doi.org/10.26398/IJAS.0035-015

MULTILINGUAL TEXTUAL DATA: AN APPROACH THROUGH MULTIPLE FACTOR ANALYSIS

Belchin Kostov
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Ramón Alvarez-Esteban[1]
Department of Economics and Statistics, Universidad de León, León, Spain

Mónica Bécue-Bertaut
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain

François Husson
Institut Agro, Université Rennes 1, CNRS, IRMAR, Rennes, France

Abstract *This paper focuses on the analysis of open-ended questions answered in different languages. Closed-ended questions, called contextual variables, are asked to all respondents in order to understand the relationships between open-ended and closed-ended responses across samples, as the latter are likely to influence word choice. We have developed "Multiple Factor Analysis on Generalised Aggregated Lexical Tables" (MFA-GALT) to examine together open-ended responses in different languages through the relationships between word choice and the variables that drive that choice. MFA-GALT investigates whether the variability between words is structured in the same way as the variability between variables, and vice versa, from one sample to another. An application to an international satisfaction survey shows the easy-to-interpret results proposed.*

Keywords: *Correspondence analysis, Lexical tables, Textual and contextual data, Multiple factor analysis, Generalised aggregated lexical table*

1. INTRODUCTION

Socio-economic surveys benefit from the introduction of open-ended questions alongside closed-ended questions because they are mutually enriching. Closed-ended questions may inform the interpretation of open-ended questions, as the meaning of words is related to the speaker's characteristics or opinions. For example, customers in a satisfaction survey are asked to rate certain aspects of the

[1]Ramón Alvarez-Esteban, ramon.alvarez@unileon.es. ORCID 0000-0002-4751-2797

product and then to give their free opinion on which aspects could be improved, which is clearly linked to the ratings. In a survey that includes the question "What does health mean to you?", closed-ended questions such as gender, age, education and health status are very helpful for exploring how definitions of health vary with these variables. In the case of international surveys, which is our framework, these open-ended questions raise the issue of analysing responses from different samples in different languages.

For a single language, textual statistics (Benzécri, 1981; Lebart et al., 1998) provide multidimensional tools for processing free responses. Separately for each sample, the free responses are coded in the form of respondents $\times$ words, called a lexical table (LT). A standard methodology is to apply correspondence analysis to this LT (CA-LT; direct analysis) and to use the closed information as a complement. It is also common to group the responses of the categories of a closed question (e.g. age crossed with gender or education level, called a contextual variable), and to create a frequency table of words $\times$ categories, known as an aggregated lexical table (ALT), which can also be analysed by CA (CA-ALT).

These approaches are extended to multiple quantitative or qualitative contextual variables by using linearly constrained CA methods (Takane et al., 1991). Balbi and Giordano (2001) deal with textual data including external information; Balbi and Misuraca (2010) propose a double projection strategy by involving external information on both documents and words; while Spano and Triunfo (2012) apply canonical correspondence analysis (CCA; ter Braak (1986, 1987)) to textual data. In line with these works, Bécue-Bertaut et al. (2014) and Bécue-Bertaut and Pagès (2015) propose the CA method on a generalised aggregated lexical table (CA-GALT). The GALT is analysed by means of a CCA adapted to textual data. In CA-GALT, as in any CA, the variability of the vocabulary is explained by the variability of the variables, and the variability of the variables is explained by the variability of the vocabulary. This fits perfectly with the perspective we have chosen here.

In the case of multilingual surveys, we propose to analyse simultaneously the different GALTs, one for each monolingual sample, using a multiple factor analysis (MFA; (Escofier and Pagès, 2016; Pagès, 2014)) adapted to processing a multiple GALT. This produces the Multiple Factor Analysis for Generalised Aggregate Lexical Tables (MFA-GALT). This paper outlines how to adapt MFA reasoning to handle a multiple GALT, and details its properties and graphical representations.

The aim of MFA-GALT is to jointly study the open-ended responses from several samples in different languages through the relationships between the choice

of words and the variables that motivate this choice. These relationships may or may not have similar structures. In other words, MFA-GALT examines whether the variability between words is structured in the same way as the variability between variables, and vice versa, across samples.

The paper is organised as follows: Section 2 presents the data structure and notation; Section 3 recalls the principles of CA-GALT and MFA, the methods that form the basis of our approach; Section 4 is devoted to MFA adapted to multiple GALTs (MFA-GALT); and Section 5 presents the properties of the method. Finally, MFA-GALT is used in a full-scale application (Section 6) to demonstrate its capabilities. The main conclusions are presented in Section 7.

2. DATA STRUCTURE AND NOTATION

L samples answered a questionnaire with closed questions, either quantitative or categorical, all of the same type; these constitute the contextual data. They also answered an open-ended question in different languages, the answers to which are the source of the textual data set. The l sample has I_l respondents who all together use J_l different words in the l language. From these answers we build the $(I_l \times J_l)$ table $\mathbf{Y_l}$, respondents $\times$ words; N_l is the grand total for this table.

The closed questions are common to all samples. The answers are coded in the $(I_l \times K)$ table $\mathbf{X_l}$, whose columns correspond to either quantitative or dummy variables encoding the categories of one or more categorical variables. Regardless of the type, k and K denote the column-variable k and the total number of column-variables respectively. The term *variable* is henceforth used for both types. From $\mathbf{Y_l}$, the proportion table $(I_l \times J_l)$ is calculated $\mathbf{P_l} = \mathbf{Y_l}/N_l$.

If we consider only the sample l, the respondents' weights are taken from the margin of the rows of $\mathbf{P_l}$ — thus proportional to the number of occurrences of words in their free answers — and stored in the $(I_l \times I_l)$ diagonal matrix $\mathbf{D_l}$. The total weight of the respondents belonging to the same sample is equal to 1. The weights of the words are similarly obtained from the margin of the columns of $\mathbf{P_l}$, thus proportional to their counts, and stored in the $(J_l \times J_l)$ diagonal matrix $\mathbf{M_l}$. The total weight of the words used by the same sample is equal to 1. $\mathbf{X_l}$ is centred and possibly normalised in the case of quantitative variables, using the weighting system $\mathbf{D_l}$. The $(J_l \times K)$ table $\mathbf{Q_l} = \frac{\mathbf{Y_l^T X_l}}{\mathbf{N_l}} = \mathbf{P_l^T X_l}$ is the data structure containing the relations between words and variables. $\mathbf{Q_l}$ is called a generalised aggregated lexical table.

Note. The name **Generalised Aggregated Lexical Table** and the acronym **GALT** are used to emphasise the close similarity between this table and the classi-

cal **Aggregated Lexical Table** (**ALT**) developed in the case of a single categorical variable (Lebart et al., 1998).

The calculation is exactly the same in both cases. What changes is only the expression of the matrix **X** itself. An ALT consists of the dummy variables corresponding to the categories of a single categorical variable.

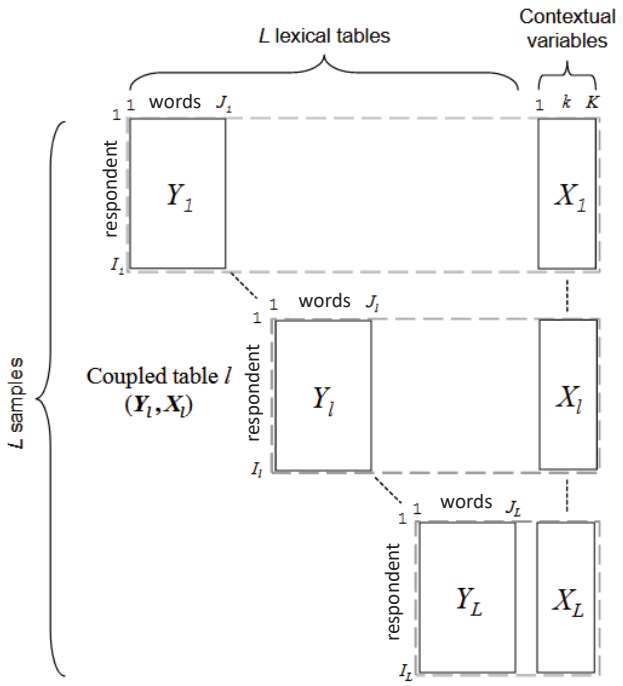


Figure 1: Sequence of L coupled tables

In the global analysis of the L samples, we have to deal with $I = \sum_l I_l$ respondents who have used $J = \sum_l J_l$ different words in the $N = \sum_l N_l$ occurrences that they have pronounced in all their free responses. The respondent and word weights are rescaled so that both totals are equal to 1 for the I respondents and J words respectively. To do this, the respondent and word weights in sample l are multiplied by N_l/N. The global weights of the respondents are stored in the $(I \times I)$ diagonal matrix **D**. The global weights of the words are stored in the $(J \times J)$ diagonal matrix **M**.

The $(I \times K)$ global table **X** is obtained combining by rows the L tables $\mathbf{X_l}$, centred by set. Table **X** is therefore also centred for weighting system **D**.

We assume $K < J$. The symbols I, I_l, J, J_l, K, L henceforth refer to both the

set and its cardinality.

3. METHODS USED AS THE BASIS OF OUR APPROACH

3.1. DEALING WITH ONE SAMPLE

In this section, we deal with only one sample and therefore consider it unnecessary to use index l.

3.1.1. CA-GALT method

We want to analyse the GALT matrix $\mathbf{Q}$ following a CA-like approach as far as possible. We therefore use the CA-GALT method (Bécue-Bertaut and Pagès, 2015; Bécue-Bertaut et al., 2014), as summarised below.

Let the $(K \times K)$ matrix $\mathbf{C} = (\mathbf{X^T D X})$ be the weighted correlation/covariance matrix of the variable-columns of the matrix $\mathbf{X}$. We compute the $(J \times K)$ matrix $\mathbf{Z}$, the double standardised form of the matrix $\mathbf{Q}$:

$$\mathbf{Z} = \mathbf{M^{-1} Q C^{-1}}. \tag{1}$$

If $\mathbf{C}$ is not invertible, $\mathbf{C^{-1}}$ is replaced by the Moore-Penrose pseudoinverse $\mathbf{C}^-$.

CA-GALT is then performed by principal component analysis (PCA) in two metrics: $\mathbf{C}$ in the row space, and $\mathbf{M}$ in the column space, i.e. PCA$(\mathbf{Z}, \mathbf{C}, \mathbf{M})$. This involves computing the S $(S \leq K)$ eigenvalues and eigenvectors of

$$\mathbf{Z^T M Z C}. \tag{2}$$

The eigenvalues are stored in the $(S \times S)$ diagonal matrix Λ, and the eigenvectors in the $(K \times S)$ matrix $\mathbf{U}$.

CA-GALT is a dual-projected analysis (Bécue-Bertaut and Pagès, 2015) that explains the variability of the words according to the variability of variables and, the variability of the variables according to the variability of the words.

Note. Metric $\mathbf{C^{-1}}$ (or $\mathbf{C}^-$) performs a multivariate standardisation that not only standardises the columns of $\mathbf{X}$ separately, but also makes them uncorrelated (Brandimarte, 2011; Härdle and Simar, 2012).

3.2. MFA GENERAL SCHEME

Multiple factor analysis (Escofier and Pagès, 2016; Pagès, 2014) analyses the multiple table combining by columns either quantitative or categorical tables. It has also been extended to frequency tables (Bécue-Bertaut and Pagès, 2004). This

method analyses a set of rows described by different sets of columns. The core of MFA is a PCA with specific weights and metrics applied to the multiple table containing quantitative tables as in PCA, categorical tables as in Multiple Correspondence Analysis (MCA), and frequency tables, especially lexical tables, as in CA. The specific approach to each type of table is obtained by coding the initial data and choosing the appropriate weights and metrics.

In order to balance the influence of the sets on the first factorial dimension, the initial weights of the columns in a given set are divided by the first eigenvalue resulting from the separate analysis of the corresponding table (PCA, MCA or CA depending on its type). The highest axial inertia of each set is thus normalised to 1. MFA identifies the main directions of variability in the data from a description of the rows by all the different sets of columns but balancing the importance of these sets, and provides the classic results of principal component methods. The characteristics and interpretation rules of PCA, MCA and CA are preserved for the quantitative, categorical and frequency sets. MFA also offers graphical tools for comparing the different sets, such as the superimposed global and partial representation of the rows as induced by all the sets or by each set separately, as well as a synthetic representation of the sets where each one is represented by only one point. These graphical results allow us to compare the typologies provided by each set in a common reference space.

4. MFA ON MULTIPLE GALT

Below, we adapt MFA to the case where the separate tables are GALTs built from the various samples, i.e. from the different coupled tables $(\mathbf{Y_l}, \mathbf{X_l})$ $(l = 1, .., L)$. The GALTs and their analysis are integrated into this approach by means of CA-GALT.

As described above, MFA is usually applied to a set of rows described by several sets of columns. We now need to analyse several sets of row-words described by one set of column-variables. However, here we are in a CA-like context where the roles of rows and columns are interchangeable. We could do this without changing the results. In the following sections we present the MFA-GALT method in a direct way.

MFA-GALT is performed in two steps exactly like a classic MFA. First, each sub-table — here a GALT — is analysed separately by applying the appropiate factorial method for its type, here CA-GALT. In the second step, a global factorial analysis is performed on all sets of multiple tables, treating each set as in the separate analyses, but taking into account the reweighting used to balance the

influence of the sets so the different sets of rows have a similar influence on the first global axis. This reweighting consists of dividing the weights of the rows of set l by the first eigenvalue obtained in the separate analyses of this set, so the highest axial inertia of each set is standardised to 1. Among the properties of this reweighting of the rows, it should be noted that the within-sets structures are not modified and that except for very special cases, the first axis of the global analysis is common to several sets and cannot therefore be generated by a single table. These two steps are described in more detail below.

First step: separate analyses

Separate CA-GALTs are performed in each set on the GALT $\mathbf{Q_l}$ according to the method in Section 3.1.1, with the exception of the metric used in the row space (and the weighting system in the column space). In this case, the covariance/correlation matrix computed from all the respondents, i.e. $\mathbf{C} = (\mathbf{X^T D X})$, is used in all the separate analyses instead of the matrices $\mathbf{C_l} = (\mathbf{X_l^T D_l X_l})$ as all row sets must be located in the same metric space. $\mathbf{C^{-1}}$ (or $\mathbf{C}^-$, if $\mathbf{C}$ is not invertible) is therefore used to standardise $\mathbf{Q_l}$. In this first step $\mathbf{Z_l} = \mathbf{M_l^{-1} Q_l C}^-$ is analysed by means of PCA($\mathbf{Z_l}$,$\mathbf{C}$,$\mathbf{M_l}$). The L first eigenvalues λ_1^l are used in the second step.

Second step: global analysis

The row weighting system is updated to balance the influence of each set in the global analysis. By construction, the matrix $\mathbf{M}$ is divided into L blocks. Block l corresponds to the J_l words used in sample l. The weights of the words in block l are divided by λ_1^l, the first eigenvalue of the separate analysis of sub-table l. The resulting weights are stored in the $(J \times J)$ matrix $\mathbf{M}_\lambda$.

The $(J \times K)$ multiple table GALT $\mathbf{Q}$ combines by rows the L matrices $\mathbf{Q_l}$ but resized by multiplying them by coefficient N_l/N ($\mathbf{Q_l} \times N_l/N$). A double standardisation of $\mathbf{Q}$ on the rows and the columns produces the $(J \times K)$ table $\mathbf{Z} = \mathbf{M}_\lambda^{-1} \mathbf{Q C^{-1}}$. If $\mathbf{C}$ is not invertible, $\mathbf{C^{-1}}$ is replaced by the Moore-Penrose pseudoinverse $\mathbf{C}^-$. MFA-GALT is then performed by a non-standardised weighted PCA on the multiple table $\mathbf{Z}$, with $\mathbf{M}_\lambda$ as row weights and metric in the column space and $\mathbf{C}$ as column weights and metric in the row space, i.e. PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{M}_\lambda$).

5. MAIN PROPERTIES OF MFA-GALT

MFA-GALT provides the classic outputs of the principal components methods, in this case a specific MFA performed on the double standardised GALT multiple table crossing words (in rows) and categories or quantitative variables (in columns). In particular, we obtain:

- coordinates, contributions and qualities of representation of word-rows
- coordinates and qualities of representation of category-columns or quantitative variable-columns.

Respondents could be reintroduced into the analysis either as supplementary rows (and thus represented on the basis of the values they take for the contextual variables) or as supplementary columns (and thus represented on the basis of the words they use). This is not further explored in this paper.

In addition, MFA outputs are provided as a partial representation of the variables, a synthetic representation of the sets, and a measure of the similarity between the sets.

5.1. REPRESENTATION OF ROW-WORDS AND COLUMN-VARIABLES

PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{M}_\lambda$) involves the diagonalisation of the matrix $\mathbf{Z^T M}_\lambda \mathbf{ZC}$. The principal axis of rank s corresponds to the eigenvector $\mathbf{u_s}$ ($\|\mathbf{u_s}\|_\mathbf{C}$=1) associated with the eigenvalue λ_s:

$$\mathbf{Z^T M}_\lambda \mathbf{ZCu_s} = \lambda_s \mathbf{u_s}. \tag{3}$$

The eigenvalues λ_s are stored in the $(S \times S)$ diagonal matrix Λ and the eigenvectors u_s — the dispersion axes — are stored in the columns of the $(K \times S)$ matrix $\mathbf{U}$.

By factor s we mean the vector of coordinates on axis s of either the word-rows (denoted F_s) or the variable-columns (denoted G_s) (Benzécri, 1973; Pagès, 2014). The values of the S row factors are stored in the columns of the $(J \times S)$ matrix $\mathbf{F}$, calculated as follows:

$$\mathbf{F} = \mathbf{ZCU}. \tag{4}$$

The row factors place the words in the direction of either the categories of respondents who use them frequently or the quantitative variables for which the respondents who use them have high values.

The values of the S column factors are stored in the columns of the $(K \times S)$ matrix $\mathbf{G}$. The matrix $\mathbf{G}$ is computed by using the transition relations between the

row and column factors, as in any PCA:

$$\mathbf{G} = \mathbf{Z}^{\mathbf{T}}\mathbf{M}_{\lambda}\mathbf{F}\Lambda^{-1/2}. \tag{5}$$

Thus, these scores are equal to the weighted covariances, or weighted correlation coefficients, between the standardised row factors and the doubly standardised columns of the multiple GALT.

5.2. SUPERIMPOSED REPRESENTATION OF THE *l* CLOUDS OF VARIABLES

According to the L sets of row-words, the column-variable k of $\mathbf{Z}$ can be divided into L sub-columns, called partial variables and denoted k^l. It is useful to represent simultaneously the L partial scatterplots, each made up of the corresponding K partial variables, on the same axes of reference. We successively consider the L matrices $\mathbf{Z_l}$ of dimension $(J_l \times K)$ issued from the matrix $\mathbf{Z}$ by retaining only the row-words belonging to the set l. From these matrices, the $(J \times K)$ matrices $\tilde{\mathbf{Z}}_l$ are built by completing $\mathbf{Z_l}$ with zeroes to have the same dimension as $\mathbf{Z}$. In order to be represented on the global axes, the K partial variables corresponding to the set l are considered as supplementary columns in the global analysis. Their coordinates are calculated using the transition relations and stored in the $(K \times S)$ matrix $\mathbf{G^l}$:

$$\mathbf{G^l} = \tilde{\mathbf{Z}}_l{}^{\mathbf{T}}\mathbf{M}_{\lambda}\mathbf{F}\Lambda^{-1/2}. \tag{6}$$

Therefore, the coordinates of the partial variables corresponding to set l can be calculated from the coordinates of the words used by sample l only. Thanks to the structure of the matrix $\tilde{\mathbf{Z}}_l$ which contains only 0 except for the rows belonging to set l, this relation for the partial variable k^l is expressed very simply :

$$G_s(k^l) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^l}} \sum_{j \in J_l} z_{jk} m_{jj} F_s(j). \tag{7}$$

In Eq.7, $[z_{jk}]$ denotes the generic term of $\mathbf{Z}$ and $\frac{1}{\sqrt{\lambda_1^l}} m_{jj}$ denotes the generic term of the matrix $\mathbf{M}_{\lambda}$, where m_{jj} is the initial weight of the word j (see Section 2).

According to Eq.7, the partial variables relative to set l are on the side of the words in this sample that are overused by respondents with high values for these contextual variables.

All the "partial" variables can usually be represented on the same scatterplot, thus providing information about the similarities/dissimilarities between the samples.

5.3. GLOBAL REPRESENTATION OF THE SETS

Another result is the representation of the L groups on the same graph, each of which is represented by a point (Pagès, 2014). To this end, the Lg coefficient (the formula of wich will be reminded below), the linkage measurement between one variable and one set of variables, is applied here to measure the linkage between each axis retained and each set of variables. First, the $(K \times K)$ matrix of scalar products $\mathbf{W_l}$ between the K column-variables of set l is computed as

$$\mathbf{W_l} = \mathbf{Z_l^T M_{\lambda l} Z_l}. \tag{8}$$

where the diagonal matrix $\mathbf{M_{\lambda l}}$, as block l of the matrix $\mathbf{M_\lambda}$, contains the weights of the variables of set l, equal here to $\frac{1}{\lambda_l^1}$.

$Lg(l, \mathbf{u_s})$ is then calculated as follows:

$$Lg(l, \mathbf{u_s}) = \langle \mathbf{W_l C}, \mathbf{u_s C} \rangle = trace(\mathbf{W_l C u_s u_s^T C}). \tag{9}$$

$Lg(l, \mathbf{u_s})$ will be used as a coordinate to place set l on the axis of rank s. This coordinate always has a value between 0 and 1. This produces a map of all the sets, each represented by one point. This map also shows the similarity, i.e. the proximity between the structures in the L sets.

5.4. MEASURE OF THE ASSOCIATION BETWEEN VOCABULARY AND CONTEXTUAL VARIABLES

Our proposal also includes the measurement of the association between vocabulary and contextual variables, firstly to select the variables that actually play a role, and secondly to interpret the results. The measures carried out successively for each sample are described in detail in Bécue-Bertaut and Pagès (2015).

Briefly, vocabulary is said to be associated with a variable if words differ significantly in the values taken by the people using them. The association between a categorical variable and vocabulary is evaluated with the classic chi-square test on the frequency table crossing words and categories (=lexical table).

A one-way analysis of variance (ANOVA) is considered in the case of a quantitative variable. The data table is reorganised as shown in Figure 2 before computing the one-way ANOVA: each row corresponds to one occurrence of a word (there are as many rows as the total number of occurrences in the corpus). The score variable and the words variable have as many values as occurrences. The one-way ANOVA is then performed between the score and the words to detect any relationships between vocabulary and scores.

Individuals	Score variable	Words		
		word A	word B	word C
ind 1	4	2	0	1
ind 2	6	1	0	0
ind 3	3	0	1	1

Individuals	Words	Score
ind 1	word A	4
ind 1	word A	4
ind 1	word C	4
ind 2	word A	6
ind 3	word B	3
ind 3	word C	3

Figure 2: Reorganization of the data for the one-way ANOVA measuring the association between vocabulary and a contextual variable.

It should be noted that since the occurrences are not independent, the usual assumptions of ANOVA are not satisfied and it is better to use permutation tests.

6. REAL DATA APPLICATION: INTERNATIONAL SURVEY

A railway company conducted a survey to determine how satisfied passengers were with its night trains. Passengers were asked to rate their satisfaction with 13 aspects related to comfort (general, cabin, bed, seat), cleanliness (common areas, cabin, toilet), staff attention (welcome attention, trip attention, language skills) and others (cabin room, air conditioning, general aspects). Each aspect was rated on a 11-point Likert scale, from 0 (very poor) to 10 (excellent). An open-ended question was added asking about the aspects that needed improvement. This question required spontaneous answers, in this case expressed in English or Spanish. The data are stored in the data structure shown in Figure 3.

The pre-processing of the data includes a careful correction of the spelling of the free answers. Stop-words are removed and the words used at least ten times are then kept for the Spanish corpus (=all answers given in Spanish), while the threshold for the English corpus is five (Lebart et al., 1998; Murtagh, 2005). Finally, 977 respondents from the Spanish sample and 283 from the English sample have no empty answers. The average length of free answers is 3.1 occurrences in both cases. The Spanish corpus contains 3029 occurrences corresponding to 88 different words and the English corpus has 871 occurrences corresponding to 68 different words.

Missing values have been imputed for the score variables. It should be noted that the rating scale has been inverted to make the graphs easier to read. The highest scores correspond to the highest levels of dissatisfaction.

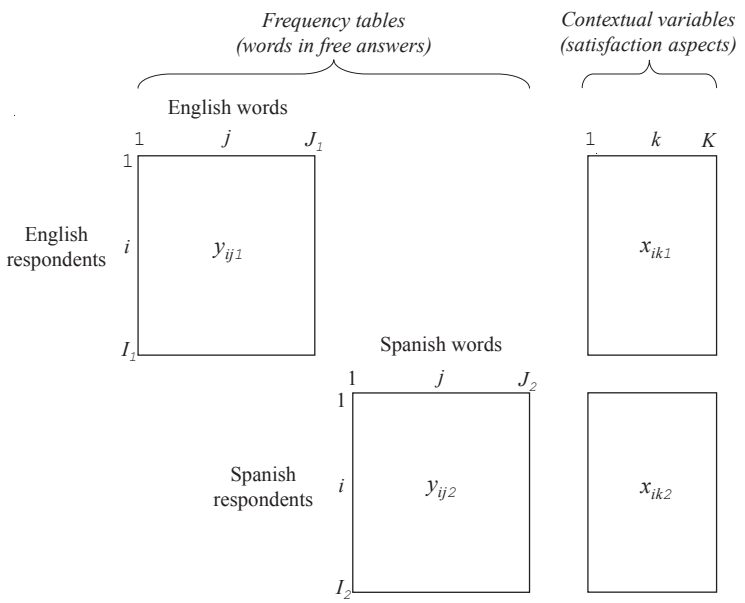


Figure 3: The dataset. On the left, the lexical tables; on the right, the contextual variables. In the example, I_1 = 283 (English respondents), I_2 = 977 (Spanish respondents), J_1 = 68 (English words), J_2 = 88 (Spanish words), K = 13 (satisfaction aspects).

Table 1: Mean satisfaction scores and association with vocabulary ratios

	Spanish respondents		English respondents	
Satisfaction aspects	mean (SD)	ass.ratio (p-value)	mean (SD)	ass.ratio (p-value)
General comfort	6.82 (1.80)	0.062 (<0.001)	6.66 (1.95)	0.091 (0.148)
Cabin comfort	6.37 (2.07)	0.063 (<0.001)	6.37 (2.03)	0.121 (0.010)
Cabin room	5.33 (2.43)	0.089 (<0.001)	5.71 (2.35)	0.136 (<0.001)
Bed comfort	6.70 (1.98)	0.050 (<0.001)	6.72 (2.03)	0.063 (0.918)
Seat comfort	6.10 (2.20)	0.059 (<0.001)	5.99 (2.38)	0.123 (0.010)
Air conditioning	6.55 (2.55)	0.107 (<0.001)	6.51 (2.71)	0.226 (<0.001)
Common areas cleanliness	7.41 (1.92)	0.043 (<0.001)	7.54 (1.86)	0.082 (0.548)
Cabin cleanliness	7.59 (1.88)	0.056 (<0.001)	7.59 (1.81)	0.116 (0.036)
Toilet cleanliness	6.21 (2.55)	0.090 (<0.001)	6.29 (2.40)	0.150 (<0.001)
Staff welcome attention	7.99 (1.92)	0.040 (0.018)	7.29 (2.45)	0.108 (0.062)
Staff trip attention	8.07 (1.85)	0.038 (0.048)	7.34 (2.29)	0.092 (0.294)
General aspects	7.77 (1.65)	0.038 (0.034)	7.48 (1.91)	0.079 (0.590)
Staff language skills	7.72 (2.08)	0.052 (<0.001)	7.14 (2.52)	0.154 (<0.001)

6.1. INITIAL FINDINGS

The most frequent words give a preliminary overview of the complaints, which are similar in both languages and expressed with homologous words. *Espacio/space* is too reduced, no place for *maletas/luggages*. *Cabinas/cabins* and *asientos/seats* lack *comodidad/comfort*, while *aseos/toilets* would benefit from more *limpieza/cleanliness*. The *Aire acondicionado/Air conditioning* seems to be causing problems. In the English sample, the words *staff* and *English* are frequently mentioned. Aspects that were not asked about are mentioned, such as *precio/price*.

Table 1 provides a first insight with the means and standard deviations of the satisfaction scores. *Staff trip attention* obtains the highest score (8.07) from Spanish-speaking respondents while English-speaking respondents gave the highest score to *Cabin cleanliness* (7.59). The lowest score is for *Cabin room* for both Spanish (5.33) and English-speaking respondents (5.71). It is worth noting that the three aspects related to staff (*Staff welcome*, *Staff trip attention* and *Staff language skills*) are significantly less valued by English-speaking than by Spanish-speaking respondents.

The association between vocabulary and a contextual variable (see Table 1, columns *ass.ratio (p-value)*) shows that *Air conditioning* receives the highest ratio for both Spanish (0.107) and English-speaking respondents (0.226). *Toilet cleanliness* is the second most important indicator for Spanish-speaking respondents (0.090), while *Staff language skills* is second with 0.154 for English respondents, although closely followed by *Toilet cleanliness* in third place (0.150). It should

be noted that Spanish-speaking respondents rank the *Staff language skills* only eighth. *Cabin room* is ranked third for Spanish-speaking and fourth for English-speaking respondents.

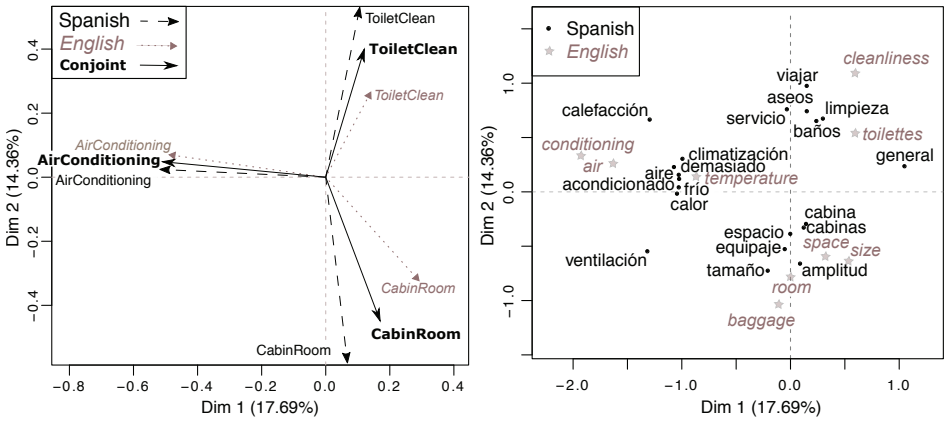
6.2. MFA-GALT ON THE MULTILINGUAL DATASET

The ranking aspects from the association-with-vocabulary ratio do not coincide with the score-average ranking. This shows that the information from the free comments differs from the closed questions, and that these two types of information are complementary, implying that the aspects the passengers believe should be improved do not match the aspects with which they are less satisfied. This justifies the interest in collecting information through open-ended questions, as this information is different and complementary.

MFA-GALT is applied on the multiple generalised aggregated lexical table. The total inertia is equal to 9.91. The first eigenvalue (1.75 corresponding to 17.69% of the total inertia) is close to the number of sets, which means that the two sets share the dispersion direction corresponding to the first global axis. The second (1.42, 14.36% of the total inertia) and third eigenvalue (1.23, 12.39% of the total inertia) are close, but the following eigenvalues are much smaller, so we focus only on the first three axes. To avoid over-emphasising the example, we will only interpret the first two axes. For a more detailed description of the results, and particularly the third dimension, the reader can refer to the thesis of Kostov (2015).

6.2.1. Global representation of the satisfaction scores and words

MFA-GALT provides graphical results in which each variable (each score) points to the words associated with it. It thus indicates the shortcomings of the scored aspect, whether or not they are common to both languages. Figure 4a shows the best represented satisfaction scores on the first MFA-GALT principal plane through their covariances with the axes. To avoid overloading the graphs, only the scores that are well represented are shown (in this case, those that have a square cosine sum on the two axes over 0.5). We first look at only the global representations of the scores, which has a three-polar structure. The three poles refer to inconveniences associated with *Air conditioning*, lack of *Toilet cleanliness* and problems related to *Cabin room*. This is in line with what the association-with-vocabulary ratios suggested. Figure 4b shows the Spanish and English words that contribute more than twice to the average contribution. We can then see words that are strongly associated with *air conditioning*, showing its shortcomings: *air*/*aire*,



(a) Best represented satisfaction scores (b) Most contributive words

Figure 4: MFA-GALT: Representation of variables (i.e. scores) and words on the plane (1,2)

conditioning/*acondicionado*, *temperature*, *frío* (=cold) *climatización* (=air conditioner), *ventilación*/*ventilation* and *calefacción* (=heating). On the positive part of the second axis, the lack of *Toilet cleanliness* is characterised by *cleanliness*/*limpieza*, *toilettes*/*aseos*/*baños*. On the negative part, the problems with *Cabin room* are described using the words *size*/*espacio* and *cabins*/*cabina(s)*.

In this example, axis 3 is specific to only the English set, which points to problems with the staff speaking poor English. This may seem like trivial information, but it shows that the method works and offers the possibility of highlighting information specific to only one sub-population, and also makes the transport company aware that language difficulties are a real problem highlighted by English-speaking respondents, unlike, for example, in air transport.

6.2.2. Partial representation of the satisfaction scores

Figure 4a shows the superimposed representation of the global and partial representations of the satisfaction scores on the plane (1,2) and highlights the similarities and differences between the two sets in terms of the association between words and scores. *Air conditioning* behaves similarly in both sets on the first axis. On the second axis, *Toilet cleanliness* and *Cabin room* are more strongly associated with Spanish than with English vocabulary, which translates into higher covariances; the complaints using English vocabulary appear more accentuated

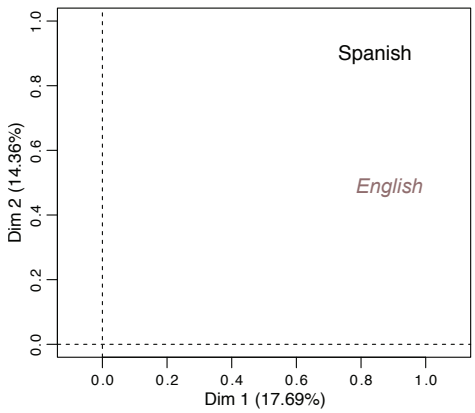


Figure 5: Representation of the sets.

and give rise to more words. In the third dimension, only English-speaking respondents complain about the lack of *Staff language skills*.

6.2.3. Representation of the sets

Similarity measures confirm that both sets share some dispersion directions. The value of the RV coefficient, multivariate generalisation of the squared Pearson correlation coefficient, equal to 0.74 ($p<0.001$), confirms that the partial configurations are relatively close but not homothetic.

According to the representation of the sets on the first dimension, the coordinate of the Spanish sample is 0.85 while the coordinate of the English sample has a slightly higher value (0.91) (Figure 5). This means that the first axis provided by MFA-GALT is of major importance for both sets and is therefore a common axis dispersion, while the Spanish set has a much larger coordinate on the second axis (0.91 vs. 0.51). The second MFAGALT axis is thus very important for Spanish-speaking respondents, and not so much for the English-speakers, while the opposite is observed for the third axis (0.42 for Spanish vs. 0.81 for English).

7. CONCLUSION

This paper proposes an original principal component method to deal with open-ended questions answered in different languages. This type of textual and contextual data produces a sequence of coupled tables, each comprising one frequency table (=lexical table) and one quantitative/qualitative table. We approach these

data through the relationships between the words and the contextual variables. Two methods — CA-GALT and MFA — are combined, hence the name of the new method: *Multiple Factor Analysis on Generalised Aggregated Lexical Tables* (MFA-GALT). The first places the words of the different sets in the same space generated by the variables, resulting in the construction of the GALTs; while the second allows the simultaneous analysis of these tables in a way that preserves the MFA properties.

An international survey with open questions answered in different languages was analysed with MFA-GALT, making it possible to study similarities among words from the same language, similarities among homologous words from different languages, associations between words and satisfaction scores, similarities between satisfaction score structures (partial representations) and similarities between groups. The results of this application show that MFA-GALT provides a good synthesis of the data and produces outputs that are easy to interpret.

The R package `Xplortext` includes the `LexGalt` function, which enables the implementation of the CA-GALT and MFA-GALT methods.

References

Balbi, S. and Giordano, G. (2001). A factorial technique for analysing textual data with external information. In S. Borra, R. Rocci, M. Vichi and M. Schader, eds., *Advances in Classification and Data Analysis*, 169–176. Springer, Berlin, Heidelberg.

Balbi, S. and Misuraca, M. (2010). A doubly projected analysis for lexical tables. In C.H. Skiadas, ed., *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13–19. Birkhäuser, Boston.

Bécue-Bertaut, M. and Pagès, J. (2004). A principal axes method for comparing contingency tables: MFACT. In *Computational Statistics and Data Analysis*, 45 (3): 481–503.

Bécue-Bertaut, M. and Pagès, J. (2015). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. In *Advances in Data Analysis and Classification*, 9: 125–142.

Bécue-Bertaut, M., Pagès, J. and Kostov, B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. In *Statistics and Operations Research Transactions*, 38: 285–302.

Benzécri, J.P. (1973). *Analyse des Données*. Bordas, Paris.

Benzécri, J.P. (1981). *Pratique de l'Analyse des Données. Tome 3, Linguistique & Lexicologie*. Bordas, Paris.

Brandimarte, P. (2011). *Quantitative Methods: an Introduction for Business Management*. John Wiley & Sons, New Jersey.

Escofier, B. and Pagès, J. (2016). *Analyses Factorielles Simples et Multiples*. Dunod, Paris, 5th edn.

Härdle, W. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*. Springer Verlag, Heidelberg, Berlin.

Kostov, B. (2015). *A principal Component Method to Analyse Disconnected Frequency Tables by Means of Contextual Information.* Ph.D. dissertation, UPC, Departament d'Estadística i Investigació Operativa. URL `https://upcommons.upc.edu/handle/2117/95759`.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall / CRC Press, New York.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Chapman and Hall/CRC, New York.

Spano, M. and Triunfo, N. (2012). La relazione sulla gestione delle società italiane quotate sul mercato regolamentato. In A. Dister, D. Longrée, and G. Purnelle, eds., *Actes de 11ème Journées d'analyse de données textuelles*. URL `http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm`.

Takane, Y., Yanai, H. and Mayekawa, S. (1991). Relationships among several methods of linearly constrained correspondence analysis. In *Psychometrika*, 56: 667–684.

ter Braak, C.J.F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. In *Ecology*, 67: 1167–1179.

ter Braak, C.J.F. (1987). *Canoco–A FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis (Version 2.1).* ITI-TNO Institute of Applied Computer Sciences, Wageningen.