Calidoscópio
18(1): 225-246 janeiro-abril 2020
ISSN 2177-6202
Unisinos - doi: 10.4013/cld.2020.181.13

Linguistic-computational Approach to the Tales of the Portuguese Oral Tradition: Lexicon and Values

Abordagem linguístico-computacional dos contos da tradição oral portuguesa: léxico e valores

Carlos Assunção[1]
University of Trás-os-Montes and Alto Douro
cassunca@utad.pt
https://orcid.org/0000-0002-5739-0754

Isabel Moreira[2]
Ministry of Education of Portugal
patrisaisabel@gmail.com
https://orcid.org/0000-0001-7577-5634

Abstract: The intrinsic relationship between oral tradition tales and the transmission of values is generally taken for granted. However, all publications focusing on this topic simply mention the important role that oral tradition literature plays in the transmission of values, and there is a very large gap in its objective study. The aim of this study is to identify how the lexicon linked to values, read from the Corpus Linguistics using a computational analysis of tales, is present in the oral tradition, contextualizing its occurrences either from a global perspective or from a comparative-contrastive perspective, in order to allow the critical reflection of the contents, their articulation with the necessary intervention in society and the vocabulary richness that the texts present. The authors who have contributed the most to Corpus Linguistics since the second half of the last century to the present will be identified in the Introduction. This work is the first reflection that is made, for what is known, between the Corpus Linguistics and the lexicon of the values of the tales of oral tra-

[1] Professor at the Department of Letters, Arts and Communication at the University of Trás-os-Montes and Alto Douro. Member of the Center for the Studies in Letters of the University of Trás-os-Montes and Alto Douro, a research unit funded by the Foundation for Science and Technology (UID/LIN/00707/2020).
[2] Basic Education Teacher at the Ministry of Education of Portugal. Member of the Center for the Studies in Letters of the University of Trás-os-Montes and Alto Douro, a research unit funded by the Foundation for Science and Technology (UID/LIN/00707/2020).

dition. The methodology used will consist of an analysis using automatic text analysis programs and the conclusions will be in line with the objectives described below.

Keywords: corpus linguistics; lexicon; values.

Resumo: É genericamente assumida a intrínseca relação entre os contos da tradição oral e a transmissão de valores. Contudo, todas as publicações dedicadas ao tema limitam-se a referir o importante papel que a literatura de tradição oral desempenha na transmissão de valores, existindo uma lacuna muito grande no que respeita ao seu estudo objetivo. Pretende-se com este estudo identificar de que forma o léxico ligado aos valores, lido a partir da Linguística de Corpus com recurso a uma análise computacional dos contos, está presente na tradição oral, contextualizando as suas ocorrências quer sob uma perspetiva global quer sob uma perspetiva comparativo-contrastiva, de forma a permitir a reflexão crítica dos conteúdos, a sua articulação com a necessária intervenção em sociedade e a riqueza vocabular que os textos apresentam. Serão identificados na introdução os autores que mais têm contribuído para a Linguística de Corpus desde a segunda metade do passado século à atualidade. Este trabalho é a primeira reflexão que se faz, pelo que se conhece, entre a Linguística de Corpus e o léxico dos valores dos contos da tradição oral. A metodologia utilizada consistirá numa análise a partir da utilização de programas de análise automática de textos e as conclusões estarão em consonância com os objetivos abaixo descritos.

Palavras-chave: linguística de *corpus*; léxico; valores.

Introduction

The objectives of this study are four: to make a new approach to the tales of the oral tradition from a linguistic perspective based on corpus linguistics, using the methods of linguistics and not of literature; check the connection between the lexicon and the values that the stories convey; extract from the corpus lexical fields and thematic fields, but also associative fields, including all the lexicon related or associated with values; assess the vocabulary richness of the tales.

The core hypothesis of this work is based on this question: is it possible to verify the values in the tales of the oral tradition from the corpus linguistics knowing that a corpus does not provide us with any information at this level? The answer to this question will be given in a reasoned way throughout the article because there are computer tools, namely at the software level, that applied to a corpus, offer us this possibility.

Currently, the use of automatic text analysis programs is not yet a common practice in this type of analysis, being limited to restricted research circles. However, this is an area whose growing importance and potential for the study of language in oral literature and even in the teaching of languages fully justify the whole effort of dissemination, so that more people become interested in investing and making it an added value in educational practice.

In fact with words, we build ourselves and make ourselves known. It is also with words that we reveal to the world who we are. Tales are also, and essentially, "revealed" by the words used and by the frequency with which they occur:

> A linguagem é um sistema probabilístico, cuja face mais notável é a frequência do uso das palavras […]. A frequência de uso (alta, baixa, intermediária), atributo inseparável da palavra, pois revela a sua ocorrência observada, tem um papel definidor da palavra, fornecendo um traço tão Ininseparável quanto o sentido [language is a probabilistic system, whose most remarkable characteristic is the frequency of words used [...]. The frequency of use (high, low, intermediate), an inseparable attribute of the word since it reveals its observed occurrence, has a defining role to play regarding the word, producing a feature as inseparable as the meaning]. (Berber Sardinha, 2004, p. 162).

Kennedy (1998), in *An Introduction to Corpus Linguistics*, recognizes the contribution of manual analyses of texts over the centuries, especially for lexicography, but clearly speaks of the advantage of computer use and the reliability of the results obtained in this second method over the first.

> The analysis of huge bodies of text 'by hand' can be prone to error and is not always exhaustive or easily replicable [...]. The Corpus Linguistics is thus now inextricably linked to the computer, which has introduced incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data. (Kennedy, 1998, p. 5)

Corpus linguistics has been the subject of several theoretical and theoretical-practical studies over the course of time. Researchers such as Stubbs (1993), Sinclair (2004), Leech (1992), Kennedy (1998), Tognini-Bonelli (2001), McEnery and Wilson (1996), Halliday (1967; 2006), Chomsky (1956), Teubert (2005), Meyer (2002), Bowker and Pearson (2002), McEnery, Xiao and Tono(2006), Berber Sardinha (2004), Gries (2010), Fadanelli and Monzón (2017), and Silberztein (2004; 2015) carried out studies on corpus linguistics and on the possibilities of its application to languages.

Approaching the tales of oral tradition from a linguistic perspective, using the tools of Computational Linguistics, allows us to pursue both the study of language and the linguistic study of language. In fact, to paraphrase Halliday, in talking of the linguistic approach to literary texts, we are not merely referring to the study of language, but rather to the study of such texts according to the methods of linguistics. There is a difference between making textual or linguistic ad hoc statements about literature that are personal in nature and arbitrarily selective—such as those that may appear in support of a preformulated literary thesis—and describing a text based on a general linguistic theory (Halliday, 1967).

A corpus on its own does not provide us with any information; it is only "a store of used language" (Hunston, 2002, p. 3). However, the software available for corpora analysis gives us the possibility to "re-arrange that store so that observations of various kinds can be made [...]. A corpus does not contain new information about language, but the software offers us a new perspective on the familiar" (Hunston, 2002, p. 3).

In order to better understand the axiological question, let us focus on the questions that Cabral Moncada asked in the preface to the translation of Johannes Hessen's work, *Filosofia dos Valores*:

> What are values? What kind of being are they? What is the ontic structure of this class of ideal objects, similar to numbers, that populate our spiritual consciousness and that seem not to be merely subjective but to have an objectivity of their own? What species and categories are there of values and what is their respective dignity in hierarchy? What means of knowledge do we have to apprehend them? Finally, in what relation do they find themselves with man, with life, with the spirit, with God?
> Here is the whole philosophical theme, largely metaphysical, which today is encompassed under the generic name of Axiology. (Hessen, 1980, p. 8, our translation)[3]

According to Hessen (1980, p. 107–110), values are classified from a double point of view: *formal* and *material*. From a *formal* point of view, values are divided as follows: *positive* and *negative*. A positive value is one we most commonly call the pure and simple expression of "value." The concept of "value" is generally used in a double sense: sometimes this word means value in general, regardless of the value–devalue polarity, as a neutral concept; and at other times we just take its positive rather than its negative aspect. The negative value is then called "devaluation." This polarity belongs to the essential structure of the axiological order, which is thus fundamentally distinguished from the order of the being that is strange to such a structure: people's values and values of things, or personal and real values. People's values, or personal values, are those that can only belong to people, such as ethical values. Real values (of *res*) are those that adhere to objects or impersonal things, such as those things said to be valuable, more generally known by the expression "goods"—values-in-themselves, or autonomous, and values derived from others or dependents. The value in itself (*Selbstwert*) lies in its very essence; it has this nature that is independent of all other values. It does not depend on them; it is not a means to them. It is counterbalanced by the *derived* value. This second value no longer owes to itself its valuable nature but draws it from another value. Its specific feature is to always be related to another, or to others. If it were not for these, it would cease to be a value. The values to which it refers are the values in themselves.

The author goes on to detail the classification that he developed regarding values, specifying those he considers likely to be included in sensitive values and spiritual values. In the former, he includes the subcategories of values of enjoyment and pleasure, vital values and utility values; in the latter, he establishes the subcategories of logical values, ethical values, aesthetic values, and religious values (Hessen, 1980, p. 110–120).

The classification developed by Hessen could be schematized as shown in Table 1.

[3] Que são os valores? Que espécie de ser lhes corresponde? Qual a estrutura ôntica desta classe de objectos ideais, parecidos com os números, que povoam a nossa consciência espiritual e que parecem não ser meramente subjectivos mas ter uma objectividade própria? Que espécies e categorias há de valores e qual a sua respectiva dignidade em hierarquia? Que meios de conhecimento temos para os apreender? Finalmente, em que relação se acham eles com o homem, com a vida, com o espírito, com Deus?
Eis toda a temática filosófica, em grande parte metafísica, que hoje se engloba sob a designação genérica de Axiologia.

Table 1. Values presented by Hessen. Systematization 1.

VALUES	Formal	positive and negative	
		personal and real	
		autonomous and dependent	
	Material	Sensitive	delightful and pleasurable
			Vital
			Useful
		Spiritual	Logical
			Ethical
			Aesthetic
			Religious

One of the most commonly used ways to classify values is that which is typically used in teaching values in philosophy, and can be schematized as in Table 2.

Table 2. Values presented by Hessen. Systematization 2.

VALUES	Ethical	referring to norms and behaviors such as solidarity, honesty, truth, loyalty, kindness, and altruism
	Aesthetic	values of expression, such as harmony, beauty, ugliness, sublimity, and tragedy
	Religious	relating to ideas of transcendence, such as sacredness, purity, and sanctity
	Political	examples include justice, equality, impartiality, citizenship, and freedom
	Vital	exemplified by health and strength

Let us move on to the methodological procedures and the results of the study.

Lexicon and values: Methodological procedures

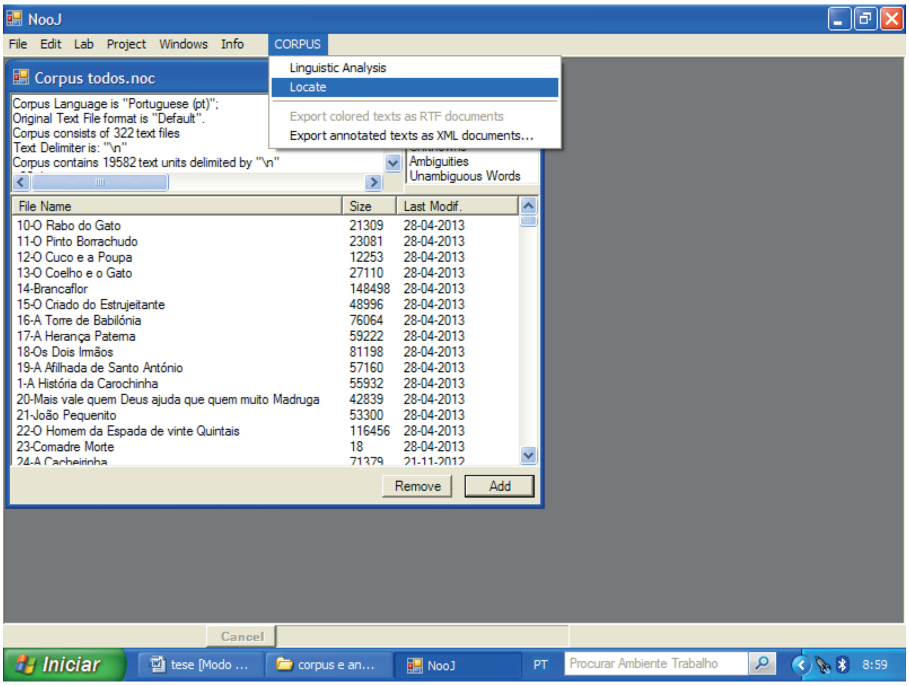
In light of the above, we begin this extraction process taking into account not only lexical fields and thematic fields, but also associative fields, including all lexicons related to or associated with what we seek to list, because otherwise we would not contemplate lexicons associated with value, that is, all references related or associated with values. A first one, related to the semantic field, having Lyons (1968) as its source, refers to the associative field astTerm used by some linguists to designate a set of lexical units that presents a particular similarity of forms and meanings among its components. And a second one, quoting Galisson and Coste (1983, p. 17, our translation), is related to the notional field:

> Term that expands the Saussurean notions of associative relations and associative chains and that, according to Bally and other linguists, designates the entire associative chains of a term or set of terms. The associative fields, which are linked to affective, intellectual and cultural factors and to each individual's experience, vary from speaker to speaker

> and can take place according to the dominant or exclusive axes, and they may even seem completely random from the linguistic point of view.][4]

In order to avoid the intuitive process of reading and interpreting it, we proceeded to the full lemmatization of the corpus,[5] aided by NooJ,[6] which allowed us to locate any word in context and thus assess its meaning there and in Excel. In the first part of this step, we opened the digital file of our lemmatization and we selected and grouped the lexicon of values, with positive and negative polarization, in a file created in Excel, containing two columns, one for the lemmas and another for the different forms of each lemma. In situations of uncertainty, we compared the occurrences in context, assessing their meaning, by opening the corpus in NooJ and checking each word in the corresponding text. For this purpose, we used the LOCATE function available in the program, as exemplified in Figures 1 and 2.

Figure 1. NooJ applied to the corpus: occurrences.

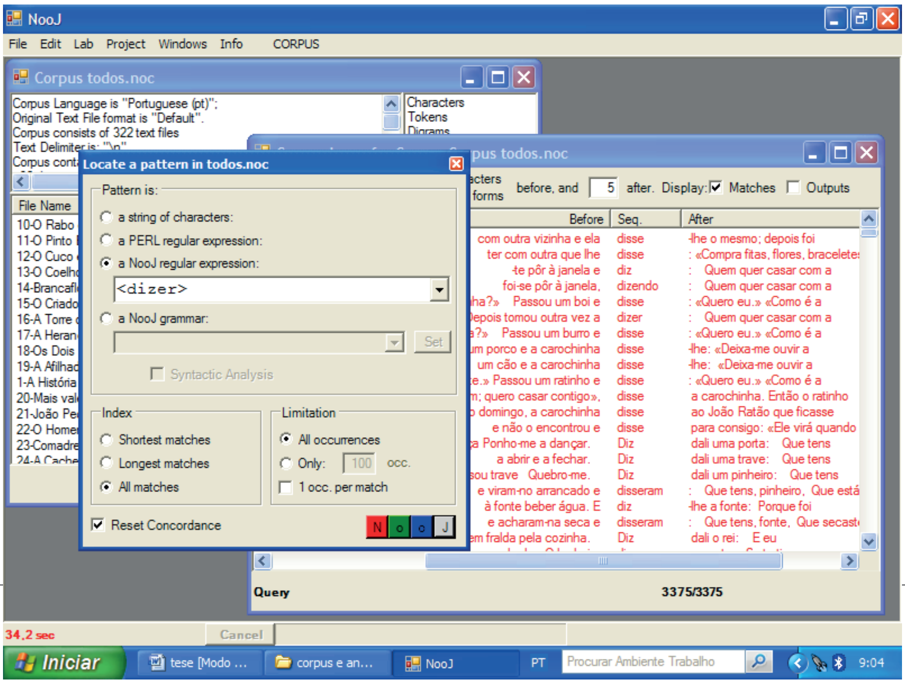


NooJ
File Edit Lab Project Windows Info CORPUS
Corpus todos.noc
Linguistic Analysis
Locate
Export colored texts as RTF documents
Export annotated texts as XML documents...
Ambiguities
Unambiguous Words
Remove
Add
Cancel
Iniciar
tese [Modo ...
corpus e an...
NooJ
PT
Procurar Ambiente Trabalho
8:59

Corpus Language is "Portuguese (pt)";
Original Text File format is "Default".
Corpus consists of 322 text files
Text Delimiter is: "\n"
Corpus contains 19582 text units delimited by "\n"

File Name	Size	Last Modif.
10-O Rabo do Gato	21309	28-04-2013
11-O Pinto Borrachudo	23081	28-04-2013
12-O Cuco e a Poupa	12253	28-04-2013
13-O Coelho e o Gato	27110	28-04-2013
14-Brancaflor	148498	28-04-2013
15-O Criado do Estrujeitante	48996	28-04-2013
16-A Torre de Babilónia	76064	28-04-2013
17-A Herança Paterna	59222	28-04-2013
18-Os Dois Irmãos	81198	28-04-2013
19-A Afilhada de Santo António	57160	28-04-2013
1-A História da Carochinha	55932	28-04-2013
20-Mais vale quem Deus ajuda que quem muito Madruga	42839	28-04-2013
21-João Pequenito	53300	28-04-2013
22-O Homem da Espada de vinte Quintais	116456	28-04-2013
23-Comadre Morte	18	28-04-2013
24-A Cacheirinha	71379	21-11-2012

Figure 2. NooJ applied to the corpus: lemmas



Corpus Language is "Portuguese (pt)";
Original Text File format is "Default".
Corpus consists of 322 text files
Text Delimiter is: "\n"
Corpus cont

File Name
10-O Rabo
11-O Pinto
12-O Cuco
13-O Coelho
14-Brancafl
15-O Criado
16-A Torre
17-A Heran
18-Os Dois
19-A Afilhad
1-A História
20-Mais val
21-João Pe
22-O Homer
23-Comadre
24-A Cache

Pattern is:
○ a string of characters:
○ a PERL regular expression:
● a NooJ regular expression:
<dizer>
○ a NooJ grammar:
□ Syntactic Analysis

Index
○ Shortest matches
○ Longest matches
● All matches

Limitation
● All occurrences
○ Only: 100 occ.
□ 1 occ. per match

☑ Reset Concordance

acters
forms before, and 5 after. Display: ☑ Matches □ Outputs

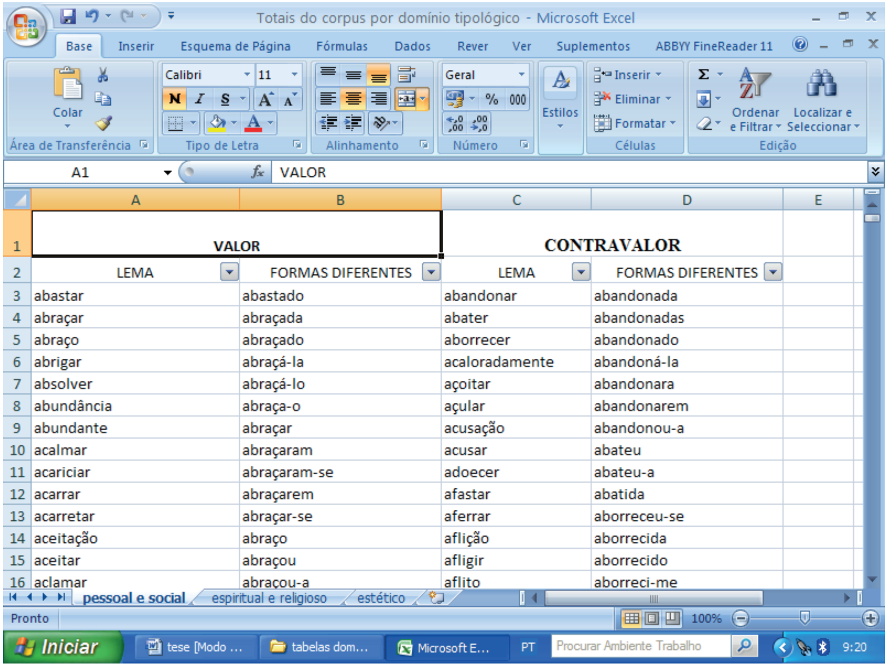
Before	Seq.	After
com outra vizinha e ela	disse	-lhe o mesmo; depois foi
ter com outra que lhe	disse	: «Compra fitas, flores, bracelete
-te pôr à janela e	diz	: Quem quer casar com a
foi-se pôr à janela,	dizendo	: Quem quer casar com a
ha?» Passou um boi e	disse	: «Quero eu.» «Como é a
epois tomou outra vez a	dizer	: Quem quer casar com a
a?» Passou um burro e	disse	: «Quero eu.» «Como é a
m porco e a carochinha	disse	-lhe: «Deixa-me ouvir a
um cão e a carochinha	disse	-lhe: «Deixa-me ouvir a
e.» Passou um ratinho e	disse	: «Quero eu.» «Como é a
m; quero casar contigo»,	disse	a carochinha. Então o ratinho
o domingo, a carochinha	disse	ao João Ratão que ficasse
e não o encontrou e	disse	para consigo: «Ele virá quando
a Ponho-me a dançar.	Diz	dali uma porta: Que tens
a abrir e a fechar.	Diz	dali uma trave: Que tens
sou trave Quebro-me.	Diz	dali um pinheiro: Que tens
e viram-no arrancado e	disseram	: Que tens, pinheiro, Que está
à fonte beber água. E	diz	-lhe a fonte: Porque foi
e acharam-na seca e	disseram	: Que tens, fonte, Que secast
em fralda pela cozinha.	Diz	dali o rei: E eu

3375/3375

34,2 sec

We searched for lemmas, in between angle brackets, so that the software would automatically list all occurrences of the lemma.

Figure 3. Corpus polarization



	A	B	C	D	E
1	VALOR		CONTRAVALOR		
2	LEMA	FORMAS DIFERENTES	LEMA	FORMAS DIFERENTES	
3	abastar	abastado	abandonar	abandonada	
4	abraçar	abraçada	abater	abandonadas	
5	abraço	abraçado	aborrecer	abandonado	
6	abrigar	abraçá-la	acaloradamente	abandoná-la	
7	absolver	abraçá-lo	açoitar	abandonara	
8	abundância	abraça-o	açular	abandonarem	
9	abundante	abraçar	acusação	abandonou-a	
10	acalmar	abraçaram	acusar	abateu	
11	acariciar	abraçaram-se	adoecer	abateu-a	
12	acarrar	abraçarem	afastar	abatida	
13	acarretar	abraçar-se	aferrar	aborreceu-se	
14	aceitação	abraço	aflição	aborrecida	
15	aceitar	abraçou	afligir	aborrecido	
16	aclamar	abraçou-a	aflito	aborreci-me	

Having made this first listing, which was a detailed process that took us a long time to complete given the length of our corpus, we moved to a second part. Here we created, in an Excel file, three additional sheets, one for each typological category, with two columns per sheet, one for the values with positive polarization, which we just called values, and another for the values with negative polarization (counter-values). Additionally, in each column pertaining to the values and counter-values, we opened two more columns, one for the lemmas and another for the different forms, as we had already done in the previous step. We then proceeded to select and group the selected lexicon of values into typological categories, distinguishing its polarization. This is illustrated in Figure 3.

During this selection process, we were faced with a problem regarding the column where we should insert some forms, because, depending on the context, they could be inserted in more than one, as they were simultaneously part of more than one category and even polarization. Having to make decisions so that we would not run the risk of repetition, which would result in changes in the final outcomes at the level of occurrences, we decided to place these forms according to their main context. Once this procedure was completed, the next step presented comparative results, accomplished by carrying out the appropriate procedures to match the lexicon that we had already obtained and listed, from the whole corpus, with each of the authors. One way to do this would be to use NooJ and go through each lemma, and each form, using the LOCATE function, to identify, for each author in the corpus, whether the form was present and the total number of occurrences. This brings us to the procedure described below.

To determine the occurrences of each form, after dividing these forms into categories, database tables were used. These tables were imported into the spreadsheet, which already included three sheets (one per category). Thus there were five sheets: full-text corpus data and each of the four authors.

Because, due to the ambiguity of some forms, these are found in different lines in the tables produced by NooJ (for example, for the lemma “abraço” [hug] there is a line for the noun and another for the verb), we created a formula capable of retrieving all this information and counting up the total occurrences of each word.Example:

=SOMA.SE(All.B2:B21451;B3;All.C2:C21451)

This formula uses the SOMA.SE function, which sums the value of the selected cells if the stated condition occurs, and within the parentheses are three arguments, divided by the “;” symbol. The first indicates the cell or range of cells to be searched (in this case, named “Todos” [All] in the B2:B21451 range of cells. The second indicates the condition to be checked, in this case the equality of content between cell B3 in this sheet that contains a form of the category under analysis and each of the cells in the range indicated in the first argument. The third indicates the cell, cells, or range of cells whose values must be summed up, provided that the condition stated in the range indicated is met.

Inserted in cell C3, this formula searches “Todos” [All] in the sheet (containing the full corpus data), in the range that covers cells from B2 to B21451 (column B in this sheet contains the listing of the different forms, which take 21,450 rows); it compares the content of each cell in this range and the content of cell B3 on the sheet in which the formula was entered (containing the word belonging to the category under analysis) and, if it finds one or more matches, it displays the sum of the respective occurrences, in the same row, on the right (column C, from cells C2 to C21451). If there is no match, the formula displays the value 0 (zero).

In cell C4, the formula will be similar in all respects, with only the second argument varying, which leads to another form of the list, and so on:

=SOMA.SE(All.B2:B21451;B4; All.C2:C21451)

To get data for each author, a variation of the same formula is used, with the same function, with only the references being changed. For example, to get data for Braga, the formula used is =SOMA.SE(TBraga.B2:B7086;B3;TBraga.C2:C7086), in which the name of the sheet to be searched and the range of cells to be analyzed were changed (because the different forms for this author only take 7,085 rows, it would be an unjustified overload of computer system resources to indicate more rows in the range than these).

Lexicon and values: Results

After completing all the procedures described in the previous item, we now present the results of our study, first in general terms and then more specifically, for each author, and also establishing comparisons among them. A first possible result is the one related to the lexicon and values (henceforth named “values”) in the corpus under study, the total number of occurrences related to values, the total number of lemmas, and the total of different forms, which we present in Table 3, comparing with the same totals in relation to the total lexicon (labeled as “general”).

Table 3. General and values: occurrences, lemmas, and different forms

GENERAL AND VALUES, CORPUS TOTALS			
	Occurrences	**Lemmas**	**Different forms**
General	196,970	4,526	12,473
Values	54,436	2,188	6,488

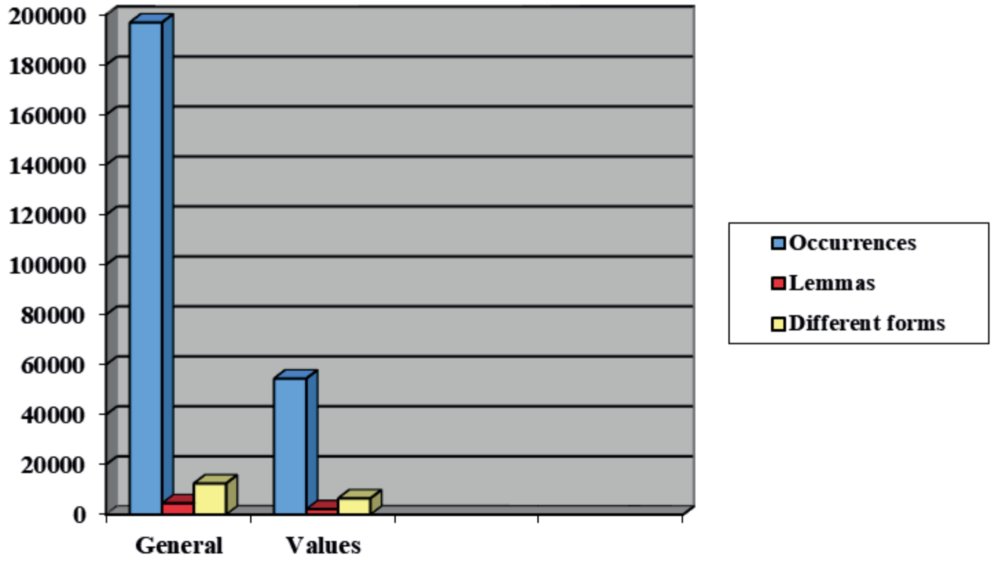
The total of different forms (12,473 for general, 6,488 for values) and the total of lemmas (4,526 for general, 2,188 for values) are but a small part of their overall occurrences (196,790 for general, 54,436 for values). However, as shown in Table 4, if we focus on the percentage results of the relation between the lemmas and the different forms with respect to the overall occurrences, for both general and values, we notice significantly higher percentages in the relationship between lemmas and different forms with the total number of occurrences, regarding both lexicon and values.

Table 4. General and values: percentages for lemmas and different forms

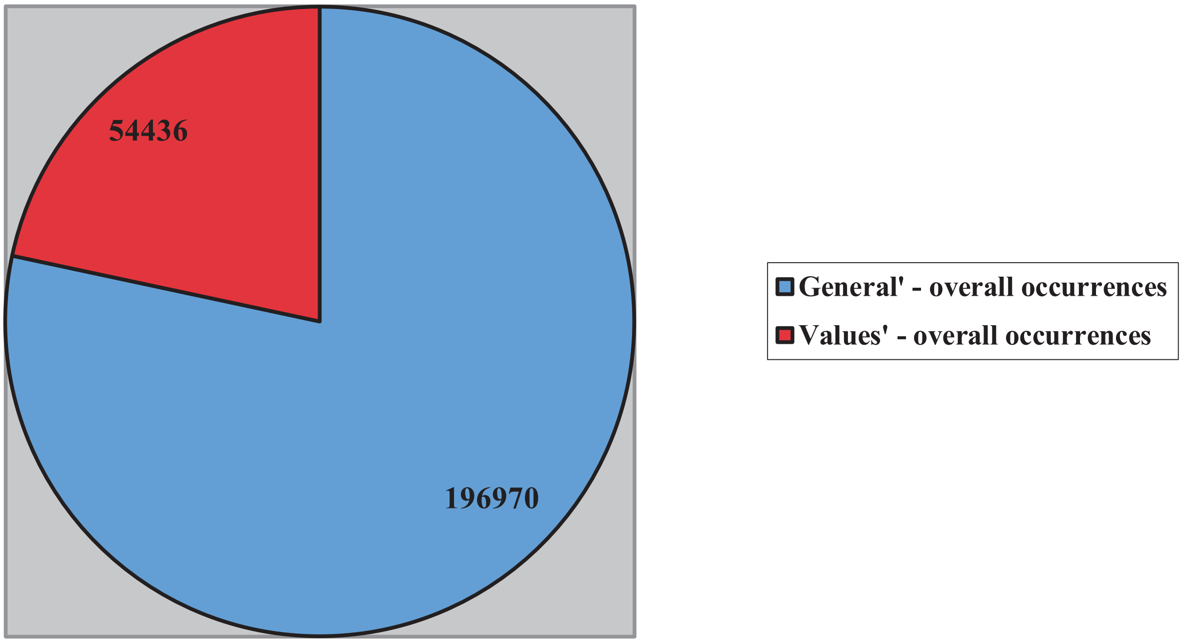
GENERAL AND VALUES, CORPUS TOTALS				
	General	**Average**	**Values**	**Average**
Occurrences	196,970		54,436	
Lemmas	4,526	2.298 %	2,188	4.019%
Different forms	12,473	6.332 %	6,488	11.919%

This percentage results will be referred to again in the section on lexical richness, but for now let us just focus on the discussion of results. The results of Tables 3–4 can be found in Graphs 1–2.

Graph 1. General and values: occurrences, lemmas, and different forms



Graph 2 illustrates the occurrences related to “values” versus all occurrences.



Comparing the percentages for total occurrences (values and general), in the corpus and per author, in order to assess the importance of the lexicon of values in the texts, we obtained the results shown in Table 5:

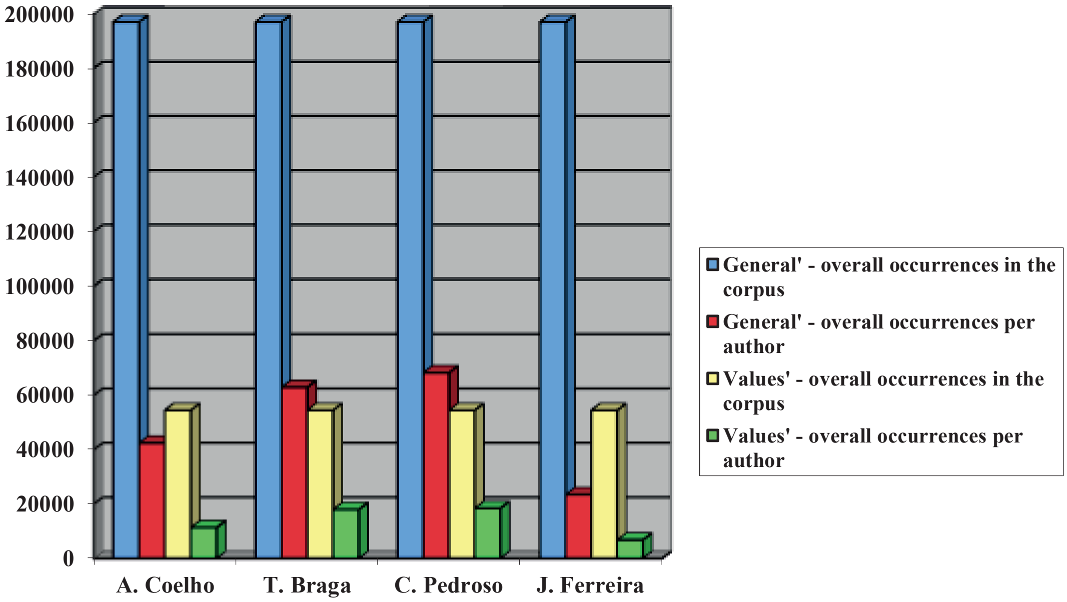
Table 5. General and values: overall occurrences in the corpus and by author

CORPUS AND AUTHORS, TOTALS			
	General occurrences	**Values Occurrences**	**Percentage**[7]
Corpus	196,970	54,436	27.64%
A. Coelho	42,382	11,312	26.69%
T. Braga	62,880	17,961	28.56%
C. Pedroso	68,244	18,397	26.96%
J. Ferreira	23,464	6,766	28.84%

The percentage results are reasonably close, both in relation to the corpus and to each author, ranging from 26.69% (Coelho) to 28.83% (Ferreira). These results, all slightly above one-quarter of each total, show the importance of the lexicon associated with values in the texts we studied. The variation of 2.14% shows that this behavior is similar among the authors.

Graph 3 shows a bar graph of the results from Table 5.

Graph 3. General and values: overall occurrences in the corpus and by author



Each author's contribution to the total percentage of values in the corpus is shown in Table 6 and Graph 4.

[7] Percentages for occurrences of values in relation to occurrences of general, in the corpus and per author. Since we do not have the same numeric value as a reference, we cannot make comparisons among them.

Table 6. General and values: total number of occurrences per author

GENERAL AND VALUES: OVERALL OCCURRENCES PER AUTHOR				
AUTHOR	**General occurrences**	**Values Occurrences**	**Percentage1**[8]	**Percentage2**[9]
A. Coelho	42,382	11,312	5.74%	20.78%
T. Braga	62,880	17,961	9.12%	32.99%
C. Pedroso	68,244	18,397	9.34%	33.80%
J. Ferreira	23,464	6,766	3.44%	12.43%
Corpus, grand total	**196,970**	**54,436**	**27.64%**	**100%**

Lexicon and values: Positive and negative polarization

Results concerning polarization are shown in terms of numbers and percentages (Table 7) and as a pie chart (Graph 5).

Table 7. Corpus totals for values: positive and negative polarization

VALUES, CORPUS TOTALS		
POLARIZATION	**Corpus**	**Percentage**[10]
Positive	40,106	73.68%
Negative	14,330	26.32%
GRAND TOTAL	54,436	100%

The results show the supremacy of the values with positive polarization over the values of negative polarization, as shown in Graph 5.

Table 8. Positive and negative polarization totals regarding values: occurrences, lemmas, and different forms

VALUES, CORPUS TOTALS			
POLARIZATION	**Occurrences**	**Lemmas**	**Different forms**
Positive	40,106	1,212	4,051
Negative	14,330	976	2,437
GRAND TOTAL	54,436	2,188	6,488

As to polarization in terms of the authors under study, the results are shown in Table 9 and Graph 6.

[8] Calculated in terms of the total number of occurrences in the corpus.

[9] Calculated in terms of the total number of occurrences for values.

[10] Calculated to determine the relationship between positive and negative polarisation values in the corpus.

Table 9. Values totals per author: occurrences, positive and negative polarization

VALUES, TOTALS PER AUTHOR: occurrences, positive and negative polarization								
	Coelho	**%[11]**	**Braga**	**%**	**Pedroso**	**%**	**Ferreira**	**%**
Positive polarization	8,456	74.75	13,305	74.08	13,442	73.07	4903	72.47
Negative polarization	2,856	25.25	4,656	25.92	4,955	26,93	1863	27.53
Grand total	11,312	100%	17,961	100%	18,397	100%	6766	100%

Similarly to what we discussed when focusing on the corpus, here we also see the supremacy of values with positive polarization over values with negative polarization. In percentage results, values with positive polarization ranged between 72.47% (Ferreira) and 74.75% (Coelho), and, obviously, the percentages indicating negative polarization are reversed and range between 25.25% (Coelho) and 27.53% (Ferreira). A very interesting finding is the percentage difference in terms of polarizations displaying minimum and maximum values: 49.5% (Coelho) and 44.94% (Ferreira). In other words, a century later, the supremacy of values with positive polarization over those with negative polarization is reduced by 4.56 percentage points. The differences between polarizations, by author, are shown in Table 10.

Table 10. Authors' totals: percentages for polarization

VALUES TOTALS, POSITIVE POLARIZATION, BY AUTHOR				
	Coelho	**Braga**	**Pedroso**	**Ferreira**
	PERCENTAGES			
	49.5%	48.16%	46.16%	44.94%

The results show that the differences between positive and negative polarizations in our corpus was gradually reduced over time, reaching a significant reduction of 4.56 percentage points in Ferreira, as mentioned above. There is clearly a desire to interpret these results (especially regarding Ferreira) as an indication of societal transformation, through an inversion of values. However, this may risk jumping to a hasty conclusion. Nevertheless, a difference in the supremacy of values with positive polarization over values with negative polarization, over time, is very interesting.

Lexicon and values: Strands of meaning or typological categories—social/personal, spiritual/religious, and aesthetic

Table 11 shows a distribution of results by typological categories.

[11] Calculated to determine the relationship between positive and negative polarization values for each author.

Table 11. Values totals by category: occurrences, lemmas, and different forms

VALUES TOTALS IN THE CORPUS ACCORDING TO CATEGORY			
CATEGORIES	Occurrences	Lemmas	Different forms
Personal and social	51,118	1,965	6,076
Spiritual and religious	1,207	102	159
Aesthetic	2,111	121	253
CORPUS TOTAL	54,436	2,188	6,488

The personal and social category has a much higher frequency than the others, totaling 51,118 occurrences, 1,965 lemmas, and 6,076 different forms. Next in frequency is the aesthetic category (2,111 total occurrences, 121 lemmas, 253 different forms), followed by the spiritual and religious category (1,207 total occurrences, 102 lemmas, 159 different forms). These are expressed in percentage form in Table 12.

Table 12. Values totals by category: number of occurrences and corresponding percentages

VALUES TOTALS IN THE CORPUS ACCORDING TO CATEGORY		
CATEGORY	Occurrences	Percentage
Personal and social	51,118	93.9%
Spiritual and religious	1,207	2.22%
Aesthetic	2,111	3.88/
CORPUS TOTAL	54,436	100%

Expressed in percentage terms, the occurrences for the personal and social category (93.9%) far surpass those for the aesthetic category (3.88%) and the spiritual and religious category (2.22%). While the difference from personal/social to the other categories is vast, the gap between the spiritual/religious and aesthetic categories is small—a difference of 1.66 percentage points.

Comparatively, the categorical distribution by author is shown in Table 13.

Table 13. Values totals by author and category

VALUES TOTALS, AUTHOR AND CATEGORY			
AUTHOR	CATEGORY		
	Personal and social	Spiritual and religious	Aesthetic
Coelho	10,709	279	324
Braga	16,863	480	618
Pedroso	17,216	351	830
Ferreira	6,330	97	339
TOTAL CORPUS	51,118	1,207	2,111

The categorical distribution by author corresponds to the results for the total corpus. The personal and social category has the highest number of occurrences for all authors, followed by the aesthetic category and finally the spiritual and religious category. Percentages are displayed in Table 14.

Table 14. Values totals by author and category, as percentages

VALUES TOTALS BY AUTHOR ACCORDING TO CATEGORY								
	Coelho	**%**	**Braga**	**%**	**Pedroso**	**%**	**Ferreira**	**%**
Personal and social	10,709	94.67	16,863	93.89	17,216	93.58	6,330	93.56
Spiritual and religious	279	2.47	480	2.67	351	1.91	97	1.43
Aesthetic	324	2.86	618	3.44	830	4.51	339	5.01
Total	11,312	100	17,961	100	18,397	100	6,766	100

Also among the authors, expressed in percentage terms, we notice the clear superiority of the personal and social category over the other categories. The values in this field range from 94.67% (Coelho) to 93.56% (Ferreira). The percentages for the spiritual and religious category vary from 2.67% (Braga) to 1.43% (Ferreira). And the aesthetic category ranges from 5.01% (Ferreira) to 2.86% (Coelho). Table 15 makes these differences more obvious.

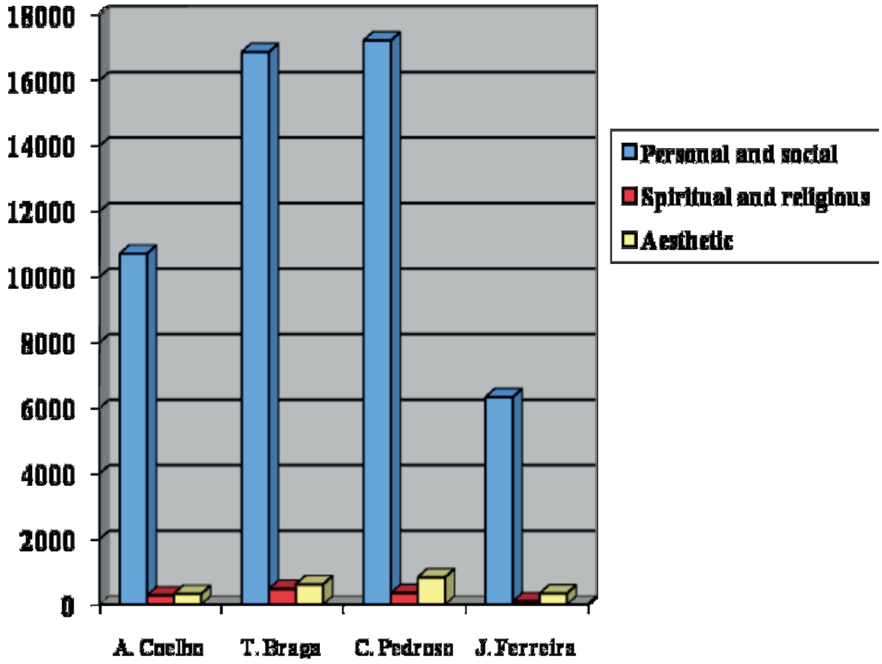
Table 15. Values totals by author and category, as percentage point differences

VALUES TOTALS BY AUTHOR ACCORDING TO CATEGORY				
	Coelho	**Braga**	**Pedroso**	**Ferreira**
CATEGORIES	**PERCENTAGE POINT DIFFERENCE**			
Personal and social vs. spiritual and religious	92.2	91.22	91.67	92.13
Personal and social vs. aesthetic	91.81	90.45	89.07	88.55
Aesthetic vs. spiritual and religious	0.39	0.77	2.61	3.58

If, regarding the difference in terms of polarization in the authors under study, we have already had the opportunity to note the change that has taken place over time, we now note a temporal change in relation to the percentage difference among the categories. Let us focus on the percentage values regarding the difference between the personal and social category and the aesthetic one and also between the aesthetic category and the spiritual and religious category. Over time, the differential between the aesthetic category and the personal and social category is shortened by 3.26 percentage points, and the differential between the aesthetic category and the spiritual and religious category is shortened by 3.19 percentage points. Thus the aesthetic category has increased in strength by 6.45 percentage points.

Interpreting the results regarding polarizations and categories, we can say that for our corpus the values with negative polarization have increased in percentage over time and the aesthetic category has become more prevalent compared to the other categories. Graph 4 illustrates the distribution of occurrences by typological category and author.

Graph 4. Values totals by author and typological category



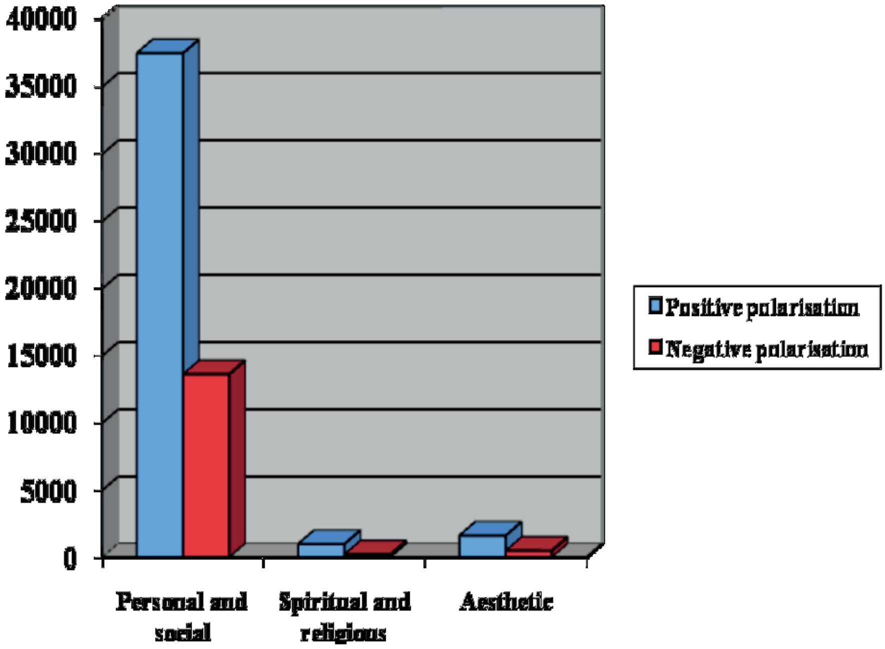
In distribution by category, it is important to know the polarization of the reported values. The results are shown in Table 16.

Table 16. Values totals by category and polarization: occurrences, lemmas, and different forms

VALUES TOTALS IN THE CORPUS BY CATEGORY AND POLARIZATION						
CATEGORIES	Positive Polarization			Negative Polarization		
	Occurrences	Lemmas	Different forms	Occurrences	Lemmas	Different Forms
Personal and social	37,498	1,037	3,727	13,620	928	2,349
Spiritual and religious	989	91	144	218	11	15
Aesthetic	1,619	84	180	492	37	73
CORPUS TOTALS	40,106	1,212	4,051	14,330	976	2,437

Graph 5 depicts the total number of occurrences by category and polarization.

Graph 5. Values totals: number of occurrences by category and polarization



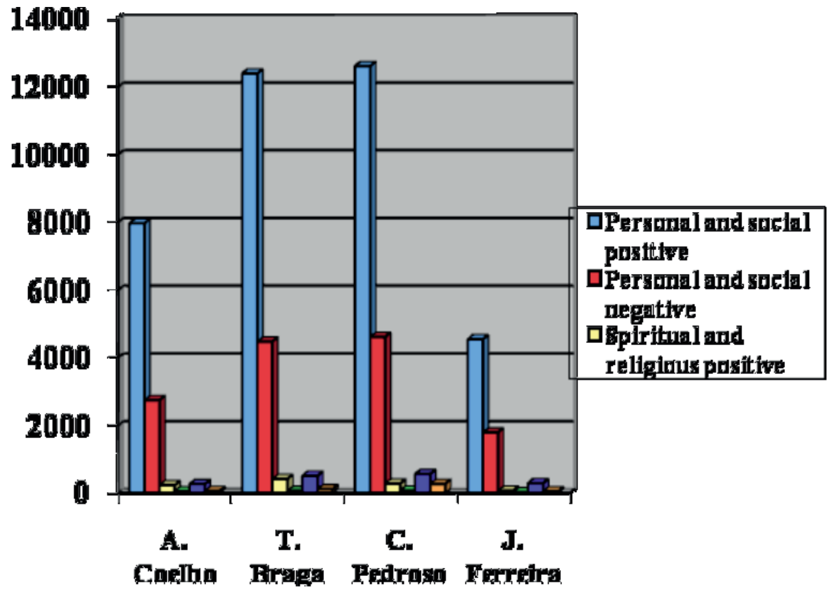
If we compare the results obtained in each author, the data are shown in Table 17.

Table 17. Values totals: number of occurrences by category, polarization, and author

VALUES TOTALS: NUMBER OF OCCURRENCES BY CATEGORY, POLARIZATION, AND AUTHOR						
AUTHOR	Personal and social		Spiritual and religious		Aesthetic	
	Positive polarization	Negative polarization	Positive polarization	Negative polarization	Positive polarization	Negative polarization
Coelho	7,964	2,745	230	49	262	62
Braga	12,387	4,476	413	67	505	113
Pedroso	12,605	4,611	277	74	560	270
Ferreira	4,542	1,788	69	28	292	47
Total	37,498	13,620	989	218	1,619	492

The distribution by category, polarization, and author is illustrated in Graph 6.

Graph 6. Values totals: number of occurrences by category, polarization, and author



The procedures regarding polarization and categories followed the same steps, but with referential values for each of them, as expected. Results can be seen in Table 18.

Table 18. Lexical richness in the corpus, for values

LEXICAL RICHNESS IN THE CORPUS, FOR VALUES					
CORPUS		LEMMAS	DIFFERENT FORMS	Repetition values	
				Lemmas	Different forms
Values totals		L = 2188 P = 54436 f = = 4,019%	F = 6488 P = 54436 f = = 11,919 %	95.981%	88.081%
Polarization	Positive	L = 1212 P = 40106 f = = 3,023%	F = 4051 P = 40106 f = = 10,101%	96.977%	89.899%
	Negative	L = 976 P = 14330 f = = 6,811%	F = 2437 P = 14330 f = = 17,006%	93.189%	82.994%

Categories	Personal and social	L = 1965 P = 51118 f = = 3,844%	F = 6076 P = 51118 f = = 11,886%	96.156%	88.114%
	Spiritual and religious	L = 102 P = 1207 f = = 8,451%	F = 159 P = 1207 f = = 13,173%	91.549%	86.827%
	Aesthetic	L = 121 P = 2111 f = = 5,734%	F = 253 P = 2111 f = = 11,985%	94.266%	88.015%

The percentage values above clearly illustrate the reduced frequency of the different forms and lemmas when compared to referential totals. As to the total lexicon of values, the percentage of repetitions concerning lemmas is 95.981%, and 88.081% for different forms. In regard to lexical richness, measured according to polarization and typological categories, knowing that we have to take into account the fact that the referential totals are different among themselves, we notice that, as far as polarization is concerned, the number of repetitions is greater in the positive polarization, both in lemmas and in different forms, which results in greater lexical richness in the negative polarization. Considering typological categories, the number of repetitions is greater in the personal and social categories, both in lemmas and in different forms, and lower in the spiritual and religious categories, in lemmas and also in different forms, and therefore there is greater lexical richness in the aforementioned category.

Conclusion

The study of literary texts using computer tools dates back to the beginning of computer science itself, which took place in the 1940s and evolved over the following decades, reaching an important development from the 1990s onwards. The use of technologies, particularly of software for the automatic analysis of texts, is a valuable resource to be used in the study of corpora, whose use makes it possible to work in areas that would otherwise be very difficult to access. The use of new technologies enables more reliable and systematic results, in a very short period of time, when compared to manually collected and processed corpora.

We share Berber Sardinha opinion (2004), who, right in the preface to his work *Lingüistica de Corpus*, is quite clear to assume corpus linguistics as an area of fundamental importance that has opened the door to a new way of seeing linguistics, also opening up the way to further research in different areas. Hunston (2002), among others, also highlights the advantages of this new way of working in linguistics and adds new data, establishing that the difference between linguistics and applied linguistics is not simply that one deals with theory and the other with the application of those theories. Rather, applied linguistics has tended to develop language theories of its own, theories that are more relevant to the questions applied linguistics seeks to answer than those developed by theoretical linguistics. Increasingly, corpora have added to the development of those applied views of language.

In order to study the lexical richness, and within it the values, of a corpus, we need to calculate the vocabulary/occurrence ratio, so that we get acquainted with the variety of forms used and, therefore, the richness expressed in percentages. In Berber Sardinha (2004, p. 94, our translation) words, "in practice, the form/item ratio indicates the lexical richness of the text. The higher its value, the more different words the text will contain. In contrast, a low value will indicate a high number of repetitions, which may indicate a less rich or varied text from the point of view of its vocabulary"[12].

We examined the lexical richness of the corpus, as far as the lexicon and values are concerned, in terms of three benchmarks: the values totals, those concerning polarization, and those referring to categories, making comparisons among them. We observed the lexical richness of the corpus, in relation to lexicon and values, through the mean frequency (f) of the different forms and the lexical lemmas pertaining to values, in relation to the total number of occurrences of values (form/occurrence ratio and lemma/occurrence ratio). To obtain this, we divided the total of different forms, represented by F, by the total number of occurrences, represented by P, and we divided the total of lemmas, represented by L, by the total of occurrences P (dividing the partial totals by the overall totals, in percentage terms).

As this is an essentially statistical study, also in terms of the computational linguistics and corpus tools used, elaborated, made known and demonstrated, its possibilities have not been exhausted; quite the contrary, as the possibilities for research continue.

References

BERBER SARDINHA, T. 2004. *Lingüística de corpus*. São Paulo, Manole, 218 p.

BOWKER, L.; PEARSON, J., 2002. *Working with specialized language*: a practical guide to using corpora. London, Routledge, xiii+242 p. https://doi.org/10.4324/9780203469255

BRAGA, T. 2002. *Contos tradicionais do povo Português*: v. 1 e 2. 6ª ed., Lisboa, Publicações Dom Quixote, 480 p.

CHOMSKY, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* IT-2, **2**(3):113-124. https://doi.org/10.1109/TIT.1956.1056813

COELHO, A. 2005. *Contos populares portugueses*. 8ª ed., Lisboa, Publicações Dom Quixote, 290 p.

FADANELLI, S.B.; MONZÓN, A.J. 2017. Gêneros textuais datasheet e artigo científico em aulas de

[12] "[a] prática, a razão forma/item indica a riqueza lexical do texto. Quanto maior o seu valor, mais palavras diferentes o texto conterá. Em contraposição, um valor baixo indicará um número alto de repetições, o que pode indicar um texto menos rico, ou variado do ponto de vista de seu vocabulário".

ESP: levantamentos léxico-estatísticos para fins educacionais. *Domínios de Lingu@gem*, **11**(2):351-378. https://doi.org/10.14393/DL29-v11n2a2017-5

FERREIRA, J.A. 1999. *Literatura popular de trás-os-montes e alto douro*: lendas e contos infantis. Vila Real, CMVR, 1219 p.

GALLISON, R.; COSTE, D. 1983. *Dicionário de didáctica das línguas*. Coimbra, Almedina, 763 p.

GRIES, S. 2010. Corpus linguistics and theoretical linguistics: a love–hate relationship? Not necessarily. *International Journal of Corpus Linguistics* **15**(3):327-343. https://doi.org/10.1075/ijcl.15.3.02gri

HALLIDAY, M.A.K. 1967. The linguistic study of literary texts. *In*: S. CHATMAN; S.R. LEVIN (eds.), *Essays on the language of literature*. Boston, Houghton-Mifflin, p. 217-223.

HALLIDAY, M.A.K. 2006. *Computational and quantitative studies*. London, British Library, 312 p.

HESSEN, J. 1980. *Filosofia dos valores*. Coimbra, Armênio Amado, 256 p.

HUNSTON, S. 2002. *Corpora in applied linguistics*. Cambridge, Cambridge University Press, 241 p. https://doi.org/10.1017/CBO9781139524773

KENNEDY, G. 1998. *An introduction to corpus linguistics*. London, Longman, 328 p.

LEECH, G. 1992. Corpora and theories of linguistic performance. *In*: J. SVARTVIK (ed.), *Directions in corpus linguistics*: proceedings of nobel symposium. Berlin, Mouton de Gruyter, p. 105-122.

LYONS. J. 1968. *Introduction to theorical linguistics*. Cambridge, Cambridge Press, 519 p.

MCENERY, T.; WILSON, A. 1996. *Corpus linguistics*. Edinburgh, Edinburgh University Press, 209 p.

MCENERY, T.; XIAO, R.; TONO, Y., 2006. *Corpus-based language studies*: an advanced resource book. London, Routledge, 386 p.

MEYER, C.F. 2002. *English corpus linguistics*: an introduction. Cambridge, Cambridge University Press, xvi+ 168 p. https://doi.org/10.1017/CBO9780511606311

PEDROSO, C. 2000 *Contos populares Portugueses.* 7ª ed., Lisboa, Vega, 583 p.

SILBERZTEIN M. 2004. NooJ: a cooperative object oriented architecture for NLP. *In: INTEX pour la linguistique et le traitement automatique des langues.* Cahiers de la MSH Ledoux. Besançon, Presses Universitaires de Franche-Comté, p. 351-362.

SILBERZTEIN M. 2015. *La formalisation des langues*: l'approche de NooJ. Paris, ISTE, 426 p.

SINCLAIR, J. 2004. Trust the text language, corpus and discourse. London, Routledge, 224 p.

STUBBS, M. 1993. British traditions in text analysis: from firth to sinclair. *In:* M. BAKER; F. FRAN-

CIS; E. TOGNINI-BONELLI (eds.), *Text and technology*: in honour of John Sinclair. Amsterdam, John Benjamins, p. 1-46. https://doi.org/10.1075/z.64.02stu

TEUBERT, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics,* **10**(1):1-13. https://doi.org/10.1075/ijcl.10.1.01teu

TOGNINI-BONELLI, E. 2001. *Corpus linguistics at work*. Amsterdam, John Benjamins, 224 p. https://doi.org/10.1075/scl.6

Submetido: 30/10/2019
Aceito: 30/03/2020