**FRONTINI — Linguistic**

<1 image>

**Linguistic Pattern Extraction and Analysis for Classic French Plays**

Frontini, F., Boukhaled, M. A. and Ganascia, J. G.

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 and Lynch (2008) compare individual Shakespeare characters against the whole play or even against all plays of the same author.

This paper proposes a methodology for the study characterisation in French plays.1 The tools developed are meant to support textual analysis by

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

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

We report here some investigations conducted on plays by Molière, 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 pattern frequencies in each character and for the visual representation of such differences.

**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 et al. (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.2 For example the sentence ‘J'aime ma maison où j’ai grandi.’ is first mapped to a sequence of PoS tags, <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 on 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 (Lebart et al., 1998). CA allows us to represent the characters as well as 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- or 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.

**Results**

We present here some preliminary results and plots derived using the R module FactoMiner (Husson et al., 2013). Patterns were extracted from the texts of four memorable Molière protagonists (Harpagon, *Avare*; Dom Juan, *Dom Juan*; Scapin, *Les fourberies de Scapin*; Sganarelle, *Le medecin malgré lui*), separating them from the rest of their respective plays.

Different types of extraction can be used to focus on different aspects. In the present experiment, extracted patterns are 3-4-5grams of simplified PoS tags (no morphology information), with at most one gap. This setup focuses mostly on the syntactic differences between characters.

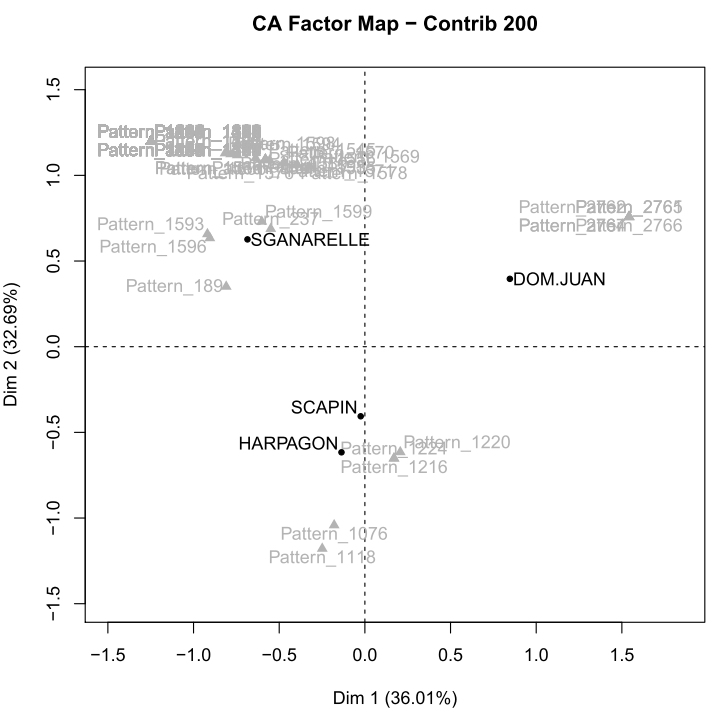


Figure 1 : Correspendance analysis of four memorable Moliére protagonists

Figure 1. Correspondence analysis of four memorable Molière 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 (example 1).

Example 1: Instances of the pattern [PRO:PER] [VER:pres] [KON] [\*] [NOM]:3

**il arrive que ces vapeurs** . . . Ossabandus, nequeys, nequer, potarinum, quipsa milus [it happens that such vapours . . .]

**je tiens que cet empêchement** de l’ action de sa langue est causé par de certaines humeurs . . . [I hold that this impediment in her tongue is caused by certain humours . . .]

**il se trouve que le poumon**, que nous appelons en latin armyan, . . . [it happens that the lung, which we call in latin armyan . . .]

**on voit que l’ inégalité** de leurs opinions dépend du mouvement oblique du cercle de la lune . . . [we can see that the inequality of their opinions depends on the oblique movement of the moon orbit . . .]

In other cases (example 2), the pattern groups assertions have a performative function, which are used first to (vainly) try to clear the misunderstanding, then to assure people of his assertions, and finally, once discovered, to confess.

Example 2: [PRO:PER] [PRO:PER] [\*] [KON]

je te dis que . . . [I tell you that . . .]

Je vous promets que . . . [I promise you that . . .]

Je vous jure que . . . [I swear that . . .]

je vous dis que . . . [I tell you that . . .]

Je vous assure que . . . [I assure you that . . .]

Je vous assure que . . . [I assure you that . . .]

je vous apprends que . . . [I inform you that . . .]

Je vous apprendrai que . . . [I shall assure you that . . .]

je vous avoue que . . . [I confess you that . . .]

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. This is also evident from one of the few patterns that are strongly associated with him (example 3), which captures the overuse of subordinate clauses.

Example 3: [KON] [PRO:PER] [\*] [VER:pres] [\*] [PRP]

sachez **que je n’ ai point d’**autre dessein que de vous épouser , , , [ know that I have no other design than to marry you ]

elle va vous dire **que je lui ai promis de** l'épouser [ . . . she is going to tell you that I promised her to marry her ]

Vous soutenez également toutes deux **que je vous ai promis de** vous prendre pour femmes [You both claim that I have promised you to marry you]

. . . et **que je sais me servir de** mon épée quand il le faut [and that I know how to use my sword when needed]

. . .

Finally Scapin and Harpagon, two comical characters, are both characterised by patterns of lower syntactic complexity; especially in the case of Harpagon (examples 4 and 5), some patterns convey the image of a self-centered person, who wants to have things his way, and subject to violent disappointments (especially when money is concerned).

Example 4: [PRO:PER] [PRO:PER] [VER:pres] [VER:pper]

on m' a privé . . . [they have deprived me . . .]

on m' a dérobé . . . [they have robbed me . . .]

on m' a volée . . . [they have stolen me . . .]

on m' a pris . . . [they have taken me . . .]

. . .

Example 5: [KON][PRO:PER][\*][KON]

que je veux que . . . [that I want that . . .]

et il faut que . . . [and it is necessary that . . .]

et vous verrez qu’ . . . [and you will see that . . .]

. . .

These syntactic patterns have a slightly different function in Scapin, the clever servant who interacts with several characters in order to try to carry out his plan. In example 6 we see the same pattern as in example 4, but used mostly to report events.

Example 6 [PRO:PER][PRO:PER][VER:pres][VER:pper]

**Je l’ ai trouvé** tantôt tout triste [I have found him so sad]

**nous nous sommes allés** promener sur le port [we have gone to walk in the harbour]

. . .

It is worth noticing how such structures in the past tense are underrepresented in the character of Sganarelle, whose discourse is prevalently in the present tense; while Dom Juan, Sganarelle, and Scapin are all actively lying in their respective plots, the use of past tense in Scapin may be more reflective of conscious scheming.

Experiments are currently ongoing with different types of pattern extraction (in terms of length, gaps, presence of lexical items or not), in order to determine the optimal setups for investigating characterization from different perspectives (e.g., syntactic vs lexical complexity). Recent developments in discourse analysis studies (Biber and Conrad, 2009) may also provide important insights in the mapping between form and communicative function.

**Preliminary Conclusions**

The proposed methodology offers a useful instrument to facilitate literary analysis and criticism; not only does it calculate and represent the distances between characters (which may be possible using other clustering techniques), but this methodology 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. The methodology can be extended to any language and literary genre, provided a reliable automatic PosTagging for that language/genre is available. 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 or works of different authors.

**Notes**

1. 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, and by a Fernand Braudel scholarship from the Fondation ‘Maison Sciences de l’Homme’, Paris.

2. PosTagging was performed automatically using the TreeTagger tool (French parameters described here: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html).

3. ‘\*’ stands for a gap. It can be instantiated by any PoS tag.

**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. and Conrad, S.** (2009). *Register, Genre, and Style.* Cambridge University Press, Cambridge.

**Conesa, G.** (1983). *Le dialogue moliéresque.* Presses Universitaires de France.

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

**Lebart, L., Salem, A. and Berry, L.** (1998). *Exploring Textual Data*, Vol. 4. Springer Science & Business Media.

**Mahlberg, M.** (2012). *Corpus Stylistics and Dickens’s Fiction*. Vol. 14. Routledge.

**Quiniou, S., Cellier, P., Charnois, T. and Legallois, D. (2012).** What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In *Computational Linguistics and Intelligent Text Processing,* Springer, pp. 166–77.

**Vogel, C. and Lynch, G.** (2008). Computational Stylometry: Who’s in a Play? In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction,* Springer, pp. 169–86.