Cultura e Tradução v. 5, n. 1 (2017) ISSN: 2238-9059
Disponível em: http://periodicos.ufpb.br/ojs2/index.php/ct

ESTUDOS DA TRADUÇÃO, ACD E LINGUÍSTICA DE *CORPUS*: DIÁLOGOS POSSÍVEIS

Priscila de Oliveira Novais LIMA (UFPB)

RESUMO: O presente trabalho tem como objetivo apresentar uma metodologia de pesquisa que promove uma área de convergência entre os Estudos da Tradução, a Análise Crítica do Discurso (doravante ACD) e a Linguística de *Corpus* (doravante LC). A LC é uma área da Linguística que se dedica à coleta e exploração de *corpora*, ou conjuntos de dados textuais (textos ou transcrições de fala) armazenados em arquivos de computador (SARDINHA, 2006). A LC tem revolucionado a maneira como se analisa a linguagem, pois conta com ferramentas computacionais especialmente criadas para investigar uma grande quantidade de dados linguísticos, fornecendo ao linguista os meios para analisá-los nos mais diversos níveis. A referida metodologia foi utilizada em um estudo cujo escopo foi o de investigar recontextualizações da prática social que ficou conhecida como "Manifestações", ocorridas no Brasil no ano de 2013. Mais especificamente, no estudo citado foram investigadas as formas através das quais ocorreram as representações de Manifestantes e de Agentes Públicos em textos jornalísticos escritos em português e em inglês, tendo como principal embasamento teórico a Teoria de Representação de Atores Sociais, proposta por Van Leeuwen (1997) . Os aludidos textos foram retirados dos jornais *The Chicago Tribune*, *The New York Times*, *The Guardian*, *Correio Braziliense, A Folha de São Paulo* e *Jornal do Brasil.* Considerando a quantidade expressiva de material textual a ser analisado – em torno de cinco mil palavras – e a grande quantidade de categorias de análise encontradas na teoria que embasou a pesquisa, observou-se a necessidade de se empregar uma metodologia que simplificasse a busca pelos dados linguísticos nos textos. Nesse sentido, seguindo os pressupostos da LC, foi utilizado um *software* chamado de Concordanciador, que é um programa que extrai todas as ocorrências de uma palavra buscada em um *corpus*. No entanto, a utilização deste programa no estudo em questão demandou alguns procedimentos metodológicos específicos, tais como: coleta de textos segundo o critério estabelecido; confecção de etiquetas de análise, as quais reuniram informações quanto ao nome do jornal, bem como quanto ao grupo de atores em questão e sua classificação segundo o sistema RAS (1997); a limpeza (retirada das imagens e hiperlinks) e preparação do *corpus* em formato específico; a anotação do *corpus* com as etiquetas elaboradas; utilização do Concordanciador para fazer pesquisas específicas dentro do *corpus* e a análise e discussão crítica e comparativa acerca das representações observadas dos atores sociais investigados nos textos. Em linhas gerias, a utilização do ferramentário metodológico advindo da LC se mostrou profícua em estudos que, como o citado neste trabalho, tenham uma grande quantidade de material textual a ser investigada e/ou abarquem uma ampla gama de categorias de análise, pois fornecem a sistematização necessária para tanto. Um exemplo de aplicação prática do sustentáculo metodológico da LC e suas contribuições aos Estudos da Tradução e à ACD serão apresentados neste trabalho.

Palavras-Chave: Análise crítica do discurso; Linguística de c*orpus*; Representação de atores sociais.

ABSTRACT: *The present work aims to present a research methodology that promotes an area of convergence between Translation Studies, Critical Discourse Analysis (hereafter CDA) and Corpus Linguistics (hereinafter CL). CL is an area of Linguistics that is dedicated to the collection and exploitation of corpora, or sets of textual data (text or speech transcriptions) stored in computer files (SARDINHA, 2006). CL has revolutionized the way language is analyzed, for it has computational tools specially designed to investigate a large amount of linguistic data, providing the analist with the means to analyze such data at the most diverse levels. The mentioned methodology was used in a study which scope was to investigate recontextualizations of the social practice that became known as "Manifestations", occurred in Brazil in the year 2013. More specifically, in the study, it has been investigated the ways through which the representations of Manifestants and of Public Agents*

occurred in journalistic texts written in Brazilian Portuguese and English, having as its main theoretical support the Theory of Representation of Social Actors, proposed by Van Leeuwen (1997). The aforementioned texts were taken from The Chicago Tribune, The New York Times, The Guardian, Correio Braziliense, Folha de São Paulo and Jornal do Brasil. Considering the expressive amount of textual material to be analyzed - around five thousand words - and the great amount of categories of analysis found in the theory that underpinned the research, it was observed the need to use a methodology that simplified the search for the data in the texts. In this sense, following the CL assumptions, we used a software called Concordancer, which is a program that extracts all occurrences of a word searched in a corpus. However, the use of this program in the study required some specific methodological procedures, such as: collection of texts according to the established criteria; production of analysis labels, which gathered information on the name of the newspaper, as well as the group of actors in question and their classification according to the RSA system (1997); cleaning (removal of images and hyperlinks) and preparation of the corpus in a specific format; annotation of the corpus with created labels; use of the Concordancer to do specific research within the corpus, and analysis as well as critical and comparative discussion about the observed representations of the social actors investigated in the texts. The use of the methodological tools from CL has proved itself to be fruitful in studies which, like the one cited in this study, have a large amount of textual material to be investigated and / or cover a wide range of analysis categories, for they provide the necessary systematization. An example of practical application of the methodological support of CL and its contributions to the Translation Studies and to the CDA will be presented in this work.

Keywords: *Critical Discourse Analysis; Corpus Linguistics; Representation of Social Actors*

INTRODUÇÃO

Este artigo visa apresentar uma metodologia de pesquisa que promove uma interface entre a ACD e a Linguística de *Corpus*. A metodologia a que estamos no referindo foi utilizada em um trabalho de dissertação de Novais (2015), intitulado *A representação de manifestantes e agentes públicos como atores sociais em textos sobre os protestos no brasil em 2013*, cujo escopo foi o de investigar recontextualizações da prática social que ficou conhecida como "Manifestações", ocorridas no Brasil no ano de 2013.

Na referida pesquisa, foram analisadas as formas através das quais os atores sociais envolvidos nos protestos acontecidos no Brasil em junho de 2013 foram representados em textos jornalísticos escritos em português e em inglês. Sendo assim, o *corpus* de pesquisa foi formado a partir de textos em duas línguas – inglês e português –, retirados dos jornais *The Chicago Tribune*, *The New York Times*, *The Guardian*, *Correio Braziliense, A Folha de São Paulo* e *Jornal do Brasil*.

O *corpus* do estudo em questão foi montado a partir de critérios linguísticos e de representatividade: os textos em português foram selecionados por se tratarem de publicações advindas de jornais de grande circulação no Brasil. Da mesma forma, os textos selecionados em inglês foram publicados em jornais de grande circulação em língua inglesa, independentemente de qual variação da língua inglesa o jornal adota. Importa destacar que, no referido estudo, não foram feitas caracterizações individuais acerca dos posicionamentos políticos e ideológicos dos jornais investigados, nem dos países onde foram publicados, devido ao caráter sociológico que uma pesquisa assim denotaria.

Deste modo, a fim de possibilitar a comparação entre os padrões de representação dos atores sociais nos textos da pesquisa em questão, foi necessário que os participantes fossem agrupados sob denominadores comuns, os quais funcionam como âncoras na análise. Nesse sentido, os atores sociais foram agrupados sob os denominadores *Manifestantes* e *Agentes Públicos* e foram assinalados nos textos como pertencentes a um ou a outro grupo.

O aludido estudo teve como sustentáculo teórico a Teoria de Representação de Atores Sociais (doravante RAS), teoria que deriva da Linguística Sistêmico-Funcional e que vem se consolidando como ferramenta de análise linguística dentro dos Estudos Críticos do Discurso. Essa teoria apresenta um inventário sociossemântico das formas através das quais os atores sociais podem ser representados no discurso e, por essa razão, contempla uma grande variedade de categorias linguísticas, como veremos mais adiante.

Em decorrência da referida necessidade de sistematização para a aplicação do sistema RAS a análises linguísticas, recorremos à Linguística de *Corpus* (doravante LC), que é uma área da Linguística que se dedica à coleta e exploração de *corpora*, ou conjuntos de dados textuais (textos ou transcrições de fala) armazenados em arquivos de computador (SARDINHA, 2006). A LC tem revolucionado a maneira como se analisa a linguagem, pois conta com ferramentas computacionais especialmente criadas para investigar uma grande quantidade de dados linguísticos, fornecendo ao linguista os meios para analisá-los nos mais diversos níveis.

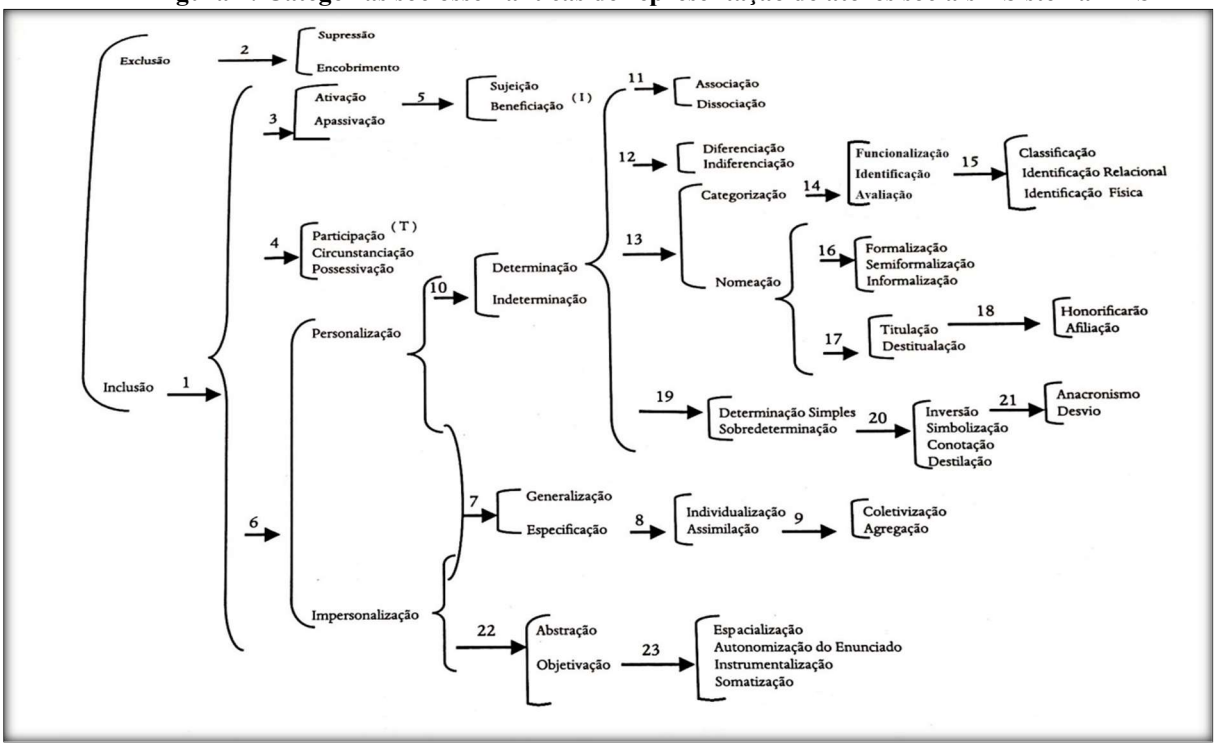
Este trabalho visa apresentar o uso das ferramentas da LC no estudo em questão e suas possíveis contribuições às pesquisas situadas nos Estudos Críticos do Discurso. Na seção seguinte, apresentaremos a Teoria de Representação de Atores Sociais e a LC.

PRESSUPOSTOS TEÓRICO-METODOLÓGICOS

Na Teoria de Representação de Atores Sociais, van Leeuwen (1997) apresenta um inventário das formas através das quais os atores sociais podem ser representados no discurso. O autor ainda ressalta que, diferentemente de outras modalidades de Análise Crítica do Discurso orientadas pela LSF, sua teoria parte das formas como os atores sociais estão representados, para, posteriormente, relacioná-las às realizações linguísticas, realçando, assim, a relevância sociológica e crítica de suas categorias.

Para uma melhor visualização da explicação sobre o sistema RAS, faz-se necessário apresentar sua sistematização conforme proposto por van Leeuwen (1997), como veremos na FIG 1:

Figura 1: Categorias sociossemânticas de representação de atores sociais – Sistema RAS



Fonte: Novais (2015, p. 29)

A partir da FIG. 1, observamos que o sistema RAS tem 23 sistemas, cada um com, no mínimo, 2 subsistemas. No entanto, neste trabalho, não pretendemos detalhar o sistema RAS, nem demonstrar o grau de especificidade de suas categorias[96]. Pretendemos tão somente apresentá-lo, apontado a quantidade de categorias de análise e enfatizando a necessidade de uma sistematização quando da sua aplicação para a análises linguísticas.

O referido inventário se biparte em dois sistemas: Exclusão e Inclusão, dos quais derivam todos os outros subsistemas. O sistema de Exclusão apresenta os subssistemas Supressão e o Encobrimento, ambos realizados no discurso por elementos linguísticos específicos. O sistema de Inclusão se segmenta em outros subssistemas, como Ativação e Apassivação; Participação, Circunstanciação e Possessivação; Personalização e Impersonalização, cada um deles com respectivas subdivisões, como é possível observar na FIG. 1, reproduzida a partir de Novais (2015, p. 29).

Em decorrência da quantidade expressiva de material textual a ser analisado – em torno de cinco mil palavras – e a grande quantidade de categorias de análise encontradas no sistema RAS, observou-se a necessidade de se empregar uma metodologia que simplificasse a busca pelos dados linguísticos nos textos. Nesse sentido, a LC forneceu o aparato metodológico necessário não apenas para a busca, mas também para a comparação e quantificação dos dados linguísticos.

Uma das ferramentas advindas da LC e à disposição do linguista analista é o Concordanciador, que, na definição de Tagnin (2004), é um programa que extrai todas as ocorrências de uma palavra de busca num *corpus* juntamente com seu cotexto, apresentando-as na forma de linhas de concordância. Na realidade, o Concordanciador é um conjunto de programas reunidos em um só. Conforme Sardinha (2006), o Concordanciador

[96] Para maior detalhamento e acesso a aplicação do sistema RAS a análises linguísticas, ler Novais (2015).

> (...) permite fazer análises baseadas na frequência e na co-ocorrência de palavras em *corpora*. Além disso, ele permite pré-processar os arquivos do *corpus* (retirar partes indesejadas de cada texto, organizar o conjunto de arquivos, inserir e remover etiquetas, etc.), antes da análise propriamente dita. (SARDINHA, 2006, p. 6)

O Concordanciador utilizado na metodologia da pesquisa investigada neste estudo é o AntConc 3.2.4,[97] desenvolvido por Laurence Anthony, o qual permite, entre outras funções, explorar *corpora* cujos conteúdos foram anotados. Hunston (2002) explica que anotação é o processo de acrescentar informações a um *corpus* com o objetivo de interpretá-lo linguisticamente com o auxílio de ferramentas eletrônicas. No entanto, o levantamento dos dados do *corpus* a partir do Concordanciador é a fase final de um processo que se inicia na compilação dos textos que serão analisados.

Na seção seguinte, apresentaremos o processo de criação de etiquetas para anotação no *corpus* e a busca pelos dados linguísticos a partir delas.

DISCUSSÃO DO OBJETO

Sardinha (2006, p. 112) explica que *tags* ou etiquetas são códigos de anotação do *corpus* que servem para muitas funções, como, por exemplo, identificação de classe de palavra, nomeação dos falantes, especificação de divisões do texto, etc. A metodologia da LC foi utilizada na pesquisa exatamente pelo fato de as etiquetas possibilitarem um único item linguístico ser classificado quanto a uma serie de parâmetros, de modo que o acesso a essa classificação foi facilitado pelas ferramentas eletrônicas de busca do AntConc.

Após o procedimento de coleta e seleção, o primeiro procedimento com os textos que compuseram o *corpus* da pesquisa investigada, consistiu na anotação manual com etiquetas, cujo processo de criação tomou como parâmetro as categorias do sistema RAS, a partir da seguinte grade de marcação:

[97] http://www.antlab.sci.waseda.ac.jp/software.html

Quadro 1: Grade de Marcação do *Corpus*

A ***Corpus***	**B** **Inclusão/ Exclusão**	**C** **Person./ Imperson.**	**D** **Determ./ Indeterm.**	**E** **Person./ Imperson.**	**F** **Grupo de atores**
1 Correio Braziliense	1 Inclusão	1 Personalização	1 Determinação	1 Nomeação	1 Manifestantes
2 Folha de São Paulo	2 Exclusão	2 Impersonalização	2 Indeterminação	2 Funcionalização	2 Rep. Do Governo
3 Jornal do Brasil		0 Não se aplica	0 Não se aplica	3 Classificação	
4 Chicago Tribune				4 Id. Relacional	
5 New York Times				5 Id. Física	
6 The Guardian				6 Id. p/ Vestuário	
				7 Avaliação	
				1 Abastração	
				2 Espacialização	
				3 Aut. do Enunciado	
				4 Instrumentalização	
				5 Somatização	
				6 Institucionalização	
				7 Ficcionalização	
				8 Sobrenaturalização	
				9 Primitivização	
				0 Não se aplica	

Fonte: Novais (2015, p. 48)

A coluna A identifica os jornais no *corpus*, sendo 1 para o *Correio Braziliense*, 2 para o *Folha de São Paulo*, 3 para o *Jornal do Brasil*, 4 para o *Chicago Tribune,* 5 para o *The New York Times* e 6 para o *The Guardian*. A coluna B identifica se o ator social foi incluído ou excluído da representação, sendo 1 para Inclusão e 2 para Exclusão. A coluna C identifica se os atores sociais foram personalizados ou impersonalizados, sendo 1 para Personalização, 2 para Impersonalização e 0 para quando a classificação não se aplicar, no caso de terem sido representados por Exclusão. A coluna D aborda as formas de representação por Determinação

(1) e Indeterminação (2) e, quando a classificação não for aplicável (0). A coluna E elenca as diversas formas pelas quais os atores sociais podem ser personalizados ou impersonalizados, sendo os primeiros códigos de 1 a 7 referentes às categorias de Personalização, e os últimos códigos de 1 a 9 referentes às formas de Impersonalização. O código 0 é para quando a classificação não se aplicar, no caso de uma Exclusão ou de uma Personalização por Indeterminação. Os códigos não se confundem pelo fato de estarem vinculados à coluna C (1=Personalização e 2=Impersonalização). Finalmente, a coluna F diz respeito ao grupo de atores investigados, sendo 1 para Manifestantes e 2 para Agentes Públicos.

As etiquetas foram inseridas imediatamente após o termo analisado, entre parênteses angulares < >. Esse tipo de marcação foi utilizado para que a etiqueta pudesse ser identificada como tal pelo concordanciador AntConc, embora o programa ofereça a possibilidade de alteração do símbolo identificador da etiqueta na ferramenta *Tag Settings*, na aba *Global Settings*. Essa ferramenta também possibilita acessar os textos considerando ou não as etiquetas.

A título de exemplificação, tomemos o seguinte fragmento anotado, retirado do *corpus*:

> Exemplo (1): "Quase 2 milhões de brasileiros <111131> fizeram manifestações <120001> pela redução das passagens do transporte público, contra os gastos com as obras da Copa do Mundo, pelo aumento dos recursos para a saúde e educação e contra a corrupção e a impunidade"

Nesse exemplo, temos que o primeiro termo anotado se encontra no jornal 1 (*Correio Braziliense*), o ator social foi representado por Inclusão, por Personalização, por Determinação, por Classificação e que se insere no denominador Manifestantes. Já o segundo termo anotado também se encontra no jornal *Correio Braziliense*, o ator social foi excluído da representação e se insere no denominador Manifestantes. Os três "0" indicam que as categorias de análise em questão não se aplicam àquela anotação.

A opção por códigos numéricos para as etiquetas criadas para anotação do *corpus* se deve ao fato de estas contemplarem seis parâmetros de classificação para cada item linguístico anotado. Ou seja, seria necessário inserir muitas informações em cada um dos referidos itens. A princípio, foram pensadas etiquetas com as iniciais de cada classificação, as quais foram, contudo, descartadas, por conta do tamanho que a etiqueta teria. Ao final, foi optado pelo código numérico, por se apresentar como uma alternativa viável e que demandaria um menor espaço para anotação.

Após termos apresentado a grade de marcação do *corpus*, exemplos de etiquetas e o processo de anotação no *corpus* da pesquisa investigada, apresentaremos os nódulos de busca por meio dos quais foram obtidos os dados linguísticos, a partir da ferramenta *busca*, no concordanciador AntConc.

Quadro 2: Nódulos de busca no AntConc

	MANIFESTANTES	AGENTES PÚBLICOS
Realizações	<?????1>	<?????2>
Exclusão	<?2???1>	<?2???2>
Inclusão	<?1???1>	<?1???2>
Personalização	<?11??1>	<?11??2>
Impersonalização	<?12??1>	<?12??2>
Determinação	<?111?1>	<?111?2>
Indeterminação	<?112?1>	<?112?2>
Espacialização	<?12021>	<?12022>
Autonomização do Enunciado	<?12031>	<?12032>
Instrumentalização	<?12041>	<?12042>
Institucionalização	<?12061>	<012062>
Nomeação	<?11111>	<?11112>
Funcionalização	<?11121>	<?11122>
Classificação	<?11131>	<?11132>

Fonte: Novais (2015, p. 50)

Embora a grade de marcação apresentada contemple todas as subcategorias de Personalização e Impersonalização descritas pelo sistema RAS, os nódulos de busca acima descritos fazem referência apenas às realizações encontradas no *corpus* da pesquisa analisada. O símbolo ? é um caractere curinga que representa qualquer caractere no Concordanciador AntcConc. Importa destacar que as pesquisas apenas no *subcorpus* em português foram feitas adicionando-se os códigos 1, 2 ou 3 à coluna A nos nódulos de busca ou a partir da seleção apenas da pasta "*Subcorpus* em português". Analogamente, as pesquisas no *subcorpus* em inglês foram feitas adicionando os códigos 4, 5 ou 6 à coluna A nos nódulos de busca ou a partir da seleção apenas da pasta "*Subcorpus* em inglês".

Os nódulos de busca servem para o (a) linguista analista especificar qual item ele (ela) deseja ter acesso dentro do *corpus* já anotado. Por exemplo, se o interesse for procurar por representação de Agentes Públicos que foram Institucionalizados, o nódulo de busca será <012062>; se o interesse for procurar por representação de Manifestantes que foram nomeados, o nódulo de busca será <?11111>.

A seguir, faremos algumas considerações a respeito da utilização da metodologia apresentada e suas contribuições para a ACD.

CONSIDERAÇÕES FINAIS

O presente trabalho apresentou uma metodologia de pesquisa que promove uma área de convergência entre a ACD e a LC. Mais especificamente, foi apresentada a metodologia utilizada na dissertação intitulada *A representação de manifestantes e agentes públicos como atores sociais em textos sobre os protestos no brasil em 2013*.

Primeiramente, fizemos uma contextualização do estudo aqui proposto, apontando para um área de convergência entre a Teoria de Representação de Atores Sociais e a Linguística de *Corpus*. Salientamos o potencial desta última em fornecer subsídio para a realização de análises linguísticas mais sistematizadas.

Em seguida, apresentamos os subsistemas do sistema RAS e observamos que, em decorrência da grande quantidade de categorias de análise encontradas nele, sua aplicação prática seria otimizada se realizada em conjunto com alguns preceitos LC, alguns dos quais foram apresentados, tais como o uso do Concordanciador, a criação de etiquetas e o processo de anotação no *corpus*.

Em seguida, apresentamos a grade de marcação, a partir da qual foram geradas as etiquetas de anotação no *corpus*. O processo geração das etiquetas se mostrou complexo, na medida em que cada etiqueta contemplava seis parâmetros de classificação para cada item linguístico anotado, consequência direta da grande quantidade de categorias de análise previstas no referencial teórico adotado.

A geração de etiquetas para anotação no *corpus* de nada serviria se não existisse a necessidade de pesquisa pelos itens linguísticos anotados. O processo de anotação no *corpus* possibilitou a pesquisa segundo critérios muito específicos, como foi possível observar pela quantidade de nódulos de busca apresentados no QUADRO 2.

Nesse sentido, na pesquisa em questão, foi possível ao analista pesquisar, por exemplo, realizações linguísticas somente no *subcorpus* em português; ou realizações linguísticas em que os participantes tenham sido Personalizados e Funcionalizados, ou Determinados, ou Personalizados e Nomeados, etc. Em outras palavras: tornou-se possível pesquisar categorias de maneira isolada.

Por fim, concluímos que a utilização do Concordanciador e de todas as outras ferramentas da LC envolvidas no uso deste *software*, se mostrou profícua e de expressiva relevância às pesquisas situadas na ACD, na medida em que contribui para facilitar o trabalho de busca pelos dados linguísticos. No entanto, a despeito do uso de *softwares* para se obter os dados linguísticos, a interpretação desses dados ainda é um trabalho que compete ao analista.

REFERÊNCIAS BIBLIOGRÁFICAS

ASSIS, R. C. **A representação de europeus e de africanos como atores sociais em Heart of Darkness (O Coração das Trevas) e em suas traduções para o português**: uma abordagem textual da tradução. Belo Horizonte, Universidade Federal de Minas Gerais, 2009. Tese.

HUNSTON, S. Methods in Corpus Linguistics: beyond the concordance line. In: **Corpora in Applied Linguistics**. Cambridge: Cambridge University Press, 2002. p.36-95.

Novais, P.A.O. **A representação de manifestantes e agentes públicos como atores sociais em textos sobre os protestos no brasil em 2013.** João Pessoa, Universidade Federal da Paraíba, 2015. Dissertação.

SARDINHA, T. B. **Lingüística de Corpus.** Barueri, SP: Manole, 2004.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: Vander Viana; Stella E. O. Tagnin. (Org.). **Corpora no ensino de línguas estrangeiras.** 1 ed. São Paulo: HUB Editorial, 2010, p. 357-361.

VAN LEEUWEN, T. A representação de atores sociais. In: Pedro, E. R. (Org.) **Análise crítica do discurso.** Lisboa: Editorial Caminho S.A., 1997. p.169-222.