ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

Procedimentos metodológicos para identificação e análise de unidades fraseológicas em *corpora* de seriados televisivos

Joel Victor Reis Lisboa[1]

Resumo: Haja vista o crescente número de pesquisas no Brasil voltadas ao léxico dos seriados televisivos (ROCHA; ORENHA-OTTAIANO, 2012; BANG; FROMM, 2013; PEIXOTO, 2014; MURAD, 2015; YAMAMOTO; LISBOA, 2019, dentre outros), o presente artigo tem como objetivo descrever os procedimentos metodológicos realizados no escopo de uma pesquisa voltada à identificação e análise de unidades fraseológicas (UFs) no seriado televisivo *Game of Thrones* (GOT). Portanto, serão apresentados os princípios e critérios de compilação de *corpora* à luz da metodologia/abordagem da Linguística de *Corpus* (LC), os procedimentos de seleção de arquivos, compilação, limpeza, conversão e padronização do *corpus* de estudo da referida pesquisa, bem como as ferramentas utilizadas e procedimentos realizados para identificação e análise de UFs em GOT. Esperamos demonstrar as possibilidades de análises lexicais em legendas de seriados televisivos por meio da metodologia/abordagem da LC, evidenciar a proficuidade do mundo ficcional desses seriados para a realização de estudos lexicais e, por fim, contribuir para embasar e despertar interesse em novas pesquisas terminográficas e fraseológicas voltadas ao mundo ficcional de seriados televisivos.

Palavras-chave: Linguística de *Corpus*. Unidades Fraseológicas. Seriados televisivos.

Methodological procedures for identification and analysis of phraseological units in *corpora* of TV series subtitles

Abstract: Considering the growing number of Brazilian researches focused on TV series' lexicon (ROCHA; ORENHA-OTTAIANO, 2012; BANG; FROMM, 2013; PEIXOTO, 2014; MURAD, 2015; YAMAMOTO; LISBOA, 2019, and others), this paper aims to describe methodological procedures used in the scope of a research focused on the identification and analysis of phraseological units (PUs) in the TV series Game of Thrones (GOT). Therefore, the principles and criteria of *corpora* compilation based on *Corpus* Linguistics (CL), the procedures of data selection, compilation, cleaning, file conversion and standardization of this research's study *corpus* as well as the tools and strategies used for identifying and analyzing PUs in GOT will be presented. We hope to demonstrate the possibilities of lexical analysis on TV series subtitles through the CL methodology/approach, highlight the advantages that fictional worlds of TV series provide for lexical studies and, finally, contribute to base and arouse interest on new terminographical and phraseological researches focused on TV series' fictional worlds.

Keywords: *Corpus* Linguistics. Phraseological Units. TV series.

1 INTRODUÇÃO

O número de pesquisas voltadas à análise lexical do mundo ficcional dos seriados televisivos vem crescendo exponencialmente no Brasil. Pesquisadores como Rocha & Orenha-

[1] Mestrando no Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Uberlândia. E-mail: joelvictorlisboa@gmail.com.

Ottaiano (2012), Bang & Fromm (2013), Peixoto (2014), Murad (2015), Yamamoto & Lisboa (2019), dentre outros, realizaram estudos no âmbito da Terminografia e/ou Fraseologia utilizando *corpora* de legendas de seriados como *House M. D.*, *Law and Order* e *Star Trek*.

Dentre as aplicabilidades e resultados dessas pesquisas, estão: (i) o treinamento de novos pesquisadores para o desenvolvimento de pesquisas em análise e descrição lexical; (ii) a divulgação da Terminologia e Fraseologia por meio dos estudos lexicais em seriados televisivos, haja vista a ampla popularidade desses seriados; (iii) a corroboração da existência de vocabulário próprio do mundo ficcional e sua influência na língua corrente; (iv) a produção de glossários/vocabulários bilíngues, disponibilizados gratuitamente *online*, que podem ser utilizados no âmbito do ensino de línguas e da tradução, dentre outras contribuições.

A facilidade e rapidez na compilação de *corpora* devido a existência de diversos repositórios de legendas *online*, a facilidade do formato em que os arquivos de legenda são disponibilizados para processamento por programas e/ou ferramentas de análise lexical e a ampla popularidade e interesse pelo mundo ficcional dos seriados televisivos são algumas das vantagens da exploração do léxico dos seriados por meio da análise de *corpora* de legendas.

É relevante evidenciar que, como o propósito da pesquisa apresentada nesse artigo, assim como das pesquisas mencionadas anteriormente, não é averiguar a qualidade dos procedimentos de tradução e/ou legendagem, mas utilizar o material linguístico disponibilizado nas legendas para a realização de pesquisas lexicais, as legendas dos seriados televisivos, apesar de serem, em geral, produzidas por equipes de legendagem não profissionais, consistem em um objeto de estudo proveitoso para as pesquisas em questão.

Isto posto, objetivamos apresentar os procedimentos metodológicos realizados no âmbito de uma pesquisa voltada à identificação e análise de unidades fraseológicas[2] (doravante UFs) em um *corpus* bilíngue (inglês e português) de legendas do seriado televisivo *Game of Thrones*. Por meio da descrição dos procedimentos realizados, visamos demonstrar as possibilidades de análises lexicais mediante a exploração de *corpora* legendas de seriados televisivos, evidenciar a proficuidade do mundo ficcional desses seriados para a realização de estudos lexicais e, por conseguinte, contribuir para embasar e despertar interesse em novas pesquisas lexicais voltadas ao mundo ficcional dos seriados televisivos.

[2] Concebemos unidades fraseológicas como padrões de coocorrência léxico-sintática convencionalizados, mais ou menos fixos, e que podem apresentar graus de idiomaticidade (TAGNIN, 2013).

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

As seguintes seções estão estruturadas da seguinte maneira: primeiramente introduziremos brevemente a Linguística de *Corpus*, metodologia/abordagem utilizada no escopo da referida pesquisa e, em seguida, apresentaremos a Fraseologia, o campo de estudos no qual essa pesquisa está inserida. Posteriormente, os procedimentos de seleção dos arquivos de legenda, compilação, limpeza, conversão e padronização do *corpus* de estudo serão apresentados, bem como os procedimentos de identificação e análise das UFs.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Linguística de *Corpus*

A Linguística de *Corpus* (doravante LC) é uma metodologia de pesquisa de base empírica e estatística, que se ocupa da compilação e análise de dados linguísticos autênticos, armazenados em formato eletrônico, por meio de ferramentas computacionais (BERBER SARDINHA, 2004). As pesquisas desenvolvidas com base na LC compartilham algumas características, como a criteriosidade na metodologia de coleta e arquitetura dos *corpora*, a exploração linguística com auxílio computacional e, principalmente, a primazia da empiricidade nas análises linguísticas (BERBER SARDINHA, 2004).

No âmbito da LC, um *corpus* é concebido como uma coleção de textos ou porções textuais (originalmente escritos ou transcrições de fala), legíveis por computador, criteriosamente compilada e estruturada de modo que sirva de amostragem de linguagem autêntica para fins de estudos linguísticos (BERBER SARDINHA, 2004).

Segundo Berber Sardinha (2004), Aluísio & Almeida (2006) e Parodi (2010), dentre os principais requisitos para compilação de *corpora* no âmbito da LC, estão:

(i) Autenticidade: os textos ou porções textuais que compõem os *corpora* devem ser produzidos por seres humanos em situações reais de comunicação, excluindo-se, portanto, linguagens de programação, expressões matemáticas, e textos criados com propósitos pedagógicos ou de pesquisa linguística.

(ii) Criteriosidade: os *corpora* devem ser compilados e estruturados criteriosamente para servirem para fins específicos, conforme à demanda das pesquisas para os quais foram compilados.

(iii) Representatividade: para servirem como amostras da língua, variação ou campo de conhecimento que se pretende analisar, os *corpora* devem ser representativos em termos de extensão e dos gêneros que os compõem.

(iv) Formato eletrônico: de modo a realizar análises de grandes quantidades de textos de maneira mais rápida e precisa, os *corpora* devem estar em formato legíveis por programas e ferramentas de análise lexical, pois utiliza-se meios computacionais para extrair e organizar os dados linguísticos.

É relevante enfatizar que nem todos os *corpora* passam pelos mesmos procedimentos de compilação e arquitetura e nem compartilham de todas as características, pois cada *corpus* é compilado para determinado propósito conforme os objetivos da pesquisa. Na pesquisa descrita neste artigo, por exemplo, haja vista que o objetivo é explorar UFs recorrentes no mundo ficcional do seriado *Game of Thrones*, nosso *corpus* é formado por arquivos de legenda criteriosamente compilados, limpos, padronizados e convertidos em formato legível por programas e ferramentas de análise lexical. Outrossim, apesar de ser um *corpus* menor do que os *corpora* compilados para outros tipos de pesquisa, ele é o mais representativo possível para a pesquisa em questão, pois é composto por todos os episódios do seriado que pretende-se analisar.

A LC concebe a língua como um sistema probabilístico em que nem tudo o que é teoricamente possível no sistema linguístico tem a mesma probabilidade de acontecer (BERBER SARDINHA, 2004). Ademais, compreende a variação de frequência em função dos contextos como não aleatória, ou seja, a língua é padronizada (SINCLAIR, 1991; BERBER SARDINHA, 2004) e a observação de padrões é impulsionada pela análise de *corpora* por meio de ferramentas computacionais que, em geral, utilizam técnicas estatísticas para identificação desses padrões.

Isto posto, e considerando que UFs são, em suma, padrões de coocorrência léxico-sintática, a LC consiste em uma metodologia fundamental para pesquisas fraseológicas, pois lança mão de programas e ferramentas de base estatística que extraem e apresentam padrões linguísticos oriundos de *corpora* de grandes quantidades de textos, cuja precisão e rapidez seriam impossíveis de serem alcançadas manualmente. Por esse motivo, na presente pesquisa seguiu-se princípios e critérios da LC para a realização dos procedimentos desde a compilação do *corpus* até a análise das UFs.

Na próxima subseção, apresentamos brevemente a disciplina que se ocupa da análise e descrição das UFs, a Fraseologia.

2.2 Fraseologia

Fraseologia pode ser concebida como termo hiperônimo que abarca os variados tipos de UFs ou como disciplina linguística que se ocupa da análise e descrição das UFs. A concepção de Fraseologia como disciplina independente da Lexicologia ainda não é unânime entre os linguistas, não obstante, consideramos tal discordância como percurso natural para subdisciplinas em processo de independência. Acreditamos que, a depender de seu desenvolvimento teórico-metodológico, a Fraseologia virá a se tornar uma disciplina independente, como ocorreu com a Lexicologia e Terminologia, por exemplo (YAMAMOTO; LISBOA, 2019).

Segundo Monteiro-Plantin (2014), estabelecer os limites da Fraseologia é uma tarefa complexa, pois não há consenso entre os linguistas no que tange à delimitação e categorização dos agrupamentos léxico-sintáticos passíveis de serem considerados UFs. Todavia, as UFs compartilham de algumas características que, de certa forma, possibilitam seu agrupamento sob um mesmo conjunto de objetos de estudo. Em suma, de acordo com Corpas Pastor (1996), Tagnin (2013) e Monteiro-Plantin (2014), dentre as características estão:

(i) Polilexicalidade: as UFs são constituídas por no mínimo dois elementos armazenados na memória e utilizados como uma só unidade.

(ii) Convencionalidade: diz respeito ao que é consolidado pelo uso ou prática linguística dos falantes, à nível sintático, semântico e/ou pragmático, e está estreitamente ligada à frequência de utilização das UFs e coocorrência de elementos constituintes. A convencionalidade e frequência estão comumente vinculadas a determinados meios de (re)produção e manutenção de UFs, como a bíblia, a mitologia, a literatura e os meios de comunicação em massa.

(iii) Fixidez: concerne à estabilidade formal das UFs e, em geral, é concebida como um *continuum*, ou seja, há UFs totalmente fixas e outras que permitem a substituição de constituintes, contudo, podendo ocorrer alterações semânticas. A fixidez está intimamente ligada à idiomaticidade, pois, via de regra, UFs idiomáticas serão eventualmente fixas, ou seja, convencionalizadas sintática e semanticamente.

(iv) Idiomaticidade: corresponde à não composicionalidade, não transparência, convencionalidade ou opacidade semântica. Em outras palavras, diz respeito à impossibilidade de apreender o sentido da UF como unidade a partir da soma dos significados de cada um de seus constituintes. Conforme o posicionamento teórico, a idiomaticidade é vista como uma escala gradual.

É necessário considerar integralmente as características apresentadas na determinação de um agrupamento lexical como UF, pois elas são interrelacionadas. Ademais, é relevante pontuar que nem todas as UFs compartilham igualmente de todas as características mencionadas, portanto, a depender da característica eleita como determinante, as concepções de UF variarão naturalmente, o que justifica a divergência tangentes à delimitação e categorização das UFs (MONTEIRO-PLANTIN, 2014).

Dentre os agrupamentos léxico-sintáticos que compartilham, em certa medida, das características evidenciadas e, portanto, são considerados UFs pelos estudos fraseológicos estão as colocações, expressões idiomáticas, frases feitas, locuções fixas, parêmias, pragmatemas, dentre outros[3]. Ademais, Monteiro-Plantin (2014, p. 65) destaca outras formas linguísticas cuja classificação é complexa, pois compartilham relativamente das características das UFs "sem poderem ser incluídas integralmente na categoria, e nem totalmente descartadas", que são os estereótipos, clichês, bordões e *slogans*.

Apresentada a metodologia/abordagem utilizada e o campo de estudos no qual a referida pesquisa se insere, na seguinte seção os procedimentos metodológicos de seleção de arquivos, compilação, limpeza, conversão e padronização do *corpus* de estudo serão apresentados.

3 PROCEDIMENTOS METODOLÓGICOS DE COMPILAÇÃO E PREPARAÇÃO DO *CORPUS* DE ESTUDO

O *corpus* de legendas de *Game of Thrones* é constituído por 146 arquivos e conta com o total de 1.228.428 *tokens*[4] e 24.361 *types*[5]. Esse *corpus* é composto por dois *subcorpora*,

[3] Ver mais em Corpas Pastor (1996), Tagnin (2013), Monteiro-Plantin (2014), dentre outros.
[4] Número total de palavras no *corpus*.
[5] Número total de palavras diferentes no *corpus*, independentemente do número de ocorrências de cada palavra.

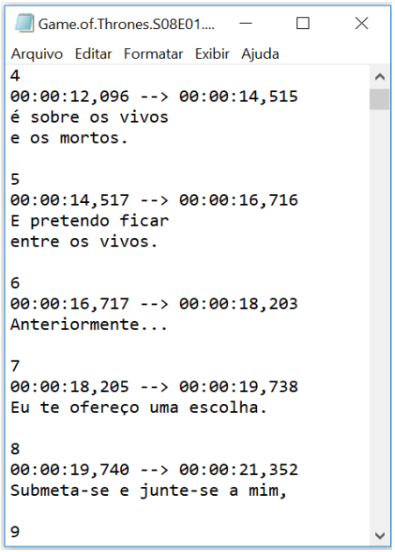
ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

um em língua inglesa e outro em língua portuguesa[6]. Cada *subcorpus* é constituído por 73 arquivos de legenda devidamente limpos e padronizados, sendo cada arquivo referente a um episódio do seriado televisivo. O *subcorpus* em língua inglesa conta com 638.786 *tokens* e 9.714 *types*, ao passo que o *subcorpus* em língua portuguesa conta com 589.642 *tokens* e 15.187 *types*.

3.1 Seleção dos arquivos e compilação do *corpus*

O *corpus* foi compilado a partir de repositórios que disponibilizam legendas de seriados televisivos gratuitamente *online*. Elegeu-se como critério de compilação dos arquivos a seleção de no mínimo duas fontes de legendagem diferentes, de modo que fosse possível fazer um contraste prévio entre os arquivos de legenda, objetivando delimitar com maior exatidão o início e fim de cada episódio. Esse critério foi utilizado, pois, em geral, os primeiros e últimos minutos de cada episódio são retrospectivas dos episódios anteriores e apresentação do que virá a ocorrer em episódios futuros do seriado, como é possível verificar na Figura 1 a seguir:

Figura 1 – Visão parcial do arquivo de legenda do primeiro episódio da oitava temporada de *Game of Thrones* extraído do *site* Legendas.TV



Fonte: elaborada pelo autor

No arquivo de legenda apresentado na Figura 1, a legendagem é iniciada desde o primeiro segundo do episódio em questão. Entretanto, observamos que as transcrições das

[6] Neste artigo utilizamos língua portuguesa ou português para nos referirmos especificamente ao português de variedade brasileira, haja vista que essa foi a variedade das legendas pesquisadas.

porções de fala dos personagens realmente só são iniciadas aos 6 minutos e 22 segundos. Dessarte, todos os diálogos apresentados anteriormente à essa marcação de tempo são, na verdade, uma retrospectiva das últimas temporadas do seriado.

Só foi possível identificar o início e o fim exato deste episódio a partir da análise contrastiva entre diferentes arquivos de legenda do mesmo episódio, entre o arquivo em português e em inglês, bem como, principalmente, pela redobrada atenção à marcação de tempo das legendas. Quando a comparação entre os arquivos não foi suficiente para definir o início e/ou fim de determinado episódio, a ferramenta Concord do programa de análise lexical WordSmith Tools 6.0 (SCOTT, 2012) foi utilizada para verificar se alguma porção de fala inicial ou final se repetiu em episódios anteriores e, a partir disso, confirmar se a cena em questão realmente faz parte do episódio sob análise. A ferramenta Concord será apresentada posteriormente neste artigo.

Haja vista que um dos meios de identificação de UFs no *corpus* é segundo o critério de frequência de ocorrência, se determinada UF for utilizada por um personagem e a cena em que foi utilizada se repetir em dois ou mais episódios, os resultados estarão comprometidos. Apesar de a ferramenta Concord possibilitar a identificação do arquivo no qual a palavra de busca foi utilizada, seria demasiadamente exaustivo voltar e realizar novamente a limpeza de cada arquivo comprometido. O Concord também permite a remoção de linhas de concordância duplicadas, entretanto, apesar de ser possível realizar a identificação e análise das UFs na referida pesquisa sem maiores problemas, a utilidade do *corpus* para futuras pesquisas estaria inviabilizada e seria impossível mensurar sua quantidade exata de *tokens*.

Isto posto, o repositório *online* utilizado para *download* das legendas em inglês foi o TVsubtitles[7], por três principais razões: 1) pela disponibilização de arquivos de legenda de todos os episódios e temporadas de *Game of* Thrones, 2) pela possibilidade de *download* em massa das legendas de todos os episódios de cada temporada por meio de arquivos compactados e 3) pela disponibilização de legendas produzidas por diferentes equipes de legendagem no mesmo arquivo compactado.

No repositório em questão, há a possibilidade de *download* de episódios separadamente ou em conjunto (*all episodes*) em até nove idiomas. Devido ao fato de nem todos os episódios possuírem legendas em língua portuguesa de variedade brasileira, foi feito o

[7] Disponível em: http://www.tvsubtitles.net/. Acesso em: 23 mar. 2020.

download em massa das legendas em língua inglesa. Cada arquivo compactado é constituído por legendas produzidas por no mínimo duas equipes de legendagem, fator que vai de encontro aos critérios estabelecidos para realizar a delimitação do início e fim de cada episódio de maneira mais exata, assim como a seleção do arquivo de legenda mais limpo, facilitando o procedimento de limpeza.

As legendas em português foram compiladas no repositório *online* Legendas.TV[8] devido à disponibilização de legendas em português de todos os episódios e temporadas de *Game of Thrones*, sendo possível também fazer o *download* em massa e obter todos os episódios de cada temporada em arquivos compactados. A única diferença em relação à compilação do *subcorpus* de legendas em inglês se deu devido à não disponibilização de legendas produzidas por diferentes equipes de legendagem no mesmo arquivo compactado, o que resultou em *downloads* de arquivos separadamente, entretanto, oriundos do mesmo repositório.

Concomitantemente à compilação dos *subcorpora* e contrastação dos arquivos de legenda, foram realizados os procedimento de limpeza e conversão dos arquivos, que serão descrito na próxima subseção.

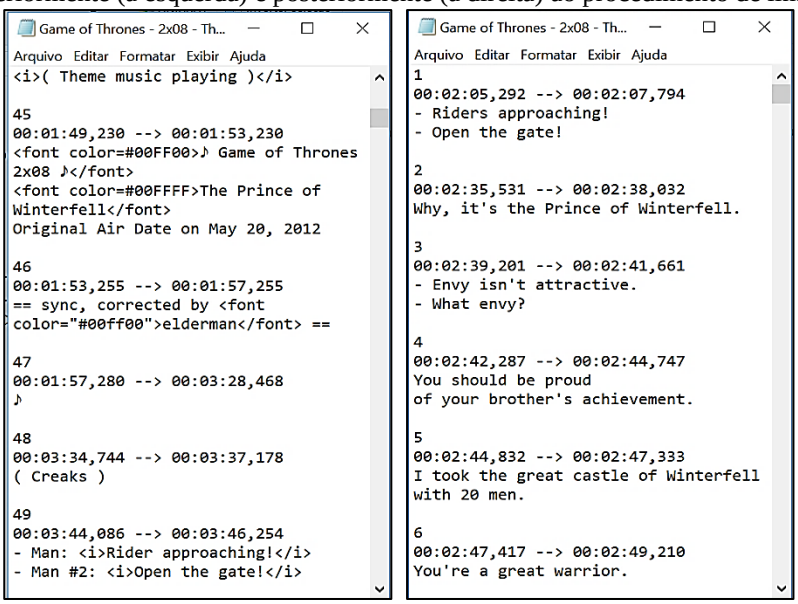
3.2 Limpeza e conversão dos arquivos de legenda

A limpeza do *corpus* foi realizada por meio do bloco de notas do *Windows* simultaneamente aos procedimentos de compilação do *corpus*, contrastação dos arquivos e definição do início e/ou fim de cada episódio. Não foi necessário fazer a limpeza da marcação de tempo das legendas, pois o programa de análise lexical utilizado na pesquisa contabiliza apenas palavras. Entretanto, foi preciso examinar exaustivamente cada arquivo de legenda, pois é comum encontrar em meio às legendas o nome do episódio, a identificação da equipe de legendagem, dos revisores, do responsável pela sincronização e/ou do próprio repositório que as disponibiliza. A Figura 2 a seguir demonstra o procedimento de limpeza:

[8] Disponível em: http://legendas.tv/. Acesso em: 23 mar. 2020.

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

Figura 2 – Visão parcial do arquivo de legenda do oitavo episódio da segunda temporada de *Game of Thrones* anteriormente (à esquerda) e posteriormente (à direita) ao procedimento de limpeza



Fonte: elaborada pelo autor

Com o intuito de realizar a limpeza do *corpus* de maneira mais rápida, analisamos alguns arquivos de legenda e percebemos a recorrência de símbolos como cerquilha (#), arroba (@) e aspas angulares (< >), e códigos como *font color*, *www*, *.net* e *.org* próximos à parte textual que não fazia parte das transcrições de porções de fala dos personagens e que, portanto, deveriam ser excluídas. Sendo assim, abrimos cada arquivo de legenda por meio do bloco de notas e fizemos a busca desses símbolos e códigos por meio do atalho Ctrl+F. Em sequência, identificamos e excluímos as porções de texto irrelevante para a pesquisa em questão.

Concomitantemente à realização da limpeza, cada arquivo foi salvo em formato *.txt* por esse ser geralmente o formato mais eficiente para processamento de *corpora* em programas de análise lexical. Posteriormente aos procedimentos de limpeza e conversão de ambos os *subcorpora*, iniciou-se o processo de padronização, sendo esse o último procedimento anterior ao processamento dos *subcorpora* e posterior identificação e análise das UFs. Apresentamos na subseção seguinte o procedimento de padronização dos *subcorpora* da referida pesquisa.

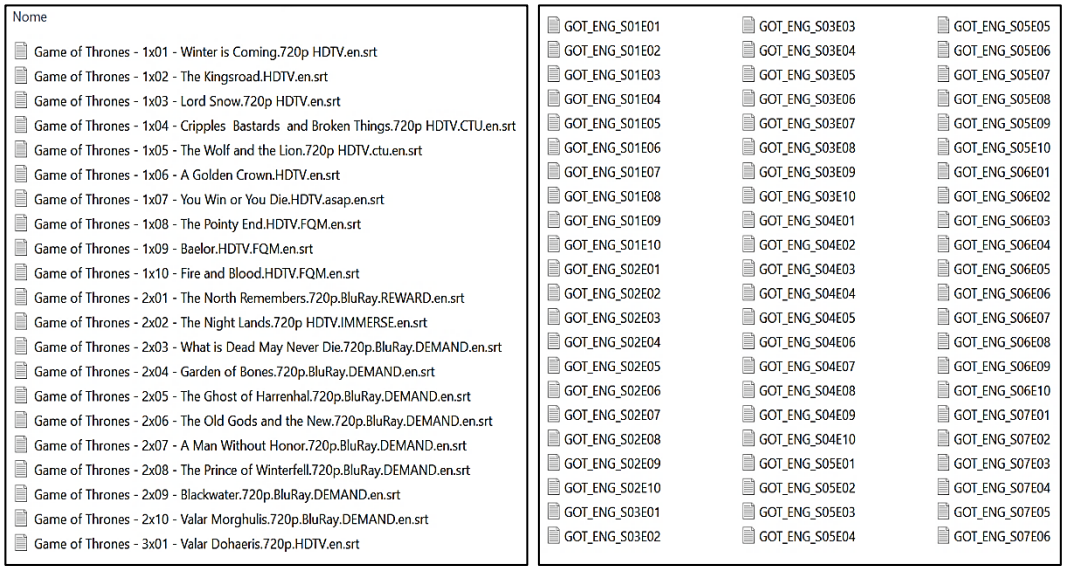
3.3 Padronização

Feita a limpeza e conversão dos *subcorpora*, foram criados diretórios separados para o *subcorpus* em inglês e o *subcorpus* em português, e, em seguida, realizou-se a

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

padronização da nomeação de todos os 73 arquivos que compõem cada um dos *subcorpora*. Os arquivos foram renomeados segundo um código estabelecido pelo pesquisador, tendo em vista facilitar a identificação e recuperação dos arquivos em português a partir da identificação das UFs no *subcorpus* em inglês, procedimento ilustrado na Figura 3 a seguir:

Figura 3 – Visão parcial do *subcorpus* em inglês de *Game of Thrones* anteriormente (à esquerda) e posteriormente (à direita) ao procedimento de padronização



Fonte: elaborada pelo autor

A imagem à esquerda apresenta o *subcorpus* em inglês anteriormente ao procedimento de padronização. A imagem à direita, por sua vez, apresenta o mesmo *subcorpus* após o procedimento em questão. De modo a simplificar a identificação dos arquivos, o código atribuído identifica simultaneamente o nome do seriado, a língua em que se encontra, a temporada e o episódio de cada arquivo.

Tomemos como exemplo o código GOT_ENG_S01E10. Esse código identifica o nome do seriado (GOT – *Game of Thrones*), o idioma do arquivo (ENG – *English*/Inglês), a temporada (S01 – *Season* 01/Temporada 01) e o episódio a que se refere o arquivo (E10 – *Episode* 10/Episódio 10). Optou-se pela nomenclatura em inglês (*season* e *episode*) para ambos os *subcorpora* devido a esse ser o padrão de nomeação de grande parte dos repositórios de legendas de seriados televisivos. Em virtude disso, a única diferença entre a nomenclatura dos arquivos que compõem cada *subcorpus* consiste na alteração do código ENG (*English)* para

POR (Português). Isto posto, o arquivo no *subcorpus* em português correspondente ao apresentado como exemplo neste parágrafo está nomeado como GOT_POR_S01E10.

Após os procedimentos de compilação, limpeza, conversão e padronização de ambos os *corpora*, deu-se início à identificação e análise das UFs. Na seção seguinte serão apresentadas algumas ferramentas e recursos relevantes para a identificação e análise das UFs no âmbito da pesquisa em questão, alguns utilizados na referida pesquisa e outros que podem ser utilizados para outros tipos de pesquisa.

4 PROCEDIMENTOS E FERRAMENTAS PARA IDENTIFICAÇÃO E ANÁLISE DE UFs: POSSIBILIDADES E APLICAÇÕES

O programa de análise lexical utilizado foi o WordSmith Tools 6.0 (SCOTT, 2012), doravante WST[9], cuja versão 4.0 é disponibilizada gratuitamente para *download*. O WST é uma suíte instalável constituída por diversas ferramentas e recursos para análise lexical. Dentre as principais ferramentas estão o Concord (Concordanciador), a KeyWords (Palavras-chave) e a WordList (Lista de Palavras). Essas três ferramentas foram utilizadas na referida pesquisa e serão apresentadas ao longo dessa seção.

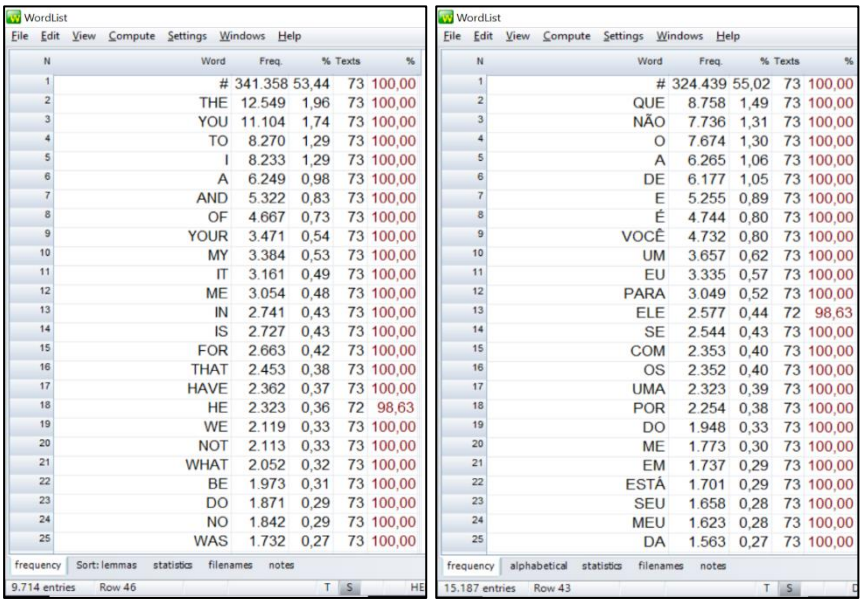
A primeira ferramenta utilizada foi a WordList, pois ela disponibiliza uma lista de todos os *types* do *corpus*, bem como apresenta a quantidade de vezes e o número de textos em que cada *type* ocorreu. Essa lista de palavras pode ser ordenada alfabeticamente, por ordem de frequência ou pelo número de textos em que cada *type* ocorreu (em ordem crescente ou decrescente).

Haja vista que o objetivo dessa pesquisa consistiu em analisar UFs recorrentes no *corpus* de legendas de *Game of Thrones*, optou-se pela apresentação da lista de palavras em ordem decrescente de frequência no *corpus*, como ilustrado na Figura 4 a seguir:

[9] Disponível em: https://www.lexically.net/wordsmith/. Acesso em: 23 mar. 2020.

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

Figura 4 – Visão parcial das listas de palavras de ambos os *subcorpora* geradas pela ferramenta WordList



WordList — File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	#	341.358	53,44	73	100,00
2	THE	12.549	1,96	73	100,00
3	YOU	11.104	1,74	73	100,00
4	TO	8.270	1,29	73	100,00
5	I	8.233	1,29	73	100,00
6	A	6.249	0,98	73	100,00
7	AND	5.322	0,83	73	100,00
8	OF	4.667	0,73	73	100,00
9	YOUR	3.471	0,54	73	100,00
10	MY	3.384	0,53	73	100,00
11	IT	3.161	0,49	73	100,00
12	ME	3.054	0,48	73	100,00
13	IN	2.741	0,43	73	100,00
14	IS	2.727	0,43	73	100,00
15	FOR	2.663	0,42	73	100,00
16	THAT	2.453	0,38	73	100,00
17	HAVE	2.362	0,37	73	100,00
18	HE	2.323	0,36	72	98,63
19	WE	2.119	0,33	73	100,00
20	NOT	2.113	0,33	73	100,00
21	WHAT	2.052	0,32	73	100,00
22	BE	1.973	0,31	73	100,00
23	DO	1.871	0,29	73	100,00
24	NO	1.842	0,29	73	100,00
25	WAS	1.732	0,27	73	100,00

frequency | Sort: lemmas | statistics | filenames | notes

9.714 entries | Row 46

WordList — File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	#	324.439	55,02	73	100,00
2	QUE	8.758	1,49	73	100,00
3	NÃO	7.736	1,31	73	100,00
4	O	7.674	1,30	73	100,00
5	A	6.265	1,06	73	100,00
6	DE	6.177	1,05	73	100,00
7	E	5.255	0,89	73	100,00
8	É	4.744	0,80	73	100,00
9	VOCÊ	4.732	0,80	73	100,00
10	UM	3.657	0,62	73	100,00
11	EU	3.335	0,57	73	100,00
12	PARA	3.049	0,52	73	100,00
13	ELE	2.577	0,44	72	98,63
14	SE	2.544	0,43	73	100,00
15	COM	2.353	0,40	73	100,00
16	OS	2.352	0,40	73	100,00
17	UMA	2.323	0,39	73	100,00
18	POR	2.254	0,38	73	100,00
19	DO	1.948	0,33	73	100,00
20	ME	1.773	0,30	73	100,00
21	EM	1.737	0,29	73	100,00
22	ESTÁ	1.701	0,29	73	100,00
23	SEU	1.658	0,28	73	100,00
24	MEU	1.623	0,28	73	100,00
25	DA	1.563	0,27	73	100,00

frequency | alphabetical | statistics | filenames | notes

15.187 entries | Row 43

Fonte: elaborada pelo autor

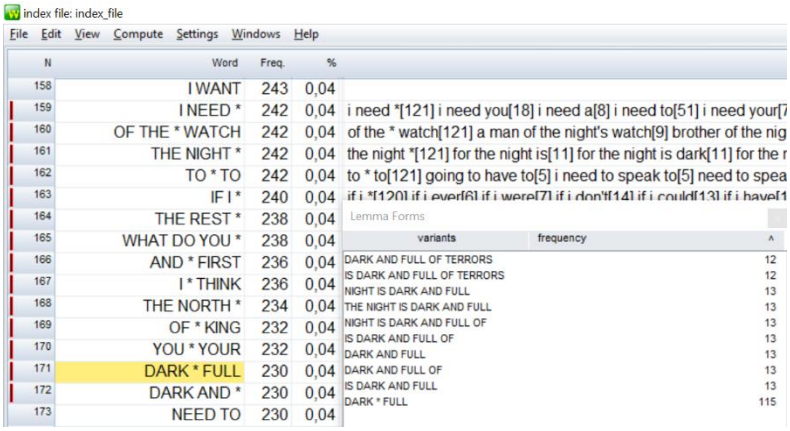
A partir das listas de palavras geradas pela ferramenta WordList (Figura 4), é possível identificar as palavras mais recorrentes em ambos os *subcorpora* e, por conseguinte, facilita a identificação de UFs frequentes nos *subcorpora*. As listas de palavras apresentadas na Figura 4 são constituídas por 9.714 e 15.187 *types*, respectivamente, organizados em ordem decrescente de ocorrência (primeira e segunda coluna de cada imagem). A terceira, quarta, quinta e sexta colunas exibem a frequência de cada *type*, a porcentagem de todas as ocorrências de cada *type* em relação ao *subcorpus* de estudo, o número de arquivos que contêm ao menos uma ocorrência de cada *type* e a porcentagem do número de arquivos em relação ao total de arquivos do *subcorpus*.

Ademais, é possível utilizar *stop lists* concomitantemente à geração das listas de palavras, lista de palavras-chave e linhas de concordância, conforme os critérios e objetivos de cada pesquisa. As *stop lists* atuam como filtros no processamento dos *corpora*, sendo possível eliminar dados linguísticos não relevantes para o propósito da pesquisa, como, por exemplo, preposições, conjunções, partículas adverbiais, dentre outros. Não obstante, como o objetivo da pesquisa em questão consistiu em analisar, além das UFs, as coligações de regência, coligações prepositivas e *phrasal verbs*, optamos por não utilizar uma *stop list* para as palavras gramaticais, pois dificultaria a identificação de candidatos a coligações.

Conforme o objetivo da pesquisa, também é possível utilizar *lemma lists*, listas geralmente utilizadas para simplificar a apresentação das listagens de palavras e facilitar a busca por variações da mesma palavra. A *lemma list*, possibilita o agrupamento de *types* que compartilham de determinadas características, como prefixos, sufixos ou variações de uma mesma palavra. A *lemma list* pode ser gerada automaticamente pela ferramenta, tendo como critério apenas prefixos e sufixos em comum, ou pode ser elaborada no bloco de notas e recuperada por meio da aba configurações avançadas da Wordlist anteriormente ao processamento do *corpus*. Como o objetivo da referida pesquisa não dependia necessariamente da lematização de *types*, optamos por não realizar esse procedimento, pois seria um processo consideravelmente demorado.

É também possível configurar a WordList para gerar uma lista de frequência de *clusters* (padrões de coocorrência entre itens lexicais) encontrados no *corpus* sob análise, recurso que facilita a identificação de UFs. O número de palavras que compõem cada *cluster* e a frequência mínima de ocorrência são definidos pelo pesquisador, sendo oito o número máximo de itens computados para cada *cluster*. É possível realizar o levantamento de *clusters* a partir de todos os *types* do *corpus* ou buscar por *clusters* que contenham palavras específicas determinadas pelo pesquisador. A Figura 5 ilustra um dos *clusters* identificado por meio da lista de *clusters* gerada pela WordList:

Figura 5 – *Cluster* identificado na lista de *clusters* gerada pela ferramenta WordList



Fonte: elaborada pelo autor

Na Figura 5, a primeira coluna exibe a ordem decrescente de frequência dos *clusters* identificados pela ferramenta, ao passo que a segunda apresenta os itens lexicais e/ou

gramaticais que os constituem, e é seguida pela terceira coluna que apresenta a quantidade de vezes que os itens constituintes coocorreram no *subcorpus* de estudo. A quarta coluna apresenta a porcentagem que esse *cluster* representa em relação ao *subcorpus* e a quinta coluna apresenta as variações de cada *cluster*, tendo como parâmetro o número de itens previamente definidos.

À título de exemplo, o *cluster* selecionado na Figura 5 (*dark * full*) é o 171º *cluster* mais frequente no *corpus*, ocorrendo 230 vezes no total e representando 0,4% em relação ao *corpus* sob análise. A partir de um duplo clique na quinta coluna referente às variações dos elementos constituintes dos *clusters*, a ferramenta exibe uma janela *pop-up* que apresenta os padrões de coocorrência que contêm *dark* e *full* como elementos constituinte e o número de vezes que cada padrão ocorreu.

A partir disso, já é possível identificar que *night is dark and full of terrors* se constitui como um forte candidato à UF no *subcorpus* de estudo devido à sua alta frequência de ocorrência, bem como de coocorrência entre os itens lexicais que o constitui. Para verificar se o referido *cluster* se constitui realmente como uma UF no *subcorpus* em questão, assim como para analisar suas propriedades combinatórias, morfossintáticas e semânticas, o próximo passo foi utilizar a ferramenta Concord, que será apresentada posteriormente nessa seção.

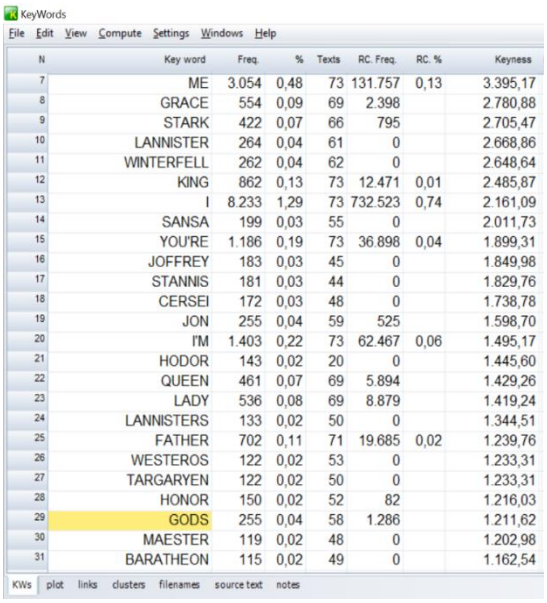
A lista de palavras gerada pela WordList também pode ser utilizada para dar início ao processo de identificação de palavras-chave por meio da ferramenta KeyWords. Essa ferramenta é essencial para estudos voltados à identificação de termos em diversos campos de conhecimento, bem como no mundo ficcional dos seriados televisivos.

Anteriormente à geração da lista de palavras-chave, é necessário recorrer à lista de palavras de um *corpus* de referência que seja aproximadamente cinco vezes maior que o *corpus* de estudo. Esse procedimento deve ser realizado, pois a ferramenta KeyWords realiza uma comparação entre a frequência relativa (em porcentagem) de cada palavra no *corpus* de estudo e a frequência relativa de cada palavra no *corpus* de referência. Por fim, a ferramenta identifica no *corpus* de estudo tanto palavras-chave positivas (de frequência relativa é expressivamente maior) como palavras-chave negativas (cuja frequência relativa é muito menor).

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

A Figura 6 demonstra a lista de palavras-chave obtida por meio da ferramenta KeyWords a partir da contrastação entre a lista de palavras do *subcorpus* em inglês com a lista de palavras do *British National Corpus*[10], o *corpus* de referência utilizado na referida pesquisa:

Figura 6 – Visão parcial da lista de palavras-chave do *subcorpus* de em inglês gerada pela ferramenta KeyWords


KeyWords

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness
7	ME	3.054	0,48	73	131.757	0,13	3.395,17
8	GRACE	554	0,09	69	2.398		2.780,88
9	STARK	422	0,07	66	795		2.705,47
10	LANNISTER	264	0,04	61	0		2.668,86
11	WINTERFELL	262	0,04	62	0		2.648,64
12	KING	862	0,13	73	12.471	0,01	2.485,87
13	I	8.233	1,29	73	732.523	0,74	2.161,09
14	SANSA	199	0,03	55	0		2.011,73
15	YOU'RE	1.186	0,19	73	36.898	0,04	1.899,31
16	JOFFREY	183	0,03	45	0		1.849,98
17	STANNIS	181	0,03	44	0		1.829,76
18	CERSEI	172	0,03	48	0		1.738,78
19	JON	255	0,04	59	525		1.598,70
20	I'M	1.403	0,22	73	62.467	0,06	1.495,17
21	HODOR	143	0,02	20	0		1.445,60
22	QUEEN	461	0,07	69	5.894		1.429,26
23	LADY	536	0,08	69	8.879		1.419,24
24	LANNISTERS	133	0,02	50	0		1.344,51
25	FATHER	702	0,11	71	19.685	0,02	1.239,76
26	WESTEROS	122	0,02	53	0		1.233,31
27	TARGARYEN	122	0,02	50	0		1.233,31
28	HONOR	150	0,02	52	82		1.216,03
29	GODS	255	0,04	58	1.286		1.211,62
30	MAESTER	119	0,02	48	0		1.202,98
31	BARATHEON	115	0,02	49	0		1.162,54

KWs plot links clusters filenames source text notes

Fonte: elaborada pelo autor

A Figura 6 apresenta a lista de palavras-chave organizada em ordem decrescente pela chavicidade (*keyness*)[11]. A primeira coluna exibe a ordenação das palavras (de acordo com a chavicidade), e a segunda apresenta as palavras-chave em questão. A terceira coluna exibe o número de vezes que cada palavra-chave ocorreu no *subcorpus* de estudo, a quarta apresenta a porcentagem que as ocorrências de cada palavra-chave representa em relação ao total de *tokens* do *subcorpus* e a quinta coluna exibe a quantidade de textos em que cada palavra-chave ocorreu. A sexta e sétima coluna são referentes ao *corpus* de referência, onde constam, respectivamente,

[10] O *British National Corpus* (BNC) conta com 100 milhões de *tokens* de inglês da variedade britânica escrito e falado. O BNC foi escolhido como *corpus* de referência devido à disponibilização de sua lista de palavras gratuitamente no *site* do WST e por ser constituído de amostras da mesma variedade de língua inglesa utilizada no seriado *Game of Thrones*. A lista de palavras do BNC está disponível em: https://lexically.net/wordsmith/support/extras.html. Acesso em: 23 mar. 2020.

[11] A chavicidade (*keyness*) é o fator que indica se determinada palavra-chave é mais frequente no *corpus* de estudo ou no *corpus* de referência. A alta chavicidade indica que determinada palavra-chave acontece única ou majoritariamente no *corpus* de estudo, constituindo-se um forte candidato a termo.

o número de ocorrências de cada palavra-chave e a porcentagem desse número em relação ao total de *tokens* do *corpus* referência. Por fim, a última coluna apresenta a chavicidade.

À título de exemplo selecionamos a palavra-chave *gods* que encontra-se destacada na Figura 6. Ela é a palavra-chave de número 29 (pela ordem de chavicidade), ocorreu 255 vezes no *corpus* de estudo, representando 0,04% do total de *tokens*, e consta em 58 dos 73 arquivos que compõem o *subcorpus* em inglês. Em relação ao *corpus* de referência, a palavra-chave *gods* aparece 1.286 vezes, e, haja vista que esse *corpus* conta com 100 milhões de *tokens*, sua porcentagem é insignificante (0,0012%), portanto, possui chavicidade alta e, por esse motivo, é um candidato a termo no *subcorpus* de estudo.

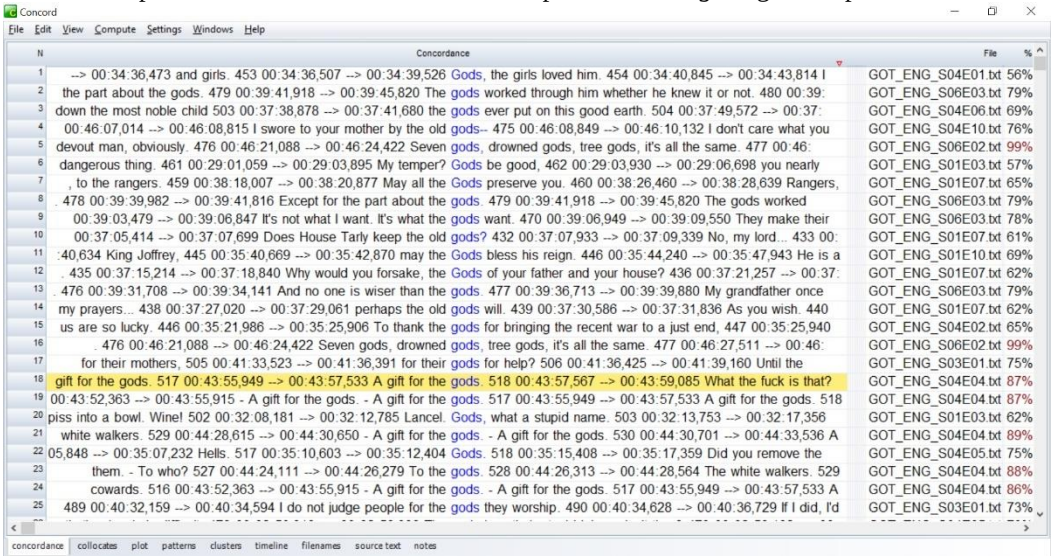
Para certificar se determinada palavra é realmente um termo ou elemento constituinte de uma ou mais UFs, ter acesso a exemplos de uso e observar características e padrões tangentes à determinada palavra, é necessário recorrer à ferramenta Concord. O Concord é uma ferramenta que busca no *corpus* de estudo todas as ocorrências da palavra ou frase selecionada pelo pesquisador. É possível utilizar essa ferramenta por meio da seleção de palavras listadas pela WordList e Keywords (seguida pelo comando Shift+Ctrl+C), assim como é possível pesquisar por palavras ou frases diretamente no ambiente de busca do Concord.

O Concord possibilita a visualização de diversas características concernentes à(s) palavra(s) de busca, como: (i) os colocados e clusters, ou seja, palavras ou agrupamento de palavras que tendem a coocorrer com a(s) palavra(s) de busca; (ii) a frequência de ocorrência da(s) palavra(s) de busca em cada arquivo que compõe o *corpus* de estudo, bem como no *corpus* em sua totalidade; (iii) identifica o nome dos arquivos e permite acessar o contexto linguístico em cada arquivo de legenda que determinada(s) palavra(s) de busca ocorre(m), dentre outros recursos.

A Figura 7 a seguir apresenta uma visão parcial das linhas de concordâncias da palavra-chave *gods* no *corpus* de estudo geradas pela ferramenta Concord:

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

Figura 7 – Visão parcial das linhas de concordância da palavra-chave *gods* geradas pela ferramenta Concord



Fonte: elaborada pelo autor

O *layout* de apresentação em linhas de concordância ilustrado na Figura 7 é também conhecido como KWIC (do inglês, *Key-Word-in-Context*, palavra-chave em contexto). Esse modelo apresenta todas as ocorrências de determinada palavra de busca – que no caso do exemplo é *gods* – ladeadas pelas palavras que a antecedem e a sucedem, cujo número máximo é definido pelo pesquisador. O *layout* das linhas de concordância é essencial para averiguar se determinada palavra é um termo ou se faz parte de uma UF, assim como para observar aspectos morfossintáticos, semânticos e padrões lexicogramaticais, justamente pela disponibilização de todas as suas ocorrências no *corpus* juntamente com os cotextos e contextos[12] linguísticos.

A partir da linha de concordância destacada na Figura 7, é possível perceber padrões de coocorrência e, portanto, um candidato a UF: *a gift for the gods*, *cluster* que ocorre quatro vezes nas linhas de concordância exibidas. Entretanto, como é possível observar na terceira coluna da Figura 7, as quatro ocorrências foram utilizadas no mesmo episódio, fator que dificulta considerar *a gift for the gods* como uma UF recorrente no *subcorpus* como um todo.

Ao utilizar a ferramenta Concord, é possível recorrer ao Concordance Sort, recurso que possibilita a reorganização da listagem de linhas de concordância conforme critérios estabelecidos pelo pesquisador no ambiente da ferramenta. Na presente pesquisa, reordenamos

[12] Neste artigo, utiliza-se o termo cotexto para fazer referência às associações entre palavras no eixo sintagmático da oração, ou seja, no "ambiente linguístico imediato" de determinada palavra de busca. Por outro lado, utiliza-se o termo *contexto* em um sentido mais amplo, não restrito somente ao nível da frase.

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

as linhas de concordância estabelecendo como critério a ordenação alfabética das palavras que ocorrem à esquerda e/ou à direita da palavra de busca, tendo em vista facilitar a identificação de UFs e a verificação do número de ocorrências e variações de cada uma. A Figura 8 a seguir apresenta as linhas de concordâncias da palavra de busca *gods* após a utilização do recurso Concordance Sort:

Figura 8 – Visão parcial das linhas de concordância da palavra-chave *gods* após a utilização do recurso Concordance Sort por meio da ferramenta Concord

Fonte: elaborada pelo autor

O recurso ilustrado na figura anterior é essencial para agilizar a identificação e análise de UFs, pois possibilita o agrupamento das linhas de concordância, facilitando a verificação de padrões, frequência e variações das UFs. Na Figura 7, identificamos quatro ocorrências do *cluster a gift for the gods*, ao passo que na Figura 8, após a reordenação das linhas de concordância, é possível verificar nove ocorrências desse mesmo *cluster*. Novamente, ao analisar a terceira coluna, percebe-se que as nove ocorrências desse *cluster* são oriundas do mesmo arquivo de legenda. Portanto, na pesquisa em questão não nos ocupamos do estudo desse *cluster* e nem o consideramos uma UF recorrente no *subcorpus* como um todo, visto que ele está presente em apenas 1,3% dos arquivos que o constitui.

Na Figura 8 também é possível perceber seis coocorrência do *cluster before the gods*, e a terceira coluna evidencia que esse *cluster* ocorreu majoritariamente na quinta temporada do seriado, mas, ao contrário do que ocorre com *a gift for the gods*, ele ocorre em seis episódios. Isto posto, consideramos o *cluster before the gods* como uma UF recorrente no

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

subcorpus em inglês, sendo o próximo passo a análise de seus cotextos e contextos linguísticos e, em seguida, a verificação de seu(s) correspondente(s) no *subcorpus* em português.

Retomaremos o exemplo do *cluster night is dark and full of terrors* identificado anteriormente na Figura 5 para apresentar a busca diretamente no ambiente de pesquisa do Concord. Ao abrir o Concord, é possível pesquisar diretamente a palavra ou *cluster* para análise e a ferramenta apresenta todas as ocorrências da(s) palavra(s) de busca. A Figura 9 apresenta os resultados obtidos por meio da busca pelo *cluster* em questão na ferramenta Concord.

Figura 9 – Visão geral das linhas de concordância de *night is dark and full of terrors* geradas por meio da ferramenta Concord

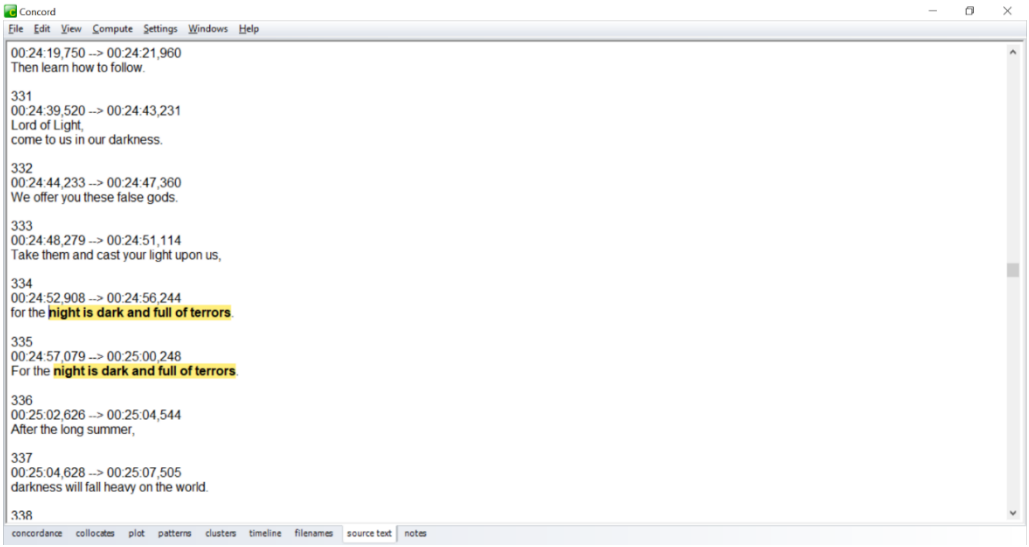
Fonte: elaborada pelo autor

As linhas de concordância apresentadas na Figura 9 estão ordenadas de acordo com o arquivo em que aparecem, e a terceira coluna exibe o código do arquivo. Ao analisar as linhas de concordância apresentadas, percebe-se que o *cluster night is dark and full of terrors* é uma UF recorrente no *subcorpus* em inglês e possui um alto nível de fixidez. Essa UF ocorre em sete episódios de seis temporadas diferentes, totalizando doze ocorrências. Também é possível perceber que via de regra, preposições e/ou artigos compõem o *cluster* e, desse modo, ele é mais extenso do que se observou anteriormente: (*for the*) *night is dark and full of terrors*.

Por meio de um duplo clique em determinada linha de concordância, o Concord direciona o pesquisador para o momento exato em que a(s) palavra(s) de busca ocorre(m) nos arquivos que compõem o *corpus* de estudo, disponibilizando, portanto, seus contextos linguísticos. A Figura 10 a seguir demonstra o contexto linguístico da primeira linha de concordância apresentada na Figura 9.

ISSN. 1807-9342 ITINERARIUS REFLECTIONIS V. 16 | Nº 2 | 2020
Revista Eletrônica de Graduação e Pós-Graduação em Educação

Figura 10 – Contexto linguístico da primeira linha de concordância de *night is dark and full of terrors* disponibilizado pela ferramenta Concord



Fonte: elaborada pelo autor

O acesso ao contexto linguístico viabilizado pelo Concord (Figura 10) permite realizar, em certa medida, análises semânticas e/ou pragmáticas da palavra ou *cluster* sob análise. Ademais, a ferramenta viabiliza a exploração de todo o arquivo de legenda, possibilitando analisar minutos anteriores e posteriores à utilização de cada palavra ou *cluster* de busca. Portanto, essa ferramenta é relevante tanto para pesquisas fraseológicas como terminográficas.

À título de exemplo, a partir da análise do contexto apresentado na Figura 10, é possível inferir que a UF *for the night is dark and full of terrors* é utilizada no âmbito religioso, especificamente em encerramento de preces direcionadas ao Senhor da Luz (*Lord of Light*). Fora do mundo ficcional do seriado, a convencionalidade à nível pragmático dessa UF corresponderia ao encerramento de determinadas preces, como por exemplo, algumas preces cristãs que são finalizadas com "em nome do pai, do filho e do espírito santo".

É possível inferir também que *dark* nesse contexto é utilizado como algo negativo, que pode ser uma metáfora para as aflições e angústias vivenciadas por aqueles que intercedem ao Senhor da Luz, bem como pode estar relacionada ao inverno que virá após o longo verão[13], haja vista que no mundo ficcional de *Game of Thrones*, o verão e o inverno, que por sinal é

[13] *After the long summer, darkness will fall heavy on the world* (excerto extraído do contexto apresentado na Figura 10 do arquivo GOT_ENG_S02E01).

muito rigoroso, duram décadas e ao longo do enredo o temor pelo longo inverno é compartilhado por grande parte dos personagens. Por fim, a partir da análise dos contextos linguísticos das outras dez ocorrências de *the night is dark and full of terrors* será possível restringir ou estender a descrição do significado e usos dessa UF, assim como elaborar uma definição.

A partir da identificação e análise das UFs no *subcorpus* em inglês, partiu-se para o *subcorpus* em português para analisar seus correspondentes. Esse procedimento foi realizado por meio de buscas por possíveis equivalentes tradutórios, tanto das UFs quanto das palavras que ocorrem em seus cotextos, na ferramenta Concord. Quando não foi possível achá-los, identificamos o arquivo e a marcação de tempo em que cada UF sob análise foi utilizada no *subcorpus* em inglês e acessamos manualmente os arquivos do *subcorpus* em português em busca pelos correspondentes, baseando-nos na marcação de tempo das legendas.

Como apresentado nessa seção, são várias as ferramentas, recursos e percursos passíveis de utilização e replicabilidade para a identificação e análise de UFs em *corpora* de seriados televisivos. Além das ferramentas e recursos apresentados nessa seção, o WST ainda possui diversos outros que são essenciais para pesquisas lexicais. O WST é apenas um dos numerosos programas de análise lexical que dispõem de ferramentas proveitosas para a realização de estudos linguísticos. Além do WST, há outros ambientes, programas e ferramentas para estudos lexicais, como AntConc[14], Sketch Engine[15], TermoStat[16] e Voyant Tools[17], alguns disponíveis gratuitamente, outros pagos.

5 CONSIDERAÇÕES FINAIS

No presente artigo, apresentamos detalhadamente os procedimentos realizados e ferramentas utilizadas desde a compilação do *corpus* até a análise das UFs em uma pesquisa fraseológicas voltada às UFs do seriado televisivo *Game of Thrones*, buscando, sempre que possível, exemplificar os procedimentos por meio de figuras e utilizando UFs e candidatos a UFs identificados no escopo da pesquisa em questão. Alguns dos resultados preliminares dessa

[14] Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 23 mar. 2020.
[15] Disponível em: https://www.sketchengine.eu/. Acesso em: 23 mar. 2020.
[16] Disponível em: http://termostat.ling.umontreal.ca/. Acesso em: 23 mar. 2020.
[17] Disponível em: https://voyant-tools.org/. Acesso em: 23 mar. 2020.

pesquisa estão publicados em um artigo anterior (YAMAMOTO; LISBOA, 2019), que objetivou apresentar um panorama das concepções e taxonomias de UFs segundo diversos autores e realizar uma análise sintática de UFs oriundas de algumas pesquisas voltadas ao léxico dos seriados televisivos.

Esperamos que este artigo sirva como uma introdução às pesquisas lexicais voltadas aos seriados televisivos, assim como aos procedimentos e critérios para a realização de análises lexicais por meio de *corpora* de legendas. Ademais, esperamos ter demonstrado o campo frutífero do mundo ficcional dos seriados para os estudos lexicais. Por fim, esperamos que o artigo venha a incentivar e auxiliar futuras pesquisas voltadas ao léxico do mundo ficcional dos seriados televisivos, haja vista que os procedimentos metodológicos descritos podem ser replicados.

REFERÊNCIAS

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. **Calidoscópio**, São Leopoldo, v. 4, n. 3, p. 159-178. 2006. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 23 set. 2019.

BANG, M.; FROMM, G. Terminologia em série: House M. D. **EntreLetras**, Araguaína, v. 4, n. 2, p. 114-133. 2013. Disponível em: https://sistemas.uft.edu.br/periodicos/index.php/entreletras/article/view/995/533. Acesso em: 24 set. 2019.

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

CORPAS PASTOR, G. **Manual de fraseología española**. Madrid: Gredos, 1996.

MONTEIRO-PLANTIN, R. S. **Fraseologia**: era uma vez um patinho feio no ensino de língua materna. Fortaleza: Imprensa Universitária, 2014.

MURAD, C. R. R. O. O léxico da série *Law and Order*: uma análise inicial baseada em corpus paralelo. **TradTerm**, São Paulo, v. 25, n. 1, p. 169-197. 2015. Disponível em: http://www.revistas.usp.br/tradterm/article/view/103247/101676. Acesso em: 24 set. 2019.

PARODI, G. **Lingüística de Corpus**: de la teoría a la empiria. Madrid: Iberoamericana/Vervuert, 2010.

PEIXOTO, L. M. Identificação de unidades fraseológicas no vocabulário de Star Trek: abordagens corpus-driven e corpus-based. **Domínios de Lingu@gem**, Uberlândia, v. 8, n. 2, p. 139-163. 2014. Disponível em: http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/27630/15768. Acesso em: 24 set. 2019.

ROCHA, J. M. P.; ORENHA-OTTAIANO, A. Colocações especializadas na área médica extraídas a partir do corpus House M.D. **Cadernos do IL**, Porto Alegre, n. 44, p. 295-318. 2012. Disponível em: https://seer.ufrgs.br/cadernosdoil/article/view/28051/18849. Acesso em: 24 set. 2019.

SCOTT, M. **WordSmith Tools**. Versão 6, 2012. Disponível em: https://www.lexically.net/wordsmith/downloads/. Acesso em: 22 nov. 2019.

SINCLAIR, J. **Corpus, Concordance, Collocation**. Oxford: Oxford University Press, 1991.

TAGNIN, S. E. O. **O jeito que a gente diz**: combinações consagradas em inglês e português. Barueri: Disal, 2013.

YAMAMOTO, M. I.; LISBOA, J. V. R. Corpus Linguistics and phraseologies on TV series: a successful experience. **Itinerarius Reflectionis**, Jataí, v. 15, n. 2, p. 1-19. 2019. Disponível em: https://www.revistas.ufg.br/rir/article/view/58713/33207. Acesso em: 24 set. 2019.