Moliere's Raisonneurs: a quantitative study of distinctive linguistic patterns

Francesca Frontini, Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

Moliere's Raisonneurs: a quantitative study of distinctive linguistic patterns

Francesca Frontini*,
Mohamed Amine Boukhaled,**
Jean Gabriel Ganascia**
*ILC-CNR, Pisa
**Labex OBVIL - LIP6 UPMC, Paris
francesca.frontini@ilc.cnr.itmohamed.boukhaled@lip6.fr
jean-gabriel.ganascia@lip6.fr

1 Introduction and approach

Great authors of plays and novels are often renowned for the ability to create memorable characters that take on a life of their own and become almost as real as living persons to their readers/audience. The study of characterization, that is, of how it is that authors manage to achieve this, has become a well-researched topic in corpus stylistics: for instance (Mahlberg, 2012) attempts to identify typical lexical patterns for memorable characters in the work of Dickens by extracting those lexical bundles that stand out (namely are over-represented) 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 give each of them with a "distinct" voice. For instance (Vogel and Lynch, 2008) compare the dialogues of individual characters in a Shakespeare play against the rest of the play or even against all plays in the Shakespearean corpus.

In Frontini et al (2015), we propose a methodology for the extraction of significant patterns that enables literary critics to verify the degree of characterization of each character with respect to the others and to automatically induce a list of linguistic features that are significant and representative for that character. The proposed methodology relies on sequential data mining for the extraction of linguistic patterns and on correspondence analysis for the comparison of pattern frequencies in each character and the visual representation of such differences.

We chose to apply this analysis to Moliere's plays and the protagonists of those plays. In this work we focus on the figure of the *raisonneurs*, characters who take part in discussions with comical protagonists providing a counterpart to their follies. Such characters were interpreted at times as spokesmen for Moliere himself, and the voice of reason, at other times as comical characters themselves and no less foolish than their opponents. Hawcroft's essay *Reasoning with fools* (2007) highlights the differences between five of these characters based on their role in the plot. Using this analysis as guidance, we compare significant linguistic patterns in order to see how these differences are marked by the author. We do this by adapting the discourse traits of each of them to the communicative function they needs to fulfill (Biber and Conrad 2009).

2 Syntactic pattern extraction and ranking

In our study, we consider a syntagmatic approach based on a configuration similar to that proposed by (Quiniou et al. 2012). The text is first segmented into a set of sentences, and then each sentence is mapped into a sequence of part of speech (PoS) tags. Tagging is automatically performed using TreeTagger (Schmid, H. 1995)[1]. For example the sentence

J'aime ma maison où j'ai grandi.

is first mapped to a sequence of PoSTagged words;

```
<J'        PRO:PER><aime         VER:pres><ma
DET:POS><Maison      NOM><où      PRO:REL><j'
PRO:PER><ai   VER:pres><   grandi   VER:pper
SENT>
```

Then sequential patterns of 3 to 5 elements are extracted. Patterns can be made of PoS Tags only, or of a mix of PoS Tags and recurring lexical elements, with possible gaps (see examples (1), (2), (3)). 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 on relatively small samples of texts.

In order to identify the most relevant patterns for each of the four characters we used correspondence analysis (CA), which is a multivariate statistical technique developed by (Benzécri, 1977) and used for the analysis of all sorts of data, including textual data (Lebart et al. 1998). CA allows us to represent both Moliere's characters and the (syntactic) 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/under-represented - that is, more distinctive - for each character or group of characters.

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

[1]For a description of the French tagset see here: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

patterns.

3 Analysis and results

CA was performed with the R module *FactoMiner* (Husson et al. 2013) on five characters from five different plays (see Table 1):

Play	Raisonneur	Counterpart
Ecole des femmes	Chrysalde	Arnolphe
Ecole des maris	Ariste	Sganarelle
Tartuffe	Cléante	Orgon
Mysantrope	Phylinte	Alceste
Malade imaginaire	Béralde	Argan

Table1: Characters and plays.

Figure 1 shows the result of the correspondence analysis, with the five *raisonneurs* printed as blue dots, the patterns printed as red triangles, and the 10 patterns with the highest contribution labeled with their identifiers. Filtering by contribution is crucial in our technique, which extracts over 9500 patterns, most of which are common to all characters (see central cloud in the plot) and thus not so interesting for our study.

The relative distances between the characters seem to match what is already known from literary criticism; first of all Béralde, who is the only character to express himself in prose, is isolated on the right of the x axis. In fact, it is not advisable to compare characters without distinguishing for prose and verse, but we have retained the example of Béralde to show how the proposed technique can easily identify differences in genre. As for the other characters, Hawcroft stresses the difference in the roles of Ariste, Philinte and Chysalde on the one hand and of Cléante on the other. The latter is a more pro-active character, more crucial to the plot; he is also less accommodating than the other three, who are depicted mostly as loyal friends and brothers, trying to help the hero to avoid the consequences of his foolish actions and beliefs. Instead, Cléante has also to worry about his sister's wellbeing: having to face not only the besotted brother in law, Orgon, but also the man who has duped him, Tartuffe.

In order to confirm this intuition, it is necessary to turn our attention to what it is that exactly causes the spatial distribution, namely the high contribution patterns, we find above. Our technique allows us not only to find the corresponding pattern for each identifier on the plot, but also to extract all underlying instances in the texts. Due to space constraints, only a brief demonstrative analysis will be performed.

Phylinte and Chrysalde are strongly associated with patterns containing prepositional phrases separated by commas. Such patterns are used in contexts where the characters give advice in a very cautious, indirect way. The overuse of punctuation itself, in these two characters, seems to be an indication that the character should be played as a

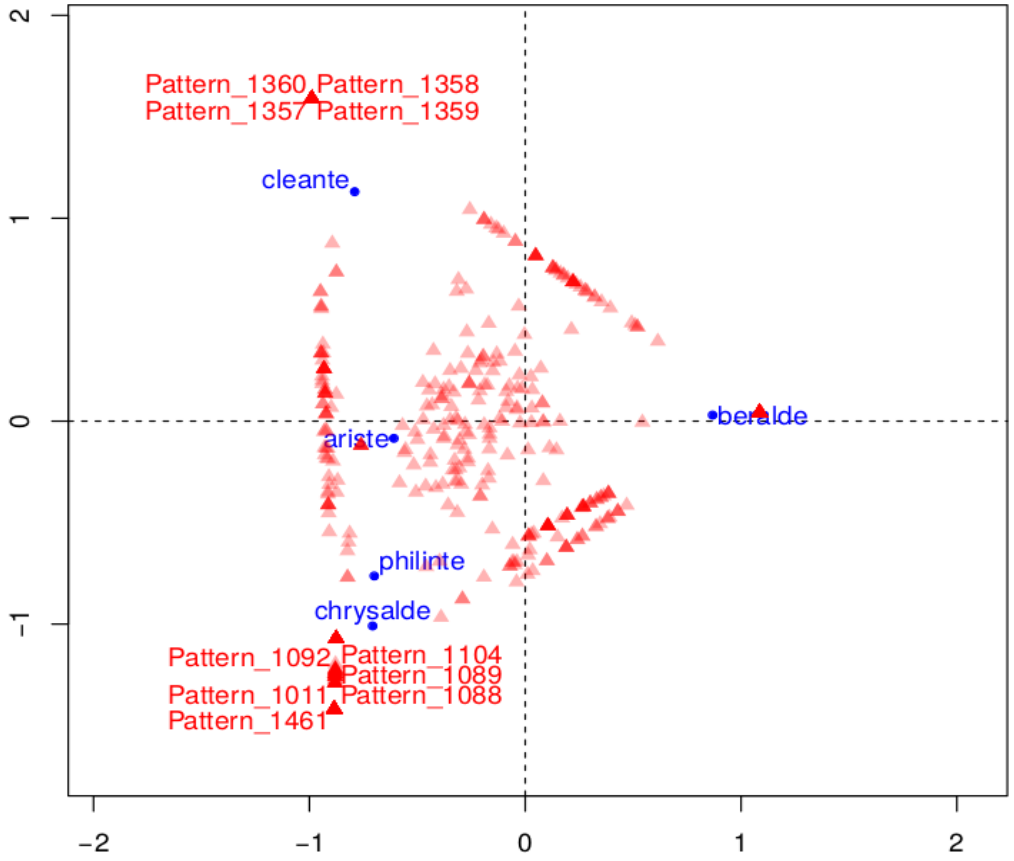

Figure 1Correspondence analysis plot, with first 10 patterns for contribution.

soft spoken person, who is fond of his friend and careful not to offend, e.g.:

(1) Pattern 1011
`[,] [any word] [PRP] [any word] [NOM]`

Instances from **Chrysalde**:

- `Entre ces deux partis il en est un honnête` **`, Où dans l' occasion`** `l' homme prudent s' arrête [...]`
- `Il faut jouer d' adresse` **`, et d' une âme`**`réduite , Corriger le hasard par la bonne conduite [...]`

Instances from **Phylinte**:

- **`, Et pour l' amour`** `de vous , je voudrais , de bon cœur , Avoir trouvé tantôt votre sonnet meilleur .`

On the other hand, the patterns most associated with Cléante contain modal constructions, and are indicative of a more direct way of advising, and of stronger arguments, e.g.

(2) Pattern 1360
`[ PRO:PER ] [any word] [VER:infi] [PRP]`

- `Les bons et vrais dévots , qu'` **`on doit suivre à`** `la trace , Ne sont pas ceux aussi qui font tant de grimace .`
- `Et s' il` **`vous faut tomber dans`** `une extrémité , Péchez plutôt encore de cet autre côté .`

Finally, the patterns extracted for Béralde are indicative of the greater simplicity and repetitiveness of his prose, and of the stereotypical role he has in the play, which is that of a man concerned with his brother, as in:

(3) Pattern 865 `[,][DET:POS][any word][PUN]`

- `Oui , mon frère , puisqu' il faut parler à cœur ouvert ,`

From this experiment it is therefore possible to conclude that the method described above is a promising one, as it not only verifies known facts about the characters in question, but also ground them on corpus based evidence.

4 Preliminary conclusions

Clustering techniques are commonly used in computer aided literary criticism. In order to prove that clusters are significant, statistical analysis can be later applied to verify that resulting clusters are significant. The strength of CA lies in the fact that it allows users to easily identify the reasons for certain texts to group together or to diverge. This helps to overcome the lack of transparency in the presentation of results which often disappoints experts when experimenting with similar techniques, thus making it a useful hermeneutical tool, in the sense of Ramsey (2011)'s algorithmic criticism.

5 Acknowledgements

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02, as well as by a scholarship from the *Fondation Maison Sciences de l'Homme*, Paris.

References

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.

Biber, D., & Conrad, S. 2009. Register, genre, and style. Cambridge University Press.

Frontini, F., Boukhaled, M. A., & Ganascia, J. G. 2015 Linguistic Pattern Extraction and Analysis for Classic French Plays. Presentation at the CONSCILA Workshop, Paris.

Hawcroft, M. 2007. Molière: reasoning with fools. Oxford University Press.

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

Lebart, L., Salem, A., & Berry, L. 1998. Exploring textual data (Vol. 4). Springer.

Leech, G. N., & Short, M. 2007. *Style in fiction: A linguistic introduction to English fictional*. Pearson Education.

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. Springer, (166–177).

Ramsay, S. 2011. Reading machines: Toward an algorithmic criticism. University of Illinois Press.

Schmid, H. 1995. Treetagger| a language independent part-of-speech tagger.Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.

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