# Categorical Data Analyses for Studying Acquisition of Relative Clauses in Linguistics

**Sergio Camiz**
Dipartimento di Matematica, Sapienza Università di Roma
E-mail: sergio.camiz@uniroma1.it
**Gastão Coelho Gomes**
Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro
E-mail: gastao@im.ufrj.br
**Ana Cristina Baptista de Abreu**
Departamento de Linguística, Universidade Federal do Rio de Janeiro
E-mail: anacristina.abreu@hotmail.com
**Christina de Abreu Gomes**
Departamento de Linguística, Universidade Federal do Rio de Janeiro
E-mail: christina-gomes@uol.com.br

**Abstract**

This paper aims at studying the ability of some descriptive and predictive qualitative data analysis methods to check the results of a linguistic experiment. In order to study the acquisition of relative clauses in preschool children of Rio de Janeiro, an experiment was carried out by Abreu (2013), that submitted to 47 children a series of repetition tests according to a Latin square design. Through correspondence analyses, log-linear models, and multinomial logistic models, the consistency of the sampling design resulted as well as the dependence of the response accuracy on some of the design factors.

**Keywords:** Correspondence Analysis, Multiple Correspondence Analysis, Log-linear models, Multinomial logistic models, Linguistics, Acquisition of Relative Clauses.

## 1 Introduction

The domain of linguistics is an interesting framework in which qualitative data analyses may be fruitfully applied: both multidimensional descriptive methods (such as *Correspondence Analysis*) and descriptive and predictive models (such as *Log-linear* and *Multinomial Logistic models*) find there their utility, see Lebart and Salem (1994) and both McEnery and Wilson (1996) and McEnery and Hardie (2012) for examples. In this work we use them to check the results of a work carried out by Abreu (2013) that concerns the acquisition of relative clauses in preschool children. As the children were tested according to a well-designed experiment, we applied the said methods to check the independence of the factors considered in the design and on the other side the dependence of the relative character upon the said factors.

Abreu (2013)'s research is based on studies about the speech of people living in Rio de Janeiro, in analogous way of those which focus on Portuguese language in both Brazilian and European Portuguese varieties. According to her, these works revealed that relative structures are variable together with the fact that they are passing through a process of change (Tarallo, 1993; Mollica, 2003). Abreu refers to Diessel and Tomasello (2000, 2005) researches in which the structural frequency and similarity revealed to be essential information to the acquisition process of these structures and from which the methodological bases for the elaboration of an experiment of induced production were taken. Data on infant speech were obtained from recordings of spontaneous speech of 47 children. One aspect of the methodology of the study was to develop a test of repetition of relative clauses, considering some variants. These stimuli were divided into three test versions, according to the Latin square sampling design (Kuehl, 2000), so that regardless of the version of the test, every child was exposed to the same types of structure and variants in the same quantity, without being exposed to variants of the same sentence Abreu (2013). Indeed, the chi-square test proved that no significant difference resulted from the different tests proposed. Three other independent factors constitute the experimental design and a dependent character:

- *Age*, the age of children, with $I45$ = 4 or 5 years old , $I67$ = 6 or 7 years old;
- *Type of relative*, the linguistic variants observed for the speakers of Brazilian Portuguese, with $TrP$ = standard variant, $TrR$ = Piedpiping-chopping, $TrD$ = Resumptive variant;
- *Phrase structure*, defined according to the syntactic function of the relative pronoun in the clause, with $EA$ = adverb, $EI$ = indirect object; $EG$ = genitive, $ES$ = subject, $EO$ = direct object.
- *Accuracy*, the dependent character, the way the children repeated the proposed test sentence, a scale measure of the acquired linguistic knowledge, with $Ac0$ = error, $Ac1$ = variation (socio-linguistic hit), and $Ac2$ = good (total hit).

Indeed, the possible combinations of *Type of relative* and *Phrase structure* are those reported in the following examples:

```
- ES - TrP:  Esta é a boneca que fala.
     - TrD:  Esta é a boneca que ela fala.
- EO - TrP:  Este é o trem que a tia deu.
     - TrD:  Este é o trem que a tia deu ele.
- EI - TrP:  Esta é a estória de que a menina falou.
     - TrD:  Esta é a estória que a menina falou dela.
     - TrR:  Esta é a estória que a menina falou
- EA - TrP:  Aquela é a praia em que o menino foi.
     - TrD:  Aquela é a praia que o menino foi nela.
     - TrR:  Aquela é a praia que o menino foi.
- EG - TrP:  Essa é a boneca cujo cabelo eu cortei.
     - TrD:  Essa é a boneca que eu cortei o cabelo dela
     - TrR:  Essa é a boneca que eu cortei o cabelo.
```

Thus, depending on the clauses' structure, there are either two or three possible relatives: for *ES* and *EO*, only *TrP* and *TrD* are possible, whereas for *EA, EI,* and *EG* all three variants *TrP, TrD,* and *TrR* occur in Portuguese. Note that the type of relative does not depend on the phase structure: some constructs simply do not appear.

In this work we use the data by Abreu (2013) to ascertain the independence of the factors and the dependence of Accuracy on the factors. The aim is to check to what extent the techniques adopted, Correspondence Analysis, Log-Linear Models, and Multinomial Logistic Models are suitable for this task. In this paper, we use them both as exploratory tools, in particular to check the quality of the experimental design, and interpretative tools, to understand to what extent one may identify a dependency of one relative character from other explanatory ones.

## 2 Theoretical framework

The study of two- and multi-way contingency tables, is nowadays performed through their decomposition. This means that each table entry, assumed to be the result of a Poisson process, depends upon a set of concurrent effects. This is true both for the table's description, considering both all co-occurring crossing characters, and for the study of the dependence, that is the relative of the relative character based on the others' levels.

Two different points of view may be taken into account:

- exploratory, essentially based on *Correspondence Analysis* (*SCA* Greenacre, 1983),
- model oriented, that is based on both Log-Linear and Multinomial Logistic Models (Goodman, 1985; Christensen, 1997).

## 2.1 Correspondence Analysis

The *Correspondence Analysis* (*CA*, Greenacre, 1983) is an eigenanalysis technique, whose aim is to decompose at the best a contingency data table, in order to put in evidence additive factors on which table's marginal values may depend. *CA* is based on *Generalized Singular Value Decomposition* (*GSVD*, Greenacre, 1983; Abdi, 2007), that we remind shortly here, together with a couple of useful theorems.

**Theorem 1.** *(Singular Value Decomposition, SVD) Any real matrix $X$ may be decomposed as $X = U\Lambda^{1/2}V'$, with $\Lambda$ the diagonal matrix of the real non-negative eigenvalues of $XX'$, $U$ the orthogonal matrix of the corresponding eigenvectors, and $V$ the matrix of eigenvectors of $X'X$ (with the same eigenvalues), with both constraints $U'U = I$ and $V'V = I$.*

The demonstration may be found in Abdi (2007). This theorem corresponds to the reconstruction formula of an $r$-rank matrix with $r$ 1-rank matrices

$$x_{ij} = \sum_{\alpha=1}^{r} \sqrt{\lambda_\alpha}\, u_{i\alpha} v_{j\alpha},$$

since the product of each $\alpha$-th couple of vectors $\sqrt{\lambda_\alpha}\, u_\alpha v'_\alpha$ produces a 1-rank matrix. On this theorem, the well known Eckart and Young (1936) theorem is based:

**Theorem 2.** *(Eckart and Young) The $s$-rank reconstruction of any real matrix $X$, with $s < r$, the rank of $X$, once its singular values are sorted in decreasing order,*

$$x_{ij} \approx \sum_{\alpha=1}^{s} \sqrt{\lambda_\alpha} \;\; u_{i\alpha} \; v_{j\alpha} = \tilde{x}_{ij,s}$$

*is the best one in the least-squares sense.*

This means that, for every $s < r$, the $s$ 1-dimensional matrices $\tilde{X} = (\tilde{x}_{ij,s})$ solves the problem to approximate a matrix $X$ by another matrix $H$ of lower rank at the best in the least-squares sense, thus by minimizing

$$\sum_{i=1}^{r} \sum_{j=1}^{c} (x_{ij} - h_{ij})^2 = \text{trace}\left((N - H)(N - H)'\right) \tag{1}$$

For *CA*, we shall adopt the Generalized Singular Values Decomposition (*GSVD*, Greenacre, 1983; Abdi, 2007), in which two other matrices are involved:

**Theorem 3.** *Given two real positive definite matrices $M_r$ and $M_c$, any real matrix $X$ may be decomposed as $X = \widetilde{U}\Lambda^{1/2}\widetilde{V}'$, under constraints $\widetilde{U}'M_r\widetilde{U} = I$ and $\widetilde{V}'M_c\widetilde{V} = I$.*

The solution is given by the $SVD$ of the matrix $\widetilde{X} = M_r^{1/2} X M_c^{1/2} = F\Lambda^{1/2}G'$, with $F'F = I$, $G'G = I$, $\widetilde{U} = M_r^{-1/2}F$, and $\widetilde{V} = M_c^{-1/2}G$. It results that $\widetilde{U}\widetilde{U}' = M_r^{-1}$ and $\widetilde{V}\widetilde{V}' = M_c^{-1}$ respectively.

The exploratory analysis paradigm states that the most relevant information is tied to the largest eigenvalues and the non-relevant to the least ones. Thus, one is led to study the relevant first and stop the study according to several thumbnail rules.

Let $N$ an $r \times c$ contingency table, with $n = n_{..}$ the table grand total, $\vec{r} = (p_{1.}, ..., p_{r.})'$ the vector of row marginal profile (with $p_{ij} = n_{ij}/n$), $\vec{c} = (p_{.1}, ..., p_{.c})'$ the vector of column marginal profile, and $D_r = \operatorname{diag}(\vec{r}), D_c = \operatorname{diag}(\vec{c})$ the corresponding diagonal matrices. The $SCA$ of $N$ results from the application of $GSVD$ to the contingency table $N$ with the constraints given by the diagonal matrices $D_r$ and $D_c$. As a result, the reconstruction formula of $N$ is:

$$n_{ij} = nr_ic_j \left(1 + \sum_{\alpha=1}^{\min(r,c)-1} \sqrt{\lambda_\alpha}\; f_{i\alpha}\; g_{j\alpha}\right). \qquad (2)$$

where 1 is the first trivial eigenvalue, that ties the origin to the centroid of data and represents the independence. This time, as both $D_r$ and $D_c$ are diagonal and represent the table marginal frequencies of rows and columns respectively, the minimization problem (1) is simplified and takes the interesting aspect

$$\begin{aligned} \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-h_{ij})^2}{e_{ij}} &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-h_{ij})^2}{nr_ic_j} \\ &= n^{-1}\operatorname{trace}\left(D_r^{-1}(N-H)D_c^{-1}(N-H)'\right) \qquad (3) \end{aligned}$$

that is the sum of squared deviations of the approximated values from the observed ones divided by the expected ones under independence.

In $SCA$ the eigenvalues sum, up to the grand total, to the table chi-square, namely

$$\chi^2 = n \sum_{\alpha=1}^{\min(r,c)-1} \lambda_\alpha = \sum_{\alpha=1}^{\min(r,c)-1} \chi^2_\alpha,$$

where each $\chi^2_\alpha$ is the share of the total deviation from expectation due to the corresponding 1-dimensional layer. Thus, this feature allows to test for significance both each 1-dimensional layer and the corresponding residual. Indeed, the partial chi-square associated to each eigenvalue, $\chi^2_\alpha = n_{..}\lambda_\alpha$, may be checked for significance with $(r + c - 2\alpha - 1)$ degrees of freedom (Kendall and Stuart, 1961), to detect if the corresponding linear ordinations of both rows and column levels explain the found deviation from expectation (Orlóci, 1978). In analogous way, the residuals may be tested by using the classical test for goodness of fit (Kendall and Stuart, 1961), simplified according to Malinvaud (Ben Ammou and Saporta, 2003). Indeed, he suggests to approximate the estimated frequency in the fraction's denominator with the expected one under independence. Thus, for each $\alpha$-dimensional partial reconstruction, the residuals correspond to

$$Q_\alpha = \sum_{ij}\frac{(n_{ij}-\widetilde{n}_{ij,\alpha})^2}{\widetilde{n}_{ij,\alpha}} \approx \widetilde{Q}_\alpha = \sum_{ij}\frac{(n_{ij}\ -\widetilde{n}_{ij,\alpha})^2}{nr_ic_j}\ = \chi^2 - \sum_{\beta=1}^{\alpha}\chi^2_\beta = n\sum_{\gamma=\alpha+1}^{\min(r,c)-1}\lambda_\gamma,$$

asymptotically chi-square-distributed with $(r - \alpha - 1) \times (c - \alpha - 1)$ degrees of freedom. In the formula, $\widetilde{n}_{ij,\alpha} = \sqrt{\lambda_\alpha}\; f_{i\alpha}\; g_{j\alpha}$ is the cell value estimated by the $\alpha$-dimensional solution, and the table chi-square test results when $\alpha$= 0 and $\widetilde{n}_{ij,0} = \frac{n_{i.}\ n_{.j}}{n_{..}}$ is the expected value under independence. Thus, for each eigenvalue, the first test indicates if the corresponding linear ordination is significant, thus deserves being taken into account and the second, if significant, suggest to examine further dimensions, since their content is not random.

Let us consider now a qualitative data table $X$ with $n$ observations, $Q$ nominal characters and $J$ the total number of levels, that is $J = \sum_{i=1}^{Q} l_i$ where $l_i$ is the number of levels of the $i$-th character. It is well known that *Multiple Correspondence Analysis* (*MCA*, Greenacre, 1983) of such a matrix is but a generalization of *CA* applied to the so-called Burt's table $B$ that gathers all contingency tables obtained by cross-tabulating all the $J$ levels of the $Q$ considered characters. Note that this also includes the diagonal tables obtained by crossing each variable with itself, an information of no practical interest for the analysis. Thus, *MCA* is the *GSVD* of $Q^{-2}D_r^{-1}BD_r^{-1}$. Here we may calculate the inertia in form of a deviation from expectation, and it results:

$$\chi_B^2 = \sum_\alpha \mu_\alpha^2 = 2\sum_{i=1}^{Q}\sum_{j=1}^{i-1}\chi_{ij}^2 + n(J-Q).$$

Note that we noted $\mu_\alpha^2$, as we shall consider their square roots as eigenvalues. Indeed, $\chi_B^2$ is not a true chi-square statistics, since it is twice the sum of the chi-squares of the $Q(Q-1)/2$ tables that cross all characters one by one plus the $Q$ diagonal tables ones, whose inertia $n(J-Q)$ has no meaning. Indeed, such diagonal matrices, that "theoretically" would indicate maximum deviation, in this case are just the expected ones, as they cross each character with itself. Nevertheless, they cause a dramatic inflation of the inertia without interest.

Details concerning this specific point may be found in Camiz and Gomes (2013): suffice here to say that it is suggested to to limit attention to the first eigenvalues larger than their mean, since the mean represents the expectation of all eigenvalues under independence. Thus, to correctly identify the significant dimension, Ben Ammou and Saporta (2003) propose to estimate the significance of the eigenvalues $\mu$ of *MCA* according to their distribution. If the characters are independent, they prove that the expectation of the variance $S_\mu^2$ of the eigenvalues is

$$\sigma^2 = E[S_\mu^2] = \frac{1}{n_{..}Q^2(J-Q)}\sum_{i\neq j}(l_i-1)(l_j-1).$$

Roughly, one may assume that the interval $\frac{1}{Q}\pm 2\sigma$ should contain about 95% of the eigenvalues. Indeed, since the kurtosis of the set of eigenvalues is lower than for a normal distribution, the actual proportion is larger than 95%.

For what concerns the amount of inertia explained by the factors whose eigenvalue is larger than the mean, $\mu \geq \overline{\mu}_\alpha = \frac{1}{Q}$, Greenacre (1988) suggests to re-evaluate it according to

$$\rho\left(\nu_\alpha\right) = \left(\frac{Q}{Q-1}\right)^2 \left(\mu_\alpha - \bar{\mu}\right)^2, \mu_\alpha \geq \overline{\mu} = \frac{1}{Q}. \qquad (4)$$

and to compare it to the off-diagonal inertia, that is the sum of squared (non-re-evaluated) eigenvalues minus the diagonal inertia: that is

$$\frac{Q}{Q-1}\left(\sum_{\mu_\alpha > 1/Q}\mu_\alpha^2 - \frac{J-Q}{Q^2}\right).$$

This does not affect the interpretation of the factors, but the use of $rho\left(\nu_\alpha\right)$ of (4) in place of $\lambda_\alpha$ in the reconstruction formula (2) increases the explained inertia and also fixes a dramatic bias of the partial reconstruction of the off-diagonal tables, as is has been empirically proved by Camiz and Gomes (2013).

## 2.2 Log-linear models

Unlike correspondence analysis, whose modeling capacity is essentially additive, in the sense that it is able to identify a limited number of factors whose effects are added one by one to estimate the table cells, the *log-linear* models (*LLM*, Goodman, 1985; Christensen, 1997) are multiplicative.

As we shall see further, this is particularly interesting to understand the interaction among the crossing characters, otherwise difficult to understand. Given an $r \times c$ matrix $N$, with grand total $n_{..}$, if $p_{ij} = \frac{n_{ij}}{n_{..}}$ is the $(i, j)$-th cell relative frequency, that we may consider a probability, thus it may be described by the *saturated log-linear model* as follows, for $i = 1, ..., I$ and $j = 1, ...J$ the number of levels of the two crossing characters:

$$log(p_{ij}) = \alpha + \beta_{1,i} + \beta_{2,j} + \gamma_{12,ij}$$

that is the sum of an intercept, two sets of parameters corresponding to the effects of both $i$-th and $j$-th levels of the corresponding character, and one set corresponding to the interaction between $I$ and $J$ levels. In general, the coefficients are taken as contrasts in respect to one level, so that the total parameters are exactly the number of cells and no residual results. In the case of a three-way contingency table, with a third character with $k = 1, ..., K$ levels, the saturated model is:

$$log(p_{ijk}) = \alpha + \beta_{1,i} + \beta_{2,j} + \beta_{3,k} + \gamma_{12,ij} + \gamma_{13,ik} + \gamma_{23,jk} + \delta_{123,ijk}$$

where both 2- and 3-way interactions are taken into account. For multi-way tables the model generalizes accordingly, see Goodman (1985); Christensen (1997) for further explanation.

Indeed, it is interesting to find more parsimonious models, able to describe at the best the table. This may be done through the progressive withdrawal of some of the effects, on condition to keep all those that appear in the more complex interactions. Thus, in practice, starting from the saturated model, the highest level interactions are removed and removing continues from highest levels to lowers as far as a reasonable fitting of the table is still obtained. In any case, no lower effect may be removed if the corresponding character is present in some higher level interaction. In order to decide whether it is convenient to withdraw which effect or not, a stepwise procedure is implemented that, starting from the saturated model removes first the highest level interactions, and tests whether such loss of information is acceptable or not in favor of a more parsimonious model. The test is performed through the Akaike (1974)'s AIC coefficient, that is

$$AIC = 2k - 2logL$$

with $k$ the number of parameters of the model and $L$ is the maximized value of the likelihood function for the model. At the following steps, all possible reductions, compatible with the hierarchy, are tested and the effect is removed that causes the minimum $AIC$ value, provided it is less than the current one. The stepwise process stops when no reduction of $AIC$ is possible. Here, the removal of some effect indicates that its variation does not influence significantly the cells entries, so that the meaning of the whole table is simplified.

## 2.3 Multinomial logistic models

A model similar to the log-linear one may be taken into account in the case in which one wishes to predict one character level of a multi-way table according to the other character's ones. These are *multinomial logistic models* ($MLM$): they estimate the probability of each level of the relative character according to each combination of the others. For every observation $Y_i, i = 1, \ldots, n$ of the criterion character, the probability to get any level $k = 1, ..., K$ is given by

$$Pr(Y_i = k) = \frac{e^{\beta_k X_i}}{\sum_{k=1}^{K} e^{\beta_k X_i}}$$

with $\beta$ and $X$ vectors of coefficients and of explanatory characters respectively (Venables and Ripley, 2002). The choice for the best model is done again hierarchically through the $AIC$ coefficient. Eventually, the found model may be used to check for correct attribution of the observations to the levels and new observations may be classified accordingly. Indeed, attention must be paid on the fact that an observation is always attributed to the level with the highest predicted probability, given by the observed combination of levels of the explanatory characters.

## 3 Results

As said, the data taken by Abreu (2013) consist of a total of 1206 observations, according to the characters *Accuracy*, *Type of relative*, *Phrase structure*, and *Child's age*, with the corresponding frequencies of their respective levels reported in Table 1.

*Table 1: Frequency tables of the characters under study.*

| *Accuracy* | | | *Type of relative* | | | *Phrase structure* | | | | | *Child's age* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Ac0* | *Ac1* | *Ac2* | *TrD* | *TrP* | *TrR* | *EA* | *EG* | *EI* | *ES* | *EO* | *Id45* | *Id67* |
| 243 | 228 | 735 | 470 | 454 | 282 | 282 | 266 | 282 | 188 | 188 | 590 | 616 |

In order to check for the consistency of the experiment's design, an *MCA* was run on the three explanatory characters, that is the design factors: unlike expected, one eigenvalue resulted significant, with a value 0.457, larger than the upper interval limit 0.36 fixed by the Ben Ammou and Saporta test. Indeed, in Table 2 results that two *Phrase structures*, namely *ES* and *EO*, were not crossed with the *Type of relative TrR* in the submitted tests: this depended upon the incompatibility between these structures and *TrR*.

*Table 2: Cross-tabulation of Type of relative with Structure.*

| | *EA* | *EG* | *EI* | *ES* | *EO* |
|---|---|---|---|---|---|
| *TrD* | 94 | 94 | 94 | 94 | 94 |
| *TrP* | 94 | 78 | 94 | 94 | 94 |
| *TrR* | 94 | 94 | 94 | 0 | 0 |

$\chi^2 = 168.221$, df $= 8$, p-value $< 2.2\text{e-}16$

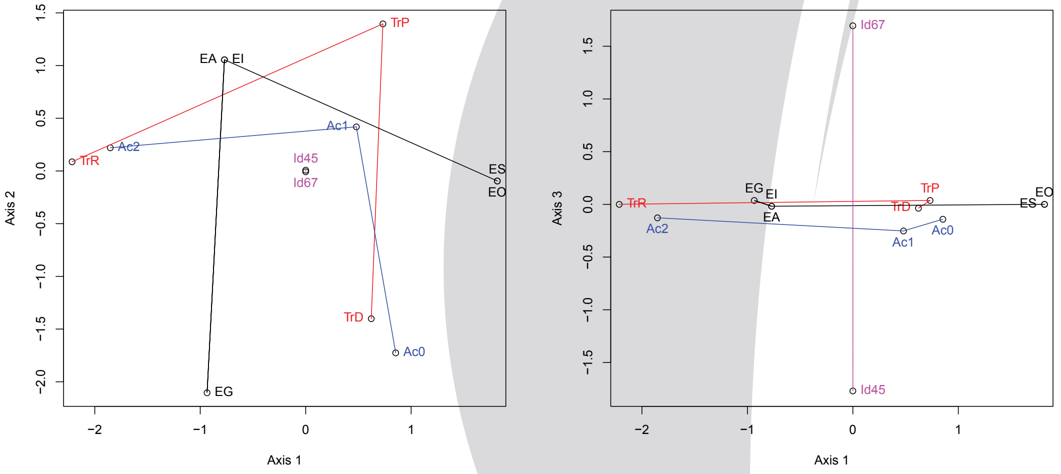

*Figure 1: The levels of all characters on the factor planes of MCA on the overall data table. Left: factors 1 and 2; Right: factors 2 and 3.*

In analogous way, the crossing of *TrP* with *EG* has been limited to 78 tests instead than 94, due to a submission error. Thus, the table's grand total (1206) results in the 13 tests submitted twice to each of 47 children minus the 16 that resulted erroneous. Both problems appear very clearly in the first graphic of Figure 1, in which *TrR* is opposed to both *ES* and *EO* on the first axis (horizontal) and *EG* is opposed to *TrP* on the second.

Indeed, applying the log-linear model, the backward elimination according to the *AIC* led to the elimination of the *Child's age*, since the other crossings where not influenced by these two levels; on the opposite, the interaction *Type of relative - Structure* resulted significant, due to the said incompatibility of *TrR* with *ES* and *EO*.

On this *MCA*, the *Accuracy* levels were projected as supplemental on the factor planes, since all tables crossing *Accuracy* with the other characters were highly significant, as proven by the $\chi^2$ tests shown in Table 3. On the graphics of Figure 1 the closeness to levels of the active explanatory factors may be considered as an influence of the latter to the level of accuracy.

*Table 3: Cross-tabulation of Accuracy with the other characters.*

| | *TrD* | *TrP* | *TrR* | | |
|---|---|---|---|---|---|
| *AC0* | 82 | 101 | 60 | | |
| *AC1* | 76 | 144 | 8 | | |
| *AC2* | 312 | 209 | 214 | | |
| $\chi^2 = 111.8392$, df = 4, p-value = 0.00000 | | | | | |
| | *EA* | *EG* | *EI* | *ES* | *EO* |
| *AC0* | 75 | 107 | 29 | 15 | 17 |
| *AC1* | 79 | 19 | 73 | 36 | 21 |
| *AC2* | 128 | 140 | 180 | 137 | 150 |
| $\chi^2 = 172.9834$, df = 8, p-value = 0.00000 | | | | | |
| | *Id45* | *Id67* | | | |
| *AC0* | 150 | 93 | | | |
| *AC1* | 112 | 116 | | | |
| *AC2* | 328 | 407 | | | |
| $\chi^2 = 21.3811$, df = 2, p-value = 0.00002 | | | | | |

On the first factor plane in Figure 1 left the pattern of *Accuracy* levels follows approximately the one of the *Type of relative*, namely *TrD - TrP - TrR* progressively reducing the quality. More complex results its relation with the *Phrase structure*: *Ac0* in on the same side of *ES* and *EO*, and of *EG* on the second factor; *Ac1* far from *EG*, and *Ac2* opposite to *ES* and *EO* . For what concerns the *Child's age*, its levels are distinguished on the third axis (Figure 1 right), along which no noticeable variation of *Accuracy* results. Note that, albeit the second factor is not significant, the position of the *Accuracy* levels may be interpreted in the usual way, as their position is at the weighed centroid of the other character's levels.

In order to get a more clear vision of the dependence of *Accuracy* on the other characters, we ran both *SCA*s on the three contingency tables crossing *Accuracy* with the other characters and another one on the table obtained by juxtaposing the three tables crossing.

The *SCA* of the table crossing *Accuracy* with *Type of relative* has only one factor highly significant, along which over 97% of total inertia is distributed. Looking at the pattern of the levels in Figure 2 left it is weird to note that the accuracy levels are not in order: the sequence is *Ac1, TrP, Ac0, TrD, Ac2, TrR*, in agreement with cross data of Table 3 above. The *SCA* of the table crossing *Accuracy* with *Phrase structure* has both factors highly significant. Looking at the pattern of the levels in Figure 2 right an evident arch effect is visible, along which the sequence is *Ac0, EG, EA, Ac1, EI, ES, Ac2, EO*, in agreement with cross data of Table 3 center. Eventually, the one-dimensional *SCA* of the table crossing *Accuracy* with *Child's age* is highly significant and the sequence along the only factor is *Ac0, Id45, Ac1, Id67, Ac2*, in agreement with cross data of Table 3 below: as the solution is one-dimensional, no graphic is possible. It is evident that the behavior of *Accuracy*, ordered in respect with both *Phrase structure* and

*Child's age* is different in respect to *Type of relative*. Thus, the examination of *SCA* of the table crossing *Accuracy* with the other characters should explain this complicated pattern.

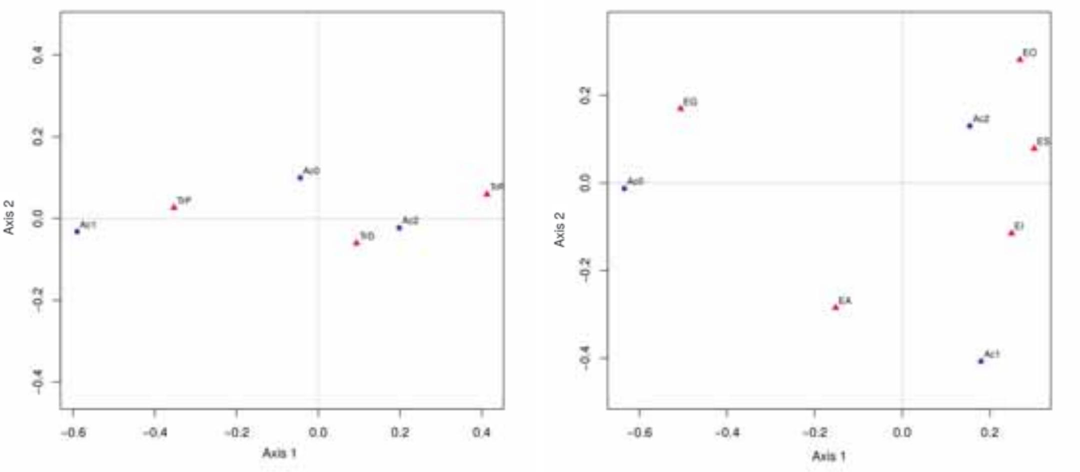

*Figure 2: The first factor planes issued from SCAs of two-way cross-tables. Left: Accuracy with Type of relative; right: Accuracy with Phrase Structure. Symmetric graphics.*

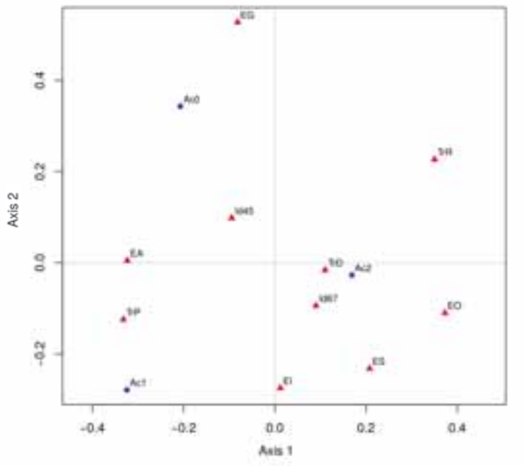

*Figure 3: The first factor plane of the SCA of the table crossing Accuracy with the three other characters Type of relative, Phrase structure, and Child's age. Symmetric graphics.*

Indeed, in this case too both factors are highly significant and of comparable magnitude. Looking at Figure 3 an arch effect is clearly visible, along which the correct sequence of *Accuracy* is clearly visible, as well as the one of *Phrase structure*. On the opposite, the *Child's age* levels are opposed, *Id45* on the side of *Ac0* and *Id67* on the side of *Ac2*, due to the little prevalence of the smallest children in answering poorly. Orthogonal to this is the sequence of the *type of relative*, whose sequence *TrP, TrD, TrR* is nearly aligned along the main diagonal, with *TrP* close to *Ac1*, *TrD* to *Ac2*, and *TrR* far from all, in particular orthogonal to *ES* and *EO*.

This may help in understanding the general structure of the data, but we may not ignore the unbalance of the design, due to the incompatibility of *TrR* with two structures. Thus, in order to try to model the relation between *Accuracy* and the other characters, we had rather carry out two different studies, one limited to the structures *ES, EO* and another to *EA, EG, EI*. In Table 4 the multiple contingency tables on which the studies will be carried out are reported and In Figure 4 the first principal planes of the *SCAs* of the two studies are shown.

*Table 4: Multiple contingency tables of the characters involved in the two separate analyses. Left: analysis on ES - EO; right: analysis on EA - EG - EI.*

| | | | *Accuracy* | | |
|---|---|---|---|---|---|
| *T.relative* | *Structure* | *Age* | *Ac0* | *Ac1* | *Ac2* |
| *TrD* | *ES* | *Id45* | 4 | 19 | 23 |
| | | *Id67* | 4 | 15 | 29 |
| | *EO* | *Id45* | 7 | 8 | 31 |
| | | *Id67* | 6 | 11 | 31 |
| *TrP* | *ES* | Id45 | 4 | 1 | 41 |
| | | Id67 | 3 | 1 | 44 |
| | *EO* | *Id45* | 3 | 1 | 42 |
| | | *Id67* | 1 | 1 | 46 |

| | | | *Accuracy* | | |
|---|---|---|---|---|---|
| *T.relative* | *Structure* | *Age* | *Ac0* | *Ac1* | *Ac2* |
| *TrD* | *EA* | *Id45* | 18 | 10 | 18 |
| | | *Id67* | 11 | 3 | 34 |
| | *EG* | *Id45* | 15 | 2 | 29 |
| | | *Id67* | 6 | 1 | 41 |
| | *EI* | *Id45* | 7 | 6 | 33 |
| | | *Id67* | 4 | 1 | 43 |
| *TrP* | *EA* | *Id45* | 11 | 29 | 6 |
| | | *Id67* | 9 | 32 | 7 |
| | *EG* | *Id45* | 32 | 2 | 4 |
| | | *Id67* | 26 | 13 | 1 |
| | *EI* | *Id45* | 12 | 27 | 7 |
| | | *Id67* | 0 | 37 | 11 |
| *TrR* | *EA* | *Id45* | 14 | 4 | 28 |
| | | *Id67* | 12 | 1 | 35 |
| | *EG* | *Id45* | 19 | 1 | 26 |
| | | *Id67* | 9 | 0 | 39 |
| | *EI* | *Id45* | 4 | 2 | 40 |
| | | *Id67* | 2 | 0 | 46 |

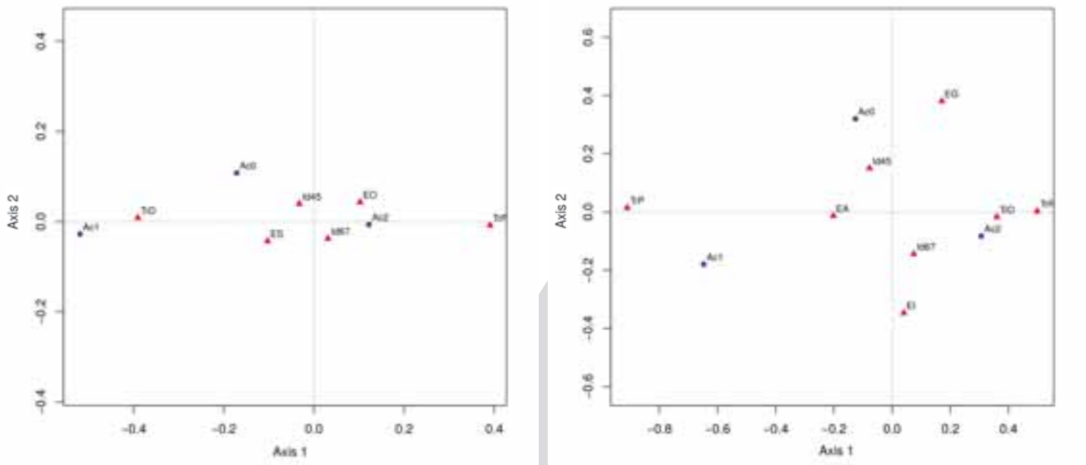

*Figure 4: The first factor planes issued from SCAs of the two reduced studies. Left: Phrase structures ES - EO; right: Phrase structures EA - EG - EI. Symmetric graphics.*

## 3.1 Analysis on ES - EO

In this study, 376 tests were involved and the frequency tables are reported in Table 5.

*Table 5: Frequency tables of the characters involved in the Analysis on ES - EO,*

| *Accuracy* | | | *type of relative* | | *Structure* | | *Age* | |
|---|---|---|---|---|---|---|---|---|
| *Ac0* | *Ac1* | *Ac2* | *TrD* | *TrP* | *ES* | *EO* | *Id45* | *Id67* |
| 32 | 57 | 287 | 188 | 188 | 188 | 188 | 184 | 192 |

In Figure 4 left the first factor plane of *SCA* is shown, but only the first dimension is significant. It is evident the same complicated pattern of *Accuracy* that was found in the overall study: indeed, the most relevant aspect is the higher frequency of *Ac1* in connection with *TrD*. The construction of a multinomial logistic model resulted very small, namely describing *Accuracy* only through *Type of relation* and *Structure*. Indeed, the quality of correct attributions, over 76%, resulted in an attribution table in which all tests had been attributed to *Ac2*, clearly unacceptable. Looking at the multiple contingency table, shown in Table 4 left, it is clear the reason: this is by far the most likely level in any case.

### 3.2 Analysis on EA - EG - EI

In this study, 830 tests were involved and the corresponding frequency tables are reported in Table 6. In Figure 4 right the first factor plane of *SCA* is shown, with both dimensions significant.

*Table 6: Frequency tables of the characters involved in the Analysis on EA - EG - EI.*

| *Accuracy* | | | *type of relative* | | | *Structure* | | | *Age* | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Ac0* | *Ac1* | *Ac2* | *TrD* | *TrP* | *TrR* | *EA* | *EG* | *EI* | *Id45* | *Id67* |
| 211 | 171 | 448 | 282 | 266 | 282 | 282 | 266 | 282 | 406 | 424 |

Here it seems that the *Accuracy* levels are vertices of a triangle, with the proximity of the other levels indicating their influence. Thus, *Id45, EA*, and *EG* are close to *Ac0*, *TrP* is close to *Ac1*, and the others to *Ac2*. The multinomial logistic model this time resulted much more complicated, as it involved both interactions of *Type of relation* with *Phrase structure* and *Child's age*. In this case, the table of correct attributions results much more interesting, albeit the percentage of correct attributions is lower, 71.69%, than in the previous case (Table 7).

*Table 7: Cross-tabulation of observed and attributed Accuracy in the analysis on ES -EO. Rows: observed Accuracy; column: predicted.*

| | *Ac0* | *Ac1* | *Ac2* |
|---|---|---|---|
| *Ac0* | 27.49 | 15.17 | 57.35 |
| *Ac1* | 8.77 | 73.10 | 18.13 |
| *Ac2* | 1.12 | 6.92 | 91.96 |

Indeed, *Ac2* is predicted very well by the model, with nearly 92% of correct attributions, and *Ac1* is well predicted too with over 73%: the problems result with *Ac0*, that is correctly attributed only in 27.5% of cases. This is due to the fact that the probability of *Ac0* is higher than the other levels only for the interaction *TrP-EG*, as it may be seen by the multiple Table 4 right: thus, only in this case it would be correctly predicted.

## 4 Conclusions

This first attempt to deal with both Correspondence Analysis, Log-linear and Multinomial Logistic Models, showed a possible synergy among the techniques, when appropriately applied. Indeed, in this study still in progress, we used both *MCA* and *LLM* to check the quality of the experimental design and both *MCA* and *MLM* to try to model the obtained *Accuracy*. The possible dependence among answers given by the same individual (random effect) was not taken into account, as no correspondence results in *MCA*. Whereas the quality of the design was clearly assessed, the results of *MLM*s did not cope with our expectations: on one side they were difficult to interpret (their interpretation is still in progress) and on the other prediction resulted limited. Indeed, this depends upon the data structures involved, but in general, through *MLM* only probability distributions are estimated. Thus, a correct prediction of a level may be obtained only if the probability of each criterion's level is the highest for some combination of the explanatory ones. Eventually, attention may be paid to random effects, not yet taken into account in the present study: this could help in deepening the subject.

**Acknowledgements**

This work was mostly carried out during the reciprocal visits of the two authors in the framework of their bilateral agreement between Sapienza Università di Roma and Universidade Federal do Rio de Janeiro. All institutions grants are gratefully acknowledged.

## 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.

Abreu, A.C.B. (2013). *Aquisição de orações relativas no português brasileiro*. Dissertação de mestrado, Linguistica, UFRJ.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), 716-623.

Ben Ammou, S., Saporta G. (2003). On the connection between the distribution of eigenvalues in multiple correspondence analysis and log-linear models. *REVSTAT-Statistical Journal*, 1.

Camiz, S., Gomes, G.C. (2013). Multiple and Joint Correspondence Analysis: Testing the True Dimension of a Study. *Modulad*, 44, 1-21.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. New York, Springer.

Diessel, H., Tomasello, M. (2000), The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, 11, p.131-151.

Diessel, H., Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, 81(4), 882-906.

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

Goodman, L.A. (1985). The Analysis of Cross-classified Data having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables with or without Missing Entries. *The Annals of Statistics*, 13(1), 10-69.

Greenacre, M.J. (1983). *Theory and Application of Correspondence Analysis*. London, Acad. Pr.

Greenacre, M.J. (1988). Correspondence analysis of mutlivariate categorical data by weighted least squares. *Biometrika*, 75, 457-467.

Kendall, M.G., Stuart, A. (1961). *The Advanced Theory of Statistics*, vol. 2. London, Griffin.

Kuehl, R.O. (2000). *Design of Experiments: Statistical principles in Research Design and Analysis*. Pacific Grove (CA), Duxbury Press.

Lebart, L., Salem, A. (1994). *Statistique Textuelle*. Paris, Dunod.

McEnery, T., Hardie, A. (2012). *Corpus Linguistics. Method, Theory and Practice*. Cambridge, Cambridge University Press.

McEnery, T., Wilson, A. (1996). *Corpus Linguistics*. Edinburgh, Edinburgh University Press.

Mollica, M.C. (2003). *Relativas em tempo real no português contemporâneo*. In: Paiva, M.C., Duarte, M.E.L. (Eds.), *Mudança Linguística em Tempo Real*. Rio de Janeiro, Contracapa, 129-138.

Orlóci, L. (1978). *Multivariate Analysis in Vegetation Research*, 2nd ed. Den Haag, Junk.

Tarallo, F. (1993). *Diagnosticando uma Gramática Brasileira: O português d'aquém e d'além mar ao final do século XIX*. In: Roberts, I., Kato, M. (Eds.), *Português Brasileiro: uma viagem diacrônica. Homenagem a Fernando Tarallo*. Campinas, Unicamp, 69-100.

Venables, W.N., Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York, Springer.