DOI: 10.14393/DLv17a2023-35
e1735

Artigo

Caminhos possíveis: pensando a abordagem da Linguística de Corpus aplicada à Análise de Discurso Crítica[1]

Potential tracks: thinking Corpus Linguistics approach to Critical Discourse Analysis

Bianca Mara GUEDES de SOUZA[*]

RESUMO: Neste artigo buscamos apresentar caminhos iniciais e possibilidades analíticas do casamento entre a abordagem da Linguística de Corpus com a teoria da Análise de Discurso Crítica. O texto procura demonstrar para analistas críticos do discurso as vantagens e desvantagens de empenhar programas computacionais de análise de corpus em suas pesquisas. Nesse sentido, detalhamos procedimentos metodológicos realizados em dois softwares - o primeiro é o Sketch Engine para análise lexical, e o segundo é o UAM Corpus Tool, um programa para etiquetagem de corpora. Como resultado compreendemos que a relação beneficia os analistas de discurso enormemente contra as críticas que normalmente enfrentam quanto ao viés dos pesquisadores influenciarem a pesquisa, quanto a verificação e quantificação de dados e triangulação dos dados pesquisados.
PALAVRAS-CHAVE: Linguística de Corpus. Análise de Discurso Crítica. Metodologia em Linguística. Sketch Engine. UAM Corpus Tool.

ABSTRACT: In this paper we seek to present initial paths and analytical possibilities for the marriage between the approach of Corpus Linguistics and the theory of Critical Discourse Analysis. The text attempts to demonstrate to critical discourse analysts the advantages and disadvantages of engaging computer programs for corpus analysis in their research. In this sense, we detail methodological procedures carried out in two software - the first is Sketch Engine for lexical analysis, and the second is UAM Corpus Tool, a program for corpora tagging. As a result, we understand that the relationship benefits discourse analysts enormously against the criticism they usually face regarding the researchers' bias influencing the research, regarding the verification and quantification of data and triangulation of the researched data.
KEYWORDS: Corpus Linguistics. Critical Discourse Analysis. Methodology in Linguistics. Sketch Engine. UAM Corpus Tool.

Artigo recebido em: 10.03.2023
Artigo aprovado em: 22.07.2023

[1] Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior -Brasil – CAPES –Código do financiamento 001.
[*] Mestra em Estudos Linguísticos (UFU), doutoranda em Estudos Linguísticos pela UFU, biancamgsouza@gmail.com

1 Introdução

Neste artigo pretendemos apresentar a abordagem da Linguística de Corpus (LC) pensando em sua aplicação metodológica às pesquisas desenvolvidas com o aparato teórico-metodológico da Análise de Discurso Crítica (ADC)[2] também chamada de Análise de Discurso Textualmente Orientada (ADTO)[3] na corrente faircloughiana. A LC oferece uma coleção de princípios metodológicos para a pesquisa de qualquer domínio linguístico, fornecendo suporte à pesquisa da língua em uso a partir de corpus linguístico tendo como ferramentas de base a tecnologia computacional e software (Parodi, 2010). Tagnin (2005, p. 21) explica que a abordagem entende corpus como "uma coletânea de textos, necessariamente em formato eletrônico, compilados e organizados segundo critérios ditados pelo objetivo de pesquisa a que se destina". Dessa forma, o formato eletrônico possibilita o estudo dos textos para que esses sejam parcialmente analisados automaticamente propiciando a perspectiva macro e quantitativa da qual a ADC carece e pode se beneficiar.

Um dos maiores expoentes da ADC é Norman Fairclough; o autor, em inúmeras ocasiões, argumenta pelo uso de ferramentas da LC na análise discursiva. No entanto, ele não se dedica a esmiuçar o relacionamento possível entre sua teoria e a metodologia da LC. Parte de sua proposta para análise de discursos perpassa a análise de vocabulário, colocações e padrões de coocorrência de palavras – em outros termos, léxico e fraseologia. Para Fairclough (2003), um léxico especializado pode estar associado a múltiplos discursos de um certo domínio da vida social, que podem ser parcialmente diferentes – por exemplo, discursos da direita e esquerda política –, que,

[2] A tendência de aproximação entre a LC e ADC já é uma realidade para pesquisadores fora do Brasil, como podemos notar ao ler os títulos e resumos de artigos publicados em quase todos os volumes após o ano 2000 da revista Discourse & Society, o maior e mais tradicional periódico internacional sobre ADC.
[3] A abordagem dialético-relacional, também nomeada Análise de Discurso Textualmente Orientada (ADTO), segundo Magalhães (2005, p. 237), "uma característica dessa forma de fazer análise do discurso é o foco na análise detalhada dos textos como se fossem janelas a iluminar as práticas sociais".

no entanto, se sobrepõem significativamente. Nesse sentido, o autor, argumenta que, para identificar discursos distintos, devemos nos focar nas relações semânticas, nas quais podemos observar diferenças. Isto posto, Fairclough apresenta que uma maneira de chegar a essa diferença relacional é observando colocações e padrões de coocorrência de palavras em textos. Para isso, ele mesmo afirma que "a maneira mais eficaz de explicar padrões de colocação é por análise de corpus assistida por computador em grandes corpora textuais" (Fairclough, 2003, p. 131)[4].

Dessa forma, buscando estreitar os laços entre a ADC e a LC, neste estudo objetivamos apresentar dois softwares – a saber, o Sketch Engine (Kilgarriff; Rychlý, 2003) e o UAM Corpus Tool (O'Donnell, 2019) – a pesquisadores da ADC que tem pouco ou nenhum conhecimento das ferramentas de análise computadorizadas. Para isso, nos empenhamos a exemplificar possíveis caminhos metodológicos da aplicação de mencionados programas em pesquisas de análise do discurso. No próximo tópico, apresentamos brevemente alguns conceitos iniciais da LC e mostramos que a aproximação da LC com a ADC e outras pesquisas de caráter qualitativo não é inédita.

2 Conceitos iniciais

Pesquisas que lançam mão da LC como metodologia valorizam a empiricidade e a análise de fenômenos linguísticos em quantidades abrangentes de textos, necessariamente empregando ferramentas computacionais e técnicas analíticas (semi)automáticas e interativas, além de realizarem análises quantitativas e qualitativas (Berber Sardinha, 2004). A LC pode ser compreendida tanto como metodologia, quanto uma abordagem metodológica, "pois em alguns tipos de pesquisa, nomeadamente 'baseadas em corpus' ou 'informadas por corpus', ela é

[4] Todas as citações originalmente em outro idioma foram traduzidas pela autora. No original: "But the most effective way of exploring collocational patterns is through computer-assisted corpus analysis of large bodies of text" (Fairclough, 2003, p. 131).

utilizada como um instrumental para a obtenção de dados que exemplifiquem, confirmem ou refutem determinada teoria ou hipótese" (Lisboa, 2021, p. 65). Portanto, nesse caso o pesquisador "seleciona a teoria com a qual trabalhará anteriormente à análise dos dados" (Lisboa, 2021, p. 65).

Berber Sardinha (2004) explica que a LC compreende e estuda a língua a partir da abordagem empírica e, consequentemente, concebe a linguagem como sistema probabilístico, no qual nem tudo que é possível de ser realizado se concretiza efetivamente na língua. Dessa forma, "Na lingüística, empírico significa primazia aos dados provenientes da observação da linguagem, em geral, reunidos sob a forma de um corpus" (Berber Sardinha, 2004, p. 30). Essa concepção de língua aproxima a LC com a Linguística Sistêmico-Funcional (LSF) de Halliday, teoria basilar para os estudos em ADC, na corrente dialético-relacional de Fairclough. Ademais, também aproxima teoricamente a LC à Teoria da Avaliatividade (Martin; White, 2005), muitas vezes aplicada às análises discursivas críticas, derivada da LSF. Fica claro, que

> A conexão existe porque, embora de inclinação empirista, Halliday não se denomina lingüista de *corpus*. A formulação das teorias de Halliday, na forma da lingüística sistêmico-funcional, não se pauta pela exigência de um *corpus* ou do instrumental comumente empregado pelos lingüistas de *corpus*. Entretanto, a sua visão de linguagem se encaixa perfeitamente nos preceitos da Lingüística de Corpus e serve como arcabouço teórico maior no qual ela se pode incluir (Berber Sardinha, 2004, p. 34-35).

É através dessa série de aproximações teóricas, que os analistas do discurso podem e devem se apossar dos conhecimentos metodológicos e analíticos da LC.

Hardt-Mautner (1995), uma das pioneiras nas pesquisas que partem da ADC e utilizam LC, argumenta que a metodologia qualitativa usada em ADC é insuficiente para pesquisas com um corpus de grande extensão. Nesse contexto, há uma "incompatibilidade entre a estrutura escolhida e a natureza dos dados que levaram ao desenvolvimento de um procedimento analítico alternativo, combinando o uso de

programas de concordância com a análise qualitativa tradicional da ADC" (Hardt-Mautner, 1995, p. 1)[5]. Além disso, podemos defender o emprego de LC em pesquisas da ADTO, pois a LC fornece técnicas para uma "base confiável e verificável para análises quali-quantitativas, bem como por possibilitar a organização, a observação e a verificação de evidências linguísticas de maneira mais célere e precisa" (Lisboa, 2021, p. 68).

Aliás, a LC fortalece metodologicamente a pesquisa, quando consideramos a possibilidade de verificação dos dados coletados e maior clareza quanto as decisões e interpretações realizadas pelo pesquisador na condução da análise qualitativa interpretativista prezada pela ADC (Magalhães; Martins; Resende, 2017). Para mais, Lisboa (2021, p. 69) salienta que "um corpus é projetado para que reflita comportamentos da língua ou domínio que se pretenda analisar e, portanto, ele deve ser construído em conformidade com os objetivos da pesquisa para o qual foi compilado", assim, é necessário que objetivos da pesquisa e a coleta de corpus estejam alinhados.

Hardt-Mautner (1995), Mautner (2009) e Kopf (2019) explicam que o casamento entre as metodologias de LC e ADC surge da preocupação com a melhor forma de analisar os dados. Para Mautner (2009), a LC permite que a análise crítica do discurso trabalhe com maior volume de dados, quando comparada a metodologias puramente manuais. Além disso, a autora ressalta o valor da LC, que "pode ajudar a reduzir o viés do pesquisador, lidando assim com um problema ao qual a ADC dificilmente é mais propensa do que outras ciências sociais, mas para a qual recebeu críticas duras e

[5] "It was this mismatch between the chosen framework and the nature of the data that led to the development of an alternative analytical procedure, combining the use of concordance programmes with CDA's traditional qualitative analysis." (Hardt-Mautner, 1995, p. 1).

persistentes" (Mautner, 2009. p. 139)[6]. A intenção é que a LC complemente a análise aprofundada da ADC, portanto, os dados quantitativos, como análises de frequências de palavras, avaliação de ocorrências individuais de palavras no corpus, exame qualitativo de seus ambientes colocacionais, descrição de padrões semânticos e identificação das funções do discurso permitem, em contexto, insights para a análise qualitativa fina da ADC (Mautner, 2009; Kopf, 2019). A metodologia da LC, também possibilita o entrecruzamento de dados e estatísticas derivadas da análise.

Por outro lado, salientamos que o valor prático dos usos da LC na pesquisa em ADC estão fundamentalmente ligados às capacidades dos softwares de suprir as demandas geradas pelas questões de pesquisa. Mautner (2009, p. 140-141)[7] ressalta que

> se o fenômeno linguístico no qual você está interessado está de fato ligado a, ou pelo menos se cristaliza em torno de itens lexicais distintos, então é provável que você ache este método uma benção tanto como uma economia de tempo prática e eficiente quanto como uma ferramenta heurística poderosa ajudando a abrir caminhos para a descoberta. Se, por outro lado, o fenômeno a ser focalizado for aquele que se desenrola em um palco textual mais amplo e com realizações lexicais variáveis e imprevisíveis, então os métodos da LC serão de pouca ou nenhuma ajuda.

Nada impede, porém, que o analista crítico do discurso empenhe ferramentas da LC como complementares ao trabalho de análise fina de fenômenos linguísticos que se desdobram ao longo do texto. Afinal, alguns softwares, como o UAM Corpus Tool,

[6] "corpus linguistics can help reduce researcher bias, thus coping with a problem to which CDA is hardly more prone than other social sciences but for which it has come in for harsh and persistent criticism" (Mautner, 2009, p. 139).

[7] It follows that if the linguistic phenomenon you are interested in is in fact tied to, or at least crystallizes around, discrete lexical items, then you are likely to find this method a boon both as a practical and efficient time-saver, and as a powerful heuristic tool helping to clear pathways to discovery. If, on the other hand, the phenomenon to be focused on is one that is played out on a larger textual stage, and with varying and unpredictable lexical realizations, then corpus linguistic methods will be of little or no help. (Mautner, 2009, p. 140-141).

sobre o qual falaremos neste artigo, foram pensados para permitir a leitura completa dos textos e oferecer a visão do "palco textual mais amplo" de Mautner (2009).

3 Etapas metodológicas

Neste artigo, objetivos apresentar dois softwares de análise lexical e anotação para pesquisadores da ADC. Para isso, empenhamos os seguintes procedimentos metodológicos: a. seleção do material teórico sobre o qual nos apoiamos; b. escolha de dois softwares para apresentação; c. compilação de *corpus* para uma análise simplificada, buscando mostrar as ferramentas do Sketch Engine (SE); e. escolha por apresentar o UAM Corpus Tool (UAM), a partir do exemplo de seu uso na dissertação (Souza, 2022). Para trabalhar com o SE, um software pago, escolhemos a opção de testar por 30 dias sem custo, com direito ao uso de um milhão de *tokens*[8]. Já o UAM é gratuitamente disponibilizado na *web* para *download*.

3.1 Compilação e limpeza de *corpus*

Os procedimentos para trabalhar com software de análise lexical ou anotação se iniciam na coleta de textos, já que, esses realizam a leitura dos textos em formato eletrônico específico, a saber, o formato text file (txt). Esse procedimento é necessário, pois "o ideal é que os arquivos estejam no formato TXT, o que significa que contém somente caracteres do teclado (letras, números e símbolos ortográficos), sem códigos de formatação específicos para certos programas" (Berber Sardinha, 2004, p. 51).

Para a compilação de um mini *corpus* que servirá de exemplo, coletamos os primeiros seis meses de publicação do site de jornalismo *O joio e o trigo*. Seguimos os seguintes critérios: estar no recorte temporal dos primeiros seis meses – a saber de 27/10/2017 a 29/03/2018, cujo idioma fosse português brasileiro. *O joio e o trigo* é um site

[8] *Tokens* são o total de ocorrências de palavras do *corpus* (Berber Sardinha, 2009).

de jornalismo especializado em alimentação, saúde e poder. Nesta pesquisa, fizemos a coleta do site de origem diretamente no formato .txt na codificação UTF-8. Durante a etapa de compilação, também realizamos o processo de limpeza do *corpus*, que consiste na eliminação de itens irrelevantes para a pesquisa em questão, tais como, cabeçalho do site, hiperlinks, imagens, gráficos, entre outros elementos. O *corpus* de exemplo é composto por 42 textos, com 81.239 *tokens* (itens) e 10.586 *types* (formas)[9].

4 Apresentando as possibilidades

Neste tópico, apresentamos de forma resumida e concisa algumas ferramentas que compõem os softwares.

4.1 Uma breve explicação sobre análise de *corpus* com o Sketch Engine

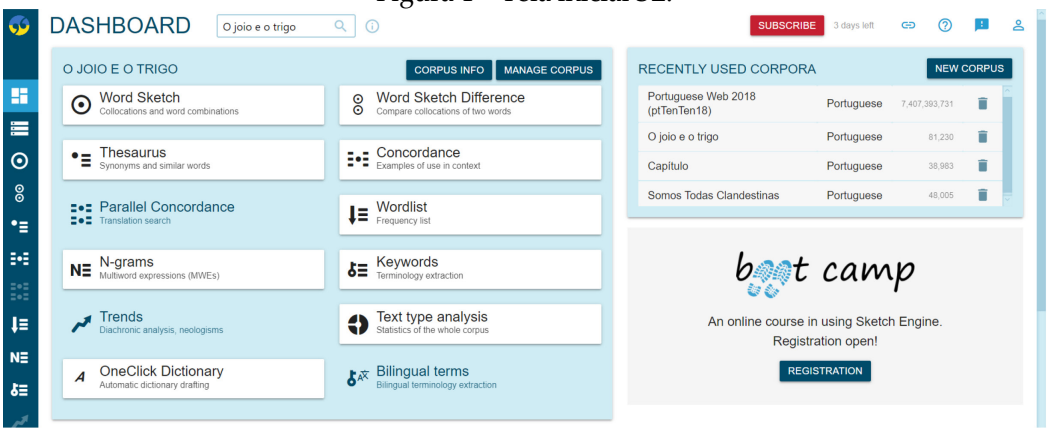
O Sketch Engine é um software criado por Adam Kilgarriff e Pavel Rychlý e desenvolvido pela Lexical Computing Ltd. O SE é um programa pago que permite a análise de textos *online*. Fromm *et al*. (2020, p. 1197) explicam que

> O objetivo dessa ferramenta é o de propiciar pesquisas, por meio de corpora, em torno do funcionamento de diversas línguas. Para atingir tal propósito, há vários recursos, como o tesauro, que encontra palavras com significados semelhantes ou que aparecem em contextos similares, as listas de frequências e a compilação e gestão de corpora.

A Figura 1, a seguir, apresenta a tela principal do SE. Na Figura 1, podemos ver as ferramentas mais proeminentes no SE, a saber: Word Sketch, Concordance, Wordlist e Keywords.

[9] *Types* são o total de palavras dos arquivos sem considerar as repetições (BERBER SARDINHA, 2009).

Figura 1 – Tela inicial SE.



Fonte: *print* da autora.

Na Figura 2 apresentamos a ferramenta Wordlist do SE. A Wordlist é usada para gerar lista de palavras por frequência. Essa ferramenta é acionada clicando no botão Wordlist, apresentado na Figura 1. Na Figura 2, vemos que na coluna *Word* as palavras são ordenadas por ondem por frequência, com a quantidade de ocorrências indicada na coluna *Frequency*. Na lista gerada como exemplo abaixo, aplicamos uma lista de exclusão de palavras, para que essas não fossem contabilizadas pelo SE. A intenção com o uso da lista de exclusão de palavras foi excluir palavras gramaticais, para privilegiar a observação de palavras lexicais[10].

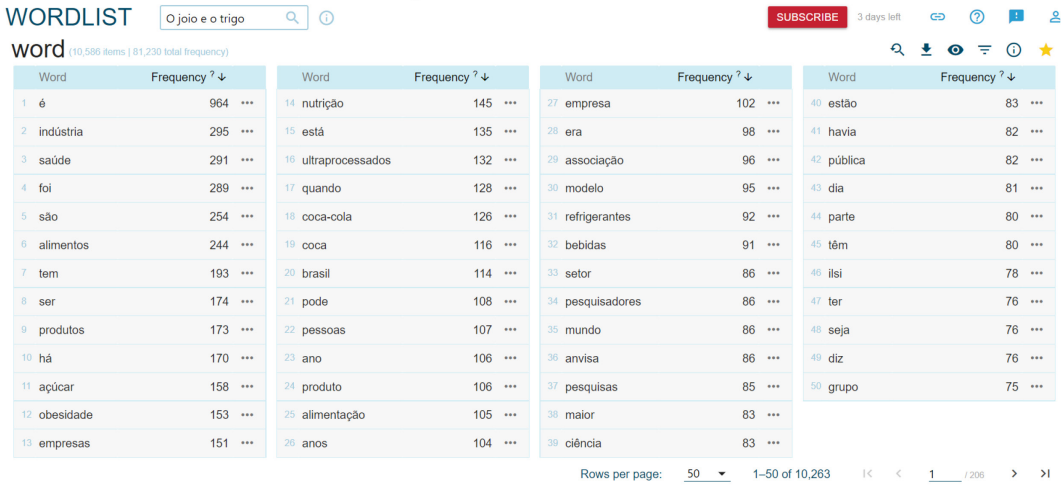
A Wordlist é uma ferramenta que pode ser bastante útil aos analistas críticos do discurso, uma vez que através da análise de seus resultados podemos observar as palavras de maior frequência e já começar a esboçar aquilo que é mais recorrente e, em certo grau, mais importante no *corpus* analisado. No exemplo apresentado na Figura 2, percebemos que os nomes[11] **indústria**, **saúde**, **alimentos**, **produtos**, **açúcar**, **obesidade** e **empresas** figuram entre as mais frequentes nas reportagens compiladas,

[10] No texto *Como as palavras se organizam*, Maria Helena de Moura Neves explica a diferença entre as palavras lexicais, "aquelas que trazem em si alguma representação do mundo (real ou fantasiado), um valor não apenas gramatical", e as palavras gramaticais, que "umas são peças da organização oracional, outras são peças definidas na semântica textual e na organização interacional" (Neves, [s.d], p. 18).

[11] Adotamos neste artigo a nomenclatura da corrente funcionalista de Halliday, onde substantivos são chamados nomes, verbos são processos e grupos adverbiais são circunstâncias.

além de várias formas dos processos **ser**, **ter** e **haver**. Com esse resultado quantificável, já podemos inferir que para um estudo crítico dos discursos presentes no periódico *O joio e o trigo* seria interessante discutir as representações construídas sobre a indústria, sobre o conceito de saúde e obesidade abordado no jornal, e possivelmente as avaliações realizadas sobre esses atores sociais. Pesquisar avaliações de diferentes atores sociais com auxílio de ferramentas computacionais em estudos baseados em *corpus* já se provaram produtivas como em Viana (2007), Soares (2021) e Marra (2022).

Figura 2 – Tela Wordlist.



Fonte: *print* da autora.

A segunda ferramenta que pode ser muito útil aos analistas de discurso é a Keywords, ilustrada na Figura 3, que apresenta lista de palavras-chave. No caso do SE, o usuário pode acessar lista de palavras-chave tanto de palavras únicas, quanto de termos de multipalavras. Para gerar essas listas, é necessário que haja um *corpus* de referência, em outros programas o pesquisador deve ele mesmo coletar esse *corpus* de referência ou buscá-lo em bancos disponíveis *online*. Ao usar o SE, o pesquisador não precisará realizar essa tarefa, já que o programa tem acesso ao Portuguese Web 2018, um *corpus* geral do português brasileiro atualizado.

Figura 3 – Tela Keywords - Multi-word terms.

Fonte: *print* da autora.

Na Figura 3, notamos alguns destaques nas palavras-chave, entre elas a segunda colocada **indústria de ultraprocessados**, reafirmando a necessidade de olhar com maior detalhe para as representações e avaliações sobre a indústria no *corpus* exemplo.

Para isso, lançamos mão da ferramenta única ao SE, a Word Sketch, ilustrada na Figura 4, em seguida. A Word Sketch processa os colocados de palavras específicas e outras palavras ao redor, dessa forma,

> [...] pode ser usado como um resumo de uma página do comportamento gramatical e colocacional da palavra. Os resultados são organizados em categorias, denominadas relações gramaticais, como palavras que servem de objeto do verbo, palavras que servem de sujeito do verbo, palavras que modificam a palavra, etc. (Lexical Computing Ltd., 2022)[12].

[12] It can be used as a one-page summary of the word's grammatical and collocational behavior. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc. (LEXICAL COMPUTING LTD., 2022).

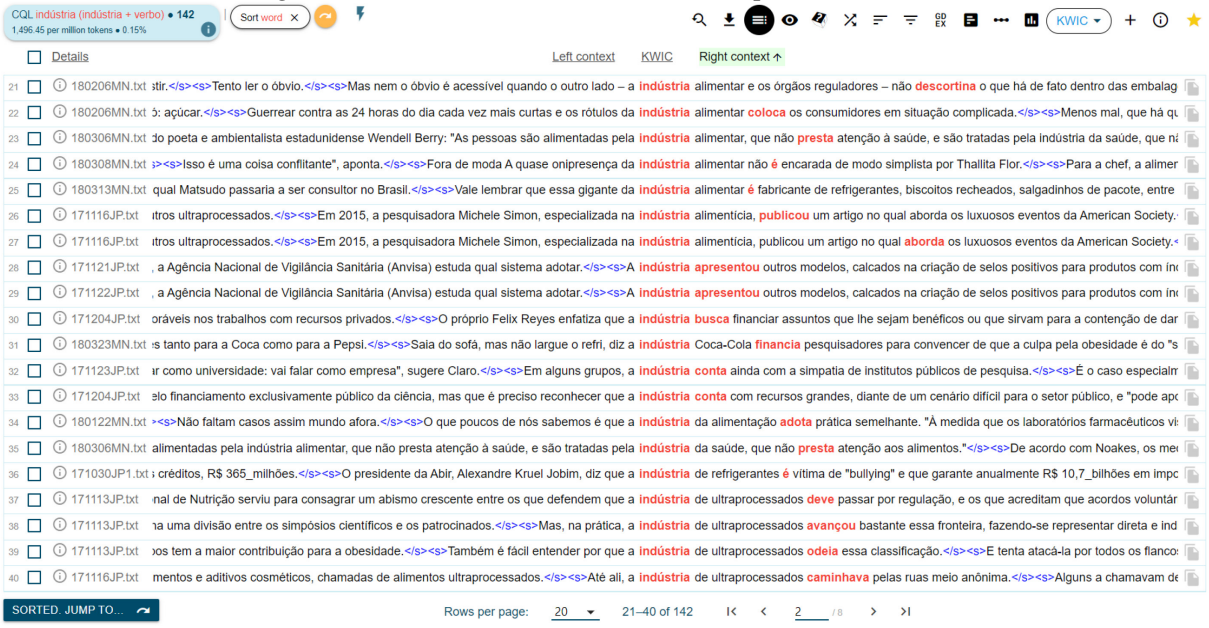
Figura 4 – Tela Word Sketch com a palavra *indústria*.

Fonte: *Print* da autora.

Na Figura 5 vemos o resultado da ferramenta Concordance com o lemma indústria. A Concordance gera linhas de concordância, que são "listagens de ocorrências de um item específico (chamado de termo de busco ou nódulo, que pode ser formado por uma ou mais palavras) acompanhado do texto aos seu redor (co-texto)" (Berber Sardinha, 2009, p. 83). Na Figura 5, vemos os resultados das linhas de concordância para indústria + verbo; a leitura dessas linhas, já esboça um começo de análise fina que pode indicar as formas como essa indústria é representada no jornal *O joio e o trigo*, e, fica explícito que tais representações são, em sua maioria, negativas. Ademais, a indústria é múltiplas vezes representada como ser de emoções e ações, que **odeia**, **caminha**, **é vítima**, entre outras realizações.

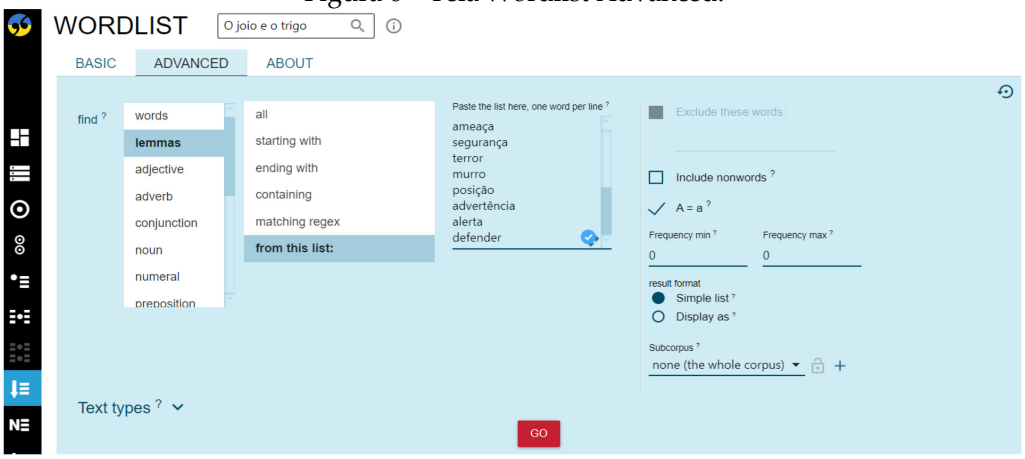
Outra possibilidade interessante e produtiva para analistas do discurso também é dentro da ferramenta Wordlist, nesse caso com a aba *Advanced* da ferramenta. Após a leitura completa da lista de palavras por frequência, exemplificadas na Figura 2, notamos uma série de nomes e processos que compõem um vocabulário relacionado à guerra. Nesse sentido, decidimos buscar a partir de lemas específicos na aba *Advanced*, mostrado na Figura 6, a saber: ataque; força; bomba; conflito; defesa; debate; pressão; lutar; luta; ameaça; segurança; terror; posição; advertência; alerta; defender; murro.

Figura 5 – Resultados Concordance com a palavra *indústria*.



Fonte: *print* da autora.

Figura 6 – Tela Wordlist Advanced.



Fonte: *print* da autora.

Na Figura 7, a seguir, vemos resultados por frequência dos lemas buscados. Esse resultado aponta que a leitura da Wordlist (Fig. 2) não só foi positiva, mas também pode resultar em uma análise específica do uso desse vocabulário de guerra em contexto.

Figura 7 - Resultados pesquisa de Wordlist Advanced.

	Lemma	Frequency
1	debate	70
2	defender	52
3	advertência	51
4	conflito	44
5	alerta	32
6	defesa	31
7	pressão	29
8	posição	26
9	ataque	20
10	segurança	15

	Lemma	Frequency
11	força	13
12	ameaça	11
13	luta	8
14	lutar	3
15	bomba	2
16	terror	2
17	murro	1

Fonte: *print* da autora.

No próximo tópico falamos sobre o UAM, um software que oferece ferramentas muito diferentes para anotação de *corpora*.

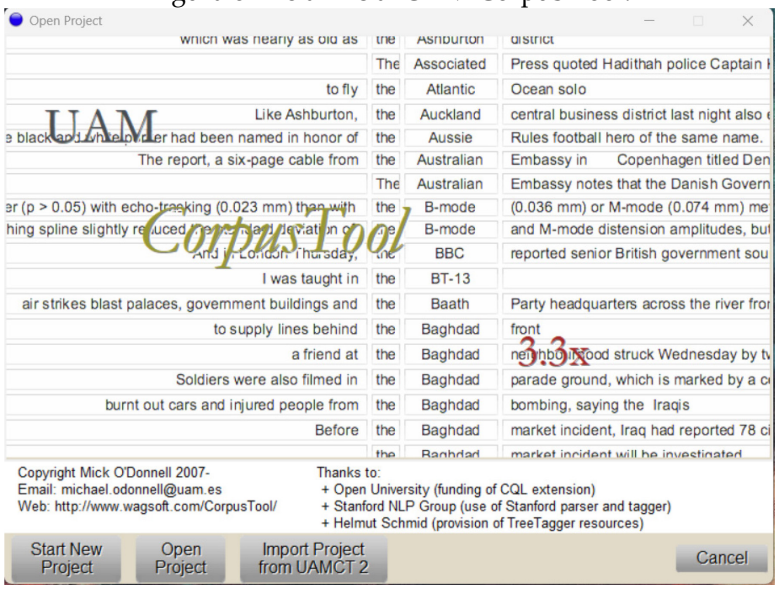
4.2 Anotações no UAM Corpus Tool

O UAM Corpus Tool é um software criado por Michael O'Donnell disponibilizado gratuitamente na internet. O programa tem como proposta auxiliar pesquisadores que já realizavam análises através de linhas de concordância em outros softwares, mas não conseguiam fazer anotações de outras características linguísticas neles. Ele é, em suma, "um sistema que permite ao usuário aplicar *tags* a segmentos de texto" (O'Donnell, 2008, p. 1434). A Figura 8 apresenta a tela inicial do UAM, na versão 3.3x.

O UAM é, antes de tudo, uma ferramenta de anotação para o pesquisador. O processo de anotação, também chamado etiquetagem, diz respeito a algo extremamente familiar para quem faz ADTO – é o momento de marcação dos textos que podemos realizar manualmente, em folhas impressas com o uso de caneta e de marcadores. A diferença entre realizar o procedimento de anotação no UAM ou realizá-lo manualmente está na facilidade de criar uma análise de diferentes categorias

em um mesmo documento e acessá-las separadamente – nomeada análise em camadas.

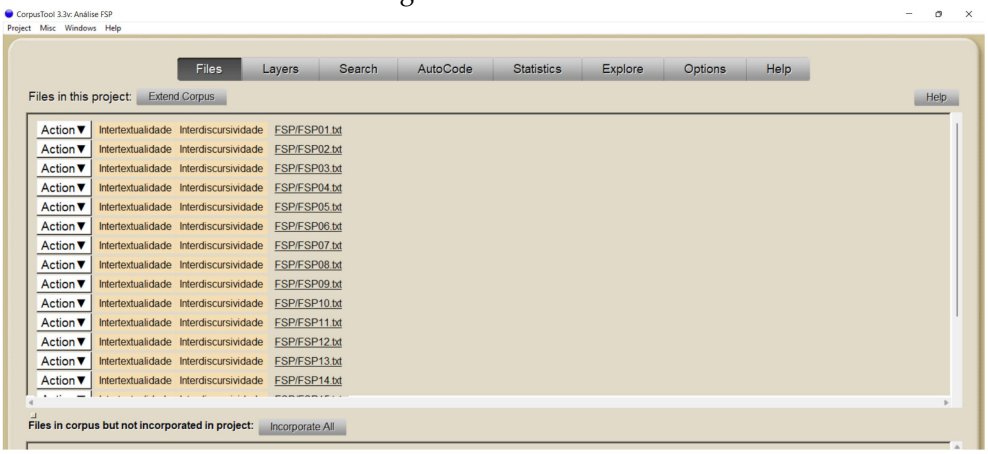
Figura 8 - Tela inicial UAM Corpus Tool.



Fonte: *print* da autora.

Em resumo, o UAM possibilita a utilização ou a criação e/ou adaptação de categorias de análise dentro dele, viabilizando a organização de todos os textos do *corpus* em uma mesma tela, a leitura desses textos e a marcação dessas categorias pelo pesquisador diretamente no software, como podemos ver na Figura 9, a seguir, que apresenta exemplo da tela *Files* da dissertação de Souza (2022). A tela principal do programa é uma interface de gestão dos textos inseridos com as camadas decididas pelo usuário. Nessa tela "o usuário pode adicionar e deletar arquivos, abrir arquivos para anotação e definir camadas de anotação para aplicar nesses arquivos" (O'Donnell, 2008, p. 1435).

Figura 9 - Tela *Files*.

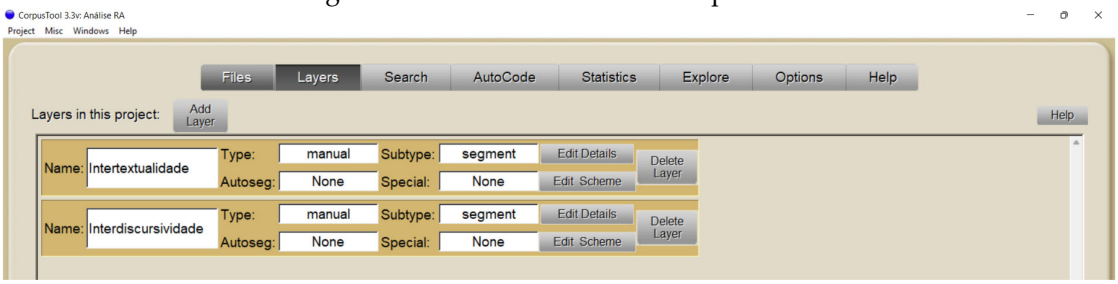


Fonte: Souza (2022).

Além disso, o software permite ao pesquisador o acesso organizado de todas as anotações. Na pesquisa de dissertação em Souza (2022), realizamos marcações sobre as vozes e textos acionados por discurso direto nas reportagens analisadas; o programa permite ao analista selecionar uma aba de pesquisa e fazer a seleção por essa categoria, previamente incluída e categorizada, para, então, apresentar-nos os resultados de todas as marcações de discurso direto realizadas. Ademais, conseguimos também gerar dados estatísticos das anotações realizadas por categorias de análise. A organização e os dados gerados pelo UAM colaboram para uma análise fina mais rica, com facilidade de acesso às marcações feitas no *corpus*.

A personalização de categorias de análise é uma das possibilidades oferecidas pelo UAM, que permite ao usuário a definição de quantas camadas (*layers*) são necessárias à sua pesquisa, além de facilitar a organização hierárquica da análise de cada camada através da intuitividade da criação das camadas. Na Figura 10, a seguir, observamos a aba *layers* do programa.

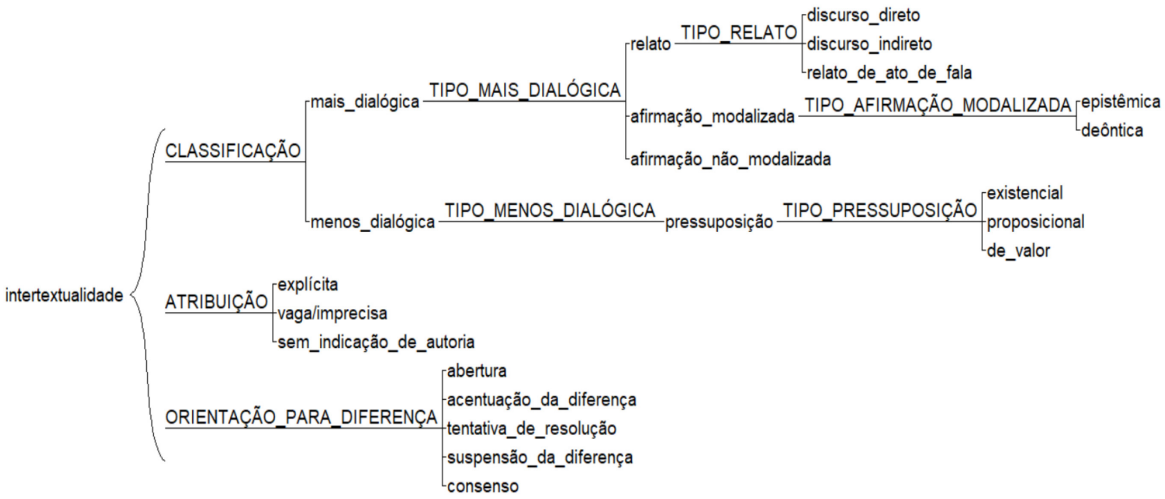
Figura 10 – Camadas no UAM Corpus Tool



Fonte: Souza (2022, p. 79).

A possibilidade de personalização de categorias de análise no UAM é uma das estrelas do software para quem faz análise crítica do discurso. Afinal, conseguimos adaptar categorias analíticas propostas por autores como Fairclough e van Leeuwen, em esquemas de análise práticos, que em ordenação lógica podem ser etiquetados no *corpus*. Souza (2022) traz como proposta metodológica a adaptação da categoria de análise intertextualidade de Fairclough (2003) em esquema de anotação no programa UAM, como podemos observar na Figura 11.

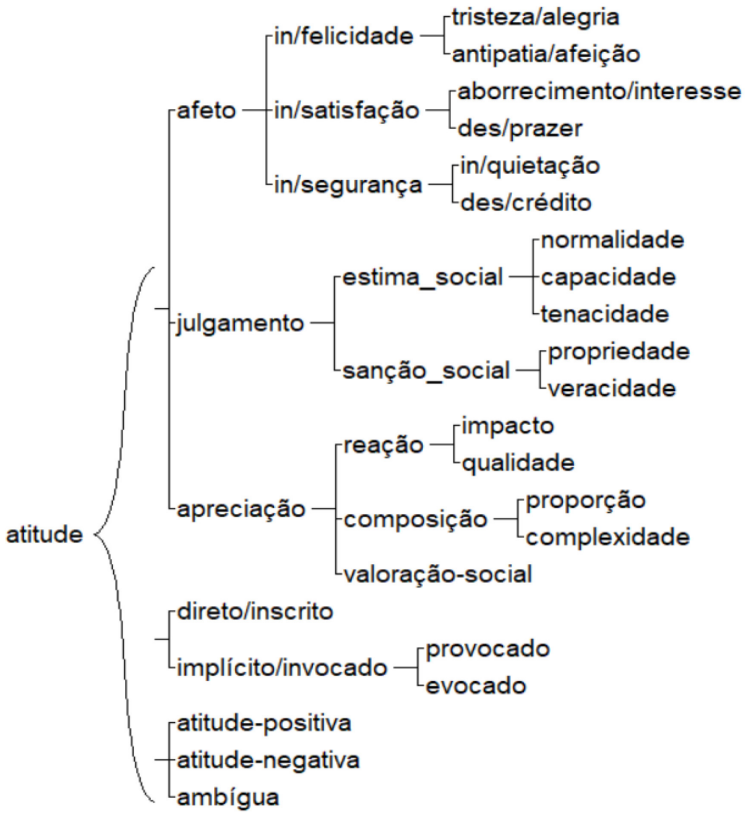
Figura 11 – Esquema analítico para análise de intertextualidade.



Fonte: Souza (2022, p. 78).

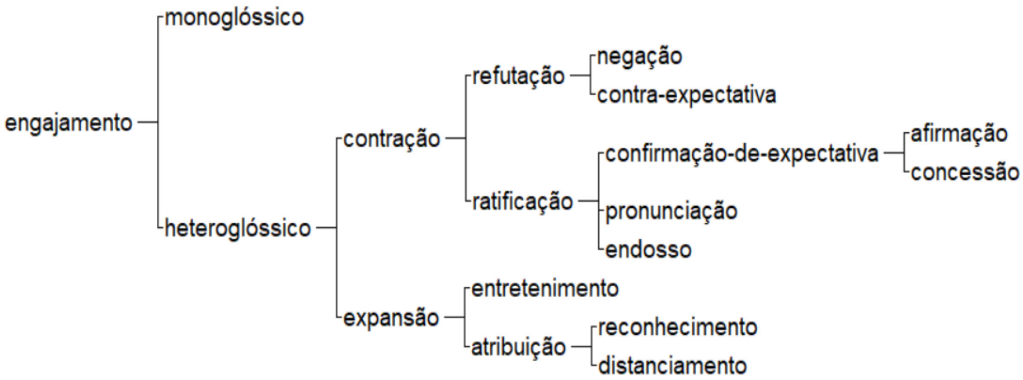
O próprio UAM apresenta em suas configurações a *layer* de análise para a Teoria da Avaliatividade (Martin; White, 2005) em inglês, que traduzimos e adaptamos para pesquisas futuras conforme mostrado nas Figuras 12, 13 e 14 a seguir.

Figura 12 – Esquema analítico para análise de Atitude da Teoria da Avaliatividade.



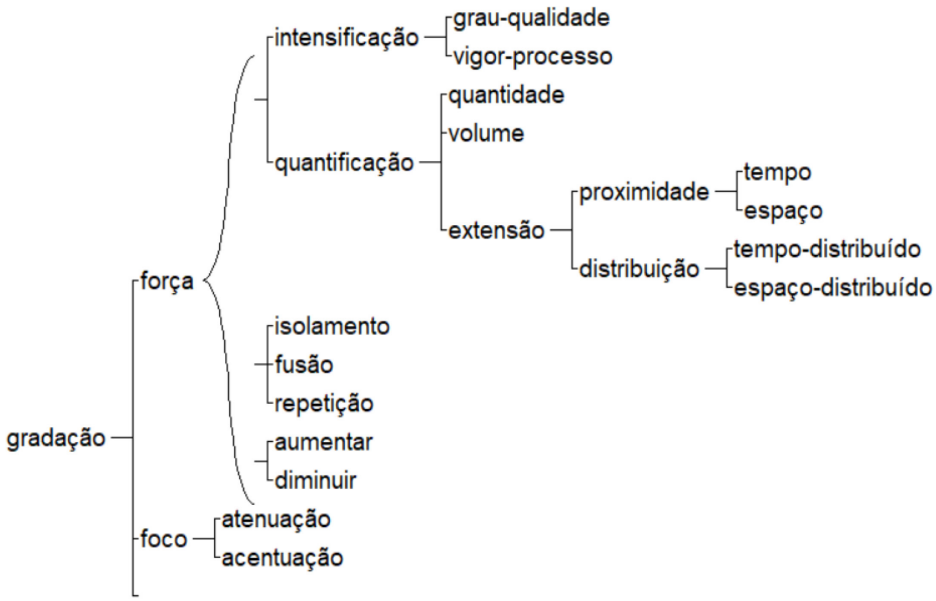
Fonte: Traduzido e adaptado no programa UAM Corpus Tool (O'Donnell, 2019), com base em Martin e White (2005).

Figura 13 – Esquema analítico para análise de Engajamento da Teoria da Avaliatividade.



Fonte: Traduzido e adaptado no programa UAM Corpus Tool (O'Donnell, 2019), com base em Martin e White (2005).

Figura 14 – Esquema analítico para análise de Gradação da Teoria da Avaliatividade.

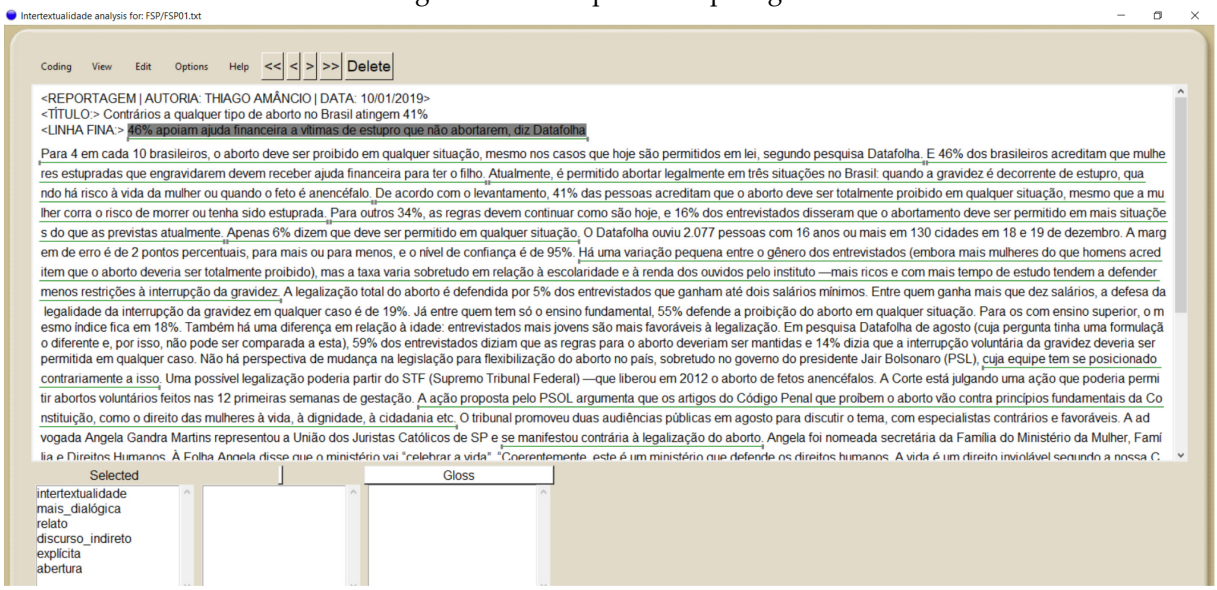


Fonte: Traduzido e adaptado no programa UAM Corpus Tool (O'Fonnell, 2019), com base em Martin e White (2005).

Em Souza (2022), também trabalhamos com *layer* para marcação da interdiscursividade, a partir dos discursos notados nas leituras das reportagens, no entanto, marcação dos discursos no programa foi realizada de forma mais livre – como é característico da ADTO, o esquema de anotação foi editado a medida em que novos discursos eram percebidos durante a análise.

A Figura 15, a seguir, ilustra como funciona o procedimento de anotação nos textos. Todos os trechos sublinhados em verde são partes nas quais identificamos algum tipo de intertextualidade, na figura selecionamos uma delas para demonstrar a etiquetagem feita, a saber: a linha fina (subtítulo) "46% apoiam ajuda financeira a vítimas de estupro que não abortarem, diz Datafolha", que classificamos seguindo o esquema analítico para intertextualidade apresentado anteriormente, que começa em tipo, nesse caso mais dialógica/relato/discurso indireto, depois para atribuição que é explícita, por último a orientação para diferença, que entendemos como de abertura.

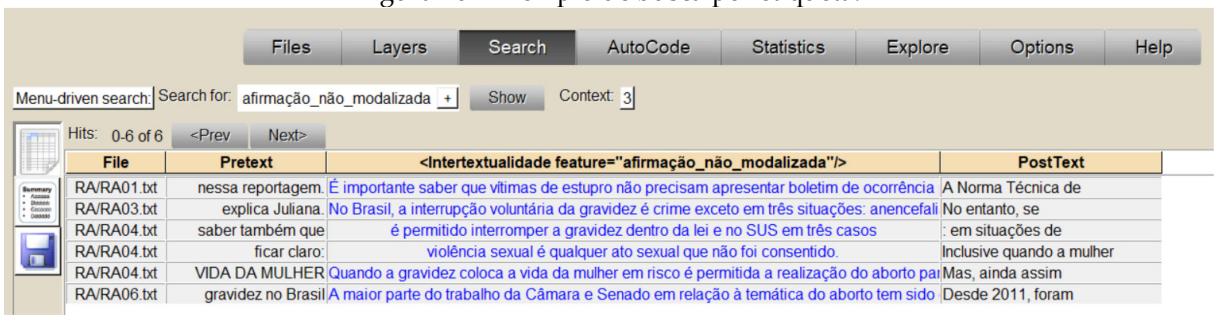
Figura 15 - Exemplo de etiquetagem.



Fonte: Souza (2022, p. 80).

Para a análise, é possível realizar buscas na aba *Search* do programa. Essa aba permite a busca específica por trechos, seja usando a função *text-driven search* para pesquisa a partir de palavras ou conjuntos específicos de palavras, ou na função nomeada *menu-driven search*. A busca *menu-driven* funciona buscando por etiqueta; por exemplo, em Souza (2022) se precisávamos acessar resultados de todas as afirmações não-modalizadas, selecionávamos essa opção. A Figura 16, a seguir, exemplifica essa aba.

Figura 16 - Exemplo de busca por etiqueta.



Fonte: Souza (2022, p. 81).

No próximo tópico fazemos um apanhado geral das reflexões de análises nas quais já empenhamos a aproximação entre LC e ADC.

5 Reflexões

Neste artigo, apresentamos brevemente dois softwares que podem ser aplicados na pesquisa em ADC. Ambos têm abordagens muito distintas sobre como podemos empenhá-los para as análises.

Por um lado, o SE permite a criação de listas de palavras, listas de palavras-chave, linhas de concordância e observação de colocações e padrões de coocorrência. O SE é um programa de análise lexical pago de uso relativamente simples, com manuais bem elaborados e suporte técnico. Uma vantagem clara diz respeito ao funcionamento do programa ser todo *online*, o que resulta em uma menor dependência de computadores específicos para executar a análise. Como a maioria dos softwares empenhados na LC, o ideal com o SE é que o analista trabalhe com grandes quantidades de textos – não muito comuns para quem faz ADC.

Já o UAM oferece como possibilidade à analistas do discurso realizarem suas marcações (etiquetagem) da análise fina de textos em suas ferramentas, incentivando seu uso mesmo com *corpora* pequenos. Como software gratuito, o UAM está atualmente em sua versão 6.2, atualizada em fevereiro de 2022, sem todas suas ferramentas. A última versão completa do UAM é 3.3x, que apresentamos neste texto. Um ponto negativo do programa é que seu último manual publicado em inglês é sobre a versão 2.8, e em português pt sobre a versão 2.6.

Uma desvantagem clara para aqueles que querem usar esses softwares diz respeito à língua. Ambos os programas estão em língua inglesa e demandam conhecimento dela para serem usados, para mais o acesso aos seus manuais também depende desse conhecimento.

Os argumentos a favor da aproximação entre a ADC e a LC são múltiplos (Baker, 2006; Baker *et al.*, 2008; Mautner, 2009, 2022) e deixam claro que é uma relação que beneficia os analistas de discurso enormemente contra as críticas que normalmente enfrentam quanto ao viés dos pesquisadores influenciarem a pesquisa, quanto a verificação e quantificação de dados e triangulação dos dados pesquisados.

Referências

BAKER, P. **Using Corpora in Discourse Analysis**. London: Continuum, 2006. DOI https://doi.org/10.5040/9781350933996

BAKER, P. *et al*. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. **Discourse & Society**, v. 19, p. 273–306. DOI https://doi.org/10.1177/0957926508088962

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

BERBER SARDINHA, T. **Pesquisa em Lingüística de Corpus com WordSmith Tools**. Campinas: Mercado das Letras, 2009. 272 p.

FAIRCLOUGH, N. **Analysing Discourse**: textual analysis for social research. London: Routledge, 2003. DOI https://doi.org/10.4324/9780203697078

FROMM, G. *et al*. Wordsmith Tools e Sketch Engine: um estudo analítico-comparativo para pesquisas científicas com uso de corpora. **Revista de Estudos da Linguagem**, [S.L.], v. 28, n. 3, p. 1191-1248, 27 maio, 2020. DOI https://doi.org/10.17851/2237-2083.28.3.1191-1248

HARDT-MAUTNER, G. "**Only Connect**": Critical Discourse Analysis and Corpus Linguistics. Lancaster: UCREL, 1995.

KILGARRIFF, A.; RYCHLÝ, Pl. **Sketch Engine**. East Sussex: Lexical Computing Limited, 2003. Disponível em: http://www.sketchengine.eu. Acesso em: 10 dez. 2022.

KOPH, S. Corpus-Assisted Critical Discourse Studies? Marrying Critical Discourse Studies and Corpus Linguistics: Über den Brückenschlag zwischen Kritischen

Diskursstudien und Korpuslinguistik. **Diskurse**, Mannheim, v. 1, n. 3, p. 92-110, 2019. Disponível em : https://majournals.bib.uni-mannheim.de/diskurse-digital/article/view/99. Acesso em: 13 dez. 2022.

LEXICAL COMPUTING LTD. **Word Sketch**: collocations and word combinations. 2022. Disponível em: https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/#top. Acesso em: 16 dez. 2022.

LISBOA, J. V. R. **Proposta de harmonização da terminologia designativa de área e subáreas do português como língua não materna baseada em corpus**. 2021. 215f. Dissertação (Mestrado) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Federal de Uberlândia, Uberlândia, 2021. Acesso em: 31 ago. 2022.

MAGALHÃES, I.; MARTINS, A. R.; RESENDE, V. M. **Análise de Discurso Crítica**: um método de pesquisa qualitativa. Brasília: Editora da Universidade de Brasília, 2017. 230 p. DOI https://doi.org/10.7476/9788523013370

MARRA, M. N. A. Avaliatividade em corpus de comentários: um olhar sobre o feminino. *In*: NOVODVORSKI, A.; LISBOA, J. V. R.; CARNEIRO, R. M. O. **Estudos exploratórios em linguística de corpus 2**. Araraquara: Letraria, 2022. p. 79-94. Disponível em: https://www.letraria.net/estudos-exploratorios-em-linguistica-de-corpus-ii/. Acesso em: 18 dez. 2022.

MARTIN, J. R.; WHITE, P. R. R. **The language of evaluation** - appraisal in English. Londres: Palgrave/Macmillan, 2005.

MAUTNER, G. Checks and balances: how corpus linguistics can contribute to CDA. *In*: WODAK, R.; MEYER, M. **Methods of Critical Discourse Analysis**. 2 ed. London: Sage, 2009. p. 138-160.

NEVES, M. H. M. **Como as palavras se organizam**. Disponível em: https://www.museudalinguaportuguesa.org.br/wp-content/uploads/2017/09/Como-as-palavras-se-organizam-em-classes.pdf. Acesso em: 11 dez. 2022.

O'DONNELL, M. **UAM Corpus Tool**. Versão 3.3, 2019. Disponível em: http://www.corpustool.com/download.html. Acesso em: 10 jan. 2020.

O'DONNELL, M. The UAM CorpusTool: Software for corpus annotation and exploration. **Proceedings of the XXVI Congreso de AESLA**, Almeria, Spain, 2008.

PARODI, G. ¿Qué es la lingüística de corpus? (Re)Surgimiento, definiciones y antecedentes. *In*: PARODI, G. **Lingüística de Corpus**: de la teoría a la empiria. Madrid: Iberoamericana, 2010. p. 14-35. DOI https://doi.org/10.31819/9783865278715

SOARES, L. C. O afeto em depoimentos sobre adoção: uma análise da avaliatividade com subsídios da Linguística de Corpus. *In*: NOVODVORSKI, A.; LISBOA, J. V. R.; CARNEIRO, R. M. O. **Estudos exploratórios em linguística de corpus**. Araraquara: Letraria, 2021. p. 61-70. Disponível em: https://www.letraria.net/estudos-exploratorios-em-linguistica-de-corpus/. Acesso em: 18 dez. 2022.

SOUZA, B. M. G. **Vozes hegemônicas e vozes insurgentes**: uma análise discursiva crítica sobre a representação do aborto na mídia. 2002. 173f. Dissertação (Mestrado) - Programa de Pós-graduação em Estudos Linguísticos, Universidade Federal de Uberlândia, Uberlândia, 2022. DOI:

TAGNIN, S. E. O. **O jeito que a gente diz**: expressões convencionais e idiomáticas. São Paulo: Disal, 2005. 117p.

VIANA, V. Utilizando o programa Wordsmith Tools na pesquisa sobre apreciação: uma sugestão metodológica. **Revista Intercâmbio**, v. 16, p. 1-21, 2007. São Paulo: LAEL/PUC-SP. Disponível em: https://revistas.pucsp.br/index.php/intercambio/article/view/3611. Acesso em: 07 dez. 2022.