Multifactorial Exploratory Approaches

Guillaume Desagulier

Multifactorial Exploratory Approaches

Guillaume Desagulier

Abstract This chapter presents four methods that are designed to explore and summarize large and complex data tables by means of summary statistics: correspondence analysis, multiple correspondence analysis, principal component analysis, and exploratory factor analysis. These methods help generate hypotheses by providing informative clusters using the variable values that characterize each observation.

Key words: correspondence analysis, exploratory factor analysis, multiple correspondence analysis, principal component analysis

1 Introduction

Once corpus linguists have collected sizeable amounts of observations and described each observation with relevant variables, they look for patterns in the data. When the data set is too large, it becomes impossible to summarize the table with the naked eye and summary statistics are needed. This is where exploratory data analysis steps in.

Exploring a data set means separating meaningful trends from the noise (i.e. "random" distributions).[1] In theory, exploratory data analysis is used to generate hypotheses because the linguist does not yet have any assumption as to what kinds of trends should appear in the data. In practice, however, linguists collect observations in the light of specific variables precisely because they expect that the latter influence the distribution of the former.

Guillaume Desagulier
MoDyCo — Université Paris 8, CNRS, Université Paris Nanterre, Institut Universitaire de France
e-mail: gdesagulier@univ-paris8.fr

[1] I am using scare quotes because, as Kilgarriff (2005) puts it, "language is never, ever, ever, random".

I present four multifactorial exploratory techniques: correspondence analysis (henceforth CA), multiple correspondence analysis (henceforth MCA), principal component analysis (henceforth PCA), and exploratoty factor analysis (henceforth EFA). These techniques rely on dimensionality reduction, i.e. an attempt to simplify complex multidimensional datasets to facilitate interpretation.

In Sect. 2, I explain the foundations of the four techniques, showing what they have in common and to what extent they differ. In Sect. 3, I illustrate each method with a case study. In Sect. 4, I show how to run each technique with R. For CA, MCA, and PCA, I rely mainly on the `FactoMineR` package because its functions and arguments are simple to understand, and because the package is widely documented and supported. I also make reference to alternative packages. For EFA, I rely on the `factanal()` function, which is part of base R.

Regarding dimensionality-reduction methods, principal component analysis is presented first in most textbooks because it came first and inspired the other methods. In this chapter, I make an exception and present the methods in increasing order of complexity with respect to the kind of data involved, starting with correspondence analysis.

2 Fundamentals

CA, MCA, PCA, and EFA are multifactorial methods because they are meant for the exploration of phenomena whose realizations are influenced by several factors at the same time. Once operationalized by the researcher, these multiple factors are captured by means of several independent variables. When observations of a phenomenon are captured by several variables, the analysis is multivariate. For this reason, multifactorial methods are also considered multivariate.

2.1 Commonalities

The challenge that underlies the visualizations obtained with dimensionality-reduction methods is the following: we seek to explore a cloud of points from a data set in the form of a *rows* $\times$ *columns* table with as many dimensions as there are columns. Like a complex object in real life, a data table has to be rotated so as to be observed from an optimal angle. Although the dimensions of a data table are eventually projected in a two-dimensional plane, they are not spatial dimensions. If the table has K columns, the data points are initially positioned in a space $\mathbb{R}$ of K dimensions. To allow for easier interpretation, dimensionality-reduction methods decompose the cloud into a smaller number of meaningful planes.

All the methods covered in this chapter summarize the table by measuring how much variance there is and decomposing the variance into proportions. These pro-

portions are eigenvalues in CA, MCA, and PCA. They are loadings in EFA (and a special kind of PCA not covered in this chapter).[2]

All four methods offer graphs that facilitate the interpretation of the results. Although convenient, these graphs do not replace a careful interpretation of the numerical results.

2.2 Differences

The main difference between these methods pertain mainly to the kind of data that one works with. CA takes as input a contingency table, i.e. a table that one cross-classifies observations on a number of categorical variables. Entries in each cell are integers, namely the number of times that observations (in the rows) are seen in the context of the variables (in the columns). Table 1 is an example of a contingency table. It displays the frequency counts of four types of nouns (rows) across three corpus files from the BNC-XML (columns).

Table 1: An example of a contingency table (Desagulier 2017, 153)

	A1J.xml	A1K.xml	A1L.xml	row totals
NN0	136	14	8	158
NN1	2236	354	263	2853
NN2	952	87	139	1178
NP0	723	117	71	911
column totals	4047	572	481	5100

MCA takes as input takes as input a table of nominal data such as Tab. 2. The table consists of *i* individuals or observations (rows) and *j* variables (columns). Historically, MCA was developed to explore the structure of surveys in which informants are asked to select an answer from a list of suggestions. For example, the question "According to you, which of these disciplines best describe the hard sciences: physics, biology, mathematics, computer science, or statistics?" requires informants to select one category.

PCA takes as input a table of data of *i* individuals or observations (rows) and *j* variables (columns). PCA handles continuous and nominal data. The continuous data may consist of means, reaction times, formant frequencies, etc. The categorical/nominal data are used to tag the observations. Table 3 is a table of 6 kinds of mean frequency counts further described by 3 kinds of nominal information.

Like PCA, EFA takes as input a table of continuous data. However, it does not commonly accommodate nominal data. Typically, Tab. 3 minus the 3 columns of nominal data can serve as input for EFA.

[2] See Baayen (2008, Sect. 5.1.1).

Table 2: A sample input table for MCA (Desagulier 2017, 36)

corpus file	mode	genre	exact match	intensifier	syntax	adjective
KBF.xml	spoken	conv	*a quite ferocious mess*	quite	preadjectival	*ferocious*
AT1.xml	written	biography	*quite a flirty person*	quite	predeterminer	*flirty*
A7F.xml	written	misc	*a rather anonymous name*	rather	preadjectival	*anonymous*
ECD.xml	written	commerce	*a rather precarious foothold*	rather	preadjectival	*precarious*
B2E.xml	written	biography	*quite a restless night*	quite	predeterminer	*restless*
AM4.xml	written	misc	*a rather different turn*	rather	preadjectival	*different*
F85.xml	spoken	unclassified	*a rather younger age*	rather	preadjectival	*younger*
J3X.xml	spoken	unclassified	*quite a long time*	quite	predeterminer	*long*
KBK.xml	spoken	conv	*quite a leading light*	quite	predeterminer	*leading*

Table 3: A sample data frame (Lacheret-Dujour et al., to appear)

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
…	…	…	…	…	…	…	…	…	…

2.3 Exploring is not predicting

The methods presented in this chapter are exploratory, as opposed to explanatory or predictive. They help find structure in multivariate data thanks to observation groupings. The conclusions made with these methods are therefore valid for the corpus only. For example, we shall see that middle-class female speakers aged 25 to 59 display a preference for the use of *bloody* in the British National Corpus (Sect. 3.2). This finding should not be extended to British English in general. Indeed, we may well observe different tendencies in another corpus of British English. Neither should the conclusions made with exploratory methods be used to make predictions. Of course, exploratory methods serve as the basis for the design of predictive modeling, which uses the values found in a sample to predict values for another sample.

Nowadays, many linguists jump to powerful predictive methods (such as logistic regression or discriminant analysis) without going to the trouble of exploring their data sets first. This is a shame because the point of running a multifactorial exploratory analysis is to generate fine research hypotheses, which the far more powerful predictive methods can only benefit from.

2.4 Correspondence analysis

Correspondence analysis (henceforth CA) is used to summarize a two-dimensional contingency table. The table is a matrix M of counts that consists of i individuals or observations (rows) and j variables (columns). The foundations of CA were laid out by Hirschfeld (1935) and Benzécri (1984). The method gets its name from what it aims to show, namely the correspondence between what the rows and the columns represent.

CA reveals the correspondence between the rows and the columns of the table. Incidentally, it also shows the correspondence between the rows and the correspondence between the columns. The basic idea is to group the rows and columns that share identical profiles.

Remember that, the linguist makes no assumption as to what kinds of groupings are to be found in the data. In practice, however, you compile a table of data because you expect to find meaningful groupings. Therefore, if you find no meaningful grouping, this is because your rows and your columns are independent. The chances are that you might want to rethink the design of your study, especially your choice of explanatory variables.

To determine whether rows and columns are independent, CA relies on the χ^2 test. It tests the significance of the overall deviation of the table from the independence model. The test computes the contribution of each cell to χ^2 and sums up all contributions to obtain the χ^2 statistic. Because we are interested in determining whether two variables are interdependent, we formulate the hypotheses as follows:

H_0: the distribution of row variables and the column variables are independent;
H_1: the distribution of row variables and the column variables are interdependent.

One calculates the χ^2 value of a cell in the i^{th} row and the j^{th} column as follows:

$$\chi^2_{i,j} = \frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}} \tag{1}$$

where $E_{i,j}$ is the expected frequency for cell i,j and $O_{i,j}$ is the expected frequency for cell i,j. The χ^2 statistic of the whole table, χ^2 is the sum of the χ^2 values of all cells.

$$\chi^2 = \sum_{i=1}^{n} \frac{(E-O)^2}{E} \tag{2}$$

Because the χ^2 score varies greatly depending on the sample size, it cannot be used to assess the magnitude of the dependence. This is measured with Cramér's V, which one obtains by taking the square root of the χ^2 statistic divided by the product of the sum of all observations and the number of columns minus one:

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{N(k-1)}}. \tag{3}$$

Central to CA is the concept of profile. To obtain the profile of a row, each cell is divided by its row total. Table 4 displays the row profiles of Tab. 1. The row profiles add up to 1. Likewise, one obtains the profile of a column by dividing each column frequency by the column total (Tab. 5). Again, the column profiles add up to 1.

Table 4: The row profiles of Tab. 1

	A1J.xml	A1K.xml	A1L.xml	row total
NN0	0.8608	0.0886	0.0506	1
NN1	0.7837	0.1241	0.0922	1
NN2	0.8081	0.0739	0.1180	1
NP0	0.7936	0.1284	0.0779	1
column average	0.7935	0.1122	0.0943	1

Table 5: The column profiles of Tab. 1

	A1J.xml	A1K.xml	A1L.xml	row average
NN0	0.0336	0.0245	0.0166	0.0310
NN1	0.5525	0.6189	0.5468	0.5594
NN2	0.2352	0.1521	0.2890	0.2310
NP0	0.1787	0.2045	0.1476	0.1786
column total	1	1	1	1

Comparing row profiles to their average row profile leads to the same conclusions as comparing column profiles to their average column profile. For example, if we compare the profile of singular common nouns (NN1) observed in file A1K.xml (0.1241) to the profile of all nouns used in file A1K.xml (0.1122) in Tab. 4, we obtain a ratio of $\frac{0.1241}{0.1122} \approx 1.1063$. This implies that the use of singular common nouns in A1K.xml is slighty over the average profile of nouns. In Tab. 5, the profile of singular common nouns in the same file is 0.6189, whereas the average profile of singular common nouns is 0.5594. We obtain the same ratio as above, i.e. 1.1063. These relative frequencies have special geometric features, which CA converts into coordinates and visualizes as points on a two-dimensional map.

Distances between profiles are measured with inertia. It is with the total inertia of the table (ϕ^2) that CA measures how much variance there is. ϕ^2 is obtained by dividing the χ^2 statistic by the sample size. CA interprets inertia geometrically to assess how far row/column profiles are from their respective average profiles. The larger ϕ^2, the more the data points are spread out on the map.

Each column of the table contributes one dimension. The more columns in your table, the larger the number of dimensions. When there are many dimensions, summarizing the table becomes very difficult. To solve this problem, CA decomposes

ϕ^2 along a few dimensions that concentrate as much information as possible. This is measured in terms of eigenvalues, which are proportions of inertia.

On top of the coordinates of the data points, two descriptors help interpret the dimensions: contribution, and quality of projection (cos^2). If a data point displays a minor contribution to a given dimension, its position with respect to this dimension must not be given too much relevance. The quality of the projection of a data point onto a dimension is measured as the percentage of inertia associated with this dimension. Usually, projection quality is used to select the dimension in which the individual or the variable is the most faithfully represented.

Individuals and variables can be declared as active or supplementary/illustrative, as is the case with multiple correspondence analysis and principal component analysis (see below). These supplementary rows and/or columns help interpret the active rows and columns. As opposed to active elements, supplementary elements do not contribute to the construction of the dimensions. Supplementary information is mostly used to help interpret the results.

The standard graphic output of CA is a symmetric biplot in which both row variables and column variables are represented in the same space using their coordinates. In this case, only the distance between row points or the distance between column points can be interpreted accurately (Greenacre 2007, 72). Only general observations can be made about the distance between row points and column points, when these points appear in the same part of the plot with respect to the center of the cloud of points (François Husson, pc). Assessing the inter-distance between rows and columns accurately is possible in either an asymmetric biplot or a scaled symmetric biplot. In an asymmetric biplot, either the columns are represented in row space or the rows are represented in a column space. In a scaled symmetric biplot, neither the row metrics nor the column metrics are preserved. Rows and columns are scaled to have variances equal to the square roots of eigenvalues, which allows for direct comparison in the same plot.[3]

2.5 Multiple correspondence analysis

Because MCA is an extension of CA, its inner workings are very similar. For this reason, they are not repeated here.

As pointed out in Sect. 2.2, it takes as input a table of nominal data. For MCA to yield manageable results, it is best if the table is of reasonable size (not too many columns), and if each variable does not break down into too many categories. Otherwise, the contribution of each dimension to ϕ^2 is small, and a large number of dimensions must be inspected. There are no hard and fast rules for knowing when there are too many dimensions to inspect. However, when the eigenvalue that corresponds to a dimension is low, we know that the dimension is of little interest (the chances are that the data points will be close to the intersection of the axes in the

[3] This possibility is not offered in `FactoMineR`. It is offered in the `factoextra` (Kassambara and Mundt 2017) and `ca` (Nenadic and Greenacre 2007) packages.

summary biplot). Michael Greenacre's `mjca()` function from the `mjca` package offers a solution to this issue.

2.6 Principal component analysis

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.

There are two main kinds of PCAs. As exemplified in Baayen (2008, Sect. 5.1.1), PCA is based on loadings. As exemplified in this chapter, PCA is based on the inspection of correlations between the variables and the principal components. Although the two kinds are similar (both are meant to normalize the coordinates of the data points), the latter is more flexible because it allows for the introduction of supplementary variables.

Before one runs a PCA, one should consider standardizing (i.e. centering and scaling) the variables. 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, Lê, and Pagès 2010, 45).

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. In the graph of variables, each variable is represented as an arrow. The circle is known as the circle of correlations. The closer the end of an arrow is to the circle (and the farther it is from where the axes intersect at the center of the graph) the better the corresponding variable is captured by the two components, and the more important the components are with respect to this variable.

2.7 Exploratory factor analysis

EFA is very close PCA. Although popular in linguistics,[4] EFA is not as comprehensive as PCA. Furthermore, the number of relevant components, which are called factors, is not determined automatically. It must be chosen before we run the analysis. It is therefore common to run a PCA beforehand to see how many meaningful components there are to inspect. The number of meaningful components in PCA is used to determine the number of factors in EFA.

[4] See Biber's studies on register variation (Biber 1991, 1995)

One added value of EFA is that "an error term is added to the model in order to do justice to the possibility that there is noise in the data" (Baayen 2008, 127). Factor analysis of mixed data (FAMD) accommodates data sets containing both continuous and nominal data (Pagès 2014, Chap. 3). In this respect, it should be considered an interesting alternative to standard EFA. For reasons of space, however, this chapter focuses on 'plain' EFA. Another added value of EFA is that it can identify certain unobservable factors which are not explicit in the data. This is an aspect that PCA is not designed to show.

Central to EFA is the concept of uniqueness. It is the variance that is 'unique' to the variable, i.e. that is not shared with other variables. It is equal to 1 minus the variance that is shared with other variables. Notice that the higher the uniqueness score of the variable, the lower its relevance in the factor model.

Also central to EFA is rotation, a procedure meant to clarify the relationship between variables and factors. As its name indicates, it rotates the factors to align them better with the variables. The two most frequent rotation methods are varimax and promax. With varimax, the factor axes are rotated in such a way that they are still perpendicular to each other. The factors are uncorrelated and the production of 1s and 0s in the factor matrix is maximized. With promax, the factor axes are rotated in an oblique way. The factors are correlated. With promax and the resulting model provides a closer fit to the data than with varimax. In either case, the goal is to arrive at a few common meaningful factors. Rotation is optional as it does not modify the relationship between the factors and the variables.

3 Case studies

Below, each of the four methods is illustrated by means of a case study. Because PCA and EFA are very similar, both methods are applied to the same data set.

3.1 Correspondence analysis

Leitner (1991) reports a study by Hirschmuller (1989) who compares the distribution of complex prepositions in three corpora of English: the Brown Corpus, the LOB Corpus, and the Kolhapur Corpus. The Brown Corpus is a corpus of American English (Francis and Kučera 1964). The LOB Corpus is the British counterpart to the Brown Corpus (Leech, Johansson, and Hofland 1978; Leech et al. 1986). The Kolhapur Corpus is a corpus of Indian English (Shastri, Patilkulkarni, and Shastri 1986).

Complex prepositions are multiword expressions (i.e. expressions that consist of several words): *ahead of*, *along with*, *apart from*, *such as*, *thanks to*, *together with*, *on account of*, *on behalf of*, or *on top of*. In Hirschmüller's data, 81 prepositions consist of two words and 154 of three and more out of a total of 235 complex prepo-

sitions. He observes a higher incidence of complex prepositions in the Kolhapur Corpus than in the other two corpora. He also observes that the most complex prepositions (i.e. prepositions that consist of three words and more) are over-represented in the corpus of Indian English.

3.1.1 Research questions

Leitner (1991, 224) interprets Hirschmüller's results in the light of the following assumption:

> "Their use is often associated with the level of formality (Quirk et al. 1985) or regarded as bad style. Since non-native Englishes are often claimed to use a more formal register than native Englishes, complex prepositions provide a little studied testing ground."

Following Following Leitner (1991, 224), we replicate Hirschmüller's study based on a two-fold assumption:

- complex prepositions are likely to be over-represented in the Kolhapur corpus;
- within the corpus, complex prepositions are likely to be over-represented in the more formal text categories.

3.1.2 Data

The data are gathered in a contingency table. It displays the number of times each preposition type is found in a certain context. Its structure is the following:

- 257 rows (one row per preposition type);
- 19 columns (one column per variable).

Table 6 is a snapshot of the contingency table.

Table 6: The preposition data set (10 lines out of 257 and 8 columns out of 19 are displayed)

	Brown	Kolhapur	LOB	adventure, western, fiction	belles lettres	general fiction	…	preposition length
on account of	14	23	16	2	9	1	…	3
atop	4	3	1	0	1	2	…	1
as to	147	55	122	3	47	10	…	2
owing to	2	14	19	0	4	0	…	2
save	34	43	47	8	16	19	…	1
aside from	13	1	0	0	2	0	…	2
to	23357	9771	22538	2740	8806	3734	…	1
through	810	706	623	127	329	165	…	1
but	3844	2260	2958	518	1581	805	…	1
over	973	762	953	229	322	266	…	1
…	…	…	…	…	…	…	…	…

Each column stands for a context where the preposition is found. There are three kinds of columns. The first three columns correspond to the three corpora. The next

15 columns correspond to the text categories. The nineteenth column specifies the word length of the prepositions.

3.1.3 Method

The first three columns will be declared as active. The last sixteen columns will be illustrative. Running a CA involves the following steps:

- inspecting the χ^2 score to decide whether the table deviates from independence;
- determining how many dimensions there are to inspect by means of the proportions of eigenvalues;
- interpreting the CA graph.

3.1.4 Results

The χ^2 score is very high (10061.31) and it is associated with the smallest possible p-value (0). The deviation of the table from independence is beyond doubt. Admittedly, the assumptions of the χ^2 test are not all met. One of them stipulates that 80% of the sample size should be greater than 5. Our table contains many cells whose values are smaller than 5. Therefore, it does not meet the assumption. While this should be kept in mind, it does not preclude the fact that the choice of a preposition and the variety of English are globally interdependent, no matter what given the importance of the score. Furthermore, the χ^2 test is used in an exploratory context, not a hypothesis-testing context. Just because its conditions are not fully met does not mean it is irrelevant.

The intensity of the relationship is definitely small, but non negligible for this sort of data: Cramér's $V = 0.111$. A score of 1 would be unrealistic as it would attest an exclusive association between the use of prepositions and the dialect of English.

Because the input table is simple and because the number of active variables is low, there are only two dimensions to inspect (Tab. 9). Indeed, the first two dimensions represent 100% of the variance of the table. In most cases, however, we should expect to inspect more than two dimensions. Our decision is based on the cumulative percentage of variance.

Table 7: Eigenvalues and percentages of variance explained by the two dimensions

	eigenvalue	percentage of variance	cumulative percentage of variance
dim1	0.020	82.34	82.34
dim2	0.004	17.66	100

The inertia (i.e. the sum of eigenvalues) is low (0.0248). This means that there is not much variance in the table and that the data points are not going to cluster near

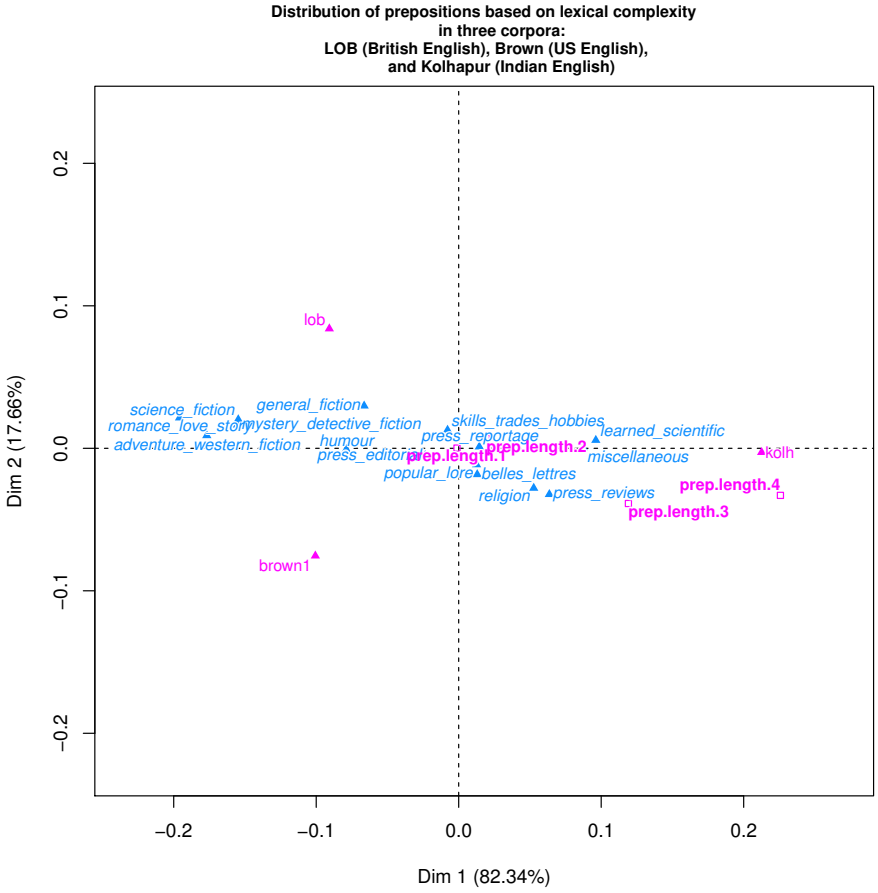

Fig. 1: CA biplot: a plane representation of individuals and variables (active and illustrative)

the intersection of the horizontal and vertical axes. In other words, the tendencies that we are about to observe are subtle.

Hirschmüller observed the following: (1) complex prepositions cluster in non-fictional texts, a preference that is amplified in the Kolhapur Corpus; (2) learned and bureaucratic writing shows a more pronounced pattern in the Kolhapur Corpus than in the British and American corpora. The CA plot reflects these tendencies (Fig. 1).

The first dimension (along the horizontal axis) accounts for 82.29% of the variance. It shows a clear divide between Brown and LOB (left) and Kolhapur (right). Large complex prepositions (three words and more: `prep.length.3` and `prep.length.4`) are far more likely to occur in Indian English than in British or US English. No such preference is observed for one-word and two-word prepo-

sitions (`prep.length.1` and `prep.length.2`). Very formal text categories cluster to the right, along with the Kolhapur corpus: `learned_scientific`, `press_reviews`, and `religion, miscellaneous` (governmental documents, foundation reports, industry reports, college catalogue, industry in-house publications). All in all, complex prepositions are specific to the Kolhapur Corpus, especially in formal contexts.

3.2 Multiple correspondence analysis

Schmid (2003) provides an analysis of sex differences in the 10M-word spoken section of the British National Corpus (BNC). Schmid shows that women use certain swear-words more than men, although swear-words which tend to have a perceived 'strong' effect are more frequent in male speech. Schmid's study is based on two subcorpora, which are both sampled from the spoken section of the BNC. One subcorpus is marked as all utterances produced by men, and the other by women. The subcorpora amount to 8,173,608 words. The contributions are not equally shared among men and women since for every 100 word spoken by women, 151 are spoken by men. To calculate the distinctive lexical preferences of men and women, while taking the lack of balance in the contributions into account, Schmid uses a coefficient formula borrowed from Leech and Fallon (1992, 30) and Hofland and Johansson (1982). This formula is based on normalized frequencies per million words. Its score ranges from -1 (if a word occurs more frequently in women's utterances) to 1 (if a word occurs more frequently in male speech). Absolute frequencies are used to calculate the significance level of the differences using the hypergeometrical approximation of the binomial distribution. With respect to swear-words, Schmid's conclusion is that both men and women swear, but men tend to use stronger swear-words than women.

3.2.1 Research questions

We repeat Schmid's study and explore the distribution of swear-words with respect to gender in the BNC-XML. Our objective is to see if:

- men swear more than women;
- some swear-words are preferred by men or women;
- the gender-distribution of swear-words is correlated with other variables: age and social class.

3.2.2 Data

Unlike Schmid, and following Rayson, Leech, and Hodges (1997), we extract our data from the demographic component of the BNC-XML, which consists of spon-

taneous interactive discourse. The swear-words are: *bloody*, *damn*, *fuck*, *fucked*, *fucker*, *fucking*, *gosh*, and *shit*. We include two exploratory variables, age and social class.[5] Tab. 8 is a snapshot of the data set.

Table 8: A snapshot of the data

word	gender	age	social class
bloody	m	Ag2	C1
shit	m	Ag1	C2
gosh	f	Ag5	AB
damn	m	Ag0	AB
bloody	m	Ag4	C1
shit	f	Ag0	DE
bloody	f	Ag2	C2
bloody	f	Ag4	C2
gosh	m	Ag4	AB
bloody	f	Ag2	AB
...	...	...	...

The data set contains 293289 swear-words. These words are described by three categorical variables (nominal data):

- gender (2 levels: male and female)
- age (6 levels: Ag0, Ag1, Ag2, Ag3, Ag4, Ag5)
- social class (4 levels: AB, C1, C2, DE)

Age breaks down into 6 groups:

- `Ag0`: respondent age between 0 and 14;
- `Ag1`: respondent age between 15 and 24;
- `Ag2`: respondent age between 25 and 34;
- `Ag3`: respondent age between 35 and 44;
- `Ag4`: respondent age between 45 and 59;
- `Ag5`: respondent age is 60+.

Social classes are divided into 4 groups:

- `AB`: higher management: administrative or professional.
- `C1`: lower management: supervisory or clerical;
- `C2`: skilled manual;
- `DE`: semi-skilled or unskilled.

3.2.3 Method

As in CA, we can declare some variables as active and some other variables as supplementary/illustrative in MCA. We declare the variables corresponding to swear

[5] `http://www.natcorp.ox.ac.uk/docs/catRef.xml`

words and gender as active, and the variables age and social class as illustrative. Running a MCA involves the following steps:

- determining how many dimensions there are to inspect;
- interpreting the MCA graph.

3.2.4 Results

The number of dimensions is rather large and the first two dimensions account for only 42.47% of ϕ^2. To inspect a significant share of ϕ^2, e.g. 80%, we would have to inspect at least 4 dimensions. This issue is common in MCA.

Table 9: Eigenvalues and percentages of variance explained by the dimensions

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.56	22.47	22.47
dim 2	0.50	20.00	42.47
dim 3	0.50	20.00	62.47
dim 4	0.50	20.00	82.47
dim 5	0.44	17.53	100

The eigenvalues can be vizualized by means of a scree plot (Fig. 2). Ideally, we would want to see a sharp decrease after the first few dimensions, and we would want these first few dimensions to account for as much share of ϕ^2 as possible. Here, no sharp decrease is observed.

In the MCA biplot (Fig. 3), each category is the color of its variable. Let us focus first on the first dimension (the horizontal axis) and ignore the second dimension (the vertical axis). Strikingly, the most explicit swear words (f-words) cluster in the rightmost part of the plot. These are used by mostly by men. Female speakers tend to prefer a softer swear word: *bloody*. Next, we focus on the second dimension and ignore the first. Words in the upper part (*gosh* and *shit*) are used primarily by upper-class speakers. F-words, *bloody*, and *damn* are used by lower social categories.

Note that speakers belonging to age groups are positioned close to the intersection of the axes. This is a sign that the first two dimensions bring little or no information about them.

Combining the two dimensions, the plot is divided into four corners in which we observe three distinct clusters:

cluster 1 (upper-right corner) *gosh* and *shit*, used by upper class speakers;
cluster 2 (lower-left corner) *bloody*, used by female middle-class speakers;
cluster 3 (lower-right corner) f-words and *damn*, used by lower-class speakers.

A divide exists between male (`m`) and female (`f`) speakers. However, as the combined eigenvalues indicate, we should be wary of making final conclusions based on the sole inspection of the first two dimensions. The relevance of age groups becomes more relevant if we inspect dimensions 3 and 4 together (Fig. 4).

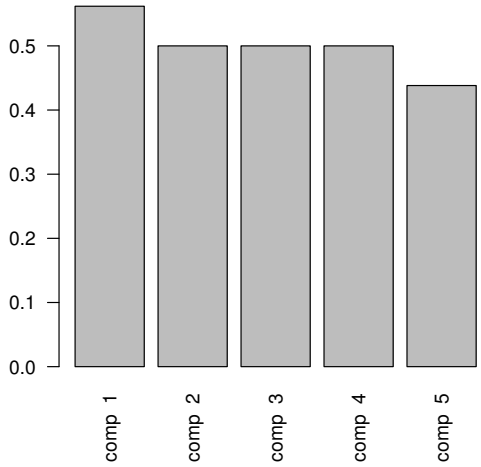

Fig. 2: A scree plot showing the eigenvalues associated with each dimension

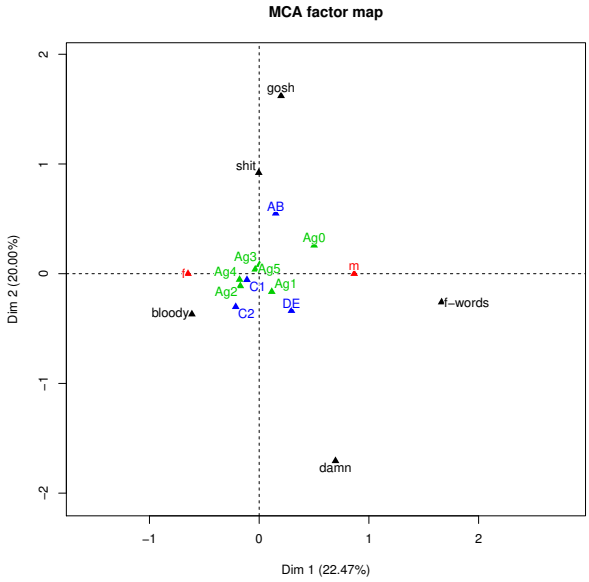

Fig. 3: MCA biplot: a plane representation of individuals and categories

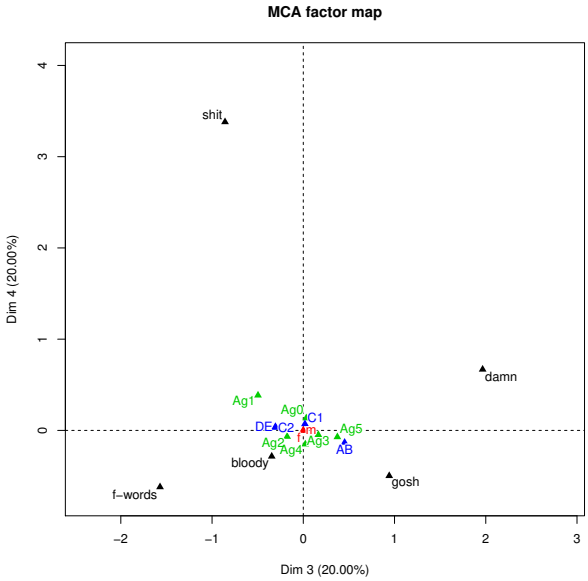

Fig. 4: MCA biplot: a plane representation of individuals and categories (dimensions 3 and 4)

3.3 Principal correspondence analysis

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'. To determine the semantic profile of each preposition, Gréa examines their preferred and dispreferred nominal collocates. He uses an association measure known as *calcul des spécificités* (Habert 1985; Labbé and Labbé 1994; Salem 1987).

3.3.1 Research question

We want to compare the semantic profiles of the prepositions. We answer this question by examining preferred and dispreferred nominal collocates of the prepositions. We want to summarize the table graphically instead of interpreting the data table directly.

3.3.2 Data

Tab. 10 displays a snapshot of the data. The rows contain the nominal collocates and the columns the prepositions. The cells contain the association scores. A posi-

tive score indicates attraction between the NP and the preposition. A negative score indicates repulsion. The assumption is that the semantic profiles of the prepositions will emerge from the patterns of attraction/repulsion.

Table 10: A snapshot of the data

NP	*au centre de*	*au cœur de*	*au milieu de*	*parmi*	*au sein de*
stratégie	17.19	511.62	-59.07	-156.38	-164.09
développement	3	481.19	-34.88	-121	-161.3
forêt	-14.4	453.48	141.63	-227.09	-352.59
enjeux	2.68	450.4	-43.43	-18.15	-288.95
campagne	22.2	389.82	5.44	-166.48	-242.42
vignoble	-9.52	385.49	0	-98.02	-153.59
vie	127.22	380.15	-0.09	-228.69	-256.61
métier	-8.96	376.05	-31.05	-87.76	-83.78
vallée	6.95	369	-3.8	-119.06	-159.48
projet	34.34	367.76	-158.71	-436.78	-1.05
...	...	...	...	...	...

3.3.3 Method

As in CA and MCA, we can declare some variables as active and some other variables as supplementary/illustrative in PCA. Here, however, we decide to declare all variables as active. 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.

3.3.4 Results

The table contains measurements in the same unit. We standardize them to avoid giving each variable a weight proportional to its variance. Perhaps some prepositions attract most nouns more than others.

We run the PCA and see than the first two components are representative enough. As evidenced in Tab. 11 and Fig. 5, the third component is also worth inspecting. We do not do it here for reasons of space.

We plot the graph of variables and the graph of individuals side by side (Fig. 6). Three main profiles appear:

- *au sein de* (upper left corner);
- *au centre de* and *au cœur de* (upper right corner);
- *au milieu de* and *parmi* (lower right corner).

The affinities between *au centre de* and *au cœur de* on the one hand and *au milieu de* and *parmi* on the other are due to similar collocational behaviors. *Au sein de* is

Table 11: Eigenvalues and percentages of variance explained by the components

	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

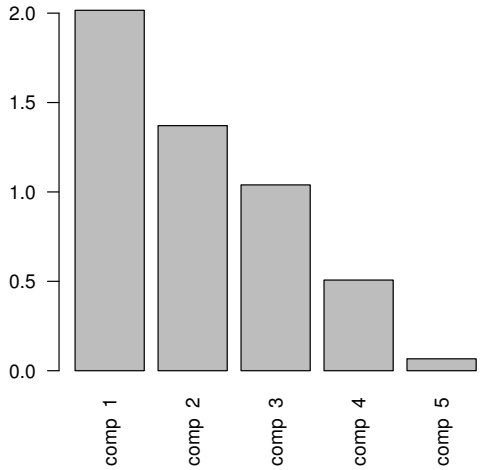

Fig. 5: A scree plot showing the eigenvalues associated with each component

the odd one out. Most NPs clutter around where the two axes intersect, a sign that their distribution is of little interest, at least with respect to our understanding of the prepositions. More interesting are those NPs that appear in the margins of the plot.

Admittedly, the graph of individuals is cluttered. 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. Fig. 7 displays four versions of the plot of individuals of Fig. 6. 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 contribution on the chosen dimensions are plotted;[6]

[6] The contribution is a measure of how much an individual contributes to the construction of a component.

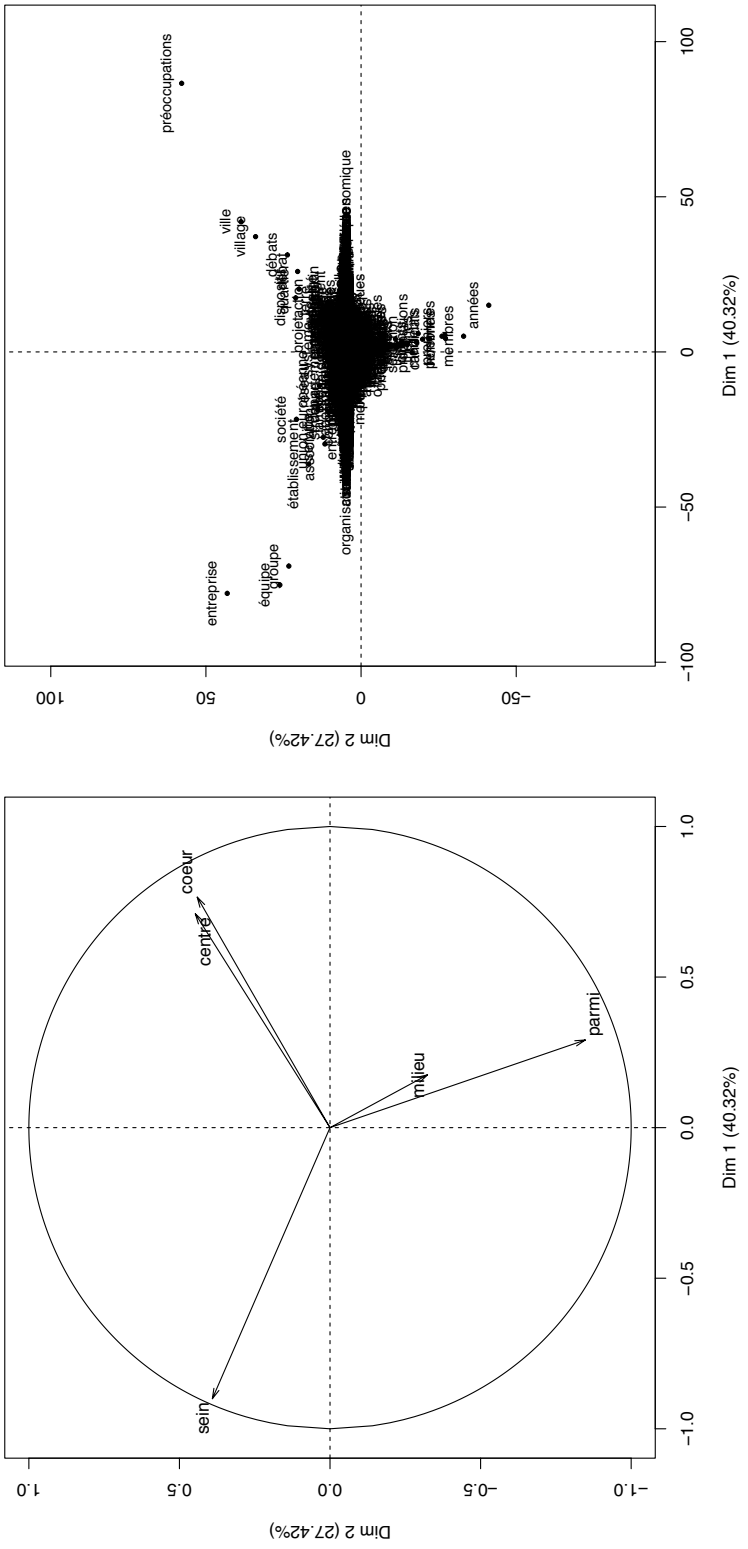

Fig. 6: PCA plots

- with `select="cos2 5"`, only the labels of the five individuals that have the highest squared-cosine score on the chosen dimensions are plotted;[7]
- 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.

In Sect. 4.3, I show how to implement these options with R.

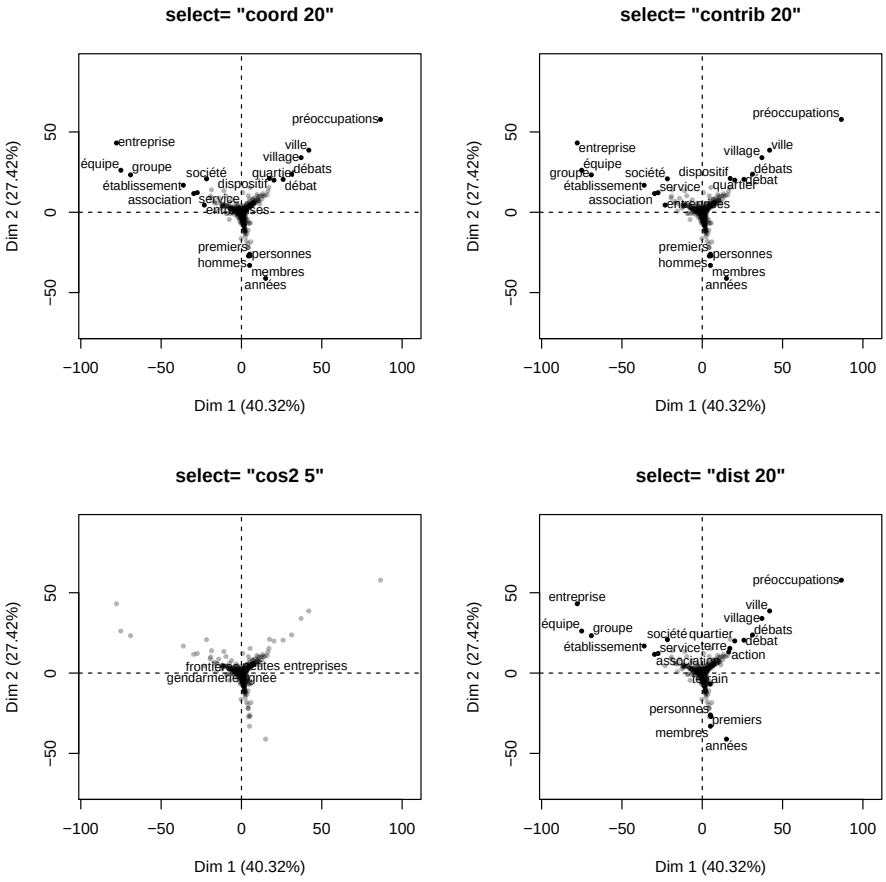

Fig. 7: Selecting NPs with `select`

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.);

[7] The squared cosine (cos^2) is a measure of how well an individual is projected onto a component.

- 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.

The graph displaying the first two components does a good job at grouping prepositions based on the nominal collocates that they have in common and revealing consistent semantic trends. However, it does not show what distinguishes each preposition. For example, *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. In constrast, *au cœur du conflit* 'at the heart of the conflict' denotes the peak of the conflict, either spatially or temporally. This issue has nothing to do with the PCA. It has to do with 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.

3.4 Exploratory factor analysis

The same data set serves as input for EFA, which is performed with `factanal()` (see Sect. 4.4). According to Fig. 5, we should specify 3 factors. Unfortunately, this is not going to work because 3 factors are too many for 5 variables in the kind of EFA that `factanal()` performs. Therefore, we set the number of required factors to 2. A χ^2 test reports whether the specified number of factors is sufficient. If the p-value is smaller than 0.05, more factors are needed. If it is greater than 0.05, no more factors are needed. The test reports that the χ^2 statistic is 12667.73 on 1 degree of freedom and that the p-value is 0. Although a third factor is required, we have no choice but stick to 2 factors. This means that we should be careful when we interpret the results.

The summary reports the uniqueness for the prepositions (Tab. 12). For example, 84.9% of the variance in *au milieu de* is not shared with other prepositions in the overall factor model. In contrast, *parmi* has low variance not accounted for by other variables (0.5%).

Table 12: EFA: uniqueness

centre	*cœur*	*milieu*	*parmi*	*sein*
0.655	0.436	0.849	0.005	0.005

The factor loadings are listed in Tab. 13 (the loadings that are too close to zero are not displayed). Factor loadings are the weights and correlations between the variables and the factors. The higher the loading the more relevant the variable is in

explaining the dimensionality of the factor. If the value is negative, it is because the variable has an inverse impact on the factor. *Au milieu de*, *au centre de*, and *au cœur de* define the first factor. *Parmi* defines the second factor. It seems that *au sein de* defines both.

Table 13: EFAs: loadings

	Factor1	Factor2
centre	0.587	
coeur	0.750	
milieu	0.389	
parmi	-0.147	0.987
sein	-0.740	-0.669

The proportions of variance explained by the factors (i.e. eigenvalues) are listed in Tab. 14. A factor is considered worth keeping if the SS loading (i.e. the sum of squared loadings) is greater than 1. Two factors are retained because both have eigenvalues over 1. Factor 1 accounts for 32.5% of the variance. Factor 2 account for 28.5% of the variance. Both factors account for 66.9% of the variance.

Table 14: EFA: eigenvalues

	Factor1	Factor2
SS loadings	1.626	1.424
Proportion Var	0.325	0.285
Cumulative Var	0.325	0.610

Figure 8 is a plot of the loadings of the prepositions on the two factors with varimax rotation. Figure 9 is the same plot of the loadings with promax rotation.

The distinctive profiles we obtain with EFA are similar to those we obtained with PCA. The only major difference is the proximity of *au milieu de* with *au centre de* and *au cœur de*. This may be due to the fact that only two factors are retained in the analysis. As far as this data set is concerned, PCA is clearly a better alternative, all the more so as individuals are not taken into account in the graphic output of this kind of EFA.

4 Practical guide with R

In this section, I show how to run the code to perform CA, MCA, PCA with `FactoMineR`. The package should therefore be downloaded and installed beforehand.

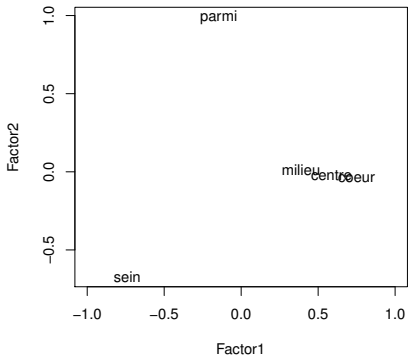

Fig. 8: loadings with varimax rotation

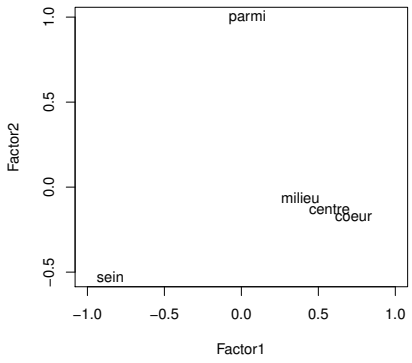

Fig. 9: loadings with promax rotation

```
> install.packages("FactoMineR")
```

EFA is run with `factanal()`, which is part of base R. Therefore, it does not require any extra package.

4.1 Correspondence analysis

The code below was used to run CA on the preposition data set (Sect. 3.1).[8] After clearing R's memory, we load `FactoMineR`, and we import the data file into R `all.preps.rds` (see companion files).[9]

```
> # clear R's memory
> rm(list=ls(all=TRUE))
> # load FactoMineR
> library(FactoMineR)
> # choose all.preps.rds
> dfca <- readRDS(file.choose())
```

The data set has been imported as a data frame.

```
> # inspect the structure of the table
> str(dfca)
'data.frame': 257 obs. of  19 variables:
 $ brown1                   : int  16 1470 207 14 106 246 0 925 523 2 ...
 $ kolh                     : int  8 1249 179 11 148 111 1 873 560 0 ...
 $ lob                      : int  12 1492 199 16 87 234 0 770 507 0 ...
 $ adventure_western_fiction: int  3 258 34 0 0 92 0 137 114 1 ...
 $ belles_lettres           : int  11 627 67 7 45 64 0 382 271 0 ...
 $ general_fiction          : int  3 465 38 4 4 72 0 235 97 0 ...
 $ humour                   : int  1 99 3 0 4 9 0 59 24 0 ...
```

[8] On top of `FactoMineR`, several packages contain a dedicated CA function, e.g. `ca` (Nenadic and Greenacre 2007), and `anacor` (de Leeuw and Mair 2009).

[9] I explain in detail how I extracted the data in this blog post: `https://corpling.hypotheses.org/284`.

```
 $ learned_scientific      : int  1 454 168 8 86 53 0 339 177 0 ...
 $ miscellaneous           : int  0 204 48 7 19 11 0 136 64 0 ...
 $ mystery_detective_fiction: int  2 299 18 3 4 75 1 153 74 1 ...
 $ popular_lore            : int  7 348 50 5 43 59 0 250 146 0 ...
 $ press_editorial         : int  0 188 21 0 17 15 0 116 136 0 ...
 $ press_reportage         : int  3 336 27 2 50 30 0 296 227 0 ...
 $ press_reviews           : int  0 142 14 0 7 4 0 75 49 0 ...
 $ religion                : int  0 120 16 1 27 9 0 65 53 0 ...
 $ romance_love_story      : int  3 350 28 3 3 48 0 145 78 0 ...
 $ science_fiction         : int  1 48 6 1 3 4 0 14 7 0 ...
 $ skills_trades_hobbies   : int  1 273 47 0 29 46 0 166 73 0 ...
 $ prep.length             : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 1 1 1 1 1 ...
> # inspect the first lines of the table
> head(dfca)
            brown1 kolh  lob adventure_western_fiction belles_lettres general_fiction humour
aboard          16    8   12                        3             11               3      1
about         1470 1249 1492                      258            627             465     99
above          207  179  199                       34             67              38      3
absent          14   11   16                        0              7               4      0
according to   106  148   87                        0             45               4      4
across         246  111  234                       92             64              72      9
            learned_scientific miscellaneous mystery_detective_fiction popular_lore
aboard                       1             0                         2            7
about                      454           204                       299          348
above                      168            48                        18           50
absent                       8             7                         3            5
according to                86            19                         4           43
across                      53            11                        75           59
            press_editorial press_reportage press_reviews religion romance_love_story
aboard                    0              3             0        0                   3
about                   188            336           142      120                 350
above                    21             27            14       16                  28
absent                    0              2             0        1                   3
according to             17             50             7       27                  3
across                   15             30             4        9                  48
            science_fiction skills_trades_hobbies prep.length
aboard                    1                     1           1
about                    48                   273           1
above                     6                    47           1
absent                    1                     0           1
according to              3                    29           2
across                    4                    46           1
```

The table consists of 257 lines (one line per preposition type) and 19 columns (one column per variable). Note that the last column (`prep.length`) is loaded as a factor because it contains nominal data (for this reason, it is said to be qualitative).

Columns 4 to 18 are quantitative and declared as supplementary (`col.sup=4:18`). These 15 columns correspond to the 15 text categories. Column 19, which corresponds to the complexity of the preposition, is qualitative and therefore supplementary (`quali.sup=19`). By default, the `CA()` function produces a graph based on the first two dimensions. For the time being, we do not generate these plots yet (`graph=FALSE`). Each graph will be plotted individually later, with specific parameters.

```
> library(FactoMineR)
> ca.object <- CA(dfca, col.sup=4:18, quali.sup=19, graph=F)
```

The output of CA is in `ca.object`. The first lines of the ouput give the χ^2 score and the associated p-value.

```
> ca.object
**Results of the Correspondence Analysis (CA)**
The row variable has  257  categories; the column variable has 3 categories
The chi square of independence between the two variables is equal to 10053.43 (p-value =  0 ).
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$col"            "results for the columns"
3  "$col$coord"      "coord. for the columns"
4  "$col$cos2"       "cos2 for the columns"
5  "$col$contrib"    "contributions of the columns"
6  "$row"            "results for the rows"
7  "$row$coord"      "coord. for the rows"
8  "$row$cos2"       "cos2 for the rows"
```

```
 9  "$row$contrib"     "contributions of the rows"
10 "$col.sup$coord"   "coord. for supplementary columns"
11 "$col.sup$cos2"    "cos2 for supplementary columns"
12 "$quali.sup$coord" "coord. for supplementary categorical var."
13 "$quali.sup$cos2"  "cos2 for supplementary categorical var."
14 "$call"            "summary called parameters"
15 "$call$marge.col"  "weights of the columns"
16 "$call$marge.row"  "weights of the rows"
```

The `eig` object allows to see how many dimensions there are to inspect.

```
> ca.object$eig
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.020398336              82.34156                          82.34156
dim 2 0.004374495              17.65844                         100.00000
```

In case there are more than two dimensions to inspect, a scree plot is useful.

```
> barplot(ca.object$eig[,2], names=paste("dimension", 1:nrow(ca.object$eig)),
+         xlab="dimensions",
+         ylab="percentage of variance")
```

To interpret the dimensions numerically, there are two options. One is to use `dimdesc()`. By default, this function describes the first three dimensions. Because we have only two dimensions to inspect, the `axes` argument of `dimdesc()` is set to `1:2`. Since there is a large number of row variables, the output of `dimdesc()` is too long and not printed here.

```
> dimdesc(ca.object, axes=1:2)
```

Another option is to use `summary()`. It displays:

- the χ^2 score and its associated p-value;
- the eigenvalues;
- the properties of the first 10 row and column variables (active and supplementary) in terms of inertia, coordinate, contribution, and quality of projection along the first two axes.

```
> summary(ca.object)

Call:
CA(X = dfca, col.sup = 4:18, quali.sup = 19, graph = F)

The chi square of independence between the two variables is equal to 10053.43 (p-value =  0 ).

Eigenvalues
                       Dim.1   Dim.2
Variance               0.020   0.004
% of var.             82.342  17.658
Cumulative % of var.  82.342 100.000

Rows (the 10 first)
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
aboard                  |     0.004 | -0.194  0.016  0.812 | -0.093  0.018  0.188 |
about                   |     0.025 | -0.030  0.045  0.368 |  0.039  0.358  0.632 |
above                   |     0.000 | -0.010  0.001  0.307 |  0.015  0.007  0.693 |
absent                  |     0.002 | -0.089  0.004  0.472 |  0.094  0.021  0.528 |
according to            |     0.061 |  0.265  0.289  0.967 | -0.049  0.047  0.033 |
across                  |     0.103 | -0.265  0.502  0.995 |  0.020  0.013  0.005 |
afore                   |     0.005 |  1.488  0.027  0.999 | -0.042  0.000  0.001 |
after                   |     0.037 |  0.062  0.118  0.655 | -0.045  0.290  0.345 |
against                 |     0.032 |  0.090  0.155  0.975 |  0.014  0.019  0.025 |
agin                    |     0.009 | -0.704  0.012  0.276 | -1.141  0.147  0.724 |

Columns
                        Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
brownl                  |     5.653 | -0.101 17.736  0.640 | -0.075 46.518  0.360 |
kolh                    |    14.049 |  0.213 68.862  1.000 | -0.003  0.056  0.000 |
lob                     |     5.071 | -0.091 13.402  0.539 |  0.084 53.426  0.461 |

Supplementary columns (the 10 first)
```

```
                              Dim.1   cos2    Dim.2   cos2
adventure_western_fiction | -0.177  0.131 |  0.009  0.000 |
belles_lettres            |  0.013  0.010 | -0.018  0.020 |
general_fiction           | -0.066  0.032 |  0.030  0.006 |
humour                    | -0.079  0.085 | -0.001  0.000 |
learned_scientific        |  0.096  0.146 |  0.006  0.001 |
miscellaneous             |  0.087  0.060 |  0.001  0.000 |
mystery_detective_fiction | -0.155  0.119 |  0.016  0.001 |
popular_lore              |  0.013  0.013 | -0.011  0.011 |
press_editorial           | -0.024  0.016 | -0.005  0.001 |
press_reportage           |  0.014  0.006 |  0.001  0.000 |

Supplementary categorical variables
                              Dim.1   cos2 v.test    Dim.2   cos2 v.test
prep.length.1             | -0.001  0.984 -5.242 |  0.000  0.016  0.671 |
prep.length.2             |  0.019  0.885  1.482 |  0.007  0.115  0.533 |
prep.length.3             |  0.119  0.904  5.660 | -0.039  0.096 -1.841 |
prep.length.4             |  0.226  0.979  4.701 | -0.033  0.021 -0.688 |
```

Finally, we plot the CA graph with the `plot.CA()` function. We do not plot the rows (`invisible="ind"`), which would clutter the graph with prepositions. The prepositions can be plotted together with the column variables by removing `invisible="ind"`. To prevent the labels from being overplotted, we set `autoLab` to `"yes"`. By setting `shadowtext` to `TRUE`, we create a background shadow that facilitates reading. We adjust the font size of the labels to 80% of their default size (`cex=0.8`). The active column variables are in magenta (`col.col="magenta"`) whereas the supplementary column variables are in Dodger blue (`col.col.sup="dodgerblue"`). Finally, we include a title (`title=`) whose font size is 80% of the default size (`cex.main=.8`).

```
> plot.CA(ca.object,
+         invisible="row",
+         autoLab="yes",
+         shadow=TRUE,
+         cex=.8,
+         col.col="magenta",
+         col.col.sup="dodgerblue",
+ title="Distribution of prepositions based on lexical complexity
+  in three corpora:\n LOB (British English), Brown (US English),
+  and Kolhapur (Indian English)",
+ cex.main=.8)
```

4.2 Multiple correspondence analysis

The data file for this case study is `swear.words.bnc.txt` (see companion files).[10] We load the data.

```
> # clear R's memory
> rm(list=ls(all=TRUE))
>
> #load FactoMineR
> library(FactoMineR)
>
> # choose swear.words.bnc.txt
> df <- read.table(file=file.choose(), header=TRUE, sep="\t")
```

As we inspect the structure of the data frame with `str()`, we see that the number of levels is going to be a bit high with respect to `word` (8).

[10] The code for the extraction was partly contributed by Mathilde Léger, a third-year student at Paris 8 University, as part of her end-of-term project.

```
> str(df)
'data.frame': 293289 obs. of  4 variables:
 $ word     : Factor w/ 8 levels "bloody","damn",..: 2 2 7 7 7 2 7 2 7 7 ...
 $ gender   : Factor w/ 2 levels "f","m": 2 2 2 2 2 2 2 2 2 2 ...
 $ age      : Factor w/ 6 levels "Ag0","Ag1","Ag2",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ soc_class: Factor w/ 4 levels "AB","C1","C2",..: 1 1 1 1 1 1 1 1 1 1 ...
> table(df$word)

 bloody    damn    fuck  fucked  fucker fucking    gosh    shit
 146203   32294    9219      11     467   23487   60678   20930
```

We group *fuck*, *fucking*, *fucked*, and *fucker* into a single factor: *f-words*. With `gsub()`, we replace each word with the single tag `f-words`.

```
> df$word <- gsub("fuck|fucking|fucker|fucked", "f-words", df$word, ignore.case=TRUE)
> table(df$word)

 bloody    damn f-words    gosh    shit
 146203   32294   33184   60678   20930
```

We convert `df$word` back to a factor.

```
> df$word <- as.factor(df$word)
```

We run the MCA with the `MCA()` function. We declare `age` and `soc_class` as supplementary (`quali.sup=c(3,4)`). We do not plot the graph yet (`graph=FALSE`).

```
> mca.object <- MCA(df, quali.sup=c(3,4), graph=FALSE)
> mca.object
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 293289 individuals, described by 4 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. of the categories"
4  "$var$cos2"          "cos2 for the categories"
5  "$var$contrib"       "contributions of the categories"
6  "$var$v.test"        "v-test for the categories"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$quali.sup"         "results for the supplementary categorical variables"
12 "$quali.sup$coord"   "coord. for the supplementary categories"
13 "$quali.sup$cos2"    "cos2 for the supplementary categories"
14 "$quali.sup$v.test"  "v-test for the supplementary categories"
15 "$call"              "intermediate results"
16 "$call$marge.col"    "weights of columns"
17 "$call$marge.li"     "weights of rows"
```

Again, the `eig` object allows us to see how many dimensions there are to inspect.

```
> round(mca.object$eig, 2)
       eigenvalue percentage of variance cumulative percentage of variance
dim 1       0.56                 22.47                             22.47
dim 2       0.50                 20.00                             42.47
dim 3       0.50                 20.00                             62.47
dim 4       0.50                 20.00                             82.47
dim 5       0.44                 17.53                            100.00
```

The corresponding scree plot is obtained as follows.

```
> barplot(mca.object$eig[,1],
+         names.arg=paste("comp ", 1:nrow(mca.object$eig)), las=2)
```

The descriptors for the first two dimensions are obtained with `dimdesc()`.

```
> dimdesc(mca.object, axes = 1:2)
$`Dim 1`
$`Dim 1`$quali
                  R2 p.value
word      0.56180343       0
gender    0.56180343       0
```

```
age       0.04527337       0
soc_class 0.03599473       0

$`Dim 1`$category
           Estimate       p.value
DE       0.19797553  0.000000e+00
AB       0.09055594  0.000000e+00
Ag0      0.34641567  0.000000e+00
m        0.56768214  0.000000e+00
f-words  0.95542445  0.000000e+00
damn     0.23038750  0.000000e+00
Ag1      0.05696397 1.754549e-123
Ag5     -0.05588133  2.331308e-16
C2      -0.18267652  0.000000e+00
C1      -0.10585495  0.000000e+00
Ag4     -0.16303530  0.000000e+00
Ag2     -0.15770030  0.000000e+00
f       -0.56768214  0.000000e+00
gosh    -0.14135352  0.000000e+00
bloody  -0.75055721  0.000000e+00


$`Dim 2`
$`Dim 2`$quali
                  R2 p.value
word      1.00000000       0
age       0.01620229       0
soc_class 0.12601448       0

$`Dim 2`$category
           Estimate       p.value
AB       0.41473018  0.000000e+00
Ag0      0.17686269  0.000000e+00
shit     0.62142079  0.000000e+00
gosh     1.11696670  0.000000e+00
Ag3      0.05207258 4.285106e-105
Ag5      0.02195358  1.406829e-19
Ag4     -0.04385357  2.582515e-51
C1      -0.01381071  1.601818e-87
Ag2     -0.08493256 2.040075e-192
Ag1     -0.12210271 4.146754e-251
DE      -0.21303043  0.000000e+00
C2      -0.18788904  0.000000e+00
f-words -0.21265143  0.000000e+00
damn    -1.23558964  0.000000e+00
bloody  -0.29014642  0.000000e+00
```

We plot the MCA map with the `plot.MCA()` function. Each category is the color of its variable (`habillage="quali"`). We remove the title (`title=""`).

```
> plot.MCA(mca.object,
+          invisible="ind",
+          autoLab="yes",
+          shadowtext=TRUE,
+          habillage="quali",
+          title="")
```

To plot dimensions 3 and 4, we add the argument `axes=c(3,4)` in the `plot.MCA()` call.

```
> plot.MCA(mca.object,
+          axes=c(3,4),
+          invisible="ind",
+          autoLab="yes",
+          shadowtext=TRUE,
+          habillage="quali",
+          title="")
```

4.3 Principal correspondence analysis

Several packages and functions implement PCA in R : e.g. `princomp()` and `prcomp()` from the `stats` package, `ggbiplot()` from the `ggbiplot` pack-

age (which is itself based on `ggplot2`), `dudi.pca()` from the `ade4` package, and `PCA()` from the `FactoMineR` package. Mind you, `princomp()` and `prcomp()` perform PCA based on loadings.

First, we load the data set (`inclusion.txt`).

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

As we inspect the data frame with `str()`, we see that 22,397 NPs were found.

```
> str(data)
'data.frame': 22397 obs. of  5 variables:
 $ centre: num  -281 -274 -219 -128 -114 ...
 $ coeur : num  -651 -913 -933 -368 -330 ...
 $ milieu: num  -545 -432 -270 -237 -173 ...
 $ parmi : num  -1685 -1129 -1072 -678 -499 ...
 $ sein  : num  4226 3830 3490 1913 1522 ...
```

Next, load the `FactoMineR` package and run the PCA with the `PCA()` function. The variables are standardized by default.

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

We make sure than the first two components are representative enough.

```
> 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
```

We plot these eigenvalues (Fig. 5).

```
> barplot(pca.object$eig[,1],
+         names.arg=paste("comp ",1:nrow(pca.object$eig)), las=2)
```

We plot the graph of variables and the graph of individuals side by side.

```
> # tell R to display the two plots side by side
> par(mfrow=c(1,2))
> # graph of variables
> plot.PCA(pca.object, choix="var", title="")
> # graph of individuals
> plot.PCA(pca.object, cex=0.8, autoLab="auto", shadowtext = FALSE, title="")
```

The `select` argument of the `PCA()` function allows the user to filter out unwanted individuals by selecting only the desired ones (see Fig. 7 in Sect. 3.3.4).

```
> 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")
```

4.4 Exploratory factor analysis

In base R, we run EFA with `factanal()`.[11]

[11] The `FactoMineR` package includes several extensions of factor analysis. Multiple factor analysis (MFA) is used to explore datasets where variables are structured into groups. Like PCA, it can

We set the `factors` argument to 2. By default, the varimax rotation applies. The output displays uniqueness, loadings, the proportions of variance explained by the factors, and the χ^2 test.

```
> fa.object <- factanal(data, factors=2)
> fa.object

Call:
factanal(x = data, factors = 2)

Uniquenesses:
centre  coeur milieu  parmi   sein
 0.655  0.436  0.849  0.005  0.005

Loadings:
        Factor1 Factor2
centre  0.587
coeur   0.750
milieu  0.389
parmi  -0.147   0.987
sein   -0.740  -0.669

               Factor1 Factor2
SS loadings      1.626   1.424
Proportion Var   0.325   0.285
Cumulative Var   0.325   0.610

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 12667.73 on 1 degree of freedom.
The p-value is 0
```

The code below is used to plot the loadings of the prepositions on the two factors with varimax rotation (Fig. 8).

```
> loadings <- loadings(fa.object)
> plot(loadings, type="n", xlim=c(-1,1))
> text(loadings, rownames(loadings))
```

To produce a plot with promax rotation (Fig. 9), we run `factanal()` again but set `rotation` to `promax`.

```
> fa.object2 <- factanal(data, factors=2, rotation="promax")
> loadings2 <- loadings(fa.object2)
> plot(loadings2, type="n", xlim=c(-1,1))
> text(loadings2, rownames(loadings2))
```

5 Key readings

More details about the inner workings of CA can be found in Greenacre (2007), Nenadic and Greenacre (2007), Husson, Lê, and Pagès (2010, Chap. 2), Glynn (2014), and Desagulier (2017, Sect. 10.4). For specific applications of CA in corpus linguistics, see Desagulier (2014, 2015b). For more details on MCA, see Greenacre and Blasius (2006), Husson, Lê, and Pagès (2010, Chap. 3), Le Roux (2010), and Desagulier (2017, Sect. 10.5). For a specific application of MCA in corpus linguistics, see Desagulier (2015b). For more details on PCA, see Husson, Lê, and Pagès (2010, Chap. 1) and Desagulier (2017, Sect. 10.2). For PCA based on loadings, see

handle continuous and/or categorical variables simultaneously (Pagès 2014). MFA further breaks down into hierarchical multiple factor analysis (Le Dien and Pagès 2003) and dual multiple factor analysis (Lê and Pagès 2010). Although commonly used in sensorimetrics, these methods are rare in linguistics.

Baayen (2008, Sect. 5.1.1). For a specific application of PCA in corpus linguistics, see Desagulier (2015a). To know more about how EFA works, see Baayen (2008, Sect. 5.1.2). For sociolinguistic applications, see Biber (1991, 1995).

There are, of course, far more exploratory techniques than those I have presented here.[12] Very similar to CA is multidimensional scaling (MDS), which can be run with the `cmdscale()` function in R.[13] Like CA, MDS hinges on Euclidean representations. Unlike CA, the main input of MDS is not a contingency table, but a similarity table.[14]. Linguistic applications of MDS include Croft and Poole (2008), Szmrecsanyi (2010), and Hilpert (2013). Baayen (2008, 5.1.4) explains how to run MDS in R.

Finally, (hierarchical) configural frequency analysis (CFA) explores contingency tables (or tables of categorical data converted into contingency tables) to detect two kinds of patterns in the data: types and antitypes (Lienert 1968; Von Eye 2003; Gries 2009). It does so by creating log-linear combinations of factors to predict cell frequencies. If, for a given linguistic phenomenon, the observed frequency is greater than the expected frequency, the configuration is a type. If, for the same linguistic phenomenon, the expected frequency is greater than the observed frequency, the configuration is an antitype. Note than CFA can be exploratory or explanatory. You can run CFA by using the `cfa()` function from the `lavaan` package.

References

Baayen, R Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Benzécri, Jean-Paul. 1984. *Analyse des correspondances, exposé élémentaire*. Vol. 1. Pratique de l'analyse des données. Paris: Dunod.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

– . 1991. *Variation across speech and writing*. Cambridge University Press.

Cadoret, Marine, Sébastien Lê, and Jérôme Pagès. 2011. "Multidimensional Scaling Versus Multiple Correspondence Analysis When Analyzing Categorization Data." In *Classification and Multivariate Analysis for Complex Data Structures*, ed. by Bernard Fichet et al., 301–308. Berlin, Heidelberg: Springer Berlin Heidelberg.

Croft, William, and Keith T Poole. 2008. "Inferring universals from grammatical variation: Multidimensional scaling for typological analysis." *Theoretical linguistics* 34 (1): 1–37.

[12] The inventory I have proposed is influenced by the French school of data analysis founded by Jean-Paul Benzécri.

[13] See Le Roux and Rouanet (2004, 2.4.3) for a comparison between MDS and CA, and Cadoret, Lê, and Pagès (2011) for a comparison between MDS and MCA.

[14] However, MDS can also take as input a contingency table.

de Leeuw, Jan, and Patrick Mair. 2009. "Simple and Canonical Correspondence Analysis Using the R Package anacor." *Journal of Statistical Software* 31 (5): 1–18. http://www.jstatsoft.org/v31/i05/.

Desagulier, Guillaume. 2015a. "A lesson from associative learning: asymmetry and productivity in multiple-slot constructions." *Corpus Linguistics and Linguistic Theory*. doi:10.1515/cllt-2015-0012.

– . 2017. *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Quantitative Methods in the Humanities and Social Sciences. New York: Springer.

– . 2015b. "Forms and meanings of intensification: a multifactorial comparison of *quite* and *rather*." *Anglophonia* 20 (2). doi:10.4000/anglophonia.558. http://anglophonia.revues.org/558.

– . 2014. "Visualizing distances in a set of near synonyms: *rather*, *quite*, *fairly*, and *pretty*." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, ed. by Dylan Glynn and Justyna Robinson, 145–178. Amsterdam: John Benjamins.

Francis, W. Nelson, and Henry Kučera. 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University.

Glynn, Dylan. 2014. "Correspondence analysis: Exploring data and identifying patterns." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, ed. by Dylan Glynn and Justyna Robinson, 443–485. Amsterdam: John Benjamins.

Gréa, Philippe. 2017. "Inside in French." *Cognitive Linguistics* 28 (1): 77–130.

Greenacre, Michael J. 2007. *Correspondence Analysis in Practice*. Vol. 2. Interdisciplinary statistics series. Boca Raton: Chapman Hall/CRC.

Greenacre, Michael J., and Jorg Blasius. 2006. *Multiple correspondence analysis and related methods*. Boca Raton: Chapman Hall/CRC.

Gries, Stefan Th. 2009. *Statistics for linguistics with R: A practical introduction*. 1st edition. Berlin: De Gruyter Mouton.

Habert, Benoît. 1985. "L'analyse des formes «spécifiques» [bilan critique et propositions d'utilisation]." *Mots* 11 (1): 127–154.

Hilpert, Martin. 2013. *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge ; New York: Cambridge University Press.

Hirschfeld, Hermann O. 1935. "A connection between correlation and contingency." In *Mathematical Proceedings of the Cambridge Philosophical Society*, 31:520–524. 4. Cambridge University Press.

Hirschmuller, Helmut. 1989. "The use of complex prepositions in Indian English in comparison with British and American English." In *Englische Textlinguistik und Varietätenforschung*, ed. by Gottfried Graustein and Wolfgang Thiele, 69:52–58. Linguistische Arbeitsberichte. Leipzig: Karl Marx Universität.

Hofland, Knut, and Stig Johansson. 1982. *Word frequencies in british and american english*. Norwegian computing centre for the Humanities.

Husson, François, Sébastien Lê, and Jérôme Pagès. 2010. *Exploratory Multivariate Analysis by Example Using R*. London: CRC press.

Husson, Francois, et al. 2015. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*. R package version 1.31.4. `http://CRAN.R-project.org/package=FactoMineR`.

Kassambara, Alboukadel, and Fabian Mundt. 2017. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5. `https://CRAN.R-project.org/package=factoextra`.

Kilgarriff, Adam. 2005. "Language is never, ever, ever, random." *Corpus linguistics and linguistic theory* 1 (2): 263–276.

Labbé, Cyril, and Dominique Labbé. 1994. "Que mesure la spécificité du vocabulaire ?" *Lexicometrica* 3:2001.

Lacheret-Dujour, Anne, et al. to appear. "The distribution of prosodic features in the Rhapsodie corpus." Chap. 17 in *Rhapsodie: A prosodic and syntactic treebank for spoken French*, ed. by Anne Lacheret-Dujour and Sylvain Kahane, 315–338. Studies in Corpus Linguistics 89. John Benjamins. `https://benjamins.com/catalog/scl.89.18lac`.

Lê, Sébastien, and Jérome Pagès. 2010. "Dmfa: Dual multiple factor analysis." *Communications in Statistics—Theory and Methods* 39 (3): 483–492.

Le Dien, S, and J Pagès. 2003. "Hierarchical multiple factor analysis: Application to the comparison of sensory profiles." *Food quality and preference* 14 (5-6): 397–403.

Le Roux, Brigitte. 2010. *Multiple correspondence analysis*. London: SAGE.

Le Roux, Brigitte, and Henry Rouanet. 2004. *Geometric data analysis: from correspondence analysis to structured data analysis*. Dordrecht: Kluwer Academic Publishers.

Leech, Geoffrey, and Roger Fallon. 1992. "Computer corpora–what do they tell us about culture." *ICAME journal* 16.

Leech, Geoffrey, Stieg Johansson, and Knut Hofland. 1978. *The LOB Corpus, original version (1970–1978)*. Lancaster, Oslo, Bergen.

Leech, Geoffrey, et al. 1986. *The LOB Corpus, POS-tagged version (1981–1986)*. Lancaster, Oslo, Bergen.

Leitner, Gerhard. 1991. "The Kolhapur Corpus of Indian English: Intra-varietal description and/or intervarietal comparison." In *English Computer Corpora*, ed. by Stieg Johansson and Anna-Brita Stenström, 215–232. Topics in English Linguistics. Berlin: Mouton de Gruyter.

Lienert, Gustav Adolf. 1968. "Die "Konfigurationsfrequenzanalyse" als Klassifikationsmethode in der klinischen Psychologie." *Bericht über den* 26:244–253.

Nenadic, Oleg, and Michael J. Greenacre. 2007. "Correspondence Analysis in R, with two- and three-dimensional graphics: The `ca` package." *Journal of Statistical Software* 20 (3): 1–13. `http://www.jstatsoft.org`.

Pagès, Jérôme. 2014. *Multiple factor analysis by example using R*. Boca Raton: Chapman / Hall/CRC.

Rayson, Paul, Geoffrey N Leech, and Mary Hodges. 1997. "Social differentiation in the use of English vocabulary: some analyses of the conversational component of

the British National Corpus." *International Journal of Corpus Linguistics* 2 (1): 133–152.

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

Schmid, Hans Jörg. 2003. "Do men and women really live in different cultures? Evidence from the BNC." In *Corpus Linguistics by the Lune*, ed. by Andrew Wilson, Paul Rayson, and Tony McEnery, 185–221. Lódź Studies in Language. Frankfurt: Peter Lang.

Shastri, S. V., C. T. Patilkulkarni, and Geeta S. Shastri. 1986. *The Kolhapur Corpus*. Kolhapur, India.

Szmrecsanyi, Benedikt. 2010. "The English genitive alternation in a cognitive sociolinguistics perspective." *Advances in cognitive sociolinguistics*: 141–166.

Von Eye, Alexander. 2003. *Configural frequency analysis: Methods, models, and applications*. Mahwah, NJ: Lawrence Erlbaum.