Proceedings of DGA 2013, pp. 1 – 5.

USING CORRESPONDENCE ANALYSIS AND ITS DISTANCE TO EVALUATE THE COMPONENTS OF A NAMING TEST FOR STUDYING APHASIA

Gastão Coelho Gomes,[1] Sergio Camiz,[2] Christina Abreu Gomes,[3] and Fernanda Duarte Senna[4]

[1] *DME– IM– UFRJ, Caixa Postal 68530 cep:21945-970, RJ* gastao@im.ufrj.br

[2] *Dipartimento di Matematica–Sapienza Università di Roma, Italia* sergio.camiz@uniroma1.it

[3] *Departamento de Linguística, UFRJ, Cidade Universitária, RJ* christina-gomes@uol.com.br

[4] *Doutoranda Programa de Pós-Graduação de Linguística, UFRJ, RJ* fonofernandasenna@gmail.com

Resumo:
Análises exploratórias de dados Multidimensionais foram usadas para avaliar dois componentes de um teste de nomeação para o estudo do acesso lexical em pacientes com afasia, ou seja, o acordo de nomeação de imagens e a idade de aquisição de nomes relacionados a um teste de origem internacional. Para o estudo ser confiável as imagens devem ser inequivocamente reconhecidas por qualquer indivíduo usando a mesma palavra/imagem. Suposições teóricas sobre as palavras afirmam que palavras adquiridas mais tarde na aprendizagem tendem a serem as primeiras a serem perdidas devido a um dano cerebral em afasia. Assim, estas duas variáveis são importantes preditoras da recuperação de palavras. Primeiro selecionamos as imagens de reconhecimento com juízes normais; então, as classificamos com relação a sua primitividade. Estas imagens foram submetidas a dois conjuntos de juízes, que tiveram que responder de acordo com duas diferentes escalas. Os dados foram analisados com técnicas de Análises Exploratórias Multidimensionais, incluindo Análises de Correspondências Simples e Múltiplas, Análise de Componentes Principais e Análise Fatorial Múltipla. Com esta comparação não observamos diferença significativa devidos ao uso das duas escalas.

Este trabalho foi feito durante a visita do prof. Camiz a UFRJ, em abril 2013, apoiada pela FAPERJ, processo APV-E-26/110.018/2013, foi enviado para o "WORKSHOP ON DISTANCE GEOMETRY AND APPLICATIONS - DGA'2013" e será publicado em suas atas.

Abstract **Exploratory Multidimensional Data Analyses were used to manage two components of a naming test for studying lexical access in aphasic patients, i.e., the naming agreement of images and age of acquisition of the names themselves from an original international test. In order to be reliable the images should be easily and unequivocally named by any subject using the same word. Theoretical assumptions about word learning states that words acquired later tend to be the first to be lost due to brain damage in aphasia. Thus, these two variables are important predictors of the patient's word retrieval. We first selected the images according to normal judges recognition agreement; then, to range them based on their primitiveness, these images were submitted to two sets of judges, that had to answer according to two different scales. Data were analyzed with several exploratory multidimensional techniques, including Simple and Multiple Correspondence Analyses, Principal Component and Multiple Factor Analyses. A comparison suggested that no mayor differences existed due to the two scales' differences.**

Keywords: **Chi-square Distance, Correspondence Analysis, Factor Analysis, Linguistics, Aphasia**

1. Introduction

This study is the continuation of a previous one [4] and concerns the evaluation of a set of 260 images [8] internationally used to test the lexical access in aphasia, i.e. the loss of some abilities related to language

production and/or comprehension due to brain damage. The test, that aims at measuring to what extent the disease affects the word retrieval, is based on the recognition of familiar objects submitted as images to the patients and their consequent verbalization. In order to be reliable, we considered that a selection of these images ought to be done to suit the Brazilian reality and we based it on two criteria: *i)* the images should be easily and unequivocally recognizable, and *ii)* the *primitiveness* of the word, say its age of acquisition, should be measured. Indeed, it is theoretically assumed (based on word learning) that words acquired later tend to be the first to be lost in aphasic patients affected by brain damage. Thus, these two characters are important predictors of the patient's word retrieval. Thus, we selected randomly three groups of non-affected people to act as judges, and asked one to identify the images and the other two to estimate the degree of primitivity of the corresponding names. As our study was based mainly both on judges and scale evaluation, we show in the following how we dealt with the judges' reliability and the scale definition.

2. The data

- **For the selection of the images, all of them (260) were submitted to a panel of 38 judges randomly selected among non-affected people. The answers have been coded as 1 = recognized, 0 = not recognized. From this selection, 161 images resulted.**
- **To measure primitiveness, we asked 128 non-affected judges to estimate how primitive were the 161 represented objects, according to their personal experience. This estimation was based on two different scales: *i)* the first panel, with 60 judges, labelled *E*, has been 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*, has been asked a measure based on a scale 1-7 as well, but this time based on age classes: the classes are: 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.**

3. Theoretical Framework

The consistency of judges is of high importance in several frameworks, as in sensorial analysis. For this task, specific estimation methods have been developed (see, e.g. [7]). Here, we preferred to consider the problem on another point of view, as no objective primitiveness may be measured, but only identify a central tendence stastistics. Thus, we only removed those judges whose results appeared clearly far from all others. For what concerns the scale definition, we tried to compare two possible scales: a free one and one based on age intervals, and we studied their agreement. Thus, we considered of interest to use for our study *exploratory multidimensional analysis methods*, since their graphical representations allowed a visual inspection of most of the questions that we might ask.

An interesting feature of the analyses that we adopted is that they are all based on the same principle, the *Singular Value Decomposition* (*SVD*, [2]) of some transformation of the original data matrix $T : X \to A = T(X)$. The *SVD* of a matrix A is given by $A = U\Lambda^{1/2}V'$, with U and V the matrices of the (vertical) eigenvectors of $A'A$ and AA' respectively, and Λ the (diagonal) matrix of their corresponding eigenvalues, sorted in descending order. The theorem states the highest importance, in terms of represented inertia, of the first generated axes in respect to the following ones.

According to the data at hand, the analyses have been submitted to *Simple Correspondence Analysis* (*SCA*, [1], [5]) to identify both judges and items with critical recognition behavior, and *Multiple Correspondence Analysis* (*MCA, ibid.*) to identify those judges with biased evaluation of primitiveness in respect to others. *Multiple Factor Analysis* (*MFA*, [3]) has been used to compare the primitiveness of the words given by the two panels of judges according to the two different scales, and eventually *Principal Component Analysis* (*PCA, ibid.*, see also [6]) has been used to define the primitivity index of our interest.

The data transformations, according to the different methods may be described as follows:

PCA $x_{ij} \to z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{n}\sigma_j}$ **standardization**

MFA $x_{ij_k} \to z_{ij_k} = \frac{x_{ij_k} - \overline{x}_{j_k}}{\sqrt{\lambda_k^1}\sqrt{n}\sigma_{j_k}}$ **std. adjusted to group's coherence**

SCA $x_{ij} \to s_{ij} = \frac{x_{ij}}{\sqrt{x_{i.}x_{.j}}} - \frac{\sqrt{x_{i.}x_{.j}}}{x_{..}}$ **deviation from independence**

MCA $x_{ij_q} \to s_{ij_q} = \frac{1}{\sqrt{Q}}\left(\frac{x_{ij_q}}{\sqrt{x_{i.}}} - \frac{\sqrt{x_{i.}}}{x_{..}}\right)$ **deviation from average profile**

4. Selecting Images

The results of first submission reported 10 images that no judge could identify, so that we withdrew them immediately. On the other side, 66 images have been recognized by all judges, thus automatically included. Therefore, we applied *SCA* to the remaining images to get a graphical representation of the pattern of both judges and images on factor planes. According to Figure 1(a) below, six judges, P13, P17, P25, P32, P34, and P36, appear further from the origin than all others, whose central pattern seems homogeneous, thus they have been withdrawn.

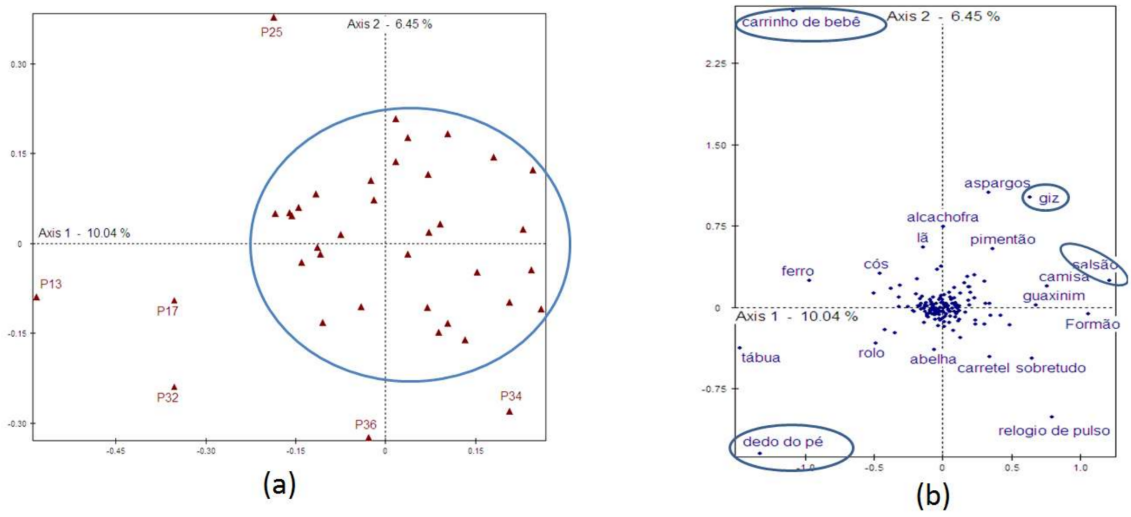

Figure 1. Analysis for the selection of the images. The items on the first factor plane of SCA: (a) The judges, (b) The names.

As well, some items, such as *baby stroller, toe, celery,* and *chalk*, were identified by no more than 5 judges. They are located at the border of the cloud as can be seen on Figure 1(b) above. We re-ran *SCA* with only 32 judges and also all the items whose frequency of correct identification was lower than 50%. From the homogeneous results we could conclude that no further removal of judges seemed necessary. Eventually, we decided to keep all the images that were correctly identified by at least 90% of judges. Based on 32 judges: 97 images were identified by all of them, 26 by only 31 (97%), 24 by 30 (94%), 14 by 29 judges (91%), summing up to 161 images.

5. Defining word primitiveness

In order to examine first the homogeneity of the judges, we started by running *MCA*s on each of two tables. Their behavior is represented by a trajectory that connects the seven levels of the scale. Observing the two graphics in Figure 2, one may observe that the trajectories of the judges that measured freely (Figure 2(a)) are much longer than those of judges based on age (Figure 2(b)). This may be explained by a reduced use of the first levels by the latter.

The pattern of trajectories on both tables on the first factor plane is very homogenous among both sets of judges: Only five of them (*E12, E23, E59, I2*, and *I58*) showed very strange trajectories (see them in Figure 3), thus were removed.

Then, we ran a *MFA*, considering the two groups of reduced judges (57 that used the free scale (*E*) and 66 with age-scale (*I*). A specific advantage of *MFA* in respect to *PCA* it its ability to represent on factor planes not only the global units, but also the partial ones, that is, in our case, 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, distances between partial words are a measure of their dissimilarity according to the two sets of measurements and they may be decomposed according to the different axes. The words with highest negative differences along the first factor are *burro, gravata, lâmpada, mala*, and *patins* and those with highest positive ones are *borboleta, cigarro, cinzeiro, escada, galinha, ônibus*, and *vestido*. Thus, the first might be words judged more primitive by the free-scale judges, whereas the second might be judged more primitive by the age-scale ones. Here, we deal only with the first axis that clearly represents primitivity of words (51.51% of total inertia), since the following explain too little inertia to deserve being taken into account (the second only 3.64%). In Figure 4 all words are represented both totally and partially, with the total units at the centroid of the respective partials. Looking at

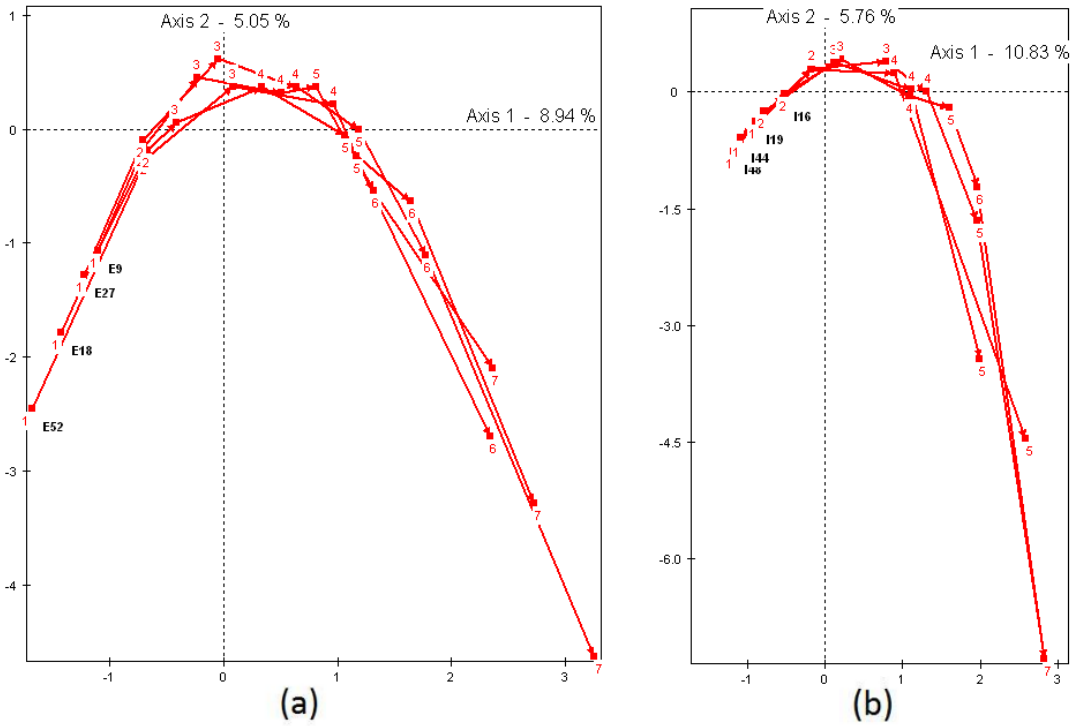

Figure 2. Analysis of the primitiveness judgements. The judges' trajectories represented on the first factor plane of *MCA*: (a) free judgements, (b) judgements based on age intervals.

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: this reflects the small rotation of the first factors of partial tables, but does not deserve a true interest for our purposes.

As the partial first factors of the two tables where most correlated among each other (.98) and with the *MFA* one (over .99), we decided to merge the two data sets , so that as measure of the words' primitivity was taken the first principal component of this unified table's *PCA*.

6. Conclusions

The study aiming at both selecting images with high naming agreement and measuring the degree of primitiveness of their correspondent words, has been carried out using only exploratory multidimensional data analyses. This allowed to withdraw judges with a clearly biased behavior in respect with the others and select

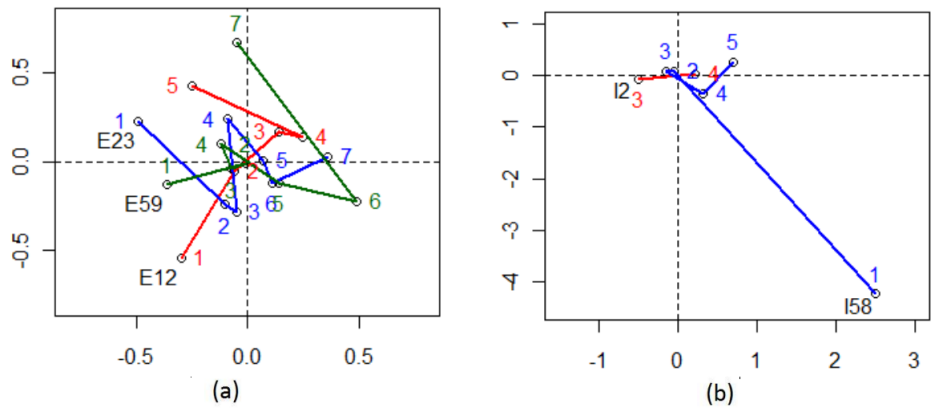

Figure 3. Analysis of the primitiveness judgements. The outlier judges' trajectories represented on the first factor plane of the respective *MCA*: (a) free judgements, (b) judgements based on age intervals.

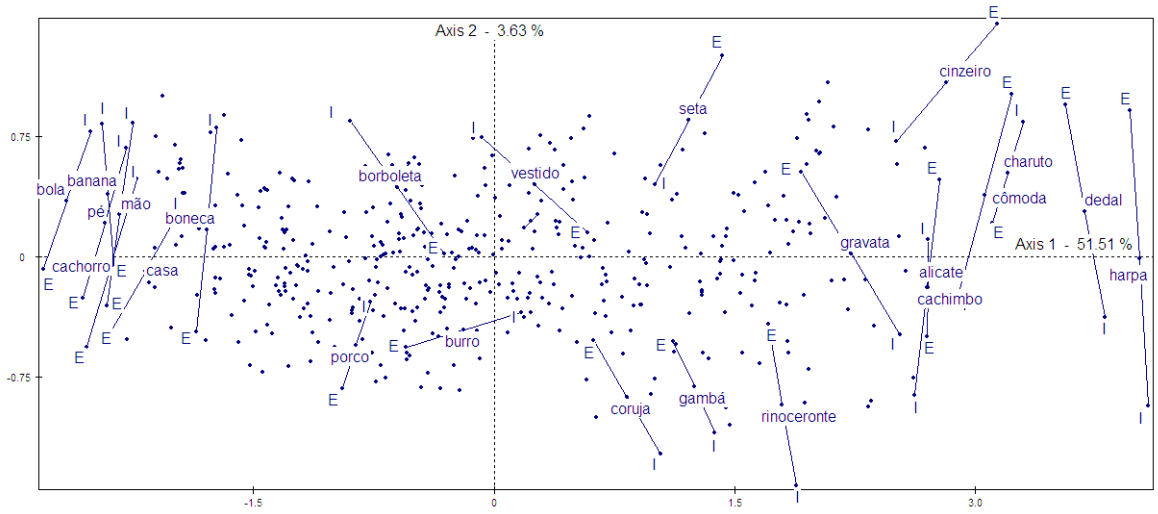

Figure 4. Analysis of the age of acquisition judgements. All the words represented on the plane spanned by the first two factors of *MFA*. Only the words with the largest trajectories are labelled, with the word (the compromise) and either *E* or *I* the partial ones.

a set of words that have been recognized by nearly the totality of judges. The free scale resulted a little better performing than the other, since it allowed a more instinctive estimate.

References

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

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

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

[4] Camiz S., G.C. Gomes, F.D. Senna, 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

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

[6] Jolliffe I.T. (2002), *Principal Components Analysis*, Berlin, Springer.

[7] Rust R.T. and B. Cooil (1994), Reliabilty Measures for Qualitative Data: Theory and Implications. *Journal of Marketing Research*, Vol. 31: pp. 1–14.

[8] Snodgrass J.G. and M. Vanderwart (1980), A standardized set of 260 pictures: Norms for name agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, Vol. 6(2): pp. 174–215.