



**HAL**  
open science

# Multifactorial Exploratory Approaches: principal component analysis

Guillaume Desagulier

► **To cite this version:**

Guillaume Desagulier. Multifactorial Exploratory Approaches: principal component analysis. École thématique. United Kingdom. 2019. halshs-02908483

**HAL Id: halshs-02908483**

**<https://shs.hal.science/halshs-02908483>**

Submitted on 29 Jul 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Multifactorial Exploratory Approaches

## principal component analysis

Guillaume Desagulier<sup>1</sup>

<sup>1</sup>MoDyCo (UMR 7114)  
Paris 8, CNRS, Paris Nanterre  
Institut Universitaire de France  
gdesagulier@univ-paris8.fr

Corpus Linguistics Summer School 2019  
June 25<sup>th</sup>, 2019  
University of Birmingham

# outline

- 1 introduction
- 2 principles
- 3 case study

# PCA

As in CA and MCA, the total variance of the table is decomposed into proportions in PCA.

There is one minor terminological difference: the dimensions are called **principal components**.

For each component, the proportion of variance is obtained by dividing the squared standard deviation by the sum of the squared standard deviations.

# reminder

This is the kind of data that PCA accommodates.

**Table 1:** A sample data frame (Lacheret-Dujour et al. 2019)

corpus sample	fPauses	fOverlaps	fFiller	fProm	fPI	fPA	subgenre	interactivity	planning type
D0001	0.26	0.12	0.14	1.79	0.28	1.54	argumentation	interactive	semi-spontaneous
D0002	0.42	0.11	0.10	1.80	0.33	1.75	argumentation	interactive	semi-spontaneous
D0003	0.35	0.10	0.03	1.93	0.34	1.76	description	semi-interactive	spontaneous
D0004	0.28	0.11	0.12	2.29	0.30	1.79	description	interactive	semi-spontaneous
D0005	0.29	0.07	0.23	1.91	0.22	1.69	description	semi-interactive	spontaneous
D0006	0.47	0.05	0.26	1.86	0.44	1.94	argumentation	interactive	semi-spontaneous
...	...	...	...	...	...	...	...	...	...

# principal components

PCA is based on the inspection of correlations between the variables and the principal components.<sup>1</sup>

---

<sup>1</sup>A second kind of PCA is based on loadings (Baayen 2008, Sect. 5.1.1). Loadings are correlations between the original variables and the unit-scaled principal components. The two kinds of PCA are similar: both are meant to normalize the coordinates of the data points. The variant exemplified in this chapter is more flexible because it allows for the introduction of supplementary variables.

# standardizing the variables

Before one runs a PCA, one should consider **standardizing** the variables (i.e. centering and scaling).

**If a table contains measurements in different units, standardizing the variables is compulsory.**

If a table contains measurements in the same unit, standardizing the variables is optional. However, even in this case, failing to standardize means giving each variable a weight proportional to its variance.

Standardizing the variables guarantees that equal weights are attributed to the variables (Husson et al. 2010, p. 45).

# steps

Running a PCA involves the following steps:

- determining how many components there are to inspect;
- interpreting the graph of variables and the graph of individuals.



## active vs. supplementary

As in CA and MCA, we can declare some variables as active and some other variables as supplementary/illustrative in PCA.

# inclusion in French

Gréa (2017) compares five prepositions that denote inclusion in French:

- *parmi* 'among'
- *au centre de* 'at the center of',
- *au milieu de* 'in the middle of'
- *au cœur de* 'at the heart of', and
- *au sein de* 'within'/'in'/'among'.

## inclusion in French

To determine the semantic profile of each preposition, Gréa examines their preferred and dispreferred nominal collocates in the FrWaC corpus. He uses an association measure known as *calcul des spécificités* (Habert 1985; C. Labbé and D. Labbé 1994; Salem 1987), which is based on the hypergeometric distribution.

# inclusion in French

## *calcul des spécificités*

A positive value indicates that the word is over-represented in the construction. The higher the value, the more the word is over-represented. A negative value indicates that the word is under-represented in the construction. The smaller the value, the more the word is under-represented Gréa (2017, Sect. 2.2).

# inclusion in French

First, we load the data set (`inclusion_FrWaC.txt`). As we inspect the data frame with `str()`, we see that 22,397 NPs were found. The rows contain the nominal collocates and the columns the prepositions. The cells contain the association scores. The assumption is that the semantic profiles of the prepositions will emerge from the patterns of attraction/repulsion.

```
> # clear R's memory
> rm(list=ls(all=TRUE))
> # load the data (inclusion_FrWaC.txt)
> data <- read.table(file=file.choose(), header=TRUE, row.names=1, sep="\t")
```

# inclusion in French

We load the FactoMineR package and run the PCA with the `PCA()` function. We declare all variables as active.

```
> library(FactoMineR)
> pca.object <- PCA(data, graph=F)
```

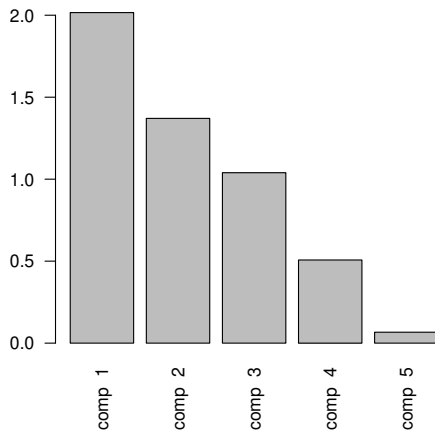
# inclusion in French

We make sure that the first two components are representative.

```
> round(pca.object$eig, 2)
      eigenvalue percentage of variance cumulative percentage of variance
comp 1      2.02           40.32           40.32
comp 2      1.37           27.42           67.74
comp 3      1.04           20.79           88.52
comp 4      0.51           10.14           98.67
comp 5      0.07            1.33          100.00
```

For this kind of analysis, the first two components should represent a cumulative percentage of variance that is far above 50%. The more dimensions there are in the input data table, the harder it will be to reach this percentage.

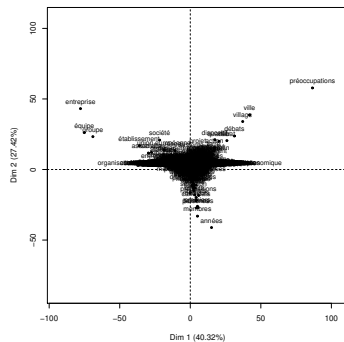
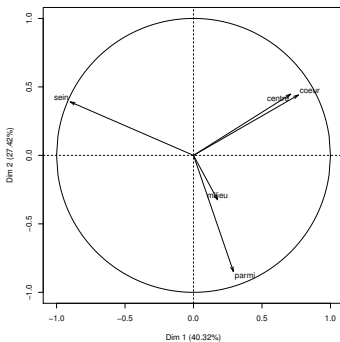
# inclusion in French





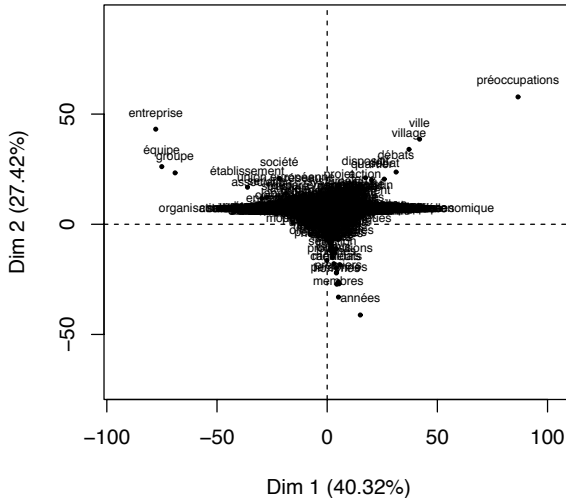
# inclusion in French

In PCA, the variables and the individuals and categories are plotted separately. The graph of variables serves as a guide to interpret the graph of individuals and categories.



# inclusion in French

Admittedly, the graph of individuals is cluttered.



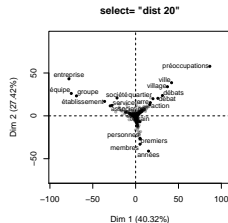
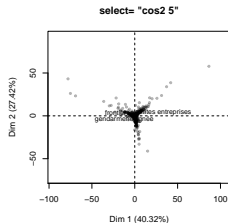
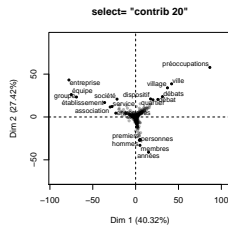
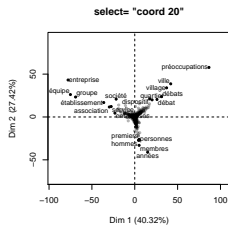
## inclusion in French

This is due to the very large number of NP types that cooccur with the prepositions.

We filter out unwanted individuals by selecting only the desired ones.

```
> plot.PCA(pca.object, select="coord 20")  
> plot.PCA(pca.object, select="contrib 20")  
> plot.PCA(pca.object, select="cos2 5")  
> plot.PCA(pca.object, select="dist 20")
```

# inclusion in French



# inclusion in French

Here is what the title of each plot means:

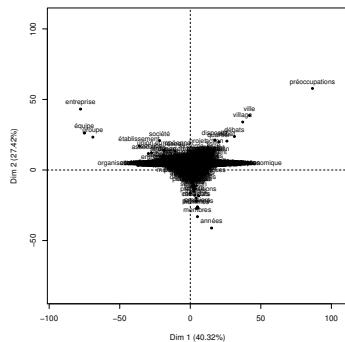
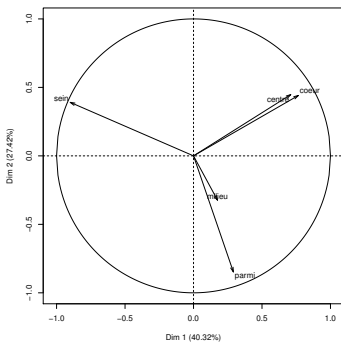
- with `select="coord 20"`, only the labels of the twenty individuals that have the most extreme coordinates on the chosen dimensions are plotted;
- with `select="contrib 20"`, only the labels of the twenty individuals that have the highest contributions on the chosen dimensions are plotted;<sup>2</sup>
- with `select="cos2 5"`, only the labels of the five individuals that have the highest squared-cosine scores on the chosen dimensions are plotted;<sup>3</sup>
- with `select="dist 20"`, only the labels of the twenty individuals that are the farthest from the center of gravity of the cloud of data points are plotted.

---

<sup>2</sup>The contribution is a measure of how much an individual contributes to the construction of a component.

<sup>3</sup>The squared cosine ( $\cos^2$ ) is a measure of how well an individual is projected onto a component.

# inclusion in French



# inclusion in French

Clear trends emerge:

- the *au sein de* construction tends to co-occur with collective NPs that denote groups of human beings (*entreprise* 'company/business', *équipe* 'team', *établissement* 'institution/institute', etc.);

# inclusion in French

Clear trends emerge:

- the *au sein de* construction tends to co-occur with collective NPs that denote groups of human beings (*entreprise* 'company/business', *équipe* 'team', *établissement* 'institution/institute', etc.);
- the *au centre de* and *au cœur de* constructions tend to co-occur with NPs that denote urban areas (*ville* 'city/town', *village* 'village', *quartier* 'district') and thoughts or ideas (*préoccupations* 'concerns/issues', *débat* 'debate/discussion/issue');



# inclusion in French

Clear trends emerge:

- the *au sein de* construction tends to co-occur with collective NPs that denote groups of human beings (*entreprise* 'company/business', *équipe* 'team', *établissement* 'institution/institute', etc.);
- the *au centre de* and *au cœur de* constructions tend to co-occur with NPs that denote urban areas (*ville* 'city/town', *village* 'village', *quartier* 'district') and thoughts or ideas (*préoccupations* 'concerns/issues', *débat* 'debate/discussion/issue');
- the *au milieu de* and *parmi* constructions tend to co-occur with plural NPs that denote sets of discrete individuals (*hommes* 'men', *personnes* 'persons', *membres* 'members'), among other things.

# inclusion in French

- the PCA graph does a good job at grouping prepositions based on the nominal collocates that they have in common and revealing consistent semantic trends.
- however, some finer semantic phenomena are harder to capture.

# inclusion in French

## e.g. *centre* vs. *cœur*

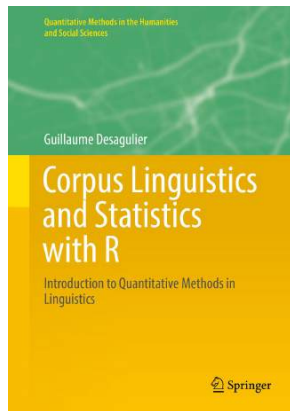
- *au centre du conflit* 'at the center of the conflict' profiles a participant that is either the instigator of the conflict or what is at stake in the conflict.
- *au cœur du conflit* 'at the heart of the conflict' denotes the peak of the conflict, either spatially or temporally.

The kind of collocational approach exemplified in the paper, which does not aim to (and is not geared to) reveal fine-grained semantic differences by itself.

# *Practical Handbook of Corpus Linguistics*








Guillaume Desagulier (to appear). “Multifactorial exploratory approaches.” In: *Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Thomas Gries. New York: Springer

# *Corpus Linguistics and Statistics with R*



Section 10.2 – (Desagulier 2017)

# Bibliography I

-  Baayen, R Harald (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
-  Desagulier, Guillaume (to appear). "Multifactorial exploratory approaches." In: *Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Thomas Gries. New York: Springer.
-  – (2017). "Clustering Methods." In: *Corpus Linguistics and Statistics with R*. New York, NY: Springer, pp. 239–294.
-  Gréa, Philippe (2017). "Inside in French." In: *Cognitive Linguistics* 28.1, pp. 77–130.
-  Habert, Benoît (1985). "L'analyse des formes «spécifiques» [bilan critique et propositions d'utilisation]." In: *Mots* 11.1, pp. 127–154.
-  Husson, François, Sébastien Lê, and Jérôme Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. London: CRC press.
-  Labbé, Cyril and Dominique Labbé (1994). "Que mesure la spécificité du vocabulaire ?" In: *Lexicometrica* 3, p. 2001.

## Bibliography II



Lacheret-Dujour, Anne et al. (2019). “The distribution of prosodic features in the Rhapsodie corpus.” In: *Rhapsodie: A prosodic and syntactic treebank for spoken French*. Ed. by Anne Lacheret-Dujour and Sylvain Kahane. Studies in Corpus Linguistics 89. John Benjamins. Chap. 17, pp. 315–338.



Salem, André (1987). *Pratique des segments répétés: essai de statistique textuelle*. Paris: Klincksieck.