# Analyse et extraction des motifs linguistiques dans un corpus théâtral

Journée CONSCILA - Paris le 16 janvier 2015

Francesca Frontini**, Amine Boukhaled, Jean-Gabriel Ganascia
Laboratoire d'informatique de Paris 6 -Labex OBVIL
**Boursière Fernand Braudel de la *Fondation Maison Sciences de l'homme* - Paris




# Cadre général

**Analyse stylistique computationnelle de pièces theâtrales classiques (Molière).**

Notre contribution se situe dans le cadre d'une **approche syntagmatique** destinée à l'étude des motifs distinctifs du discours de chaque personnage.

# Approches précédentes

**Mahlberg (2012)** - étude des personnages de Dickens, basée sur les fréquences des patrons textuels (linguistique de corpus).

**Vogel & Lynch (2008)** - étude des personnages de Shakespeare, basée sur l'analyse statistique des séquences de lettres (identification d'auteurs).

# Notre approche

**Objectifs:**

1) Identifier la cohésion du discours des personnages
2) Identifier les traits stylistiques importants de ces personnages

**Méthodologie**:

Analyse statistique et visualisation afin de déterminer les séquences syntagmatiques les plus typiques et significatives pour chaque personnage.

# Types de motifs

La **fouille de données séquentielles** (Agrawal 1993, Fournier-Viger et al 2014) nous permet d'extraire les répétitions au niveau de la forme du mot, du lemme et de la classe syntagmatique.

- Motifs syntaxiques avec trous
  - ex: (PRO:PER) (VER:pres) (KON) (*) (NOM)

- Motifs lexico-syntaxiques avec trous
  - ex: (je PRO:PER)(veux VER:pres)(*)(PRO:PER)

# Corpus

Initiative de numérisation des textes littéraires au sein du Labex OBVIL **(TEI, bonne qualité, marquage des tours de parole).**

**Pièces de Molière:**

4 personnages (pièces en prose) :

- **Harpagon** - L'Avare (6145 mots)
- **Dom Juan** - Dom Juan (66133 mots)
- **Scapin** - Les Fourberies de Scapin (6079 mots)
- **Sganarelle** - Le Médecin malgré lui (3854 mots)

# Extraction des motifs

- motif 3-5 grams
- 1 trou consécutif au maximum
- motif syntaxique
- fréquence relative des motifs  >1% des phrases du texte
- fréquence absolue des motifs >5 occurences

Inconvénient: la fouille de motifs séquentiels produit une **grande quantité de motifs** même sur des échantillons de textes relativement petits (2768 dans notre cas).

# La méthode d’analyse des motifs

**Analyse des correspondances** (Benzécri, 1977): technique d’analyse statistique multivariée

- Projection des personnages et de leurs motifs dans un espace bi-dimensionnel
- Classement des motifs en fonction de leur *contribution* combinée sur les deux axes, en permettant ainsi au chercheur de filtrer les motifs les moins intéressants, c’est-à-dire ceux qui ont une faible contribution.

# La méthode d’analyse des motifs (2)

Pour l’analyse des correspondences, on utilise **FactoMiner** - Logiciel R pour l’analyse des attributs dans une base de données (Husson et al., 2013).

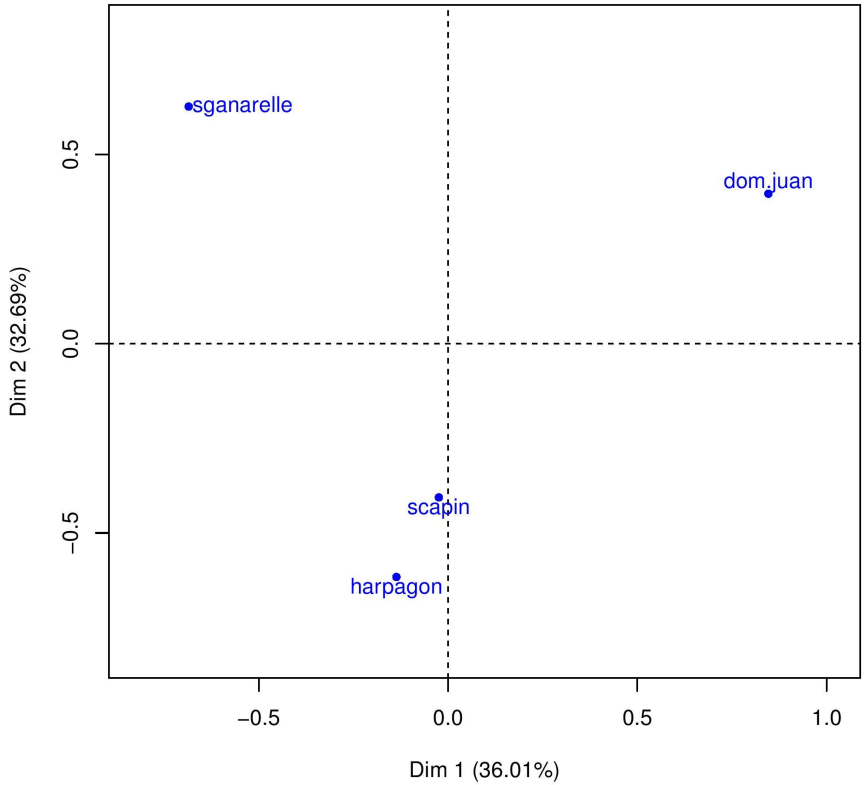
Chaque personnage est representé par un vecteur des frequences normalisées des motifs qui apparaissent dans son discours.

**ex: Sganarelle**

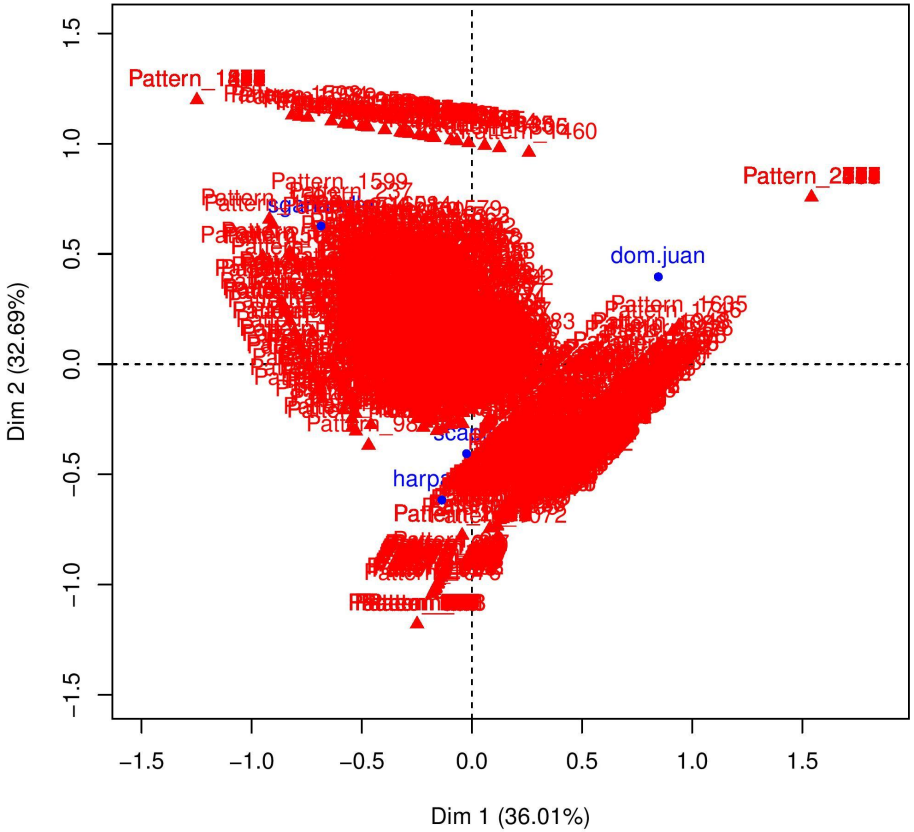
<motif-1, motif-2, ..... , motif-n>

<0,12 , 0,07 , ..... , 0,003>

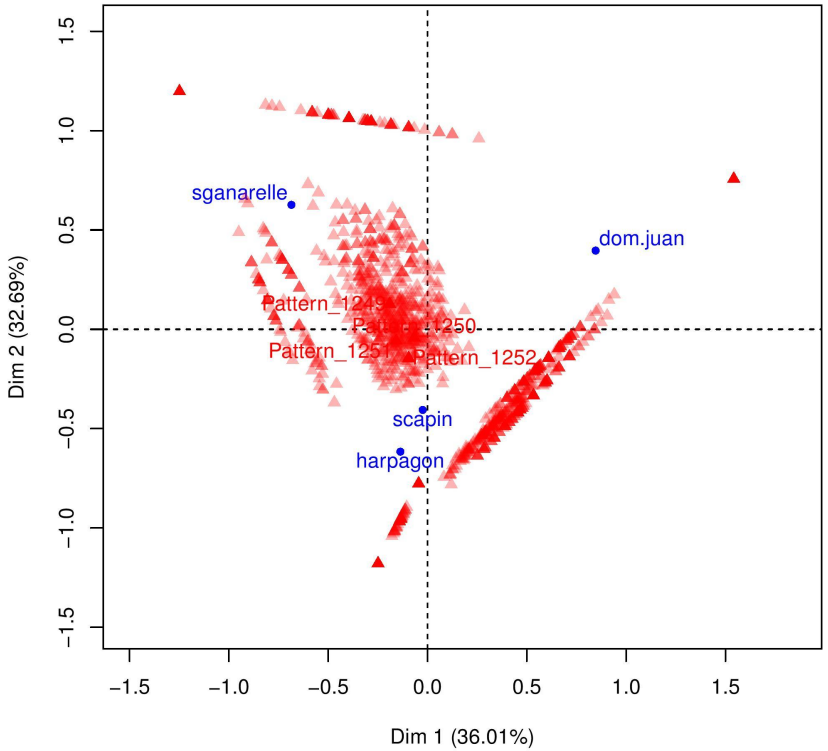
# Projection des personnages



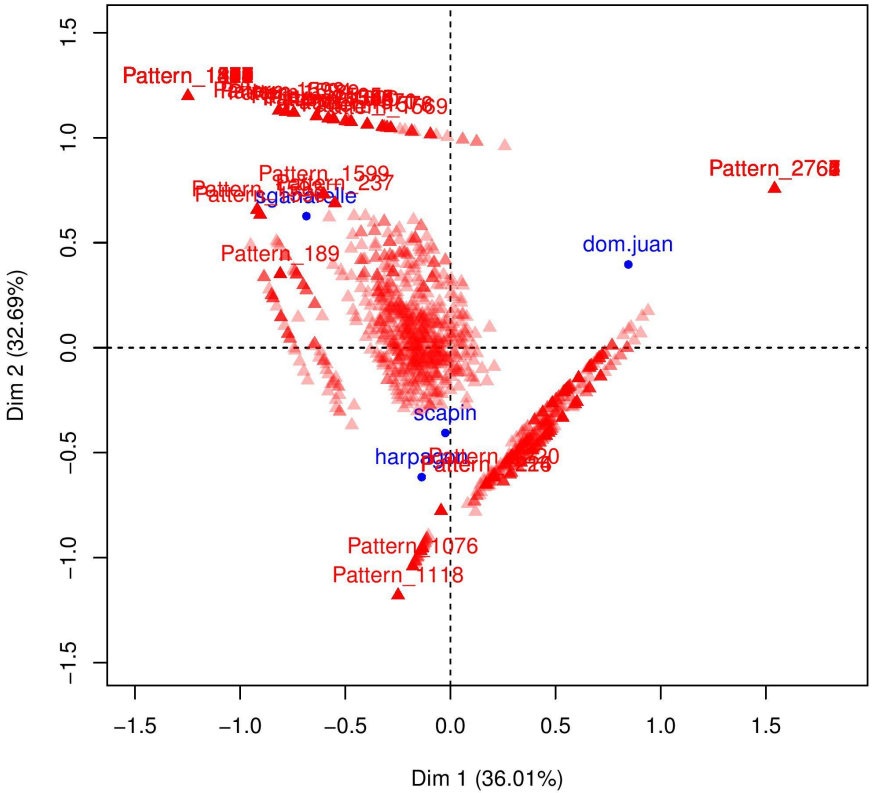
# Projection des motifs (sans filtrage)



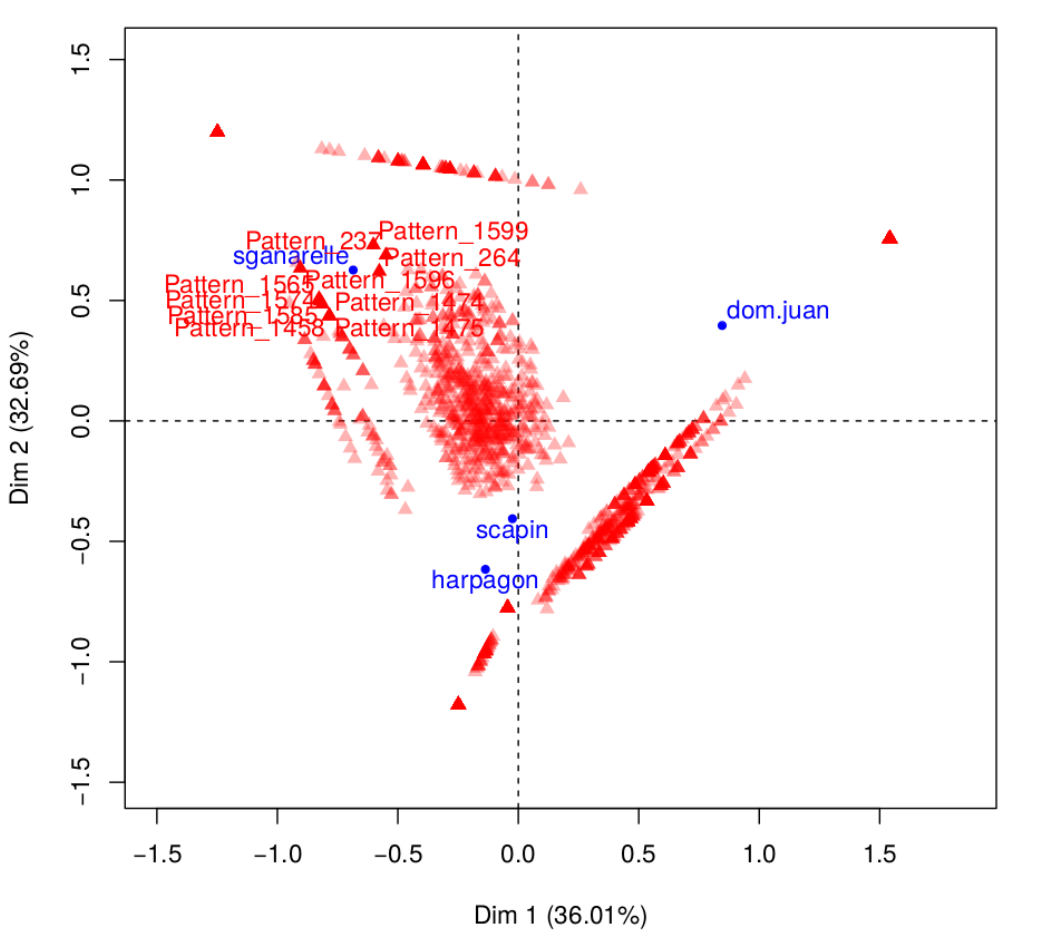
# Projection des motifs (plus fréquents)



# Filtrage par contribution (motifs top 200)



# Sganarelle


Pattern_1599
Pattern_237
sganarelle
Pattern_264
Pattern_1596
Pattern_1565
Pattern_1574
Pattern_1474
Pattern_1585
Pattern_1458
Pattern_1475
dom.juan
scapin
harpagon
Dim 2 (32.69%)
Dim 1 (36.01%)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

# Sganarelle: Exemple 1

Le caractère le plus isolé semble être Sganarelle, le protagoniste d'une pièce dans laquelle un homme simple est forcé par les circonstances à faire semblant d'être un grand médecin. Son discours est nettement différent du point de vue syntaxique.

1) [PRO:PER] [VER:pres] [KON] [*] [NOM] - **Dans les diagnostics**, par exemple:
   - .. **il arrive que ces vapeurs** ... Ossabandus , nequeys , nequer , potarinum , quipsa milus
   - **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, …
   - **on voit que l'inégalité** de leurs opinions dépend du mouvement oblique du cercle de la lune

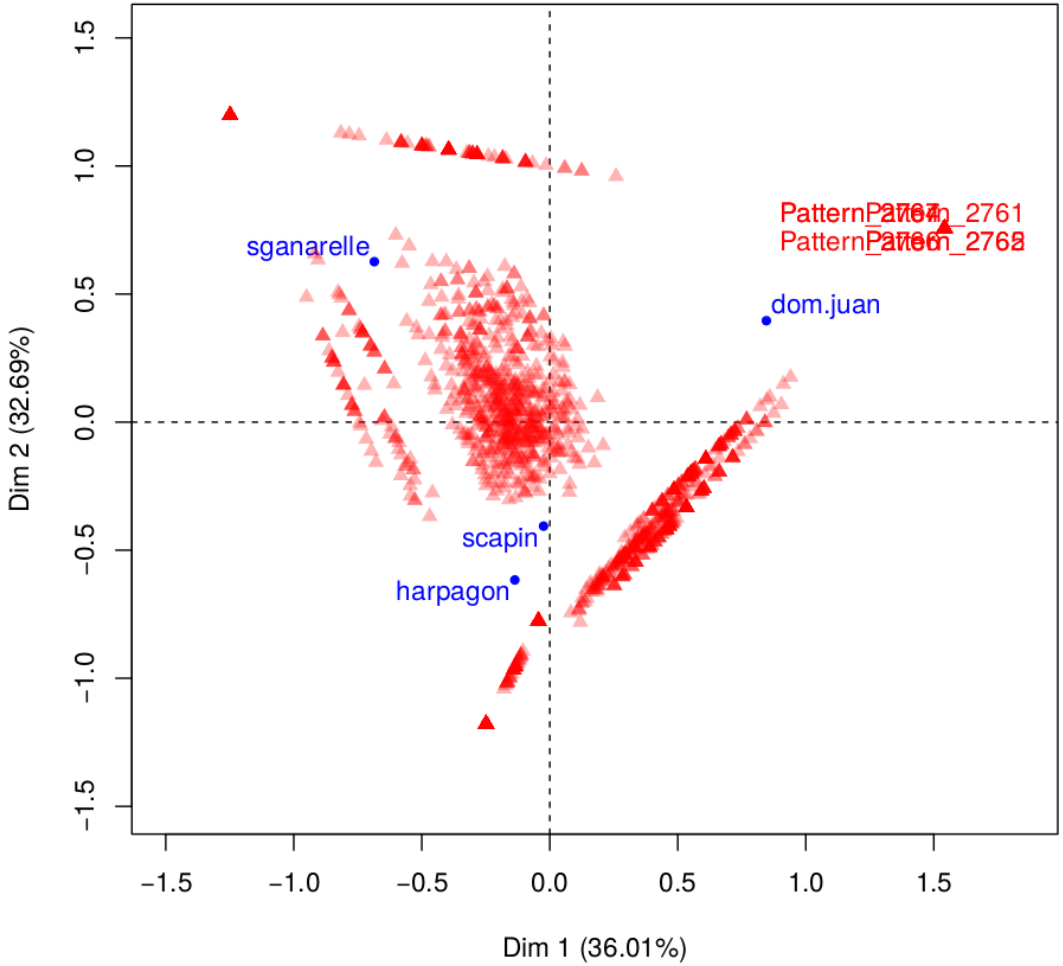
# Sganarelle: Exemple 2

Dans d'autres cas, les motifs sont utilisés pour dans un premier temps pour tenter de dissiper un malentendu, et une fois sa vraie identité est découverte, pour avouer la vérité.

2) [PRO:PER] [PRO:PER] [*] [KON] - **phrases à fonction illocutive**, par exemple:

- je te dis que …
- Je vous promets que …
- Je vous jure que …
- je vous dis que …
- Je vous assure que …
- je vous apprends que …
- je vous avoue que …

# Dom Juan


sganarelle
dom.juan
scapin
harpagon
Dim 1 (36.01%)
Dim 2 (32.69%)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

# Dom Juan: Exemple 1

Personnage complexe, Dom Juan est un noble qui reste isolé surtout par sous-représentation (moins de motifs distinctifs), ce qui peut signifier que son langage est moins répétitif et peut-être plus élaboré.

Cela ressort de manière significative dans les motifs fortement associés à son discours.

3) [KON] [PRO:PER] [*] [VER:pres] [*] [PRP] - phrases complexes

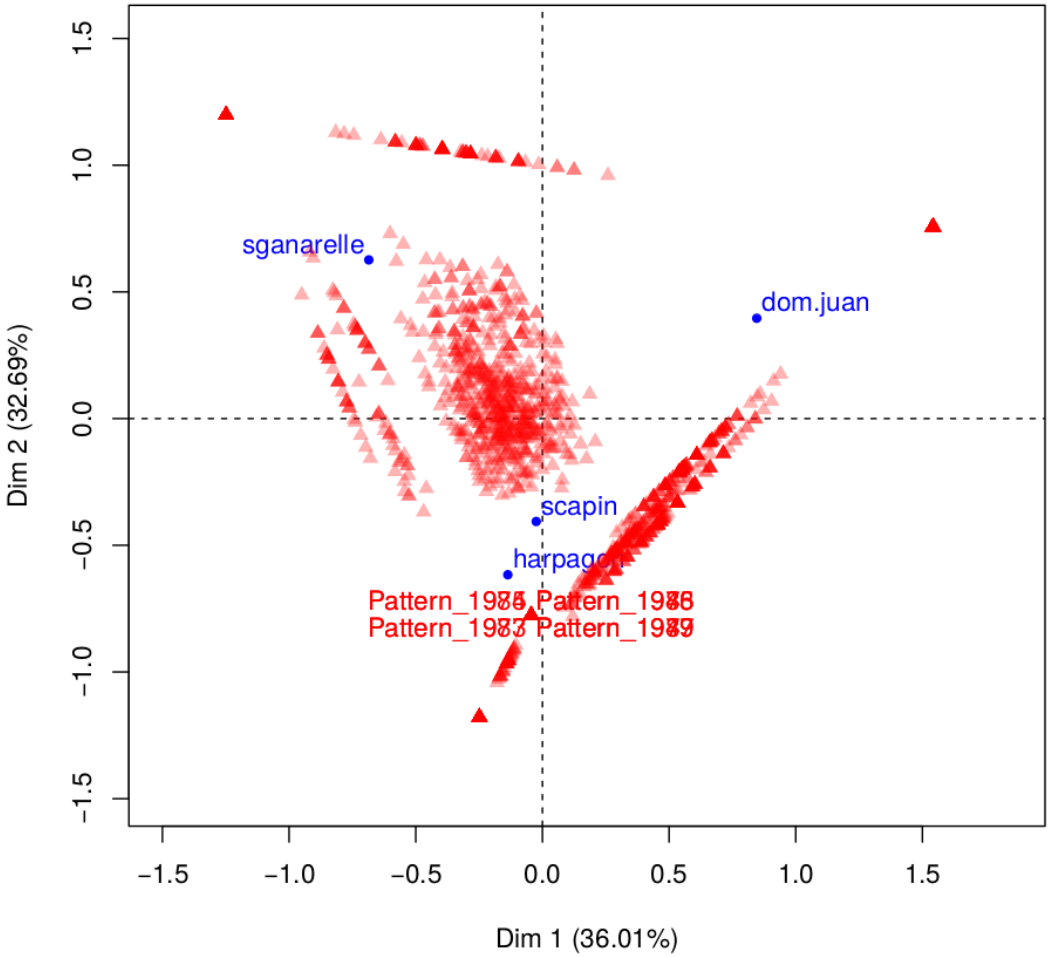
• sachez **que je n'ai point** d'autre dessein que de vous épouser …

• elle va vous dire **que je lui ai promis de** l'épouser

• Vous soutenez également toutes deux **que je vous ai promis de** vous prendre pour femmes

• … et **que je sais me servir de** mon épée quand il le faut

# Harpagon


sganarelle
dom.juan
scapin
harpagon
Pattern_1976
Pattern_1983
Pattern_1989
Dim 2 (32.69%)
Dim 1 (36.01%)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

# Harpagon

Harpagon et Scapin sont des personnages comiques à basse complexité syntaxique.

Certes, les motifs d'Harpagon sont typiques d'un personnage égoïste, autoritaire et qui peut facilement se mettre en colàre, surtout à propos d'argent.

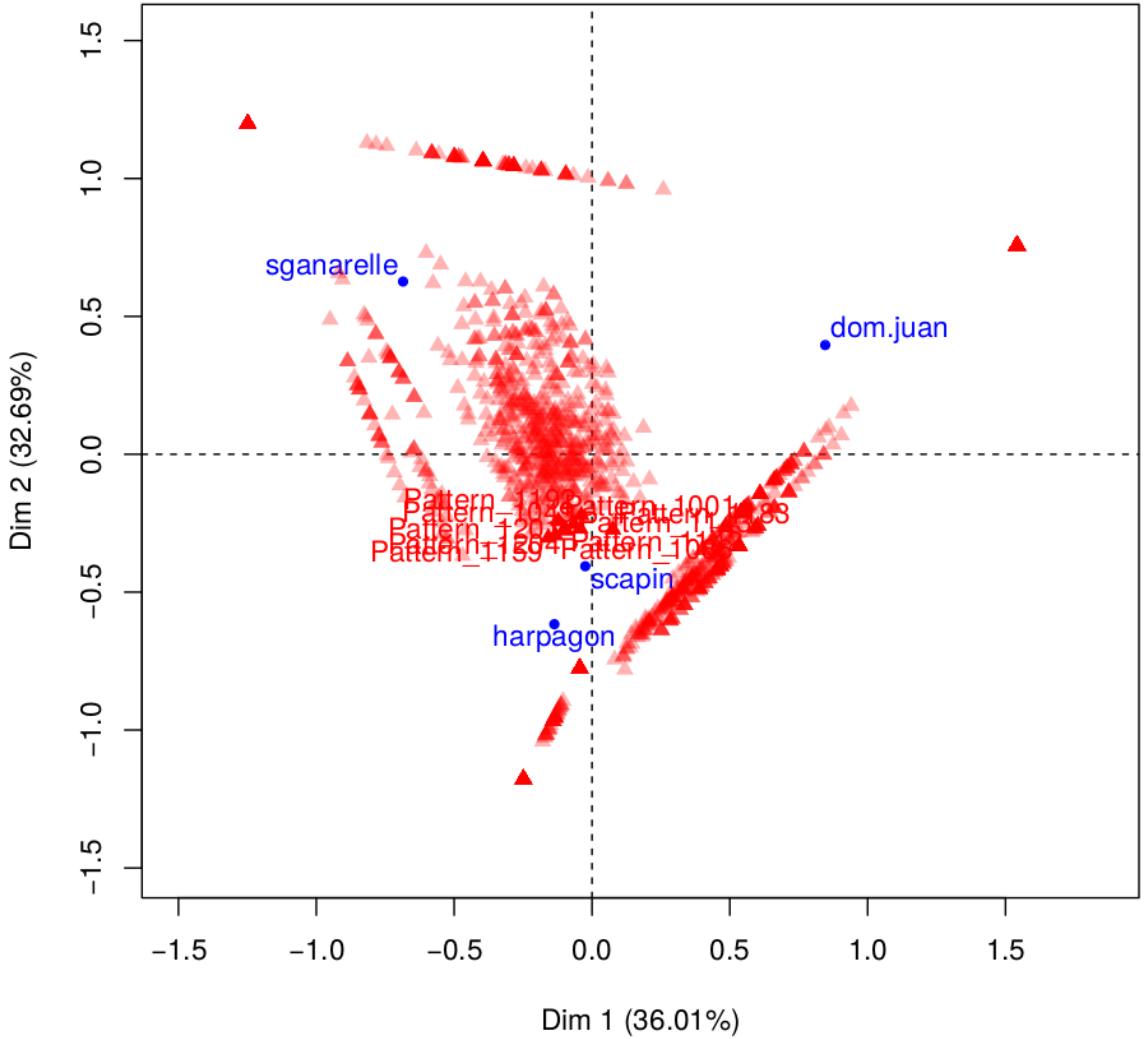
4) [PRO:PER] [PRO:PER] [VER:pres] [VER:pper]

on m' a privé … on m' a dérobé … on m' a volé … on m' a pris …

5) [KON][PRO:PER][*][KON]

- que je veux que …
- et il faut que …
- et vous verrez qu' …
- …

# Scapin


sganarelle
dom.juan
scapin
harpagon
Pattern_1159
Dim 1 (36.01%)
Dim 2 (32.69%)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

# Scapin

Les patrons propres à Harpagon remplissent une fonction différente chez Scapin, le serviteur fourbe qui doit interagir avec plusieurs personnages pour faire aboutir son plan.

En particulier, il se retrouve souvent en train de **raconter des événements**, véridiques ou inventés, à d'autres personnages.

4') [PRO:PER][PRO:PER][VER:pres][VER:pper]

- **Je l' ai trouvé** tantôt tout triste ...
- **nous nous sommes allés promener** sur le port .
- ...

# Interprétation des motifs significatifs

(Biber & Conrad, 2009) - Il est important de relier un motif significatif avec une **fonction communicative** ou un trait **psycholinguistique** des personnages.

Nous avons remarqué une certaine corrélation entre un personnage et la structure syntaxique de son discours.

Néanmoins, les motifs sont, à un certain degré, très spécifiques aux personnages  et donnent ainsi à chaque personnage donné une voix qui lui est propre.

# Réflexions sur l'analyse morpho-syntaxique

L'usage d'outils entraînés sur des corpus journalistiques pose des problèmes:

Conventions orthographiques produisant des anomalies

- **capitalisation** (M. Jourdain), apostrophe, …

Segmentation:

- **"c'est" présentatif** => [c'] [est] vs [c'est]

Etiquetage: **KON** (coordination / subordination)

Le choix d'une étiquette donnée peut considérablement affecter les résultats et produire des différences fictives entre les personnages.

Merci pour votre attention!

# Bibliographie


- 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.
- Mahlberg, M. (2012). Corpus stylistics and Dickens's fiction (Vol. 14). Routledge.
- Husson, F., Josse, J., Le, S., & Mazet, J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R, R package version 1.24.
- Fournier-Viger, P., Gomariz, Gueniche, T., A., Soltani, A., Wu., C., Tseng, V. S. (2014). SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR), 15: 3389-3393.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. Journal of statistical software, 25(1), 1-18.
- 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 Berlin Heidelberg.
- 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 Berlin Heidelberg.

# L’outil d’extraction

Nous travaillons sur l’implémentation d’un outil d’extraction de motifs :

- paramétrisation
  - syntaxe / lexico-syntaxe
  - l’ensemble simplifié et complet des classes syntaxiques de TreeTagger
  - choix de la taille du motif et des trous
- filtrage des motifs pour éliminer la redondance