Multifactorial Exploratory Approaches: correspondence analysis

Guillaume Desagulier

Multifactorial Exploratory Approaches

correspondence analysis

Guillaume Desagulier[1]

[1]MoDyCo (UMR 7114)
Paris 8, CNRS, Paris Nanterre
Institut Universitaire de France
gdesagulier@univ-paris8.fr

Corpus Linguistics Summer School 2019
June 25th, 2019
University of Birmingham

outline

CA

- 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).

CA

- 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.

the χ^2 test

- 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

the χ^2 test

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.

the χ^2 test

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

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

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 is the sum of the χ^2 values of all cells.

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

profile

- central to CA is the concept of profile
- to obtain the profile of a row, each cell is divided by its row total

row profiles

Table 5 displays the row profiles of Table 3. The row profiles add up to 1.

Table 1: The row profiles of Table 3

	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

column profiles

Likewise, one obtains the profile of a column by dividing each column frequency by the column total (Table 4). Again, the column profiles add up to 1.

Table 2: The column profiles of Table 3

	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

symmetry

CA performs an analysis of rows and columns that is both simultaneous and symmetric.

Table 3: An example of a contingency table

	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

symmetry: column analysis

A column analysis consists in interpreting the column profiles using the rows as reference points on the basis of a table such as Table 4. 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), it appears that these noun types are slightly over-represented in A1K.xml by a ratio of $\frac{0.6189}{0.5594} \approx 1.1063$.

Table 4: The column profiles of Table 3

	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

symmetry: row analysis

A row analysis consists in interpreting the row profiles using the columns as reference points on the basis of a table such as Table 5, 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. $\frac{0.1241}{0.1122} \approx 1.1063$.

Table 5: The row profiles of Table 3

	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

inertia

- 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

dimensions and eigenvalues

- 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

contribution and quality of projection

Two descriptors help interpret the dimensions:

- contribution: 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
- quality of projection: used to select the dimension in which the individual or the variable is the most faithfully represented

active or supplementary/illustrative individuals or variables

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.

complex prepositions

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, Garside, et al. 1986; Leech, Johansson, and Hofland 1978). The Kolhapur Corpus is a corpus of Indian English (Shastri et al. 1986).

complex prepositions

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*.

complex prepositions

Hirschmüller 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.

complex prepositions

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.

complex prepositions

We run CA on the preposition data set:
`prepositions_brownlobkolh.txt`

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

complex prepositions

```
> 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 ...
 $ 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              : int  1 1 1 1 2 1 1 1 1 1 ...
```

complex prepositions

The data set has been imported as a data frame. 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 3 kinds of columns:

- the first 3 columns correspond to the three corpora.
- the next 15 columns correspond to the text categories.
- The 19th 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).

complex prepositions

The first three columns are declared as active. Columns 4 to 18 are quantitative and declared as supplementary (`col.sup=4:18`). 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)
```

complex prepositions

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 eigenvalues;
- interpreting the CA graph.

complex prepositions

```
> 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 =  (
*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"
```

complex prepositions

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.

```
> 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
```

complex prepositions

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.)

complex prepositions

The CA graph is plotted with the `plot.CA()` function

```
> 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)
```

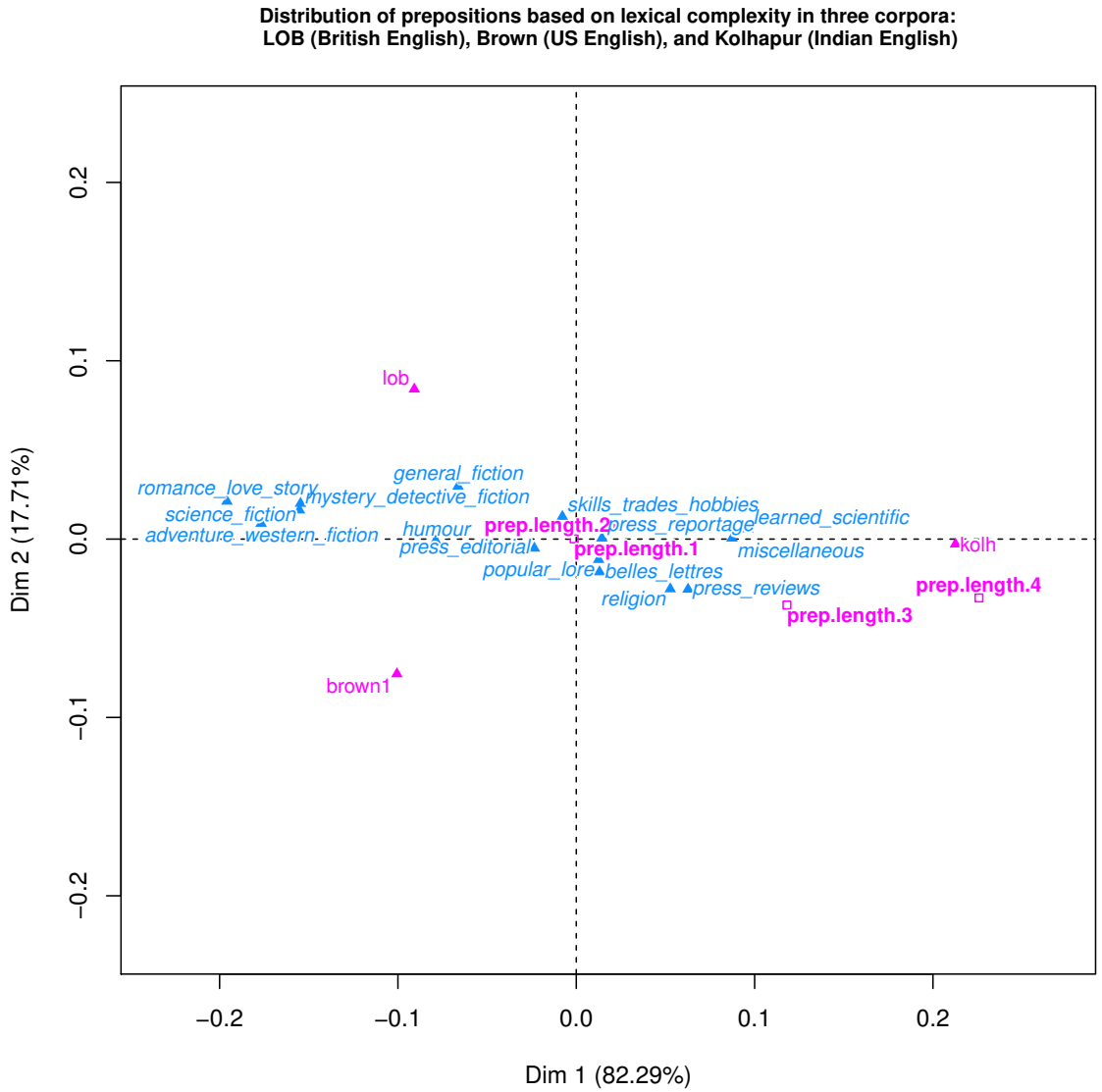
complex prepositions

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...

complex prepositions



complex prepositions

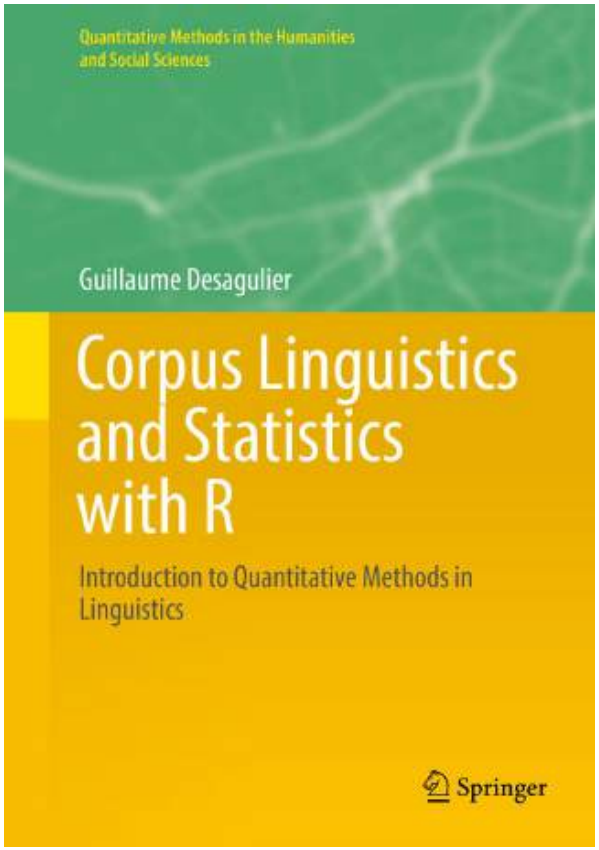
summary

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 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). All in all, complex prepositions are specific to the Kolhapur Corpus, especially in formal contexts.

Practical Handbook of Corpus Linguistics

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

Corpus Linguistics and Statistics with R



Section 10.4 – (Desagulier 2017)

Bibliography I

- Benzécri, Jean-Paul (1984). *Analyse des correspondances, exposé élémentaire*. Vol. 1. Pratique de l'analyse des données. Paris: Dunod.
- Desagulier, Guillaume (to appear). "Multifactorial exploratory approaches." In: *Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Thomas Gries. New York: Springer.
- – (2017). "Clustering Methods." In: *Corpus Linguistics and Statistics with R*. New York, NY: Springer, pp. 239–294.
- Francis, W. Nelson and Henry Kučera (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Brown University. Providence, Rhode Island.
- Greenacre, Michael J. (2007). *Correspondence Analysis in Practice*. Vol. 2. Interdisciplinary statistics series. Boca Raton: Chapman Hall/CRC.

Bibliography II

- Hirschfeld, Hermann 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, 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. Vol. 69. Linguistische Arbeitsberichte. Leipzig: Karl Marx Universität, pp. 52–58.
- Leech, Geoffrey, Stieg Johansson, Roger Garside, et al. (1986). *The LOB Corpus, POS-tagged version (1981–1986)*. Lancaster, Oslo, Bergen.
- Leech, Geoffrey, Stieg Johansson, and Knut Hofland (1978). *The LOB Corpus, original version (1970–1978)*. Lancaster, Oslo, Bergen.

Bibliography III


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. Topics in English Linguistics. Berlin: Mouton de Gruyter, pp. 215–232.


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