Multivariate Exploratory Approaches

Guillaume Desagulier

Chapter 19
Multivariate Exploratory Approaches

Guillaume Desagulier

Abstract This chapter provides both a theoretical discussion of what multivariate exploratory approaches entail and step-by-step instructions to implement each of them with R. Four methods are presented: correspondence analysis, multiple correspondence analysis, principal component analysis, and exploratory factor analysis. These methods are designed to explore and summarize large and complex data tables by means of summary statistics. They help generate hypotheses by providing informative clusters using the variable values that characterize each observation.

19.1 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.

I present four multivariate exploratory techniques: correspondence analysis (henceforth CA), multiple correspondence analysis (henceforth MCA), principal compo-

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".

nent analysis (henceforth PCA), and exploratory factor analysis (henceforth EFA). These techniques rely on dimensionality reduction, i.e. an attempt to simplify complex multivariate datasets to facilitate interpretation.

19.2 Fundamentals

CA, MCA, PCA, and EFA 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.

19.2.1 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 proportions are eigenvalues in CA, MCA, and PCA. They are loadings in EFA (and another kind of PCA that is 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 numeric results.

19.2.2 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 cross-classifies observations on a number of categorical variables (see Chap. 20). Entries in each cell are integers, namely the number of times that observations (in the rows)

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

are seen in the context of the variables (in the columns). Table 19.1 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 19.1 An example of a contingency table (Desagulier 2017, p. 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 a case-by-variable table such as Table 19.2. 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.

Table 19.2 A sample input table for MCA (Desagulier 2017, p. 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*

PCA takes as input a table of data of i individuals or observations (rows) and j variables (columns). The method 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 19.3 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, Table 19.3 minus the three columns of nominal data can serve as input for EFA.

Table 19.3 A sample data frame (Lacheret-Dujour, Desagulier, Fleury, and Isel 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
…	…	…	…	…	…	…	…	…	…

19.2.3 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. 19.3.2). 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. Expanding on Gries (2006), Glynn (2014) finds that usage features and dictionary senses are correlated with dialect and register thanks to two multivariate exploratory techniques (correspondence analysis and multiple correspondence analysis). To confirm these findings, Glynn (2014) turns to logistic regression. This confirmatory multivariate technique allows to specify which of the usage features and dictionary senses are significantly associated with either dialect or register, and determine the importance of the associations.

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

19.2.4 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. Incidentally, CA 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.

It should be remembered that the linguist makes no assumption as to what kinds of groupings are to be found in the data. In practice, however, a table of data is compiled because meaningful groupings are expected to be found. Therefore, if no meaningful grouping is found, this is because the rows and the columns are independent. In this case, it is advisable to rethink the design of the study, especially the 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 distributions of row variables and column variables are independent;
H_1: the distributions of row variables and 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:

$$\chi^2_{i,j} = \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{19.1}$$

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

$$\chi^2 = \sum_{i=1}^{n} \frac{(O-E)^2}{E} \tag{19.2}$$

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:

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{N(k-1)}}. \tag{19.3}$$

Central to CA is the concept of profile. To obtain the profile of a row, each cell is divided by its row total. Table 19.4 displays the row profiles of Table 19.1. The row profiles add up to 1. Likewise, one obtains the profile of a column by dividing each column frequency by the column total (Table 19.5). Again, the column profiles add up to 1.

CA performs an analysis of rows and columns that is both simultaneous and symmetric. A column analysis consists in interpreting the column profiles using the rows as reference points on the basis of a table such as Table 19.5. For example, the value in the A1K.xml column for singular common nouns (NN1) is 0.6189. Comparing this value with the average proportion of NN1 in the sample (0.5594),

Table 19.4 The row profiles of Table 19.1

	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 19.5 The column profiles of Table 19.1

	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

it appears that these noun types are slightly over-represented in A1K.xml by a ratio of $\frac{0.6189}{0.5594} \approx 1.1063$. A row analysis consists in interpreting the row profiles using the columns as reference points on the basis of a table such as Table 19.4, in which the same cell displays a value of 0.1241. In other words, of all the singular common nouns that occur in the corpus files, 12.41% occur in A1K.xml. On average, A1K.xml contains 11.22% of the nouns found in the sample. The ratio is the same as above, i.e. $\frac{0.1241}{0.1122} \approx 1.1063$.

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 large a proportion of inertia as possible. These proportions of inertia are known as eigenvalues.

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 generally redundant. Its main function is to help interpret the results by providing relevant groupings. Whether a group of individuals or variables should be declared as active/illustrative depends on what the linguist considers are primary or secondary in the exploration of the phenomenon under study.

19.2.5 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. 19.2.2, MCA takes as input a table of nominal data. For this method 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 summary plot).

19.2.6 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.

As exemplified in this chapter, PCA is based on the inspection of correlations between the variables and the principal components.[3] Before one runs a PCA, one should consider standardizing (i.e. centering and scaling) the variables (see Chap. 17). 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 giv-

[3] 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.

ing 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, p. 45).

19.2.7 Exploratory factor analysis

EFA was made popular in linguistics by Biber's studies on register variation as part of the multidimensional (MD) approach (Biber 1988; Biber 1995). The goal of the MD approach is to detect register differences across the texts and text varieties of a corpus based on groups of linguistic features that co-occur significantly. Technically, this approach starts with a large number of linguistic variables and relies on factor analysis to reduce this number to a few basic functional dimensions that account for differences between texts. MD analysis is featured in a vast number of synchronic and diachronic studies on various discourse domains such as eighteenth century English (Biber 2001), blogs (Grieve, Biber, Friginal, and Nekrasova 2010), academic English (Biber and Gray 2016), etc. It has been applied to languages other than English such as Somali (Biber and Hared 1992) or Korean (Kim and Biber 1994). For an overview, see Biber and Conrad (2001).

Although close to PCA, EFA differs with respect to the following. The number of relevant components, which are called factors, is not determined automatically. It must be chosen beforehand. EFA is designed to identify patterns of joint variation in a number of observed variables. It looks for variables that are highly correlated with a group of other variables. These intercorrelated variables are assumed to measure one underlying variable. This variable, which is not directly observed, but inferred, is latent. It is known as a factor. This is an aspect that PCA is not designed to show. 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, p. 127).[4]

Representative study 1

D. Glynn (2014). "The many uses of *run*." In: *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Ed. by D. Glynn and J. A. Robinson. Vol. 43. Human Cognitive Processing. John Benjamins, pp. 117–144

[4] 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.

Research questions

Glynn (2014) examines the semasiological variation of *run* in the light of sociolinguistic variables. The study posits that "even for a lexeme as culturally 'simple' and as socially 'neutral' as *run*, one must account for the social dimension of language in semantic analysis" (Glynn 2014, p. 124).

Data

Glynn's study is based on 500 occurrences of *run* in British and American English (250 occurrences for each variety). The occurrences break down into conversation and online personal diaries. The diary examples were extracted from the LiveJournal corpus, developed by Dirk Speelman (University of Leuven). The conversation examples were extracted from the British National Corpus and the American National Corpus.

Method

Each entry was annotated for dictionary sense, register, and dialect. The data were submitted to correspondence analysis.

Results

The first two dimensions of CA account for 87% of ϕ^2, which means that the conclusions based upon their inspection only are reliable. In American conversation, *run* tends to mean 'increase', 'diffuse', and 'motion into difficulty'. In the American diary genre, *run* is characterized by the following dictionary senses: 'campaign', 'copy' and, to some extent, 'metaphoric motion'. Although specific to American English, 'meet' and 'extend space' is used in either register. In British English, *run* is highly and distinctly associated with'flow' and 'extend time'. A relative association with British English is also found with senses such as 'use up', 'cause motion' and 'escape'. To further explore the detail of the sociolinguistic variation at work with *run*, Glynn resorts to multiple correspondence analysis.

Representative study 2

G. Desagulier (2015). "A lesson from associative learning: asymmetry and productivity in multiple-slot constructions." In: *Corpus Linguistics and Linguistic Theory*. DOI: 10.1515/cllt-2015-0012

Research questions

This paper addresses a claim made by Kay (2013) that only fully productive constructions should count as constructions. Desagulier (2015) posits that even patterns that are not fully productive often have subregularities that are. He shows that the *A as NP* construction (*stiff as a board*, *cool as a cucumber*, *flat as a pancake*), which Kay had argued was simply idiomatic, licenses productive coinages when used with particular adjectives or nouns (e.g. *black as NP*, *A as hell*).

Data

All the occurrences of *A as NP* were extracted from the BNC-XML, amounting to 1,819 tokens. Only instances of *A as NP* where the adjective is intensified were kept. Examples involving a literal comparison and no intensification were discarded. Each adjective and noun appearing in *A as NP* was assigned a range of mean scores based on the following measures: an asymmetric association measure (ΔP), a symmetric association measure (collostruction strength indexed on the log-likelihood statistic), type frequency (V), the frequency of hapax legomena ($V1$), potential productivity ($\mathscr{P}$), and global productivity (P^*).

Method

The individuals consist of all adjective and NP types of *A as NP* tokens. Each of the 1,278 individuals (402 adjective types and 876 NP types) is examined in the light of four active variables: collostruction strength, the difference $\Delta P_{NP|A} - \Delta P_{A|NP}$, $\mathscr{P}$, and P^*. Three supplementary quantitative variables were also included to verify that no counterintuitive result was obtained with respect to the computation of hapax-based measures: V, $V1$, and construction frequency. The data table was submitted to PCA.

Results

Three clusters stand out. Globally productive individuals and those that belong to highly associated pairs (i.e. characterized by high collostruction strength and low ΔP) cluster along the horizontal axis (first principal component). The former appear in the upper-right corner of the plot of individuals whereas the latter cluster in the lower-right corner. Individuals that are productive according to $\mathscr{P}$ cluster along the vertical axis (second principal component). In other words, individuals that belong to highly associated pairs are among the least potentially productive and the most globally productive.

Individuals with extreme values for the first component are mostly nouns (*day*, *night*, *snow*, *sheet*, etc.). Most nouns with the highest P^* values denote paragons whose semantic relation with the adjective can be easily accessed despite its conventional nature, e.g. day is bright in *bright as day*, sheets are white in *white as a sheet*. Most nouns with the highest collostruction strength denote paragons whose semantic relation with the adjective is less obvious. These lexemes belong to highly conventionalized expressions (*bold as brass*, *safe as houses*, etc.). Globally productive individuals are more likely to be used in new *A as NP* formations than individuals belonging to strongly associated pairs.

The most productive individuals according to $\mathscr{P}$ belong to weakly associated pairs. The most productive subschemas are indexed on adjectives. These adjectives denote basic properties such as colors and shades (*black*, *white*, *red*, *clear*, *bright*, *pale*), texture and constitution (*big*, *sharp*, *strong*, *thick*, *stiff*, *light*), and temperature (*cold*). There are fewer productive subschemas indexed on nouns. With respect to the most productive subschema, *A as hell*, the NP has lost its literal meaning to the benefit of an exclusively intensifying function.

As we move down from the upper-left to the bottom-right part of the plot, productivity declines and conventionalization and autonomy increase. In this study, PCA helps spot distinct loci of constructional productivity at subschematic levels. In other words, productivity is by no means an all or nothing affair.

Representative study 3

D. Biber (1988). *Variation across Speech and Writing*. Cambridge University Press

Research questions

The aim of this work is to spot the patterns of linguistic variation among registers in a corpus of English texts. This landmark study implements an intuition formerly formulated by sociolinguists according to which linguistic features that co-occur significantly can discriminate among registers.

Data

Biber combines the London-Lund and the Lancaster-Oslo-Bergen corpora to obtain a large and varied corpus that contains a wide variety of spoken and written texts (Biber 1988, Appendix I).

Method

Sixty-seven linguistic features are included in the analysis (Biber 1988, pp. 73–75). These are grouped into sixteen classes: (a) tense and aspect markers, (b) place and time adverbials, (c) pronouns and pro-verbs, (d) questions, (e) nominal forms, (f) passives, (g) stative forms, (h) subordination features, (i) prepositional phrases adjectives and adverbs, (j) lexical specificity, (k) lexical classes, (l) modals, (m) specialized verb classes, (n) reduced forms and dispreferred structures, (o) coordination, and (p) negation.

First, the corpus is tagged for linguistic features. Next, the frequency counts of all linguistic features are extracted, normalized, and standardized. This guarantees a fair comparison of frequency distributions across texts of unequal lengths. Then, factor analysis is used to identify the dimensions, where each dimension captures a pattern of underlying co-occurrence patterns among linguistic features. A factor loading indicates the extent to which a given feature is representative of the dimension underlying a factor. Dimension scores are calculated for each text sample by adding up standardized frequencies with salient positive loadings and subtracting salient negative loadings on a dimension. Finally, each dimension is interpreted in functional terms. This correspondence between dimensions and functions is facilitated by promax rotation.

Results

Because dimensions have a functional basis, each of them is associated with a distinctive pattern of register variation and assigned an interpretive label. In Biber (1988), five major dimensions are found:

1. involved versus informational production;
2. narrative discourse;
3. situation-dependent versus elaborated reference;
4. overt expression of argumentation;
5. impersonal/asbtract style.

Each dimension is captured by a distinction between positive and negative features. For example, the positive features of the first dimension are: verbs, pronouns, adverbs, dependent clauses, and other (contractions, discourse particles, clause coordination, etc.). The negative features of the same dimension are: nouns, long words, prepositional phrases, attributive adjectives, and lexical diversity. Subsequent studies have confirmed that the underlying dimensions of variation and the relations among registers display similar configurations across languages.

19.3 Practical guide with R

In this section, I show how to run the code to perform CA, MCA, and PCA with `FactoMineR`. The package should therefore be downloaded and installed beforehand. EFA is run with `factanal()`, which is part of base R. Therefore, it does not require any extra package.

19.3.1 Correspondence analysis

Leitner (1991) reports a study by Hirschmüller (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, Johansson, Garside, and Hofland 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 preposi-

tions consist of two words and 154 of three and more, out of a total of 235 complex prepositions. 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. Leitner (1991, p. 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 Leitner (1991), 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.

With the code below, we run CA on the preposition data set.[5] After clearing R's memory, we load `FactoMineR` and import the data file into R (`19_prepositions_brownlobkolh.rds`, see companion files).[6]

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

The data set has been imported as a data frame. To inspect it, enter `str(dfca)` and/or `head(dfca)`. It displays the number of times each preposition type is found in a certain context. The table consists of 257 lines (one line per preposition type) and 19 columns (one column per variable). 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 fifteen columns correspond to the text categories. The nineteenth column specifies the word length of the prepositions. This last column (`prep.length`) is loaded as a factor because it contains nominal data (for this reason, it is said to be qualitative).

The first three columns are declared as active. 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`).

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

By default, the `CA()` function produces a graph based on the first two dimensions. For the time being, these plots are not generated yet (`graph=FALSE`). Each graph will be plotted individually later, with specific parameters.

[5] 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).

[6] Details on how the data were extracted can be found in this blog post: `https://corpling.hypotheses.org/284` (accessed 9 June 2019).

The output of CA is in `ca.object`. The first lines of the ouput give the χ^2 score and the associated p-value. The χ^2 score is very high (10,053.43) 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 cells should display expected frequencies that are greater than 5. Our table contains many cells whose expected 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, 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.

```
> 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. Because the input table is simple and because the number of active variables is low, there are only two dimensions to inspect. Indeed, the first two dimensions represent 100% of the variance of the table. In most other studies, however, we should expect to inspect more than two dimensions. Our decision is based on the cumulative percentage of variance. 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 tendencies that we are about to observe are subtle.

```
> 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")
```

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, p. 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 (Husson, p.c.). Assessing the inter-distance between rows and columns accurately is possible in either an asymmetric plot 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.[7]

The CA graph is plotted with the `plot.CA()` function. The rows are made invisible to avoid cluttering the graph with prepositions (`invisible="ind"`). The prepositions can be plotted together with the column variables by removing `invisible="ind"`. To prevent the labels from being overplotted, `autoLab` is set to `"yes"`. By setting `shadowtext` to `TRUE`, a background shadow facilitates reading. The font size of the labels is adjusted 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, a title is included (`title=`). Its 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)
```

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. 19.1).

The first dimension (along the horizontal axis) accounts for 82.34% 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 prepositions (`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). The second dimension (along the vertical axis) accounts for 17.66% of the variance. It distinguishes the LOB corpus (upper part of the plot) from the

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

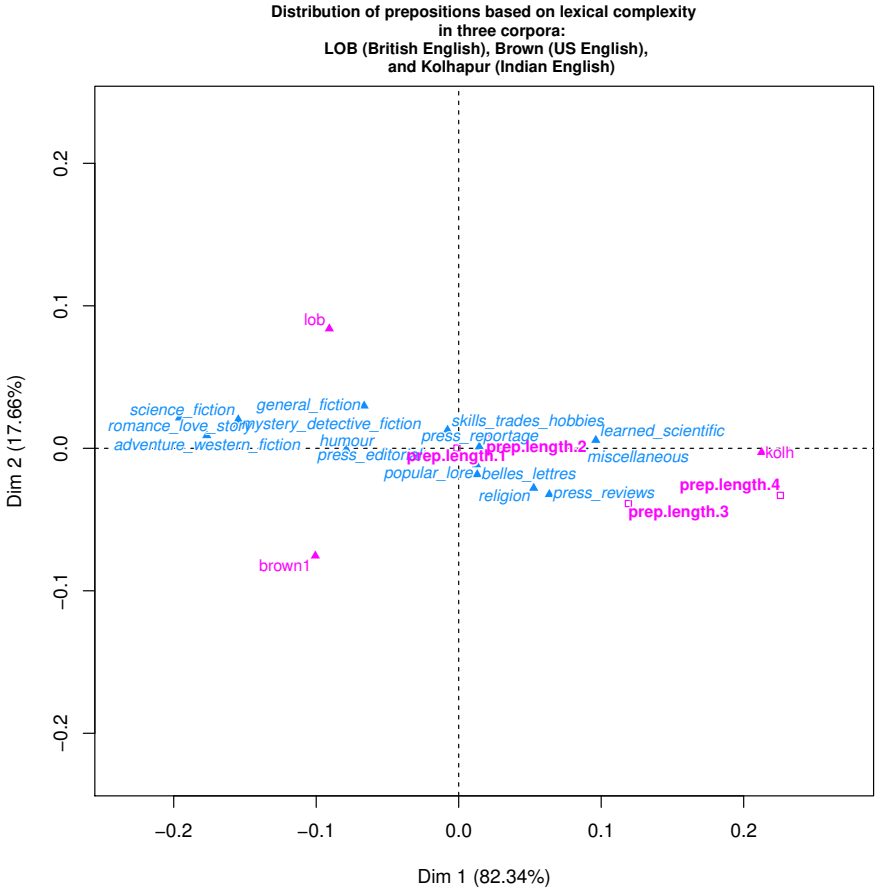

Fig. 19.1 CA biplot: a plane representation of individuals and variables (active and illustrative)

Brown corpus (lower part). All in all, complex prepositions are specific to the Kolhapur Corpus, especially in formal contexts.

19.3.2 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. 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's measures rely on the difference coefficient, borrowing the formula from Leech and Fallon (1992, p. 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 hypergeometric 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.

Schmid's study is repeated here in order to explore the distribution of swear-words with respect to gender in the BNC-XML. The goal 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.

The data file for this case study is `19_swearwords_bnc.txt` (see companion files).[8] Unlike Schmid, and following Rayson, Leech, and Hodges (1997), the data are extracted from the demographic component of the BNC-XML, which consists of spontaneous interactive discourse. The swear-words are: *bloody*, *damn*, *fuck*, *fucked*, *fucker*, *fucking*, *gosh*, and *shit*. Two exploratory variables are included in addition to gender: age and social class.[9]

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

The data set contains 293,289 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;

[8] 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.

[9] `http://www.natcorp.ox.ac.uk/docs/catRef.xml` (accessed 9 June 2019).

- `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.

As we inspect the structure of the data frame with `str()`, it is advisable to keep an eye on the number of levels for each variable and see if any can be kept to a minimum to guarantee that inertia will not drop.

```
> 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
```

The variable `word` has eight levels. We can 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
```

The number of levels has been reduced to five. We convert `df$word` back to a factor.

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

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-words and gender as active, and the variables age and social class as supplementary/illustrative.

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)
```

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 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. The eigenvalues can be vizualized by means of a scree plot (Fig. 19.2). It is obtained as follows.

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

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

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

In the MCA biplot (Fig. 19.3), 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 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. 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 male and female 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 male lower-class speakers.

A divide exists between male (`m`, right) and female (`f`, left) 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 dimensions 3 and 4 are inspected together (Fig. 19.4). To do so, the argument `axes=c(3,4)` is added in the `plot.MCA()` call.

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

With respect to dimensions 3 and 4, the male/female distinction disappears (both variables overlap where the two axes intersect). A divide is observed between *f*-words and *bloody* (left), used mostly by younger and middle-aged speakers, and *gosh* and *damn* (right), used mostly by upper-class speakers from age groups 3 and

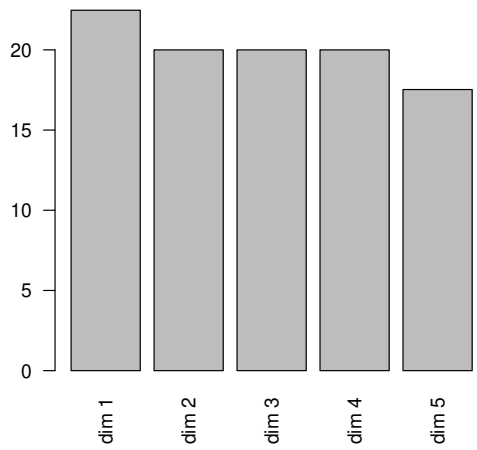

Fig. 19.2 A scree plot showing the eigenvalues associated with each dimension

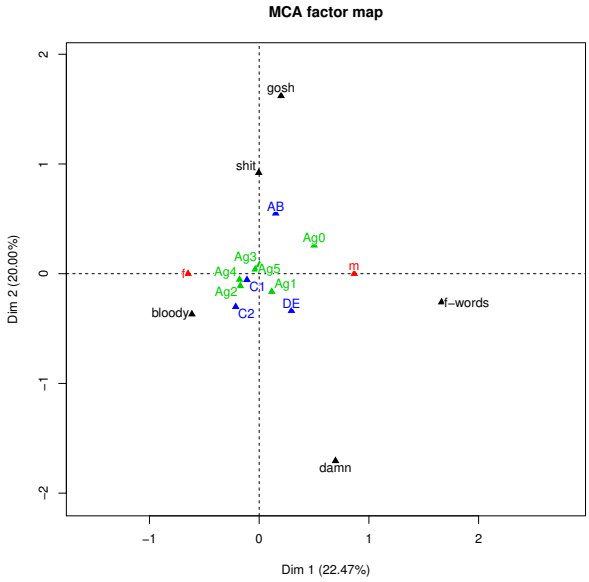

Fig. 19.3 MCA biplot: a plane representation of individuals and categories

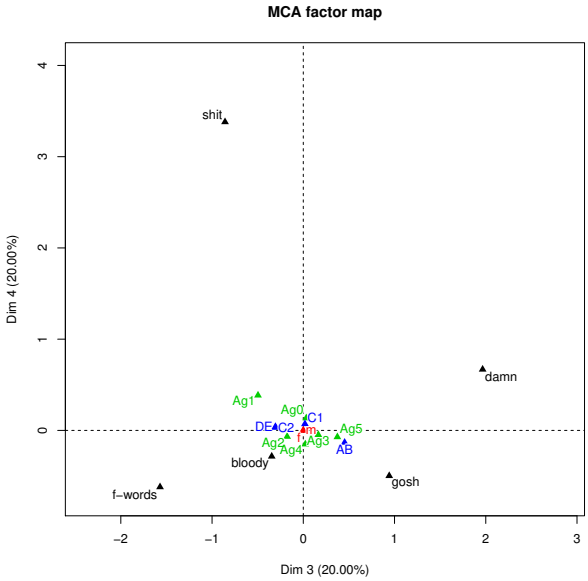

Fig. 19.4 MCA biplot: a plane representation of individuals and categories (dimensions 3 and 4)

5. The most striking feature is the outstanding position of *shit* in the upper-left corner. Although used preferably by male and female upper class speakers (Fig. 19.3), it is also used, although to a lesser degree, by much younger speakers from lower social classes.

19.3.3 Principal component 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), which is based on the hypergeometric distribution. 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).

To compare the semantic profiles of the prepositions, the preferred and dispreferred nominal collocates of the prepositions are examined in the FrWaC corpus.

The goal is to summarize the table graphically instead of interpreting the data table directly.

First, we load the data set (`19_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.

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. We load the `FactoMineR` package and run the PCA with the `PCA()` function.[10] The table contains measurements in the same unit. Standardizing them avoids giving each variable a weight proportional to its variance. Perhaps some prepositions attract most nouns more than others. The variables are standardized by default.

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

We make sure that the first two components are representative.[11] These eigenvalues are plotted in Fig. 19.5.

```
> 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
```

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.

```
> 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 (Fig. 19.6).

```
> # 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="")
```

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

[11] 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.

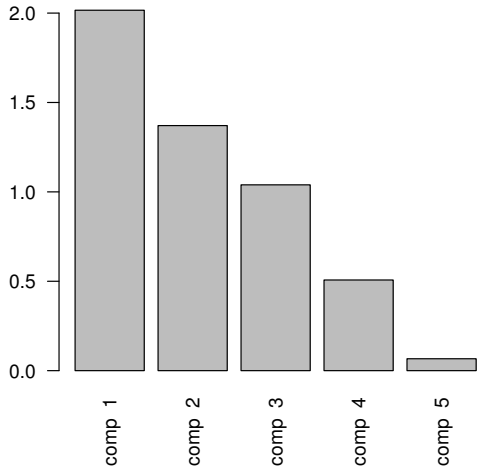

Fig. 19.5 A scree plot showing the eigenvalues associated with each component

Three main profiles stand out:

- *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 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. 19.7 displays four versions of the plot of individuals of Fig. 19.6.

The `select` argument of the `PCA()` function allows the user to 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")
```

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;

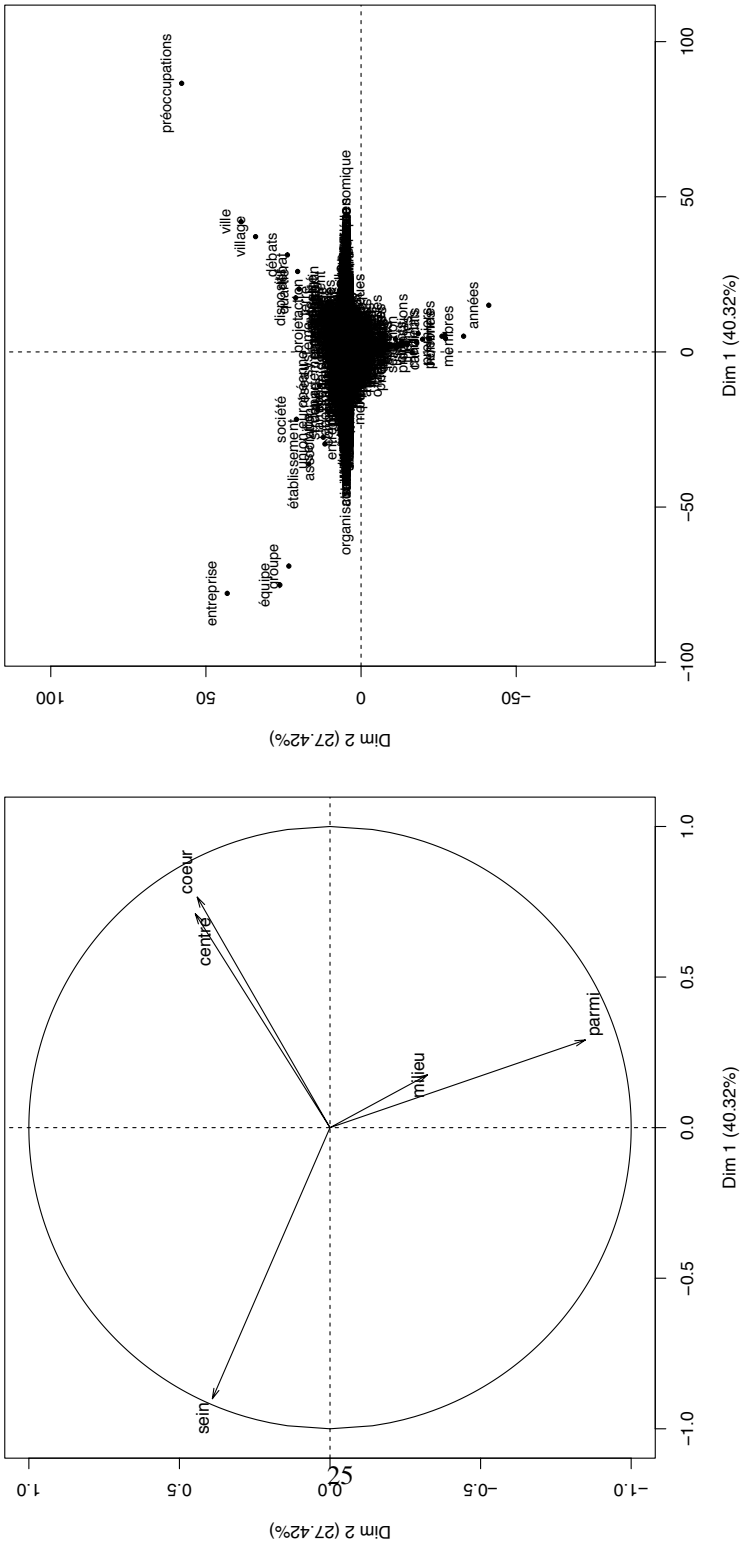

Fig. 19.6 PCA plots

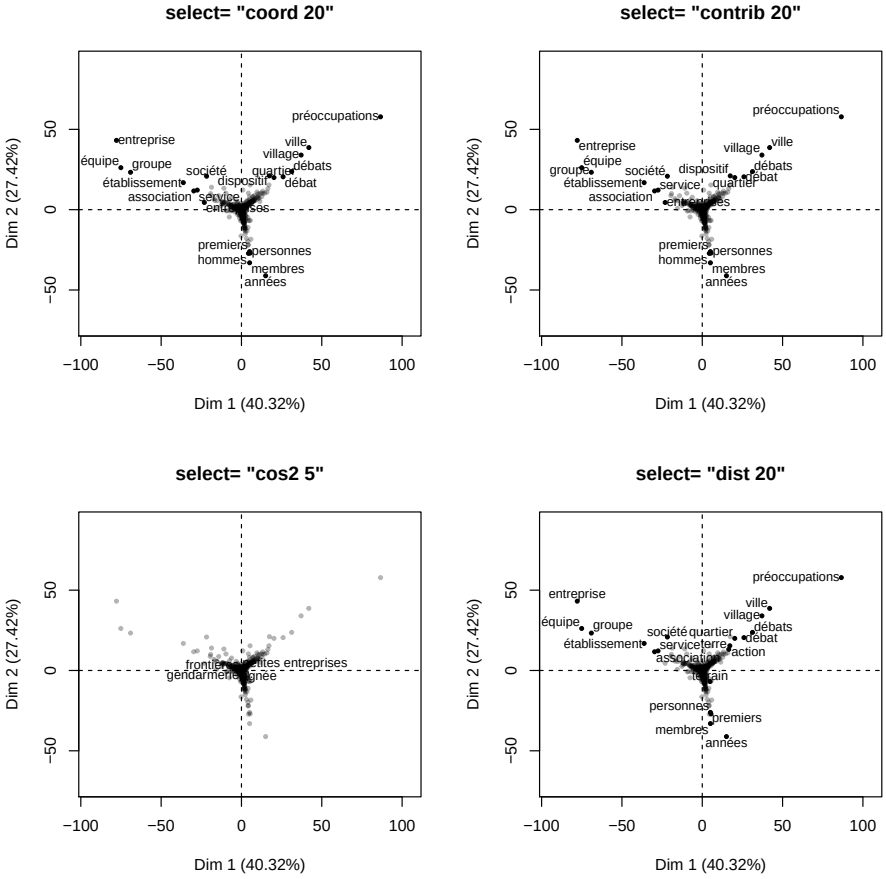

Fig. 19.7 Selecting NPs with `select`

- with `select="contrib 20"`, only the labels of the twenty individuals that have the highest contributions on the chosen dimensions are plotted;[12]
- with `select="cos2 5"`, only the labels of the five individuals that have the highest squared-cosine scores on the chosen dimensions are plotted;[13]
- 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.

Clear trends emerge:

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

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

- 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.

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.

19.3.4 Exploratory factor analysis

The same data set serves as input for EFA, which is performed with `factanal()`. According to Fig. 19.5, which shows that three principal components are worth investigating, we are tempted to 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.[14] 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 12,667.73 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.

[14] How many factors are considered worth keeping involves a choice based a metric known as SS loadings, as explained below.

In base R, we run EFA with `factanal()`.[15] The `factors` argument is set to 2. By default, the varimax rotation applies.

```
> 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 output displays uniqueness, loadings (the loadings that are too close to zero are not displayed), the proportions of variance explained by the factors, and the χ^2 test. 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.

The proportions of variance explained by the factors (i.e. eigenvalues) are listed under the factor loadings. A factor is considered worth keeping if the corresponding 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.

In EFA, rotation is 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, 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. Figure 19.8 is a plot of the load-

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

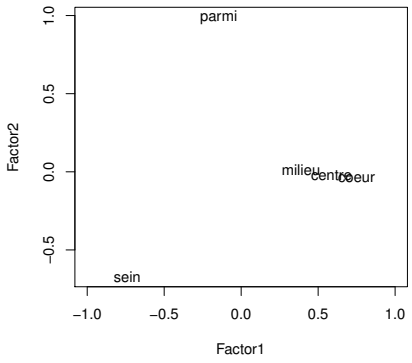

Fig. 19.8 loadings with varimax rotation

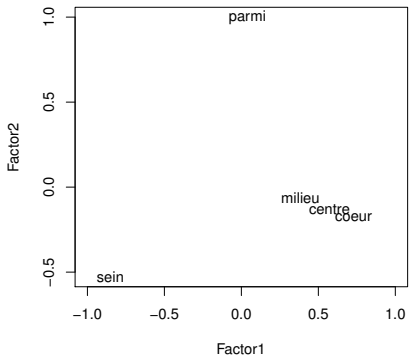

Fig. 19.9 loadings with promax rotation

ings of the prepositions on the two factors with varimax rotation. Figure 19.9 is the same plot of the loadings with promax rotation.

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

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

To produce a plot with promax rotation, 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))
```

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.

19.3.5 Reporting results

When reporting the results of CA, MCA, or PCA, the following elements should be included:

- the cumulative percentage of variance explained by each dimension/component;
- the graph and its interpretation.

Additionally, numeric descriptors such as contribution and quality of projection can be reported.

Each methods has its specificities. In CA, it is customary to report the χ^2 test result to see if the table deviates from independence. This result is part of the default output of the `CA()` function of the `FactoMineR` package (see Section 19.3.1).

In MCA, the eigenvalues associated with the first dimensions are often much lower than in CA and PCA. This means that it is often necessary to take more dimensions into account in the analysis. When the dimensionality of a dataset is high, the representation quality of a variable on a given plane is bound to be poor. However, how much a variable contributes to a given dimension is not affected by the high-dimensional nature of the data. Although optional, taking a look at the contribution and reporting the scores might be a good idea. In Section 19.3.2, the contribution scores of the variables are accessed by entering the following:

```
> mca.object$var$contrib
              Dim 1        Dim 2        Dim 3        Dim 4        Dim 5
bloody   1.668529e+01 6.782887e+00 5.964242e+00 4.032829e+00 1.668529e+01
damn     4.740002e+00 3.204420e+01 4.254051e+01 4.924310e+00 4.740002e+00
f-words  2.784165e+01 7.606179e-01 2.789159e+01 4.350051e+00 2.784165e+01
gosh     7.329551e-01 5.436848e+01 1.835285e+01 5.123951e+00 7.329551e-01
shit     1.019655e-04 6.043815e+00 5.250816e+00 8.156886e+01 1.019655e-04
f        2.141147e+01 1.164802e-21 5.381496e-23 1.268778e-21 2.141147e+01
m        2.858853e+01 1.524969e-21 7.838154e-23 1.686021e-21 2.858853e+01
```

In PCA, there are two graphs to inspect: the graph of variables and the graph of individuals (see Section 19.3.3). The graphs produced with CA, MCA, and PCA should be interpreted by focusing first on the horizontal axis and then on the vertical axis.

The output of `fa.object` in Section 19.3.4 is typical of how the results of an EFA should be reported. Therefore, it can conveniently be copied and pasted into the results section of a paper. Biber (1988) offers an excellent example of how the linguist can make sense of the EFA numeric indicators. See Chapter 26 for more general information on how to report the results in a quantitative corpus-based study.

19.4 Further reading

R. H. Baayen (2008). ***Analyzing Linguistic Data: A Practical Introduction to Statistics using R*****. Cambridge University Press**

Well-known to quantitative linguists, this textbook explains, among many other methods, how to run CA and PCA with R. It also shows how to run EFA. The data sets are directly relevant to linguistics and provided as part of the `languageR` package.

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

Chap. 10 of this book presents in greater detail three of the four methods covered in this chapter: CA, MCA, and PCA. Each method is illustrated with a detailed linguistic case study. The corresponding data sets and R scripts are provided in the form of companion files.

M. J. Greenacre (2007). *Correspondence Analysis in Practice*. Vol. 2. Interdisciplinary statistics series. Boca Raton: Chapman & Hall/CRC

This textbook focuses on CA and its variants (joint correspondence analysis, canonical correspondence analysis, co-inertia analysis, co-correspondence analysis) as well as multiple correspondence analysis. Although the book gives priority to practice, the theoretical and mathematical aspects of CA are presented in two appendices (A and B, respectively). The book can be read in combination with the documentation of the `ca` package (Nenadic and Greenacre 2007).

F. Husson, S. Lê, and J. Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. London: CRC press

Like Greenacre (2007), this book's main thrust is toward practice while making room for the theoretical and mathematical underpinnings of multivariate exploratory methods. It shows how to implement CA, PCA, and MCA with the `FactoMineR` package featured in the present chapter.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Benzécri, J.-P. (1984). *Analyse des Correspondances: Exposé Élémentaire*. Vol. 1. Pratique de l'Analyse des Données. Paris: Dunod.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D. (2001). "Dimensions of variation among eighteenth-century registers." In: *Towards a History of English as a History of Genres*. Ed. by H.-J. Diller and M. Görlach. Heidelberg: C. Winter, pp. 89–110.

Biber, D. and S. Conrad (2001). *Variation in English: Multi-dimensional studies*. London: Longman.

Biber, D. and B. Gray (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press.

Biber, D. and M. Hared (1992). "Dimensions of register variation in Somali." In: *Language Variation and Change* 4.1, pp. 41–75.

de Leeuw, J. and P. Mair (2009). "Simple and Canonical Correspondence Analysis Using the R Package *anacor*." In: *Journal of Statistical Software* 31.5, pp. 1–18.

Desagulier, G. (2015). "A lesson from associative learning: asymmetry and productivity in multiple-slot constructions." In: *Corpus Linguistics and Linguistic Theory*. DOI: 10.1515/cllt-2015-0012.

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Quantitative Methods in the Humanities and Social Sciences. New York: Springer.

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

Glynn, D. (2014). "The many uses of *run*." In: *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Ed. by D. Glynn and J. A. Robinson. Vol. 43. Human Cognitive Processing. John Benjamins, pp. 117–144.

Gréa, P. (2017). "Inside in French." In: *Cognitive Linguistics* 28.1, pp. 77–130.

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

Gries, S. T. (2006). "Corpus-based methods and cognitive semantics: The many senses of *to run*." In: *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Ed. by S. T. Gries and A. Stefanowitsch. Mouton de Gruyter, pp. 57–99.

Grieve, J. et al. (2010). "Variation among blogs: A multi-dimensional analysis." In: *Genres on the Web*. Ed. by A. Mehler, S. Sharoff, and M. Santini. Vol. 42. Text, Speech and Language Technology. Springer, pp. 303–322.

Habert, B. (1985). "L'analyse des formes «spécifiques» [bilan critique et propositions d'utilisation]." In: *Mots* 11.1, pp. 127–154.

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

Hirschmüller, H. (1989). "The use of complex prepositions in Indian English in comparison with British and American English." In: *Englische Textlinguistik und Varietätenforschung*. Ed. by G. Graustein and W. Thiele. Vol. 69. Linguistische Arbeitsberichte. Leipzig: Karl Marx Universität, pp. 52–58.

Hofland, K. and S. Johansson (1982). *Word Frequencies in British and American English*. Norwegian computing centre for the Humanities.

Husson, F., S. Lê, and J. Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. London: CRC press.

Kassambara, A. and F. Mundt (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5.

Kay, P. (2013). "The Limits of (Construction) Grammar." In: *The Oxford Handbook of Construction Grammar*. Ed. by T. Hoffmann and G. Trousdale. Oxford: Oxford University Press.

Kilgarriff, A. (2005). "Language is never, ever, ever, random." In: *Corpus Linguistics and Linguistic Theory* 1.2, pp. 263–276.

Kim, Y.-J. and D. Biber (1994). “A corpus-based analysis of register variation in Korean.” In: *Sociolinguistic Perspectives on Register*. Ed. by D. Biber and E. Finegan. New York: Oxford University Press, pp. 157–181.

Labbé, C. and D. Labbé (1994). “Que mesure la spécificité du vocabulaire ?” In: *Lexicometrica* 3, p. 2001.

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

Lê, S. and J. Pagès (2003). “Hierarchical multiple factor analysis: Application to the comparison of sensory profiles.” In: *Food Quality and Preference* 14.5-6, pp. 397–403.

Lê, S. and J. Pagès (2010). “Dmfa: Dual multiple factor analysis.” In: *Communications in Statistics—Theory and Methods* 39.3, pp. 483–492.

Leech, G. and R. Fallon (1992). “Computer corpora–what do they tell us about culture.” In: *ICAME journal* 16.

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

Leech, G., S. Johansson, and K. Hofland (1978). *The LOB Corpus, original version (1970–1978)*. Lancaster, Oslo, Bergen.

Leitner, G. (1991). “The Kolhapur Corpus of Indian English: Intra-varietal description and/or intervarietal comparison.” In: *English Computer Corpora*. Ed. by S. Johansson and A.-B. Stenström. Topics in English Linguistics. Berlin: Mouton de Gruyter, pp. 215–232.

Nenadic, O. and M. J. Greenacre (2007). “Correspondence Analysis in R, with two- and three-dimensional graphics: The `ca` package.” In: *Journal of Statistical Software* 20.3, pp. 1–13.

Pagès, J. (2014). *Multiple Factor Analysis by Example using R*. Boca Raton: Chapman & Hall/CRC.

Rayson, P., G. N. Leech, and M. Hodges (1997). “Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus.” In: *International Journal of Corpus Linguistics* 2.1, pp. 133–152.

Salem, A. (1987). *Pratique des Segments Répétés: Essai de Statistique Textuelle*. Paris: Klincksieck.

Schmid, H. J. (2003). “Do men and women really live in different cultures? Evidence from the BNC.” In: *Corpus Linguistics by the Lune*. Ed. by A. Wilson, P. Rayson, and T. McEnery. Lódź Studies in Language. Frankfurt: Peter Lang, pp. 185–221.

Shastri, S. V., C. T. Patilkulkarni, and G. S. Shastri (1986). *The Kolhapur Corpus*. Kolhapur, India.