Linguistic Pattern Extraction and Analysis for Classic French Plays

Francesca Frontini, Mohamed Amine Boukhaled, Jean-gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS / OBVIL
{francesca.frontini, mohamed.boukhaled, jean-gabriel.ganascia}@lip6.fr

1. Introduction and approach

Great authors of fiction and theatre have the capacity of creating memorable characters that take life and become almost as real as living persons to the readers/audience. The study of characterization, namely of how this is achieved, is a well-researched topic in corpus stylistics: for instance (Mahlberg, 2012) attempts to identify typical lexical patterns for memorable Dickens' characters by extracting those lexical bundles that stand out (namely are overrepresented) in comparison to a general corpus. In other works, authorship attribution methods are applied to the different characters of a play to identify whether the author has been able to provide each of them with a "distinct" voice. For instance (Vogel & Lynch, 2008) compare individual Shakespeare characters against the whole play or even against all plays of the same author.

The purpose of this paper is to propose a methodology for the study characterization of several characters in French plays of the classical period. The tools developed are meant to support textual analysis by:

1) Verifying the degree of characterization of each character with respect to others.

2) Automatically inducing a list of linguistic features that are significant, representative for that character.

Preliminary investigations have been conducted on plays by Moliere, cross-comparing four protagonists from four different plays. The proposed methodology relies on sequential data mining for the extraction of linguistic patterns and on correspondence analysis for comparison of patterns frequencies in each character and for the visual representation of such differences.

2. Syntactic pattern extraction and ranking

In our study, we consider a *syntagmatic approach* based on a quite similar configuration to the one proposed by (Quiniou, Cellier, Charnois, & Legallois, 2012) . The text is first segmented into a set of sentences, and then each sentence is mapped into a sequence of syntactic (POS-tag) items. For example the sentences "J'aime ma maison où j'ai grandi." is first mapped to a sequence of PoSTags, <PRO:PER VER:pres DET:POS NOM PRO:REL PRO:PER VER:pres VER:pper SENT>; then sequential patterns of a determined length are extracted. A minimal filtering is applied, removing patterns with less than 5% of support; nevertheless sequential pattern mining is known to produce (depending on the window and gap size) a large quantity of patterns even relatively small samples of texts.

In order to identify the most relevant patterns for each of the four characters we thus used correspondence analysis (CA), which is a multivariate statistical technique developed by (Benzécri, 1977) and used for data analysis. CA allows us to represent both the characters and the patterns on a bi-dimensional space, thus making it visually clear not only which characters are more similar to each other but also which patterns are over/underrepresented - that is more distinctive - for each character or group of characters. Moreover patterns can be

ranked according to the combined contribution on both axes, and those with the highest contribution can be retained, thus enabling the researcher to filter out less interesting patterns.

3. Results

We present here some preliminary results and plots derived using the R module FactoMiner (Husson, Josse, Le, & Mazet, 2013).
Patterns were extracted from the texts of four memorable Moliere protagonists (Harpagon - 'Avare'; Dom Juan - 'Dom Juan'; Scapin - "Les fourberies de Scapin"; Sganarelle - "Le medecin malgré lui") which have been extracted separating them from the rest of their respective plays. The extracted patterns are 3-4-5grams of PoS only, with at most one gap.

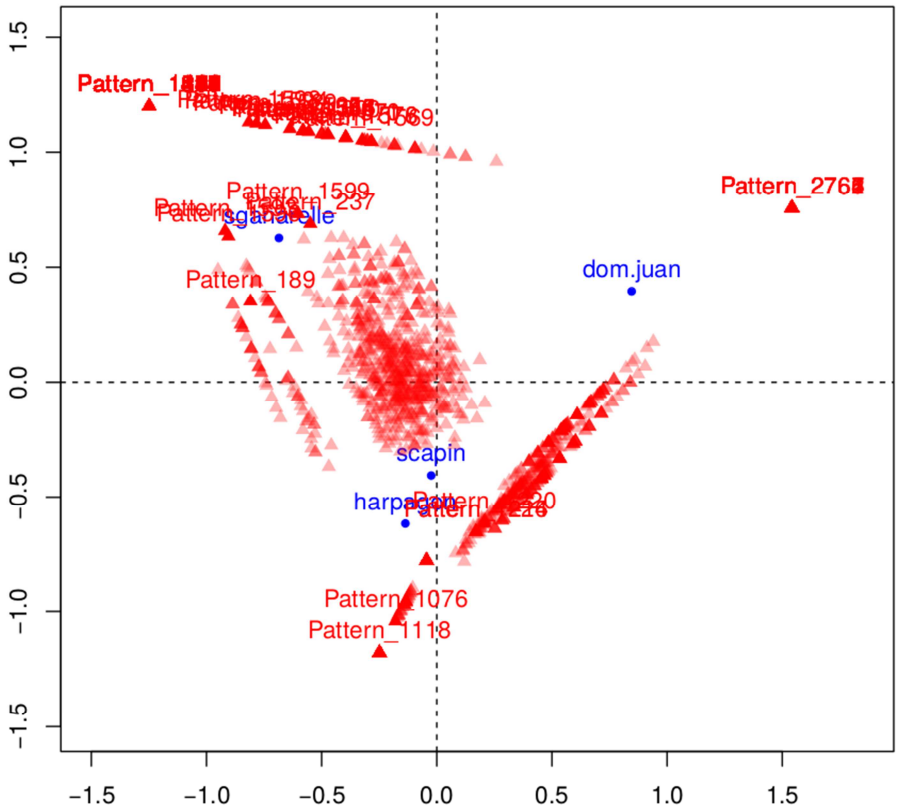

Figure 1 : Correspondance analysis of four memorable Moliére protagonists

The plot (see Figure 1) shows the relative distances of the four characters according to CA; all patterns are in shadow, except for the first 200 by contribution. The most isolated character seems to be *Sganarelle*, the protagonist of a piece in which a simple man is forced by circumstances to pretend to be a great doctor. His language is namely different from the syntactic point of view from the others. Among his most significant patterns we find syntactic structures that are typically used to express diagnosis (see example 1), and other ones that are used to apologize, as he eventually finds himself discovered.
Example 1 (instances of the pattern [PRO:PER] [VER:pres] [KON] [*] [NOM]) :[1]

- Qui est causée par l' âcreté des humeurs engendrées dans la concavité du diaphragme , **il arrive que ces vapeurs** ... Ossabandus , nequeys , nequer , potarinum , quipsa milus

[1] '*' stand for a gap, it can be instantiated by any POS tag.

- Pour revenir donc à notre raisonnement , **je tiens que cet empêchement** de l' action de sa langue est causé par de certaines humeurs ...
- **il se trouve que le poumon** , que nous appelons en latin armyan , ayant communication avec le cerveau
- d' autant que l' incongruité des humeurs opaques qui se rencontrent au tempérament naturel des femmes étant cause que la partie brutale veut toujours prendre empire sur la sensitive , **on voit que l' inégalité** de leurs opinions dépend du mouvement oblique du cercle de la lune

Dom Juan, a nobleman and a complex character, is instead isolated by underrepresentation, in that he has less distinctive patterns, which may mean that his language is less repetitive and, possibly more elaborated. Finally *Scapin* and *Harpagon*, two comical character, share patterns of low syntactic complexity, used to interact with other characters or to complain (especially in the case of Harpagon) about what happens to them.
In the final papers examples of patterns will be discussed and different types of pattern extraction (in terms of length, gaps, presence of lexical items or not) will be compared to investigate characterization from different perspectives (e.g. syntactic vs lexical complexity ...).

4. Preliminary conclusions

The proposed methodology offers a useful instrument to facilitate literary analysis and criticism; not only it calculates and represents the distances between characters (which may be possible using other clustering techniques) but also provides a way to motivate and explain this difference based on the extraction of significant and distinctive sets of patterns for each character, which is a strong requirement for all computational stylistics methods. Generally speaking CA offers a statistically well founded way of measuring differences in the occurrence of linguistic patterns in different texts, and may be successfully applied to the study of stylistics in general, for instance by comparing whole plays/ works of different authors.

5. Bibliography

Benzécri, J.-P. (1977). Histoire et préhistoire de l'analyse des données. Partie V: l'analyse des correspondances. *Cahiers de L'analyse Des Données*, *2*(1), 9–40.

Husson, F., Josse, J., Le, S., & Mazet, J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R, R package version 1.24.

Mahlberg, M. (2012). *Corpus stylistics and Dickens's fiction* (Vol. 14). Routledge.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). What about sequential data mining techniques to identify linguistic patterns for stylistics? In *Computational Linguistics and Intelligent Text Processing* (pp. 166–177). Springer.

Vogel, C., & Lynch, G. (2008). Computational Stylometry: Who's in a Play? In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction* (pp. 169–186). Springer.