ON THE SELECTION OF IMAGES FOR A RECOGNITION TEST USING CORRESPONDENCE 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
Cidade Universitária, RJ
fonofernandasenna@gmail.com

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

RESUMO

Esse estudo consiste na avaliação de imagens a serem usadas em um teste com pacientes afásicos, no qual eles são solicitados a nomear objetos apresentados através de imagens específicas. A produção dos afásicos obtida através de testes provê uma maneira de avaliar a busca lexical e a extensão do comprometimento do conhecimento fonológico dos afásicos em função do dano cerebral. As imagens, disponíveis em Snodgrass e Vanderwarts (1980), precisam ser facilmente e inequivocamente identificadas e um amplo consenso precisa ser estabelecido para as imagens usadas com a finalidade mencionada. Análise de Correspondência foi utilizada para avaliar o julgamento das imagens por 38 sujeitos, para identificar até que ponto eles identificaram a imagem com o mesmo nome definido pelo pesquisador.

PALAVARAS CHAVE. Anáise de Correspondêcias, Anáise Exploratóia, Linguítica

ABSTRACT

This study concerns the evaluation of images to be used in a test with aphasic patients, in which these are asked to verbalize objects submitted to them through specific images. The results of the aphasic obtained by this test provides a mean to evaluate their word retrieval and the extent of the compromising of their phonological knowledge, due to brain damage. The images, taken from Snodgrass and Vandewarts (1980), must be easily and unequivocally identified and a broad consensus must exist on the set of images used for the mentioned purpose. Correspondence Analysis was applied to evaluate the images' judgments by 38 subjects, in order to understand to what extent they identify the image with the same name defined by the researcher.

KEYWORDS. Correspondence Analysis. Exploratory Analysis. Linguistics.

1. Introduction

In order to build a test for the association of names to images, a pilot study was carried out by selecting 260 images, whose name was defined by the researcher, and submitted to a panel of 38 judges, in order to understand to what extent they identify the image with the same name defined by the researcher.

From the answers provided by the 38 judges a data table was built, composed by one, if the name given by the judge was the same given by the researcher, and zero, if not. In this work we try to identify which of the 266 images, with the associated name, could be used for a subsequent study, aiming at testing the ability of afasic patients to identify and correctly tell the images' names.

The figures were selected among the ones available in Snodgrass ans Vanderwarts (1980). The authors proposed a set of 260 pictures for use in experiments investigating differences and similarities in the processing of pictures and words. The pictures were selected according to a set of rules that provide consistency of pictorial representation. In Snodgrass and Vanderwatts experiment, 4 groups of subjects were presented to four different sequences of the 260 slides and performed four different tasks in order to evaluate the drawings. They were asked to identify each picture as briefly and unambiguously as possible by writing only one name. They also judged the familiarity of each picture "according to how usual or unusual the object is in your realm of experience." (p. 183), the visual complexity of the image using a scale of 5-points in which 1 indicated very simple and 5 indicated very complex, with the complexity defined as the amount of detail or intricacy of lines in the picture. Subjects provided judgments about image agreement in order to establish how closely each picture resembled their mental image of the object. As a result the pictures have been standardized on these four variables of central relevance to memory and cognitive processing, name agreement, image agreement, familiarity, and visual complexity, and have been used in different tasks that associate pictures and words.

The results of the present evaluation of the figures will provide a selection of the best figures to be used in a naming task to be applied to aphasic individuals and to a control group. Failure in word retrieval is a pervasive characteristic of aphasic patients. Unsuccessful attempts result in substitutions that correspond to phonological changes in the target, neologisms that retain some phonological similarities with the target or with little relationship with the target, semantic substitution, morphological substitution and non-related substitutions. There are some evidences that there is a relationship between the amount of error type and the depth of the lesion (Dell et al., 1997). Studies have observed the degree of phonological overlapping between target and error observed in different types of aphasia (cf. Bose at al., 2007). This investigation evaluates an index of phonological overlapping between target and error that detects the degree of segmental and syllabic similarities taking into consideration the linear order of these phonological units. The purpose is to capture the phonological complexity of substitutions in accessing phonological information in the lexicon. Considering a lexicon as proposed by Usage-based Models, organized as a network of lexical relationships based on phonetic and semantic similarities, and phonological grammar as a ladder of levels emerging from the word-forms stored in the lexicon (Pierrehumbert, 2003), the phonological distance between target and error will show the information kept from the activated target which can be seen as an indication of the mechanism used in the retrieval.

2. Methodology

Simple Correspondence Analysis (Benzécri et coll., 1973-82; Greenacre, 1983; Langrand and Pinzón, 2009) is a special kind of Principal Component Analysis, specifically developed to study contingency data tables crossing two qualitative variables, but nowadays applied to any table of positive values. We do not step here into discussion of such choices: suffice here to say that some of its features find their rationale (and their utility) in the decomposition of the chi-square in

independent components, that loses much of its sense once applied to tables whose chi-square is not significant or in which the chi-square statistics loses its sense (as in our case).
SCA is based on the so-called *generalized singular value decomposition* (*GSVD*, Greenacre, 1983; Abdi, 2007), that allows to set conditions to both rows and columns of a matrix A through two positive definite square matrices. If M and W are such matrices, thus A is decomposed as $A = \tilde{U}\tilde{\Lambda}^{1/2}\tilde{V}'$ with the constraints $\tilde{U}'M\tilde{U} = I$ and $\tilde{V}'W\tilde{V} = I$. Such generalization is necessary to cope with the methods' special metrics, the chi-square, whose adoption leads to its decomposition. *GSVD* of A is obtained through the *SVD* of the matrix $\tilde{A} = M^{1/2}AW^{1/2}$ giving $\tilde{A} = U\Lambda^{1/2}V'$, and setting $\tilde{U} = M^{-1/2}U$ and $\tilde{V} = W^{-1/2}V$ to get the searched solution. In case of *SCA* of a contingency table $F = f..P$ crossing a m-levels character by row with a q-levels one by column, with $f..$ the table grand total and P a $m \times q$ matrix of probabilities (such that $\sum\sum p_{mq} = 1$), it results in defining two diagonal matrices D_m and D_q containing the marginal row- and column-profiles of F respectively. Then the *GSVD* of F results by solving $P = \Phi\Lambda^{1/2}\Psi$, with $\Phi' D_m^{-1}\Phi = I$ and $\Psi' D_q^{-1}\Psi = I$ or, that is the same, $\Phi'\Phi = D_m$ and $\Psi'\Psi = D_q$ respectively. As it may be proved that the highest singular value equals 1, after some manipulation the reconstruction formula of F results from that of P as $f_{ij} = f..p_{ij} = n..p_{i.}p_{.j}\left(1 + \sum_{\alpha=1}^{min(m,q)-1}\lambda_\alpha^{-1/2}\Phi_{i\alpha}\psi_{j\alpha}\right)$.

As the λ_α are sorted in decreasing order, the contribution to the reconstruction of the corresponding factors $\Phi_{i\alpha}\psi_{j\alpha}$ to the chi-square becomes negligible, so that the first few factors may be taken into account. This leads to a graphical representation of the levels of the two crossing variables on reduced dimensional spaces, whose proximity is usually interpreted in terms of *association*: the closest the levels on the space of representation the more similar is their behaviour in respect to the other levels, thus the more associated they are.

3. The Analysis

3.1 A simple approach

To cope with our aims, we first sorted both rows and columns of the data table according to both rows and columns total of 1s. It resulted that 66 words summarized all 38 correct wording, whereas 10 were always wrong. We also identified three judges, namely 13, 17, and 34, that gave the same name to less than 75% of the images, and we withdrew them. Then, we re-sorted the words considering only the 35 remaining judges. It resulted that 90 images were correctly identified by all judges, 25 by all but one, 24 by all but two, and 13 by all but three. Thus, 115 images were correctly identified by more than 95% of the 35 selected judges (indeed, 97%) and 152 were correctly identified by more than 90% of the judges (91%). The list of the selected items is reported in Appendix 1.

3.2. A better insight

For a more insightful selection, we applied *SCA* to the data table in order to take advantage of the method's features. Indeed, *SCA* can show on optimal planes of representation the scattering of both judges and names. This allows to better identify which judges to withdraw, as those whose behavior might be considered in some way different from the others. We tried through graphics to identify which of these may be considered deviant.
For this first *SCA* we removed both the 66 items universally recognized (certainly accepted in the selection) and the 10 never recognized (certainly rejected), as no information could be extracted by them. As a result, three factors resulted sufficiently larger than the following to be taken into

account, summarizing 22.78% of total variation. On the plane spanned by the first two factors (Figure 1) six judges appear further from the origin than all others, namely *P13, P17, P25, P32, P34, and P36*, thus showing a behavior somehow different from the others, whose pattern seems otherwise homogeneous. No particular remark results concerning the plane spanned by the factors 1 and 3, so that we limit the further considerations on the first factor plane.

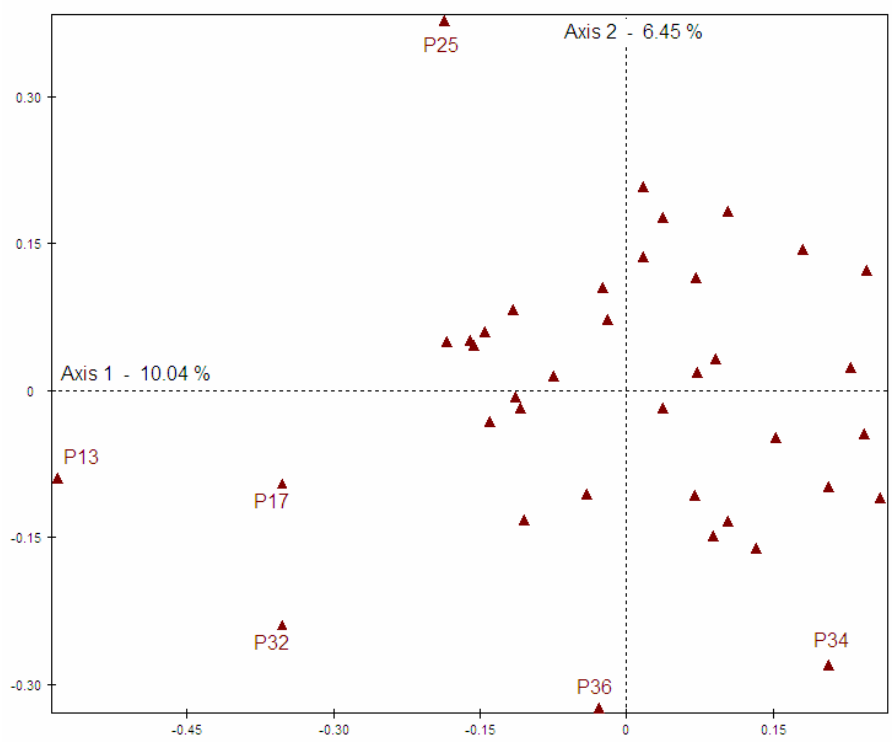

Figure 1 - The judges on the first factor plane of first SCA. Only the outliers are labeled.

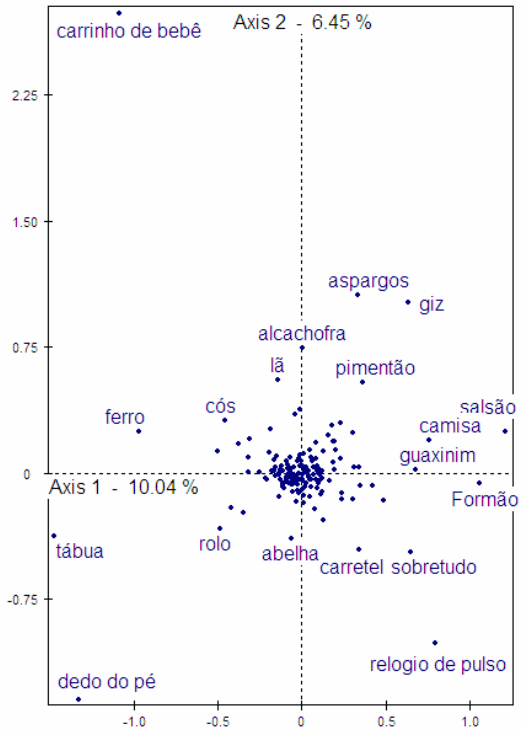

Figure 2 - The words on the first factor plane of the first SCA. Only the external are labeled.

The inspection of the scatter of the names on the same first plane (Figure 2) shows a pattern in which those further apart from the central cloud correspond to items very rarely identified: some of them, as *carrinho de bebé, dedo do pé, salsão,* and *gíz,* were identified by no more than 5 judges (1, 2, 3, and 5, respectively), but many others rarely identified resulted on the border of the cloud. This allowed us to interpret the *SCA* results as a progressive furthering from the center of the cloud of both judges and items related to the reduced number of correct identifications. Indeed, the directions in which the items are scattered may be interpreted in some way, but we did not deepened this issue, since it could not help in our choice.

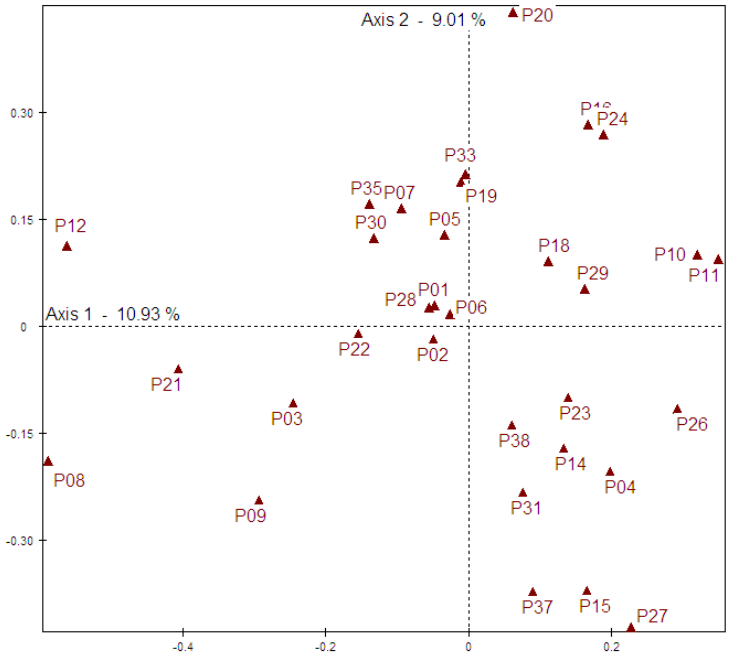

Figure 3 - The judges on the plane spanned by the first two factors of the second SCA.

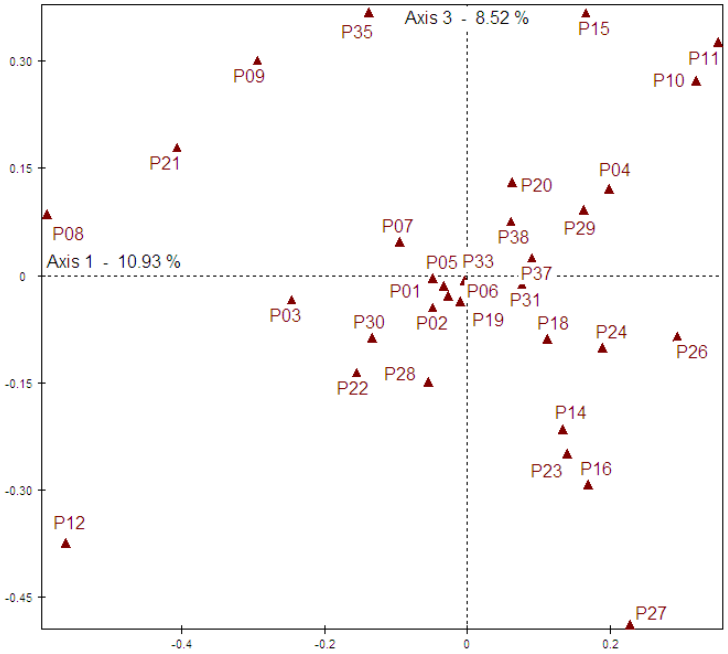

Figure 4 - The judges on the plane spanned by the factors 1 and 3 of the second SCA.

We had better re-run *SCA* once removed the six deviant judges, so that only 32 remained. For this task, we removed consequently the items to which all 32 judges answered correctly, but also all the items whose frequency of correct identification by these judges was lower than 50%. Thus only 55 items were considered, their correct identification ranging within 17 and 31 judges. The resulting *SCA* shows three main factors summarizing 28.46% of total inertia. The pattern of the judges on the first factor plane (Figure 3) this time is much more homogeneous, without evident outliers, little present only on the plane spanned by the factors 1 and 3 (Figure 4) but not so evidently.

As well, the scatter of the items on the same factor planes (Figures 5 and 6) do not show evident departures from a homogeneous scattering. For this reasons we may conclude that no further removal of judges seems necessary.

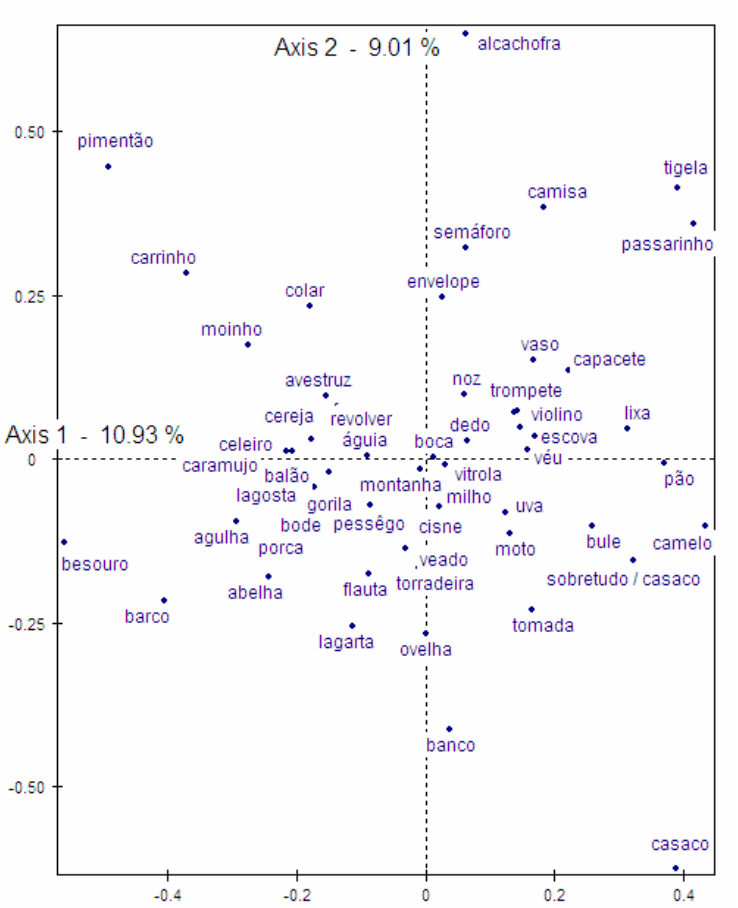

Figure 5 - The names on the plane spanned by the first two factors of the second SCA.

On the opposite, a choice was necessary to define which items to use in the further study, but we did not think of interest to use the *SCA* results to carry it out. Thus, we decided to keep all the items that were correctly identified by at least 90% of them. As, based on 32 judges, 97 were identified by all of them, 26 by only 31 (97%), 24 by 30 (94%), 14 by 29 judges (91%), this 161 items were selected. They are listed in Appendix 2.

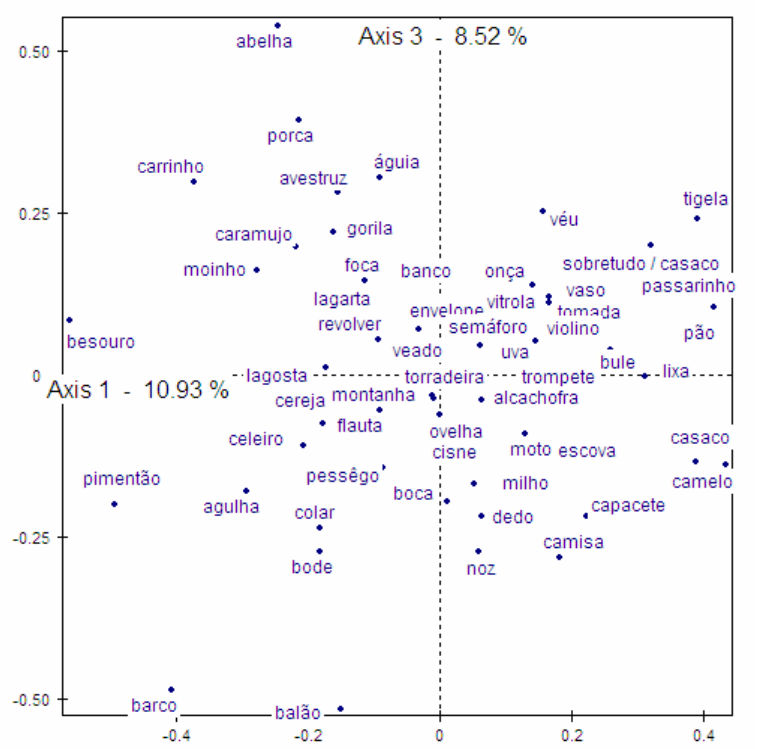

Figure 6 - The names on the plane spanned by the factors 1 and 3 of the second SCA.

4. Final remarks

Correspondence Analysis results allowed to identify some judges whose behavior was really different from all others. Thus we could remove them from the further study. On the opposite, the final selection of the items to take into account for further studies was simply based on a generalized consensus. We only allowed a small variation that in our opinion was not too important.

The choice that results in an important tool for selecting the words that will be used for the subsequent tests. Indeed, the naming tests for aphasic patients needs a background of general agreement on the name to associate to an image proposed to a patient. In fact, it is on this basis that word retrieval and phonological knowledge may give information on the patient's brain damage.

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.

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.

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

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

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.

Appendix 1 – The first selection

35 correct answers		91	flor	205	saia	***33 correct answers***	
173	abacaxi	5	formiga	25	sino	181	abobora
132	abajur	55	galinha	67	sofá	14	bola
4	âncora	32	garrafa	222	sol	74	boneca
187	anel	49	gato	197	tesoura	174	cachimbo
255	apito	185	geladeira	236	tomate	162	calça
212	aranha	103	girafa	21	urso	242	caminhão
241	arvore	232	gravata	78	vestido	45	canhão
2	avião	120	helicóptero	111	violão	122	casa
16	banana	57	igreja	70	xícara	219	fogão
180	batata	3	jacaré	260	zebra	133	folha
27	bicicleta	257	janela			175	jarro / jarra
178	bolsa	33	laço	***34 correct answers***		6	maçã
40	borboleta	138	lâmpada	37	alface	221	mala
41	botão	140	leão	90	bandeira	81	pato
116	cabide	30	livro	42	bolo	134	perna
143	cadeado	146	lua	31	bota	171	piano
53	cadeira	106	luva	7	braço	253	poço
22	cama	115	mão	73	cachorro	98	raposa
167	caneta	114	martelo	35	caixa	60	relogio
126	canguru	211	meia	59	cigarro	8	seta
47	carro	252	melancia	26	cinto	244	tartaruga
121	cavalo	220	morango	150	cogumelo	228	televisão
157	cebola	155	nariz	104	copo	68	vaca
48	cenoura	105	óculos	191	galo	37	vassoura
87	cerca	86	olho	97	garfo		
118	chapeu	39	ônibus	117	harpa	***32 correct answers***	
10	cinzeiro	83	orelha	168	lápis	18	barril
209	cobra	94	pé	77	maçaneta	113	cabelo
182	coelho	89	peixe	63	palhaço	58	charuto
247	colete	65	pente	198	parafuso	128	chave
215	colher	166	pera	149	rato	79	comoda
119	coração	238	pião	251	regador	230	dedal
69	coroa	161	pincel	186	rinoceronte	130	faca
52	corrente	169	pinguim	204	sapato	206	gambá
160	coruja	129	pipa	100	sapo	189	patins
84	elefante	172	porco	227	telefone	195	sanduiche
131	escada	76	porta	44	vela	258	taça
216	esquilo	151	prego			80	tambor
217	estrela	192	regua			240	trem

Note: the numbers refer to the original order in the data base.

Appendix 2 – The second selection

32 correct answers

173	abacaxi
132	abajur
4	âncora
187	anel
255	apito
212	aranha
241	arvore
2	avião
16	banana
180	batata
27	bicicleta
178	bolsa
40	borboleta
41	botão
116	cabide
143	cadeado
53	cadeira
35	caixa
22	cama
167	caneta
126	canguru
47	carro
121	cavalo
157	cebola
48	cenoura
87	cerca
118	chapeu
59	cigarro
26	cinto
10	cinzeiro
209	cobra
182	coelho
150	cogumelo
247	colete
215	colher
104	copo
119	coração
69	coroa
52	corrente
160	coruja
84	elefante
131	escada
216	esquilo
217	estrela
91	flor
5	formiga
55	galinha
191	galo
32	garrafa
49	gato
185	geladeira
103	girafa
232	gravata
120	helicóptero
57	igreja
3	jacaré
257	janela
33	laço
138	lâmpada
140	leão
30	livro
146	lua
106	luva
115	mão
114	martelo
211	meia
252	melancia
220	morango
155	nariz
105	óculos
86	olho
39	ônibus
83	orelha
198	parafuso
94	pé
89	peixe
65	pente
166	pera
238	pião
161	pincel
169	pinguim
129	pipa
172	porco
76	porta
151	prego
192	regua
205	saia
25	sino
67	sofá
222	sol
197	tesoura
236	tomate
21	urso
78	vestido
111	violão
70	xícara
260	zebra

31 correct answers

137	alface
90	bandeira
42	bolo
31	bota
7	braço
174	cachimbo
73	cachorro
162	calça
242	caminhão
206	gambá
97	garfo
117	harpa
175	jarro / jarra
168	lápis
77	maçaneta
221	mala
63	palhaço
189	patins
171	piano
149	rato
251	regador
186	rinoceronte
204	sapato
100	sapo
227	telefone
68	vaca
44	vela

30 correct answers

181	abobora
176	alicate
18	barril
14	bola
74	boneca
113	cabelo
45	canhão
122	casa
58	charuto
223	cisne
230	dedal
130	faca
219	fogão
133	folha
6	maçã
81	pato
134	perna
253	poço
98	raposa
60	relogio
8	seta
244	tartaruga
228	televisão
37	vassoura

29 correct answers

165	amendoim
225	balanço
75	burro
128	chave
79	comoda
101	frigideira
135	limão
12	machado
93	mosca
164	pavão
195	sanduiche
258	taça
80	tambor
240	trem

Note: the numbers refer to the original order in the data base.