period. There is nothing wrong with it, except that it mocks well-known authors. Many great writers, for instance Marcel Proust, began by pastiches for the fun and to improve their style. Their detection is close to the identification of literary style (Dinu, Niculae, & Sulea, 2012), which requires capturing the essence of an artist's style or of a period. Halfway from detection of plagiarisms and identification of pastiches, the recognition of textual reuses and citations helps to track the literary influences and the spirit of the epoch. Some of the textual reuses and citations are conscious, other not. They may correspond to explicit – or implicit – and more or less distorted quotations. Usually, textual reuses proceed by transforming a piece of text, while citations are verbatim, but it's not always the case. When a sufficient part of the original text is kept, the fragments can be recognized. This is exactly what we attempt to do automatically here. Reuses and approximate citations are far more difficult to detect than plagiarisms, because the original fragments of text may be distorted, but far less than pastiches. Anyway, their detection could be of great interest for scholars interested in intertextuality.

3 Criteria

As previously said, text reuse and citation discovery is inspired from plagiarism detection, but it has to take into account all the alterations that may have transformed the initial text. To precise the type of distortions that affect a text, we started from a hand made study realized by Tania Duclos who shows in (Duclos, 2013) [cf. *figure 1*] how some parts of the Human Comedy (Balzac, 1976-1981) reuse fragments of texts from Théophile Gautier.

For instance, some of the passages highlighted in *figure 1* are identical, while others are somehow different. For instance *“en brocatelle à plis soutenus et puissants, s'entouraient de fraises godronnées”* becomes *“de brocatelle aux plis soutenus et puissants, les hautes fraises godronnées”* and *“des manches à crevés et à sabots de dentelles d'où la main sortait comme une fleur de sa capsule”* becomes *“les manches à crevés et à sabots de dentelles, dont la main sort comme le pistil du calice d'une fleur”*. Lastly, some fragments look far more difficult to identify, because they are composed of isolated words or even different words (e.g. *“diamants”* and *“pierreries”* or *“tableau”* and *“gravures”*) of which meanings are closed. Here, we try to detect string homologies where some words may be missing, especially stop words, i.e. articles, pronouns or prepositions. It may also happen that the number and the genre of nouns, adjectives or verbs change as when *“d'une main mignonne frappée de fossettes”* is transformed in *“des mains mignonnes frappées de fossettes”*.



Béatrix (H. Balzac)	Jenny Colon – Portraits (Th. Gautier)
« Si elle pouvait par un artifice quelconque porter le costume dues ~~anci~~ temps où les femmes avaient des corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes en brocatelle à plis soutenus et puissants, s'entouraient de fraises godronnées cachaient leurs bras dans des manches à crevés et à sabots de dentelles d'où la main sortait comme une fleur de sa capsule, et qui rejetaient ~~leurs~~ les mille boucles de leur chevelure ~~sur leurs épaules~~ au delà d'un chignon ficelé de pierreries, elle lutterait avec avantage ~~cont~~ avec les beautés les plus célèbres que vous voyez vêtues ainsi dit-elle en montrant un tableau à Calyste, ### debout, ~~devant un~~ tenant ~~une m~~ un papier et chantant avec un seigneur brabançon, pendant qu'un nègre verse dans un verre à patte du vieux vin d'Espagne et qu'~~une~~ la vieille femme de charge arrange des biscuits. »	« Les costumes romanesques de Piquillo conviennent beaucoup au type de beauté de Mlle Colon; les grandes robes de lampas ou de brocatelle aux plis soutenus et puissants, les hautes fraises godronnées et frappées à l'emporte-pièce, comme on en voit dans les dessins de Romain de Hooge; les manches à crevés et à sabots de dentelles, dont la main sort comme le pistil du calice d'une fleur, les feutres à ganse de perles, à plumes crespelées, les chaînes et les rivières de diamants écaillant d'étincelles papillotantes la blancheur mate de la poitrine, les corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes: - toute la toilette abondante et fantasque du seizième siècle s'adapte merveilleusement à la physionomie de Mlle Colon, que l'on prendrait, dans un de ses costumes capricieux, pour un de ces belles dames des gravures d'Abraham Bosse, qui marchent gravement une tulipe à la main, suivies du petit page nègre qui porte leur queue, leur chien et leur manchon, dans les allées bordées de buis d'un parterre du temps de Louis XIII. »

Figure 1: *example of hand coded comparison* (Duclos, 2013) *between a fragment of Béatrix* (Balzac, 1976-1981) *on the left and a fragment of Théophile Gautier* (Gautier, Portraits contemporains, 1874) *on the right.*

4 Detection of Fragments with Holes

Among the more efficient existing plagiarism detection techniques, many are based on fingerprints built with the hash coding of character strings (Potthast, Eiselt, Barron-Cedeño, Stein, & Rosso, 2011), (Potthast, Stein, Barron-Cedeño, & Rosso, 2010), (Burrows, Tahaghoghi, & Zobel, 2006). Other techniques evaluate the statistical distribution of vocabulary with a vector space model of the texts and a cosine similarity measure that evaluate their closeness, however they don't seem to be appropriate to our purpose, even if we use the distribution of words.

We have implemented and adapted the fingerprint method and we have evaluated it on the reuses isolated by Tania Duclos. It helped us to optimize the values of the different parameters. To do this, we have first eliminated the "stop words", i.e. article, preposition, pronoun, auxiliary verbs, etc. We have also used the Snowball (Porter, 2001) (Tomlinson, 2004) stemmer to reduce the words to their root, which allows being independent from the inflected forms used in the text. For instance, the words "fishing", "fished", "fish", "fishes" and "fisher" are reduced to the same root word "fish".

The second part consists in extracting sequences of words characterized by their minimal size, i.e. by the minimal number of consecutive non-"stop words" they contain, which we call the window size. In addition, because we want to allow missing words, we also introduce possible holes. This means that a window of size 4 does not necessarily correspond to 4 consecutive words.

Once the similar fragments are discovered, they are adjoined end to end, which build blocs. Lastly, we have manually defined what we call "weak words", which are not very significant, and we filter the blocs of similar words of which number of non-"weak words" is bigger than a minimal threshold, for instance 4. This allows eliminating noise, without loosing a lot of information.

5 Obtained Results

The program has been implemented in SWI-Prolog (cf. (SWI-Prolog's home)) using an external table to store hash-coded texts. It is quite efficient, for instance it took less than 10 minutes to index all the Balzac's "Human Comedy" (Balzac, 1976-1981) that contains more than 25 millions of characters on a 2GHz MacPro. Then, it takes a couple of minutes to discover text reuses on entire novels.

Using this program, we were able to retrieve all the handed coded reuses of (Duclos, 2013), except the "yellow" one (see *figure 1*). We have also detected many interesting citations and reuses, for instance a reuse of the Gautier's Novel entitled "Mademoiselle de Maupin" (Gautier, Romans, contes et nouvelles, 2002) in the Balzac's Novel "Modeste Mignon" (Balzac, 1976-1981), which has not been mentioned before, or a citation of Lyttleton both in "Delphine" (de Staël, 1869) and in "Ursule Mirouët" (Balzac, 1976-1981). We also tested the system between Lautréamont's work (Lautréamont, 2009) and Buffon one the one hand and the French moralists like Pascal, La Rochefoucauld or La Bruyère on the other. We have retrieved many text reuses among which some interesting distortions like, for instance the Pascal aphorism *"Nous naissons injustes; car chacun tend à soi: cela est contre tout ordre."* that has been rewritten in *"Nous naissons justes. Chacun tend à soi. C'est envers l'ordre."*

6 Perspectives

For the near future, we plan to extensively use our system in many fields of literature, especially on the 19th century French literature, with Balzac's work, which is the aim of the PHOEBUS project funded by the CNRS. More precisely, PHOEBUS is intended to investigate the textual reuses between the Balzac's youth novels and the "Human Comedy", and between Balzac's work and his contemporaries' work like Théophile Gautier, Benjamin Constant, George Sand etc. We also plan to digitalize the journals where many authors published either under their own names, or anonymously and to compare them with the "Human Comedy". Lastly, we will conduct a thorough comparison with similar approaches.

References

Büchler, M., Crane, G., Mueller, M., Burns, P., & Heyer, G. (2011). One Step Closer To Paraphrase Detection On Historical Texts: About The Quality of Text Re-use Techniques and the Ability to Learn Paradigmatic Relations. (G. K. Thiruvathukal, & S. E. Jones, Éds.) *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* .

Balzac, H. (1976-1981). *La comédie humaine* (Vol. I-XII). (C. L. Pléiade, Éd.) Paris: Gallimard.

Burrows, S., Tahaghoghi, S., & Zobel, J. (2006). Efficient Plagiarism Detection for Large Code Repositories. *Software - Practice and Experience , 37*, 151-175.

de Staël, G. (1869). *Delphine.* Paris: Garnier frères.

Dinu, L. P., Niculae, V., & Sulea, O.-M. (2012). Pastiche detection based on stopword rankings. Exposing impersonators of a Romanian writer. *EACL 2012 - Workshop on Computational Approaches to Deception Detection* (pp. 72-77). Avignon: Association for Computational Linguistics.

Duclos, T. (2013). *L'intertextualité dans une Fille d'Eve et Béatrix d'Honoré de Balzac.* Paris-Sorbonne University.

Gautier, T. (1874). *Portraits contemporains.* Paris, France: Charpentier et Cie.

Gautier, T. (2002). *Romans, contes et nouvelles* (Vol. I-II). (C. d. Pléiade, Éd.) Paris: Gallimard.

Lautréamont. (2009). *Œuvres complètes.* (C. d. Pléiade, Éd.) Paris: Gallimard.

Porter, M. (2001, October). *Snowball: A language for stemming algorithms*. Consulté le October 22, 2012, sur Snowball: http://snowball.tartarus.org/texts/introduction.html

Potthast, M., Eiselt, A., Barron-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. Dans V. Petras, P. Forner, & P. D. Clough (Éd.), *Notebook Papers of CLEF 11 Labs and Workshops.*

Potthast, M., Stein, B., Barron-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. Dans C.-R. Huang, & D. Jurafsky (Éd.), *23rd International Conference on Computational Linguistics (COLING 10)*, (pp. 997-1005). Stroudsburg, Pennsylvania.

Roe, G. R. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research . *Digital Humanities.* Hamburg.

SWI-Prolog's home. (s.d.). Consulté le Octobre 22, 2012, sur SWI-Prolog: http://www.swi-prolog.org/

Tomlinson, S. (2004). Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer TM at CLEF 2003. Dans C. Peters (Éd.), *Working Notes for the CLEF 2003 Workshop.* Springer.