Analysis of multilingual free responses

Mónica Bécue[1], Jérôme Pagès[2], Campo-Elias Pardo[3]

[1]EIO. Universitat Politècnica de Catalunya – 08028 Barcelona – Spain.
monica.becue@upc.es

[2]ENSA/INFSA – 65 rue de Saint-Brieuc, CS 84215– 35042 Rennes cedex – France
jerome.pages@agrorennes.educagri.fr

[3]Departamento de Estadística. Universidad Nacional de Colombia – Bogotá – Colombia
cepardot@unal.edu.co

Abstract

The analysis of international survey data leads to deal with cross language free responses. To conserve a CA-like approach, which is a reference methodology to analyse open-ended questions, we propose to apply multiple factor analysis for contingency tables. This methodology allows for representing the whole of the lexical tables in a same reference space, keeping CA-like features, in particular transition formulae and interpretation rules. To illustrate the interest of this approach, we use an example extracted from a large international survey in four countries. [1]

Keywords: textual similarity, texts visualisation, cross language open-ended questions.

1. Introduction

The analysis of international survey data with close and open-ended questions leads to deal with free responses expressed in different languages by different samples. A reference methodology to analyse open-ended questions is correspondence analysis (CA) (Lebart *et al*., 1998), applied to lexical tables, aggregated or not (categories_of_individuals × words or individual × words tables). In this work, we want to extend this kind of methodology to several languages in order to tackle multiple aggregated lexical tables.

In textual statistics field, previous works by Lebart (Akuto and Lebart, 1992; Lebart, 1995 and 1998) have proposed a principal axes method to tackle free responses in different languages, allowing for superimposed representations of the categories, as induced by every sample, in a same referential space.

In information retrieval field, various methods have been developed to deal with multilingual corpus. We can cite CL-LSI (Littman *et al.*, 1998) and kernel correlation analysis (KCAA) (Vinokourov *et al*., 2003) which combines LSI (Latent Semantic Analysis which operates a SVD applied to the rough document × term matrix) or ICA (Independent Component Analysis) and canonical correlation analysis for training multilingual retrieval information tools on aligned corpus. These methods allow for representing and comparing the configurations of the words, as induced in every corpus, and searching for equivalent words, in terms of translation, through local similarities among the configurations.

[1] This work has been partially supported by the European project NEMIS (IST- 2001 – 37574/ Information Society Technology Program) and by a grant from National University of Bogotá.

We propose here to use multiple factor analysis for contingency tables (MFACT) which combines features from principal axes methods, canonical analyses and Procustes analysis (Bécue and Pagès, 2004), allowing us to deal with the whole of the responses without translation, relegating this operation to a further step. As in CA, this method offers a visualisation of the distances between categories and also a visualisation of the distances between the whole of the words, meaningful even when the words belong to different languages, being both representations linked by transition rules. Concerning the categories representation, a global representation, as induced by the whole of the answers, and partial representations, as induced by every sample, are obtained, being these representations situated in a same referential space.

2. Data

The data are extracted from a large international survey (Hayashi *et al*., 1992)[2]. People from four countries (Great Britain, France, Italy, Japan) are asked several closed questions and, moreover, the following three open-ended questions:

1 - " What is the most important thing to you in life?"

2 - " Anything else ?"

3 - " What does the culture of your country mean to you?"

The Japanese answers are romanised. We analyse the answers to the first two questions, gathered and considered as a unique response.

In each country, the free answers are grouped into 18 category-documents by crossing gender (male, female), age (into three categories: 18-34, 35-44, 55 and over) and education level (into three categories: low, medium and high). Then, for each country, from the count of words in the whole answers, the lexical table arises by crossing the 18 documents and the most frequent words. Only the words used at least 20 times are kept, this low threshold having been chosen in respect to the 18 categories of respondents.

3. Objectives

The general objective is to compare the structures induced on the categories-rows by the word-columns in the different countries and to detect which categories are similar whatever the countries or not.

Although the word-columns are different for each sample, many of these are the translation of a same word in the different languages. So, we have a special interest in studying their mutual position.

4. Methodology

4.1. *Notation*

The answers given in language t are codified as a (I, J_t) table Y_t in which the element at row i, column j is the relative frequency of word j in the category-document i, as calculated in respect to the sum of terms of the whole of the tables. Word-columns are not the same through the tables. We have T tables (T=4). These tables are juxtaposed row-wise, into a multiple lexical table Y. We denote f_{ijt} the relative frequency, in table Y_t (=country t; t =1,...T), with which row i (=document i ; i=1,...I) is associated with column j (=word j ; j=1,...J_t;

[2] We thank Profesor Lebart to put at our disposal this data set.

$\sum_t J_t = J$). Thus, $\sum_{ijt} f_{ijt} = 1$. We note $f_{i..} = \sum_{jt} f_{ijt}$ the row margin of the table Y, $f_{.jt} = \sum_i f_{ijt}$ the column margin of the table Y, $f_{i.t} = \sum_j f_{ijt}$ the row margin of the subtable Y_t, $f_{..t} = \sum_{ij} f_{ijt}$ the sum of the terms of table Y_t. In the example, T=4; I=18; J_1=106, J_2=96, J_3=55, J_4=48.

4.2. *Analysis of the multiple lexical table*

4.2.1. *Separate analyses*

First, CA is applied to each table Y_t in order to have a first information about common features in the four structures (Figure 1). Then, for every t, t=1,..,T, separate modified margins CA are performed, imposing the row margin $\{f_{i..}, i=1,...,I\}$ and the column margin $\{f_{.jt}, j=1,...J\}$. The first eigenvalue of these pseudo-separate CA, denoted by $\lambda_1{}^t$, will be used in the global analysis to balance the influence of the groups. Performing this CA, with modified margins, is equivalent to perform a general principal axes method on the table with general term given by:

$$\frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) \cdot f_{.jt}}{f_{i..} f_{.jt}} \qquad (1)$$

using the weights ($f_{i..}$) for the rows (and metric in the columns space) and the weights ($f_{.jt}$) for the columns (and metric in the rows space). In such a way, the rows keep the same weight through all the analyses.

4.2.2. *Global analyses through MFACT*

MFACT (Bécue and Pagès, 2003) consists in a multiple factor analysis (Escofier and Pagès, 1988-1998) extended to contingency tables: a general principal axes method is performed on the juxtaposition of the T tables, with general term given by (1), giving the weight $f_{i..}$ to row i and the weight $f_{.jt}/ \lambda_1{}^t$ to the column (j,t). This method offers results:

- common to all principal axes methods, mainly a global representation of the rows and columns;
- specific to multiple tables, mainly the superimposed representation of the structures induced on the categories by the words in each country, called partial structures, and the representation, in a same reference space, of all the factors obtained in the separate analyses.

5. Results

5.1. *Brief description of the four corpus*

The whole of the free answers of every country constitutes a corpus, identified in the following using the name of the country. Table 1 summarises the main characteristics of the four corpus.

Corpus	Individuals	Corpus length	Distinct words	Kept corpus length	Kept length (%)	Kept distinct words
U. Kingdom	1043	13912	1357	10658	76.6	106
France	1009	14206	1248	11309	79.6	96
Italy	1048	6154	788	4443	72.2	55
Japan	2265	6962	697	5462	78.5	48

Table 1. Main characteristics of the four corpus (Frequency threshold equal to 20)

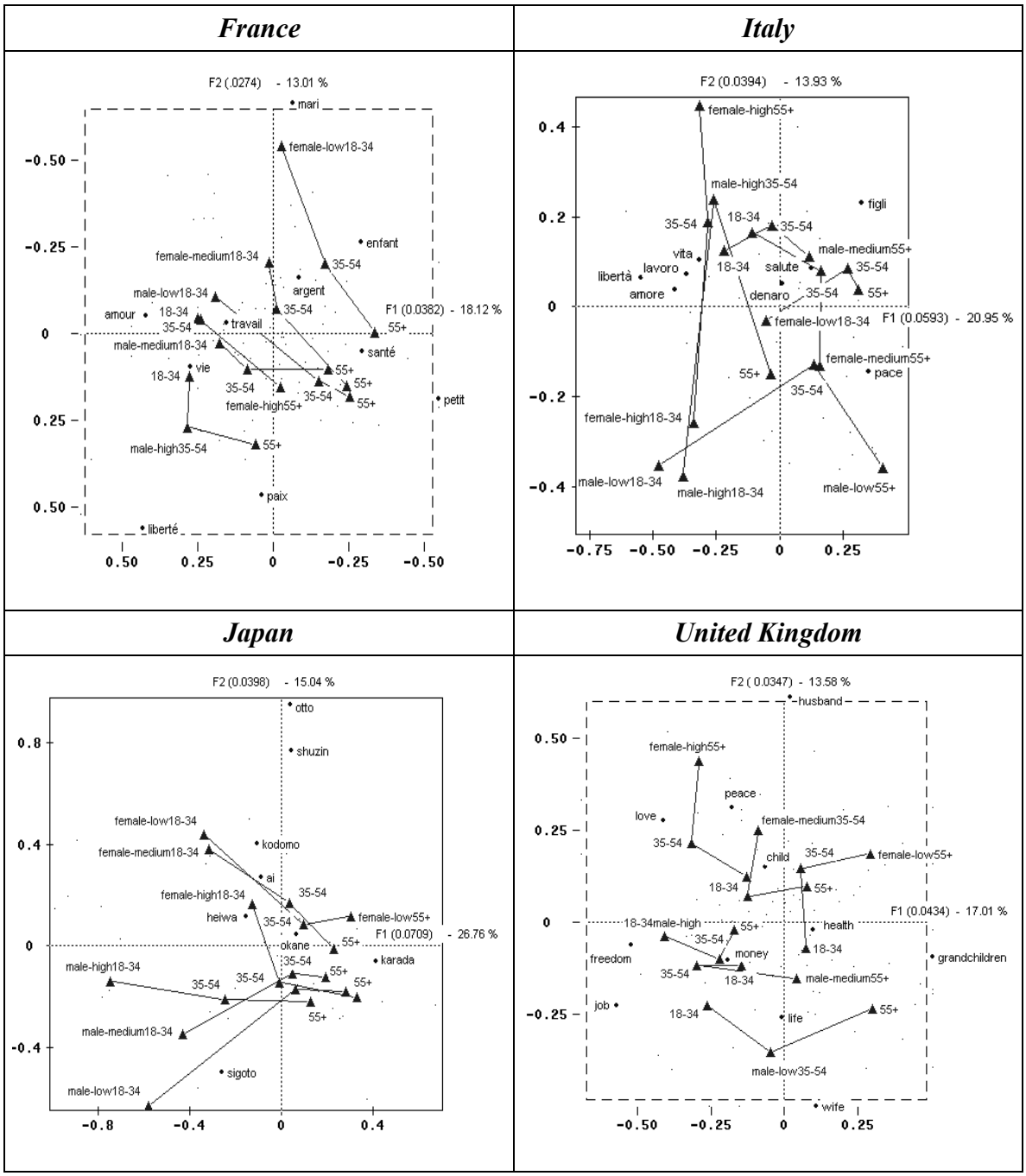

Figure 1. Separate CA in the four countries with trajectories of age, fixing gender and qualification levels

Through the reading of the four lists of words, the frequent topics are identified. Most of them are common to all the countries and are expressed using equivalent words in that sense that they correspond to a same word translated in the four languages.

5.2. *Separate CA*

Figure 1 visualises the four first principal planes, corresponding to the separate CA, in which are drawn the six trajectories of age, fixing gender and education levels.

	Inertia	First eigenvalue	Second eigenvalue	% of inertia kept on the first plane
UK	0.2546	0.0443 (17.4%)	0.0340 (13.3%)	30.7%
FR	0.2149	0.0392 (18.3%)	0.0266 (12.5%)	31.0%
IT	0.2848	0.0570 (20.0%)	0.0415 (14.6%)	34.6%
JA	0.2700	0.0730 (27.0%)	0.0401 (14.9%)	41.9%

Table 2. Inertia and first eigenvalues of the separate CA

The age × gender trajectories present common aspects through all the countries. In the cases of United Kingdom and Japan, age increases along the first axis, while the second axis opposes both genders. In the case of France, there is a structure somewhat similar, but rotated: age increases along the second bissector and the first bissector opposes both genders. Italy looks to be more peculiar, and does not offer, in this first plane, any opposition between genders.

5.3. *Global analysis*

The two first eigenvalues stand out against the following ones: λ_1= 3.52 and λ_2=2.03 (respectively, 18.2% et 10.5% of the total inertia). The high value of the first global eigenvalue (its maximum is equal to 4), indicates that the first global axis is a dispersion direction close to the first separate axis of each subtable, although it is not mixed up.

Table 3-a shows that each corpus contributes, in a balanced way, to the inertia of the first factor. Concerning the second factor, the contribution of Italy is weak, while United Kingdom and France supply the two biggest contributions. The correlations between the first global factor and the projections of the four categories-clouds are strong for all the countries. Concerning the second axis, the correlation is strong with the projections of UK, France and Japan clouds, weaker but nevertheless high in the case of Italy (Table 3.b).

	F1	F2
Total inertia	3.52	2.03
U. Kingdom	0.80	0.76
France	0.93	0.58
Italy	0.88	0.25
Japan	0.90	0.44

Table 3.a. Decomposition per country of the inertia of the two first factors

	F1	F2
U. Kingdom	0.93	0.95
France	0.98	0.93
Italy	0.97	0.81
Japan	0.96	0.90

Table 3.b. Correlations between the projections of the global cloud and the ones of the four partial clouds

It can be concluded that the two first factors are common to the four clouds. The first factor is an important dispersion direction for every country and the second factor is an important dispersion direction for UK, France and Japan, and not so important for Italy.

The visualisation of the categories obtained by MFACT is given by Figure 2. The 6 trajectories of age intervals show a rather regular structure, compromise between the representations offered by the separate CA. Age increases along the first axis, and the second axis opposes both genders, structure similar with that we found in United Kingdom, France and Japan in the separate CA. We can note that the categories with the high education degree have, on the first axis, coordinates which correspond to younger people with lower degrees.

5.3.1. *Interpretation of the proximities between words*

Proximities between words can be interpreted as a resemblance between their users. More precisely, the squared distance between word *j* (belonging to table *t*) and word *k* (belonging to table *r*) can be written in the two following ways:

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}} \right) - \left(\frac{f_{ikr}}{f_{.kr}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2 \quad (2)$$

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{ikr}}{f_{.kr}} \right) - \left(\frac{f_{i.t}}{f_{..t}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2 \quad (3)$$

Case 1: the words belong to a same table (t = r)

The proximity between two words is interpreted in term of resemblance between profiles exactly as in the usual CA.

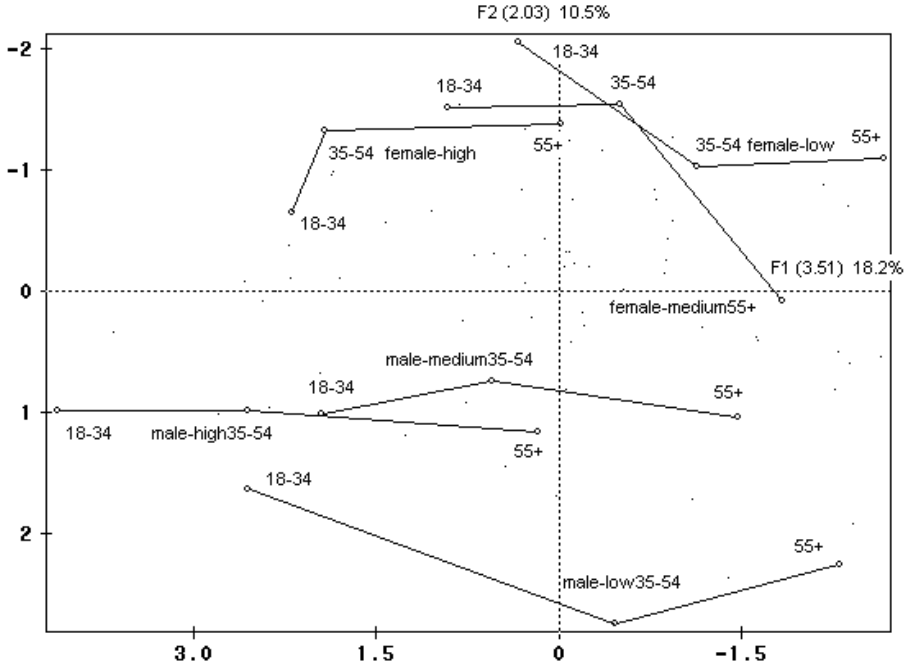

Figure 2 .Global representation of the categories

Case 2: the words belong to different tables (t ≠ r)

The word profiles are relativized by the average profiles, as shown in the two expressions of the squared distance. Expression (2) shows that the profile of a word $\{f_{ijt}/f_{.jt}\ i \in I\}$ intervenes by its deviation from the average profile of the corresponding table $\{f_{i.t}/f_{..t}\ i \in I\}$. Expression (3) shows how the differences between word profiles are relativized by the differences between average profiles, which is important when, as it is the case, the row margins differ from one subtable to another For example, in Italy, the answers of the women between 35 and 54 with a low level of education constitute 4.1% of the Italy corpus, when in United Kingdom corpus the answers of the same category constitute the 10.9% of its length. In fact, the categories have different counts in the four countries.

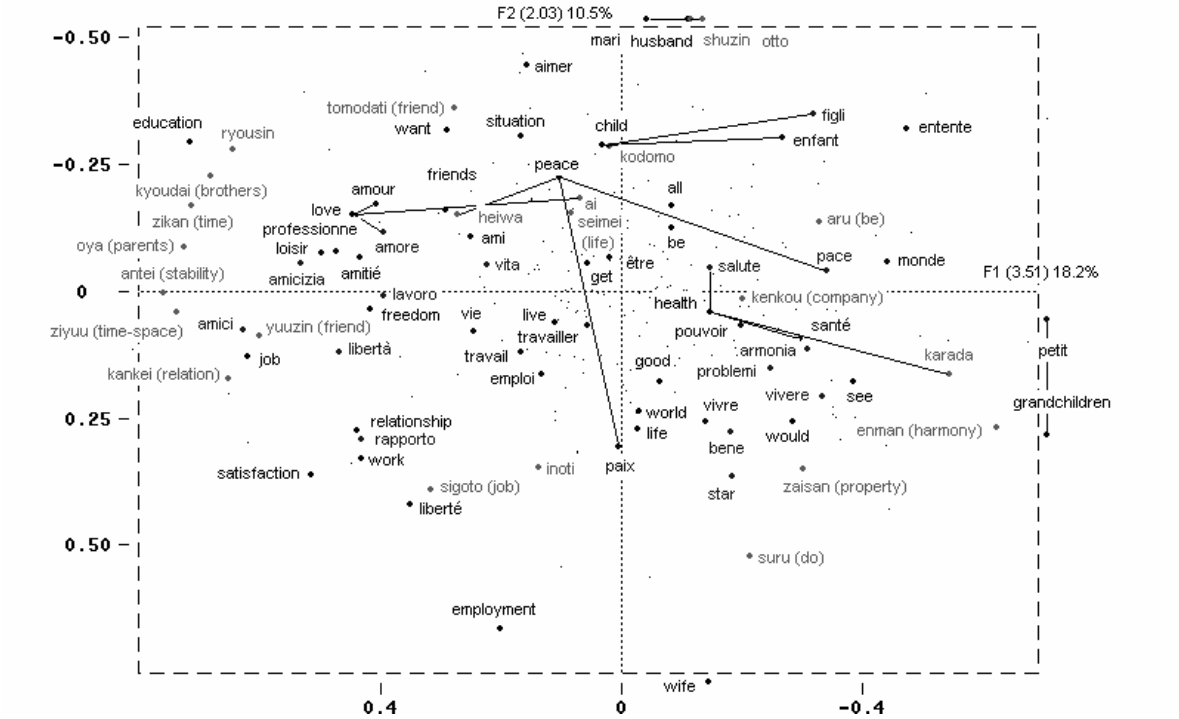

Figure 3. Global representation of the words

Thus, in the example, according to Figure 3, *love*, *amour, amore* are rather quoted by the same categories, when *ai* lies close to the centroid. Consulting the data, we can see that *love, amour, amore* are used, above all by young and medium age individuals (see Table 4) and much more fewer by individuals over 55, while *ai* has a nearly uniform distribution. On the contrary, *peace, paix, pace* and *heiwa* lie in rather distant positions, that corresponds to their different profiles (Table 4).

	Males			Females		
Words	**<35**	**35-54**	**>54**	**<35**	**35-54**	**>54**
love	9 (27.3%)	3 (09.1%)	2 (06.1%)	4 (12.1%)	9 (27.3%)	6 (18.2%)
amour	17 (19.3%)	11 (12.5%)	6 (06.8%)	27 (30.7%)	21 (23.9%)	6 (06.8%)
amore	46 (25.8%)	20 (11.2%)	13 (07.3%)	50 (28.1%)	30 (16.9%)	19 (10.7%)
ai	1 (04.3%)	4 (17.4%)	4 (17.4%)	5 (21.7%)	5 (21.7%)	4 (17.4%)
peace	12 (15.2%)	5 (06.3%)	11 (13.9%)	13 (16.5%)	21 (26.6%)	17 (21.5%)
paix	9 (20.0%)	6 (13.3%)	13 (28.9%)	6 (13.3%)	5 (11.1%)	6 (13.3%)
pace	12 (13.5%)	9 (10.1%)	18 (20.2%)	11 (12.4%)	15 (16.9%)	24 (27.0%)
heiwa	4 (05.3%)	15 (19.7%)	13 (17.1%)	14 (18.4%)	26 (34.2%)	4 (05.3%)

Table 4. Distribution of the words corresponding to love and peace

This fact suggests two kinds of interpretation: the first is linguistic natured (the translation reflects a meaning analogy but not a custom analogy), the second is sociologic natured (the different categories do not have the same values from a country to another).

5.3.2. *Superimposed representation of the partial clouds*

In order to compare the structures induced on the categories by the four clouds of words, we superpose the global description of the categories and those induced by each columns groups.

(partial categories). It allows for pointing out the convergences and divergences between the countries.

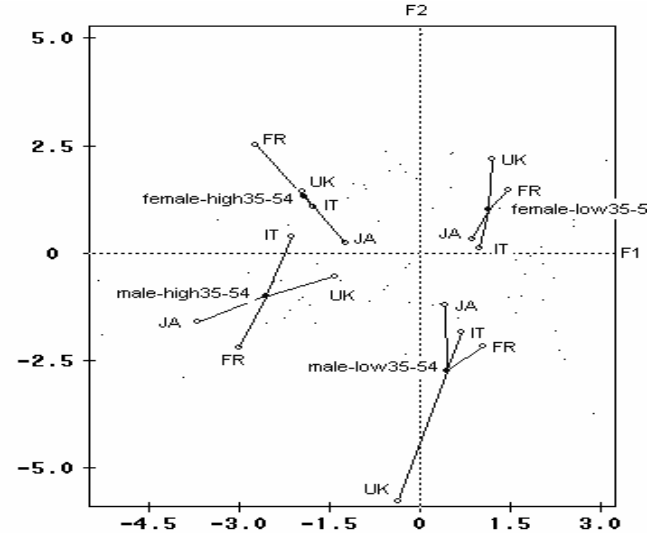

Figure 4. Excerpt of the superimposed representation

For example, the excerpt of the superimposed representation of the partial clouds presented in Figure 4 suggests that that males and females in 35-54 age interval, with high or low qualification, quasi do not differ in Italy. This fact was already pointed out on the separate first planes (Figure 1).

6. Conclusion

The analysis gives a synthetic vision of the diversity of the words used depending on age, gender and qualification. It offers cross language representations of the categories, balanced compromise between the monolingual (or separate) representations, and of the words, being both representations linked by transition rules, which allows us to relate the similarities and dissimilarities between categories to vocabulary and vice-versa. The distances between words from different countries are meaningful and can be interpreted from users profiles. Finally, the superposed representation of the partial categories on the same reference space than the global categories enables to interpret the divergences among homologous categories from a vocabulary point of view.

References

Akuto H. and Lebart L. (1992). Le repas idéal. Analyse de réponses libres en anglais, français, japonais. *Les Cahiers de l'Analyse des Données*, vol. (17/3): 327-352.

Bécue M. and Pagès J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Comp. Statistics and Data Analysis* (to be published).

Escofier B. and Pagès J. (1988-1998). *Analyses factorielles simples et multiples; objectifs, méthodes et interprétation*. Dunod.

Hayashi C., Suzuki T. and Sasaki M. (1992). *Data Analysis for Social Comparative research: International Perspective*. North Holland.

Lebart L. (1995). Assessing and comparing patterns in multivariate analysis. In Hayashi C. *et al*. (Eds), *Data Science and application*. Academic Press.

Lebart L. (1998). Text mining in different languages. *Applied Stochastic Models and Data Analysis*, vol. (14): 323-334.

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

Littman M.L., Dumais S.T. and Landauer T.K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette G. (Ed.), *Cross language information retrieval.* Kluwer.

Vinokourov A., Shawe-Taylor J. and Cristianini N. (2002). Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In *Oral presentation at the 15th NIPS.*