DETERMINING THE AGE OF ACQUISITION OF WORDS FOR A RECOGNITION TEST USING EXPLORATORY DATA ANALYSIS

Sergio Camiz
Dipartimento di Matematica Guido Castelnuovo
Sapienza Università di Roma
Piazzale Aldo Moro 5, I - 00185 Roma Italia
sergio.camiz@uniroma1.it

Gastão Coelho Gomes
DME – IM – UFRJ,
Caixa Postal 68530 cep:21945-970, RJ
gastão@im.ufrj.br

Fernanda Duarte Senna
Doutoranda Programa de Pós-Graduação em de Linguística, UFRJ
fonofernandasenna@gmail.com

Christina Abreu Gomes
Departamento de Linguística, UFRJ
Cidade Universitária, RJ
christina-gomes@uol.com.br

RESUMO

Neste trabalho discutimos os primeiros resultados de 161 palavras remanescentes de um estudo prévio de julgamento de imagens, especificamente no que concerne a comparação entre dois conjuntos de julgadores aos quais as palavras foram submetidas para serem avaliadas de acordo com duas escalas diferentemente sugeridas de idade de aquisição, uma contendo faixas de idade e outra sendo uma escala livre. Foi estudado o efeito dos tipos de escala em função das classes de idade. Análise de Correspondências, Fatorial e de Componentes Principaisforam utilizadas para a análise dos dados. Os resultados indicaram uma concordância entre as duas escalas pelo menos no que diz respeito ao primeiro componente principal que deve ser tomado como uma medida da primitividade de idade de aquisição. Isso representa em torno de 50% da informação resultante das observações, sendo o restante distribuído em outros fatores menoresque não foram estudados em detalhe até o momento.

PALAVARAS CHAVE. Anáise de Correspondêcias, Anáise Fatorial, Idade de Aquisição

ABSTRACT

In this paper we discuss the first results carried out on 161 words selected in a previous study on images judging, specifically in what concerns the comparison between two sets of judges. The words were submitted asking to score them according to two differently suggested scales of age of acquisition: a free-scale one and another with the suggested slots of ages. The effect of the scale constraints given by the suggested age-classes was studied. Principal Component Analysis, Multiple Correspondence Analysis, and Multiple Factor Analysis were used to analyse the data. The results indicate a good agreement between the two scales, at least as far as it concerns a first principal component, but with differences concerning the scale range used. This represents only around 50% of the information and may be taken as a measure of primitiveness of age acquisition.

KEYWORDS. Correspondence Analysis, Multiple Factor Analysis. Age of Acquisition.

1. Introduction

In this paper we report the results of an evaluation of the age of acquisition of words related to images, to be used in tests with aphasic patients, in which these are asked to verbalize objects submitted to them through specific images. The *age of acquisition* is an important variable related to Jacobson's hypothesis according to which linguistic structures acquired earlier tend to be kept and the structures acquired later tend to be lost under brain damage. Thus, establishing the *age of acquisition* of some words, a measure of how early in his/her life a person got in touch with the word and included it in his/her vocabulary, is an important step for the analysis. The words were selected from a previous study that considered the convergence in naming 260 images by 38 subjects, in order to understand to what extent they identify the image with the same name defined by the researcher. In this paper we discuss the first results for the 161 words selected by a previous study of judging images, specifically in what concerns the comparison between two sets of judges to which the words were submitted, to score according to two differently suggested scales of age of acquisition.

2. The data

161 words were selected from a previous study among 260 images, whose name was defined by the researcher, and whose images were submitted to a panel of 38 judges to check an agreement on their naming. Then 128 subjects were asked to estimate the degree of primitiveness of these words using two different kind of scale: *i)* the first panel, with 60 subjects, labelled *E,* was asked to measure the age of acquisition on a scale from 1 to 7 according to how early in their life each word was first known, but without specifically mentioning the age; here 1 corresponds to very early in life and 7 to most late; *ii)* the second panel, with 68 subjects, labelled *I,* was asked to indicate in which class of years of their life they acquired them, on a scale 1-7 that this time corresponded to age classes, with 1=0-2 years, 2=2-4 years, 3=4-6 years, 4=6-8 years, 5=8-10 years, 6=10-12 years and 7=13 and further. Thus the effect of the scale constraints given by the suggested age-classes ought to be studied.

3. Exploratory data analysis methods

The limits of the surveys prevented the use of statistical tests, since neither the randomness nor the representativeness requirements could be fulfilled. On the opposite, we could take advantage of the tools available in the framework of *exploratory analysis.* By this term, due to Tukey (1977) we refer to all mathematical, statistical, and computer science methods that may concur to study the structure of a data set in order to extract the contained information in an interpretable way. Their use, started by Benzécri (1973-82) and his *French school*, that claim that "*the models should follow the data, not the inverse*", are also known as *cognitive models* and are nowadays referred as *data mining* in the Anglo-Saxon world. The exploratory work is not exhaustive (Tukey, 1977; see also Camiz, 2001): it is a detective work (Diggle et al., 1994) adopted to ease the acquisition of information contained into the data, but no certainty may result nor statistical inference to a population. We refer here to one of two main sets of tools that are available in multidimensional exploratory data analysis, corresponding to the two big frames currently taken into account (Wittaker, 1967; Benzécri, 1973-82), ordination, as the other, classification did not result of interest for our current purposes. Ordination aims at finding uni- or multi-dimensional orderings in the data, that may be used to sort them accordingly. The identification of such orderings, usually independent from each-other, sometimes allows identifying them with factors that influence the objects at hand, thus causing the found data diversity. We adopted different ordination techniques, according to the kind of data at hand, as each one allows some analysis tools and prevents some others, depending upon their specific features.

3.1 Ordination

The identification of orderings, also known as *gradients* or *factors,* is based on the well known *singular value decomposition* (*SVD,* Eckart and Young, 1936; Greenacre, 1983; Abdi, 2007) and on Eckart and Young's theorem. *SVD* states that any real matrix A may be decomposed according to $A = U\Lambda^{1/2}V'$, with Λ a real quasi-diagonal matrix (that is a matrix with all off-diagonal elements equal to zero) and two real orthogonal matrices U and V (that is such that $U'U = I$ and $V'V = I$). Such *SVD* is unique, up to the ordering of elements, with U a matrix of eigenvectors of AA', V one of $A'A$ and Λ the diagonal matrix whose elements are the corresponding eigenvalues of both. Thus, the elements of A may be reconstructed by the formula $x_{ij} = \sum \lambda_\alpha^{1/2} u_{i\alpha} v_{j\alpha}$. The most interesting feature of *SVD* results from the Eckart and Young's (1936) statement that, once the eigen-elements are sorted in decreasing order, the reconstruction limited to the first r is the best r-dimensional one in the sense of least-squares: $x_{ij} \approx \sum_{\alpha=1}^{r} \lambda_\alpha^{1/2} u_{i\alpha} v_{j\alpha}$. As the total inertia of the data table is given by the sum of the squared singular values, it results that the share of total inertia explained by the r-dimensional solution is given by $\sum_{\alpha=1}^{r} \lambda_\alpha / \sum_{\alpha} \lambda_\alpha$.

If we consider a real data matrix X, with n rows and p columns, we can consider both of them as vectors spanning respectively the n- and p-dimensional spaces R^n and R^p. Thus, implicitly, the p columns and the n rows may be represented as points in the space spanned by the other ones. Due to this representation, it results that *SVD* is strictly tied to the diagonalisation of the scalar products defined in these spaces by XX' and $X'X$ respectively. It may be proved that the eigenvectors are orthogonal and that in decreasing order they maximize their scalar product with all the vector of the basis. It may be also proved that the straight lines spanned by each one maximize the inertia of the orthogonal projection of the clouds of points on them. This leads to an important paradigm of exploratory ordination, that is the higher importance of the directions of maximum inertia in respect to the others.

We shall use three different methods of ordination, *Principal Component Analysis* (*PCA*), and *Simple* (*SCA*) *and Multiple Correspondence Analysis* (*MCA*)*,* according to the nature of the data: *PCA* for measures and *MCA* for qualitative data. In addition, we adopted *Multiple Factor Analysis* (*MFA*) to compare the results of the two different scales of judgements. Indeed, the ordinal scale scores given by the judges may not easily be dealt with, as no easy ordination technique is currently available. Thus we shall use both *PCA*, considering the scales as true measures, and *MCA*, losing the scale nature, but earning the possibility to locate each scale's levels on the factor space.

3.2 Principal Component Analysis

Principal Component Analysis (Benzécri, 1973-82; Gower and Hand, 1996; Jolliffe, 2002; Langrand and Pinzón, 2009) is the most known ordination method suitable for measure data. Based on the above-mentioned *SVD*, it gives principal components as directions along which the maximum of data inertia results. Considered as new variables, it results also that they are maximally correlated with all the variables that form the vector space basis. This leads to a nice interpretation of these components, that are built as linear combinations of the variables and that represent a component common to all of them. The reciprocal relations between variables and principal components derive from the eigenvectors and result in absolute (the share of principal component variation due to the unit variation of a variable) and relative contributions (the share of the variable variation due to the unit variation of the principal component). This method will be used for the scales scores, albeit they only roughly approximate a measure.

3.3 Multiple Correspondence Analysis

Multiple Correspondence Analysis (Benzécri et coll., 1973-82; Greenacre, 1983; Gower and

Hand, 1996; Langrand and Pinzón, 2009) is considered a *PCA* for qualitative data. It is also a generalization of Correspondence Analysis to the case of several categorical variables, as it shares the same special chi-squares metrics. It may be shown that the reconstruction formula in this case becomes $f_{ij} = f_{..}p_{ij} = n_{..}p_{i.}p_{.j}\left(1+\sum_{\alpha=1}^{min(m,q)-1}\lambda_{\alpha}^{-1/2}\Phi_{i\alpha}\psi_{j\alpha}\right)$, with $\Phi_{i\alpha},\psi_{j\alpha}$ the *MCA* factors. In Gomes *et al.* (2010) the relations between *MCA* and *SCA* applied to both the indicator matrix *Z* derived from *X* and the so-called Burt's matrix $B = X'X$ are described. Indeed, the two methods give related solutions. We shall adopt this method when dealing with the scales scores, losing their scale nature, but keeping the ability to represent all levels on the factor spaces. This could reveal an improvement in respect to *PCA*, but an important Guttman effect may result.

3.4 Multiple factor analysis

A special case of data tables is the one in which the same units are observed according to several set of characters, that for some reason one wants to keep separated, or repeatedly in different occasions. These *multiple tables* deserve being analysed in a special way, in order to distinguish the structure of each table/occasion and show the differences among the tables. For this task, the current two-way methods are not suitable, because it is neither possible to identify the structure of each subtable nor show the variation of the pattern from one table to the other. Indeed, the structure of each table may be investigated by a specific *PCA,* but the following comparison of the results may be cumbersome and not effective. For a simultaneous study, *Multiple Factor Analysis* (Escofier and Pagès, 1997) was introduced, based on multiple tables with the same units and different sets of variables or the same variables observed in different occasions. Technically speaking, it is but a weighed *PCA* that is a *PCA* of a pooled table, in which all the values of each table are divided by the square root of the inverse of the first eigenvalue of the individual *PCA*, but this allows very useful developments. This rescaling is introduced to equilibrate the importance of all tables in this pooled analysis: in fact, without it, tables with stronger inner structure that is with more correlated variables, would have a higher influence on the first largest factors. In this way, instead, they are normally balanced. As well, this rescaling allows to compare the inertia of each table projected on the factors with the usual interpretation of contributions and quality of representation not only of the variables but also of each whole table. The comparison of these tables inertias on the factors allows to study the *interstructure,* that is the mutual relations among the tables; the representations on the *MFA* factor spaces allows both the representation of the *compromise,* that is the representation of the units due to all tables, and of the *intrastructure,* that is the common representation of all variables and the individual factors; eventually, on this same space, the partial units, that is those seen by the different tables, may be individually projected, allowing individual comparisons and the drawing of their *trajectories.*
The method may be used also with qualitative data, by adopting the metrics of *MCA* (Pagès, 2002). Here, we shall use it for measures, but we forecast its use for qualitative data too.

4. Results

4.1 Judges selection

In order to examine first the homogeneity of the judges, we started by running two separate *PCA*s on the tables with words in rows as units and judges in column as variables. Thus, the judges with a non-homogeneous behaviour would result in an outlier position in respect to the others. Indeed, from the graphics that represent the pattern of the judges on the circle of correlations on the plane spanned by the first two principal components, in both analyses some outliers resulted: in the first the judges *E12, E23, E59* (Fig. 1a) and in the second *I2* and *I58* (Fig. 1b). All of them resulted very poorly correlated with the first axis and much more with the following ones. On the opposite, no outlier word resulted evident from the inspection of both the graphics of the principal plane and their contribution to the following factors.

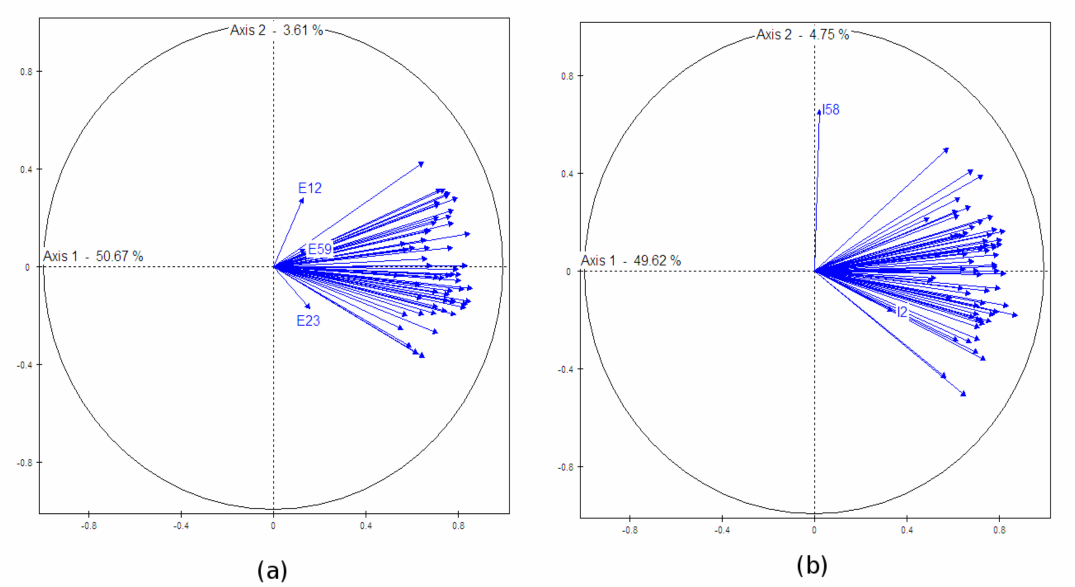

Figure 1. Analysis of age of acquisition judgements. The judges on the plane spanned by the first two principal components of PCA on the respective data: (a) the E scale-free, (b) the I age-based scale. Only the outliers are labelled.

To understand in what the behaviour of the 5 said judges was different, we ran two *MCA*s, in which the age of acquisition was taken as a qualitative level. *MCA* allowed us to represent on the plane spanned by the first two factors the trajectories of the values given by the 5 judges. Unlike the others, whose trajectories are always pretty regular along an arch pattern, they resulted very short and irregular (Figs. 2a, 2b)

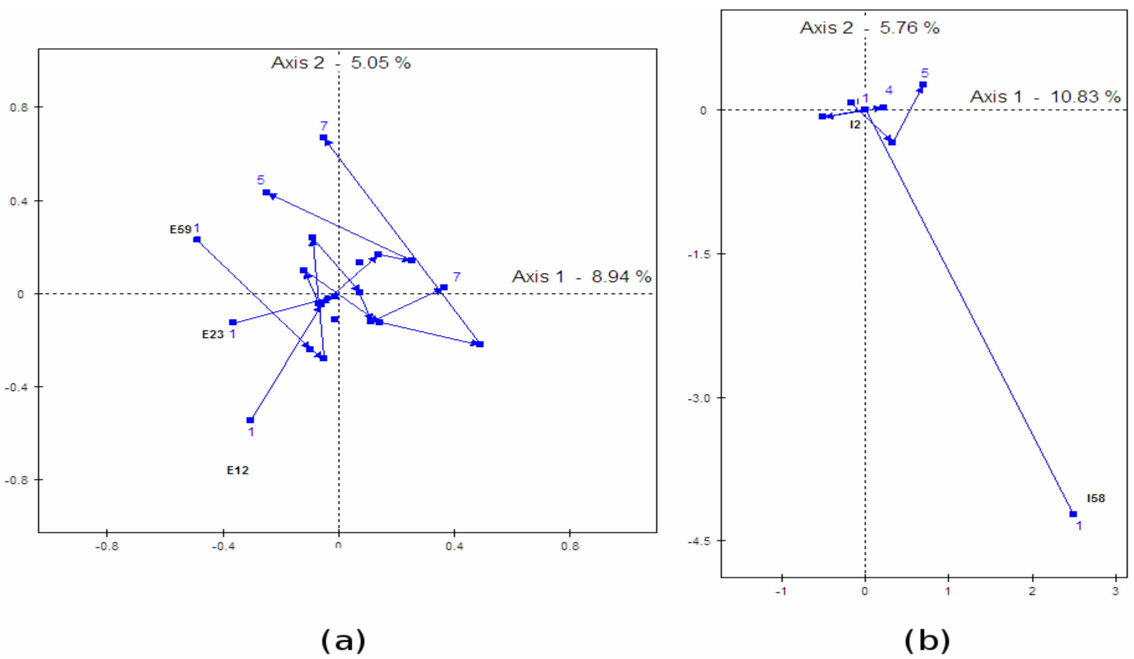

Figure 2. Analysis of the age of acquisition judgements. The outlier judges trajectories on the plane spanned by the first two factors of MCA on the respective data: (a) the E scale-free, (b) the I age-based scale. Only the extremes of the scales are labelled.

Thus, we decided to drop them and rerun the analyses. We dealt with 161 words and 123 judges (57 with free scale and 66 with age-scale; 29 females and 94 men; 17 secondary school and 106 university level).

4.2 Pooled PCA
A *PCA* run on the whole data set, composed by 161 words and 123 judges, gave a highest first eigenvalue (63.29) that alone explains 51.46% of the total inertia. The following factors are very small in comparison and explain a small share of inertia, but nonetheless the second at least deserves being taken into account, for his non negligible value (4.49, 3.65%). As with the first axis all the judges are positively correlated, this axis is easily interpreted as a size factor, with the words scaled least on the negative side (the *early acquired*) of the axis and those scaled most on the positive one (the *late* ones). The interpretation of the second axis may reflect the opposition among words, but we may observe that, unlike the judgements according to the age scale are equally scattered in both sides, the judgement with the free scale are found mostly in the positive side (39 vs. 21). The pattern of the judges is reported in Fig. 3 with only the most external ones labelled for space reasons. The scatter of the words is reported in Fig. 4. Here, the most primitive words are set on the left of the graphics, as *bola, pé, cachorro, banana, mão, casa...,* and the latest ones are set on the right, like *harpa, dedal, charuto, cômoda, cinzeiro...* Yet, it is not clear the meaning of the second axis, with oppositions of words with similar first coordinate, as *sol* vs. *cama*, *coelho*vs. *formiga*, *cobra* vs. *garrafa*, *coruja*vs. *laço*, *gambá*vs. *seta*, *rinoceronte*vs. *machado*... It's likely that this only depends upon the distribution of the judges, but the difference in the used scaling may play some role too.

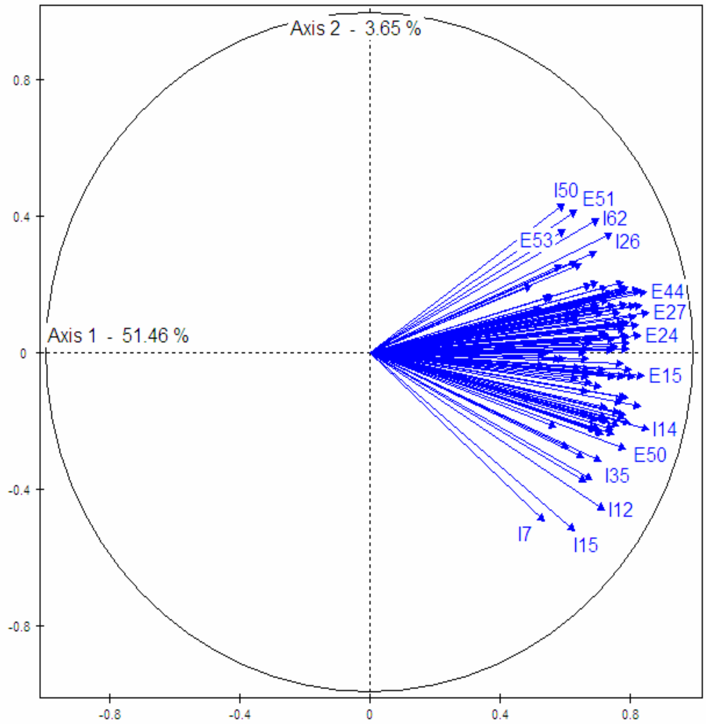

Figure 3: Analysis of the age of acquisition judgements. All the judges represented on the plane spanned by the first two principal components of PCA on the pooled data. Only the outer judges are labelled.

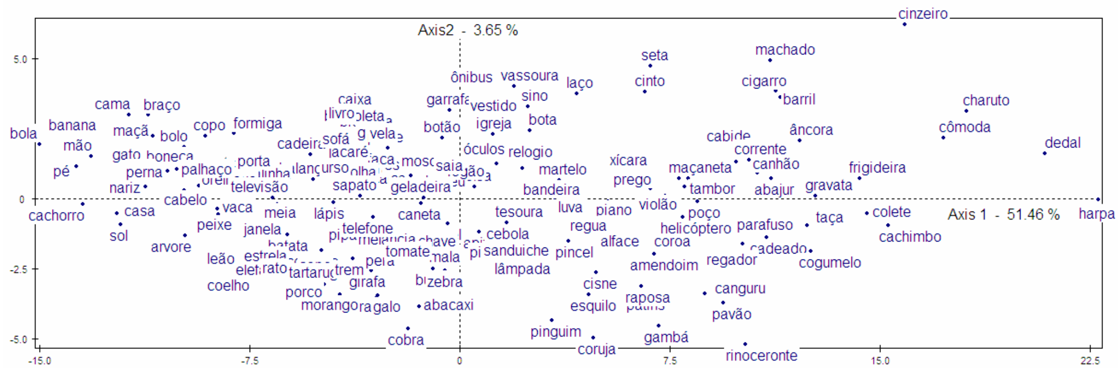

Figure 4: Analysis of age of acquisition judgements. All the words represented on the plane spanned by the first two principal components of PCA on the pooled data.

4.3 Multiple Factor Analysis

Indeed, the pooled analysis did not help very much in understanding to what extent the two scales are different and give different results. For this reason, we ran two separate analyses, followed by a *Multiple Factor Analysis.* We shall not discuss in detail the separate *PCA*: suffice here to say that both showed a first axis that explains more than half of the total variation and a probably non-negligible second (53+4=57%). In addition, the correlation between the two first factors is around 98% and that of the two second is around 67%, whereas those between the following two pairs is around 33%. Thus, the most important aim of our study, the age of acquisition, results nearly identically estimated by the two analyses. In what concerns the second axes, the high correlation suggests that their interpretation might not depend upon the judges, as they were different in the two samples, even if the same pattern of variation among judges might be found in both.

The *AFM* shows a pattern of the eigenvalues very similar to both *PCA*s: the first factor summarizes over 51% of the total inertia and the second raises the explanation up to 55%. The first axes of the individual *PCA*s of the groups are nearly identical to the first axis of *MFA* (correlations = .995 and .994 respectively), whereas the second axes are opposed along the second axis of *MFA*, with correlations .869 and -.938 respectively (Fig. 5). Thus, we can state that the second axis of *MFA* marks a distinction between the two sets of judges: this was suspected during the pooled analysis and now it appears more clearly. The position of the words on the first factor plane of *MFA* may thus be read as follows: the first coordinate marks an average measure of primitiveness, from left, most primitive, to right, least primitive. It appears that the most primitive words are (in order) *bola, banana, cachorro, pé, mão, casa, cama, sol, braço, gato, nariz...* and the least *harpa, dedal, charuto, cômoda, cinzeiro, cachimbo, alicate, colete, frigideira...*

Then, the second coordinate indicates which group of judges gave a higher value in average: on the negative side the *I* age-scale based ones, and on the positive side the *E* free-scale based. Thus, even looking at Fig. 4, based on the pooled analysis, *seta, machado,* and *cinzeiro* have been certainly considered least primitive by the free-judges than by the age-based, that on the opposite considered least primitive *cobra, coruja, rinoceronte….*

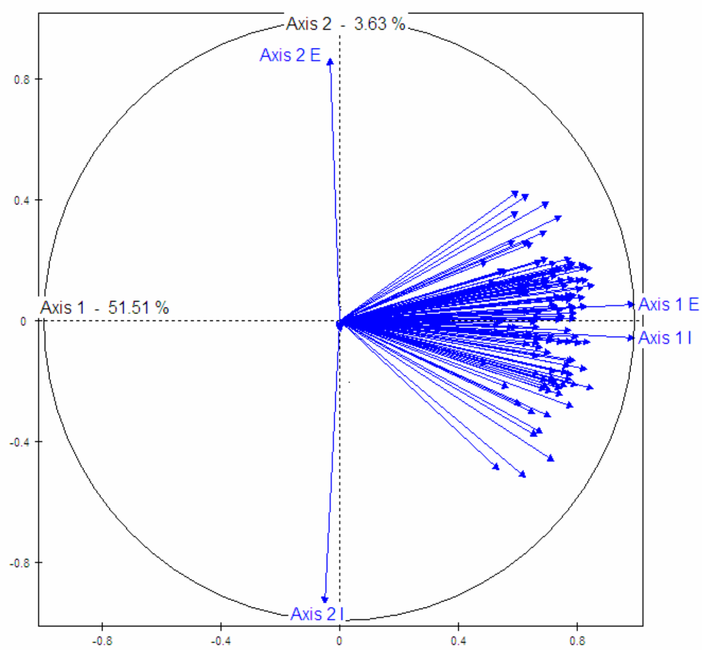

Figure 5. Analysis of the age of acquisition judgements. All the judges are represented on the plane spanned by the first two factors of MFA. Only the partial principal components are labelled: E, the scale-free, I the age-based ones.

MFA returns not only the words coordinates, but also the partial words, that is the projection of the words seen by either group of judges. Indeed, the total word is situated on the centroid of the two partial words. Therefore, we can compute the differences between partial coordinates, that will show a kind of dissimilarity between each pair of partial words. Hence, we could sort the words according to the differences between the first, second coordinates of the corresponding partials, and to their Euclidean distances. We get as highest negative differences along the first factor *burro, gravata, lâmpada, mala, patins* and with highest positive differences *borboleta, cigarro, cinzeiro, escada, galinha, ônibus, vestido.* Thus, the first might be words that were judged more primitive by the free-scale judges, whereas the second might be judged more primitive by the age-scale ones. On the first *AFM* factor plane (Fig. 6) all units are represented through both as seen separately by the two groups and as a compromise. The three points are tied by a line and labelled with the word (the centroid) and either *E* or *I* (the partial). Looking at the extreme of the first axis it is interesting to find the words with the largest differences on the second axis and in particular a reverse behaviour: on the negative side the *E* points lower than the *I*, whereas on the positive side the *I* points are lower than the *E* ones. This is an indication that the *I* judges made a smaller use of the extreme values of the scale than the *E* ones.

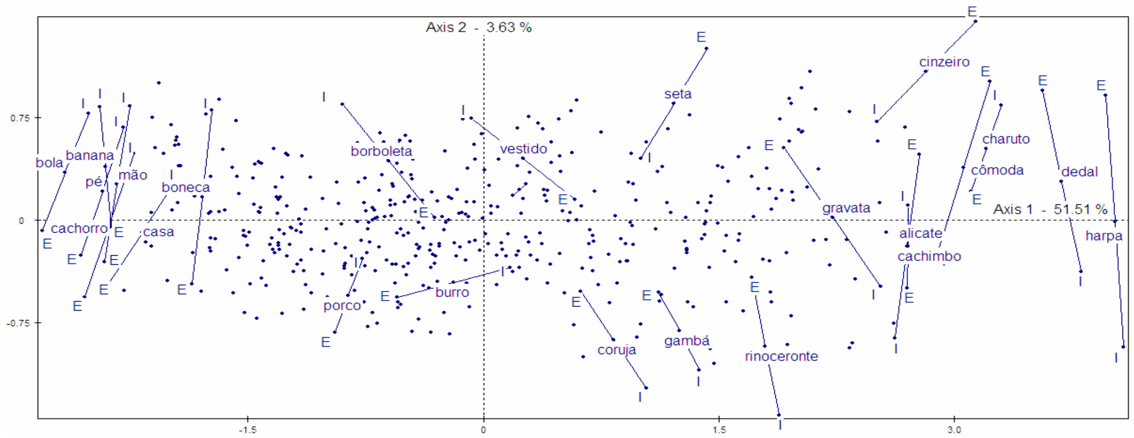

Figure 6. Analysis of the age of acquisition judgements. All the words represented on the plane spanned by the first two factors of MFA. Only the words with the largest trajectories are labelled, with the word (the compromise) and either E or I the partial ones.

5. Discussion and conclusion

As already said, the judges were asked to define the age of acquisition of the submitted words according to different ways of measuring. Thus, a comparison of the responses to the two scales is what we needed. Based on the *AFM* results, we started inspecting the differences in evaluation of the judges belonging to the two groups. Indeed, it seems that the free-scale judges took into account most of the seven given levels: only eleven out of 60 started the scaling from two, wqhereas six of them limited to six and three of them to five the upper bound. On the opposite, the age-class judges limited very often their scale, in particular in what concerns the upper levels: whereas only 4 did not consider the 0-2 years a period of wording, only 23 used the scale up to the highest level. Hence, this quick observation claims in favour of the free scale, in comparison with the age-class based, but a further study is necessary.

The work reported in this paper, the first part of a study in progress, concerning the estimate of the age of acquisition of a word, allows us to establish a noticeable difference between two ways of proposing the test: free-scale and age-based scale. Indeed, a very good agreement between the two scaling results, at least as far as it concerns a first principal component, that may be taken as a measure of primitiveness of age acquisition. The second factor seems to give information concerning the difference in the mean of the distributions according to the two scales as well as the different width of the profiles. On the other end, this represents only around 55% of the overall information, so that it is likely that the other part would include more details concerning both the different scales effects and some specific characteristics of the judges. Further studies will concern these specific aspects.

The limits of the study performed so far are a kind of abuse of *PCA* on data that only roughly may be considered measures, whereas their proper nature is only of ordered data. No well-established methods exist so far in literature for this kind of data and the alternative is to lose the order and treat these data as qualitative. Thus *MCA* will be adopted in the next studies. With it order will be lost as a specific feature, but it can be recovered by tying the subsequent scale levels of each word on the factor space. To get further information on the scale levels, e think also to the study of the words profiles, obtained by building frequency tables of levels according to either groups of judges. The study of profiles may be carried out through *SCA* summarizing evaluations of the judges, with more synthetic and descriptive results. For further comparison between the two sets of judgements, a version of *MFA* based on correspondence analysis metrics (Pagès, 2002), could be adopted for a joint representation. This will be the subject of a further study, albeit a strong Guttman effect (Guttman 1953; see also Camiz, 2005) is expected. For this reasons, methods to derive a linear representation will be adopted.

References

Abdi H. (2007), *Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD)*, in: N. Salkind (ed.), *Encyclopedia of Measurement and Statistics,* Thousand Oaks (CA), Sage.

Benzécri J.P. *and coll.*(1973-82), *L'analyse des données,* 2 voll., Paris, Dunod.

Bose A., O. Raza, and L. Buchanan (2007*). Phonological relatedness between target and error in neologistic productions. Brain and Language, 103: 120-121.*

Camiz S. (2001), «Exploratory 2- and 3-way Data Analysis and Applications». *Lecture Notes of TICMI*, Tbilisi University Press, vol. 2. http://www.emis.de/journals/TICMI/lnt/vol2/lecture.htm

Camiz S. (2005), «The Guttman Effect: its Interpretation and a New Redressing Method», (*Data Analysis Bulletin)*, 5: pp. 7-34.

Dell G.S., M.F. Schwartz, N. Martin, E.M. Saffran, and D.A. Gagnon (1997)*, Lexical Access in Aphasic and Nonaphasic Speakers. Psycological Review, 104: 801-838.*

Diggle P.J., K.Y. Liang, and S.L. Zieger (1994), *Analysis of Longitudinal Data,* Oxford, Clarendon Press.

Eckart C. and G. Young (1936), The approximation of one matrix by another of lower rank, *Psychometrika,* 1: pp. 211-218.

Escofier B. and J. Pagès (1998). *Analyses factorielles simples et multiples,* 3e ed., Paris, Dunod.

Gomes G.C., F.D. Senna, S. Camiz, and C.A. Gomes (2010). «Correspondence Analysis in a Study of Aphasic Patients», *XLII SBPO 2010,* Bento Gonçalves (RS) Brazil, 30/8-3/9. http://www.sobrapo.org.br/sbpo2010/xliisbpo_pdf/72251.pdf

Gower J.C. and D.J. Hand (1996), *Biplots,* London, Chapman and Hall.

Greenacre M. (1983), *Theory and Applications of Correspondence Analysis,* London, Academic Press.

Guttman L. (1953), «A Note on Sir Cyril Burt's Factorial Analysis of Qualitative Data», *British Journal of Statistical Psychology*, 6: pp. 21-4.

Jolliffe I.T. (2002), *Principal Component Analysis*, New York, Springer, Springer Series in Statistics, 2nd ed.

Langrand C. and L.M. Pinzón (2009), *Análisis De Datos. Métodos y ejemplos,* Bogotá, Escuela Colombiana de Ingenieria Julio Garavito.

Pagès J. (2002). «Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. » *Revue de Statistique Appliquée,* 50(4): pp. 5-37.

Pierrehunbert J. (2003). *Probabilistic Phonology: Discrimination and Robustness. In R. Bod, J. Hay and S. Jannedy (eds.) Probability Theory in Linguistics. The MIT Press, Cambridge MA: 177-228.*

Snodgrass J.G. and M. Vanderwart M. (1980)*, A standardized set of 260 pictures: Norms for name agreement, familiarity and visual complexity. Journal of Experimental Psychology: Human Learning & Memory, 6: 174-215.*

Tukey J.W. (1977), *Exploratory Data Analysis,* Reading (Mass.), Addison-Wesley.

WittakerR.H. (ed.) (1967), *Handbook of Vegetation Science - Part V: Ordination and Classification of Vegetation*. Junk, The Hague.