DOI: 10.17851/2237-2083.23.3.695-726

Topic Modeling for Keyword Extraction: using Natural Language Processing methods for keyword extraction in Portal Min@s

A modelagem de tópicos para extração de palavras-chave: o uso de métodos de processamento natural da linguagem para extração de palavras-chave no Portal Min@s

Arnaldo Cândido Junior
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil
arnaldoc@utfpr.edu.br

Célia Magalhães
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil
celiamag@gmail.com

Helena Caseli
Universidade Federal de São Carlos, São Paulo, Brasil
helenacaseli@gmail.com

Régis Zangirolami
Littera TI – Processamento de Dados, São Carlos, São Paulo, Brasil
regisrmz@gmail.com

Abstract: This article aims to evaluate the application of two efficient automatic methods for keyword extraction used by Corpus Linguistics and Natural Language Processing communities for generating keywords from literary texts: WordSmith Tools and Latent Dirichlet Allocation (LDA). These tools have their own specificities and are based on different extraction techniques; thus an analysis focused on their performance was required. This article aims to understand how each method works and to evaluate them when applied to extract keywords from literary works. To this end, we used human

analysis, with knowledge of the field of the texts used. The LDA method was used for extracting keywords through its integration with *Portal Min@s: Corpora de Fala e Escrita*, a general corpora-processing system, designed for different research in corpus linguistics. The experiment outcomes confirm the effectiveness of WordSmith Tools and LDA in extracting keywords from literary corpus. They also show that human analysis of the lists is required at a stage prior to experiments to complement the automatically generated list, crossing WordSmith Tools and LDA results, and that the linguistic intuition of a human analyst about the lists generated separately by the two methods in this study was more favorable to the use of the WordSmith Tools keyword list.

Keywords: keyword extraction, natural language processing, corpus analysis, WordSmith Tools, Latent Dirichlet Allocation, *Portal Min@s*

Resumo: Este artigo tem o objetivo da avaliar a aplicação de dois métodos automáticos eficientes na extração de palavras-chave, usados pelas comunidades da Linguística de *Corpus* e do Processamento da Língua Natural para gerar palavras-chave de textos literários: o *WordSmith Tools* e o *Latent Dirichlet Allocation* (LDA). As duas ferramentas escolhidas para este trabalho têm suas especificidades e técnicas diferentes de extração, o que nos levou a uma análise orientada para a sua performance. Objetivamos entender, então, como cada método funciona e avaliar sua aplicação em textos literários. Para esse fim, usamos análise humana, com conhecimento do campo dos textos usados. O método LDA foi usado para extrair palavras-chave por meio de sua integração com o *Portal Min@s: Corpora de Fala e Escrita*, um sistema geral de processamento de *corpora*, concebido para diferentes pesquisas de Linguística de *Corpus*. Os resultados do experimento confirmam a eficácia do WordSmith Tools e do LDA na extração de palavras-chave de um *corpus* literário, além de apontar que é necessária a análise humana das listas em um estágio anterior aos experimentos para complementar a lista gerada automaticamente, cruzando os resultados do WordSmith Tools e do LDA. Também indicam que a intuição linguística do analista humano sobre as listas geradas separadamente pelos dois métodos usados neste estudo foi mais favorável ao uso da lista de palavras-chave do WordSmith Tools.

Palavras-chave: extração de palavras-chave; processamento natural da linguagem; análise de *corpus*; WordSmith Tools; Latent Dirichlet Allocation; *Portal Min@s*.

Recebido em: 31 de junho 2015.
Aprovado em: 23 de novembro 2015.

1 Introduction

Keywords are a quick and efficient way of indicating the main topics of a text. According to Scott (1996), they are words whose frequency in a text is exceptionally high compared to some standard. Scott (1997) also defines keyword as a word that occurs with unusual frequency, either high or low, compared to a reference corpus. Thus, the analysis of keywords tend to indicate the topic addressed in a certain text or corpus. Keywords are used in scientific papers and literary works to facilitate the cataloging and organization of texts, and thus the search for them. The concept of keyword has become a lot stronger and today it is present in everyone's life, thanks to the Internet and the need to create effective search solutions in an environment where the amount of news, works, articles, blogs, among other types of text, has grown.

Corpus Linguistics (CL) and Natural Language Processing (NLP) communities apply different systems and methods to extract keywords. There are many systems available, such as Kea,[1] Maui indexer,[2] Carrot2,[3] among others. Two systems commonly used by the CL community are WordSmith Tools[4] and Wmatrix.[5] Examples of extraction methods used by the NLP community are the TF-IDF techniques (MANNING; SCHÜTZE, 2000), Latent Semantic Analysis (LSA) (LANDAUER *et al*., 1998) and Latent Dirichlet Allocation (LDA)[6] (BLEI, 2012; DREDZE *et al*., 2008). The latter two methods are also important for Information Retrieval in the context of document extraction, since they allow this extraction based on synonyms and

[1] <http://www.nzdl.org/Kea/>. Access on: Oct. 4, 2015.
[2] <https://code.google.com/p/maui-indexer/>. Access on: Oct. 4, 2015.
[3] <http://project.carrot2.org/>. Access on: Oct. 4, 2015.
[4] <http://www.lexically.net/wordsmith/version6/>. Access on: Oct. 4, 2015.
[5] <http://ucrel.lancs.ac.uk/wmatrix/>. Access on: Oct. 4, 2015.
[6] <http://www.cs.princeton.edu/~blei/lda-c/>. Access on: Oct. 4, 2015.

topics, respectively, which are approximations to search words, helping search engines.

Although there are separate analyses of the techniques used in each method, there is no study comparing methods for keyword extraction in literary works, which is the focus of this paper. There is no evaluation of topic extraction methods for this scenario either. The application of these methods to corpora of fictional texts, for example, must take into account some issues, given the nature of prose fiction. Such corpora, if comprised of novels, in addition to being larger than corpora of academic texts, for example, do not allow prior definition of a fixed vocabulary used in the texts, which happens only with prior knowledge of each text.

This article main goal is to compare two efficient automatic methods for keyword extraction used by CL and NLP communities for generating keywords from literary texts: WordSmith Tools keyword extraction and LDA keyword extraction. A secondary objective was to define whether these methods are effective and accurate when applied to the literary genre.

As both tools have their own specificities and are based on different extraction techniques, an analysis focused on their performance was required. To this end, we have applied human analysis, with knowledge of the field of the texts used. The LDA method was used for keyword extraction from its integration with *Portal Min@s: Corpora de Fala e Escrita*, a general corpora-processing system, designed for different research in corpus linguistics. The Portal has a web interface, which facilitated the efficient analysis of the data presented in this paper.

This text is organized as follows. Section 2 presents the theoretical foundation for this paper, including the tools, methods and corpora used. Section 3 details the experiments. Section 4 presents a discussion of the outcomes. Finally, Section 5 brings the final considerations.

2 Theoretical Foundation

This section provides details on the aforementioned methods, including a presentation of *Portal Min@s*, which shows the implementation of one of the extraction methods, LDA. In Section 2.1, we describe the methods for keyword extraction commonly used in the two areas of research: CL and NLP. In Section 2.2, the corpus used for the experiments. In Section 2.3, *Portal Min@s*, illustrating the use of the LDA method.

2.1 Methods for Keyword Extraction

Keywords are defined as a set of terms in a document that provides a summary of its content to readers (LIU *et al*., 2010). Several areas, such as Information Retrieval, CL and NLP, make use of keyword extraction for various tasks such as document categorization, clustering and summarization. Several methods for keyword extraction have been proposed and different software implemented. Popular methods and software include WordSmith Tools, TF-IDF, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and WMatrix.

WordSmith Tools is a software for Windows that is easy to install and user-friendly. It has three main features: concordancing, keyword extraction, and word listing. To enable extraction, you must first generate the word lists of the texts. WordSmith Tools generates keywords when comparing the list of frequent words of two corpora, being the first corpus called “study corpus” and the second, “reference corpus”. The reference corpus is used only to allow comparison, setting an average frequency at which the terms normally occur. Any terms present in the study corpus that show a significant change from the normal frequency set based on the reference corpus, for more and for less, will be highlighted as keywords (SARDINHA, 2000, p. 8). The only requirement of the tool is that the reference corpus is greater than the study corpus. According to Sardinha (2000, p. 12), it is ideal that the reference corpus is at least five times the size of the study corpus because it was observed that smaller reference corpus generates a smaller number of keywords, while those five times greater than the study corpus tend to

maintain a reasonable number of keywords. The advantage of using WordSmith Tools is that this method is vastly used by corpus linguists, thus simplifying the process of comparing results on different papers. It is also worth noting that WordSmith Tools uses mature statistical methodology for keyword extraction.

Other popular approach for keyword extraction is TF-IDF (MANNING; SCHÜTZE, 2000) is a statistical method used to show how important a word is for a text within a corpus. It is the result of multiplying two frequencies of words: the term frequency (TF) and the inverse document frequency (IDF). In practical terms, it compares the frequency at which a word occurs in a text with the frequency at which the same word occurs in the corpus as a whole, enabling the creation of a ranking of frequency of words in a text based on their frequency in the larger whole. It is a very practical method of finding the weight and importance of the words of a text in a specific field of study. It also has the advantage of not relying on a stop-words list, since the most frequent words in the corpus texts have less weight and are last in the ranking of importance. The main advantage of TF-IDF is its simplicity and easiness to implement in software, which is ideal for tool developers and experimenters with background in programming languages.

LSA (LANDAUER, *et al*., 1998) is a method that measures the level of similarity of words and excerpts of texts after analyzing a corpus. It assumes that words with the same meaning, or a similar one, must appear in similar contexts. The method uses a mathematical technique called Singular Value Decomposition (SVD) to build a matrix of word occurrence in the paragraphs and corpus texts, enabling to determine the weight of each word in a text. After creating the occurrence matrix, the LSA must find a minimum version of this matrix to reduce the analysis complexity, minimize noise, and unify the occurrence of words of similar meaning, including the frequency of these words in texts corpus. This method is also commonly known as LSI (Latent Semantic Indexing). It is indicated as suitable for uses in which part of the keywords are already known before the extraction takes place.

LDA, inspired in LSA, is a probabilistic model for generating topics. On this model, Blei (2012) states: “To this end, machine learning researchers have developed probabilistic topic modeling, a suite of

algorithms that aim to discover and annotate large archives of documents with thematic information." (p. 77). That is, this kind of model aims to analyze a corpus and define, through probabilistic algorithms, the topic addressed in the text. According to Blei (2012, p. 78), the idea behind LDA is that a text deals with multiple subjects, and these subjects can be represented by topics. A topic is formally defined as a distribution of words of a fixed vocabulary, and each document displays its topics in different proportions.

The method has been defined by Blei, Ng and Jordan (2003) as a hierarchical Bayesian model, which seeks to highlight the content of a document by representing its topics, generated through efficient approximation techniques. The document is seen as a mixture of topics following some probability distribution. The method seeks to associate each sentence with at least one topic. The same sentence can be associated with more than one topic. In this case, an estimate is made on the relation of the sentence with each topic in question. All calculations are made based on the bag of words of the document. The LDA method is applied for retrieving information in a first phase that clusters words by topic. In a second phase, the documents are clustered according to their topics (TDK TECHNOLOGIES, 2015). LDA main advantage is its ability to cluster keywords in topics.

Finally, WMatrix is a web platform that works in any system or environment, and which uses the so-called Matrix method. This method makes statistical comparisons in corpora, generating accurate keywords (RAYSON, 2002). It was developed based on a comparison of corpus analysis tools to provide a solution that has all the required and appropriate features for editing and extracting information from a corpus. The idea is not to eliminate the need for further human analysis, but to refine the most the initial analysis and allow a human being to examine a much larger volume of texts with less effort. According to Rayson (2002, p. 153), its main applications are: in the study of vocabulary according to social contexts; contrast native and non-native English speakers; and in the semantic analysis of documents that require software engineering. Wmatrix uses 21 semantic categories[7] to classify a text, including the

[7] <http://ucrel.lancs.ac.uk/usas/>. Access on: Oct. 4, 2015.

Portuguese language and others. The categories are comparable to the topics that LSA and LDA raise automatically, but are fixed. Works like McIntyre and Walker (2010) prove the effectiveness of WMatrix in corpora analysis of poetry and dramatic texts, but as far as we know, there are no empirical studies on its application to long literary texts such as novels. Thus, one of WMatrix advantages relies on the fact it has been applied to keywords extraction in less common genres such as poetry and dramatic texts.

Considering the presented methods and tools, the scope of research work was adjusted to compare WordSmith Tools and LDA, so the analysis of other strategies has been left for future works. WordSmith Tools was selected for being widely used in the linguistic community, considered as a classic tool for extracting keywords. LDA was selected for bringing an innovative approach based on topics, which allows a comparative analysis of its strengths and limitations in relation to a classic strategy.

2.2 Corpora

The corpora described in this section were the basis for the experiments described in Section 3. Two categories of corpora were used: study (literary) and reference (to support keyword extraction).

The corpus of literary texts (study corpus) used in the experiments described here was compiled with translations of *Heart of Darkness*, a Joseph Conrad novel published in 1899, with each version in Portuguese produced by a different translator (Mark Santarrita, in 1996; Regina Regis Junqueira and Hamilton Trevisan, in 1984). The corpus was digitalized from their printed versions in books. The numbers of tokens and sentences in these three documents are shown in Table 1.

Table 1 - Tokens and sentences by translation

File	Tokens	Sentences
HOD_Junqueira	41,808	2,201
HOD_Santarrita	36,681	2,335
HOD_Trevisan	37,896	2,419

The reference corpus used was PLN-BR GOLD, which comprises 1,024 texts and 338,441 tokens and was compiled in the PLN-BR project (MUNIZ *et al*., 2007), having only news and coverages for which *Folha de S.Paulo* (a Brazilian newspaper) has republishing rights. The size of this corpus represents 1% of the largest corpus in the PLN-BR project, the PLN-BR FULL, in order to proportionally preserve the distribution of this largest corpus. It is a stratified random sample proportional to the distribution of the PLN-BR FULL corpus, which covers the years 1994-2005, related to texts published by newspaper *Folha de S.Paulo*.

2.3 *Portal Min@s*

Portal Min@s[8] intends to provide a unified and systematized computational base for processing corpus compiled and made available for linguistics research. The system was motivated by the recent expansion of research in corpus linguistics, which brought large amounts of corpus demanding robust processing. The Portal, as well as other efforts, seek to meet this demand. The differential in relation to other tools is its generalist approach, seeking to act in different contexts, based on the type of corpus (for example, annotated or not) and task (for example, translation and lexicography) or area of study in corpus linguistics.

[8] <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>. Access on: Oct. 4th, 2015.

The tool is free and publicly available to all institutions interested in its use, requiring only a web-hosting server. The advantages of offering a web system are the possibility of simultaneous remote access by multiple users, without the need of installation in each user's computer.

Several features have been implemented, given the generalist focus of the Portal, treating different types of corpus, different languages and different research in corpus linguistics. Examples of features include generation of concordances, alignments for parallel texts and extraction of keywords, focus of this paper. Unlike corpus processing portals that work with a fixed list of corpus (DAVIES, 2005, 2009), the Portal allows the creation of corpus on demand and offers a module to import previously existing corpus, provided that they meet the format requirements of the Portal.

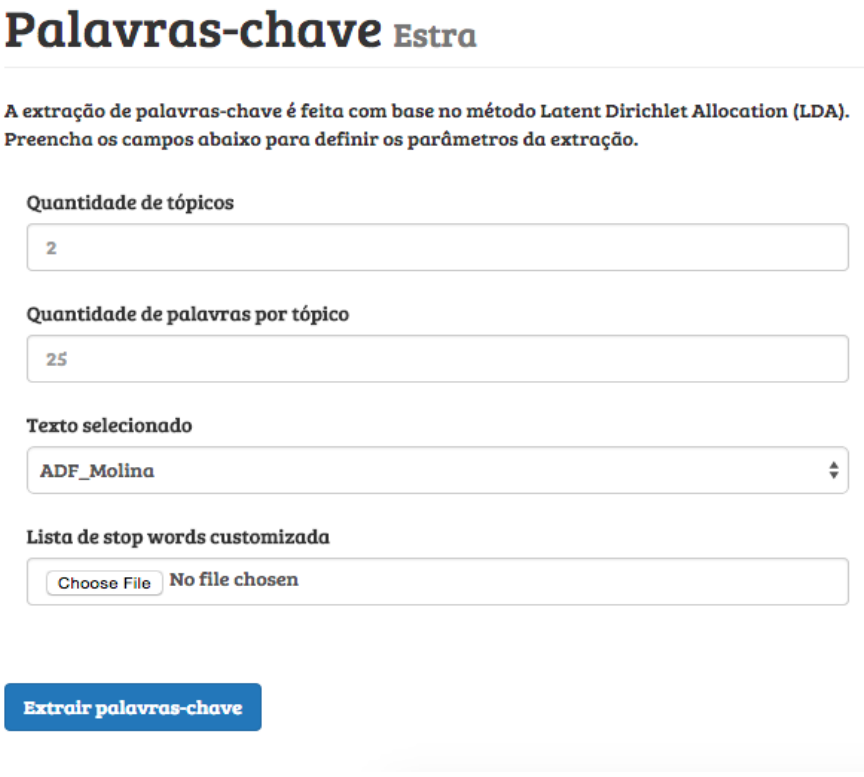
The Portal is based on eight core modules for general features to access corpora, and six support modules, created for managing and importing corpora. The core modules are: concordances, alignments, statistics and frequencies, keywords (focus of this paper), annotations, and multimodal corpus. In addition to the core modules, the Portal also has supporting modules, responsible for managing users, as well as importing and managing corpora, namely: importing module, text manager, corpus manager, subcorpus manager, tag manager and user manager.

In particular, the module used in this work allows automatic keyword generation. In this task, the Portal acts as an interface for the LDA-C tool (BLEI, 2003), an implementation of the LDA method with only command line interface available. During the extraction process, the user must inform the number of topics to be extracted. With a single topic, the LDA is similar to other methods used in keyword extraction software. From two or more topics on, however, the tool clusters keywords into coherent topics using statistical techniques.

The keyword extraction module has a practical interface and does not require any technical knowledge from the user, who can perform automatic extraction of texts with only three parameters: the number of topics, the number of keywords for each topic and the analysis text, as illustrated in Figure 1. An optional fourth parameter is a customized list

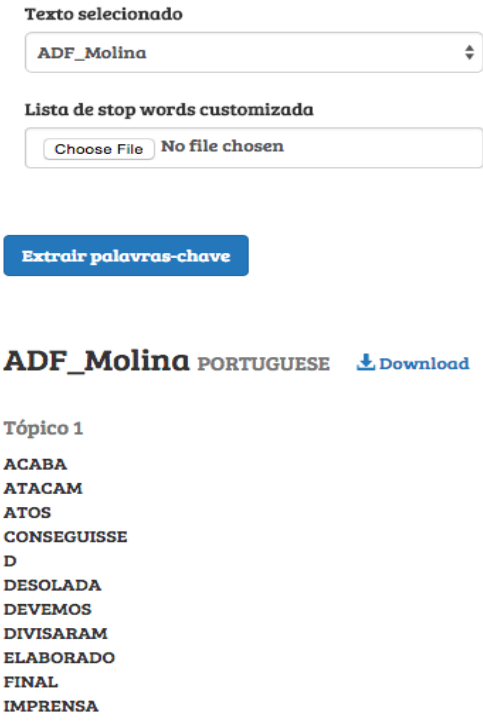
of stop words that can be sent by the user; although *Portal Min@s* has a standard list for Portuguese, English and Spanish, the user has the option of using their own list to perform extraction in a specific text. The list of extracted keywords is displayed on the screen, as shown in Figure 2, but it is also available for download in plain text format.

Figure 1 - Interface of the keyword extraction tool of *Portal Min@s*



Source: <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>.

Figure 2 - Example of keyword extraction with display of results



Source: <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>.

Section 3 presents the experiments carried out to compare LDA and WordSmith keyword extraction.

3 Experiments

The idea of the study was to generate keywords through WordSmith Tools and LDA (Section 3.1), and then make comparisons between the keywords generated by analyzing which of them are common and which are specific to each method, and thus understand how each method works and their effectiveness for literary texts (Section 3.2).

In the experiments described in this article, we used only 800 of the 1,024 texts of the reference corpus, PLN-BR GOLD, enough to obtain a subcorpus approximately 5 times greater than the study corpus, following Sardinha's recommendations (2000). The subcorpus had 226,722 tokens and 17,002 sentences.

To apply the methods and compare their keyword extraction, we carried out studies with the literary corpus described in Section 2.2.

3.1 Keyword generation via WordSmith Tools and LDA

WordSmith generates a list of keywords from each corpus' list of word frequencies. The generated list has both positive and negative keywords. Positive keywords are the words of the study corpus that occur more frequently than the average set by the reference corpus, and negative keywords are those below the average rate. For this study, we have considered only positive keywords. Thus, three documents of keywords were generated: one for each translation. The first translation had 61 keywords, the second had 66 and the third had 60.

For the second method, the Latent Dirichlet Allocation (LDA), we have used implementation in language C, provided by one of the authors, called LDA-C (BLEI, 2003). The version also brings a Python script, which helps generating the list. The existing documentation is not complete; therefore, some details of the tool operation are not clear. As a result, this experiment sought to use as parameters the default values recommended by the documentation.

WordSmith and *Portal Min@s* generate slightly different frequency lists due to variations in the tokenization process. For this reason, we made some changes in the experiments to use the WordSmith frequency list as input to LDA. Two scripts have been developed to adapt the output to the format:

[M] [term_1]:[count_1] [term_2]:[count_2] ... [term_N]:[count_N],

where [M] is the number of single terms in the document, and the number associated with each term means how many times the term occurred in the document (BLEI, 2003). These generated files enable to

run LDA-C and generate output files with the keywords. In an exploratory approach, two topics were defined for extraction through the LDA method. An additional script was made to post-process the output and generate the files to be submitted for human analysis.

After applying the tools in the three case studies, 9 files were obtained:

- 3 files containing lists of keywords generated by WordSmith Tools, one for each translation; and
- 6 files containing the topics and their keywords generated by LDA, two files for each translation.

From these 9 files, it was possible to cross data and generate new results to refine the final analysis. Two new processes were performed: the intersection between the lists of keywords generated by each tool separately, highlighting in which case study each keyword occurs (TABLE 2), as well the intersection between the final values of each tool (TABLE 3).

Table 2 - Number of keywords generated in each translation (case) for each tool

	Case 1	**Case 2**	**Case 3**	**Intersection**
WordSmith Tools	61	66	60	93
LDA (2 topics)	150	146	153	227

Table 3 - Keywords generated by each tool, considered as a whole

	WordSmith Tools	**LDA**	**Both**
Occurs in 1 case	37	87	11
Occurs in 2 cases	20	56	7
Occurs in 3 cases	36	84	22

Finally, Table 4 presents the 10 best candidates for keywords, i.e. those words generated by both tools and that occur in the three translations (3 cases).

Table 4 - The 10 best candidates for keywords

Kurtz	Homem (Man)	Meu (My)	Tão (So)	Terra (Earth)
Cabeça (Head)	Olhos (Eyes)	Homens (Men)	Marfim (Ivory)	Peregrinos (Pilgrims)

The next step was to analyze these results and come up with a hit rate for each tool. To this end, the candidates generated were analyzed by an expert in the respective text. This analysis sought to determine whether the tools studied are accurate enough to extract keywords from a literary corpus.

3.2 Analysis

The analysis of keyword candidates started with the lists generated by the Keywords module of WordSmith Tools 6.0. The method comprises the analysis of the combined keywords in all the three translations. The keywords were automatically combined into a single list using a Python script implemented specifically to this task.

In sum, the translations provided 93 unique words, 37 of which were common to all three translations. The keywords were mostly lexical, except for *parecia(m)* (the verb 'to seem' in Portuguese), forms of address and proper names. This list of words is shown in Table 5 (see APPENDIX A).

The list from the LDA keywords for the two topics of each text contained 249 unique words, 78 of which were common to both topics of all three translations. The keywords were mostly lexical, except for *parecia(m)* (the verb 'to seem' in Portuguese), forms of address and proper names. This list of words is shown in Table 6 (see APPENDIX A).

For the analysis of candidates for keywords in the lists generated by LDA, we decided to use an extra approach due to the difficulty in naming a topic for each of the two lists generated for each text, given the

similarity of the lists and adversity of words listed. The methodology used in the analysis is described below:

1) the words were clustered into microtopics;
2) some of these topics were discarded because they are more related to textual characteristics (for example, the topic description of actions, relations and utterances that integrated verbs; the topic that comprised first-person pronouns, and topics that we could not identify because they are conjunctions, interjections etc.); and
3) microtopics were clustered into macrotopics, resulting in two macrotopics, although we were not sure whether some of the words they integrated were relevant.

The results of the analysis of lists generated by LDA are presented below.

3.2.1 HOD_Trevisan

For the HOD_Trevisan_tópico1 list, with 99 words,[9] the following microtopics were obtained:

1) Narrative: *me* (me), *minha* (my), *meu* (my), *mim* (me), *meus* (my) (5);
2) Human: *homem* (man), *olhos* (eyes), *cabeça* (head), *voz* (voice), *mãos* (hands), *homens* (men), *Mr* (Mr.), *peregrinos* (pilgrims), *braço* (arm), *jovem* (young), *passos* (steps), *homens* (men), *diabo* (devil), *braços* (arms) (14);
3) Nature: *rio* (river), *floresta* (forest), *mundo* (world), *margem* (shore), *treva* (darkness), *luz* (light), *terra* (earth), *água* (water), *árvores* (trees), *mar* (sea), *sol* (sun), *coisa* (thing), *sombra* (shadow), *coisas* (things), *forma* (shape), *escuridão* (dark), *coração* (heart) (18);
4) Time: *tempo* (time), *vez* (time), *vezes* (times), *momento* (moment), *sempre* (always), *meses* (months), *instante* (instant)

[9] In this list, '*negros*' (the black) and '*selvagem*' (wild) could also be part of the 'human' topic; '*volta*' (around) could be under 'actions description'.

(7);

5) Description of action, relations and utterances: *disse* (said), *dizer* (say), *olhar* (look), *poderia* (could), *parecia* (seemed), *encontrava* (found), *fosse* (were), *pareciam* (seemed), *saber* (know), *ficou* (became), *podia* (could), *sei* (know), *ouvi* (heard), *ver* (see) (14);
6) Another description: *direção* (direction), *frente* (front), *acima* (above), *caminho* (way), *movimento* (movement), *volta* (around), *nada* (nothing), *qualquer* (any), *tão* (so), *menos* (less), *sequer* (even), *nenhum* (none), *antes* (before), *ninguém* (nobody), *grande* (large), *claro* (clear), *certo* (right), *longo* (long), *possível* (possible), *pequeno* (small), *selvagem* (wild), *negros* (black), *dois* (two), *duas* (two) (24);
7) Human artifacts: *barco* (boat), *cabine* (cabin), *porta* (door), *companhia* (company) (4);
8) Abstractions: *verdade* (truth), *impressão* (impression), *respeito* (respect), *silêncio* (silence), *existência* (existence), *poder* (power), *razão* (reason), *vida* (life), *morte* (death) (9); and
9) No topic: *oh* (oh), (*no*) *entanto* (but),[10] *porém* (however), *embora* (but) (4).

Topics 1, 5 and 9 were disregarded. Microtopics 2, 4, 7 and 8 were clustered into a
"human nature" macrotopic (34 words in total) and microtopics 3 and 6 were clustered into a 'physical nature' macrotopic (42 words in total). Based on the number of words, we can say that the 'physical nature' macrotopic is predominant in this list.

For the HOD_Trevisan_tópico2 list, with 98 words,[11] the following microtopics were obtained:

1) Narrative: *me* (me), *meu* (my), *minha* (my), *mim* (me), *minhas*

[10] The correct translation of "but" is "no entanto". As the align is based on a one-to-one correspondence of lexical items, "but" was aligned only with "entanto".

[11] In this list, '*preciso*' (accurate) and '*volta*' (around) could also belong to the 'actions decription' microtopic, etc. and '*negro*' (black) and '*selvagem*' (wild) could also belong to the 'human' microtopic.

(my), *meus* (my) (6);

2) Human: Kurtz, *homem* (man), *administrador* (administrator), *homens* (men), *voz* (voice), *pés* (feet), *espécie* (species), *peregrinos* (pilgrims), *corpo* (body), *olhos* (eyes), *Deus* (God), *rosto* (face), *alma* (soul), *Mr* (Mr.), *tom* (tone) (15);
3) Nature: *forma* (shape), *rio* (river), *terra* (earth), *mar* (sea), *ar* (air), *floresta* (forest), *sol* (sun), *selva* (jungle), *margem* (shore), *colina* (hill), *marfim* (ivory), *coisa* (thing), *coração* (heart), *luz* (light) (14);
4) Time: *tarde* (afternoon), *noite* (night), *dia* (day), *dias* (days), *momento* (moment), *tempo* (time), *instante* (instant), *vezes* (times) (8);
5) Description of action, relations and utterances: *parecia* (seemed), *disse* (said), *tivesse* (had), *dizer* (say), *pode* (can), *poderia* (could), *sabia* (knew), *vi* (saw), *preciso* (need), *exclamou* (exclaimed), *perguntei* (asked), *olhar* (look), *ver* (see), *podia* (could), *imaginar* (imagine), *sei* (know), *falar* (speak), *murmurou* (muttered), *esperar* (wait), *saber* (know), *posso* (could) (21);
6) Another description: *nada* (nothing), *tão* (so), *mal* (bad), *melhor* (best), *simples* (simple), *maior* (greater), *impossível* (impossible), *longo* (long), *selvagem* (wild), *negro* (black), *capaz* (able), *direção* (direction), *antes* (before), *dois* (two), *duas* (two), *volta* (around), *quilômetros* (kilometers), *lugar* (place), *meio* (middle), *lado* (side), *ponto* (point) (21);
7) Human artifacts: *entreposto* (customs), *barco* (boat), *posto* (post), *trabalho* (work), *cerca* (fence) (5);
8) Abstractions: *silêncio* (silence), *verdade* (truth), *ideia* (idea), *realidade* (reality), *respeito* (respect), *fim* (end) (6); and
9) No topic: (*no*) *entanto* (but), *porém* (however) (2).

Microtopics 1 and 5 were disregarded. The microtopics 2, 4, 7 and 8 (34 words) were reclustered into a macrotopic called ‘human nature’ and microtopics 3 and 6 (35 words) under the ‘physical nature’ macrotopic. The two macrotopics have a balanced number of words in this list without predominance of any of them. This list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary. An example is ‘Kurtz’, the most representative character of human degradation at the end of the

colonization in the 19th century.

For the HOD_Trevisan list, only the words common to the two previous lists were considered, clustered into the two macrotopics. In a total of 31 words, we obtained the following results:

1) Human nature: *homem* (man), *noite* (night), *olhos* (eyes), *peregrinos* (pilgrims), *barco* (boat), *momento* (moment), *dia* (day), *verdade* (truth), *respeito* (respect), *voz* (voice), *tempo* (time), *homens* (men), *instante* (instant) (13); and
2) Physical nature: *luz* (light), *margem* (shore), *direção* (direction), *coração* (heart), *nada* (nothing), *tão* (so), *rio* (river), *frente* (front), *volta* (around), *dois* (two), *forma* (shape), *antes* (before), *floresta* (forest), *longo* (long), *selvage* (wild), *sol* (sun), *terra* (earth), *mar* (sea), *coisa* (thing) (19).

The results show that the words clustered into the "physical nature" macrotopic, comparing the two lists, Topic 1 and Topic 2, have the most occurrences, confirming the dominance of this topic in the text. Words of higher frequency in this list could also be the candidates for keywords.

3.2.2 HOD_Junqueira

For HOD_Junqueira_tópico1, with 99 words,[12] we obtained the following microtopics:

10) Narrative: *me* (me), *minha* (my), *mim* (me), *meu* (my), *minhas* (my) (5);
11) Human: Kurtz, *senhor* (Mr.), *homem* (man), *cabeça* (head), *voz* (voice), *homens* (men), *Sr.* (Mr.), *gente* (people), *olhos* (eyes), *rosto* (face), *corpo* (body), *mente* (mind), *alma* (soul) (13);
12) Nature: *rio* (river), *marfim* (ivory), *mundo* (world), *coisa* (thing), *mata* (forest), *trevas* (darkness), *forma* (shape), *coração* (heart), *árvores* (trees), *sombra* (shadow), *vista*

[12] In this list, '*sombra*' (shadow) could also be part of the 'human' topic and '*vista*' (view) could be part of the 'actions description' topic, etc.

(view), *margem* (shore), *coisas* (things), *mar* (sea) (14);

13) Time: *tempo* (time), *sempre* (always), (*de*) *repente* (suddenly),[13] *vezes* (times), *noite* (night), *momento* (moment), *vez* (time), *dia* (day) (8);
14) Description of action, relations and utterances: *parecia* (seemed), *falei* (said), *falou* (said), *ver* (see), *olhar* (look), *disse* (said), *falar* (speak), *tivesse* (had), *pareceu* (seemed), *iria* (would go), *ouvir* (hear), *creio* (believe), *podia* (could), *ouvi* (heard), *fiquei* (was), *via* (saw), *ficava* (stay), *voltar* (return), *saber* (know), *ficar* (stay), *vi* (saw), *acho* (think), *ia* (went) (23);
15) Another description: *grande* (large), *certa* (right), *possível* (possible), *velho* (old), *maior* (greater), *doente* (sick), *novo* (new), *tão* (so), *qualquer* (any), *nada* (nothing), *ninguém* (nobody), *mal* (bad), *demais* (too), *maneira* (way), *nenhum* (none), *meio* (middle), *lado* (side), *redor* (around), *direção* (direction), *caminho* (way), *junto* (together), *lugar* (place) (22);
16) Human artifacts: *barco* (boat), *posto* (post), *porta* (door), *casa* (house), *ponto* (point), *sala* (room) (6);
17) Abstractions: *respeito* (respect), *dor* (pain), *vida* (life), *fim* (end), *final* (final), *verdade* (truth) (6); and
18) No topic: *porém* (but), *embora* (although) (2).

Topics 10, 14 and 18 were disregarded. Microtopics 11, 13, 16 and 17 (33 words in total) were clustered into a "human nature" macrotopic and microtopics 12 and 15 (36 words in total) were clustered into a 'physical nature' macrotopic. It can be argued that the 'physical nature' macrotopic, with the largest number of words in this list, would be the predominant topic of the list.

For HOD_Junqueira_tópico2, with 99 words,[14] we obtained the following microtopics:

10) Narrative: *me* (me), *meu* (my), *mim* (me), *meus* (my), *minha*

[13] The correct translation of "suddenly" is "de repente". As the align is based on a one-to-one correspondence of lexical items, "suddenly" was aligned only with "repente".

[14] In this list, '*negro*' (black) could belong to the 'human' microtopic.

(my), *minhas* (my) (6);
11) Human: Kurtz, *homem* (man), *gerente* (manager), *Sr* (Mr.), *peregrinos* (pilgrims), *mãos* (hands), *cabeça* (head), *gente* (people), *pés* (feet), *senhor* (Mr.), *homens* (men), *mão* (hand), *Deus* (God), *sujeito* (subject), *olhos* (eyes) (15);
12) Nature: *coisa* (thing), *rio* (river), *terra* (earth), *selva* (forest), *ar* (air), *luz* (light), *coisas* (things), *água* (water), *sol* (sun), *mundo* (world), *mata* (forest), *forma* (shape), *marfim* (ivory) (13);
13) Time: *tempo* (time), *momento* (moment), *meses* (months), *noite* (night), *dia* (day) (5);
14) Description of action, relations and utterances: *dizer* (say), *ver* (see), *fosse* (were), *olhar* (look), *disse* (said), *saber* (know), *fiquei* (became), *parecia* (seemed), *pode* (can), *sei* (know), *vi* (saw), *dar* (give), *poderia* (could), *comecei* (started), *podia* (could), *sabia* (knew), *tivesse* (had) (17);
15) Another description: *negro* (black), *tão* (so), *nada* (nothing), *qualquer* (any), *menos* (less), *nenhuma* (none), *mal* (bad), *dois* (two), *duas* (two), *primeira* (first), *meio* (middle), *frente* (front), *longe* (far), *lado* (side), *perto* (near), *fundo* (deep), *acima* (above), *próprio* (own), *alto* (high), *novo* (new), *claro* (clear), *melhor* (best), *impossível* (impossible), *profunda* (deep), *perdido* (lost), *lugar* (place), *antes* (before) (27);
16) Human artifacts: *posto* (post), *trabalho* (work), *barco* (boat), *companhia* (company) (4);
17) Abstractions: *silêncio* (silence), *vida* (life), *fato* (fact), *verdade* (truth), *respeito* (respect), *desejo* (desire), *força* (strength), *morte* (death), *ideia* (idea), *fim* (end) (10); and
18) Not identified: *embora* (although), *afinal* (finally) (2).

Microtopics 10, 14 and 18 were disregarded. Microtopics 11, 13, 16 and 17 (34 words) were reclustered into a macrotopic called 'human nature' and microtopics 12 and 15 (40 words) into the 'physical nature' macrotopic. The ‘physical nature’ macrotopic prevails over the ‘human nature’ macrotopic in this list too. It is also observed that this list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary.

For the list, only the words common to the two previous lists were considered, clustered into the two aforementioned macrotopics. In a

total of 33 words, we obtained the following results:

3) Human nature: *senhor* (Mr.), *dia* (day), *noite* (night), *olhos* (eyes), *barco* (boat), *fim* (end), *vida* (life), *posto* (post), *verdade* (truth), *gente* (people), *respeito* (respect), *tempo* (time), *homem* (man), *homens* (men), *cabeça* (head), Kurtz, *momento* (moment) (17); and
4) Physical nature: *coisa* (thing), *novo* (new), *mundo* (world), *antes* (before), *coisas* (things), *mata* (forest), *meio* (middle), *lado* (side), *rio* (river), *mal* (bad), *tão* (so), *maneira* (way), *marfim* (ivory), *forma* (shape), *lugar* (place), *qualquer* (any) (16).

The results show that the "human nature" macrotopic, comparing the two lists, Topic 1 and Topic 2, has a word more than the 'physical nature' macrotopic. In this list, we can say that the two topics are balanced, so the predominance of one or the other is not confirmed. Words of higher frequency in this list could also be the candidates for keywords.

3.2.3 HOD_Santarrita96

For HOD_Santarrita96_tópico1, with 98 words,[15] the following microtopics were obtained:

19) Narrative: *me* (me), *me* (me), *minha* (my) (3);
20) Human: Kurtz, *homem* (man), *olhos* (eyes), *Sr* (Mr.), *senhor* (Mr.), *cabeça* (head), *gerente* (manager), *homens* (men), *voz* (voice), *gente* (people), *peregrinos* (pilgrims), *mão* (hand), *pés* (feet), *espécie* (species), *pé* (foot) (15);
21) Nature: *coisa* (thing), *terra* (earth), *rio* (river), *ar* (air), *mar* (sea), *mato* (forest), *marfim* (ivory), *luz* (light), *escuridão* (dark), *trevas* (darkness), *água* (water), *floresta* (forest), *margem* (shore) (13);
22) Time: *tempo* (time), *dia* (day), *repente* (suddenly), *vezes*

[15] In this list, '*negro*(s)' (the black) could be under the 'human' topic and '*volta*' (around) could be part of the 'actions description' topic, etc.

(times), *meses* (months), *dias* (days), *noite* (night) (7);

23) Description of action, relations and utterances: *disse* (said), *dizer* (say), *parecia* (seemed), *pode* (can), *olhar* (look), *ver* (see), *sei* (know), *vi* (saw), *deve* (should), *ia* (went), *sabe* (know), *falar* (speak), *queria* (wanted), *murmurou* (muttered), *fosse* (were), *saber* (know), *sentia* (felt), *pareciam* (seemed), *ouvir* (hear), *sabem* (know), *ficava* (was), exclamou (exclaimed) (22);
24) Another description: *barulho* (noise), *nada* (nothing), *lado* (side), *tão* (so), *qualquer* (any), *acima* (above), *dois* (two), *grande* (large), *negros* (black), *meio* (middle), *simples* (simple), *lugar* (place), *duas* (two), *maior* (larger), *mal* (bad), *nenhum* (none), *perto* (near), *claro* (clear), *menos* (less), *modo* (way), *volta* (around), *própria* (own), *grandes* (large), *negro* (black), *demais* (too), *doente* (sick) (26);
25) Human artifacts: *posto* (post), *vapor* (steam), *porta* (door) (3);
26) Abstractions: *vida* (life), *silêncio* (silence), *verdade* (truth), *poder* (power), *fim* (end), *sensação* (sensation), *morte* (death), *fato* (fact), *certeza* (certainty) (9); and
27) No topic: *afinal* (finally), *oh* (oh) (2).

Topics 19, 23 and 27 were disregarded. Microtopics 20, 22, 25 and 26 (34 words in total) were clustered into a "human nature" macrotopic and microtopics 21 and 24 (39 words in total) were clustered into a 'physical nature' macrotopic. It can be argued that the 'physical nature' macrotopic, with the largest number of words in this list, would be the predominant topic of the list.

For HOD_Santarrita96_tópico2, with 100 words,[16] the following microtopics were obtained:

19) Narrative: *meu* (my), *minha* (my), *me* (me), *meus* (my), *minhas* (my), *mim* (me) (6);
20) Human: Kurtz, *homem* (man), *senhor* (Mr.), *gente* (people), *gerente* (manager), *Sr* (Mr.), *homens* (men), *voz* (voice), *Deus* (God), *cabeça* (head), *mãos* (hands), *alma* (soul), *sujeito*

[16] In this list, '*negro*' (black) could also belong to the 'human' microtopic.

(subject), *peregrinos* (pilgrims), *multidão* (crowd), *face* (face), *nome* (name) (17);

21) Nature: *coisa* (thing), *rio* (river), *coisas* (things), *coração* (heart), *margem* (shore), *selva* (jungle), *sol* (sun), *terra* (earth), *árvores* (trees), *monte* (mound), *marfim* (ivory), *sombra* (shadow), *mundo* (world), *céu* (sky), *água* (water), *luz* (light), *trevas* (darkness) (17);
22) Time: *vez* (time), *noite* (night), *vezes* (times), *momento* (moment), *repente* (suddenly), *dia* (day) (6);
23) Description of action, relations and utterances: *podia* (could), *via* (saw), *parecia* (seemed), *disse* (said), *vi* (saw), *ver* (see), *sabia* (knew), *ouvi* (heard), *creio* (believe), *parece* (seem), *sabem* (know), *pareciam* (seemed), *vi* (saw), *deu* (gave), *perguntei* (asked), *fiquei* (became), *esperava* (waited), *dizia* (said), *pareceu* (seemed) (19);
24) Another description: *grande* (large), *tão* (so), *frente* (front), *qualquer* (any), *antes* (before), *meio* (middle), *alto* (high), *ninguém* (nobody), *fundo* (deep), *modo* (way), *nada* (nothing), *primeira* (first), *certa* (right), *branco* (white), *negro* (black), *trás* (behind), *menos* (less), *claro* (clear), *dois* (two), *velho* (old), *perto* (near), *próprio* (own), *região* (region) (23)
25) Human artifacts: *vapor* (steam), *casa* (house), *trabalho* (work), *companhia* (*company*), *rebites* (revites), *Europa* (Europe) (6);
26) Abstractions: *verdade* (truth), *ideia* (idea) (2); and
27) Not identified: *jamais* (never), *embora* (although), *oh* (oh), *afinal* (finally) (4).

Microtopics 19, 23 and 27 were disregarded. Microtopics 20, 22, 25 and 26 (31 words) were reclustered into the 'human nature' macrotopic and microtopics 21 and 24 (40 words) into the 'physical nature' macrotopic. The ‘physical nature’ macrotopic prevails over the 'human nature' macrotopic in this list too. It is also observed that this list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary.

For the list, only the words common to the two previous lists were considered, clustered into the two aforementioned macrotopics. In a total of 35 words, we obtained the following results:

5) Human nature: *homens* (men), *homem* (man), Kurtz, *peregrinos* (pilgrims), *Sr* (Mr.), *cabeça* (head), *verdade* (truth), *repente* (suddenly), *dia* (day), *gente* (people), *gerente* (manager), *noite* (night), *voz* (voice), *senhor* (Mr.), *vezes* (times), *vapor* (steam) (16); and
6) Physical nature: *luz* (light), *rio* (river), *terra* (earth), *coisa* (thing), *trevas* (darkness), *menos* (less), *dois* (two), *claro* (clear), *nada* (nothing), *tão* (so), *grande* (large), *negro* (black), *qualquer* (any), *água* (water), *modo* (way), *marfim* (ivory), *margem* (shore), *perto* (near), *meio* (middle) (19).

The results show that the “human nature” macrotopic, comparing the two lists, Topic 1 and Topic 2, has a word more than the ‘physical nature’ macrotopic. In this list, it can be said that the 'physical nature' topic is predominant. Words of higher frequency in this list could also be the candidates for keywords.

Section 4 discusses the results obtained by applying both methods on the literary corpus, and comparing their performance based on a quality-oriented analysis.

4 Discussion of Results

The experiment of human analysis of the lists generated by the keyword module (WordSmith) selected 37 keywords from the first list of 93 words, and 78 from the second list of 249 words. Most candidates for keywords in these lists include lexical words like nouns and adjectives, among others, with some occurrences of grammatical words such as the verb *parecer* (to seem). The occurrence of inflected words from this lemma as keywords was expected, since the version of this lemma in English, *seem*, was also noted by Stubbs (2005) as highly frequent in the word list of the novel in English. Returning to the lists of the experiment of human analysis, among the candidates are the 10 keywords presented in Table 4, illustrating the results obtained when considering the two tools used in the first experiment; for example, Kurtz, protagonist whose search and story is told by the sailor-narrator, and ivory, object of travels

during the colonization and exploration of continents such as Africa by some of the empires in the 18th and 19th centuries. It also presents words like *cabeça* (head) and *olhos* (eyes) that indicate a fragmented representation of the native population found in the description of the narrator (MAGALHÃES; ASSIS, 2009). However, this framework does not provide list words such as *parecia* (seemed / singular), *pareciam* (seemed / plural), *escuridão* (darkness), among others, indicators of key topics in the texts, such as uncertainty (STUBBS, 2005) and difficulty to understand what you see in a very different violent clash of cultures with a very distinct power position. The experiment of human analysis of word lists generated by the LDA method allowed the clustering of these words into very general topics, such as physical and human nature. It also pointed to the fact that the words found in lists in which the physical nature topic predominates would most likely be keyword candidates. On the one hand, the two general topics, physical nature and human nature, permeate the work, which can be interpreted as a representation of the Western man's confrontation with the devastation of human and physical natures in a context of colonization. For example, some words from these lists coincide with some of those presented in Table 4, such as Kurtz, pilgrims, earth and ivory. On the other hand, these general topics leave out words generated by the LDA lists (*qualquer* [any], *nenhuma* [none], *parecia* [seemed], etc.), also constituents of other more specific topics, but fundamental in the novel, such as the uncertainty and difficulty of understanding the landscape in the novel. Other words such as *escuridão* (dark), *trevas* (darkness), *sombra* (shadow) are key indicators of the specific topics are not necessarily inferred from general topics either.

We can also say that if there are corpus-based research results which show that conjunctions have a low frequency in novels, an investigation of the occurrence of words such as '(no) *entanto*' ('however'), '*porém*' ('but'), '*embora*' ('although') (the latter if it is in fact a conjunction according to the context), the LDA topics lists could be justified in order to verify the adversative conjunction function in this novel.

On Section 5, final remarks on the study are made.

5 Final remarks

The first aim of this study was to compare the extraction of keywords from a literary corpus using two popular tools, with each of them having a specific operation and efficiency proven in other types of texts and documents. In short, we applied the WordSmith Tools and LDA in three different translations of the same literary work. The WordSmith Tools method compares the text word frequency with a reference average frequency and determines keywords according to the words whose frequency is notably larger. LDA, on the other hand, makes a probabilistic analysis of the text word frequency and generates keyword divided into topics. The lists generated by each tool were concatenated, facilitating the analysis of the results.

The second aim was validate the results of the aforementioned experiment, resulting from an experiment of human analysis of keyword lists obtained with the two tools that would indicate the keyword candidates. In this experiment, we can conclude that for long texts, such as novels, the human analysis of lists is necessary at a stage prior to experiments to complement the list automatically generated, crossing the results of the tools. A suggestion for a future experiment would be to anticipate the human analysis procedure, i.e., before the integration of the results obtained automatically with the two tools, or even before setting the parameters that allow running LDA, generating more specific and robust results for larger literary texts or corpora.

The results indicate that both methods can be applied to the literary text. However, we observed that human analysis and refining of the methods are crucial. We should also mention that the linguistic intuition of human analysts on examining lists generated by each of the two methods in this experiment was more favorable to the use of WordSmith Tools keyword lists.

In this paper, two topics were automatically generated by LDA, which tended to cluster keywords into two macrotopics from a linguistic point of view: human and physical nature. Future research work include: (a) use LDA extraction with three computational topics and analyze the macrotopics obtained; (b) use a larger number of computational topics to

see if the resulting lists tend to be clustered according to the linguistic microtopics obtained in the human analysis; and (c) include new methods and tools in the analysis, such as WMatrix.

6 Acknowledgements

The authors thank CAPES for financing the *Portal Min@s: corpora de fala e escrita* (AUX 151/2013) project, which enabled this research, and all partners from the Interinstitutional Center for Computational Linguistics (NILC) and the Experimental Translation Laboratory (LETRA) who supported this project, especially Professor (Ph.D.) Sandra Maria Aluísio, who generously participated in this article, providing relevant comments and suggestions, and researcher Thais Blauth, who collaborated in the analysis of lists generated by WordSmith Tools.

7 References

BERBER-SARDINHA, T. Comparing *corpora* with WordSmith Tools: How large must the reference corpus be? São Paulo, 2000.

BLEI, D. M. Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large documents archives. Communications of the ACM, s.l., n. 55, p. 77-84, 2012.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, s.l., n. 3, p. 993-1022, 2003.

CONRAD, J. *O coração da treva*. Translated by Hamilton Trevisan. São Paulo: Global Editora e Distribuidora Ltda., 1984.

CONRAD, J. *O coração das trevas*. Translated by Regina Régis Junqueira. Belo Horizonte: Editora Itatiaia Ltda., 1984.

CONRAD, J. *O coração das trevas*. Translated by Marcos Santarrita. Rio de Janeiro: Ediouro S. A., 1996.

DAVIES, M. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. International Journal of Corpus Linguistics 10: 301-28, 2005. DOI: <http://dx.doi.org/10.1075/ijcl.10.3.02dav>.

DAVIES, M. Relational databases as a robust architecture for the analysis of word frequency. In What's in a Wordlist?: Investigating Word Frequency and Keyword Extraction, ed. Dawn Archer. London: Ashgate, p. 53-68, 2009.

DREDZE, M.; WALLACH, H. M.; PULLER, D; PEREIRA, F. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces* (IUI '08). ACM, New York, NY, USA, p. 199-206, 2008. DOI: <http://dx.doi.org/10.1145/1378773.1378800>.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. Discourse Processes, 25:259–284, 1998. DOI: <http://dx.doi.org/10.1080/01638539809545028>.

LIU, Z.; HUANG, W.; ZHENG, Y.; SUN, M. Automatic Keyphrase Extraction via Topic Decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, p. 366–376, 2010.

MAGALHÃES, C; ASSIS, R. C. Representação de atores sociais em corpus paralelo: Heart of Darkness e suas traduções para o português. In COHEN, Maria Antonieta; LARA, Gláucia Muniz Proença. (Org.). *Linguística, tradução, discurso.* Belo Horizonte: Editora UFMG, 2009, p. 201-220.

MANNING, C. D.; H. SCHÜTZE. Foundations of statistical natural language processing. MIT Press, 2000.

MCINTYRE, D.; WALKER, B. How can corpora be used to explore the language of poetry and drama. In: O'KEEFE, A.; McCARTHY, M. (eds.). *The Routledge handbook of corpus linguistics*. London; New York: Routledge, 2010, p. 516-530.
DOI: <http://dx.doi.org/10.4324/9780203856949.ch37>.

MUNIZ, M.; PAULOVICH, F.; Minghim, R.; INFANTE, K.; Muniz, F.; VIEIRA, R.; ALUÍSIO, S. M. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: Corpus Linguistics 2007 Conference, 2007, Birmingham. Proceedings of the Corpus Linguistics 2007 Conference, 2007.

RAYSON, P. E. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Lancaster University, 2002.

SCOTT, M. WordSmith Tools Manual. Oxford: Oxford University Press, 1996.

SCOTT, M. PC analysis of keywords - and key keywords. *System*, vol 25, no. 2, 1997, p. 233-245. DOI: <http://dx.doi.org/10.1016/S0346-251X(97)00011-0>.

STUBBS, M. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, vol 14, no. 1, p. 5-24, 2005.
DOI: <http://dx.doi.org/10.1177/0963947005048873>.

TDK TECHNOLOGIES. Topic Modeling Explained: LDA to Bayesian Inference. Retrieved on: July 12, 2015, from: <https://www.tdktech.com/tech-talks/topic-modeling-explained-lda-to-bayesian-inference>.

Apêndice A – Generated Keywords

Table 5 – Keywords obtained from WordSmith Tolls

PODIA (COULD)	SELVAGEM (WILD)
LUZ (LIGHT)	SEQUER (EVEN)KURTZ
ADMINISTRADOR (MANAGER)	VERDADE (TRUTH)
MR (MR – Mister)	SR (MR)
TERRA (EARTH)	HOMENS (MEN)
SOMBRA (SHADOW)	PARECIA (SEEMED)
AR (AIR)	POSTO (POST)
MAR (SEA)	NEGRO (BLACK)
VAPOR (STEAM)	TALVEZ (MAYBE)
ÁGUA (WATER)	GERENTE (MANAGER)
ENTREPOSTO (CUSTOMS)	FLORESTA (FOREST)
BARCO (STEAMBOAT)	SENHOR (MR)
MARGEM (SHORE)	MARFIM (IVORY)
SELVA (JUNGLE)	ESCURIDÃO (DARKNESS)
MATA (FOREST)	CORAÇÃO (HEART)
SILÊNCIO (SILENCE)	PARECIAM (SEEMED)
NADA (NOTHING)	PEREGRINOS (PILGRIMS)
VOZ (VOICE)	

Table 6 – Keywords obtained from LDA

PARECIAM (SEEMED)	COLINA (HILL)
VERDADE (TRUTH)	DIABO (DEVIL)
NADA (NOTHING)	TIMONEIRO (HELMSMAN)
MATAGAL (SCRUB)	ÁRVORES (TREES)
RAZÃO (REASON)	RIBANCEIRA (BANK)
MÃO (HAND)	NAVIO (SHIP)
VULTOS (FIGURES)	ROSTOS (FACES)
MATO (JUNGLE)	ATMOSFERA (ATMOSPHERE)
PÉS (FEET)	VOCÊ (YOU)
SOLIDÃO (LONELINESS)	TREVAS (DARKNESS)
LEME (HELM)	PARECEU (SEEMED)
QUILÔMETROS (KILOMETERS)	TEMPO (TIME)
CAPIM (GRASS)	NAVIOS (SHIPS)
SELVAGEM (WILD)	CLARÃO (GLARE)
TÊNUE (FINE)	AR (AIR)
REMANSO (QUIET)	SOL (SUN)
REBITES (RIVET)	FLORESTA (FOREST)
TREVA (DARKNESS)	TOM (TONE)
MARINHEIRO (SAILOR)	QUIETUDE (QUIET)
ROSTO (FACE)	TOLO (SILLY)
NEGROS (THE BLACK)	BRAÇOS (ARMS)
MARGEM (SHORE)	TOCO (STUB)
SOMBRA (SHADOW)	NATIVOS (NATIVES)
CONVÉS (DECK)	TERRÍVEL (TERRIBLE)
VAPOR (STEAM)	CIMA (TOP)
MISTÉRIO (MYSTERY)	SOMBRAS (SHADOWS)
SELVAGENS (WILD)	ALMA (SOUL)
CÁ (HERE)	ESCURIDÃO (DARKNESS)
VOZES (VOICES)	FULGOR (GLOW)
POSTO (POST)	MARLOW
IMPRESSÃO (IMPRESSION)	GRITO (SCREAM)
PADIOLA (LITTER)	NARIZ (NOSE)
MÃOS (HANDS)	CÉU (SKY)
PROFUNDEZAS (DEPTHS)	CABINA (CABIN)
DEMÔNIO (DEMON)	CABINE (CABIN)
LENHA (FIREWOOD)	MONTE (MOUND)
NEVOEIRO (FOG)	RIO (RIVER)
MATA (FOREST)	NUS (NAKED)
HORROR (HORROR)	NEGRO (BLACK)