Acta Scientiarum
http://periodicos.uem.br/ojs/acta
ISSN on-line: 1983-4683
Doi: 10.4025/actascilangcult.v41i1.42079

A lexicometric analysis of the poems from *O Guardador de Rebanhos*

Carlos Assunção[1*] and Carla Araújo[2]

[1]Universidade de Trás-os-Montes e Alto Douro, Quinta de Prados, 5001-801, Vila Real, Portugal. [2]Instituto Politécnico de Bragança, Bragança, Portugal. *Author for conrrespondence. E-mail: cassunca@utad.pt

ABSTRACT. This study is guided by a methodology that fits into a research model related to corpus linguistics and it primarily aims at developing a lexicometric analysis of the poems from *O guardador de rebanhos*, by Alberto Caeiro (Pessoa, 2007), using a computational resource of lexical analysis, the NooJ software (Silberztein, 2015). Based on the linguistic data provided by NooJ, we begin by presenting the general features that characterise the poems from *O guardador de rebanhos*, then we analyse the list of theme words, departing from the list of tokens in descending order of frequency. In the last part, we present thematic fields, drawn from the theme words of the corresponding poems. The lexicometric analysis of the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007), may be seen as an opportunity for didactic operationalisation in secondary education, revealing NooJ as a potential didactic resource and providing evidence of the indisputable contribution of corpus linguistics to teaching.

Keywords: teaching of portuguese; corpus linguistics; nooj; theme words; thematic fields

Análise lexicométrica dos poemas de *O guardador de rebanhos*

RESUMO. Este estudo orienta-se por uma metodologia que se enquadra num modelo de investigação relacionado com a linguística de corpus e tem como objetivo central efetuar uma análise lexicométrica dos poemas de *O guardador de rebanhos*, de Alberto Caeiro (Pessoa, 2007), usando um recurso computacional de análise lexical, o programa *Nooj*. Partindo dos dados linguísticos fornecidos pelo *Nooj*, começamos por apresentar os dados gerais caracterizadores dos poemas de *O guardador de rebanhos*, depois analisamos a listagem de palavras-tema, partindo da listagem dos *tokens* por ordem decrescente de frequência. Na última parte, apresentamos campos temáticos, definidos a partir das palavras-tema dos respetivos poemas. A análise lexicométrica dos poemas de *O guardador de rebanhos*, de Alberto Caeiro (Pessoa, 2007), constitui uma possibilidade de operacionalização didática no Ensino Secundário, revelando o *Nooj* como um potencial recurso didático e demonstrando o incontestável contributo da linguística de corpus para o ensino.

Palavras-chave: ensino do português; linguística de corpus; nooj; palavras-tema; campos temáticos.

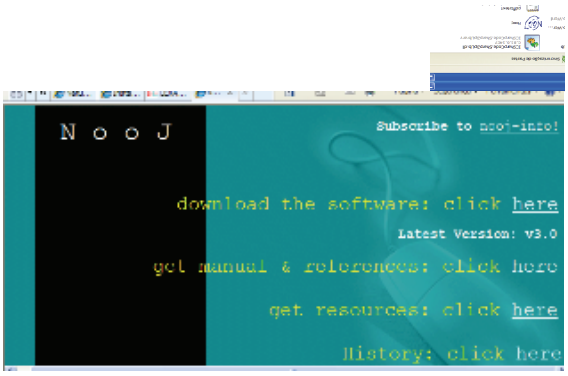
Received on March 20, 2018.
Accepted on February 13, 2019.

Introduction[1]

This article is made up of three parts and its main objective is to present a lexical study of the poems from *O guardador de rebanhos* (*The keeper of flocks*), by Alberto Caeiro. *O guardador de rebanhos* is a collection of poems written by Alberto Caeiro, a heteronym of Fernando Pessoa (2007). The poems were written in 1914 and the writer Fernando Pessoa traced their genesis to a single night when Caeiro was suffering from insomnia. Fernando Pessoa is the greatest Portuguese poet of the 20th century and his works are translated into many languages, thus being studied in many countries outside the lusophone world. This is the reason why we have decided to write this text in English, so that it can be read by a wider community of readers. The current study was undertaken using the portuguese version of *O guardador de rebanhos* and NooJ, a computer programme that performs lexical analysis. This computational resource enables the text to be tackled in portuguese, whereas the majority of the resources that are available relate to an analysis of

[1] *The keeper of flocks*, using Chris Daniels's translation (Pessoa, 2007). From here onwards, this translated title will be adopted and used in this study.

data written in English. NooJ is linguistic development environment software that, on the one hand, enables formal descriptions (dictionaries and grammar books) of a wide range of natural languages to be generated and, on the other hand, effectively applies these descriptions to long texts. The use of NooJ allows us the possibility to put into practice a diverse range of procedures (Figure 1).



NooJ is a linguistic development environment that includes large -coverage dictionaries and grammars, and parses corpora in real time. NooJ includes tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. Dictionaries and grammars are applied to texts in order to locate morphological, lexical and syntactic patterns and tag simple and compound words. NooJ can build complex concordances, with respect to all types of Finite State and Context-Free patterns. NooJ users can easily develop extractors to identify semantic units in large texts, such as names of persons, locations, dates, technical expressions of finance, and so on (Silberztein, 2015).

Figure 1. NooJ, a computational resource of lexical analysis (Silberztein, 2015).

There are another computer programs of lexicometric analysis: lexico, stablex, tropes, wordsmith, microconcord, TACT, corpus presenter, intex, lexa, concordance, protan, SALT, text analysis, and so on. AntConc, for example, is a corpus analysis toolkit designed by Laurence Anthony (Anthony, 2004). As the authors Fadanelli and Monzón 2017) noted,

> It is equipped with a Concordance program, word frequency list generators and the tool used for extraction, a keyword function (offering an on-screen list of candidates for keywords in a corpus, comparing the study corpus with another corpus, at least five times greater - called the reference corpus); the reference corpus was composed of texts extracted from the COCA (Contemporary Corpus of American English) and the BNC (British National Corpus) corpora, researching the occurrence of the 30 most frequent words in the COCA Corpus wordlist. The reference corpus has approximately 211,000 tokens, more than five times the size of the study corpus (Sardinha, 2000). The texts originate from diverse domains, academic, news, literature, Internet, and went through the same punctuation cleaning, etc. to integrate the corpus of research (Fadanelli & Monzón, 2017, p. 363, our translation)[2].

In the first part, after a definition of lexicometry, brief reference is made to its potentialities and to some of the concepts that it covers, such as keyword, theme word, word frequency counters, taggers, concordancers and concordancies. There is also a brief presentation of the NooJ programme and the definition of thematic field.

In the second part, using NooJ, we will carry out a lexicometric analysis, based on the statistical analysis of the theme words, in order to define possible thematic fields in those poems. In light of the above mentioned, we will begin by presenting the general data of the corpus and by organising a list of theme words, departing from the list of tokens in descending order of frequency. The lexicometric analysis of the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007), ends with the presentation of the thematic fields, drawn from the corresponding theme words.

In the last part of this work, we will present a pedagogical proposal that explores the potentialities of the lexicometric analysis of the poems from *The keeper of flocks* (Pessoa, 2007). The poem *O guardador de rebanhos* (Pessoa, 2007) is studied by secondary school students attending 12th grade and it is compulsory for all, regardless of the area of study. These are 17- or 18-year-old students. At university level, the poem is studied in curricular units related to literature in degrees in humanities, specifically in letters.

[2] "equipado com um concordanciador geradores de listas de frequência de palavras e a ferramenta utilizada para a extração, a função Keyword (oferece na tela a lista da candidatas a palavras-chave de um corpus,comparando o corpus de estudo com um outro corpus no mínimo cinco vezes maior - chamado de corpus de referência); o corpus de referência foi composto de textos extraídos do corpus COCA (Contemporary Corpus of American English) e do BNC (British NationalCorpus), pesquisando a ocorrência das 30 palavras mais frequentes na wordlist do Corpus COCA. O corpus de referência possui aproximadamente 211 mil tokens, mais do que cinco vezes o tamanho do corpus de estudo (Sardinha,2000). Os textos têm origens de diversos domínios, acadêmico, noticiários, literatura, internet, e passaram pela mesma limpeza de pontuação, etc. que o Corpus de pesquisa".

Lexicometry

Lexicometry consists of a set of objective, descriptive, inductive and scientific technological methods, which, due to statistical analysis programmes, enable "[...] formal reorganisations of the vocabulary (set of forms actualised in discourse, attested in a text or in a corpus of texts). The lexicometric study implies an exhaustive survey of ALL occurrences, of ALL the forms of the corpus to be studied" (Carvalho, Marques, & Silva, 1999, p. 225, our translation)[3]. This means that, besides granting us access to a detailed and rigorous linguistic register of the vocabulary of the corpus under analysis, lexicometry also gives us systematic and objective results, and this contributes to an objective presentation of the quantified linguistic data.

For example, as we notice in sections 3 and 4 of this study, through lexicometry, we can observe, in a rigorous way, the frequency with which the words from *The keeper of flocks* (Pessoa, 2007) occur within it, as well as identify the keywords, the theme words and frequency forms 1, that is, the 'hapax legomena or hapaxes', as a result of the various computational resources used in corpus linguistics, from frequency counters, programmes that provide the frequency of words, to taggers, which perform automatic analysis of the corpus and tagging, to textual engineering software as well as to concordancers.

As its name implies, the concordancers allow the extraction of concordances. According to Berber Sardinha (2004), concordance is a list of the contexts of occurrence of a given linguistic form, the search word, which appears centralised and accompanied by the linguistic forms that occur to its left and its right, that is, by its original co-text. Because the concordances grant access to the meaning and contextualised use of words, their analysis is of great pedagogical potentiality (McCullough, 2001), as we will show in section 4 of this study.

For us to be able to undertake this lexicometric analysis of the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007), we used NooJ (Silberztein, 2015), a computer programme of lexical analysis, available online at http://www.nooj-association.org/. NooJ is a linguistic engine

> [...] based on an annotation structure. An annotation is a pair (position, information) that determines that a certain position in the text has certain properties. When NooJ processes a text, it produces a set of annotations that are stored in the Text Annotation Structure (TAS) and are synchronised with it (Mota & Silberztein, 2007, p. 196, our translation).

After having presented some of the concepts that the lexicometric analysis considers within the scope of this study, we shall now refer to what we understand by theme word and by thematic field.

Regarding the concept of the theme 'What is word', Genouvrier and Peytard (1974) argued that the "[...] theme word is a word characterised by a very high frequency and that, in a list sorted by decreasing frequency of the vocabulary of an author, it belongs, for example, to the first 50 positions "(Genouvrier & Peytard 1974, p. 313, our translation)[4]. As the above-mentioned authors point out, the analysis according to theme words allows the characterisation of the "[...] author's style as deviation from a norm [...]. Both the key words and the 'legomena hapaxes', that is, the exclusive terms', are analysed as a way to characterise typical thematic and semantic areas" (Genouvrier & Peytard, 1974, p. 317-318, our translation,)[5].

Galisson and Coste (1983), relating the concepts of keyword and theme word, express their views on this issue as follows:

> The keyword is a full (non-grammatical) word, of great frequency in a work (or in the whole work) of an author; this frequency has the characteristic – when compared to the 'theme word' - of being very far from the frequency of the same word in a corpus of works of the same kind. In other words, the 'keyword' has the peculiarity of being abnormally frequent in a work or in an author (Galisson & Coste, 1983, p. 114-115, our translation,)[6].

In line with Galisson & Coste (1983), the thematic fields

[3] "reorganizações formais do vocabulário (conjunto de formas atualizadas no discurso, atestadas num texto ou num corpus de textos). O estudo lexicométrico impõe o levantamento exaustivo de TODAS as ocorrências, de TODAS as formas do corpus a estudar".

[4] "palavra-tema é uma palavra caracterizada por uma frequência muito elevada e que, numa ordenação por frequência decrescente do vocabulário de um autor, pertence, por exemplo, aos primeiros 50 lugares".

[5] "estilo do autor como desvio a partir de uma norma […]. Tanto as palavras-chave como os "hapaxes legomena", isto é os termos exclusivos", são analisados no sentido de caracterizar áreas temático semânticas típicas".

[6] "A *palavra-chave* é uma palavra plena (não gramatical), de grande frequência numa obra (ou em toda a obra) de um autor; esta frequência apresenta a característica – em relação à *palavra-tema* – de estar muito longe da frequência da mesma palavra num corpus de obras do mesmo género. Por outras palavras, a *palavra-chave* possui a particularidade de ser anormalmente frequente numa obra ou num autor".

> [...] constitute sets of terms that are functionally possible within a given thematic situation and whose internal organisation depends on a number of parameters paired with a psychosocial activity. For example, the thematic field of the 'house' would include 'building' (hall, staircase, elevator, step, etc.), 'construction' (materials, etc.), 'place of residence' (function, decoration, etc.), [...] and the organisation of these terms would depend on the activities of the individual in this thematic situation (Galisson & Coste, 1983, p. 104, our translation)[7].

Then, based on the linguistic analysis developed by NooJ, we will present the lexicometric analysis of *O guardador de rebanhos*, by Alberto Caeiro (Pessoa, 2007). In light of Silberztein's theorisation, it is important to warn against possible confusion over the usage of the terms 'Simple word', which is a type of minimal meaningful language unit, and 'Word form', which is a type of 'token': *Tokens* are the basic linguistic objects processed by NooJ. They are classified into three types: word forms are sequences of letters between two delimiters; digits; and delimiters. A 'word form' is a sequence of letters that does not necessarily correspond to a minimal language unit. A 'simple word' is, by definition, a minimal language unit, that is, the smallest non-analysable element of the vocabulary.

Lexicometric analysis of the poems from *The keeper of flocks*, by Alberto Caeiro

Before proceeding to the lexicometric analysis of Alberto Caeiro's work, let us look at a short passage from the poem *Sou um guardador de rebanhos* (Pessoa, 2013, p. 31).

Sou um guardador de rebanhos.
O rebanho é os meus pensamentos
E os meus pensamentos são todos sensações.
Penso com os olhos e com os ouvidos
E com as mãos e os pés
E com o nariz e a boca.
Pensar uma flor é vê-la e cheirá-la
E comer um fruto é saber-lhe o sentido".
[I am a keeper of sheep.
The sheep are my thoughts
And all my thoughts are sensations.
I think with my eyes and ears
And with my hands and feet
And with my nose and mouth.
To think of a flower is to see it and smell it
And to eat a fruit is to Khnow its meaning.]

Once the linguistic analysis starts, NooJ displays the general distinguishing features of the text, as shown in Table 1.

Table 1. General features of the 'corpus'.

General features of the 'corpus' – *The keeper of flocks*	
Text units	1074
No. of characters	39137 (29357 letters; 6386 blank spaces; 1394 other delimiters)
*tokens	6603
*word forms	7409
*delimiters	1394
Annotations	27473
Ambiguity	1139 different types of ambiguity
Non-ambiguous linguistic units	501

Source: Silberztein (2015).

The NooJ programme analysed the tokens and their frequencies. Tokens can be presented in descending order of their frequency and/or alphabetically.

[7] "constituem conjuntos de termos funcionalmente possíveis no interior de uma determinada situação temática e cuja organização interna depende de um certo número de parâmetros emprestados à atividade psicossocial. Ex: o campo temático da "casa" compreenderia o que diz respeito ao "edifício" (hall, escada, elevador, degrau, etc.), à "construção" (materiais, etc.), ao "lugar de habitação" (função, decoração, etc.), […] e a organização destes termos dependeria das atividades do indivíduo que se encontrasse nessa situação temática".

Through the analysis of the most frequent items, we found that, as in most corpora, the most frequent forms are functional or grammatical words, for example, in *The keeper of flocks* (Pessoa, 2007), the five most frequent tokens are the following: '*E*' / '*e*', '*que*' / '*Que*', 'a', 'o', '*de*'[8], having a frequency of 448, 373, 223, 205, 142, respectively.

In the analysis of the most frequent tokens, it is important to note that, as it happens in most lexicometric applications, NooJ distinguishes between upper and lower-case, considering, separately, each different form of the same lemma.

In this context, the token that appears in the first position of the list in descending order of frequency, '*que* [that]', has a frequency of 307, to which the frequency of the form '*Que* [That]' is added, matching 66. Similarly, the token which appears in the second position of the list in descending order of frequency, '*E* [And]', has a frequency of 237, while the form '*e* [and]' has a frequency of 211.

We selected the 60 most frequent tokens, we filtered and exported the data and copied it to microsoft word, producing a list of theme words and their frequencies, which we transcribe below. We have chosen to present the listing of tokens, having common nouns, adjectives, verbs and adverbs as our main criteria.

Table 2. List of theme words.

Frequency	Token	Frequency	Token	Frequency	Token	Frequency	Token
146	*é/É* [is/Is]	19	*ver* [see]	13	*sinto* [I feel]	11	*tinha* [had]
140	*não/Não* [no/No]	18	*sei* [I know]	13	*vezes* [times]	11	*Montes* [s/he rides / hills]
74	*eu/Eu* I [lower-case]/I [upper-case]	17	*tem* [has]	13	*ter* [to have]	11	*penso* [I think]
43	*flores* flowers	16	*está* [is]	13	*homens* [men]	10	*casa* [house]
40	*coisas* [things]	16	*olhos* [eyes]	13	*nunca/Nunca* [never/Never]	10	*flor* [flower]
31	*sol* [sun]	16	*luar* [moonlight]	12	*pedras* [stones]	10	*terra* [land]
30	*são* [are]	16	*sou* [am]	12	*alma* [soul]	10	*olhar* [look]
27	*árvores* [trees]	15	*céu* [sky]	12	*tenho* [I have]	10	*vento* [wind]
24	*há/Há* [there is/There is]	15	*Têm* [they have]	12	*gente* [people]	10	*aldeia* [village]
24	*Natureza* [Nature]	15	*toda* [all]	12	*passa* [s/he *passes*/goes by]	10	*cor* [colour]
24	*Pensar* [to think]	15	*céu* [sky]	12	*dia* [day]	9	*sentir* [to feel]
24	*ser* [to be]	14	*rio* [river]	12	*noite* [night]	9	*água* [water]
22	*coisa* [thing]	14	*rios* [rivers]	12	*era* [was]	9	*estrada* [road]
20	*Deus* [God]	13	*todos* [all]	11	*natural* [natural]	9	*vida* [life]
20	*nada* [nothing]	13	*Saber* [to know]	11	*sabe* [knows]	9	*janela* [window]

Source: The keeper of flocks (Pessoa, 2007).

In the first stage of analysis of theme words, we come across phenomena of potential ambiguity, inherent to the majority of linguistic forms, because, in the tokens list provided by NooJ, the theme words are not part of the contexts of occurrence and, thus, it was necessary to carry out the extraction of concordances in NooJ. As a result, we analysed all the contexts of occurrence of each of the theme words of the corpus and we looked up its definitions in the dictionary. The dictionary that was used is the *Dicionário Infopédia da língua portuguesa da Porto editora*® (Online version: http://www.infopedia.pt/) not only due to its availability, but also because it is the most commonly used dictionary by both teachers and students in elementary and secondary schools in Portugal.

In fact, the examination of concordances and of the dictionary allows for a distancing from the intuition inherent to the observer's point of view. As an example, we can see the theme word '*montes*', which

[8] And / and, that / That, the [feminine and masculine definite articles], of.

constitutes a phenomenon of partial homonymy, that is, '*montes*' as both 'rides' [verb to ride] and 'hills' [noun], as we notice when browsing through the dictionary[9], referring to the form of the verb 'to ride' and to the masculine noun 'hill'.

As we can see in Figure 1 from the analysis of concordances of the theme word '*montes*', we may conclude that the form '*montes*' matches the masculine plural noun 'hill' and the first of the nine meanings of this noun that are in the above-mentioned dictionary ('1. elevation of land above the earth's surface, less extensive and not as high as a mountain; 2. rhyme or set of any stacked objects, a collection; 3. 'figurative' considerable portion; 4. group; gathering; 5. games of chance using cards; 6. portion of goods as part of an inheritance, 7. playing cards; 8. 'regionalism' (from the Alentejo) estate quarters made up of several buildings around a courtyard; designation sometimes attributed to the actual farmstead; 9. 'regionalism' (from Trás-os-Montes) tract of land covered with spontaneous shrubby (usually, woody) and herbaceous vegetation'). Thus, as we notice in Table 3, the theme word '*montes'* is part of the thematic field of 'universe'.

Table 3. Thematic fields in *The keeper of flocks*.

Theme words	Thematic fields
olhos; *cor*; *olhar*; *ver*[6].	Thematic field of sight (*O ver*) [The look]
Deus; *alma*[7].	Thematic field of transcendence (*A metafísica / anti-metafísica*) [metaphysics/anti-metaphysics]
flores; *sol*; *árvores*; *Natureza*; *ser*; *luar*; céu; *rio*; *rios*; *homens*; *água*; *dia*; *noite*; *pedras*; *montes*; *natural*; *flor*; *terra*; *vento*; *coisas*; *coisa*; *gente*; *aldeia*; *vezes*; *passa*; *casa*; *cor*; *vida*; *janela*; *estrada*[8].	Thematic field of universe (*O real objetivo*) [The real objective]
sinto; *sei*; *saber*; *sabe*; *sentir*[9].	Thematic field of the senses (*O sensacionismo*) [Sensationism]
pensar; *penso*[10].	Thematic field of reflection (*O pensamento*) [The thought]
não / Não; *nunca / Nunca*; *nada*[11].	Thematic field of refusal (*A rejeição do pensamento/misticismo*) [Rejection of thought/mysticism]

Source: The keeper of flocks (Pessoa, 2007).

In addition, in NooJ, we can also view the ambiguous words (Figure 3) and the unambiguous ones (Figure 4), as we are granted access to all the annotations produced for the corpus under study.

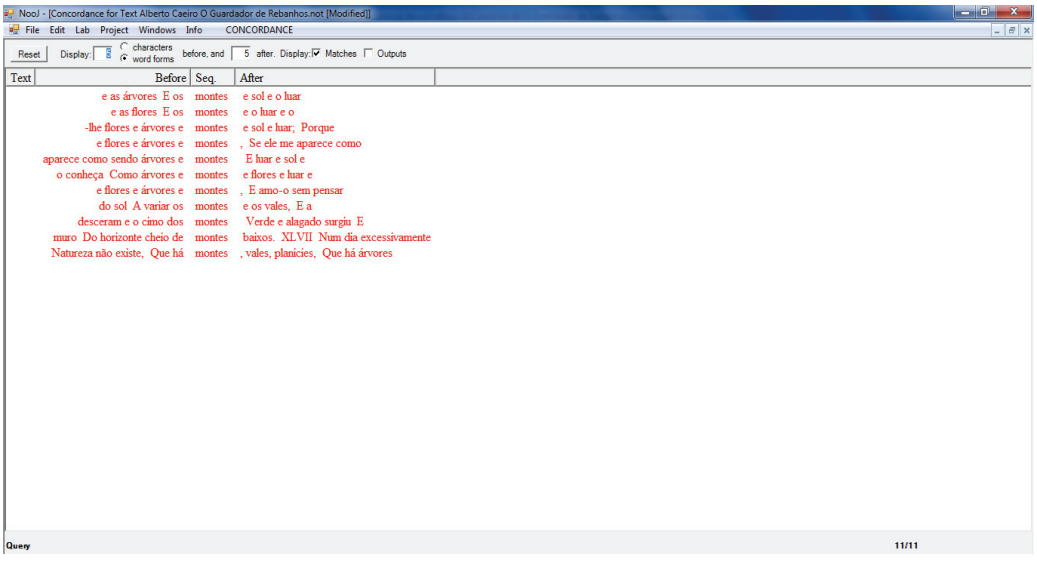


Figure 2. Concordances of the theme word '*montes*' (Pessoa, 2007).

[9] *monte* In *Dicionário Infopédia da língua portuguesa com acordo ortográfico* [online].

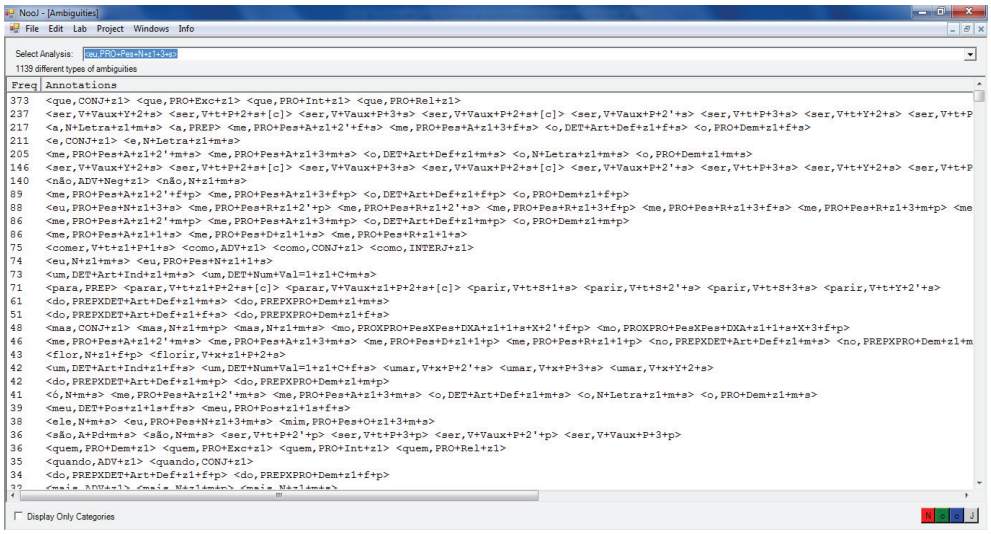


Figure 3. Ambiguous words (Pessoa, 2007).

Annotations can also be viewed, verse by verse. For example, in Figure 5, we can examine the annotations of the first verse of the poems from *The keeper of flocks*.

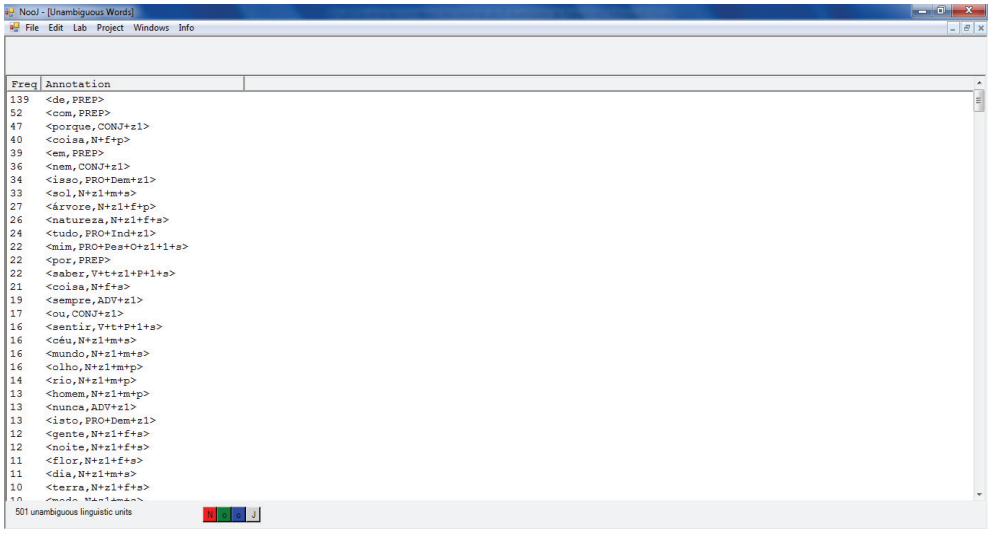


Figure 4. Unambiguous words (Pessoa, 2007).

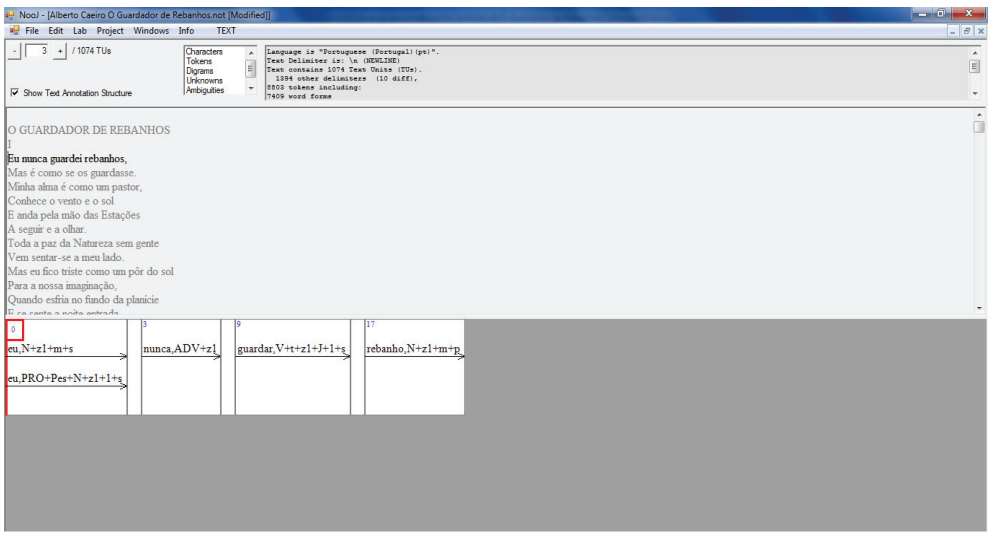


Figure 5. Annotations of the first verse of the poems from *O guardador de rebanhos* (Pessoa, 2007).

Based on this methodology, as we can see in Table 3, in the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007), we could define six thematic fields within different domains: the thematic field within the domain of sight was delimited from the theme words *olhos*, *cor*, *olhar* and *ver*[10]; the thematic field of transcendence was delimited from the theme words *Deus* and *alma*[11]; the thematic field within the domain of universe, from the theme words *flores*; *sol*; *árvores*; *Natureza*; *ser*; *luar*; *céu*; *rio*; *rios*; *homens*; *água*; *dia*; *noite*; *pedras*; *montes*; *natural*; *flor*; *terra*; *vento*; *coisas*; *coisa*; *gente*; *aldeia*; *vezes*; *passa*; *casa*; *cor*; *vida*; *janela* and *estrada*[12]; the thematic field of the senses, related to the theme words *sinto*; *sei*; *saber*; *sabe*; *sentir*[13]; the thematic field of reflection was built upon the theme words *pensar*; *penso*[14]; the thematic field within the domain of refusal was delimited from the theme words *não / Não*; *nunca / Nunca*; *nada*[15].

Pedagogical potentialities introduced by the lexicometric analysis of the poems from *The keeper of flocks*

The *Programa e metas curriculares de português do ensino secundário*[16] (Buescu, Maia, Silva, & Rocha, 2014), arguing for a "[...] global language pedagogy [...]" (Buescu et al., 2014, p. 5, our translation)[17], which implies interaction among the five domains (speaking, reading, writing, literary education and grammar), in the context of the 12th grade, in the domain of literary education, recommends the study of poems by Alberto Caeiro (Buescu et al., 2014), in particular, and of the poetry attributed to the heteronyms of Fernando Pessoa, in general (Buescu et al., 2014). In this field of action, we believe that the study of Alberto Caeiro's poems may benefit from the contribution of lexicometrics, because, just as Biber (2011) claims, we also believe that the adoption of methods from Corpus linguistics in the study of literary texts sets the scene towards a new direction for classes of portuguese, due to the fact that, as mentioned in Section 2 of this work, the data resulting from the lexicometric analysis enables rigorous and comprehensive knowledge of the author's vocabulary to arise. In light of the above, we hereby present some pedagogical approaches that promote vocabulary learning of the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007).

Because the theme words that are part of Table 2, in Section 3, challenge students with problems of lexical ambiguity, the teacher should prepare activities that guide students in the adoption of strategies to overcome those difficulties.

Students, acting as discoverers of meaning, may observe and analyse the concordances related to the theme words of the poems from *The keeper of flocks*, provided by NooJ, as well as look up their definitions in the dictionary[18], as exemplified in Table 4, which focuses on the theme word 'trees'.

Similarly, from other theme words, the teacher may suggest tasks such as the one shown in the previous table, leading students to gain access to the real meaning of the theme words from *The keeper of flocks* (Pessoa, 2007) and enabling thematic fields to be identified through them.

Based on the analysis of the theme words shown in Table 2, Section 3, students may build thematic fields in poems from *The keeper of flocks* (Pessoa, 2007). Consequently, the teacher may suggest that the students carry out tasks such as the ones presented in Table 5.

To sum up, the analysis of cases of lexical ambiguity, including both homonymy and polysemy, allows the students to discover new meanings. This finding, which results from observation, selection, verification and inference, grants the students the opportunity to develop a better representation of the meanings of words, greatly influencing the expansion of their lexical competence and, thus, it must be permanently encouraged in classes of portuguese.

Therefore, in the context of teaching and learning, the lexicometric analysis of the poems from *The keeper of flocks* (Pessoa, 2007) offers a fruitful possibility of didactic intervention, because the full access to the vocabulary of the poems from *The keeper of flocks* (Pessoa, 2007) enables the development of the students' linguistic competence.

[10] eyes, colour, look and see.

[11] God and soul.

[12] Flowers; sun; trees; Nature; to be; moonlight; sky; river; rivers; men; water; day; night; stones; mountains; natural; flower; land; wind; things; thing; people; village; times; passes/goes by; house; colour; life; window and road.

[13] [I] feel; [I] know; to know; [s/he] knows; to feel.

[14] to think, [I] think.

[15] no / No; never / Never; nothing.

[16] The national curriculum secondary programme of study for the Portuguese language.

[17] "preconizando […] uma pedagogia global da língua […]"

[18] *Dicionário Infopédia da língua portuguesa da Porto editora*® (Online version).

Finally, we should note that it is up to the teacher of portuguese to adapt the lexicometric analysis of the poems from *The keeper of flocks* (Pessoa, 2007) to each particular pedagogic and didactic context, bearing in mind the students' individual needs.

Table 4. Example of activity – definition /concordance of the theme word.

Theme word *árvores* [trees]	
Entry[19]	Concordance
1. BOTÂNICA planta lenhosa que pode atingir grandes alturas e cujo tronco se ramifica na parte superior [BOTANICS Woody plant that can reach great heights and whose trunk branches in the upper part]	*nas ruas, E a maneira como dava pelas coisas, É o de quem olha para **árvores**, E de quem desce os olhos pela estrada por onde vai andando E anda a* [the streets, And the way he had of taking things in, Was like someone looking at **trees**, Or lowering their eyes to the road where they go walking Or taking in][20]
2. Representação de alguma coisa em forma de um esquema com tronco e ramificações [Representation of something in the form of a schema with trunk and branches]	
3. MECÂNICA peça principal rotativa de uma máquina [MECHANICS main rotating part of a machine]	
4. MECÂNICA eixo, veio, fuso [MECHANICS shaft, axle, spindle]	
5. NÁUTICA mastro completo do navio [SAILING the ship's full-mast]	
6. 'figurado' pessoa muito alta ['figurative meaning' a very tall person]	
6. LINGUÍSTICA representação gráfica da estrutura de uma frase ou oração, salientando as relações de hierarquia e derivação por meio de linhas descendentes [LINGUISTICS graphical representation of the structure of a sentence or clause, emphasising relationships of hierarchy and derivation by means of descending lines]	
árvore de Natal [Christmas tree]	
árvore genealógica [Family tree]	

Source: Dicionário Infopédia da língua portuguesa da Porto editora® (2017)

Table 5. Example of activity – thematic fields in poems from *The keeper of flocks*.

Theme words	Thematic fields
Example: *olhos*; *cor*; *olhar*; *ver* [eyes; colour; look; see]	Thematic field of *Visão* [Sight]
	1. Thematic field of _____ 2. Thematic field of _____ 3. Thematic field of _____ 4. Thematic field of _____ 5. Thematic field of _____ 6. Thematic field of _____

Source: The keeper of flocks (Pessoa, 2007).

Final considerations

The contemporary framework in the process of teaching and learning of Portuguese, in Portugal, requires the development and the availability of didactic resources that may serve as the basis for the demanding and thorough pedagogical practices of the 21st-century educational conjuncture. Therefore, 'NooJ' can be seen as a proposal for a didactic approach, within the scope of the new methodologies inherent to the teaching of languages. Undoubtedly, the didactic potentialities of 'NooJ' are almost limitless. Its level of proficiency depends on the teacher's creativity and on the student's curiosity. By extracting sequences or linguistic contexts from works recommended by Programmes of portuguese for various levels, the teacher is able to

[19] *árvore [tree]* in *Dicionário Infopédia da língua portuguesa da Porto editora®* [online].
[20] Chris Daniels's translation (Pessoa, 2007, p. 18).

use a didactic approach to certain linguistic concepts (word classes, morphology, among others), in an objective, appealing and motivating way. Simultaneously, this resource also allows the teacher to acquire the necessary skills to foster within students a critical, self-reflective attitude towards the language, aiming to develop observation competencies and an analysis of the language in a process of discovery of its functioning system.

Thus, 'NooJ', in an educational setting, sets up a new teaching and learning architecture that the Portuguese language classes should incorporate and operationalise, since it is considered a relevant didactic resource. Actually, 'NooJ' can help teachers to work out another way of handling this precious and powerful tool, the Language, and, in the analysis and for the analysis of Language itself, make learning happen.

In light of the above considerations, a lexicometric analysis may have a prominent role in the process of didactic operationalisation in secondary education, revealing 'NooJ' as a potential didactic resource and setting forth the undeniable contribution of Corpus linguistics to teaching, while emphasising a more objective and scientific analysis of linguistic data, using computational tools.

From what we have discussed, we may conclude that the lexicometric analysis of the poems from *The keeper of flocks*, by Alberto Caeiro (Pessoa, 2007), paves the way for students to actively participate in linguistic discoveries, turning them into '[...] researchers and not merely 'receivers' of language" (Berber Sardinha, 2011, p 305, our translation)[21].

In fact, this didactic approach shows that the lexicometric analysis provides the teacher and the students with the stimulating opportunity to gain access to an objective and detailed inventory of all the linguistic forms of the poems from *The keeper of flocks*, enabling a clear definition of thematic fields in the literary work of the Pessoan heteronym and reaching a more objective and neutral methodology than the traditional analysis that adopts a hermeneutical stance.

We should note that, although the lexicometric analysis of the poems from *The keeper of flocks* (Pessoa 2007) distances itself from traditional analysis, as an opportunity for didactic action, within the multiplicity of possible didactic activities available to the teacher, depending on his/her creativity, this does not mean its exclusion, because the subjective points of view, resulting from traditional analysis, may be confirmed (or not) by lexicometric analysis.

Finally, we should point out that the diversity of teaching methodologies is of utmost importance in the teaching of portuguese, because, due to the heterogeneity of students, it must also be diversified and renewed, challenging the teacher of portuguese with the stimulating opportunity to adopt new alternatives for didactic action, enhanced by technological developments.

References

Anthony, L. (2004). *Design and development of a freeware corpus analysis toolkit for the technical writing classroom*. Retrieved from https://www.researchgate.net/publication/4167906_AntConc_Design_and_development_of_a_freeware_corpus_analysis_toolkit_for_the_technical_writing_classroom

Berber Sardinha, T. (2004). *Linguística de corpus*. São Paulo, SP: Manole.

Berber Sardinha, T. (2011). Como usar a linguística de corpus no ensino de língua estrangeira. In V. Viana, & S. E. O. Tagnin (Orgs.), *Corpora no ensino de línguas estrangeiras* (p. 301-356). São Paulo, SP: HUB Editorial.

Biber, D. (2011). Corpus linguistics and the study of literature: back to the future?. *Scientific Study of Literature, 1*(10), 15-23.

Buescu, H., C., Maia, L. C., Silva, M. G., & Rocha, M. R. (2014). *Programa e metas curriculares de português ensino secundário.* Lisboa, PT: Ministério da Educação e Ciência.

Carvalho, D., Marques, M. E. R., & Silva, M. F. (1999). Discurso: práticas lexicométricas. In M. A. Mota, & P. Marrafa (Orgs.), Linguística computacional: investigação fundamental e aplicações (p. 255-262). Lisboa, PT: Edições Colibri / Associação Portuguesa de Linguística.

Dicionário Infopédia da língua portuguesa da Porto editora®. (2017). Porto, PT: Porto Editora. Retrieved on July 10, 2017 from http://www.infopedia.pt/

Fadanelli, S. B., & Monzón, A. J. (2017). Gêneros textuais datasheet e artigo científico em aulas de ESP: levantamentos léxico-estatísticos para fins educacionais. *Domínios de Lingu@gem, 11*(2), 351-378.

[21] "[…]pesquisadores e não meros "receptores" da língua"

Galisson, R., & Coste, D. (1983). *Dicionário de didáctica das línguas*. Coimbra, PT: Livraria Almedina.

Genouvrier, E. & Peytard, J. (1974). *Linguística e ensino do português*. Coimbra, PT: Livraria Almedina.

McCullough, J. L. (2001). Los usos de los córpora de textos en la enseñanza de lenguas. In T. M. Parera (Ed.), *Nuevas tecnologías para el autoaprendizaje y la didáctica de lenguas* (p. 125-140). Lleida, ES: Milenio.

Mota, C., & Silberztein, M. (2007). Em busca da máxima precisão sem almanaques: O Stencil/Nooj no HAREM. In D. Santos, & N. Cardoso (Eds.), *Reconhecimento de entidades mencionadas em português: documentação e actas do* HAREM, *a primeira avaliação conjunta na área* (p. 191-208). Retrieved from http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap15-SantosCardoso2007-MotaSilberztein.pdf.

Pessoa, F. (2007). The collected poems of Alberto Caeiro (C. Daniels, Trad.). Exeter, UK: Shearsman Books. Retrieved from https://irp-cdn.multiscreensite.com/12e499a6/files/uploaded/fenando-pesssoa-colllected-poems-of-alberto-caeiro-SAMPLE.pdf.

Pessoa, F. (2013). Poemas completos de Alberto Caeiro. Retrieved from https://www.luso-livros.net/wp-content/uploads/2013/06/Poemas-de-Alberto-Caeiro.pdf

Silberztein, M. (2015). A linguistic development environment. Retrieved from http://www.nooj-association.org/