SORT 38 (2) July-December 2014, 285-302

Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach

Mónica Bécue-Bertaut[1,*], Jérôme Pagès[2] and Belchin Kostov[1,3]

Abstract

This work proposes an original textual statistical method to uncover the relationships between opinions, expressed as free-text answers, and respondents' characteristics. This method also identifies the specific links between each characteristic and certain words used in these answers. Promising results are obtained as shown by an application to real data collected to know what health means for non-experts, essential knowledge for effective public health interventions.

MSC: 62P20

Keywords: Aggregated lexical tables, textual analysis, free-text answers, social marketing.

1. Introduction

Open-ended questioning is able to capture information in the form of free-text answers which could not be observed from closed questioning. The usual statistical methodology to deal with this type of answer gives a central role to correspondence analysis (CA; Benzécri, 1973, 1981; Lebart, Salem and Berry, 1998; Murtagh, 2005). However, the direct analysis by CA of the lexical table, crossing respondents (rows) and words (columns), benefit from introducing the respondents' characteristics, such as age and education, to obtain more robust results (Lebart et al., 1998, pp. 103-104). A first and classical way of doing it consists in grouping the respondents from one categorical variable

[*] The corresponding author will be the first one; her contact address is: Mónica Bécue-Bertaut. Departament d'Estadística i Inv. Operativa. Universitat Politècnica de Catalunya. C/ Jordi Girona 1-3, 08034, Barcelona, Spain. Phone: +34934017031, Fax: +34934015855, e-mail address: monica.becue@upc.edu

[1] Universitat Politècnica de Catalunya, Barcelona, Spain.

[2] Agrocampus-Ouest, Rennes, France.

[3] Transverse group for research in primary care-IDIBAPS, Barcelona, Spain.

Received: March 2014
Accepted: June 2014

and building an *aggregated lexical table* (ALT) crossing categories (rows) and words (columns). CA on this aggregated lexical table (CA-ALT) offers a symmetric approach to the relationships between words and categories allowing for explaining the variability observed among the words by the variability observed among the categories and vice-versa. The attractions/rejections between certain words and certain categories are indicated and visualised on the principal planes.

Considering several categorical variables is frequently required to better understand the variability observed in the lexical choices. An approach similar to CA-ALT consists in building a new variable from crossing all the categories of all the selected variables. In practice, such a cross-tabulation would lead to an unwieldy number of categories when dealing with samples limited to 500 or even 1000 respondents. Furthermore, as a complex network of relationships may exist among the variables, some of these categories would be either empty or with low counts.

When crossing the variables to be considered proves impracticable, three strategies are proposed (Lebart et al., 1998; Garnier and Guérin-Pace, 2010; Cousteaux, 2010). Performing a multiple correspondence analysis on these variables and clustering the respondents from their principal coordinates enables to return to the case of a single categorical variable. The partition into clusters plays the role of a categorical variable and the aggregated lexical table crossing clusters and words is built and then analysed by CA. This strategy, called *working demographic partition* (WDP), highlights the main lexical choices related to the characteristics of the respondents (Lebart et al., 1998, pp. 188-121). However, this strategy presents two main drawbacks. The clustering requires taking several decisions which are not obvious and any direct reference to the variables and categories is lost. This hinders the interpretation of the graphics in terms of relationships between variables/categories and words. A second option consists in applying CA to the multiple aggregated table juxtaposing the aggregated lexical tables built from each categorical variable. This approach has the drawback of not cancelling the associations among the variables and hence possible confusion effects remain. Finally, a direct analysis of the free-answers, that is, CA on the respondents × words table can be performed. The projection of the categories, at the centroid of the respondents who belong to them, allows for detecting which variables (and which categories) are strongly associated with the words. However, in this case also, the effects of the different variables are merged.

We present here a methodology able to take into account several grouping variables while untangling their respective influence on the lexical choices and avoiding spurious relationships between certain categories and certain words.

The overview of this paper is as follows. Section 2 presents the motivation, based on a case study. In Section 3, the notation is listed. Section 4 recalls the classical methodology to deal with an aggregated lexical table. Section 5 is devoted to the analysis of a multiple aggregated table through classical CA and through the methodology that we propose. The effectiveness of this latter is evaluated in Section 6 on the case study. We conclude in Section 7 with some remarks.

2. Case-study based motivation

In 1989-1990 the Valencian Institute of Public Health (IVESP) conducted a survey to better know the attitudes and opinions related to health for the non-expert population. This information is essential to enhance public health policy. Effective advertising and greater dissemination concerning healthy habits are thus oriented by a deep knowledge of real lifestyles. A sample of 513 residents over 14 years of age was observed. The first question included in the questionnaire "*What does health mean to you*?" required free and spontaneous answers. A priori, the variables *Age group* (*under 21, 21-35, 36-50* and *over 50)*, *Gender* and *Health condition* (*poor, fair, good* and *very good health*) were considered as possibly conditioning the respondents' viewpoint on health. The primary objective is to uncover and describe their complex influence on the ways of defining health. Identifying the different concerns and their relationships with the respondents' characteristics is aimed at.

3. Notation

For the convenience of the reader, the main notation and terminology are listed and specified here.

N, I, J, K, L	number of occurrences, respondents, words, categories, categorical variables, respectively;
$\mathbf{X} = [\mathrm{x}_{ik}]$	$(I \times K)$ two-way two-mode data matrix describing the respondents from K dummy variables issued from coding $L(L \geq 1)$ categorical variables into a disjunctive form. So, $x_{ik} = 1$ if i belongs to category k, otherwise 0;
$\mathbf{Y} = [y_{ij}]$	$(I \times J)$ two-way two-mode data matrix describing the respondents from the frequency of the words that they used to answer an open-ended question. y_{ij} counts the occurrences of word j (column) in respondent i's answer (row). The grand total of this table is $\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} = N$, total number of occurrences of the corpus. $\mathbf{Y}$ is called the lexical table (LT);
$\mathbf{P} = [p_{ij}] = \left[\frac{y_{ij}}{N}\right]$	$(I \times J)$ proportion matrix issued from the lexical table. The row margin of $\mathbf{P}$ is the vector $(p_{.1}, \ldots, p_{.j}, \ldots, p_{.J})^{\mathsf{T}}$ with, for $j = 1, \ldots, J$, $p_{.j} = \sum_{i=1}^{I} p_{ij}$. The column margin of $\mathbf{P}$ is the vector $(p_{1.}, \ldots, p_{i.}, \ldots, p_{I.})^{\mathsf{T}}$ with, for $i = 1, \ldots, I$, $p_{i.} = \sum_{j=1}^{J} p_{ij}$;

$\mathbf{D_I} = [d_{\mathbf{I}_{ii}}] = [p_{i.}]$	$(I \times I)$ diagonal matrix. $d_{\mathbf{I}_{ii}}$ is equal to the relative frequency of occurrences corresponding to respondent i's free answer;
$\mathbf{D_J} = [d_{\mathbf{J}_{jj}}] = [p_{.j}]$	$(J \times J)$ diagonal matrix. $d_{\mathbf{J}_{jj}}$ is equal to the relative frequency of occurrences of word j in the whole set of free answers;
$\mathbf{Q} = \mathbf{D_I^{-1} P D_J^{-1}}$ $= [q_{ij}] = \left[\frac{p_{ij}}{p_{i.}\,p_{.j}}\right]$	$(I \times J)$ data matrix analysed by CA.
$\bar{\bar{\mathbf{Q}}} = \mathbf{D_I^{-1}}\left(\mathbf{P} - \mathbf{D_I 1 D_J}\right)\mathbf{D_J^{-1}}$ $= [\bar{\bar{q}}_{ij}] = \left[\frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{i\cdot}p_{\cdot j}}\right]$	$(I \times J)$ data matrix, double-centred form of $\mathbf{Q}$ which can be alternatively considered by CA. $\mathbf{1}$ denotes the $(I \times J)$ matrix with generic term the constant 1. $\bar{\bar{\mathbf{Q}}}$ describes the weighted deviation between $\mathbf{P}$ and the $(I \times J)$ independence model matrix $\mathbf{D_I 1 D_J} = [p_{i.} \cdot p_{.j}]$;
$\mathbf{Y_A} = \mathbf{Y}^{\mathsf{T}}\mathbf{X} = [y_{jk}]$	$(J \times K)$ two-way two-mode data matrix describing the categories (columns) from the frequency of the words (rows) used in the free answers of the categories' respondents. $y_{\mathbf{A}_{jk}}$ is the count of occurrences of word j in category k's answers. $\mathbf{Y_A}$ is called either *aggregated lexical table* (ALT; $L = 1$) or *multiple aggregated lexical table* (MALT; $L > 1$);
$\mathbf{P_A} = \left[p_{\mathbf{A}_{jk}}\right] = \left[\frac{y_{\mathbf{A}_{jk}}}{L \cdot N}\right]$	$(J \times K)$ proportion matrix issued from $\mathbf{Y_A}$. The row margin of $\mathbf{P_A}$ is the vector $\left(p_{\mathbf{A}_{.1}}, \ldots, p_{\mathbf{A}_{.k}}, \ldots, p_{\mathbf{A}_{.K}}\right)^{\mathsf{T}}$ with $p_{\mathbf{A}_{.k}} = \sum_{j=1}^{J} p_{\mathbf{A}_{jk}}$, $k = 1, \ldots, K$. The column margin of $\mathbf{P_A}$ is the vector $\left(p_{\mathbf{A}_{1.}}, \ldots, p_{\mathbf{A}_{j.}}, \ldots, p_{\mathbf{A}_{J.}}\right)^{\mathsf{T}}$ with $p_{\mathbf{A}_{j.}} = \sum_{k=1}^{K} p_{\mathbf{A}_{jk}}$ for $j = 1, \ldots, J$;
$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A D_K^{-1}}$	$(J \times K)$ data matrix built from $\mathbf{P_A}$ and analysed by CA;
$\mathbf{D_K} = \left[d_{\mathbf{K}_{kk}}\right] = \left[p_{\mathbf{A}_{.k}}\right]$	$(K \times K)$ diagonal matrix which gathers the terms of the row-margin of $\mathbf{P_A}$. $d_{\mathbf{K}_{kk}}$ is the relative frequency of the occurrences used by the category k's respondents;
$\mathbf{Q_A^G} = \mathbf{D_J^{-1} P_A C^{-}}$	$(J \times K)$ data matrix built from $\mathbf{P_A}$ where $\mathbf{C}^{-}$ is the generalized inverse of $\mathbf{C} = \mathbf{X}^{\mathsf{T}}\mathbf{D_I X}$. This is the data matrix analysed by the methodology that we propose.

4. Classical correspondence analysis on lexical tables

We present CA and the methodology that we developed in terms of our application field. We consider the frequency table $\mathbf{Y}$ and the contextual data matrix $\mathbf{X}$ observed on the same respondents. In this Section the columns of $\mathbf{X}$ are dummy variables corresponding to the categories of only one variable.

In textual analysis, it is usual to apply CA on the lexical table (CA-LT), and on words×categories tables, called *aggregated lexical tables* CA-ALT.

4.1. Direct analysis of the lexical table CA-LT

As any classical CA, the direct analysis of the lexical table CA($\mathbf{Y}$) can be performed in three equivalent ways:

1. As the principal component analysis (PCA) on the following $(I \times J)$ data matrix

$$\bar{\bar{\mathbf{Q}}} = \mathbf{D_I}^{-1}(\mathbf{P} - \mathbf{D_I 1 D_J})\mathbf{D_J}^{-1} \tag{1}$$

   with metric $\mathbf{D_J}$ in the row space (metric $\mathbf{D_I}$ in the column space) and weighting system $\mathbf{D_I}$ on the rows (weighting system $\mathbf{D_J}$ on the columns) (Bécue-Bertaut and Pagès, 2004; Böckenholt and Takane, 1994; Escofier and Pagès, 2008). This PCA is denoted PCA$\left(\bar{\bar{\mathbf{Q}}}, \mathbf{D_J}, \mathbf{D_I}\right)$. This formulation, besides underlining that what is analysed is the deviation of $\mathbf{P}$ from the independence model matrix, places this method in the general scheme for principal axes methods. We favour here this point of view which allows for generalisations in a more straightforward manner. Equivalently, the $(I \times J)$ data matrix

$$\mathbf{Q} = \mathbf{D_I}^{-1}\mathbf{P}\mathbf{D_J}^{-1} \tag{2}$$

   can be considered in PCA$(\mathbf{Q}, \mathbf{D_J}, \mathbf{D_I})$.
   Both PCA$\left(\bar{\bar{\mathbf{Q}}}, \mathbf{D_J}, \mathbf{D_I}\right)$ and PCA$(\mathbf{Q}, \mathbf{D_J}, \mathbf{D_I})$ lead to the same results due to the centring usually performed by a PCA.

2. As the ordinary SVD of $\mathbf{D_I}^{-1/2}\mathbf{P}\mathbf{D_J}^{-1/2} = \mathbf{D_I}^{1/2}\mathbf{Q}\mathbf{D_J}^{1/2}$ completed by further computing to obtain the row and column factors (Böckenholt and Takane, 1994; Greenacre, 1984; Lebart et al., 2006; Legendre and Legendre, 1998).
3. As the two analyses of the row and column profiles matrices through, respectively, the PCA of $\mathbf{D_I}^{-1}\mathbf{P}$, with row metric $\mathbf{D_J}^{-1}$ and weighting system $\mathbf{D_I}$, and the PCA of $\mathbf{D_J}^{-1}\mathbf{P}$,with row metric $\mathbf{D_I}^{-1}$ and weighting system $\mathbf{D_J}$ (Escofier and Pagès, 2008; Lebart et al., 2006)

4.2. Analysis of an aggregated lexical table CA-ALT

The $(J \times K)$ aggregated lexical table

$$\mathbf{Y_A} = \mathbf{Y}^\mathsf{T}\mathbf{X} \tag{3}$$

is built and transformed into the $(J \times K)$ proportion matrix

$$\mathbf{P_A} = \mathbf{P}^\mathsf{T}\mathbf{X}. \tag{4}$$

The $(K \times K)$ diagonal matrix $\mathbf{D_K}$ stores the row-margin of $\mathbf{P_A}$ whose generic term is the proportion of occurrences corresponding to category k. In this section, where only one categorical variable is considered, $\mathbf{D_K}$ is equal to $\left(\mathbf{X}^\mathsf{T}\mathbf{D_I}\mathbf{X}\right)$.

From $\mathbf{P_A}$, the $(J \times K)$ matrix

$$\mathbf{Q_A} = \mathbf{D_J}^{-1}\mathbf{P_A}\mathbf{D_K}^{-1} \tag{5}$$

is computed. Then, CA-ALT is performed through PCA $(\mathbf{Q_A}, \mathbf{D_K}, \mathbf{D_J})$.

This analysis provides good and robust results which indicate the associations (respectively, oppositions) between words to the extent that they are related to identical (respectively, different) categories of the contextual variable.

4.3. Correspondence analysis as a double projected analysis

We consider the "inflated" $(N \times K)$ matrix $\mathbf{X_N} = [x_{\mathbf{N};n,k}]$ and $(N \times J)$ matrix $\mathbf{Y_N} = [y_{\mathbf{N};n,j}]$ (Legendre and Legendre, 1998, p. 595). $\mathbf{X_N}$ and $\mathbf{Y_N}$ cross the N occurrences and, respectively, the K indicators corresponding to the column-categories of table $\mathbf{X}$ and the J words. If occurrence n corresponds to word j, $y_{\mathbf{N};n,j} = 1$; $y_{\mathbf{N};n,j} = 0$ otherwise. If occurrence n has been pronounced by a respondent who presents category k, $x_{\mathbf{N};n,k} = 1$; $x_{\mathbf{N};n,k} = 0$ otherwise. The $(N \times N)$ diagonal matrix $\mathbf{D_N}[1/N]$ corresponds to the uniform weighting system on the rows. Both the column-words of $\mathbf{Y_N}$ and the column-variables of $\mathbf{X_N}$ are in R^N space.

The proportion matrix $\mathbf{P_A}$ can be rewritten as

$$\mathbf{P_A} = \mathbf{Y}_\mathbf{N}^\mathsf{T}\mathbf{D_N}\mathbf{X_N} \tag{6}$$

and matrix $\mathbf{Q_A}$ as

$$\mathbf{Q_A} = \mathbf{D_J}^{-1}\mathbf{P_A}\mathbf{D_K}^{-1} = \left(\mathbf{Y}_\mathbf{N}^\mathsf{T}\mathbf{D_N}\mathbf{Y_N}\right)^{-1}\left(\mathbf{Y}_\mathbf{N}^\mathsf{T}\mathbf{D_N}\mathbf{X_N}\right)\left(\mathbf{X}_\mathbf{N}^\mathsf{T}\mathbf{D_N}\mathbf{X_N}\right)^{-1}. \tag{7}$$

Eq. (7) shows that the columns of $\mathbf{Q_A}$ are the $\mathbf{D_N}$-orthogonal projection of the dummy-columns of $\mathbf{X_N D_K^{-1}} = \mathbf{X_N}\left(\mathbf{X_N^T D_N X_N}\right)^{-1}$ on the subspace of R^N generated by the column-words of $\mathbf{Y_N}$. Similarly, Eq. (7) shows that the rows of $\mathbf{Q_A}$ are the $\mathbf{D_N}$-orthogonal projection of the column-words of $\left(\mathbf{Y_N D_J^{-1}}\right) = \mathbf{Y_N}\left(\mathbf{Y_N^T D_N Y_N}\right)^{-1}$ on the subspace of R^N generated by the dummy-columns of $\mathbf{X_N}$.

This viewpoint highlights that CA studies both the variability of the cloud of words, insofar as it is explained by the variability of the categories, and the variability of the cloud of categories, insofar as it is explained by the variability of the words.

CA($\mathbf{Y_A}$) is a double-projected analysis because

$$\mathbf{D_K^{-1}} = \left(\mathbf{X^T D_I X}\right)^{-1} \tag{8}$$

is equal to the inverse of the matrix of moments of the second order of $\mathbf{X}$ relative to the origin, i.e, all of the off-diagonal terms are null because the columns of $\mathbf{X}$ are orthogonal.

Note that this rationale places CA in the context of canonical analysis (Saporta, 2006, pp. 212-217).

5. Analysis of a multiple aggregated lexical table

5.1. Classical correspondence analysis on a multiple aggregated lexical table

We may be interested in a broader context, such as a set of L categorical variables $(L > 1)$. As the starting point, the multiple aggregated lexical table is built by juxtaposing row-wise the L aggregated lexical table built from the L categorical variables. From now on, $\mathbf{Y_A}$ is used to denote this multiple aggregated lexical table. We follow a rationale akin to that of the former section.

The aggregated lexical table

$$\mathbf{Y_A} = \mathbf{Y^T X} \tag{9}$$

is built and transformed into the proportion matrix

$$\mathbf{P_A} = \frac{\mathbf{Y_A}}{L \cdot N} = \frac{\mathbf{P^T X}}{L}. \tag{10}$$

Diagonal matrix $\mathbf{D_K}$ stores the row-margin of $\mathbf{P_A}$ whose general term is the proportion of occurrences corresponding to category k. From $\mathbf{P_A}$, matrix

$$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A D_K^{-1}} \tag{11}$$

is computed. Then, PCA $(\mathbf{Q_A}, \mathbf{D_K}, \mathbf{D_J})$ is performed. As in the usual CA, the first eigenvalue is equal to 1 and the corresponding axis is neglected.

The main difference with Section 4.2 is that $\mathbf{D_K}$ is no longer equal to $\left(\mathbf{X}^\mathsf{T}\mathbf{D_I}\mathbf{X}\right)^{-1}$. This latter matrix presents non-null off-diagonal terms because the column-categories of $\mathbf{X}$ are generally not orthogonal when belonging to different variables. It is no longer a double-projected analysis and hence the influence of the associations among the categories of different variables is not filtered.

5.2. CA with a modified metric on a multiple aggregated lexical table

In this section, the dummy columns of $\mathbf{X}$ are centred. To maintain a double projected analysis, the starting point consists in substituting the row space metric $\mathbf{D_K^{-1}}$ by the Moore-Penrose pseudoinverse $\mathbf{C}^-$ of

$$\mathbf{C} = \left(\mathbf{X}^\mathsf{T}\mathbf{D_I}\mathbf{X}\right) = [c_{kk'}], \tag{12}$$

Matrix $\mathbf{C}$ is the covariance matrix between the columns of $\mathbf{X}$ taking into account that the respondents are endowed with weighting system $\mathbf{D_I}$.
Note: if $k = k'$, $c_{kk'}$ is equal to the sum of weights of the respondents belonging to this category. If $k \neq k'$ and k and k' belong to the same variable then $c_{kk'} = 0$; if $k \neq k'$ and k and k' belong to different variables, then $c_{kk'}$ is equal to the sum of weights of the respondents belonging both to category k and category k'. $\mathbf{C}^-$ substitutes $\mathbf{D_K^{-1}}$ in the expression of the $(J \times K)$ data matrix

$$\mathbf{Q_A^G} = \mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{C}^-, \tag{13}$$

that will be analysed through PCA $\left(\mathbf{Q_A^G}, \mathbf{C}, \mathbf{D_J}\right)$.

Metric $\mathbf{C}^-$ operates a multivariate standardisation that not only separately standardises the columns of $\mathbf{X}$ but in addition makes them uncorrelated (Brandimarte, 2011; Härdle and Simar, 2012). To compute $(\mathbf{C}^-)^{1/2}$, $\mathbf{C}$ is diagonalised and the whole of its S_C non-null eigenvalues, all positive, are ranked in descending order and stored in the $(S_C \times S_C)$ diagonal matrix $\mathbf{\Lambda_C}$. S_C is equal to the dimension of the space spanned by the columns of $\mathbf{X}$, that is, the number of independent dummy-columns of $\mathbf{X}$. The corresponding eigenvectors are stored in the columns of the $(K \times S_C)$ matrix $\mathbf{U_C}$. The S_C columns of $\mathbf{X}(\mathbf{C}^-)^{1/2}$, with $(\mathbf{C}^-)^{1/2} = \mathbf{U_C}\mathbf{\Lambda_C^{-1/2}}$, are standardised and uncorrelated. The set of dummy columns of $\mathbf{X}$ is now taken into account through the subspace that they span. Performing PCA $\left(\mathbf{Q_A^G}, \mathbf{C}, \mathbf{D_J}\right)$ is equivalent to analyse the column-centred multiple aggregated table $\mathbf{P_A}$ through CA with a modified metric $\mathbf{C}^-$ in the row space.

We have called the multiple aggregated lexical table $\mathbf{P_A}$ *generalised aggregated lexical table* (GALT) and the methodology *correspondence analysis on a* GALT (CA-

GALT). This analysis provides the usual PCA results. The S non-null eigenvalues are ranked in descending order and stored in the $(S \times S)$ diagonal matrix $\mathbf{\Lambda}$. The factors on the row-words and column-categories are stored, respectively, in the $(J \times S)$ matrix $\mathbf{F}$ and $(K \times S)$ matrix $\mathbf{G}$.

The interpretation of the results of this specific CA follows the usual CA interpretation rules (Escofier and Pagès, 2008; Greenacre, 1984; Lebart et al., 1998). We will only emphasize here the transition relationships. The transition relationships linking $\mathbf{F}$ and $\mathbf{G}$ are expressed in Eq. 15 and Eq. 17 hereafter.

Given that

$$\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{D_I}\mathbf{X}\right)^{-}\left(\mathbf{X}^{\mathsf{T}}\mathbf{D_I}\mathbf{X}\right)=\mathbf{X}, \tag{14}$$

the matrix $\mathbf{F}$ is expressed as

$$\begin{aligned}\mathbf{F} &= \mathbf{Q_A^G}\mathbf{C}\mathbf{G}\mathbf{\Lambda}^{-1/2} = \mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{C}^{-}\mathbf{C}\mathbf{G}\mathbf{\Lambda}^{-1/2} = \mathbf{D_J^{-1}}\frac{\mathbf{Y}^{\mathsf{T}}}{N\cdot L}X\left(\mathbf{X}^{\mathsf{T}}\mathbf{D_I}\mathbf{X}\right)^{-}\left(\mathbf{X}^{\mathsf{T}}\mathbf{D_I}\mathbf{X}\right)\mathbf{G}\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{G}\mathbf{\Lambda}^{-1/2}.\end{aligned} \tag{15}$$

The matrix $\mathbf{G}$ is expressed as

$$\begin{aligned}\mathbf{G} &= \left(\mathbf{Q_A^G}\right)^{\mathsf{T}}\mathbf{D_J}\mathbf{F}\mathbf{\Lambda}^{-1/2} = \left(\mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{C}^{-}\right)^{\mathsf{T}}\mathbf{D_J}\mathbf{F}\mathbf{\Lambda}^{-1/2} = \mathbf{C}^{-}\mathbf{P_A^{\mathsf{T}}}\mathbf{D_J^{-1}}\mathbf{D_J}\mathbf{F}\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{C}^{-}\mathbf{P_A^{\mathsf{T}}}\mathbf{F}\mathbf{\Lambda}^{-1/2}.\end{aligned} \tag{16}$$

By considering the matrices $\mathbf{Y_N} = [y_{\mathbf{N};n,j}]$ and $\mathbf{X_N} = [x_{\mathbf{N};n,k}]$ defined similarly to those in Section 4.3, but $\mathbf{X_N}$ now comprising the K centred dummy columns corresponding to all the categories of the selected categorical variables, $\mathbf{G}$ can be rewritten as

$$\mathbf{G} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{D_I}\mathbf{X}\right)^{-}\frac{\mathbf{X}^{\mathsf{T}}\mathbf{Y}}{N\cdot L}\mathbf{F}\mathbf{\Lambda}^{-1/2} = \left(\left(\mathbf{X_N^{\mathsf{T}}}\mathbf{D_N}\mathbf{X_N}\right)^{-}\mathbf{X_N^{\mathsf{T}}}\mathbf{D_N}\mathbf{Y_N}/L\right)\mathbf{F}\mathbf{\Lambda}^{-1/2} = \mathbf{B}\mathbf{F}\mathbf{\Lambda}^{-1/2} \tag{17}$$

Here the $(K \times J)$ matrix $\mathbf{B} = \left(\left(\mathbf{X_N^{\mathsf{T}}}\mathbf{D_N}\mathbf{X_N}\right)^{-}\mathbf{X_N^{\mathsf{T}}}\mathbf{D_N}\mathbf{Y_N}/L\right) = [b_{kj}]$ is, except for the scaling coefficient $1/L$, the matrix of regression coefficients (strictly, analysis of variance coefficients given that the regressors are dummy variables) of all the column-words of $\mathbf{Y_N}$ on the regressor column-categories of $\mathbf{X_N}$. These coefficients are issued from the simultaneous, or multivariate, linear regression of all the column-words of $\mathbf{Y_N}$ on the column-categories of $\mathbf{X_N}$ (Finn, 1974).

Eq. 15 shows that, as in classical CA, a word is placed on axis s, up to a coefficient varying from one axis to the other, at the centroid of the categories that use it, endowing the categories with the weighting system $\left(\frac{p_{Ajk}}{p_{Aj.}}, k = 1,\ldots,K\right)$.

Eq. 17 reflects that category k is placed on axis s, up to a coefficient varying from one axis to the other, at the centroid of the words, endowing them with the weighting system $(b_{kj}, j = 1, \ldots, J)$. The weight given to word j, equal to b_{kj}, is the coefficient of category k in the regression of column-word j on all the categories. Thus, a category is placed in the direction of the words that the respondents belonging to this category tend to use, all things being equal.

6. Results

From the data presented in Section 2, a multiple aggregated lexical table is built by juxtaposing the three aggregated lexical tables issued from using *age group* (four categories), *gender* (two categories) and *health condition* (four categories) as grouping variables.

We first perform a separate CA on each of the tables involved in the analysis. Then, a classical CA is applied on the multiple aggregated lexical table. Finally, CA-GALT is performed with the three previous variables as contextual variables. The comparison of the results obtained from these last two methods allows for demonstrating the effectiveness of CA-GALT.

6.1. Pre-processing of the data

The 392 respondents having answered the open-ended question are selected. Only the words used at least 10 times are selected because a minimum threshold on the word frequency is required to make the comparisons between free answers meaningful from a statistical point of view (Lebart et al., 1998, p. 104; Murtagh, 2005, chap. 5). The final corpus is composed of 7751 occurrences (corpus length) from 126 different words (vocabulary length).

6.2. Separate correspondence analysis on the lexical table and on the aggregated tables

Table 1 summarizes the results of each analysis through classical indicators that are the global inertia, the Cramer's V^2 and the first eigenvalue. Cramér's V^2 is computed by dividing Φ^2 by $Min(I-1, J-1)$, that is with the maximum inertia that the table could present.

The intensity of the relationship between the vocabulary and either the respondents or each of the grouping variables is measured through the inertia Φ^2 (Table 1). The Cramer's V^2 allows for comparing the intensity of the relationships between the rows (either the respondents or the categories of respondents) and the columns (the words) from one table to another.

In all the cases the Cramer's V^2 value is weak. This is a usual feature when analysing a corpus of open-ended answers. The associations between words and respondents/categories develop as small variations among words selected from a common vocabulary

Table 1: *Summary of the analyses.*

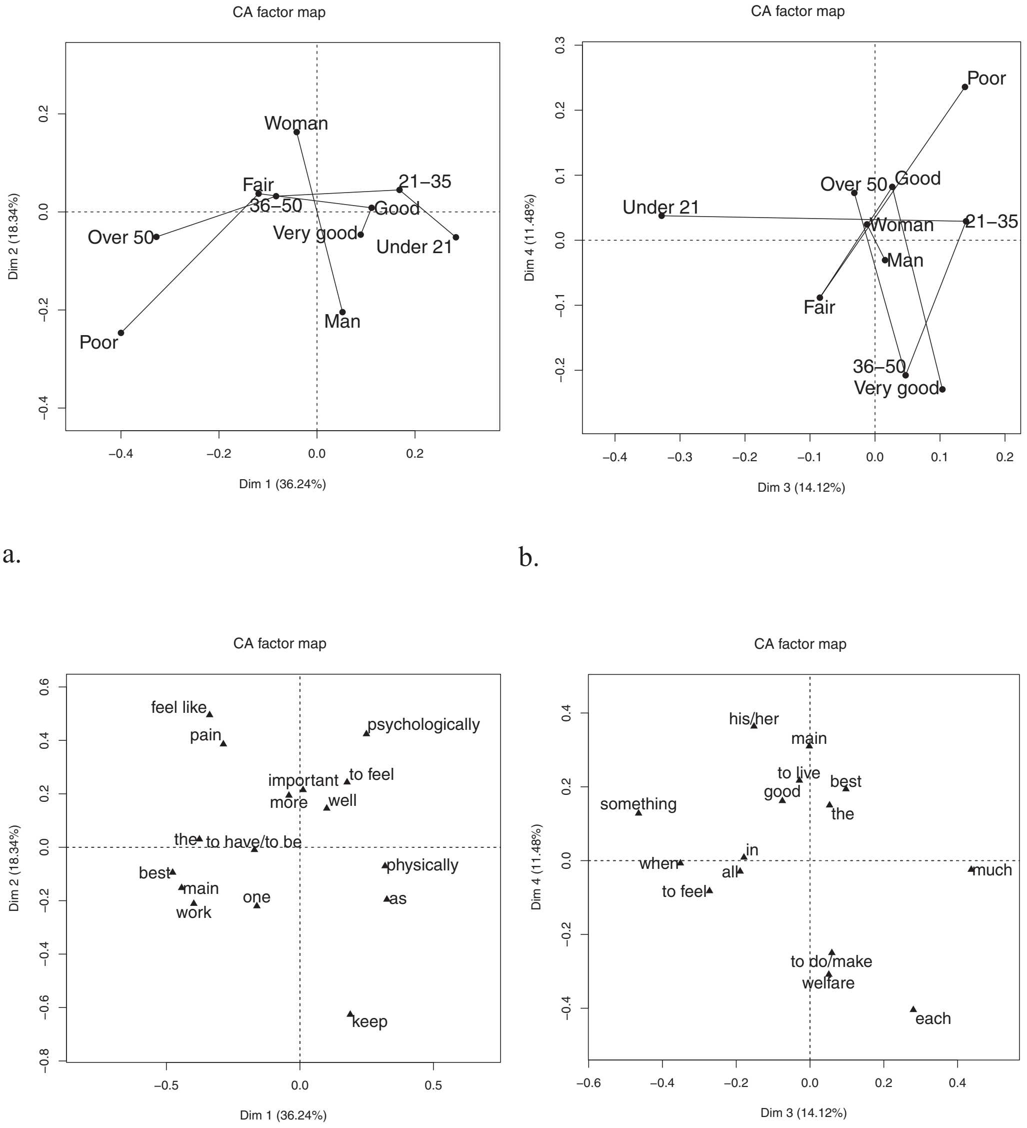
Analysis	Φ^2	Cramer's V^2	λ_1
CA on the lexical table	7.145	0.044	0.246
CA on the by *age* aggregated lexical table	0.106	0.035	0.063
CA on the by *health condition* aggregated lexical table	0.071	0.024	0.033
CA on the by *gender* aggregated lexical table	0.038	0.038	0.038

widely shared by all the speakers of the same language. The individual variability, as measured by the Φ^2, is huge but manifested through a multiplicity of loosely structured syntagamatic associations. The aggregation of the free answers leads to a weak loss in terms of intensity of the relationship, as evaluated by the Cramer's V^2, despite the huge decreasing of the inertia. What is lost is mainly the non-structured part of the inertia. Thus, the Cramer's V^2 only decreases from 0.044 to 0.038/0.035 when aggregating the free answers by *gender/age group* while the total inertia Φ^2 dramatically lessens. A slightly more pronounced lowering of the Cramer's V^2 is observed when aggregating the free answers by *health condition*.

The direct analysis visualises the relationships between respondents and words on classical CA graphs (not reproduced here). In this case, the projection of the categories at the centroid of the respondents belonging to them shows that a relationship between the three categorical variables and the vocabulary does exist. The two gender categories are opposed on the first axis while the second axis ranks *age* and *health condition* categories in their natural order. The significance of the positions of the categories is assessed through classical tests (Lebart et al., 1998, pp. 123-128). However, the strong association between *age* and *health condition* trajectories makes it difficult to untangle their real influence on the word choices. We can nevertheless report that the *age* trajectory is more elongated than *health condition* trajectory and that *poor health* lies in a position that distinguishes the *over 50* category from others. This analysis merges the non-explained individual variability, which is always huge in the case of the direct analysis of free-answers, and the variability explained by the multiple belonging to categories of several variables. Therefore it is necessary to complete this initial analysis by others focusing on possible specific associations between categories and words. That being said, this first step can be very useful to suggest interesting grouping variables.

6.3. Classical CA on the multiple aggregated table

CA is applied to the multiple aggregated table. The total inertia is equal to 0.072. *Age group*, *health condition* and *gender* contribute to this total inertia bringing, respectively 49.4%, 33.0% and 17.6% of this total inertia. The first two axes, whose inertia are respectively 0.026 and 0.013, keep together 54.6% of the total inertia.



a.

b.

c.

d.

Figure 1: *Categories and contributory words on the CA planes (1,2) and (3,4).*

Figure 1.a offers the representation of the categories on the plane (1,2). The trajectory of *age group* categories notably follows the first axis, outlining a weak arch effect. This axis ranks, in their natural order, the *health condition* categories except for the inversion between *very good health* and *good health* which lie very close. The extreme categories of this variable, very particularly *poor health*, are opposed to the intermediate categories on the second axis, indicating a more pronounced arch effect than *age group*. However, the main opposition on the second axis concerns the two *gender* categories

Table 2: *age group* × *health condition frequency of occurrence table.*

	Health condition				
Age group	Poor	Fair	Good	Very good	*margin*
Under 21	0	19	44	8	71
21-35	2	37	81	15	135
36-50	2	34	33	7	76
Over 50	21	49	36	4	110
margin	25	139	194	34	392

so that *age group* and *gender* are practically orthogonal, this in terms of the vocabulary that they use. Regarding the plane (3,4) (Figure 1.b), the third axis shows that young people (*under 21*), besides using words close to those used by the following age group as revealed by axis one, also express themselves with their own words. No clear pattern stands out on the fourth axis.

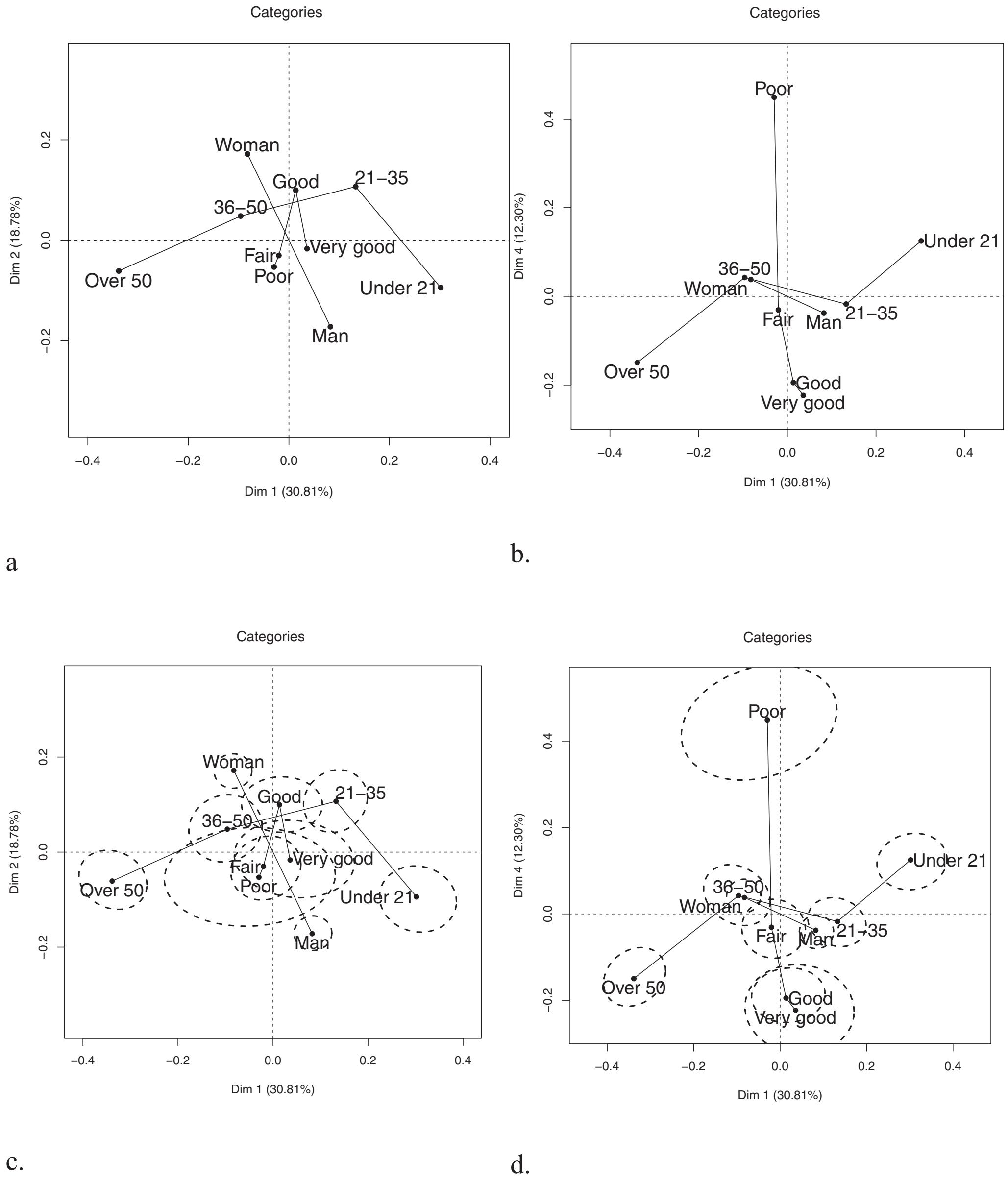
The words representation (Figures 1.c and 1.d) brings information about the meaning of the oppositions and trajectories, showing for example that the words *the*, *best*, *main* – used in expressions such as (health is) *the best*, *the main* (thing) – and *work* are words both used by the oldest and/or less healthy categories and avoided by the youngest and/or more healthy categories. However, one might wonder if the choice/rejection of these words is related to *age* or to *health condition* or to both.

Table 2 shows that *age group* and *health condition* are strongly associated but still that the association is sufficiently loose as to allow for untangling the influence of both variables on the vocabulary, provided that an adequate method is applied. Precisely, CA-GALT offers a suitable approach because the associations between the variables are cancelled.

6.4. CA-GALT on the multiple aggregated table

CA-GALT is applied on the multiple aggregated table. The total inertia is equal to 0.2067. The first two axes are moderately dominant with eigenvalues equal to 0.0636 (30.81% of the inertia) and 0.0388 (18.78%).

Figures 2.a and 3.a display, respectively, the contextual variables and the words with a high contribution on the CA-GALT first principal plane. These representations are completed by drawing confidence ellipses (Efron, 1979; Lebart et al., 2006). Only the confidence ellipses around the words *the*, *best*, *main* and *work* are represented, because these words are favoured as examples to show the effectiveness of the approach that we propose (Figures 3.c and 3.d). If all the ellipses were drawn, only those around *he/she* and *to be able*, on plane (1,2) and around *to be able* and *from* on plane (1,4) would overlap the centroid.



a

b.

c.

d.

***Figure 2:** Categories on the CA-GALT planes (1,2) and (1,4) completed by confidence ellipses.*

As in the former analysis, the trajectory of *age group* notably follows the first axis. The extreme categories of this variable, *over 50* (at the left); *under 21* (at the right) bring, respectively, 52.1% and 23.1% of this axis inertia. However, *health condition* representation differs. The categories of this variable now lie close to the centroid on the first plane and their confidence ellipses extensively overlap one another (Figure 2.c). Regarding the words, we find again *the*, *best* and *main* with high coordinates at the left

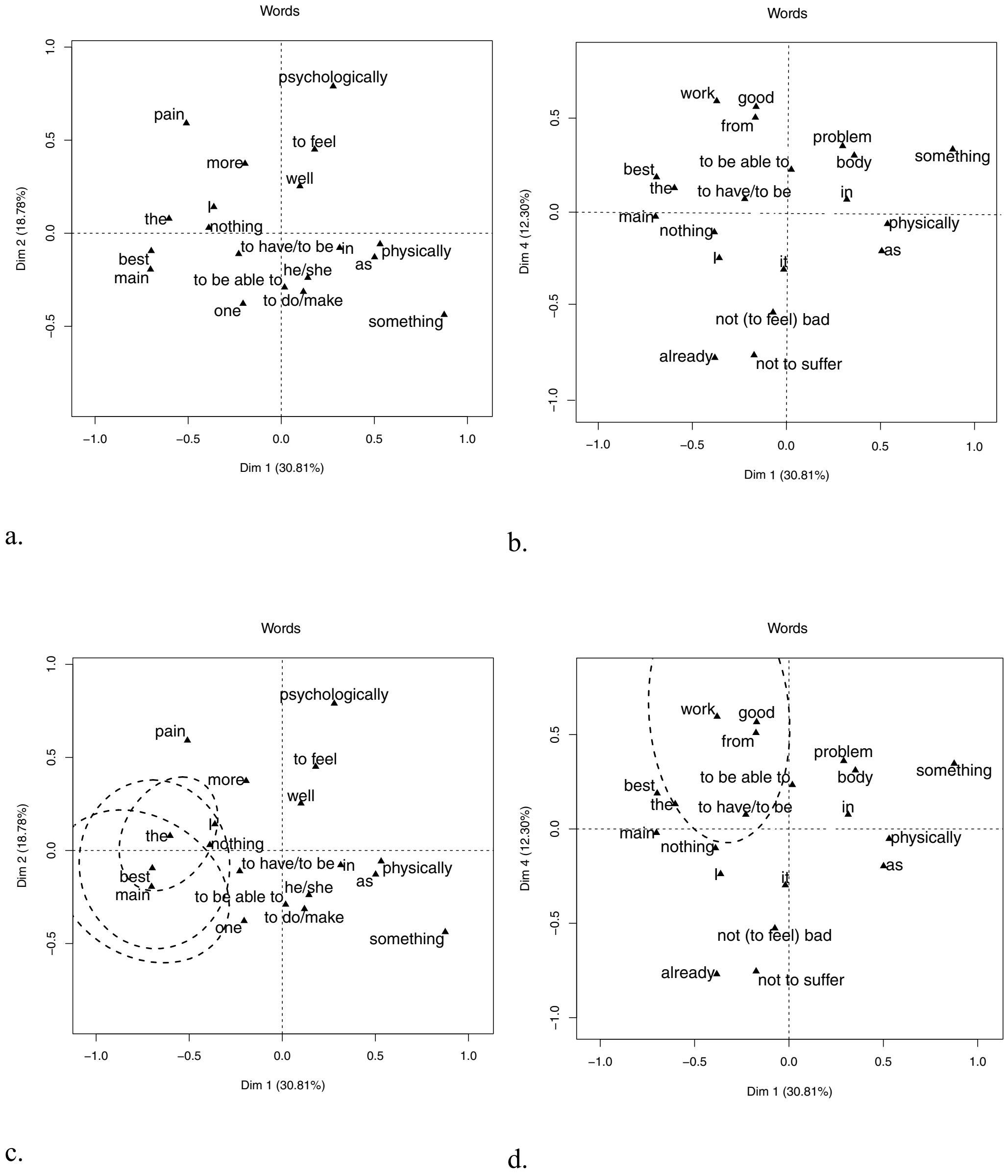


Figure 3: *Contributory words on the CA-GALT planes (1,2) and (1,4) completed by confidence ellipses.*

of the axis, indicating that they are words both very used by the oldest categories and avoided by the youngest. The word *work* is no more present on this graphic, since it is close to the centroid and thus not a key word for the oldest categories.

We detail neither the second axis, which opposes *Man* and *Woman*, nor the third (not reproduced in the graphic), which highlights the specific use of the vocabulary by the *under 21* respondents. Both axes are close to those computed in the former CA.

The fourth axis turns out to be of interest because of ranking *health condition* categories in their natural order (Figure 2.b). These categories, which provide together 75% of the axis inertia, are well separated, except for the two better health categories whose confidence ellipses overlap (Figure 2.d). The word *work* lies close to *poor health* category in the positive part of the fourth axis (Figures 2.b and 3.b), pointing out a strong association between this word and this category and also little use of this word by the most healthy categories. The word *work* contrasts on the fourth axis with *bad*, *suffer* and *already* which are associated with *good health, very good health* and *over 50.* These latter words are used in free answers where health is defined through negative expressions such as *not to feel bad*, *not to be bad*, *not to suffer, not to suffer from any disease* or *pain.*

The discrimination between the words associated with *poor health* and those associated with *over 50* that CA-GALT uncovers has to be checked in the data. The variable crossing *age group* and *health condition* is created but grouping similar categories to ensure a minimum membership in every category. This cross-variable allows for comparing the vocabulary from the *health condition* viewpoint at a same age and vice-versa. For each category, the moderate/significant under/over use of the words can be computed from using the test proposed by Lebart et al. (1998, chapter 6). We conserve not only the significant under/over used words (p-value<0.1) but also the moderate under/over used words ($0.1 < p$-value < 0.16), because the progression of the use of a word depending on *age* increasing, and of *health condition* decreasing, is also of interest. The results corresponding to the four words (*the*, *best*, *main* and *work*) are summarised in Table 3.

Table 3: *Categories under or over using the words* the, best, main *and* work.

	Significant under-use	*Moderate under-use*	Moderate over use	Significant over-use
Word				
the	<21-good/very good health <21 fair heath 21-35 fair/poor health		>50 good/very good health	>50 fair health >50 poor health
best	<21-good/very good health	21-35 fair/poor health	>50 good/very good health	>50 poor health
main	21-35 fair/poor health			>50 good/very good health >50 poor health >50 fair health
work	21-35 good health			>50 poor health 36-50 poor/fair health

Table 3 shows that the respondents who do not or barely use the words *the*, *best* and *main* (in expressions such that "*the best/ the main thing*") differ from those who overuse from the age viewpoint. These words are moderately or significantly overused only by the *over 50* category of respondents with very different health conditions. These three words usage depends on *age*, not on *health condition*, and increases with the former.

We now focus on the word *work,* significantly under-used by the *21-35* with *good health* and significantly overused by the *over 50* with *poor health* and the *36-50* with *poor* or *fair health*, which is the less healthy category for this age group. These results are not as obvious as the former results to allow us to conclude on the effect alone of *health condition* on the selection/rejection of this word. It is more difficult to untangle the influence of *age* and *health condition* in this case because the *poor health* category is almost made up of only *over 50* respondents (21 from 25). Nevertheless, the *over 50* and *36-50* not presenting the worst health condition corresponding to their own age do not over use *work*, even only moderately. This allows for concluding that *health condition* has, at least, a much stronger effect than *age* on the selection of this word.

We can finish telling that CA-GALT is able to untangle the complex influence of *age* and of *health condition* on the lexical choices from differences existing in the data through a *ceteris paribus* analysis.

7. Conclusion

The direct analysis of the lexical table offers valuable visualizations of the associations among the respondents and among the words that also indicate the relationships between respondents and words (Lebart et al., 1998). However, it is necessary to go further and identify the complex relationships between respondents' characteristics and lexical choices. The inclusion of selected categorical variables as explaining variables in the analysis highlights these relationships provided that all the main sources of variability are taken into account. This leads to consider a multiple aggregated lexical table, juxtaposing the aggregated tables built from each selected categorical variable. A specific CA, called CA-GALT, analyses this table while keeping the double projected approach that CA offers. CA-GALT studies the diversity of the vocabulary through the dispersion of the categories and the dispersion of the categories through the diversity of the vocabulary. Thus, the associations and/or oppositions between words acquire their meaning from the categories that they attract or reject and vice versa. The application of the method to a real data set has demonstrated how free-text and closed answers combine to provide relevant information. The influence of each variable on the lexical choices is visualised, avoiding "confusion effect". The words favoured by the different categories uncover the health-associated concerns related to each variable (*age, health condition* or *gender)* in a *ceteris paribus* analysis.

Software note

The R function CaGalt (Correspondence Analysis on Generalised Aggregated Lexical Table) has been developed by the authors. This function will be included in the next release of package FactoMineR (Husson et al., 2007; Lê et al., 2008). Meanwhile, this function can be requested from the authors.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness. Grant ECO2012-35584.

References

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

Benzécri, J.-P. (1973). *L'Analyse des Données. Tome I. L'Analyse des Correspondances*. Dunod, Paris.

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

Böckenholt, U. and Takane, Y. (1994). Linear constraints in correspondence analysis. In: Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*. Academic Press, London.

Brandimarte, P. (2011). *Quantitative Methods: An Introduction for Business Management*. John Wiley & Sons, Inc., Hoboken, USA.

Cousteaux, A.-S. (2010). Représentations de la santé et cycle de vie. De la recherche du bien-être au maintien des capacités. *OSC-Notes et Documents N^o 2010-01.* http://www.sciencespo.fr/osc/sites/sciencespo.fr.osc/files/nd_2010_01.pdf.

Efron, B. (1979). Bootstrap methods: another look at jackknife. *The Annals of Statistics*, 7, 1–26.

Escofier, B. and Pagès, J. (2008). *Analyses Factorielles Simples et Multiples*, 4th Ed. Dunod, Paris.

Finn, J. D. (1974). *A General Model for Multivariate Analysis.* Holt, Rinehart and Winston, New York.

Garnier, B. and Guérin-Pace, F. (2010). *Appliquer les méthodes de la statistique textuelle*. CEPED, Paris.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York. (Out-of-print, freely downloadable from www.carme-n.org)

Härdle, W. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*, 3rd Ed. Springer Verlag, Heidelberg.

Husson, F., Josse, J., Lê, S. and Mazet, J. (2007). *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.19.

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

Lebart, L., Piron, M. and Morineau, A. (2006). *Statistique exploratoire multidimensionnelle: Visualisation et inférence en fouilles de données*, 4th Ed. Dunod, Paris.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*, 2nd Ed. Elsevier Science, Amsterdam.

Lê, S., Josse, J. and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analisis. *Journal of Statistical Software*, 25, 1–18.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall, Boca Raton.

Saporta, G. (2006). *Probabilités, Analyse des Données et Statistique*, 2nd Ed. Technip, Paris.