UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE LETRAS E LINGUÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

LUCAS MACIEL PEIXOTO

O *CORPUS OF ENGLISH LANGUAGE VIDEOS*: UMA NOVA FERRAMENTA DE *CORPUS ON-LINE* PARA APRENDIZAGEM DIRECIONADA POR DADOS

UBERLÂNDIA - MG

2016

LUCAS MACIEL PEIXOTO

O *CORPUS OF ENGLISH LANGUAGE VIDEOS*: UMA NOVA FERRAMENTA DE *CORPUS ON-LINE* PARA APRENDIZAGEM DIRECIONADA POR DADOS

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Linguísticos do Instituto de Letras e Linguística da Universidade Federal de Uberlândia, como requisito parcial para obtenção do título de Mestre em Estudos Linguísticos

Área de Concentração: Estudos em Linguística e Linguística Aplicada
Linha de pesquisa: Teoria, Descrição e Análise Linguística

Orientador: Prof. Dr. Guilherme Fromm

Uberlândia - MG

2016

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

P379c
2016

Peixoto, Lucas Maciel, 1989-
O Corpus of English Language Videos : uma nova ferramenta de corpus on-line para aprendizagem direcionada por dados / Lucas Maciel Peixoto. - 2016.
114 f. : il.

Orientador: Guilherme Fromm.
Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Estudos Linguísticos.
Inclui bibliografia.

1. Linguística - Teses. 2. Linguística de corpus - Teses. 3. Língua inglesa - Ensino auxiliado por computador - Teses. 4. Língua inglesa - Estudo e ensino - Inovações tecnológicas - Teses. I. Fromm, Guilherme. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Estudos Linguísticos. III. Título.

CDU: 801

AGRADECIMENTOS

Ao caro orientador Prof. Dr. Guilherme Fromm, que me mostrou a Linguística de Corpus e a Linguística Computacional, que me ensinou grande parte do que sei e que sempre se manteve paciente, leve e de bom humor durante as orientações.

Aos mentores que proporcionaram momentos importantes da minha formação: Prof. Dr. Ariel Novodvorski, Profª. Drª. Carla Tavares, Profª. Drª. Cristiane Carvalho, Profª. Drª. Daisy Vale, Profª. Drª. Maria Clara Magalhães e Profª. Drª. Valeska Souza.

Aos professores que participaram da banca de qualificação deste trabalho, Prof. Dr. Ariel Novodvorski e Profª. Drª. Fernanda Ribas, pelos comentários construtivos.

A todos do Instituto de Letras e Linguística da UFU, especialmente aos colegas dos grupos de pesquisa GPELC e PLEX, com quem compartilhei ótimos momentos de aprendizagem e de convívio amigável e divertido durante as reuniões e congressos.

Ao Prof. Dr. Mauricio Cunha Escarpinati, ao Prof. Dr. Marcelo Keese Albertini e ao Luiz Fernando Afra Brito, docentes e discente da Faculdade de Computação da UFU que realizaram o serviço essencial de desenvolver o sistema computacional deste projeto.

A toda a minha família, especialmente aos meus pais e às minhas irmãs, pelo carinho e apoio em todos os momentos, e por terem me ensinado o valor do estudo, do trabalho, do conhecimento e da dedicação.

Ao Wagner, pela amizade de longa data e pela presença constante em minha vida, tanto nos aspectos acadêmicos quanto nos não acadêmicos.

À Carol, ao Igor, à Inês e ao Maico, pelos passeios, festas, risadas, momentos memoráveis e tudo o mais que constitui nossas amizades e me traz grande felicidade.

À Marina, por me ensinar novas maneiras de revisar meus pontos de vista e traduzir meus sentimentos.

RESUMO

O tema central deste trabalho é o uso de *corpora* no ensino-aprendizagem de língua inglesa, e seu objetivo principal foi a concepção e desenvolvimento do *Corpus of English Language Videos* (CELV), juntamente com uma ferramenta *on-line* para consultas a esse *corpus*, que inclui, entre outras funções, a geração de linhas de concordância. O principal embasamento teórico para o desenvolvimento e aplicação do *corpus* e de sua ferramenta de consulta foi a Aprendizagem Direcionada por Dados, uma abordagem ao ensino-aprendizagem de línguas que propõe o papel do aluno como investigador da língua por meio da observação empírica de dados linguísticos contidos em um *corpus*, com auxílio do professor e do concordanciador e passando por um processo cognitivo indutivo de aprendizagem. O principal embasamento metodológico para a realização do trabalho foi a Linguística de *Corpus*, a área da Linguística dedicada à compilação e análise de amostras de língua em formato eletrônico por meio de computadores, e também a Linguística Computacional, a área da Linguística destinada à criação e uso de ferramentas computacionais para o processamento de línguas. Foi compilado um *corpus* a partir de legendas de vídeos do *site* YouTube, e foi desenvolvido e publicado na *internet* um sistema computacional de busca a esse *corpus*, para uso de pesquisadores, professores e alunos interessados no ensino-aprendizagem de inglês. O *corpus* possui tamanho e variedade linguística suficientes para ser usado como fonte de exemplos de língua inglesa em uso, e sua ferramenta de consulta permite não só a observação de linhas de concordância de forma escrita, mas, também, o acesso aos vídeos originais que compõem a amostra, possibilitando que o usuário assista e ouça as palavras e expressões pesquisadas no *corpus* em forma audiovisual. A principal fonte de inspiração para a compilação do *corpus* e desenvolvimento da ferramenta foi o *Corpus of Contemporary American English* (COCA). A página do CELV na *internet* possui uma interface de pesquisa simples e de fácil uso, buscando ajudar a promover a difusão das técnicas da Linguística de *Corpus* para que mais pesquisadores, professores e alunos possam tirar proveito das vantagens do uso de *corpora* em suas pesquisas, atividades de ensino e de aprendizagem. Após a conclusão do desenvolvimento do *corpus* e da ferramenta, ambos foram testados com professores de inglês, e foram levantadas as opiniões desses participantes da pesquisa sobre a utilidade do CELV em suas atividades de ensino, recebendo, de maneira geral, avaliações positivas. Ainda, foram feitas sugestões sobre tópicos em língua inglesa cuja aprendizagem pode ser enriquecida com uso desta ferramenta. Espera-se que o produto deste trabalho possa vir a contribuir com o ensino-aprendizagem de inglês e com a literatura sobre as aplicações de *corpora* nesse contexto.

Palavras-chave: Linguística de *Corpus*. Aprendizagem Direcionada por Dados. *Corpora on-line*.

ABSTRACT

The main topic of this research is the use of corpora in the teaching and learning of the English language, and its main objective was the conception and development of the Corpus of English Language Videos (CELV), along with an on-line tool for queries in this corpus, which includes, among other functions, the generation of concordance lines. The main theoretical basis for the development and application of the corpus and its query tool was Data-Driven Learning, an approach to the teaching and learning of languages which proposes that the role of the student is that of a language investigator who conducts empiric observation of the linguistic data contained in a corpus, aided by the teacher and the concordancer and following an inductive cognitive process for learning. The main methodological basis for the implementation of the project was Corpus Linguistics, the field in Linguistics dedicated to the compilation and analysis of language samples in electronic format using computers, and Computational Linguistics, the field in Linguistics aimed at the creation and use of computer tools for the processing of languages. A corpus was compiled using YouTube video subtitles, and a computer system for searches in this corpus was developed and published on the internet, to be used by researchers, professors and students interested in the teaching and learning of English. The corpus has enough size and linguistic variation to be used as a source for examples of English language in use, and its query tool enables not only the observation of concordance lines in written form, but also the access to the original videos which compose the sample, allowing the user to watch and listen to the words and expressions searched in the corpus in audiovisual format. The main source of inspiration for the compilation of the corpus and development of the tool was the Corpus of Contemporary American English (COCA). The CELV website has a simple and easy-to-use search interface, aiming to help promote the spread of Corpus Linguistics techniques so that more researchers, professors and students may benefit from the advantages of using corpora in their researches, teaching and learning activities. After the conclusion of the development of the corpus and the tool, both were tested with English teachers, and the opinions of these participants of the research about the usefulness of CELV in their teaching activities were raised, receiving, in general, positive feedback. Additionally, suggestions about topics of the English language whose learning can be enriched by using this tool were made. It is expected that the product of this research may contribute with the teaching and learning of English and with the literature about the applications of corpora in this context.

Keywords: Corpus Linguistics. Data-Driven Learning. On-line corpora.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE GRÁFICOS

SUMÁRIO

1. INTRODUÇÃO

Desde o início da minha prática como professor de língua inglesa, tenho o hábito de usar vídeos em sala de aula com meus alunos, por esse ser um tipo de mídia com o qual tenho contato constantemente durante os meus próprios estudos do inglês, na forma de vídeos do YouTube, filmes e episódios de seriados. Vídeos são dinâmicos porque estimulam vários sentidos corporais e processos cognitivos simultaneamente: usa-se a audição e a visão para entender as informações, e é necessário interpretar rapidamente o sentido das falas enquanto se assiste. Assim, assistir a um vídeo em língua estrangeira é uma boa oportunidade para a sua prática. Com efeito, segundo Quevedo (1994), vídeos são recursos úteis para o processo de ensino-aprendizagem, pelo fato de apresentarem elementos da comunicação que não estão disponíveis por meio da escrita, como pronúncia, entonação de voz, gestos corporais, expressões e movimentos faciais.

Em busca de novas maneiras de se usar vídeos na sala de aula de inglês, elaborei uma ferramenta de *corpus on-line*: o *Corpus of English Language Videos* (CELV)[1], que permite consultas linguísticas em uma coleção de textos composta por arquivos de legenda extraídos de vídeos do YouTube, e é o tema central deste trabalho. O seguinte pensamento gerou a ideia inicial para este projeto: a aprendizagem de alunos de língua inglesa poderia ser enriquecida se fosse possível encontrar rapidamente vários trechos de vídeos que exemplificassem o uso de determinada palavra ou expressão, para que os alunos pudessem observá-la em uso. Se uma das estratégias frequentemente usadas por professores de línguas para demonstrar determinada palavra ou expressão é apresentar diversos exemplos de seu uso em frases escritas, melhor ainda seria mostrar segmentos de vídeos contendo a palavra ou expressão, pois assim se tem acesso a mais aspectos da comunicação, como pronúncia, entonação e expressões faciais.

No entanto, para trazer exemplos específicos em forma de vídeo para as aulas, seria necessário pesquisar vários vídeos (por exemplo, no YouTube) até encontrar um ou mais que contivessem as palavras ou expressões de interesse. Em seguida, se o interesse estivesse em apenas um trecho do vídeo, seria necessário navegar por ele até um momento específico, ou salvá-lo no computador e usar algum *software* de edição para cortá-lo, extraindo determinada cena ou passagem. Nem sempre é possível encontrar vídeos que correspondam ao conteúdo desejado, e nem sempre o processo de obtenção e edição de

---

[1] www.celvonline.com

vídeos está à disposição do professor. Essas limitações levaram à ideia de produzir uma ferramenta que fizesse exatamente isso: encontrar exemplos de palavras ou expressões da língua inglesa em vídeo, em grande quantidade, apresentadas de maneira a permitir o estudo de seu uso contextualizado.

A reunião de uma grande quantidade de exemplos de uso da língua inglesa foi viabilizada pela adoção da Linguística de *Corpus* (doravante, LC) como base metodológica deste trabalho. A LC é uma abordagem que se dedica à análise empírica de textos em grande quantidade e em formato eletrônico; os *corpora*, seu objeto de estudo, têm se demonstrado um recurso útil para informar o ensino e aprendizagem de línguas, auxiliando na criação de materiais didáticos empiricamente embasados e estudo da língua em uso (BERBER SARDINHA, 2004). Já o embasamento teórico para elaboração do conceito da ferramenta e de sugestões para o seu uso na aprendizagem de inglês foi buscado em estudos sobre a aplicação de *corpora* no ensino-aprendizagem de línguas, sobretudo na abordagem conhecida como Aprendizagem Direcionada por Dados (doravante, ADD), proposta por Johns (1991).

O uso da LC como metodologia deste trabalho foi essencial, porque uma das considerações a se fazer quando se discute o ensino de línguas é a importância do contato dos alunos com situações reais de comunicação. Gilmore (2004) compara a linguagem contida em atividades de alguns livros didáticos usados nas décadas de 80 e 90 com exemplos de situações comunicativas reais, demonstrando que o conteúdo linguístico analisado nesses livros é artificial e não apresenta várias das características da comunicação real, como pausas, quebras e hesitação na fala. Isso ocorre porque livros didáticos para ensino de línguas frequentemente simplificam sua linguagem com o objetivo de facilitar a compreensão dos alunos. Como destacam Moreira Filho (2007), Duarte (2011) e Acunzo (2012), essa prática pode ter consequências negativas, já que alguns alunos, mesmo tendo concluído seus estudos em cursos de inglês, podem não compreender textos reais quando os encontram.

Nesse sentido, uma das vantagens do uso de *corpora* no ensino-aprendizagem de línguas é o fato de que os textos que compõem a amostra de um *corpus* são autênticos; a autenticidade é, na verdade, um dos critérios mais importantes para compilação de *corpora* (BERBER SARDINHA, 2004). Vários tipos de texto autêntico podem ser usados para compilar um *corpus*, a depender do que se pretende analisar ou ensinar: textos de jornais e revistas, textos provindos de meios digitais como *blogs* e redes sociais, e até

mesmo programas televisivos por meio de transcrições da fala. Outro tipo de texto autêntico que tem sido usado recentemente no âmbito do nosso grupo de pesquisa em LC da Universidade Federal de Uberlândia para a construção e análise de *corpora* são legendas de vídeos, como nos trabalhos de Bang e Fromm (2013) e Peixoto (2014), que compilaram legendas de filmes e seriados com o intuito de realizar análises lexicais.

Nesta pesquisa, também utilizei legendas como fonte de dados linguísticos, mas desta vez para a compilação de uma amostra de vídeos do site YouTube com vistas à sua aplicação no ensino-aprendizagem de inglês. Para propósitos de ensino, considera-se que a linguagem contida nesses vídeos é autêntica, porque não foi produzida especificamente para fins pedagógicos, como define Peacock (1997). Ainda, pode-se acrescentar que:

> Materiais autênticos, particularmente os audiovisuais, oferecem uma fonte muito mais rica de estímulos para os aprendizes, e podem ser explorados de diferentes formas e em diferentes níveis para desenvolver a sua competência comunicativa. (GILMORE, 2007, p. 103)[2]

O uso de legendas de vídeos para compor o CELV possibilitou a criação de um novo tipo de recurso em *corpora on-line*, qual seja permitir o acesso aos vídeos originais por meio do clique sobre as linhas de concordância[3] criadas pela ferramenta, resultando em um *corpus* que pode exibir informação linguística tanto de forma escrita quanto audiovisual. A relevância desse recurso para o ensino-aprendizagem de inglês com uso de *corpora* está no acesso a informações linguísticas como a pronúncia e a entonação. *Corpora on-line* comumente apresentam informações linguísticas apenas de forma escrita, o que permite a observação de padrões lexicogramaticais, mas não de outros elementos importantes da comunicação, como os sons das palavras, gestos e expressões faciais. A possibilidade de observação e estudo de características sonoras da língua com uso de um *corpus on-line* representa uma expansão das possibilidades de uso dos *corpora* no ensino-aprendizagem de línguas.

O que pretendo com o CELV é uma forma direta de aplicação de *corpora* ao ensino de língua inglesa, pois foi concebido a fim de ser manuseado por professores e alunos de inglês seguindo a proposta da ADD, abordagem que enxerga o aluno como um

---

[2] Todas as traduções aqui contidas são de minha autoria. No original: Authentic materials, particularly audio-visual ones, offer a much richer source of input for learners and have the potential to be exploited in different ways and on different levels to develop learners' communicative competence.

[3] Linhas de concordância são fragmentos de texto extraídos de um *corpus*, e serão definidas detalhadamente no capítulo teórico desta dissertação.

pesquisador capaz de fazer suas próprias investigações linguísticas a partir da observação de linhas de concordância, em contato direto com um *corpus*, mediado pelo professor (JOHNS, 1991). Assim, *corpora* podem estar presentes na sala de aula como uma ferramenta de consulta, usada de forma semelhante à maneira como são usados os dicionários, com a vantagem de que os resultados das consultas a um *corpus* não são definições ou traduções, mas, sim, linhas de concordância. A partir delas, é possível observar padrões lexicogramaticais da língua, focando a atenção do aluno observador nas relações lexicais e semânticas que podem ser estabelecidas entre determinada palavra e as outras palavras em seu entorno.

Também existem formas indiretas de se aplicar dados provindos de *corpora* no ensino de línguas estrangeiras. Análises feitas a partir de *corpora* compostos por textos autênticos de língua nativa contribuem com o ensino ao revelarem padrões de uso que permitem a criação de materiais didáticos (gramáticas, dicionários e livros) de cunho descritivo, e não prescritivo – ou seja, embasados por evidências linguísticas empíricas, e não por intuição individual. Também podem ser compilados *corpora* a partir de textos de aprendizes, que revelam formas típicas de uso de uma língua estrangeira por alunos de diferentes nacionalidades, demonstrando possíveis influências de suas respectivas línguas maternas em sua produção linguística, o que ajuda a lidar com aspectos de aprendizagem específicos a diferentes contextos.

Sobre o uso de *corpora* nos estudos linguísticos, Sinclair (2004) afirmou que as evidências provindas das grandes coleções de textos em formato eletrônico foram ignoradas e consideradas irrelevantes por um quarto de século, apesar de sua grande importância. Desde a época dessa afirmação, observa-se uma tomada de consciência sobre a existência dos *corpora* e seus usos. Programas de análise lexical usados em estudos linguísticos com base em *corpora*, como o *WordSmith Tools* (SCOTT, 2016a) e o *AntConc* (ANTHONY, 2014), têm sido desenvolvidos, aprimorados e difundidos. No Brasil, a Linguística de *Corpus* tem ganhado visibilidade, fato evidenciado pela realização de eventos como o ELC/EBRALC[4], pela existência de grupos de estudo

[4] O Encontro de Linguística de *Corpus* e a Escola Brasileira de Linguística Computacional são eventos que acontecem juntos, bianualmente, e servem como palco para palestras, apresentações de trabalhos recentes e discussões sobre as duas áreas. O evento de 2014 foi realizado na UFU: www.elc-ebralc-2014.com.br.

dedicados à área[5] e pela publicação de periódicos com números dedicados ao tema[6]. Esse crescimento da área proporciona um contexto favorável ao desenvolvimento de pesquisas sobre seus vários aspectos, como a compilação de novos *corpora*, novas investigações linguísticas com base em *corpora* já existentes e até mesmo a criação de novas ferramentas computacionais.

Esta pesquisa trabalha com a hipótese de que o uso de uma ferramenta de *corpus on-line* com acesso a vídeos, ao possibilitar um contato empírico com a língua escrita e também oral, oferecerá ao aluno oportunidades de enriquecimento da sua aprendizagem do inglês. Para nortear o trabalho e investigar essa hipótese, busco responder às seguintes questões de pesquisa: como construir uma ferramenta de *corpus on-line* a partir de uma coleção de textos formada por arquivos de legenda? Como integrar a informação escrita dos arquivos de legenda aos vídeos originais selecionados do YouTube, possibilitando o acesso a informações linguísticas também em formato audiovisual? Como usar a ferramenta resultante desse processo no contexto de ensino-aprendizagem de inglês, considerando a abordagem da ADD?

Assim, o objetivo geral do trabalho é construir uma ferramenta de *corpus on-line* para busca textual em legendas de vídeos, concebida conforme os princípios da LC e da ADD, que traga novas possibilidades para o ensino de inglês com auxílio de *corpora*. Os objetivos específicos são: compilar um *corpus* a partir da extração de legendas de vídeos do YouTube, estabelecendo critérios para a seleção dos vídeos que comporão a amostra; desenvolver uma ferramenta de busca que permita diferentes tipos de consulta e apresentação dos dados do *corpus* em forma de linhas de concordância, e disponibilizar essa ferramenta na *internet*; e testar o funcionamento da ferramenta e levantar opiniões de professores de inglês sobre sua aplicação no ensino-aprendizagem da língua.

O capítulo 2 desta dissertação, denominado Fundamentação Teórica, será dividido em quatro subcapítulos. O subcapítulo 2.1 apresentará a Linguística de *Corpus*: seus aspectos teóricos, como a definição de *corpus*, a abordagem empírica à análise de dados linguísticos e a concepção de língua como sistema probabilístico, e seus aspectos práticos, como os critérios para compilação de *corpora*, tipologia de *corpus* e linhas de

---

[5] Alguns grupos de pesquisa em LC do Brasil: GPELC – UFU (http://gpelc.blogspot.com.br), COMET – USP (comet.fflch.usp.br), NILC – USP (www.nilc.icmc.usp.br), GELCORP-SUL – UFRGS (www.ufrgs.br/textecc), InCognito – UFMG (www.c-oral-brasil.org).

[6] Por exemplo, a revista Letras & Letras da UFU, cujo Volume 30, Número 2 do ano de 2014 foi dedicado à Linguística de *Corpus*: www.seer.ufu.br/index.php/letraseletras/issue/view/1217.

concordância. O subcapítulo 2.2 apresentará a Linguística Computacional e o Processamento de Linguagem Natural, demonstrando a parceria interdisciplinar que se estabelece entre as áreas da Linguística e da Computação para possibilitar o desenvolvimento de sistemas computacionais capazes de desempenhar tarefas de processamento linguístico como a análise lexical, tradução automática, etiquetagem e recuperação de informação. O subcapítulo 2.3 se dedicará a expor as formas diretas e indiretas de se aplicar *Corpora* no Ensino de Línguas Estrangeiras, incluindo, principalmente, a proposta da ADD. O subcapítulo 2.4 será reservado para a apresentação de algumas possibilidades recentes que têm sido exploradas em relação ao uso de vídeos e legendas como fonte de dados linguísticos para compilação de *corpora* e ensino-aprendizagem de línguas.

O capítulo 3, denominado Metodologia, explicará a trajetória metodológica percorrida para a realização do trabalho, começando pelas decisões iniciais do projeto e incluindo: (i) compilação do *corpus* a partir da extração de legendas de vídeos do YouTube, (ii) preparação e padronização dos arquivos de legenda por meio da formatação e etiquetagem morfossintática, (iii) indexação do *corpus*, (iv) desenvolvimento do sistema de busca e da interface do CELV na *internet* e (v) teste da ferramenta. As partes computacionais do trabalho foram realizadas em parceria com dois docentes e um discente do curso de Sistemas de Informação da Faculdade de Computação da Universidade Federal de Uberlândia, uma vez que eu não possuo conhecimentos de programação suficientes para o desenvolvimento de sistemas computacionais. Os detalhes desse processo de desenvolvimento em parceria com os programadores serão apresentados ao longo do texto.

O capítulo 4, Resultados, apresentará os três resultados principais deste trabalho. No subcapítulo 4.1, será descrito o *corpus* que foi compilado e disponibilizado para consulta. No subcapítulo 4.2, será descrita a interface da ferramenta de busca e de seu *site* na *internet*. No subcapítulo 4.3, serão apresentados e discutidos os dados obtidos durante o teste da ferramenta com professores de inglês. No subcapítulo 4.4, serão feitas sugestões para a aplicação do CELV em diferentes tópicos do ensino-aprendizagem de inglês.

O capítulo 5 conterá Considerações Finais sobre a pesquisa, retomando e rediscutindo a hipótese apresentada nesta introdução, demonstrando os objetivos que foram alcançados com este trabalho bem como as limitações do *corpus* e da ferramenta,

apontando direções futuras para a continuação do desenvolvimento do CELV e discorrendo sobre os efeitos positivos que este projeto surtiu na minha própria formação.

2. FUNDAMENTAÇÃO TEÓRICA

O capítulo teórico desta dissertação está dividido em quatro partes. Primeiramente, serão apresentados conceitos fundamentais da LC. Em seguida, será demonstrada a interdisciplinaridade necessária entre as áreas da Linguística e da Computação para o desenvolvimento de ferramentas computacionais de processamento de linguagem. A terceira parte detalhará as características da ADD, e pode ser considerada o ponto focal do trabalho, uma vez que essa abordagem foi o que embasou a concepção do CELV, enquanto a LC e a Linguística Computacional serviram como suportes práticos para a concretização da ideia. Por fim, na quarta e última parte da Fundamentação Teórica, serão apresentados alguns exemplos de trabalhos que conectam vídeos, legendas e *corpora* ao ensino de línguas e são dignos de menção por serem comparáveis ou similares a este trabalho.

Situar a LC e a Linguística Computacional no corpo deste texto exigiu certa reflexão, uma vez que algumas questões relacionadas a essas duas áreas poderiam ser apresentadas tanto no capítulo teórico quanto no capítulo metodológico. As contribuições de outros autores em relação a questões práticas do trabalho com *corpora* (diretrizes para compilação e etiquetagem de *corpora*, uso de *corpora on-line*, trabalho com linhas de concordância e procedimentos para processamento de linguagem) serão apresentadas no capítulo teórico, pois, embora descrevam procedimentos metodológicos, fazem parte da literatura das áreas relevantes para esta pesquisa. A forma como esses procedimentos metodológicos descritos foram aplicados à construção da ferramenta elaborada neste projeto, por sua vez, será apresentada no capítulo metodológico.

2.1 Linguística de *Corpus*

A citação a seguir tem sido frequentemente mencionada em trabalhos e apresentações acadêmicas para introduzir a LC. Por se tratar de uma descrição clara e objetiva da área, este trabalho mantém a tradição:

> A Linguística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. (BERBER SARDINHA, 2004, p. 3)

Berber Sardinha acrescenta, ainda, que um *corpus* é representativo; ou seja, seu conteúdo retrata uma amostra específica de língua, determinada pelos critérios estabelecidos durante a coleta dos textos que o compõem. A noção de representatividade do *corpus* é um conceito relativo; de forma geral, considera-se que quanto maior o *corpus*, mais representativo ele é. Contudo, também deve ser levada em consideração a variedade dos textos selecionados. Um *corpus* de tamanho grande e composto apenas por um tipo de texto é representativo apenas da variedade linguística desse tipo; portanto, conclusões retiradas a partir de análises desse *corpus* dizem respeito apenas àquela amostra de língua, e não podem ser estendidas a outros tipos de texto e outras variedades sem que se analise novas amostras.

Para se medir o tamanho de um *corpus*, usa-se um computador para contar seu número total de palavras. Nesse contexto, as palavras são chamadas de *tokens*, um termo computacional que é usado nos programas de análise lexical como o *WordSmith Tools* e o *AntConc*. Outra grandeza que pode ser medida em um *corpus* é seu número de *types*, as palavras distintas existentes na amostra, desconsiderando suas repetições (SCOTT, 2016b). Para exemplificação, observe-se a frase abaixo:

> A Linguística de *Corpus* é a área da Linguística
> responsável pela compilação e análise de *corpora*.

Essa frase possui 16 *tokens* (número total de palavras) e 13 *types* (palavras distintas). As palavras "a", "Linguística" e "de" aparecem duas vezes, cada uma; no entanto, são contadas apenas uma vez para se chegar ao número de *types*, pois suas repetições não são consideradas.

Dois fatores podem influenciar o número absoluto de *types* de um *corpus*: seu tamanho (quantidade de *tokens*) e sua variedade de textos. É interessante que um *corpus* compilado para ser usado no ensino-aprendizagem de línguas contenha uma grande variedade de *types* e de ocorrências de cada *type*, para que possa fornecer vários exemplos de uso da língua. A razão entre o número de *types* e *tokens*, ou *type/token ratio* (TTR), pode ser usada como uma indicação de sua densidade lexical. Ao se acrescentar mais textos a um *corpus*, observa-se que o aumento de seu número de *types* se torna cada vez menos significativo em relação ao aumento de seu número de *tokens*. (SCOTT, 2016b).

A contagem de *tokens* e *types* de um *corpus* serve para fornecer uma visão geral sobre o tamanho da amostra. A partir desta primeira observação quantitativa, é possível

considerar outras informações, como a frequência de ocorrência de cada palavra do *corpus*, padrões de co-ocorrência, fraseologismos, palavras-chave e exemplos contextualizados de uso de lexias. Essas informações, levantadas a partir de uma amostra autêntica de língua, permitem conclusões empiricamente embasadas sobre os fenômenos linguísticos.

2.1.1 A abordagem empírica à análise de dados linguísticos

A abordagem da LC à análise linguística é empírica, o que significa dizer que se dedica à exploração de dados reais, e não exemplos construídos artificialmente ou derivados da intuição. Sua forma de raciocínio é indutiva, e seu caminho metodológico é pautado pelo olhar sobre evidências observáveis. A análise científica empírica segue um processo de observação, hipotetização e verificação. Observam-se evidências sobre determinado fenômeno, formulam-se hipóteses sobre seu funcionamento, e verificam-se as hipóteses por meio de novas observações, repetindo o procedimento até a obtenção de conclusões com base nos padrões identificados (BORDAG, 2007).

Ao aplicar esse processo metodológico à análise linguística, a LC, por meio de ferramentas computacionais para a organização e processamento dos dados, ajuda a revelar fatos sobre a língua que seriam difíceis de se perceber por simples intuição. Ainda, cabe explicitar que a metodologia da LC atua "tanto de forma quantitativa, considerando informações sobre o número de ocorrências e co-ocorrências de padrões lexicogramaticais, quanto qualitativa, examinando os dados e propondo interpretações" (PEIXOTO, 2014, p. 161-162).

Historicamente, pode-se dizer que a abordagem da LC esteve em oposição à da Gramática Gerativa, sobretudo sua vertente transformacional, cujo berço foi o *Massachusetts Institute of Technology* (MIT), e que tem Noam Chomsky como expoente principal. Chomsky (1965) introduziu os conceitos de competência linguística e desempenho linguístico, que se referem, respectivamente, ao conhecimento mental intrínseco de um falante sobre a língua e suas regras, e ao uso real da língua em situações concretas. Enquanto a abordagem gerativa considera a língua pelo ponto de vista da competência, dedicando sua atenção àquilo que é linguisticamente possível dado um conjunto de regras gramaticais, a LC se pauta pelo desempenho, analisando a língua em uso por meio de seus fenômenos observáveis. Na visão de Chomsky, tem-se que:

> A teoria linguística se preocupa, principalmente, com um falante ideal, em uma comunidade de fala completamente homogênea, que conhece sua língua perfeitamente e não é afetado por condições gramaticalmente irrelevantes como limitações de memória, distrações, alterações de atenção e interesse e erros (aleatórios ou característicos) ao aplicar seu conhecimento da língua no desempenho real. (CHOMSKY, 1965, p. 3)[7]

Admite-se que o autor escreveu esse texto há 50 anos e novas ideias se desenvolveram no âmbito dos estudos linguísticos desde então. No entanto, a valorização do "falante ideal" inspirou várias gerações de linguistas e deu fomento ao nativismo linguístico, isto é, a concepção de que a língua e suas estruturas gramaticais estão presentes de maneira inata na mente e inseridas no código genético humano. Em contraste, Bordag (2007) explica que o empirismo está associado à rejeição de ideias inatas. No contexto dos estudos linguísticos, isso equivale a dizer que o conhecimento linguístico humano se desenvolve a partir da recepção de insumos pelos sentidos corporais, e não a partir de conhecimentos previamente existentes desde o nascimento.

Nesse sentido, a abordagem linguística empírica busca embasar seu estudo da língua em informações que devem ir além da intuição individual. Se o conhecimento linguístico de um único indivíduo é fruto de suas experiências pessoais, que diferem das dos demais, não se pode dizer que ele possui conhecimento da totalidade da língua. Assim, em busca de descrições linguísticas que pudessem alcançar maiores níveis de generalização, Sinclair (1991) propõe a criação de *corpora* monitores. Um *corpus* monitor é uma amostra a mais ampla e representativa possível da língua, que permita a obtenção de conclusões gerais sobre seu funcionamento. Quanto maior e mais representativa a amostra, mais relevantes serão as conclusões obtidas a partir dela.

Apesar da valorização de grandes quantidades de dados linguísticos, Sinclair (1991) afirma que os estudos com *corpora* não ignoram o valor da introspecção. A intuição individual do linguista deve ser usada, por exemplo, para avaliar ou conferir evidências, mas não para a criação de teorias sobre o funcionamento da língua sem o suporte de dados empíricos. Contar apenas com a intuição individual sem comprovação empírica ou apenas com evidências empíricas sem interpretação subjetiva são caminhos extremos que devem ser evitados, como colocado na passagem a seguir:

---

[7] No original: *Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shits of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance*.

> Uma separação nociva é visível na comunidade linguística entre aqueles que não se interessam por *corpora*, e veem as intuições pessoais como suficientes para a obtenção de dados, e as pessoas que chamei de fetichistas de *corpus*, que tratam todos os fatos como impuros e profanos a não ser que tenham vindo diretamente e sem edições de um *corpus*. Nos extremos, temos uma divisão entre sonhadores e contadores de *tokens* – de um lado, pessoas cujo trabalho consiste em especular sobre como princípios universais podem dar conta de nuances sutis de suas próprias reações intrínsecas a frases específicas, e, do outro lado, pessoas que pensam que contar os pronomes diferentes em dez milhões de palavras de texto e tabular os resultados é uma contribuição para a ciência. (PULLUM, 2009, p. 5)[8]

Considerando-se a polarização descrita por Pullum, o caminho metodológico ideal é aquele que leve em consideração tanto a intuição individual como as evidências empíricas, uma vez que ambos podem ser usados de maneira complementar para prover diferentes olhares em direção ao mesmo objeto de estudo, resultando em conclusões mais ricas do que se se partisse apenas de um ponto de vista.

Halliday (1991) explica que a gramática de uma determinada língua possui natureza probabilística, que pode ser observada por meio de técnicas com uso de *corpora*. Em outras palavras, a concepção linguística que subjaz a LC considera que a linguagem "é um sistema probabilístico, no qual certos traços são mais frequentes que outros" (SARDINHA, 2004, p. 23). Pode-se dizer que essa definição tomou forma a partir do momento em que foi possível observar a língua em uso com o auxílio de computadores, já que sua capacidade de processamento permite a rápida organização de dados linguísticos provenientes de grandes coleções de textos em formato eletrônico, viabilizando a observação da língua sob um ponto de vista estatístico. Conforme avanços tecnológicos possibilitaram a análise de quantidades maiores de dados e o desenvolvimento de novos tipos de ferramentas computacionais, essa natureza probabilística da língua ficou cada vez mais evidente.

Em decorrência da natureza probabilística da linguagem, observa-se que seu sistema é padronizado. Conforme explica Berber Sardinha (2004), isso é evidenciado por

---

[8] No original: *an unwholesome split is visible in the linguistics community between those who broadly want nothing to do with corpora and think personal intuitions are fine as a basis for data gathering, and the people that I have called corpus fetishists who treat all facts as unclean and unholy unless they come direct and unedited out of a corpus. At the extremes, we get a divide between dreamers and token-counters — on the one hand, people whose work consists in speculating on how universal principles might account for subtle shades of their own inner reactions to particular sentences, and on the other, people who think that counting the different pronouns in ten million words of text and tabulating the results is a contribution to science.*

análises embasadas em *corpora* que demonstram a recorrência não-aleatória de determinados padrões dentro do sistema linguístico. Essa padronização se dá sob a forma da repetição significativa de estruturas lexicogramaticais, ou, em outras palavras, a "atração" que determinadas palavras demonstram por outras. Esses fenômenos de co-ocorrência entre palavras são regulares, observáveis e comparáveis em diferentes *corpora*. Sinclair (1991) chamou essa característica da língua de princípio idiomático.

2.1.2 O princípio da livre escolha e o princípio idiomático

Sinclair (1991) explica que a produção linguística, seja ela oral ou escrita, é articulada por meio de dois movimentos: uma parte do que se produz linguisticamente consiste em palavras escolhidas livremente; a outra parte consiste em frases pré-construídas. Assim, pode-se dizer que um falante não é criativo o tempo todo durante sua produção linguística; ele também faz uso de estruturas linguísticas memorizadas, de natureza probabilística. O autor denomina de princípio da livre escolha a produção linguística sem o uso de estruturas pré-estabelecidas, e de princípio idiomático o uso regular de padrões linguísticos. Em suas palavras, o princípio idiomático significa que "um usuário da língua tem disponível para si um grande número de frases parcialmente pré-construídas, que constituem escolhas únicas, embora pareçam ser analisáveis como segmentos" (SINCLAIR, 1991, p. 110)[9].

Evidências baseadas em *corpora* demonstram que "o texto normal é amplamente deslexicalizado, e parece ser formado por exercício do princípio idiomático, com trocas ocasionais para o princípio da livre escolha" (SINCLAIR, 1991, p. 113)[10]. Erman e Warren (2009) realizaram análises em *corpora* escritos e falados, em busca da identificação e contagem de estruturas linguísticas pré-fabricadas (*prefabs*) que demonstrassem o uso do princípio idiomático. Em sua amostra, aproximadamente 55% da produção linguística se deu com o uso desse tipo de estrutura. Os autores ressaltam que identificar os *prefabs* é uma tarefa difícil, porque não são sempre estruturas totalmente pré-construídas, como, por exemplo, *of course*, que é uma frase fixa; também podem ser parcialmente pré-construídas, como a frase *a * of*, na qual apenas as palavras

---

[9] No original: *The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.*

[10] No original: *normal text is largely delexicalized, and appears to be formed by exercise of the idiom principle, with occasional switching to the open-choice principle.*

a e *of* são fixas, e o asterisco representa um espaço (*slot*) que pode ser preenchido por várias palavras. Assim, há uso do princípio da livre escolha dentro de expressões idiomáticas, o que indica que os dois princípios atuam de forma complementar.

Segundo Viana (2010), o princípio idiomático ocorre na forma de quatro fenômenos: colocação, coligação, preferência semântica e prosódia semântica. A colocação é a co-ocorrência estatisticamente significativa de palavras. Como exemplo desse fenômeno, o autor cita a palavra *surprising*, que frequentemente ocorre junto com a palavra *hardly*, formando a lexia *hardly surprising*. A coligação é semelhante à colocação, porém considera também atributos gramaticais das palavras que se associam, como diferentes posições em que uma palavra pode estar situada em uma frase. A palavra *flexible*, por exemplo, ao se associar com substantivos abstratos, costuma se posicionar antes desses substantivos; no entanto, ao se associar com sujeitos animados, costuma se posicionar após as formas nominais. A preferência semântica é a tendência de determinadas palavras se associarem a outras lexias de campos semânticos específicos. Viana (*op. cit.*) cita como exemplo a palavra *completely*, que costuma se associar a palavras que indicam ausência (*disappeared* ou *empty*) e a palavras que indicam mudança (*changed* ou *destroyed*). A prosódia semântica, por fim, considera o valor afetivo que determinada palavra assume em relação aos seus colocados. Ou seja: certas palavras frequentemente se associam a outras palavras que possuem um valor positivo, negativo ou neutro. A palavra *utterly*, por exemplo, apresenta prosódia semântica negativa, pois costuma se associar a palavras que têm valor negativo, como *burned* ou *confused*. A Figura 1, a seguir, esquematiza esses princípios da produção linguística.

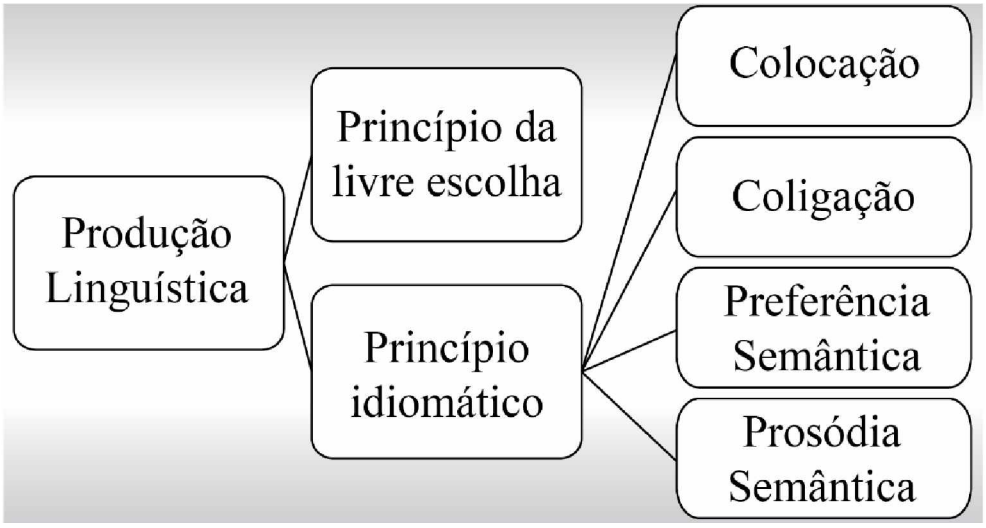


Figura 1: princípios da produção linguística conforme Sinclair (1991).
Fonte: CARNEIRO, 2013, p. 3.

Carneiro (2013) explica que os quatro fenômenos do princípio idiomático se situam em diferentes níveis linguísticos: a colocação, no nível lexical; a coligação, no nível lexicogramatical; a preferência semântica, no nível semântico; e a prosódia semântica, no nível pragmático. O estudo de estruturas linguísticas que correspondam ao princípio idiomático interessa à Linguística de *Corpus* porque permite a identificação das regularidades que formam, como visto, grande parte da produção linguística. Por serem frequentes, esses padrões podem ser observados com o auxílio de ferramentas computacionais, com o objetivo de alcançar conclusões sobre o uso que é feito da língua por seus falantes, o que possibilita a realização de estudos descritivos empiricamente embasados e possui aplicações pedagógicas que serão apresentadas no subcapítulo 2.3 deste texto, intitulado *Corpora* no ensino-aprendizagem de línguas.

2.1.3 Compilação e tipologia de *corpus*

O principal fator determinante das características de um *corpus* e, consequentemente, dos estudos linguísticos que poderão ser feitos a partir dele, são os critérios utilizados para a seleção dos tipos de texto que comporão a amostra. Essa fase de construção do *corpus* é chamada de compilação. Nesse momento, é importante ter em mente qual será o propósito do *corpus*, e é preciso encontrar formas de obter os textos que o comporão. Uma das maneiras mais práticas de se obter textos é por meio do *download* na *internet*. No entanto, alguns estudos linguísticos exigem a obtenção de textos que não estão tão facilmente disponíveis. Beilke (2014), por exemplo, compilou textos de uma variante da língua alemã, o pomerano, que é falado em algumas localidades do Brasil. Segundo a autora, para a coleta de textos, foi necessário buscar exemplos dessa variante linguística em várias fontes diferentes, incluindo visitas pessoais a comunidades pomeranas em diversos lugares do país. Segundo Beilke, há "dificuldade de encontrar textos escritos em pomerano, embora atualmente [seu] banco de dados conte com 175.545 *tokens*, o que para a realidade linguística da variedade pomerana é bastante significativo" (BEILKE, 2014, p. 188).

Assim, é necessário relativizar a noção de representatividade de um *corpus* conforme as circunstâncias de sua coleta e a amostra disponível. O número de *tokens* obtido por Beilke (*op. cit.*), por exemplo, foi suficiente para seus fins de pesquisa. Mesmo assim, o tamanho da amostra é de vital importância, pois determina a quantidade de informação linguística disponível para análise. Por isso, ao se compilar *corpora*, deve-se

buscar o maior número de *tokens* possível, dentro das limitações da pesquisa. Contudo, além do critério da quantidade, Viana (2010) explica que também devem ser consideradas as noções de diversidade e equilíbrio, especialmente para *corpora* que objetivam servir como amostras gerais de língua. Diversidade, segundo o autor, diz respeito à variedade de textos da amostra. É importante não focar apenas em um tipo de texto, mas, sim, incluir "uma ampla gama de gêneros discursivos, contextos de produção, participantes (...), entre outros" (VIANA, 2010, p. 28). Equilíbrio, por sua vez, se refere à distribuição quantitativa desses diferentes tipos de texto na amostra. Se determinado tipo de texto prevalece quantitativamente sobre outros, os resultados das análises linguísticas provenientes dessa amostra serão mais relacionados a esse tipo, e dirão pouco sobre os demais.

Cabe, neste momento, apresentar uma definição mais detalhada de *corpus*. Como visto, trata-se de uma coleção de textos em formato eletrônico, compilada criteriosamente para fins de estudo linguístico. Mais especificamente, um *corpus* deve possuir quatro características fundamentais (McENERY, WILSON, 1996, *apud* ALUÍSIO, ALMEIDA, 2006), a saber: (i) amostragem e representatividade, ou seja, um tamanho de amostra que seja suficiente para representar a variedade linguística que se pretende estudar; (ii) tamanho finito, por exemplo, 1 milhão ou 10 milhões de palavras (com exceção dos *corpora* monitores); (iii) formato eletrônico, que permite que a amostra seja analisada rapidamente e também acrescida de novos dados/informações; e (iv) referência padrão, isto é: um *corpus* pode servir como referência para diversos estudos sobre a variedade linguística que representa, podendo ser reutilizado várias vezes se disponibilizado para a comunidade científica. As três primeiras características dessa definição já haviam sido mencionadas neste texto. A quarta característica, no entanto, introduz um importante aspecto dos *corpora*:

> Entende-se que disponibilização de *corpus* compilado para futuras pesquisas é uma característica inerente ao *corpus*, de forma que todo o esforço empreendido para a sua construção não seja útil apenas para uma pesquisa, uma vez que se tem uma referência padrão de língua ou de variedade de língua que pode ser utilizada por outros pesquisadores (ALUÍSIO, ALMEIDA, 2006, p. 158).

Para se obter um *corpus* com as quatro características mencionadas, Aluísio e Almeida (*op. cit.*) sugerem três etapas metodológicas. Primeiramente, é necessário projetar o *corpus*, o que significa estabelecer previamente os fatores que definem a sua tipologia. Berber Sardinha (2004) propõe sete critérios para descrever a tipologia de um

corpus: modo (falado, escrito ou ambos), tempo (sincrônico ou diacrônico, contemporâneo ou histórico), seleção (monitor, de amostragem, dinâmico, estático, equilibrado), conteúdo (especializado, regional, dialetal, monolíngue, bilíngue, multilíngue), autoria (de aprendiz ou de língua nativa), disposição interna (paralelo ou alinhado), e finalidade (de estudo, de referência ou de treinamento). Assim, durante o projeto do *corpus*, é necessário especificar quais serão suas características e fazer previsões sobre o resultado desejado com a sua compilação.

Estabelecidos esses critérios, e segunda etapa é a compilação e manipulação do *corpus*. Nesse momento, buscam-se maneiras de se obter textos para a amostra, seja por pesquisa na *internet*, digitalização de textos impressos, transcrição de fala ou outros meios. Uma vez coletados os textos, eles devem ser manipulados para permitir sua leitura por computador, o que significa sua conversão para arquivos em formato .txt e também a limpeza e a formatação do texto, retirando informações desnecessárias. A decisão sobre as informações que devem ser mantidas ou descartadas depende do tipo de estudo que se pretende realizar. Finalmente, após a formatação, os arquivos devem ser armazenados em diretórios do computador e nomeados. Aluísio e Almeida (*op. cit.*) recomendam a criação de uma nomenclatura padronizada para todos os arquivos a fim de organizar a amostra e facilitar a busca por textos específicos.

A terceira e última etapa sugerida pelas autoras é a anotação, que é a inserção de marcações ou etiquetas no texto de forma a facilitar a identificação de determinadas informações, e pode ser estrutural ou linguística. A anotação estrutural é a inserção de informações relacionadas a autoria, origem, tamanho do arquivo e data de coleta, e também a categorização de elementos textuais que podem vir a ser objeto de estudo, como títulos, parágrafos, sentenças, palavras, nomes próprios e outros. A anotação linguística, por sua vez, é a marcação de categorias linguísticas, como classes gramaticais (anotação morfossintática) ou semânticas (anotação semântica). Um *corpus* com anotação linguística possui uma etiqueta atribuída a cada uma das palavras de sua amostra. Em *corpora* de grande tamanho, é de se esperar que a inserção manual dessas etiquetas seja inviável; no entanto, com o auxílio de ferramentas de Processamento de Linguagem Natural, a etiquetagem pode ser feita de forma automática ou semiautomática. Uma dessas ferramentas, o etiquetador morfossintático CLAWS (GARSIDE, 1996), será apresentada mais detalhadamente no subcapítulo 2.2 deste texto, intitulado Linguística Computacional e Processamento de Linguagem Natural.

2.1.4 O *Corpus of Contemporary American English*

O *Corpus of Contemporary American English* (COCA), desenvolvido por Davies (2008-), é um exemplo de *corpus* monitor do inglês dos Estados Unidos. Isso significa que seu objetivo é servir como uma amostra próxima do que seria a língua inglesa norte-americana "geral", possibilitando uma grande variedade de estudos linguísticos. Para servir a tal propósito, esse *corpus* possui mais de 520 milhões de palavras, igualmente distribuídas entre cinco gêneros textuais: língua falada, ficção, revistas, jornais e textos acadêmicos. Sua amostra também está balanceada conforme o ano de produção dos textos, totalizando 20 milhões de palavras por ano, com início em 1990 até 2015, e sendo expandido regularmente com textos mais recentes. O *corpus* está disponível para consulta por meio de uma ferramenta *on-line*[11] que possui várias funções avançadas de pesquisa, como filtros por gênero, subgênero e ano, e buscas por classes gramaticais e sinônimos de palavras. Outras funções incluem: busca por lemas (exemplo: a consulta [*break*] encontra esta palavra e suas inflexões, como *breaks*, *breaking*, *broke* e *broken*), busca com uso de etiquetas gramaticais (exemplo: a consulta [v*] *the door* encontra sintagmas formados por qualquer verbo + *the door*) e busca com uso de coringas ou *wildcards* (exemplo: a consulta *dis*ed* encontra palavras como *discovered*, *discussed* e *disappeared*, e a consulta * *the internet* encontra *on the internet*, *of the internet*, *over the internet*, etc.). Várias dessas funções também podem ser encontradas em outros *corpora on-line*, embora esses não sejam tão amplos como o COCA.

Em 2016, o COCA ganhou uma nova interface de busca. De acordo com Davies (2008-), a interface anterior possuía uma quantidade excessiva de quadros (*frames*), o que dificultava a navegação, especialmente em dispositivos móveis como *tablets* e *smartphones*. Portanto, a nova versão da página retirou os quadros, e agora os usuários podem consultar o *corpus* com mais praticidade. Além disso, houve uma simplificação da interface, por meio do ocultamento dos recursos mais avançados de busca. Como mencionado anteriormente, o COCA possui uma grande quantidade de ferramentas e funções; no entanto, nem sempre o pesquisador está interessado em usar todos esses recursos, e, por isso, não é necessário que eles apareçam na tela a todo instante, bastando que estejam acessíveis quando forem necessários.

Na minha opinião, esse tipo de atualização da interface das ferramentas de *corpora on-line* é uma medida importante, porque um dos fatores que interferem

---

[11] corpus.byu.edu/coca/

negativamente no acesso aos *corpora* por pesquisadores e alunos é a dificuldade técnica ao lidar com mecanismos complicados e com excesso de informações. A nova interface do COCA está mais próxima da encontrada em mecanismos de busca usados popularmente, como o Google, o que favorece a introdução de novos usuários aos trabalhos com *corpora*.

Esta descrição resumida do COCA é importante para este trabalho porque a interface e algumas das funções principais desse *corpus* foram a principal referência para a construção do CELV. Considerando-se que o COCA é um dos maiores e mais usados *corpora* disponíveis gratuitamente na *internet*, procurei utilizá-lo como base para decidir quais funções de busca incluir na ferramenta que permite buscas no CELV, e também como organizar a estrutura interna do *corpus* desenvolvido neste trabalho, dentro das limitações de tempo, recursos e quantidade de textos disponíveis para a realização da pesquisa. Mais detalhes sobre a influência do COCA no desenvolvimento do CELV serão apresentados no capítulo metodológico deste texto.

2.1.5 Linhas de concordância e concordanciadores

Grande parte dos estudos linguísticos com uso de *corpora* são realizados a partir da análise de linhas de concordância, que são extratos de texto retirados de uma amostra a partir de determinada palavra de busca e listados em uma tela de computador. Nesse formato de exibição, as linhas são centralizadas na palavra de busca, o que permite a observação das outras palavras em seu entorno, como pode ser observado na Figura 2.

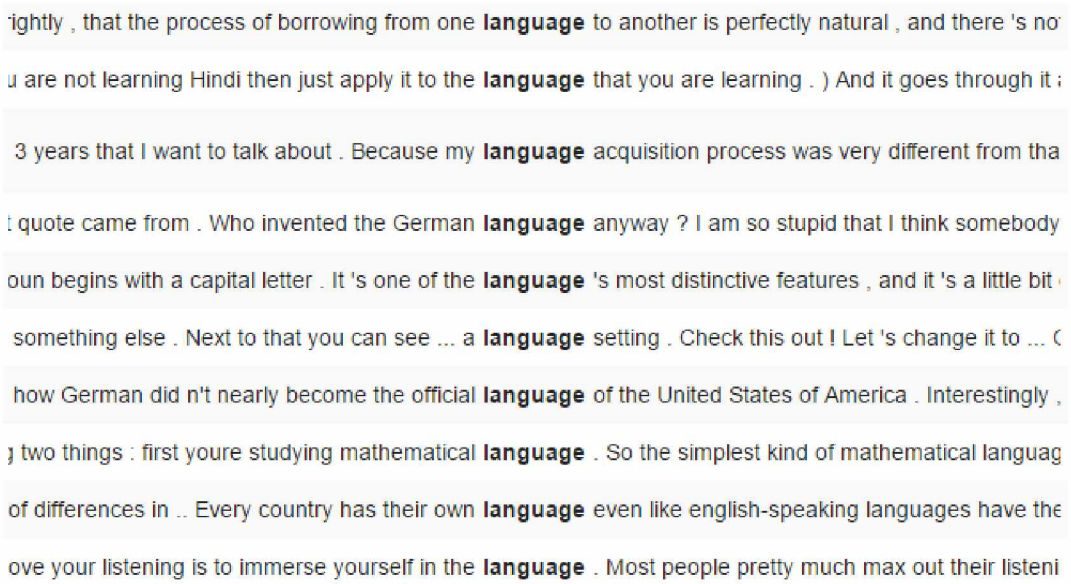


Figura 2: dez exemplos de linhas de concordância com a palavra *language*.
Fonte: captura de tela do CELV.

Como se vê pela figura, linhas de concordância apresentam a língua de uma maneira diferente da habitual, encontrada comumente nos textos. Por isso, para chegar a conclusões a respeito do funcionamento de determinada palavra ou expressão a partir da observação dessas linhas (usando sempre, também, a introspecção e análise subjetiva do pesquisador), é essencial se familiarizar com esse modo de exibição e compreender quais informações é possível obter a partir de sua análise, como explicado por Viana (2010):

> O importante a ser ressaltado na exploração [de linhas de concordância] é que não deve ser empregada uma leitura estritamente linear da esquerda para a direita, da primeira para a última linha. Esse procedimento é empregado para a compreensão de ideias, o que não é o foco nem o objetivo da observação de linhas de concordância. Em substituição, deve-se aproveitar a disposição das linhas de concordância [...] para verificar como a palavra [...] é empregada [...]. Com o posicionamento da palavra de busca no centro da linha, o pesquisador ter a possibilidade de se concentrar na observação dos exemplos em busca de algum tipo de padronização. A leitura nesse caso deve ser iniciada justamente pela palavra em posição central, verificando as palavras empregadas à esquerda e à direita (VIANA, 2010, p. 73-74).

Assim, linhas de concordância devem ser lidas do centro para as extremidades, observando as palavras que ocorrem no entorno da palavra de busca, procurando encontrar formas de uso de palavras específicas, e não necessariamente compreender o sentido de cada frase. Observando-se várias dessas linhas em uma tela de concordância, é possível perceber que emergem padrões lexicogramaticais que ajudam a descrever como determinada palavra é usada. A percepção desses padrões, facilitada por essa forma de exibição, seria mais difícil por meio de uma leitura tradicional. Pode-se dizer, então, que linhas de concordância são uma maneira alternativa de se exibir informação linguística, e possuem a vantagem de permitir a rápida observação de padrões de uso de determinadas palavras, embora retirem o foco sobre a compreensão das ideias do texto.

Sistemas computacionais capazes de gerar listas de linhas de concordância a partir de um determinado *corpus* e uma ou mais palavras de busca são chamados concordanciadores. Uma das preocupações principais deste trabalho é o uso pedagógico de concordanciadores, ou seja, seu manuseio por professores e alunos tanto dentro quanto fora do ambiente da sala de aula, com o objetivo de observar exemplos autênticos da língua inglesa a fim de aprender sobre o uso de palavras e expressões em inglês. Johns (1991) afirma que o concordanciador é a ferramenta mais importante para esse tipo de

abordagem de *corpora* no contexto de ensino-aprendizagem de línguas estrangeiras. O autor explica que o formato mais comum para a exibição das informações linguísticas em um concordanciador é o KWIC, que significa *keyword-in-context*, ou palavra-chave em contexto; esse formato nada mais é do que a já mencionada lista de linhas de concordância centralizadas em uma palavra de busca e apresentando certa quantidade de caracteres/palavras à esquerda e à direita. Viana (2010) tece alguns comentários sobre essa nomenclatura, considerando-se o estado atual dos estudos em LC. Segundo o autor, KWIC é um termo cristalizado na literatura sobre LC, embora atualmente seja mais apropriado se pensar em "palavras de busca", e não "palavras-chave", e "cotexto[12]" ao invés de "contexto", quando se está lidando com linhas de concordância.

Segundo Johns (1991), além da listagem de linhas de concordância no formato KWIC, outro recurso útil em concordanciadores é a possibilidade de ordenar e reordenar as linhas alfabeticamente, a partir das palavras à esquerda ou à direita da central, em determinada posição. Assim, é comum encontrar uma função chamada de *sort* em programas de análise lexical e em ferramentas de *corpora on-line*, que oferece opções como L1 (ordenação alfabética das palavras posicionadas 1 espaço à esquerda da central), L2 (ordenação alfabética das palavras posicionadas 2 espaços à esquerda da central), R1, R2 (similares às funções L1 e L2, mas para a direita), e assim sucessivamente. Para exemplificação, observe-se as opções da função *Concordance Sort* do *WordSmith Tools* e o resultado da ordenação de linhas de concordância, nas figuras a seguir:

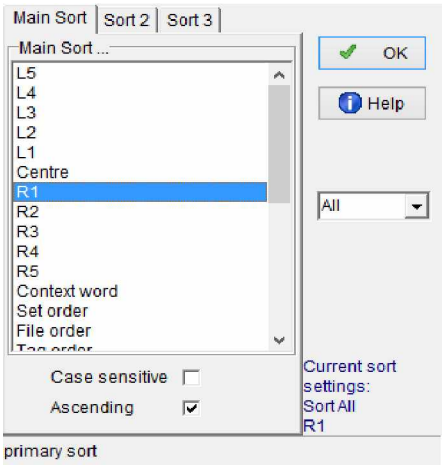


Figura 3: opções da função *Concordance Sort* do *WordSmith Tools*.
Fonte: captura de tela do *WordSmith Tools*.

[12] Na literatura sobre LC, inclusive em Viana (2010), entende-se por "cotexto" as palavras que ocorrem imediatamente ao redor da palavra de interesse. Por exemplo: as 4 palavras à sua esquerda e as 4 palavras à sua direita. A quantidade de palavras à esquerda e à direita representa o horizonte do cotexto, e pode variar de acordo com o fenômeno linguístico que se deseja estudar.

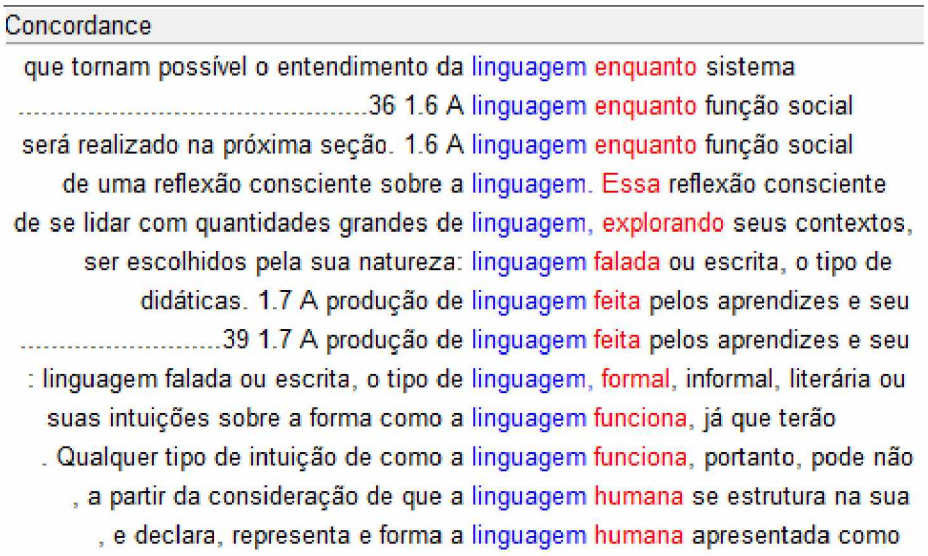


Figura 4: exemplos de concordâncias ordenadas alfabeticamente pela função *sort*.
Fonte: captura de tela do *WordSmith Tools*.

Como explicitado, o concordanciador assume papel central nas análises linguísticas que usam como base metodológica a LC. Em termos de aprendizagem de línguas com uso de *corpora*, as linhas de concordância são o principal material usado pelos aprendizes para interagir com amostras de uso da língua, como será demonstrado no subcapítulo 2.3. O capítulo metodológico desta pesquisa explicará como foi elaborado o concordanciador do CELV e suas funcionalidades.

2.2 Linguística Computacional e Processamento de Linguagem Natural

Segundo Biemann (2007), a Linguística Computacional e o Processamento de Linguagem Natural (doravante, PLN) são áreas distintas que possuem o mesmo objeto de estudo: dados linguísticos armazenados em formato eletrônico. Embora distintas, as duas áreas são complementares e pertencem a um contexto interdisciplinar entre a Linguística e a Computação, de modo que pesquisadores desses dois campos de estudo podem trabalhar cooperativamente para atingir objetivos em comum. Com efeito, segundo Fromm (2006, p. 135), "embora [as duas áreas], à primeira vista, se configurem como ciências em campos díspares (humanas e exatas), desde o final do século XX e, especialmente a partir do século XXI, têm trabalhado juntas para o aprimoramento de ambas". O autor menciona, ainda, algumas das principais contribuições que a Computação traz aos estudos linguísticos: obras lexicográficas podem ser elaboradas a partir de grandes coleções de texto em formato eletrônico, e, portanto, embasadas em exemplos concretos e não inventados; a própria estrutura dos dicionários recebe novos

recursos, como a navegação entre verbetes por meio de *hyperlinks*, aumentando a velocidade de consulta e trazendo novas possibilidades de uso; *corpora* também são usados no treinamento de diversos tipos de sistemas computacionais para processamento de língua, como os corretores ortográficos. Outras aplicações, segundo Halliday (2005), são a construção de sistemas de tradução automática de línguas, etiquetadores e ferramentas para extração de termos.

Portanto, a Linguística Computacional trabalha com dados linguísticos do ponto de vista de um especialista em Linguística, preocupando-se com a solução de questões linguísticas com uso de ferramentas computacionais; já o PLN pode ser entendido pelo ponto de vista de um especialista em computação, que se encarrega do desenvolvimento dessas ferramentas computacionais. Para um especialista em Linguística, um *corpus* é uma coleção de dados linguísticos que serve como objeto de estudo, e para um especialista em Computação, é uma amostra de dados para a construção de sistemas computacionais. Essa parceria interdisciplinar existente entre as duas áreas aparece ilustrada na Figura 5.

Figura 5: a interdisciplinaridade do trabalho com *corpora*.
Fonte: elaboração própria.

Este trabalho se interessa por um tipo específico de sistema de PLN: etiquetagem morfossintática automática de textos em língua inglesa em formato eletrônico. Como previamente mencionado, com base em Aluísio e Almeida (2006), a anotação é uma das etapas da construção de um *corpus*, e a etiquetagem morfossintática é um tipo de anotação que atribui categorias gramaticais a cada palavra de uma determinada porção de texto. Essa categorização gramatical é importante para ferramentas de *corpora on-line* porque,

sem ela, as possibilidades de consulta ao *corpus* disponibilizado seriam limitadas a pesquisas simples, como por palavras ou frases específicas (por exemplo: *house* ou *go to the cinema*), ou, no máximo, por frases contendo alguma variação de palavras (por exemplo: *go to the* *, onde o asterisco simboliza qualquer palavra). A etiquetagem morfossintática permite implementar no sistema de busca a possibilidade de consultas como *[v*] to the [nn*]*, onde os parâmetros entre colchetes simbolizam qualquer verbo e qualquer substantivo, respectivamente. Portanto, o motivo de se realizar a etiquetagem morfossintática de um *corpus* é possibilitar consultas específicas em relação às classes gramaticais das palavras contidas na amostra.

O sistema de etiquetagem morfossintática escolhido para etiquetar o CELV foi o CLAWS, porque é capaz de etiquetar amostras de texto escrito em língua inglesa com precisão a partir de 95% (ou seja, em média, 95% das palavras do texto serão etiquetadas corretamente), tendo sido usado, também, para etiquetar o COCA. As características desse sistema serão detalhadas na seção seguinte.

2.2.1 O etiquetador CLAWS

CLAWS[13] significa *Constituent Likelihood Automatic Word-tagging System*, e é um sistema de etiquetagem morfossintática (em inglês, *part-of-speech tagging*) desenvolvido na Universidade de Lancaster pelo centro de pesquisa UCREL[14] (*University Centre for Computer Corpus Research on Language*).

O CLAWS possui uma versão disponível na *internet* para uso gratuito[15]. Nessa versão, é possível inserir, no máximo, 100.000 palavras para etiquetagem. Portanto, caso se deseje usar a ferramenta para etiquetar um *corpus* de tamanho maior do que esse, é necessário dividir a amostra em partes menores do que 100.000 palavras e inserir o texto múltiplas vezes até que se complete a etiquetagem de todo o texto.

Segundo Garside (1996), a etiquetagem aplicada pelo CLAWS segue seis etapas, que podem ser resumidas da seguinte maneira:

1. O usuário insere um texto em inglês e o sistema processa o texto inserido, reconhecendo as palavras distintas (*tokens*);

---

[13] ucrel.lancs.ac.uk/claws/
[14] ucrel.lancs.ac.uk/
[15] ucrel.lancs.ac.uk/claws/trial.html

2. Atribui-se uma lista de etiquetas possíveis a cada palavra do texto. As escolhas de etiquetas que podem ser atribuídas a cada palavra são retiradas de uma lista (*lexicon*) que contém um grande número de palavras associadas às suas respectivas classificações gramaticais;
3. Para palavras que não estejam contidas no *lexicon*, o sistema segue um conjunto de regras pré-estabelecidas para determinar etiquetas aplicáveis;
4. A partir da análise do cotexto ao redor de cada palavra do texto, o sistema ajusta as listas de etiquetas possíveis atribuídas nas etapas 2 e 3, por meio da comparação com uma biblioteca de padrões lexicogramaticais previamente construída;
5. Com base em dados estatísticos, o sistema calcula a probabilidade de cada combinação de etiquetas em uma dada sequência de palavras, e seleciona a combinação mais provável;
6. Cada palavra contida no texto inicialmente inserido recebe uma etiqueta gramatical escolhida pelo sistema, e o texto é retornado ao usuário.

Cada uma dessas etapas envolve processos computacionais, cálculos estatísticos e algoritmos cuja descrição detalhada foge do escopo deste trabalho. Resumidamente, o sistema é executado a partir da etapa 1, com a inserção de dados linguísticos pelo usuário. Nas etapas 2 e 3, o sistema considera as palavras do texto isoladamente, listando as etiquetas possíveis para cada palavra. De acordo com Garside (1996), nesse momento do processo a etiquetagem está ambígua, pois há mais de uma etiqueta atribuída a cada palavra. O objetivo final do processo é escolher uma única etiqueta para cada palavra, o que é feito nas etapas 4 e 5 por meio de um procedimento de desambiguação, no qual o sistema considera o cotexto ao redor de cada palavra para decidir a combinação de etiquetas mais provável. As etiquetas mais prováveis são aceitas como corretas e retornadas ao usuário, associadas a cada palavra do texto inicialmente inserido. O autor explica que esse processo tem uma precisão de, aproximadamente, 95%, o que varia de acordo com o tipo de texto inserido.

2.3 *Corpora* no ensino-aprendizagem de línguas

Desde o início deste texto, têm sido feitas referências sobre o potencial dos estudos baseados em *corpora* para o ensino-aprendizagem de línguas. Esta seção irá se dedicar a

explicar mais detalhadamente essas aplicações, e apresentar as principais contribuições da literatura sobre o tema. Este trabalho está focado no ensino de língua inglesa; no entanto, as questões aqui discutidas podem ser aplicadas, também, a outras línguas.

De acordo com Römer (2008, p. 112):

> Durante as duas últimas décadas, *corpora* (ou seja, grandes coleções sistematizadas de língua escrita e/ou falada, armazenadas em computador e usadas em análises linguísticas) e evidências de *corpus* têm sido usados não somente na pesquisa linguística, mas também no ensino-aprendizagem de línguas [...]. Existe, agora, uma grande variedade de materiais de referência completamente baseados em *corpus* (como dicionários e gramáticas) disponíveis para aprendizes e professores, e vários pesquisadores e professores têm feito sugestões concretas sobre como concordâncias e exercícios derivados de *corpus* podem ser usados na sala de aula de ensino de línguas [...].[16]

As palavras da autora apontam para o primeiro aspecto que precisa ser apresentado quando se discutem usos pedagógicos de *corpora*: o fato de que existem aplicações indiretas (elaboração de materiais de referência) e diretas (uso de concordâncias na sala de aula) de *corpus* no contexto do ensino-aprendizagem de línguas. Para ilustrar essas duas formas de aplicação, a autora propõe o seguinte esquema:

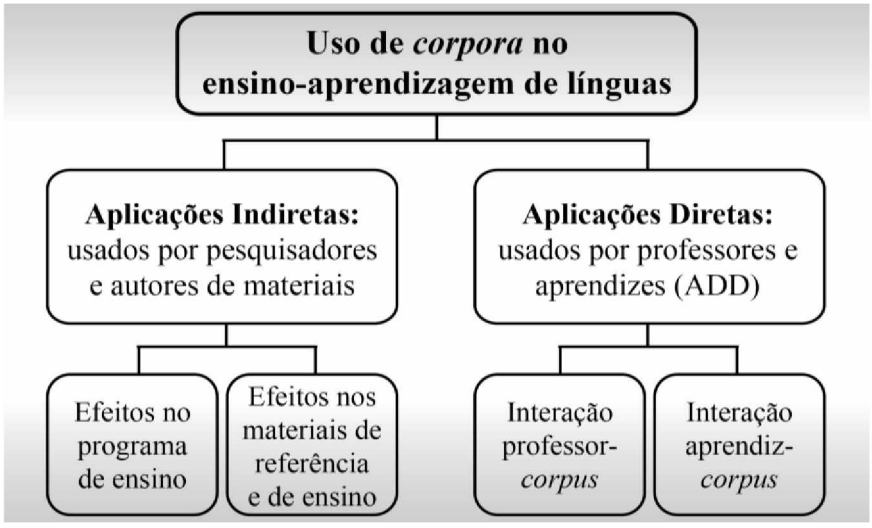


Figura 6: tipos de aplicação pedagógica de *corpus*.
Fonte: traduzido de Römer, 2008, p. 113 (original: Anexo 2).

[16] No original: *Over the past two decades, corpora (i. e. large systematic collections of written and/or spoken language stored on a computer and used in linguistic analysis) and corpus evidence have not only been used in linguistic research but also in the teaching and learning of languages [...]. There is now a wide range of fully corpus-based reference works (such as dictionaries and grammars) available to learners and teachers, and a number of dedicated researchers and teachers have made concrete suggestions on how concordances and corpus-derived exercises could be used in the language teaching classroom [...].*

Como demonstrado pela figura, as aplicações pedagógicas indiretas de *corpora* geram dois tipos de efeito: (i) no programa de ensino, já que os dados empiricamente embasados por *corpora* permitem a escolha de conteúdos linguísticos relevantes para o planejamento de cursos de línguas, a partir de informações sobre a frequência de ocorrência e co-ocorrência de palavras na língua e, também, sobre os problemas mais comuns enfrentados por aprendizes de línguas, o que pode ser analisado a partir de *corpora* de aprendizes; e (ii) nos materiais de referência e ensino, já que, a partir de exemplos de língua autêntica retirada dos *corpora*, é possível elaborar dicionários, gramáticas e livros didáticos que proporcionem aos aprendizes amostras reais da língua, ao invés de exemplos artificialmente construídos. Considera-se que uma das maiores contribuições nessas duas frentes (*corpora* na elaboração de programas e materiais de ensino) foi o projeto COBUILD (SINCLAIR, 1987), com a compilação e análise do *Collins Corpus*, que, por sua vez, levou à criação do *Bank of English*. Essas coleções de texto levaram à elaboração de vários materiais de referência utilizados até os dias de hoje.

O motivo para se apresentar as aplicações pedagógicas de *corpora* começando pelas indiretas antes das diretas é histórico. Ainda segundo Römer (2008), a LC, quando de seu surgimento[17] na segunda metade do século XX, preocupava-se, principalmente, com a pesquisa linguística. Os pesquisadores de *corpus* da época tinham o objetivo de aplicar as informações linguísticas provenientes de *corpora* à criação de obras lexicográficas, o que trouxe todos os benefícios previamente mencionados para a elaboração de materiais de referência voltados para o ensino de línguas.

No entanto, pesquisadores preocupados com a busca de aplicações mais diretas afirmam que ainda não há trabalhos e propostas suficientes para trazer as vantagens dos *corpora* para a sala de aula; Römer (2006, p. 121) afirma que "apesar do progresso que tem sido feito na área de linguística de *corpus* e ensino e línguas, a prática do ensino de língua inglesa, até o momento, não foi amplamente afetada pelos avanços da pesquisa com *corpus*"[18], e Meunier (2011, p. 461) afirma que "há, portanto, uma clara divisão entre

---

[17] Aqui, entende-se que a LC não surgiu, exatamente, na segunda metade do século XX, uma vez que linguistas já compilavam coleções de texto manualmente para análise linguística antes disso. A menção ao "surgimento" da LC feita no texto diz respeito ao estado da área como a conhecemos hoje, ou seja, auxiliada por ferramentas computacionais, permitindo a análise rápida e automática de grandes quantidades de texto.

[18] No original: *despite the progress that has been made in the field of corpus linguistics and language teaching, the practice of ELT has so far been largely unaffected by the advances of corpus research*.

[...] a introdução de dados de *corpus* em livros de referência e materiais de ensino de um lado, e as práticas cotidianas de ensino, de outro"[19].

Um dos primeiros proponentes das aplicações pedagógicas diretas de *corpora*, ou seja, a ideia de se trazer *corpora* para a sala de aula, usando concordanciadores para explorar a língua, foi Tim Johns. O autor argumentava que, assim como pesquisadores especializados em linguística podem observar os dados de um *corpus* e descobrir informações importantes sobre o funcionamento da língua, aprendizes de línguas também podem olhar para as concordâncias e chegar às suas próprias conclusões de maneira indutiva, o que seria uma forma de aprendizagem a partir dos dados. Daí surgiu o termo cunhado pelo autor, *Data-Driven Learning* (Aprendizagem Direcionada por Dados), e sua afirmação de que "a pesquisa é séria demais para ser deixada apenas com os pesquisadores" (JOHNS, 1991, p. 2)[20].

Como mencionado na introdução deste texto, o CELV foi concebido como uma aplicação direta de *corpora* no ensino-aprendizagem de línguas, isto é, como um recurso que pode ser usado diretamente por professores e aprendizes tanto dentro quanto fora da sala de aula. Por isso, a próxima seção deste texto se dedicará a explicar detalhadamente os aspectos da ADD e apresentar investigações recentes que apontam para possíveis aplicações dessa abordagem, bem como suas limitações e desafios.

2.3.1 Aprendizagem Direcionada por Dados

Antes de me aprofundar nos detalhes relacionados à ADD, julgo ser necessário tecer algumas considerações sobre a tradução do termo para o português. Em pesquisas recentes sobre o tema no Brasil, o termo original em inglês, *Data-Driven Learning*, tem recebido traduções diferentes para a língua portuguesa. Para apresentar essas diferenças de tradução, compilei um pequeno *corpus* composto por sete dissertações de mestrado e duas teses de doutorado[21] que abordam o tópico. Juntos, os nove textos somam 373.752 palavras. Para encontrar as diferentes formas como *Data-Driven Learning* foi traduzido, busquei por colocações entre as palavras Aprendizagem + Dados e Aprendizado + Dados,

---

[19] No original: *There is thus a clear divide between the exponentially growing number of publications in applied native corpus research and the introduction of corpus data in reference books and teaching materials on the one hand and everyday teaching practices on the other.*

[20] No original: *research is too serious to be left to the researchers.*

[21] As dissertações usadas foram Ferreira (2006), Moreira Filho (2007), Contrera (2010), Duarte (2011), Acunzo (2012), Tartoni (2012) e Silero (2014). As teses usadas foram Kinderman (2011) e Almeida (2014).

com uso do *WordSmith Tools*. Os resultados foram: Aprendizado Movido por Dados (13 ocorrências), Aprendizagem Dirigida por Dados (13 ocorrências), Aprendizagem Movida por Dados (10 ocorrências), Aprendizado Movido a Dados (3 ocorrências), Aprendizagem Movida a Dados (2 ocorrências), Aprendizado Dirigido por Dados (1 ocorrência) e Aprendizagem Dirigida pelos Dados (1 ocorrência). A abreviação AMD aparece 7 vezes. *Data-Driven Learning*, em inglês, ocorre 48 vezes, e sua abreviação, DDL, 81 vezes. Nas traduções, nota-se o uso alternado de *aprendizagem* e *aprendizado*, de *dirigido(a)* e *movido(a),* e das preposições *por*, *a* e *pelos*. Além disso, há a ocorrência de mais de uma dessas variações em alguns dos textos.

Esses dados indicam que ainda não há consenso sobre a tradução do termo para o português. Não bastassem essas variações, Scott (2010) traz ainda outra tradução: Aprendizagem Direcionada por Dados. Como observado no título desta pesquisa e ao longo deste texto, este trabalho concorda com essa tradução, por entender que a palavra "direcionar" remete melhor aos princípios da abordagem, como o papel ativo do aluno de língua estrangeira em seu próprio processo de aprendizagem, usando dados provindos de *corpora* como forma de direcionamento para a descoberta de padrões lexicogramaticais. Acrescento que, ao meu ver, uma padronização do termo em português seria vantajosa para estudos futuros sobre o tema, pois facilitaria a comunicação entre pesquisadores nele interessados. Como já explicado, Aprendizagem Direcionada por Dados, abreviada como ADD, me parece uma boa opção. Feito esse esclarecimento tradutório, prossigo com a descrição do tema.

O texto de Johns (1991) pode ser considerado um artigo seminal sobre a ADD, pois nele o autor descreve os principais aspectos da abordagem. O primeiro aspecto, já mencionado, é o uso direto de concordanciadores na sala de aula. Essa prática traz contribuições importantes sobre o lugar da gramática no ensino-aprendizagem de línguas. Segundo o autor, a ADD permite que o aprendiz adquira consciência gramatical, uma vez que está ativamente engajado no processo de descobrir padrões de uso da língua a partir de evidências autênticas, e não a partir de regras fornecidas previamente. A observação de vários exemplos de uso da língua permite que o aluno passe por um processo cognitivo indutivo, durante o qual ele é capaz de chegar a generalizações gramaticais, e o resultado disso é um maior entendimento sobre o funcionamento da língua. Em suas palavras, "[...] quando a descrição gramatical é o produto do próprio engajamento do aluno com as

evidências, essa descrição pode demonstrar um grau muito maior de abstração e argúcia" (JOHNS, 1991, p. 3)[22].

Para criar condições propícias para a viabilização desse processo cognitivo indutivo de aprendizagem com base em concordâncias, os papéis do professor e do aluno no processo são repensados. Segundo Johns, o professor possui, tradicionalmente, um papel de controle do processo. No contexto convencional, "o professor, geralmente, faz uma pergunta (cuja resposta já sabe) para verificar se a aprendizagem ocorreu; o aprendiz tenta responder à pergunta; e o professor dá retorno sobre a correção da resposta" (JOHNS, 1991, p. 1)[23]. Na ADD, em contrapartida, o professor atua como um agente de consulta e mediação, e não como controlador de todo o processo. Nesse sentido, uma das vantagens do uso de linhas de concordância geradas a partir de amostras de linguagem autêntica é que, frequentemente, os exemplos listados na tela do computador são inesperados e, talvez, desconhecidos pelo professor, o que faz com que ele deixe de ser o único informante sobre a língua presente no processo; o *corpus* se torna um informante capaz de apresentar uma grande quantidade de exemplos de uso não inventados, e o professor deve auxiliar os aprendizes durante a interpretação desses exemplos.

O papel do aluno, por sua vez, deve ser parecido com o de um pesquisador. Evidentemente, não se espera do aluno o domínio de técnicas específicas de pesquisa em linguística, mas a observação de vários exemplos de determinada palavra ou expressão possibilita ao aluno uma investigação indutiva de seu uso na língua. A ADD favorece o desenvolvimento de um espírito investigativo pelo aprendiz, que deve buscar, ativamente, por respostas a questões linguísticas que podem ter sido instigadas pelo professor ou por ele próprio. Johns (1991) descreve o aluno, portanto, como um descobridor da língua.

Para que os professores e alunos possam assumir os papéis acima descritos, engajando-se em um processo de ensino-aprendizagem com base nos preceitos da ADD, é necessário que adquiram certos conhecimentos básicos sobre o funcionamento dos *corpora*, suas possibilidades e suas limitações. Esse conjunto de conhecimentos pode ser chamado de letramento de *corpus* (HEATHER; HELT, 2012) ou competência de *corpus* (KREYER, 2008), e será explorado detalhadamente a seguir.

---

[22] No original: [...] *when grammatical description is the product of the learner's own engagement with the evidence, that description may show a far greater degree of abstraction and subtlety* [...].

[23] No original: *The teacher typically asks a question (answer already known) to check that learning has taken place; the learner attempts to answer the question; and the teacher gives feedback on whether the question has been successfully answered.*

2.3.1.1 Letramento de corpus

Letramento de *corpus* é o conjunto de habilidades necessárias para fazer uso de *corpora* na análise e ensino de línguas (HEATHER; HELT, 2012). Tais habilidades envolvem noções técnicas necessárias para lidar com as ferramentas e dados relacionados ao trabalho com *corpora*, como conhecimento da composição textual do *corpus* com que se está trabalhando, manuseio dos recursos disponíveis para consulta (funcionalidades e possibilidades de busca), noções básicas sobre o funcionamento de concordanciadores e programas de análise lexical e domínio de conceitos fundamentais da LC como frequência, co-ocorrência e o Princípio Idiomático proposto por Sinclair e descrito anteriormente neste texto. Assim, no contexto do ensino-aprendizagem de línguas, é importante considerar que "professores que desejam usar *corpora* com seus alunos precisam ter um bom entendimento dos aspectos multifacetados do letramento de *corpus*" (MEUNIER, 2011, p. 463)[24].

Ainda no que diz respeito a aplicações pedagógicas de *corpora*, Kreyer (2008) propõe o termo competência de *corpus*, uma expansão da noção de letramento de *corpus*, que considera, além dos conhecimentos técnicos acima mencionados, as complexas inter-relações envolvidas na análise dos dados de um *corpus*, representadas na Figura 7.

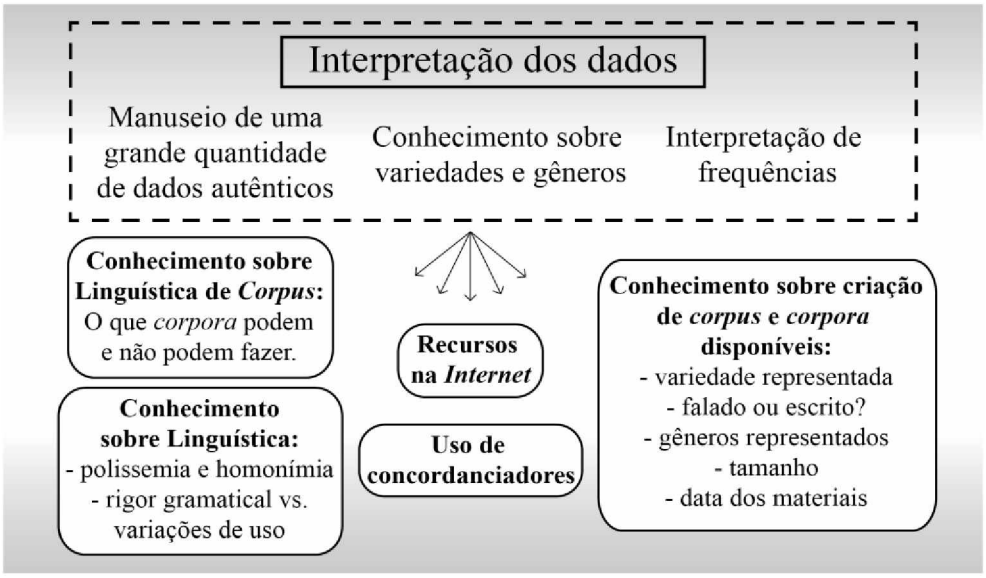


Figura 7: aspectos da competência de *corpus*.
Fonte: traduzido e adaptado de Kreyer, 2008, p. 433 (original: Anexo 1).

[24] No original: *teachers who want to use corpora with their students need to have a good understanding of the multi-faceted aspects of corpus literacy*.

Conforme ilustrado pela figura, além dos conhecimentos técnicos necessários para lidar com as diversas ferramentas envolvidas no trabalho com *corpora*, é preciso saber interpretar os dados, ter conhecimentos básicos sobre linguística geral e sobre LC, familiarizar-se com o uso de concordanciadores e ter conhecimento sobre diferentes *corpora* disponíveis para consulta na *internet*, especialmente quando se trata da aplicação de *corpora* no ensino-aprendizagem de línguas. Todas essas informações podem ser usadas para auxiliar a análise dos dados provindos de *corpora*, levando a interpretações mais precisas e embasadas. Assim, a competência de *corpus* é um requisito que precisa ser desenvolvido tanto por professores quanto por alunos para que se possa utilizar *corpora* adequadamente em contextos pedagógicos. Segundo Kreyer, "competência de *corpus* significa o conhecimento de como lidar com aspectos problemáticos no trabalho com *corpora*" (KREYER, 2008, p. 423)[25].

O desenvolvimento da competência de *corpus* ocorre conforme professores e alunos se tornam mais familiarizados com o trabalho com *corpora* e com os próprios recursos computacionais e tecnológicos usados. Para situar esse processo dentro dos estudos sobre ensino-aprendizagem de línguas em geral, a seção a seguir proporá um encaixe da ADD no construto hierárquico de abordagem, método e técnica proposto por Anthony (1963) e expandido por Almeida Filho (2011).

2.3.1.2 A ADD no construto abordagem, método e técnica

Em seu artigo seminal, Anthony (1963) buscou esclarecer para a comunidade linguística interessada no ensino-aprendizagem de línguas o significado dos termos abordagem, método e técnica, com o intuito de evitar confusões terminológicas e facilitar a comparação entre diferentes maneiras de se ensinar línguas. Assim, o autor propôs um construto hierárquico em cujo topo está situado o conceito de abordagem, seguida do método e das técnicas. Anos depois, buscando esclarecer e expandir o construto proposto por Anthony, Almeida Filho (2011) acrescenta, na base do construto, o conceito de recursos. Essa hierarquia pode ser visualizada na Figura 8.

---

[25] No original: *corpus competence means knowledge of how to deal with problematic aspects in working with corpora.*

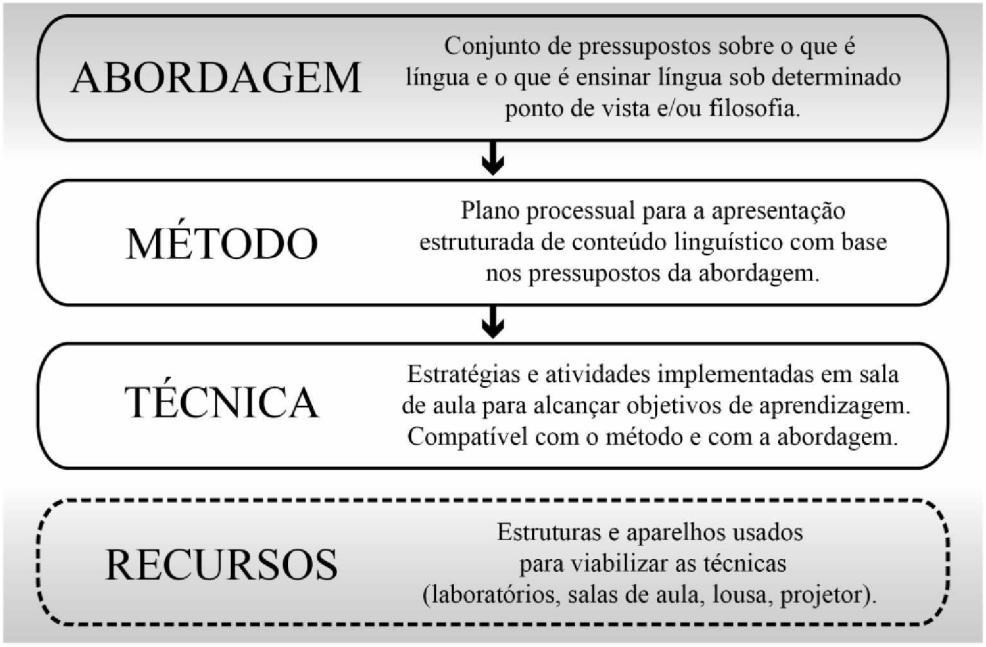


Figura 8: construto abordagem, método, técnica e recursos.
Fonte: elaborada com base em Anthony (1963) e Almeida Filho (2011).

A abordagem, situada no topo da hierarquia, é definida como um grupo abstrato e abrangente de pressupostos ou crenças sobre a natureza da língua e sobre quais seriam as melhores maneiras de ensiná-la. Representa, por exemplo, um consenso entre vários professores de línguas que concordam uns com os outros em termos de suas concepções de língua e experiências de ensino-aprendizagem, manifestando essas crenças em sua prática docente por meio do método ou métodos que adotam.

Método, por sua vez, é a estruturação concreta da abordagem por meio do estabelecimento de planos e processos que norteiem a apresentação do conteúdo linguístico a ser ensinado. Anthony (1963) esclarece que uma abordagem pode dar suporte a vários métodos, que se manifestam em situações concretas de aprendizagem em sala de aula, isto é, técnicas.

Técnicas são as atividades explicitamente realizadas pelo professor durante sua aula para cumprir seus objetivos de ensino, como exercícios contidos nos livros didáticos, tarefas individuais, tarefas em grupos, entre outros. Em outras palavras, ao se observar um professor dando sua aula, o que se vê são as técnicas usadas por ele, que estão implicitamente conectadas ao seu método e à sua abordagem.

A esses três conceitos, Almeida Filho (2011) acrescenta o termo recursos, ou seja, os materiais e ambientes que viabilizam o processo de ensino-aprendizagem. Uma característica importante dos recursos é que costumam se modificar conforme ocorrem

avanços tecnológicos e são propostas novas maneiras de organização do contexto de ensino-aprendizagem.

Conforme previsto por Anthony (1963), esse construto oferece uma maneira estruturada de se explicar uma determinada metodologia de ensino. Por isso, o usarei para buscar situar a ADD dentro dos estudos sobre ensino-aprendizagem de línguas.

Começando pelo nível da abordagem e resgatando conceitos previamente expostos neste texto, é possível listar alguns pressupostos abrangentes e abstratos que embasam a ADD:

1) A língua "é um sistema probabilístico, no qual certos traços são mais frequentes que outros" (SARDINHA, 2004, p.23), retomando Halliday (1991);

2) A produção linguística é governada pelo Princípio Idiomático e pelo Princípio da Livre Escolha, e os usuários da língua fazem uso de uma grande quantidade de estruturas pré-construídas e padronizadas. (SINCLAIR, 1991);

3) Aprender uma língua envolve, entre outros fatores, familiarizar-se com suas estruturas padronizadas e abstrair a ocorrência e co-ocorrência de seus itens lexicais dentro da lógica de seu sistema gramatical.

Esses três pontos podem ser considerados os principais pressupostos que dão suporte às práticas de ensino-aprendizagem da ADD. Percebe-se, portanto, que esta é uma forma de ensinar línguas atrelada à concepção de linguagem da LC.

No nível do método, pode-se dizer que o plano processual que norteia as ações dos professores e alunos na ADD é o engajamento do aprendiz com evidências linguísticas por meio de um processo cognitivo indutivo, que é auxiliado por ferramentas computacionais e por grandes quantidades de dados linguísticos (*corpora*).

No nível da técnica, a principal atividade de ensino-aprendizagem explicitamente realizada em um contexto de ADD é a interação com linhas de concordância por meio de concordanciadores. Essa interação deve ser direcionada pelas etapas do método empírico de análise de dados, ou seja, observação de evidências (dados linguísticos), formulação de hipóteses sobre seu funcionamento (indagações sobre o uso da língua, instigadas pelo professor ou pela curiosidade do próprio aluno) e verificação das hipóteses (por meio de confirmação ou correção, que pode vir do professor ou dos próprios dados).

Por fim, no nível dos recursos, tem-se os materiais essenciais para viabilizar o processo de trabalho com a ADD, isto é, computadores com acesso a ferramentas de

corpora, que podem ser *off-line* (programas de análise de *corpora* instalados no computador) ou *on-line*, como o COCA, o CELV e outros.

Esta proposta de inserção da ADD dentro do construto de Anthony pode ajudar a entender onde essa forma de ensinar línguas se encaixa dentro dos estudos linguísticos. No entanto, trata-se de uma simplificação, isto é, uma estruturação resumida do que vem a ser a ADD, cujo propósito é servir como referência para fins de introdução ao tema e comparação com outras formas de ensino.

Um outro paralelo proposto entre a ADD e o construto de Anthony é seguinte: "Johns frequentemente se referia à ADD como uma 'abordagem', delineando várias técnicas a ela associadas; ela não é um 'método' propriamente dito, mas pode ser integrada em vários tipos de curso, cada um com seus próprios métodos" (BOULTON, 2011, p. 573)[26]. Por esse ponto de vista, a ADD não necessariamente estipula um conjunto de prescrições e um plano processual a ser seguido para o ensino de línguas, o que seria característico de um método; ao invés disso, ela oferece um conjunto de princípios abstratos que podem ser interpretados e aplicados de diversas formas, o que a situa no nível da abordagem. Seguindo a mesma linha de raciocínio, Tartoni acrescenta que "a reflexão da língua a partir das técnicas da Linguística de Corpus deve ser apenas um dos vários momentos que compõem uma sequência didática" (TARTONI, 2012, p. 97). Em suma, portanto, a ADD é um conjunto de pressupostos sobre o ensino de línguas que possui técnicas próprias e pode ser aplicado a diversos contextos, e foca-se no empoderamento dos aprendizes para que explorem dados linguísticos por conta própria e cheguem às suas próprias conclusões.

Retomando a esquematização proposta pela Figura 7, observa-se que a ADD pode ser definida mais especificamente a partir de dois tipos de interação: entre o professor e o *corpus* e entre o aprendiz e o *corpus*. Essas duas abordagens também podem ser chamadas de abordagens *soft* e *hard*, respectivamente, e serão apresentadas a seguir.

2.3.1.3 Abordagens *soft* e *hard*

Segundo Römer (2008), a interação entre o professor e o *corpus* é uma forma controlada de se usar *corpora* pedagogicamente. Trata-se do uso que um professor pode

---

[26] No original: *Johns frequently referred to DDL as an 'approach', outlining a number of associated techniques; it is not a 'method' in its own right, but can be integrated into various types of courses each with their own methods.*

fazer das concordâncias obtidas por meio do *corpus* antes mesmo de mostrá-las aos seus alunos: o *corpus* pode servir como um informante parecido com um falante nativo da língua, e consultado pelo professor para esclarecer dúvidas linguísticas; as concordâncias também podem servir para prover exemplos de frases que o professor pode usar para elaborar exercícios; além disso, a observação dos dados do *corpus* pode enriquecer o conhecimento do próprio professor sobre a língua. Berber Sardinha (2010) explica que, ao elaborar materiais para sala de aula com base em *corpus*, o professor pode trazer as concordâncias de diversas formas, com preparação prévia do material: integral (linhas de concordância assim como aparecem no concordanciador, sem alterações), selecionada (contendo apenas as concordâncias mais relevantes, escolhidas pelo professor), editada (modificadas pelo professor para simplificar ou organizar o conteúdo) e lacunada (escondendo a palavra de busca no centro das concordâncias). O autor explica, também, que o ambiente ideal para explorar esse material é em um laboratório de informática, onde todos os alunos têm acesso a computadores com concordanciadores instalados; no entanto, caso esse tipo de ambiente não esteja disponível, o professor deve imprimir ou projetar as concordâncias, o que tem a desvantagem de não permitir sua exploração de forma dinâmica. A interação professor-*corpus* também tem sido chamada, na literatura sobre o tema, de abordagem *soft* (GABRIELATOS, 2005).

A interação entre os aprendizes e o *corpus*, por sua vez, é o contato direto dos alunos com o concordanciador, sem que os dados tenham sido previamente selecionados ou alterados pelo professor. Essa modalidade de interação também pode ser chamada de abordagem *hard* (GABRIELATOS, 2005), e tem implicações sobre a forma de se conduzir uma aula de línguas dentro do contexto da ADD:

> Quando os aprendizes têm acesso direto aos *corpora*, o foco da lição pode se tornar mais flexível para refletir seus interesses e necessidades. Em outras palavras, o professor ou os aprendizes têm a opção de modificar os objetivos e a direção da lição *in loco*, de acordo com o que emerge. [...] Se as concordâncias não oferecerem pistas suficientes, os aprendizes podem obter mais texto apenas clicando na palavra-chave ou num botão específico (dependendo do programa). (GABRIELATOS, 2005, p. 1)[27]

[27] No original: *When learners have direct access to corpora, the focus of the lesson can be made more flexible to reflect their interests and needs. In other words, the teacher or learners have the option of modifying the aims and direction of the lesson on the spot according to what emerges. [...] If the concordance or sentences do not offer enough clues, learners can get more text just by clicking either on the key word or a special button (depending on the software).*

Assim, na versão *hard* da ADD, é necessário que o aluno tenha acesso direto ao *corpus*, diferentemente da versão *soft*, na qual o contato do aluno com os exemplos de língua contidos no *corpus* é intermediado pelo professor, que seleciona, formata e adapta as amostras previamente. Além disso, na versão *hard* é preciso que o aluno possua as habilidades necessárias para manusear *corpora*, ou seja, que possua letramento de *corpus*, como previamente apresentado. Nesse contexto, cabe ao professor o papel de guia do processo, sem controlar constantemente o conteúdo apresentado e as atividades a serem realizadas, o que favorece a atuação do aluno como um investigador da língua, direcionado por sua própria curiosidade e suas indagações linguísticas. Por esse motivo, a abordagem *hard* pode, inclusive, ser utilizada com sucesso por um aluno mesmo fora do contexto da sala de aula e sem o auxílio de um professor, em seu próprio computador e conectado à *internet* para consultar ferramentas de *corpora on-line*.

Gabrielatos (2005) esclarece que existem pontos intermediários entre as duas abordagens; isto é, não é necessário optar exclusivamente por uma ou por outra, sendo possível realizar atividades centradas no professor, atividades centradas no aluno, ou uma combinação dos dois tipos, de maneira colaborativa entre professores e alunos e com auxílio dos *corpora*.

2.3.1.4 Trabalhos sobre ADD

Esta parte da dissertação será dedicada à menção de alguns trabalhos nacionais e estrangeiros que abordaram a ADD, com o intuito de apresentar brevemente diferentes maneiras como a abordagem tem sido estudada no Brasil e no mundo e as principais propostas para sua implementação no ensino-aprendizagem de línguas.

No âmbito nacional, temos: Ferreira (2006), que identificou, com auxílio da LC, palavras em língua inglesa de relação atípica entre ortografia e pronúncia que comumente causam dúvidas em aprendizes brasileiros, com vistas a orientar a elaboração de materiais didáticos embasados pela ADD com maior ênfase nessas palavras; Moreira Filho (2007), que desenvolveu um *software* para auxiliar a preparação baseada em *corpora* de aulas de inglês com ênfase na leitura; Duarte (2011), que elaborou material didático baseado em *corpus* para o ensino de inglês especificamente para aprendizes da área de Tecnologia Ambiental; Acunzo (2012), que levantou e investigou um *corpus* específico do tema Publicidade com o objetivo de propor atividades de ensino de inglês baseadas na ADD para profissionais da área; Tartoni (2012), que investigou a aplicação de linhas de

concordância no aprimoramento do uso de construções linguísticas com *to* e *for* por estudantes do 9º ano do Ensino Fundamental em uma escola pública; Almeida (2014), que examinou o uso feito por aprendizes brasileiros de colocações com verbos de alta frequência do inglês, identificado problemas de uso dessas colocações e, também, propondo atividades para seu aprimoramento com base na ADD; e Silero (2014), que comparou o uso dos quantificadores *a few* e *few* por falantes nativos do inglês com o uso destes quantificadores por aprendizes brasileiros, e sugeriu uma atividade para a identificação de padrões linguísticos contendo estes quantificadores com base na ADD.

Recorrendo novamente à estrutura proposta pela Figura 7, podemos situar mais especificamente esses trabalhos da seguinte forma: Ferreira (2006) se encaixa na categoria Aplicações Indiretas de *corpora* ao ensino de línguas, pois focou-se no estudo de dados linguísticos provindos de um *corpus* para propor a elaboração de materiais didáticos empiricamente embasados; Duarte (2011), Acunzo (2012) e Silero (2014) também se encaixam, na maior parte de seus trabalhos, na categoria Aplicações Indiretas, mas também acrescentaram sugestões de atividades de ensino baseadas nos preceitos da ADD, com o uso de concordâncias em sala de aula; o trabalho de Moreira Filho (2007), por sua vez, se encaixa na categoria Aplicações Diretas de *corpora* no ensino de línguas, mais especificamente na abordagem *soft* da ADD, pois buscou criar uma maneira de viabilizar o acesso a *corpora* por professores para a preparação de suas aulas; Tartoni (2012) e Almeida (2014), por fim, também se encaixam na categoria Aplicações Diretas, mais especificamente na abordagem *hard* da ADD, por terem usado concordanciadores diretamente com aprendizes em sala de aula com o intuito de aprimorar o uso de um determinado padrão linguístico.

Dentre os trabalhos nacionais mencionados, o de Moreira Filho (2007) é o que mais se assemelha a este trabalho, por também se tratar do desenvolvimento de um sistema computacional baseado nos preceitos da LC e da ADD e voltado para o uso de *corpora* em sala de aula. Como demonstrado, o autor focou a abordagem *soft* da ADD, enfatizando a interação do professor de língua com o *corpus* por meio da proposta de seu *software* para preparação de aulas. O que eu proponho com o CELV, como será detalhado nas seções finais deste texto, é um sistema que possa ser usado tanto por professores quanto por alunos.

Já em relação a trabalhos estrangeiros sobre a ADD, alguns exemplos são: Ma (1994), que aplicou procedimentos de ensino baseados na ADD a um grupo de 18 alunos

adultos e concluiu que o trabalho com concordâncias em sala de aula exige que professores e alunos tenham conhecimento sobre o potencial e limitações dos *corpora*, e que os aprendizes devem desenvolver estratégias específicas para o trabalho com grandes quantidades de dados linguísticos; Cobb (1997), que investigou o uso de concordâncias por estudantes universitários árabes durante um programa de ensino de inglês e concluiu que a interação direta com um *corpus* teve um impacto positivo em suas pontuações nas tarefas avaliativas do programa; Kennedy e Miceli (2010), que trabalharam com três aprendizes de italiano propondo maneiras introdutórias de manuseio de *corpora*, usando-os, inicialmente, como ferramentas de consulta gramatical e para o fomento da criatividade durante a escrita; Boulton (2011), que buscou esquematizar a evolução das técnicas de ADD desenvolvidas desde Johns (1991) até a atualidade, com o intuito de categorizar a metodologia relacionada ao tema e unificar o termo; e Çelik (2011), que investigou o efeito da ADD na aprendizagem de inglês como língua estrangeira em comparação ao uso de dicionários e concluiu que os aprendizes usuários da ADD obtiveram maior retenção dos conteúdos aprendidos.

Os trabalhos estrangeiros mencionados acima são apenas uma amostra da vasta gama de pesquisas existentes sobre o tema no mundo. Observa-se que há uma grande quantidade de trabalhos estrangeiros sobre a ADD, e no Brasil a área está começando a se desenvolver. Adicionalmente, grande parte dos trabalhos já se preocupa com propostas, desenvolvimento e testes de técnicas diretas de aplicação de *corpora* nas salas de aula. No entanto, apesar da expansão e enriquecimento das pesquisas sobre a ADD, os textos encontrados sobre o assunto sempre apontam para uma grande necessidade de se desenvolver cada vez mais sugestões de uso da abordagem nos contextos de ensino-aprendizagem, e também de se testar a abordagem com alunos de diferentes perfis. Há uma constante preocupação em se identificar as limitações da ADD, e buscar soluções para a superação dessas limitações. É desse tópico que tratará a próxima parte do texto.

2.3.1.5 Limitações da ADD

Podem ser elencadas duas limitações principais da ADD: (i) dificuldades no acesso a ferramentas computacionais e de *corpora*, por professores e alunos, seja por indisponibilidade desses recursos ou por falta de conhecimento técnico ou aversão à tecnologia por parte dos usuários; e (ii) aplicação da ADD muito restrita ao perfil específico dos estudantes universitários, que frequentemente já possuem certa

proficiência em inglês e estão em um contexto de formação acadêmica, o que implica em alto letramento.

A limitação da ADD mais comumente mencionada nos textos sobre o assunto é a dificuldade dos aprendizes se adaptarem às técnicas usadas, principalmente ao trabalho com concordâncias, que não são, para eles, uma maneira natural de visualização da língua. Acunzo (2012) menciona que alguns alunos possuem uma aversão ao uso de certas tecnologias, o que pode, inicialmente, dificultar sua inserção em um contexto de ADD. Ma (1994) explica que olhar para várias telas de concordâncias sem conseguir extrair delas nenhuma informação útil pode ser muito frustrante para os aprendizes, e por isso é essencial que eles desenvolvam estratégias eficientes de análise de dados previamente ao seu contato com um *corpus*. Moreira Filho (2007) afirma que, em sua maioria, as ferramentas computacionais de análise da língua são feitas para o uso de pesquisadores e linguistas, o que pode dificultar seu manuseio por alunos de línguas ou por professores não familiarizados com a LC. Duarte (2011) comenta que, para que possam manipular diretamente as informações de um *corpus*, é necessário que os alunos passem por algum tipo de introdução ao uso das ferramentas de *corpora* necessárias durante o processo. Kennedy e Miceli (2010) observaram que, no caso específico de seus alunos de italiano, os aprendizes demonstraram progresso com a abordagem; no entanto, as autoras admitem que nem todos os alunos são facilmente motivados a usar concordanciadores.

Outro ponto que se deve levantar sobre a aplicabilidade da ADD se refere ao fato de que grande parte das pesquisas sobre o tema – como Cobb (1997), Kennedy e Miceli (2010) e Çelik (2011) – estão focadas no ambiente universitário, que é um contexto muito específico de alunos. Neste contexto, o perfil dos aprendizes é de alto grau de letramento, idade adulta e, em alguns casos, conhecimento prévio sobre a língua estudada que já os classifica em nível intermediário ou avançado. Em contrapartida, alguns trabalhos brasileiros demonstraram bons resultados na aplicação da ADD fora do contexto universitário: Tartoni (2012) afirma que obteve resultados positivos com o uso de concordanciadores em sala de aula por alunos do ensino fundamental de nível básico de proficiência em inglês, e que é possível que o uso das técnicas da ADD, associado a outros tipos de atividade de ensino, promova a familiarização e adaptação dos aprendizes ao trabalho com linhas de concordância; Acunzo (2012) também buscou aplicar a ADD em novos contextos, aplicando suas técnicas a alunos da área de Publicidade em seus próprios locais de trabalho, e não no ambiente universitário; Moreira Filho (2007), por sua vez,

teve um foco específico no uso de seu *software* de preparação de aulas por professores de Ensino Médio de escolas públicas.

Como pôde ser observado, a literatura sobre a ADD demonstra um esforço para a identificação das limitações da abordagem. Com base no reconhecimento dessas limitações, já existem propostas, estudos e testes, como alguns dos supramencionados, para lidar com essas dificuldades e buscar maneiras de expandir e fomentar o uso dessa abordagem, o que poderá ajudar a consolidar a ADD como uma prática viável dentro do contexto do ensino-aprendizagem de línguas.

2.4 Uso de vídeos e legendas no ensino-aprendizagem de línguas

Este subcapítulo se dedicará à apresentação de alguns recursos já existentes que promovem o uso de vídeos e legendas direcionado ao ensino-aprendizagem de línguas, alguns com o auxílio da LC, e outros, não. Os recursos aqui mencionados possuem semelhanças com o CELV e, em alguns casos, me ajudaram a desenvolver ideias para a criação da ferramenta e suas funções.

2.4.1 *Movie Segments to Assess Grammar Goals*

Movie Segments to Assess Grammar Goals[28] (AZEVÊDO, 2008) é um *blog* que publica atividades para o ensino de língua inglesa com uso de trechos de filmes. Cada publicação do *blog* traz uma sugestão de tópico gramatical e uma sequência de exercícios que podem ser realizados com base em um determinado vídeo, que também é incorporado na página, além de uma categorização do nível de proficiência e faixa etária para os quais cada atividade é recomendada. A Figura 9, a seguir, exemplifica uma publicação do *blog* com os primeiros exercícios propostos em uma sequência de atividades sobre um trecho do filme *The Shining*.

[28] moviesegmentstoassessgrammargoals.blogspot.com.br

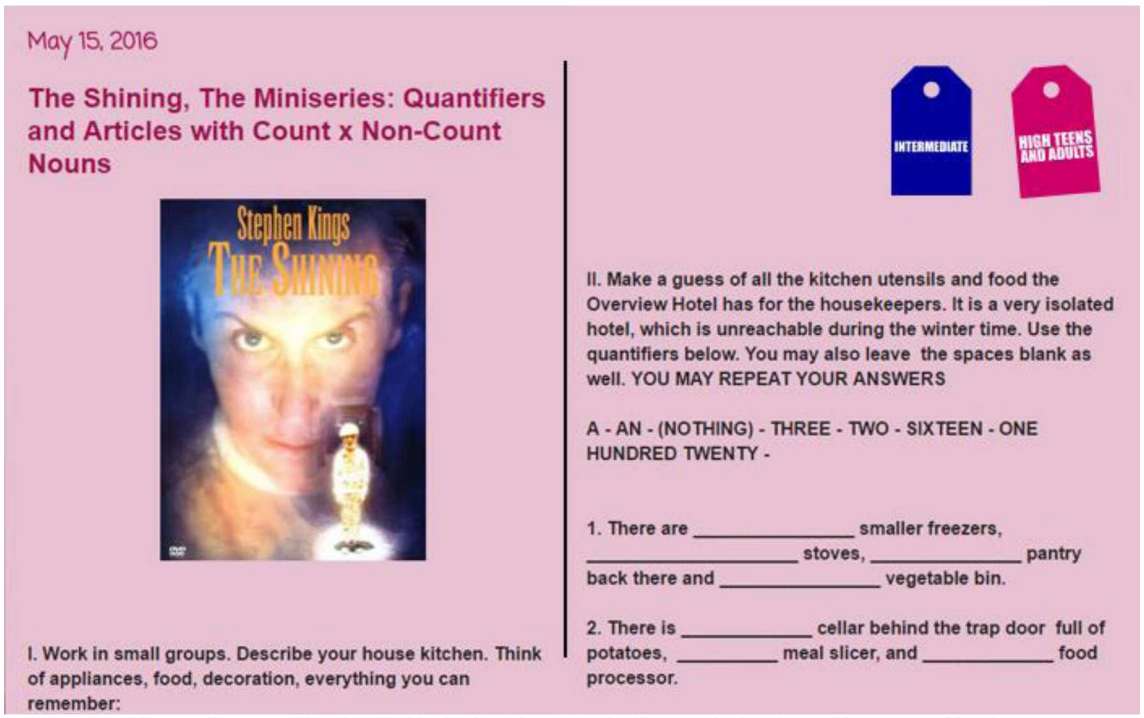


Figura 9: exemplo de atividade do *blog Movie Segments to Asses Grammar Goals*.
Fonte: captura de tela de moviesegmentstoassessgrammargoals.blogspot.com.br.

Desde o início de sua atuação, o *blog* produziu uma grande quantidade de atividades com vídeos, focadas em variados tópicos gramaticais da língua inglesa. Azevêdo (2008) descreve seus exercícios como divertidos e desafiadores, acrescentando que a presença dos vídeos no contexto de ensino-aprendizagem de inglês é inspiradora e ajuda a promover a motivação dos alunos.

Seguindo essa linha de trabalho com vídeos, *Movie Segments for Warm-ups and Follow-ups*[29] (AZEVÊDO, 2009) é outro *blog* criado pelo mesmo autor, desta vez focando no uso de trechos de filmes em *warm-ups* e *follow-ups*, isto é, atividades que podem ser usadas com o objetivo de preparar o aluno para a aprendizagem de um certo tópico (como um "aquecimento") ou revisar e finalizar o seu estudo.

Os dois *blogs*, além de fornecerem uma grande quantidade de materiais de ensino previamente preparados, também trazem inspiração para o uso de vídeos em sala de aula, porque demonstram maneiras eficazes de se conduzir esse tipo de atividade, o que pode motivar os professores usuários dos *blogs* a elaborarem suas próprias aulas com vídeos.

Há uma diferença importante entre a forma de uso de vídeos feita por esses *blogs* e a forma que proponho com o CELV. Os *blogs* incorporam, em suas atividades, cenas que possuem princípio, meio e fim (embora sejam trechos recortados de um filme), de

[29] warmupsfollowups.blogspot.com.br

maneira que seja possível ao aluno compreender o sentido geral do vídeo. O CELV, por outro lado, incorpora os vídeos em sua ferramenta a partir de linhas de concordância, que, como mencionado anteriormente, não são usadas com o objetivo de se compreender o sentido geral de um texto, mas, sim, o uso específico de uma palavra ou expressão.

2.4.2 O projeto LeViS e o projeto ClipFlair

O projeto LeViS foi uma proposta de aplicação da atividade de legendagem como forma de ensino-aprendizagem de línguas. O acrônimo LeViS significa *Learning via Subtitling*, o que pode ser traduzido como Aprendendo pela Legendagem. A página do projeto na *internet*[30] afirma que o uso de materiais audiovisuais no ensino-aprendizagem de línguas possui várias vantagens, como o desenvolvimento das capacidades de abstração e identificação de informações e a visualização da narrativa pelo aluno. Adicionalmente, a página explica que a presença de legendas em filmes e outros vídeos é um fator que promove a aprendizagem de línguas, o que ocorre especialmente em estudantes de tradução, que não só assistem aos vídeos, mas também se envolvem com a atividade de produção dessas legendas.

A partir da observação de que a legendagem de vídeos ajuda a desenvolver e praticar habilidades linguísticas, a proposta do LeViS foi desenvolver um *software* de simulação de legendagem chamado LvS, com o intuito de oferecer aos aprendizes a oportunidade de elaborar legendas em um ambiente computacional, não pela produção das legendas em si, mas pelos benefícios indiretos dessa atividade na aprendizagem da língua. Esse *software* fornece as ferramentas básicas necessárias para a inserção de legendas em um vídeo, e inclui recursos adicionais para nortear o uso pedagógico da legendagem, com instruções para a produção de legendas de boa qualidade e espaço para notas do professor e dos aprendizes, como demonstrado na Figura 10, a seguir.

[30] levis.cti.gr

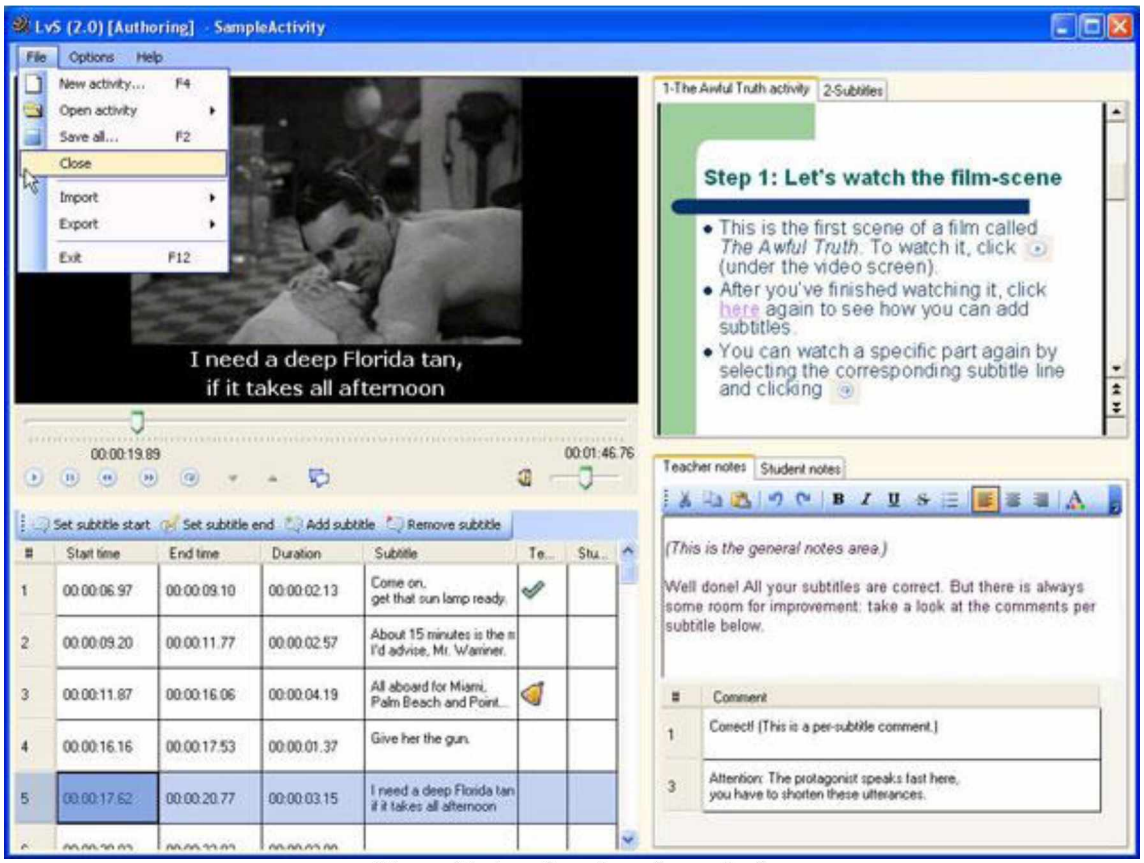


Figura 10: interface do *software* LvS
Fonte: levis.cti.gr

A página do LeViS na *internet* não atribui a autoria do projeto a uma pessoa específica, mas aponta que a instituição que o coordenou foi a *Hellenic Open University*, e que o projeto terminou em 2008, dando lugar ao ClipFlair. Esse novo projeto, localizado em uma outra página da *internet*[31], segue a linha de trabalho do LeViS, porém com uma interface atualizada que permite não apenas a adição de legendas, mas também títulos, anotações, comentários e balões de fala, e acrescenta a possibilidade da prática da habilidade oral dos alunos com o *revoicing*, isto é, gravar a própria voz e acrescentá-la aos vídeos por meio de atividades interativas como redublagem, descrição oral e karaokê.

Os dois projetos caracterizam-se pelo foco no desenvolvimento das capacidades de produção linguística dos alunos (a escrita e a fala), também chamadas de atividades de *output*, cuja prática é viabilizada pela interação com os vídeos a partir das plataformas computacionais desenvolvidas. Essa característica é uma diferença entre esses projetos e a proposta do CELV, pois nele o foco é nas atividades de compreensão da língua (a leitura e a audição), também chamadas de atividades de *input*, cuja prática é viabilizada pela

[31] clipflair.net

interação com linhas de concordância e com os vídeos nelas incorporados. No entanto, os projetos também possuem semelhanças com o CELV, pois desenvolveram sistemas computacionais que enfatizam o trabalho com vídeos e com legendas.

2.4.3 O *corpus* ELISA

Diferentemente dos *blogs* e projetos mencionados nas duas seções anteriores, o ELISA, além de incluir o uso de vídeos, é um trabalho que propõe a aplicação de *corpora* no ensino-aprendizagem de línguas. A autora do ELISA explica que ele é "um pequeno *corpus* de inglês falado contendo vídeos de entrevistas com falantes nativos sobre suas carreiras profissionais" (BRAUN, 2006, p. 27)[32], e o acrônimo ELISA significa *English Language Interview Corpus as a Second-Language Application*.

Braun explica que seguiu uma abordagem diferente da ADD para nortear o desenvolvimento do ELISA: ao invés de focar o trabalho com *corpora* na exploração de concordâncias, a autora advoga que os aprendizes tenham contato com os textos de um *corpus* em um nível discursivo, com acesso aos textos completos para que todo o seu contexto possa ser estudado dentro da situação comunicativa em que aparece, e não apenas em um nível textual e observando somente o cotexto de algumas palavras, como ocorre quando se usa linhas de concordância.

Para implementar essa proposta, segundo Braun (2006), o ELISA foi compilado como um *corpus* de tamanho pequeno (60.000 palavras), o que viabilizou uma abordagem qualitativa, e não quantitativa, dos textos da amostra durante sua compilação e uso. O número menor de palavras possibilitou a anotação manual dos textos, o que permitiu o acréscimo de marcações pedagogicamente úteis que seriam impossíveis de se acrescentar de maneira automatizada por computador, como a categorização dos tópicos descritos ao longo dos textos. Segundo a autora, os entrevistados gravados durante a coleta de textos para o ELISA seguiram uma estrutura padronizada em sua fala, de forma que cada entrevista apresenta alguns ou todos os seguintes tópicos: localização geográfica do(a) profissional entrevistado(a), o que ele(a) faz, seu histórico pessoal, como ele(a) começou sua carreira, exemplos de seus projetos, sua formação e treinamento, questões econômicas de seu trabalho, questões de negócios, rotina do trabalho, desafios e planos futuros. Essa padronização permitiu que os diferentes tópicos mencionados em cada entrevista do

---

[32] No original: *a small corpus of spoken English containing video interviews with native speakers about their professional career.*

ELISA fossem categorizados e anotados manualmente em um esquema de marcação XML[33], seguindo uma estrutura previsível que favorece sua exploração pedagógica.

Essa marcação manual, viabilizada pela abordagem qualitativa à amostra, está dentro do que Braun denominou enriquecimento pedagógico do *corpus*. Segundo a autora, um outro fator que promove enriquecimento pedagógico de um *corpus*, e que também está presente no ELISA, é a inclusão de dados audiovisuais:

> Um tipo de enriquecimento que cumpre uma vastidão de propósitos é a inclusão de materiais audiovisuais em um *corpus* (pedagogicamente relevante). Além das transcrições, o ELISA contém videoclipes das entrevistas. Uma vantagem é que as pistas visuais, gestuais e de entonação nos videoclipes ajudam muito a contextualizar e esclarecer elocuções problemáticas. Uma segunda e igualmente importante vantagem é que *corpora* audiovisuais como o ELISA viabilizam formas inteiramente novas de exploração. Eles podem ser usados, especialmente, para atividades de compreensão oral. (BRAUN, 2006, p. 39)[34]

Outra sugestão feita por Braun (2006) para o enriquecimento pedagógico de um *corpus* é a elaboração prévia de materiais para o seu uso em sala de aula, como a criação de atividades a serem realizadas antes da exploração do *corpus* para introduzir os assuntos que serão estudados, e atividades depois da exploração do *corpus*, como forma de verificar o entendimento dos alunos sobre o conteúdo estudado.

A autora esclarece que sua abordagem qualitativa não significa uma invalidação dos concordanciadores e da ADD, que têm grande utilidade no estudo de padrões linguísticos e atividades focadas no léxico. Assim, sua proposta representa um acréscimo à abordagem da ADD, e não a sua substituição por outra abordagem; em outras palavras, além de se usar concordanciadores, é interessante que se promova o enriquecimento pedagógico do *corpus*.

---

[33] XML, ou *eXtensible Markup Language*, é uma linguagem de marcação comumente utilizada na LC para o acréscimo de anotações sobre informações específicas em textos de determinada amostra.

[34] No original: *One type of enrichment which fulfils a multitude of purposes in the inclusion of audiovisual materials into a (pedagogically relevant) corpus. ELISA contains video clips of the interviews in addition to the transcripts. One advantage is that the visual, gestural and intonational clues in the video clips greatly help to contextualize and clarify problematic utterances. A second, equally important advantage is that audiovisual corpora such as ELISA open up entirely new ways of exploitation. In particular, they can be used for listening comprehension activities.*

O ELISA e suas ferramentas de pesquisa não se encontram mais disponíveis na antiga página do projeto na *internet*[35]. Ao invés disso, o trabalho de Braun serviu como embasamento para um novo projeto, o BACKBONE[36], que oferece uma quantidade maior de entrevistas, mas mantém a linha de trabalho com *corpora* de tamanho pequeno, abordagem qualitativa dos dados e preocupação com o enriquecimento pedagógico.

A diferença principal entre esses projetos e o CELV é a quantidade de dados linguísticos e a existência ou não de um tratamento pedagógico aprofundado. O ELISA e o BACKBONE, partindo da abordagem qualitativa, possuem uma amostra pequena de vídeos, com transcrições de cada vídeo contendo marcações dos tópicos abordados, e acompanhada de atividades previamente elaboradas para o uso em sala de aula. O CELV, por sua vez, usa uma abordagem quantitativa, contendo um grande número de vídeos, cada um associado à sua legenda correspondente, porém com menor tratamento pedagógico da amostra.

2.4.4 *WordSmith* 7.0 e suas novas funções de som e vídeo

Assim como o ELISA, o *WordSmith Tools* (doravante, WST) também é um recurso voltado para o trabalho com *corpora*. No entanto, diferentemente de todos os recursos supracitados, que já existiam antes do início da realização deste trabalho e influenciaram sua elaboração, a versão 7 do WST foi lançada em dezembro de 2015, quando esta pesquisa já se encaminhava para suas etapas finais. Curiosamente, essa nova versão trouxe para o programa um novo recurso que é muito similar à ideia central do CELV, isto é, permitir o acesso a pontos específicos de vídeos em uma determinada amostra a partir de arquivos de legenda.

Na versão atual do manual do WST, Scott (2016b) se refere a esse novo recurso como *video conversion*, isto é, conversão de vídeos, e explica que a nova versão do programa pode lidar com arquivos de legendas no formato .srt, convertendo-os para a codificação necessária (Unicode) para leitura no concordanciador, e sincronizando o tempo das legendas com o tempo dos vídeos, se necessário. Além disso, o concordanciador do WST (chamado de ferramenta Concord) agora é capaz de reproduzir arquivos de vídeo e áudio de formatos comuns (como .mp4 e .avi) a partir das legendas

---

[35] www.uni-tuebingen.de/elisa
[36] projects.ael.uni-tuebingen.de/backbone

em .srt. O autor descreve esse novo uso da ferramenta como citado a seguir e ilustrado pela Figura 11.

> Suponha que você faça uma concordância de "*elephant,*" e queira ouvir como essa palavra é falada de fato, em contexto. A última vogal é um *schwa*? A segunda vogal soa como "i", "e", "u" ou um *schwa*? [...] Se você tiver definido etiquetas que se refiram a arquivos multimídia, e se tais etiquetas estiverem presentes no "contexto de etiqueta" de uma determinada linha de concordância, você pode ouvir ou ver o arquivo original. A etiqueta será analisada para identificar o arquivo requerido, baixando-o de um endereço da *internet* caso necessário, e então reproduzindo-o. (SCOTT, 2016b, p. 1)[37]

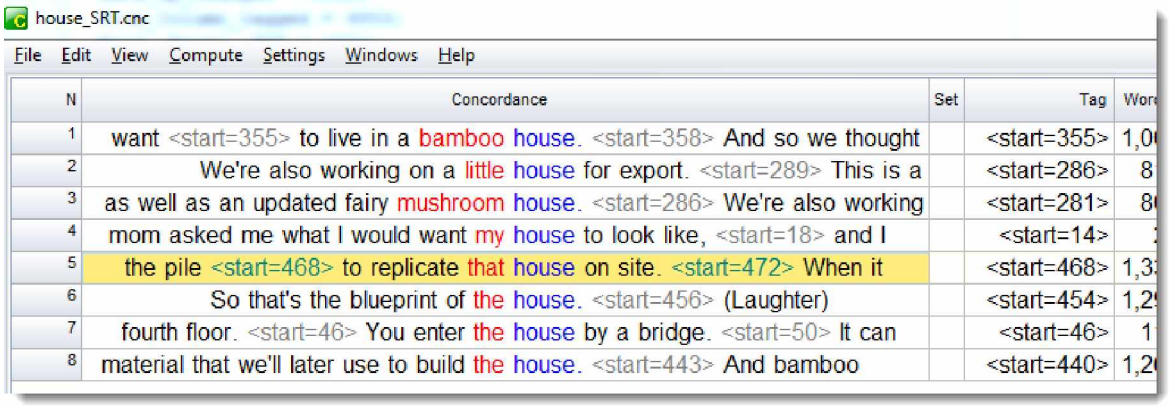


Figura 11: interface da ferramenta Concord do WST contendo legendas de vídeos.
Fonte: manual do WST 7 (SCOTT, 2016b).

A figura apresenta a forma como o WST 7 exibe linhas de concordância geradas a partir de arquivos de legenda. No caso demonstrado pela figura, etiquetas como <start=468> e <start=472> foram previamente definidas para se referir à marcação de tempo de um determinado arquivo multimídia, permitindo que tal arquivo seja reproduzido no tempo determinado pela etiqueta a partir do clique duplo na coluna *Tag* da ferramenta. Scott explica que o arquivo é reproduzido pelo *software* gratuito VLC Media Player[38], que também deve ser previamente instalado no computador.

A diferença entre essa nova função do WST e o CELV é o fato de que o CELV é um *corpus* audiovisual cujos arquivos de legenda foram compilados previamente e

---

[37] No original: *Suppose you do a concordance of "elephant" and want to hear how the word is actually spoken in context. Is the last vowel a schwa? Does the second vowel sound like "i" or "e" or "u" or a schwa? [...] If you have defined tags which refer to multimedia files, and if there are any such tags in the "tag-context" of a given concordance line, you can hear or see the source multimedia. The tag will be parsed to identify the file needed, if necessary downloading it from a web address, and then played.*

[38] www.videolan.org/vlc

disponibilizados para consulta na *internet*, com vídeos imediatamente acessíveis por meio do clique sobre as linhas de concordância da ferramenta, e cujas palavras já passaram por etiquetagem morfossintática. Em contrapartida, para se fazer uso dos novos recursos audiovisuais do WST 7, o usuário precisaria compilar seus próprios arquivos de legenda e também arquivos multimídia, seja baixando-os da *internet* ou especificando seu endereço *on-line*, e também definir etiquetas que associem os arquivos de legenda aos seus respectivos arquivos multimídia, além de implementar quaisquer outras formas de marcação que deseje utilizar. A vantagem de se usar o WST é a possibilidade de compilar um *corpus* próprio, que reflita as necessidades de estudo do pesquisador; a amostra do CELV, em contrapartida, é previamente definida, mas pronta para o uso.

Isso conclui o capítulo teórico desta dissertação. Foram apresentados os tópicos que embasaram o desenvolvimento deste trabalho, principalmente os preceitos da LC e da ADD e suas implicações para o desenvolvimento do CELV, a partir de referências tanto nacionais quanto estrangeiras sobre o assunto, além de exemplos de trabalhos que fazem uso da ADD e também de vídeos e legendas no ensino-aprendizagem de línguas. O próximo capítulo se dedicará a explicar as etapas metodológicas realizadas para a produção e teste do *corpus* e da ferramenta aqui propostos.

3. METODOLOGIA

O desenvolvimento do *corpus* CELV e também da ferramenta *on-line* que permite consultas nele se deu por meio de cinco etapas principais, que serão apresentadas e detalhadas nas próximas seções deste capítulo. No entanto, antes de explicar as etapas principais do projeto, cabe apresentar algumas das decisões iniciais da pesquisa.

Como mencionado na introdução deste texto, a motivação que levou ao desenvolvimento do CELV surgiu da minha prática como professor de inglês. A resposta para viabilizar a criação da ferramenta idealizada foi a LC, a partir do uso de arquivos de legenda. Conforme visto no capítulo teórico desta dissertação, a exibição de vários exemplos de língua na forma de linhas de concordância permite a observação de padrões linguísticos usados em contextos autênticos, o que levou à decisão de criar uma ferramenta de *corpus on-line* centrada na exibição de concordâncias.

As legendas, por sua vez, são o registro da fala dos vídeos em formato de texto escrito eletrônico, o que é necessário para que o texto possa ser lido e processado por um computador, permitindo a aplicação das técnicas de LC. Outra característica dos arquivos de legenda que foi essencial para o desenvolvimento da ferramenta é a presença das marcações de tempo, que indicam em que momento do vídeo cada legenda aparece na tela. A estrutura interna do texto em um arquivo de legenda é como no seguinte exemplo:

> 1
> 00:00:11,448 --> 00:00:16,259
> Hello once again, percussionists of the *internet*.
> In this cajon tutorial we'll be looking at a fairly simple hip hop beat.

A numeração que aparece na primeira linha estipula a sequência de legendas ao longo do vídeo. A marcação de tempo, na segunda linha, especifica o intervalo durante o qual essa legenda será exibida na tela do vídeo. As demais linhas contêm o texto propriamente dito da legenda. As marcações de tempo foram usadas no sistema do CELV para recuperar a informação sobre os momentos em que cada legenda aparece, permitindo a reprodução dos vídeos a partir dos momentos específicos de cada elocução. Esse recurso possibilita que a ferramenta *on-line* do CELV reproduza apenas os trechos de interesse de cada vídeo, resultando em um *corpus* capaz de apresentar a informação textual escrita, e, adicionalmente, reproduzi-la em vídeo. Conforme explicado por Braun (2006), esse tipo de *corpus* pode ser chamado de *corpus* audiovisual. O método usado para recuperar

a informação das marcações de tempo e aplica-la na reprodução dos vídeos será explicado mais detalhadamente nas próximas seções deste capítulo.

Uma vez definido que a ferramenta seria um concordanciador, e que o tipo de arquivo para compor o *corpus* seriam legendas de vídeos, a próxima decisão foi sobre o tipo de vídeo a partir do qual extrair legendas. Em um primeiro momento, considerei utilizar episódios de seriados televisivos e também filmes, já que estes são exemplos de vídeos comumente usados em sala de aula com os alunos, e existem páginas na *internet* especializadas em disponibilizar legendas para esses tipos de vídeo, como o Legendas.tv. No entanto, esses tipos de vídeo não permitiriam a implementação do principal recurso da ferramenta, que é a reprodução dos vídeos a partir das concordâncias, porque filmes e episódios de seriados não estão disponíveis gratuitamente na *internet* da forma necessária para a implementação da ferramenta.

Optou-se, portanto, por utilizar o site YouTube como fonte de vídeos e legendas. Essa escolha provou ser bastante prática, uma vez que, além da grande quantidade de vídeos disponíveis, o YouTube conta com vários recursos que facilitaram o desenvolvimento das funções do CELV. Desde 2007, o *site* disponibiliza para seus usuários a possibilidade de escrever legendas para os seus vídeos, e vários canais já contam com um grande acervo de vídeos legendados. Além disso, o YouTube possui um recurso próprio para reprodução de vídeos a partir de um momento específico, e permite a incorporação de vídeos em outras páginas.

Assim, as três decisões tomadas no momento inicial da pesquisa foram: (i) construção de um concordanciador para observação de exemplos de língua inglesa em uso; (ii) compilação de um *corpus* composto por arquivos de legenda para suprir exemplos para o concordanciador; e (iii) uso do YouTube como fonte de vídeos/legendas.

Determinadas essas características iniciais da ferramenta, o próximo passo foi decidir especificamente como o concordanciador deveria funcionar, isto é, sua maneira de apresentação dos dados linguísticos e suas funções de pesquisa. As duas ferramentas que serviram de embasamento para essa decisão foram o WST e o COCA, por serem recursos amplamente utilizados por pesquisadores da LC e também por terem sido os sistemas com os quais eu tive mais contato em meus estudos sobre *corpora* durante minha formação acadêmica.

Apesar de inspirada pelo WST e pelo COCA, a escolha das funções de pesquisa do CELV foi, em grande parte, subjetiva, com base nas formas com que eu mesmo utilizo

esses sistemas com maior frequência e com base, também, na minha concepção de como um *corpus* audiovisual pode ser explorado. Assim, as funções escolhidas para serem implementadas na ferramenta foram: (i) pesquisas simples, contendo apenas uma ou mais palavras em inglês; (ii) pesquisas complexas, incluindo parâmetros de pesquisa como * (busca por qualquer palavra), | (busca por qualquer uma das palavras separadas por barras verticais), e {} (busca por lemas da palavra inserida entre chaves); (iii) pesquisas com uso de etiquetas morfossintáticas; (iv) possibilidade do uso de filtros de pesquisa a partir das categorias do *corpus*; (v) apresentação dos resultados de busca em forma de lista, contendo o número de ocorrências de cada resultado entre parênteses; (vi) apresentação dos resultados de busca em forma de gráfico, contendo a distribuição de ocorrências dos resultados de busca de acordo com as categorias do *corpus*; (vii) ferramenta de ordenação alfabética (*sort*) das linhas de concordância; (viii) acesso aos vídeos originais de cada legenda por meio do duplo clique nas linhas de concordância; e (ix) possibilidade de ajuste do tempo inicial de reprodução dos vídeos. Os efeitos produzidos por cada uma dessas funções em pesquisas no CELV, bem como a maneira como foram desenvolvidas para o sistema, serão demonstrados nas seções seguintes.

O critério subjetivo permitiu-me acrescentar à ferramenta funcionalidades que acredito serem úteis para a sua aplicação no contexto de ensino-aprendizagem de inglês; no entanto, não foram incluídas funções sobre as quais eu não tinha conhecimento antes da elaboração do trabalho. A possibilidade de inclusão dessas funções aparecerá neste texto como um dos itens para desenvolvimento futuro da ferramenta, e isso será retomado no capítulo final.

Escolhidas as funções da ferramenta, foi necessário explicar essas informações para os programadores que ficaram responsáveis pelo desenvolvimento do sistema. Para isso, foi elaborada uma apresentação em formato *PowerPoint*, cuja transcrição está disponível no Apêndice 1. Essa apresentação foi mostrada aos programadores com os seguintes objetivos: (i) explicar o que é e para que serve um *corpus*; (ii) demonstrar exemplos de outras ferramentas de *corpora on-line* já existentes; e (iii) especificar cada função a ser desenvolvida para o CELV e que efeito deve produzir. Posteriormente, essa apresentação se tornou um documento de instruções, usado como referência durante o desenvolvimento da ferramenta.

A partir da demonstração dessas instruções aos programadores, o *corpus* começou a ser compilado e a ferramenta começou a ser desenvolvida, prosseguindo conforme as etapas indicadas na Figura 12.

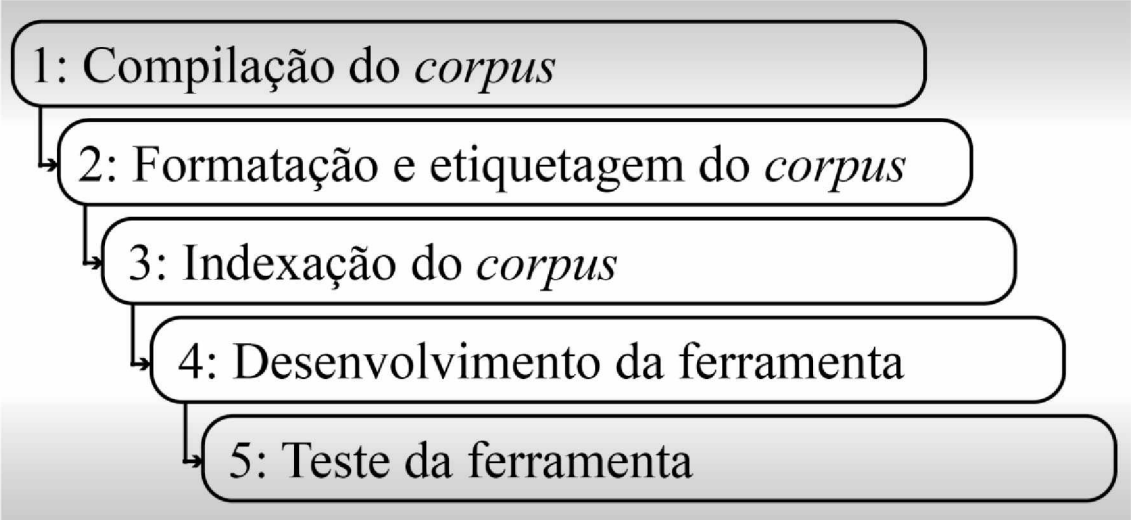


Figura 12: etapas principais da construção do CELV.
Fonte: elaboração própria.

Durante cada uma dessas etapas, principalmente nas quatro primeiras, foi necessário que houvesse comunicação constante com os programadores, para esclarecer dúvidas que surgiram durante a interpretação do documento de instruções. A natureza dessas dúvidas foi, em grande parte, sobre as funções do sistema e seus efeitos. Foi necessário, também, prover amostras de tamanho pequeno do *corpus*, mesmo antes do fim de sua compilação, para que os programadores pudessem ter acesso ao tipo de dados com os quais estavam lidando e desenvolver o sistema a partir deles. Além disso, os programadores deram algumas sugestões sobre ajustes que poderiam ser feitos na interface da ferramenta ou em alguma de suas funções, para facilitar o desenvolvimento do sistema e chegar aos resultados desejados com maior eficiência. Por isso, o resultado final da ferramenta possui algumas diferenças em relação ao documento de instruções apresentado, que serão demonstradas ao longo das próximas seções, juntamente com a explicação detalhada de cada etapa apresentada na Figura 12.

3.1 Compilação do *corpus*

Considerando-se que o YouTube possui um grande número de vídeos, e levando em conta as diretrizes para a compilação de *corpora* explicitadas no capítulo teórico deste texto, foram estabelecidos três critérios para a seleção de legendas para o CELV:

- País: foram selecionados canais originários dos países Austrália, Canadá, Estados Unidos e Reino Unido. Estes países foram selecionados porque são falantes de inglês como língua nativa, e porque há uma amostra suficiente de vídeos no YouTube de cada um deles. Embora outros países (como Nova Zelândia, Irlanda e outros) também pudessem enriquecer a amostra, não há quantidade suficiente de vídeos provenientes destes países.
- Gênero: no YouTube, existem canais especializados em produzir determinados tipos de vídeo. Foram selecionados canais especializados em produzir vídeos dos gêneros *how to* (tutoriais), *vlogs* e *talks* (palestras), por haver uma grande quantidade desses tipos de canal.
- Tema: cada um dos gêneros selecionados se especializa, ainda, em temas específicos. Assim, na amostra selecionada, existem tutoriais sobre culinária, beleza e técnicas musicais, *vlogs* sobre tópicos gerais, viagens e ciência, e palestras sobre meio ambiente, política e tecnologia.

Estabelecidos os critérios, pesquisei no YouTube por canais que atendessem à cada uma das categorias, e tivessem vídeos com legendas disponíveis. Uma vez encontrados vários canais de cada categoria, fiz a extração das legendas com uso do *software* gratuito Google2SRT[39], que extrai arquivos em formato *.srt* a partir do endereço (*link*) dos vídeos no YouTube. Buscou-se chegar a um número equilibrado de *tokens* para cada categoria do *corpus*, embora algumas categorias não tivessem legendas suficientes para chegar a esse número, o que resultou em um *corpus* que não está perfeitamente balanceado, conforme será apresentado e discutido na seção de resultados deste texto.

Após a extração dos arquivos de legenda, foi elaborado um padrão de nomenclatura de arquivo que inclui todas as informações importantes sobre cada legenda: país, gênero, tema, canal e *link* para o vídeo. Os arquivos foram armazenados em

[39] google2srt.sourceforge.net/pt-br/

diretórios no meu computador de maneira a facilitar o acesso a um texto específico, caso necessário. Na Figura 13, a seguir, é possível visualizar um exemplo de diretório contendo arquivos assim organizados.

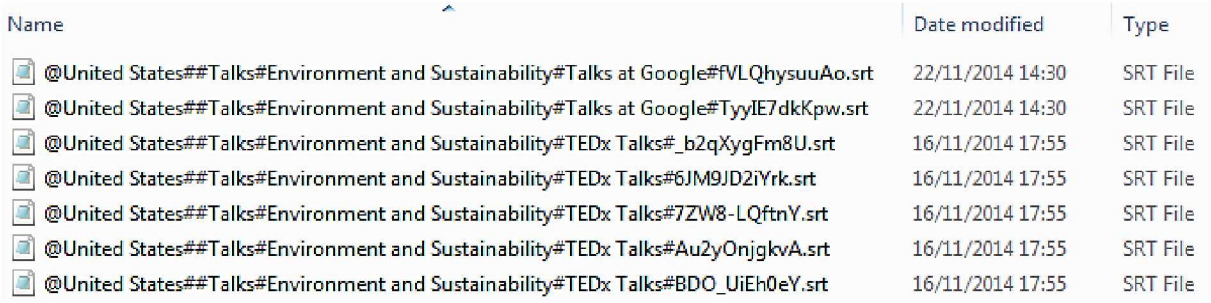


Figura 13: diretório contendo arquivos de legenda em formato *.srt* e sua nomenclatura.
Fonte: captura de tela do diretório contendo os arquivos do *corpus*.

Um desafio encontrado durante essa etapa de compilação foi a identificação do país de origem de cada canal do YouTube cujos vídeos entraram na amostra. O YouTube não possui informações claras sobre o país de origem de cada canal; por isso, foi necessário encontrar alternativas para determinar essa informação. O principal recurso utilizado foi o *site* Social Blade[40], que se especializa em prover estatísticas e informações que não são facilmente obtidas no próprio YouTube. Nesse *site*, é possível inserir o nome de um usuário do YouTube e, com isso, conferir várias informações sobre seu canal, incluindo o país de origem.

No entanto, essa não é uma garantia de que todos os vídeos contidos em determinado canal irão conter uma amostra da língua inglesa falada no país especificado, já que há outros fatores que influenciam nessa questão. Por exemplo: usuários do YouTube frequentemente recebem convidados de outros países e gravam vídeos com eles. Dessa maneira, para determinar com precisão o país de origem de cada vídeo da amostra, seria necessário assistir a todos os vídeos, um por um, o que é inviável dada a abordagem quantitativa do trabalho. O que fiz foi associar cada canal a um determinado país, partindo do pressuposto de que todos (ou a maioria) dos vídeos daquele canal conteriam a variedade do inglês proveniente daquele país. A desvantagem dessa abordagem é a possibilidade de que, por exemplo, um vídeo categorizado como proveniente dos Estados Unidos no CELV contenha uma amostra de língua inglesa falada em outro país.

---

[40] socialblade.com

A categorização de gênero e tema dos vídeos também apresenta limitações similares, decorrentes da quantidade de vídeos da amostra e impossibilidade de se categorizar cada vídeo individualmente. Por exemplo: foram categorizados como *vlogs* todos os vídeos provenientes de canais que se auto intitulam *vlogs* ou que contêm, em sua maioria, vídeos que se encaixam nesse gênero. No entanto, é comum que um usuário que possui um canal de *vlogs* decida gravar um vídeo que não se encaixe nesse gênero. O mesmo acontece com a categorização de temas. Há, portanto, certa imprecisão na categorização dos vídeos da amostra.

O resultado dessa etapa foi um *corpus* composto por cerca de 4.100.000 palavras, provenientes de cerca de 5.300 vídeos do YouTube, organizado em diretórios contendo os arquivos originais de legenda em formato SubRip. As características detalhadas do *corpus* serão expostas do capítulo de resultados.

Os arquivos originais coletados passaram, então, por um processo de formatação, para que pudessem ser interpretados pelo sistema do CELV, e, em seguida, por um processo automático de etiquetagem, para possibilitar o uso de etiquetas morfossintáticas nas buscas. Essa etapa será descrita na próxima seção.

3.2 Formatação e etiquetagem do *corpus*

As legendas extraídas na etapa anterior foram formatadas, visando simplificar e padronizar sua estrutura para que suas informações pudessem ser interpretadas pelo sistema computacional, e também preparar os arquivos para a etiquetagem do CLAWS.

Primeiramente, os arquivos foram convertidos do formato *.srt* para o formato *.txt*, com padrão de codificação UTF-8[41]. Em seguida, usei o programa gratuito TXTcollector[42] para combinar vários arquivos de legenda em arquivos únicos, contendo legendas de determinada subdivisão da amostra. Essa combinação foi feita para facilitar as etapas posteriores do processo, já que é mais fácil lidar com uma quantidade menor de arquivos unificados com grandes quantidades de texto por arquivo. O TXTcollector também insere um cabeçalho acima do texto correspondente a cada arquivo de legenda. O texto desse cabeçalho corresponde ao nome de cada arquivo (que foi padronizado na

[41] A definição da codificação UTF-8 foi feita pelos programadores responsáveis pela produção do sistema, para adequar os arquivos ao seu processo de desenvolvimento.
[42] bluefive.pair.com/txtcollector.htm

etapa anterior), e é utilizado nas etapas posteriores para recuperar informações relacionadas a cada legenda.

Nesse momento, os arquivos unificados pelo TXTcollector foram acessados usando o programa Notepad++[43], um editor de texto gratuito semelhante ao Bloco de Notas do Windows, porém com um número maior de funções, para formatar as marcações de tempo e os cabeçalhos de cada legenda. Para ilustrar essa parte do processo, considere-se o seguinte exemplo, que contém uma legenda em seu formato original, encimada por um cabeçalho inserido automaticamente pelo TXTcollector a partir do nome do arquivo, que aparece como demonstrado anteriormente pela Figura 13:

> @United States#HowTo#Cooking#Fifteen Spatulas#C6VuNiqVr6w.srt
> 1
> 00:00:00,190 --> 00:00:05,920
> Hey guys it's Joanne from FifteenSpatulas.com,
> today we are going to make some cheesecake

Após a formatação, essa legenda ficou assim:

> <‡United States†HowTo†Cooking†Fifteen Spatulas†C6VuNiqVr6w>
> <00:00:00>
> Hey guys it's Joanne from FifteenSpatulas.com,
> today we are going to make some cheesecake

Primeiramente, o cabeçalho que havia sido previamente inserido pelo TXTcollector foi colocado entre chaves anguladas (< e >). Essas chaves são necessárias porque o CLAWS, que será usado no momento da etiquetagem dos textos, é configurado para ignorar informações que estejam entre elas. Ainda, os caracteres @ e # foram substituídos por ‡ e †. Esses caracteres servem como indicadores de indexação, para apontar o início de cada cabeçalho e separar as informações de categorização de cada legenda (país, gênero, tema, canal e *link*). Foram escolhidos esses dois caracteres incomuns[44] para que, durante o processamento pelo sistema computacional, fossem facilmente identificáveis e distinguíveis do restante do texto. Os arquivos de legenda foram coletados em um computador com sistema operacional Windows, que não permite que os caracteres ‡ e † sejam usados nos nomes dos arquivos. Os caracteres @ e #, por

---

[43] notepad-plus-plus.org

[44] Poderiam ter sido escolhidos quaisquer outros caracteres incomuns, desde que não estivessem contidos nos textos que pertencem ao *corpus*.

sua vez, podem ser usados em nomes de arquivos do Windows, mas não puderam servir como indicadores de indexação, porque aparecem como texto normal em algumas das legendas. Isso foi o que justificou essa diferença de caracteres entre os nomes de arquivos e os cabeçalhos, e a necessidade dessa substituição durante a formatação.

Continuando a explicação sobre o processo de formatação, o próximo passo foi remover o número contido na primeira linha de cada legenda (no exemplo dado, aparece o número 1), que diz respeito à sequência de exibição das legendas durante sua reprodução em uma tela, mas não é uma informação importante para a ferramenta. Em relação à marcação de tempo, manteve-se a marcação inicial, que é a única necessária para se reproduzir os vídeos em momentos específicos, e removeu-se a marcação final. A marcação de tempo inicial foi, também, colocada entre chaves anguladas, para que seja ignorada pelo CLAWS.

Por fim, foi usada a ferramenta do CLAWS, disponível *on-line*[45], para etiquetar o conteúdo linguístico propriamente dito dos arquivos formatados. Como a versão *on-line* e gratuita do etiquetador suporta um número máximo de 100 mil palavras por uso, foi necessário inserir pequenas porções de texto múltiplas vezes até que toda a amostra estivesse etiquetada. O resultado da etiquetação é um texto no qual as informações entre chaves anguladas são mantidas intactas, e o restante do texto recebe etiquetas gramaticais, como no exemplo abaixo:

> <‡United States†HowTo†Cooking†Fifteen Spatulas†C6VuNiqVr6w>
> <00:00:00>
> Hey_UH guys_NN2 it_PPH1 's_VBZ Joanne_NP1 from_II FifteenSpatulas.com_NP1 ,_, today_RT we_PPIS2 are_VBR going_VVGK to_TO make_VVI some_DD cheesecake_NN1

O exemplo ilustra o formato final dos textos contidos na amostra do CELV, e contém três elementos: cabeçalho, marcação de tempo, e texto da legenda. Cada um desses elementos é de vital importância para o funcionamento da ferramenta.

O cabeçalho é o primeiro elemento, e aparece, ainda, entre as chaves anguladas inseridas com uso do Notepad++. Essas chaves foram importantes durante a etiquetagem pelo CLAWS, e não são necessárias para as próximas etapas; porém, não há motivos para que sejam descartadas, o que seria um trabalho desnecessário de formatação. Após a

---

[45] ucrel.lancs.ac.uk/claws/trial.html

primeira chave angulada, aparece o símbolo ‡, que indica para o sistema o início de um cabeçalho, e cada um dos cabeçalhos corresponde a um arquivo de legenda diferente. Após esse símbolo, estão inseridos: o país de origem da legenda que se segue (no exemplo, *United States*), seu gênero (no exemplo, *How To*), seu tema (no exemplo, *Cooking*), seu canal de origem (no exemplo, *Fifteen Spatulas*), e uma sequência de 11 caracteres que remete ao *link* de seu vídeo original no YouTube (no exemplo, *C6VuNiqVr6w*). Essas categorias são separadas umas das outras pelo símbolo †, e sempre aparecem nessa mesma ordem, para que o sistema possa identificar a que se refere cada informação. Dentro dos arquivos de texto que compõem o *corpus*, o cabeçalho aparece uma vez por legenda, em seu início, enquanto as marcações de tempo e os textos aparecem várias vezes ao longo de cada legenda, até o fim de sua extensão, quando começa uma nova legenda com novo cabeçalho, e assim sucessivamente.

A marcação de tempo é o segundo elemento, e foi obtida a partir dos arquivos de legenda originais, porém simplificada da seguinte forma: foi retirada a informação referente aos milissegundos, mantendo-se apenas as horas, minutos e segundos, e foi totalmente descartada a marcação de tempo final, mantendo-se a inicial. A marcação de tempo final não é importante para a ferramenta, cuja função é apenas reproduzir os vídeos a partir de determinado momento, mas não pausar automaticamente sua reprodução em um momento específico. A pausa fica a critério do usuário. Assim como nos cabeçalhos, foram mantidas as chaves anguladas nas marcações de tempo, não por serem relevantes para as próximas etapas, mas porque sua retirada seria desnecessária. A marcação de tempo serve para indicar ao sistema que o texto que se segue aparece em determinado momento do vídeo especificado pelo cabeçalho (no exemplo, 00:00:00).

O terceiro elemento é o texto da legenda, ou seja, a informação linguística propriamente dita. Por ser o objeto de análise dos usuários da ferramenta, é a única informação que aparece nas linhas de concordância, exceto pelas etiquetas morfossintáticas, que servem apenas para classificar cada palavra e permitir a busca por classes gramaticais, não aparecendo para o usuário final. Os cabeçalhos e marcações de tempo também não aparecem para o usuário da ferramenta, porque servem apenas para a organização interna do sistema e de suas funções.

Todas essas formatações foram feitas por mim com uso da ferramenta de substituição do Notepad++ (botões "Substituir" e "Substituir todos"), com o uso de expressões regulares, que podem ser assim definidas:

> Expressão regular é um recurso computacional que permite a identificação de padrões textuais ou sequências de caracteres. Por exemplo: as marcações de tempo nos arquivos de legenda seguem o padrão xx:xx:xx,xxx --> xx:xx:xx,xxx, onde x representa um algarismo qualquer que faz parte da informação de tempo. Esse padrão é recuperável pelo processador de expressões regulares do Notepad++, permitindo a formatação de todas as marcações de tempo em um dado documento com um único clique. (PEIXOTO, AFRA BRITO, 2015, p. 288)

Para que os textos formatados e etiquetados possam ser interpretados pela ferramenta e finalmente buscados pelo usuário final, é necessário que sejam indexados. A etapa de indexação será apresentada a seguir.

3.3 Indexação do *corpus*

As etapas 3 e 4 do desenvolvimento da ferramenta foram, em sua maioria, procedimentos computacionais, feitos por dois docentes e um discente do curso de Sistemas de Informação da Universidade Federal de Uberlândia, conforme as minhas orientações e com certa participação minha para esclarecimento de funções do sistema e, em alguns casos, reformatação dos textos. Uma vez que os detalhes estritamente computacionais do processo fogem do escopo deste texto, essas etapas serão explicadas de maneira simplificada e com foco em seus aspectos linguísticos.

A indexação é uma técnica computacional na qual uma porção de dados (neste caso, dados linguísticos na forma de texto escrito) é processada e reorganizada, de maneira que possa ser lida rapidamente por um computador. Mais especificamente, é:

> [...] a transformação de estruturas textuais em estruturas computacionais, a fim de organizar as informações de maneira otimizada para a implementação de um sistema de busca. Para exemplificar um processo de indexação, considere-se uma coleção de documentos de texto, cada um com um número variável de palavras, armazenada em um computador. A indexação consiste no isolamento de todas as palavras contidas nos documentos, reorganizando-as em uma lista enumerada (índice), na qual cada palavra aparece associada aos documentos em que ocorre. O resultado é uma listagem semelhante a um glossário do tipo que se encontra nas páginas finais de livros técnicos e científicos, no qual os termos específicos que foram usados no livro aparecem listados em ordem alfabética, seguidos das páginas em que aparecem na obra. Essa estrutura de dados permite o acesso rápido a informações específicas, possibilitando a consulta. (PEIXOTO, AFRA BRITO, 2015, p. 282)

O índice obtido nessa etapa do desenvolvimento da ferramenta, assim como os cabeçalhos, marcações de tempo e etiquetas morfossintáticas contidos nos textos, é uma informação que não aparece para o usuário final da ferramenta, e serve para a organização interna do sistema e interpretação computacional dos dados.

A etapa de indexação, por ser um procedimento estritamente computacional, foi feita, principalmente, pelos programadores do sistema; porém, a minha intervenção foi necessária em alguns momentos, porque a indexação de alguns caracteres demonstrou-se problemática. Por exemplo: alguns canais do YouTube usam em suas legendas o caractere ‘ para indicar aspas, enquanto outros usam o caractere ', ligeiramente diferente. Observou-se que o caractere ' provocou erros no Notepad++, causando a eliminação de grandes quantidades de texto durante a formatação, enquanto o caractere ‘ não apresentou nenhum problema. Por isso, foi necessário substituir todas as ocorrências de ' por ‘ em toda a amostra, corrigindo a formatação dos casos problemáticos.

Esse e outros problemas causados pela falta de padronização das legendas, uma decorrência do fato da amostra provir de uma grande variedade de canais diferentes, fizeram com que várias tentativas fossem necessárias até que a indexação ocorresse adequadamente. Quando se obteve uma amostra devidamente indexada, foi possível prosseguir para a próxima etapa do trabalho, o desenvolvimento da ferramenta.

3.4 Desenvolvimento da ferramenta

Essa etapa da metodologia envolveu a transferência do *corpus* indexado para um servidor na *internet* e a obtenção de um domínio (www.celvonline.com) para desenvolvimento da página com a interface de consulta ao *corpus*. Após sua indexação, o *corpus* do CELV foi armazenado em forma de banco de dados em um servidor do provedor *Amazon Web Services*[46]. O servidor realiza a comunicação entre o usuário e o banco de dados, processando as requisições de pesquisa no *corpus* e retornando os resultados da busca na página do CELV.

Na página, foram acrescentados elementos comumente encontrados em ferramentas de *corpora on-line*, como uma caixa para inserção do texto de busca e filtros de pesquisa, que correspondem às categorias dos textos que compõem a amostra (filtros por país, gênero, tema e canal). As consultas realizadas resultam em uma lista de

[46] Esse provedor foi escolhido por indicação dos programadores. Seu endereço é: aws.amazon.com.

frequência das palavras buscadas. É possível, então, clicar em cada palavra da lista de frequência para obter exemplos de seu uso na forma de linhas de concordância. Essa interface foi inspirada, principalmente, pelo COCA, embora de maneira simplificada.

O desenvolvimento da ferramenta e de sua página na *internet* também foram, em sua maior parte, procedimentos computacionais. Os principais detalhes dessa etapa da metodologia serão apresentados, resumidamente, a seguir.

De acordo com Peixoto e Afra Brito (2015), o CELV, como qualquer sistema computacional, funciona por meio de algoritmos, isto é, sequências de etapas especificadas por um programador para que um computador realize uma determinada tarefa. Os algoritmos do CELV foram escritos em linguagem de programação *JavaScript* e realizam as funções de contagem de frequência, geração de linhas de concordância, ordenação de linhas de concordância e criação de expressões *booleanas*, que são expressões lógicas para a realização de operações a partir de determinados operadores computacionais. Em algumas funções, os operadores lógicos interpretados pelo sistema computacional são inseridos pelo usuário por meio de caracteres específicos, que chamei de parâmetros, e estão apresentados no quadro a seguir.

Quadro 1: parâmetros de pesquisa do CELV.

Parâmetro	Descrição	Exemplo
*	Qualquer palavra.	*open the* * = encontra *open the door*, *open the eye*, *open the can*, etc.
|	Uma e/ou outra palavra.	*do a|an|the* = encontra *do a*, *do an*, e/ou *do the*.
{}	Lemas da palavra	*{break}* = encontra *break*, *breaks*, *breaking*, *broke*, e/ou *broken*.
[]	Classe gramatical	*do [nn*]* = encontra *do things*, *do exercises*, *do business*, etc.

Fonte: elaboração própria.

Escolhi alguns dos caracteres usados no COCA para desempenhar as mesmas funções, exceto no caso da lematização. Por exemplo: se inseridos no COCA, os caracteres * e | realizam as mesmas funções que no CELV. Os caracteres [], no COCA, servem tanto para a busca por lemas de uma palavra quanto para a busca com uso de etiquetas gramaticais. Já no CELV, [] servem apenas para a inserção de etiquetas gramaticais, e a busca por lemas deve ser feita com o uso de {}, conforme demonstrado pelo quadro. A diferenciação entre [] e {} no CELV foi requisitada pelos programadores para facilitar o desenvolvimento do sistema. Busquei especificar os mesmos caracteres

usados no COCA para que pesquisadores já acostumados com ele possam fazer buscas com facilidade também no CELV.

As funções de busca com os caracteres * e | foram desenvolvidas sem a necessidade de recursos linguísticos, usando apenas recursos computacionais. Já as funções de busca por classes gramaticais e por lemas precisaram de recursos linguísticos específicos para serem implementadas. A busca por classes gramaticais usa a informação da etiquetagem realizada pelo CLAWS, como previamente descrito. A busca por lemas, por sua vez, usa uma lista de lemas disponível gratuitamente[47], obtida na página do WST, que possui 14.762 grupos de lemas da língua inglesa. Embora não esteja completa e não seja atual, foi de fácil obtenção e suficiente para a implementação da lematização no CELV. O sistema do CELV utiliza a lista identificando a primeira palavra em cada grupo de lemas como a forma base, e associando as palavras seguintes como seus lemas, permitindo pesquisas como a exemplificada no Quadro 1.

As figuras a seguir ilustram cada um dos exemplos mencionados no Quadro 1.

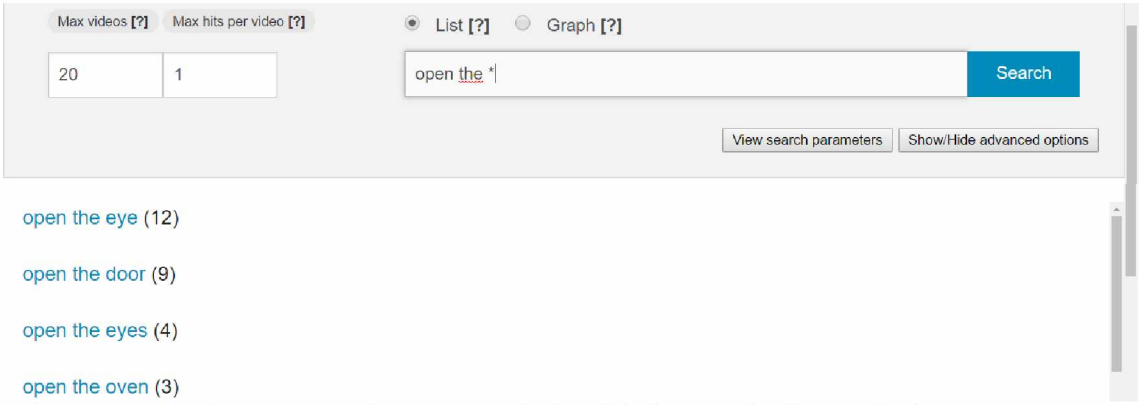


Figura 14: pesquisa por *open the* *, exibindo os 4 primeiros resultados.
Fonte: captura de tela do CELV.

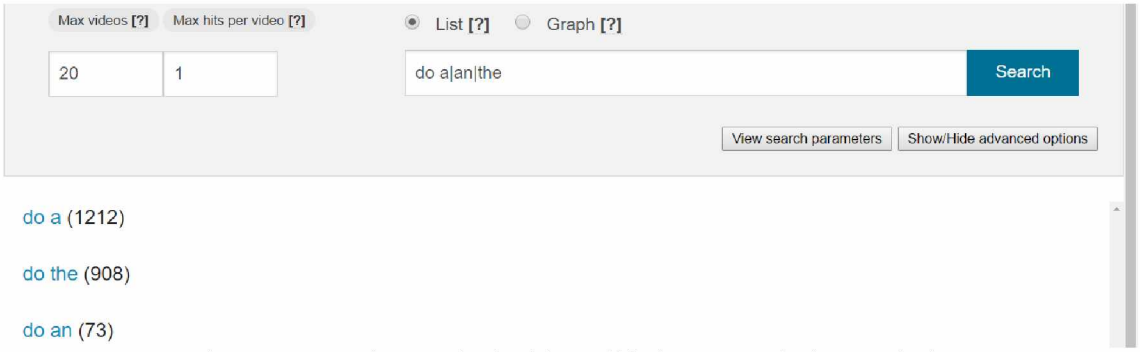


Figura 15: pesquisa por *do a|an|the*, exibindo os 3 resultados possíveis.
Fonte: captura de tela do CELV.

[47] lexically.net/downloads/BNC_wordlists/e_lemma.txt

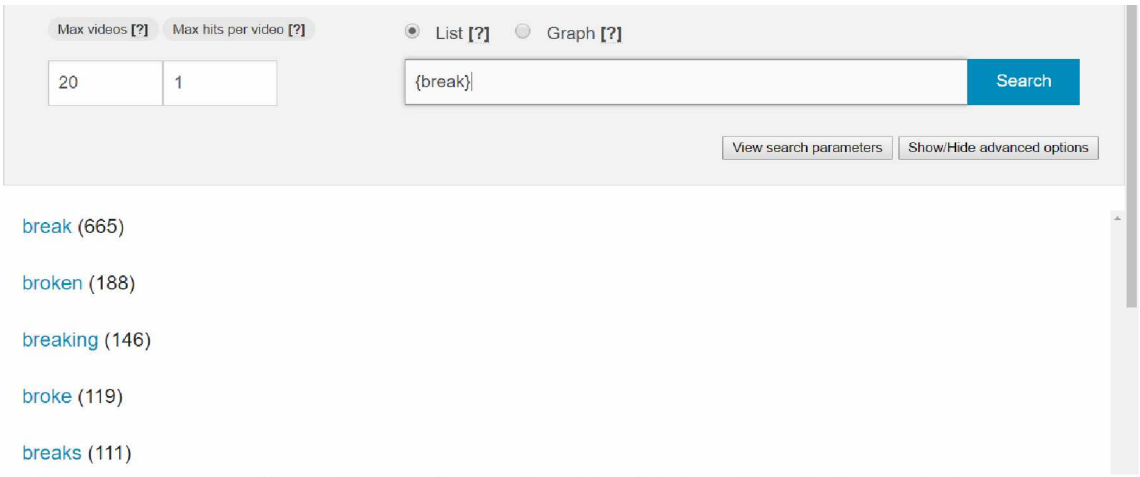


Figura 16: pesquisa por *{break}*, exibindo os 5 resultados possíveis.
Fonte: captura de tela do CELV.

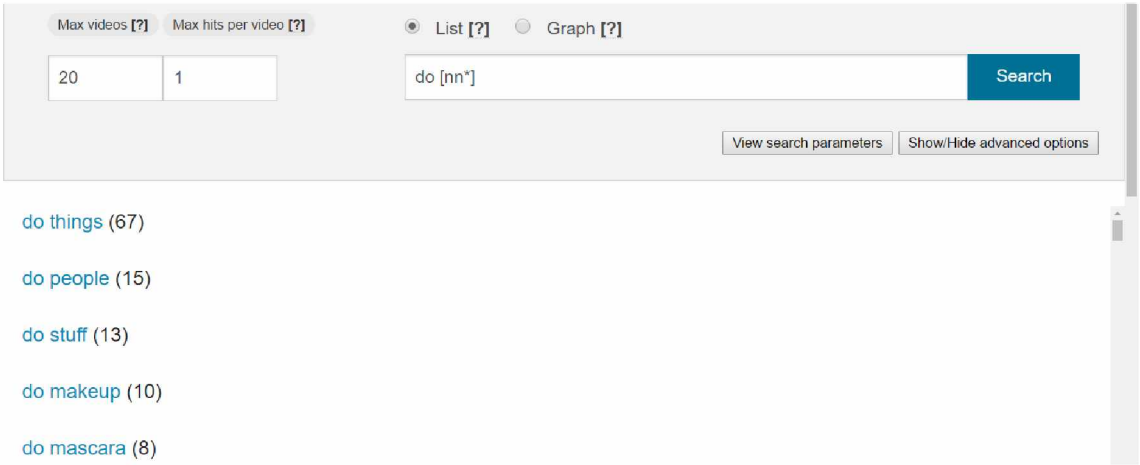


Figura 17: pesquisa por *do [nn*]*, exibindo os 5 primeiros resultados.
Fonte: captura de tela do CELV.

Outra função desenvolvida para a ferramenta foi a opção entre duas formas de pesquisa: lista e gráfico. A opção lista usa o banco de dados de legendas indexadas para realizar uma contagem de frequência dos termos de busca, apresentando-os em uma listagem com frequência decrescente. A opção gráfico, além de realizar uma contagem de frequência, agrupa os resultados da busca conforme os filtros de pesquisa selecionados pelo usuário, e retorna um gráfico de barras que apresenta os resultados da busca proporcionalmente distribuídos de acordo com os filtros. As figuras anteriores, 14, 15, 16 e 17, ilustram a opção lista. A Figura 18, a seguir, ilustra a opção gráfico, em uma pesquisa pela palavra *theatre* com todos os filtros de país selecionados.

Figura 18: pesquisa por *theatre* com a opção gráfico.
Fonte: captura de tela do CELV.

As barras do gráfico representam o uso da palavra pesquisada nos países contidos no *corpus*, de maneira proporcional ao tamanho da amostra de cada país. Nesse caso, a palavra ocorre 0,05 vezes a cada dez mil palavras na amostra da Canadá, e 0,33 vezes a cada dez mil palavras na amostra do Reino Unido. Assim, a palavra *theatre*, escrita desta maneira, é mais comum no Reino Unido do que nos demais países. Uma pesquisa usando a outra grafia da palavra, *theater*, apresentaria proporções diferentes. Esse exemplo demonstra a utilidade da opção gráfico na comparação entre diferentes variantes do inglês usadas nos países da amostra.

Uma outra função do CELV, e uma de suas principais, é permitir o acesso aos vídeos que compõem a amostra a partir do clique nas linhas de concordância. O desenvolvimento dessa função foi facilitado pela existência prévia de recursos do YouTube para reprodução de vídeos a partir de momentos específicos. Um *link* para um vídeo no YouTube possui, por exemplo, a seguinte estrutura:

www.youtube.com/watch?v=**C6VuNiqVr6w**

A parte em negrito no exemplo acima é uma sequência de 11 caracteres presente em cada *link*, e diferente em cada vídeo do YouTube, servindo para identifica-lo. Como explicado previamente, essa sequência de caracteres é uma das informações salvas nos cabeçalhos dos arquivos do *corpus* e posteriormente indexadas.

Depois da sequência de caracteres que identifica o vídeo, o *link* pode ser acrescido de vários parâmetros que especificam funções do YouTube. Por exemplo: para reproduzir o vídeo exemplificado acima a partir do tempo 1 minuto e 30 segundos, e não de seu início, pode-se usar o seguinte *link*:

www.youtube.com/watch?v=C6VuNiqVr6w**&t=1m30s**

Neste caso, a parte em negrito especifica a função do YouTube para reprodução do vídeo a partir de um determinado momento. O que o sistema do CELV faz é usar as informações contidas nos arquivos de legenda e indexadas no servidor para montar um *link* que aponta para o vídeo no momento de elocução de determinada fala de interesse. Aqui, é importante destacar que essa função que permite o acesso aos vídeos originais da amostra é a principal vantagem da ferramenta e é o que torna o CELV um *corpus* audiovisual, representando uma nova maneira de se trabalhar com *corpora*. Como exemplo da função, considere-se o início da legenda a seguir:

‡United States†2011†How To†Cooking†Stephanie Manley†ziT3KQV59p0
<00:00:09> Hi_UH ,_, this_DD1 is_VBZ Stephanie_NP1 Manley_NP1 ,_, and_CC welcome_NN1 to_II Copy_NN1 Kat_NN1 Recipe_NN1 where_CS I_PPIS1 make_VV0
<00:00:14> recipes_NN2 that_CST taste_VV0 just_RR like_CS they_PPHS2 do_VD0 in_II the_AT restaurant_NN1 ._.

Suponha-se que um usuário do CELV encontre o texto acima (sem as etiquetas morfossintáticas) em uma linha de concordância, após ter buscado, por exemplo, pela palavra *recipes*. Assim que o usuário fizer um duplo clique sobre essa linha de concordância, o sistema identificará que ela se refere ao vídeo do YouTube cujo *link* é ziT3KQV59p0, e que a fala de interesse do usuário aparece no momento <00:00:14>, e montará o seguinte *link*:

www.youtube.com/watch?v=**ziT3KQV59p0&t=14s**

A partir dessas informações, o sistema exibe o vídeo em uma nova janela dentro da própria interface do CELV, na qual o vídeo aparece incorporado e reproduzido a partir do momento especificado.

As outras funções desenvolvidas para o sistema são estritamente computacionais.

São elas: geração e ordenação alfabética de linhas de concordância, filtros de busca, ajuste do tempo inicial de reprodução de cada vídeo, ajuste da quantidade máxima de exemplos a exibir e ajuste da quantidade máxima de ocorrências por vídeo a exibir.

Por fim, foi inserido no código do *site* um *script* da ferramenta Google Analytics, que serve para monitorar o tráfego recebido nas páginas do CELV, gerando informações sobre o número de acessos, os países de origem dos usuários da ferramenta e outros detalhes. Essas informações foram armazenadas pelo Google Analytics em forma de gráficos, que serão apresentados no capítulo de resultados deste trabalho. Além disso, gravei três tutoriais em forma de vídeo com o objetivo de explicar as funções de pesquisa e ilustrar possíveis aplicações do CELV para o ensino-aprendizagem de línguas. Os vídeos possuem fala em inglês, legendas em inglês e português, e foram publicados em um canal do YouTube criado especificamente para o CELV[48], e também incorporados no *site* da ferramenta, na página intitulada Tutorial. O roteiro dos vídeos está disponível nos Apêndices 2, 3 e 4, e seu conteúdo será comentado na seção de resultados deste trabalho.

3.5 Teste da ferramenta

Como mecanismo para coleta de dados sobre a ferramenta, foi elaborada uma oficina intitulada "*Online corpus tools for English language teaching: exploring possibilities and testing the Corpus of English Language Videos*". Essa oficina foi ministrada a dez professores do programa Inglês sem Fronteiras da Universidade Federal de Uberlândia no dia 17 de julho de 2015, no laboratório de informática do curso de graduação em Letras da UFU, e teve duração de 2 horas. Durante o encontro, foram apresentados alguns conceitos fundamentais da LC e da ADD e, em seguida, os professores puderam ter contato com o COCA e outras ferramentas de *corpora on-line* e, por fim, com o CELV. A atividade final da oficina foi a elaboração de exercícios para a sala de aula que fizessem uso do CELV, levando em consideração o contexto de ensino de cada professor.

É importante esclarecer que, por limitações de tempo, foi feita apenas a oficina direcionada a professores, e não houve mecanismo de coleta de dados sobre o uso do CELV por alunos, embora a minha intenção seja que a ferramenta seja usada tanto por professores quanto por alunos. A coleta de opiniões e teste com alunos será um objetivo

[48] www.youtube.com/channel/UCWRLBQnHboBNW3EnQ8_69EA

futuro para continuação do desenvolvimento do CELV, e isso será retomando nas considerações finais deste texto.

Ao fim da oficina, foi aplicado a cada professor um questionário (Apêndice 5) contendo perguntas sobre sua opinião sobre o CELV e sobre aplicações pedagógicas de *corpora* na sala de aula. O questionário foi previamente digitado em um formulário elaborado na plataforma *Google Forms* e disponibilizado aos professores participantes da pesquisa por meio de um *link*[49]. O uso do *Google Forms* permitiu que as respostas dos professores ficassem gravadas e fossem, posteriormente, baixadas em um arquivo no formato *.xlsx*, que pôde ser explorado com uso do *Microsoft Excel* para análise dos dados.

Para realizar o teste com professores e aplicar o questionário mencionado, foi necessário submeter uma versão resumida deste projeto de pesquisa ao Comitê de Ética da Universidade Federal de Uberlândia. O projeto foi aprovado e a aplicação do questionário foi liberada. O processo ficou registrado por meio do Certificado de Apresentação para Apreciação Ética (CAAE) sob número 44388415.0.0000.5152.

As respostas ao questionário foram analisadas e comparadas aos resultados dos outros trabalhos que exploraram a ADD, apresentados previamente no capítulo teórico desta dissertação. A apresentação e discussão desses dados serão feitas no capítulo de resultados, a seguir.

[49] goo.gl/forms/OGP6GDbnFV

4. RESULTADOS

Esta pesquisa é caracterizada, principalmente, por resultar em um produto material: o *Corpus of English Language Videos*, e a sua ferramenta de consulta disponibilizada *on-line*. Assim, a natureza deste trabalho difere do modelo tradicional de pesquisa porque o resultado principal é o produto material em si, de forma que esta seção de apresentação dos resultados consistirá, em grande parte, na descrição desse produto.

No entanto, com o intuito de dar sentido à existência desse produto e nortear o seu uso por pesquisadores, professores e alunos de língua inglesa, faz-se necessário, também, tecer comentários sobre as suas possíveis aplicações linguísticas, isto é, sobre como o CELV pode vir a contribuir com o ensino-aprendizagem de língua inglesa. Por isso, como previamente mencionado, a etapa final deste trabalho se dedicou ao teste da ferramenta com professores de inglês.

Assim, os resultados alcançados pelo presente trabalho são três: o *corpus*, a ferramenta e sua página na *internet*, e os dados obtidos a partir do teste com professores. Esses resultados serão apresentados e discutidos nas seções seguintes, juntamente com uma seção especificamente dedicada a sugestões de uso do CELV para o ensino-aprendizagem de inglês.

4.1 O *corpus*

O *Corpus of English Language Videos* foi coletado ao longo do ano de 2014, e possui as seguintes características:

- **Número total de *tokens*:** 4.133.384
- **Número total de vídeos (legendas):** 5.344
- **Modo:** falado (registros escritos de fala em vídeos do YouTube)
- **Tempo:** sincrônico e contemporâneo (vídeos produzidos entre 2007 e 2014)
- **Seleção:** de amostragem, estático, não equilibrado
- **Conteúdo:** especializado (contém apenas legendas de vídeos do YouTube)
- **Autoria:** de língua nativa (Austrália, Canadá, Estados Unidos e Reino Unido)

A contagem de *tokens* do *corpus* está distribuída em subdivisões que correspondem a cada gênero, tema e país da amostra, como apresentado na Tabela 1.

Tabela 1: distribuição de *tokens* no CELV.

Gênero: Instructional (*How To*) – 2.261.912 *tokens*			
Tema: Beleza e Estilo – 772.115 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
148.691 *tokens*	22.129 *tokens*	300.761 *tokens*	300.534 *tokens*
Tema: Culinária – 934.920 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
225.220 *tokens*	109.004 *tokens*	300.462 *tokens*	300.234 *tokens*
Tema: Música – 554.877 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
52.860 *tokens*	66.832 *tokens*	232.518 *tokens*	202.667 *tokens*
Gênero: *Talks* – 388.318 *tokens*			
Tema: Meio Ambiente e Sustentabilidade – 92.582 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
12.110 *tokens*	10.854 *tokens*	15.226 *tokens*	54.392 *tokens*
Tema: Política e Sociedade – 157.753 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
29.204 *tokens*	28.636 *tokens*	46.685 *tokens*	53.228 *tokens*
Tema: Ciência e Tecnologia – 137.983 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
19.240 *tokens*	30.940 *tokens*	31.877 *tokens*	55.926 *tokens*
Gênero: *Vlogs* – 1.483.154 *tokens*			
Tema: Tópicos Gerais – 668.105 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
12.656 *tokens*	37.976 *tokens*	194.127 *tokens*	423.346 *tokens*
Tema: Científico e Educacional – 606.180 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
17.186 *tokens*	0 *tokens*	60.914 *tokens*	528.080 *tokens*
Tema: Viagem – 208.869 *tokens*			
Austrália	Canadá	Estados Unidos	Reino Unido
0 *tokens*	94.185 *tokens*	39.950 *tokens*	74.734 *tokens*
Total Austrália	Total Canadá	Total Estados Unidos	Total Reino Unido
517.167 *tokens*	400.556 *tokens*	1.222.520 *tokens*	1.993.131 *tokens*

Fonte: elaborado com base na contagem de palavras dos textos do *corpus* com uso do WST.

Percebe-se, pela distribuição de *tokens*, que o CELV não está balanceado em nenhum nível. Há uma quantidade muito maior de vídeos dos gêneros *How To* e *Vlogs* do que do gênero *Talks*. Adicionalmente, há grandes discrepâncias no número de *tokens* dentro de cada tema e país.

Apesar da importância do balanceamento de um *corpus*, a natureza do material linguístico utilizado para o CELV, isto é, vídeos do YouTube, impediu a criação de uma amostra perfeitamente balanceada. No YouTube, alguns países possuem uma produção de vídeos muito maior do que outros. Da mesma forma, certos temas são mais populares do que outros e, portanto, possuem um maior número de canais e vídeos. Nas subdivisões

do *corpus* que demonstram um número baixo ou inexistente de *tokens*, não foi encontrada amostra suficiente para que fossem equilibradas com as outras subdivisões, que correspondem a locais e tópicos com grande produção de vídeos. O gênero *How To*, por ser o mais popular, possibilitou certa aproximação de uma amostra equilibrada. Os outros gêneros possuíam uma quantidade muito pequena de vídeos provindos de determinados países, de maneira que optei por compensar com a coleta de uma quantidade muito maior em outros países. A maior dificuldade enfrentada durante a coleta foi que, mesmo quando existiam vídeos sobre os países e temas desejados, nem sempre esses vídeos possuíam legendas em inglês.

Consideradas essas limitações, procurei preencher o máximo possível cada subdivisão do *corpus*, buscando uma amostra com certa variedade linguística, o que seria vantajoso para a aprendizagem de línguas por prover exemplos de contextos variados. Ainda assim, a grande quantidade de vídeos sobre um único tema ou gênero gera certas tendências durante as buscas. No CELV, é muito mais comum encontrar palavras relacionadas a culinária, beleza, tecnologia e música do que outros temas.

Para ilustrar as palavras da amostra, a Tabela 2 compara os vinte e cinco substantivos mais frequentes no COCA e no CELV.

Tabela 2: 25 substantivos mais frequentes no COCA e no CELV.

	COCA			CELV		
Posição	Palavra	Quantidade absoluta	Frequência relativa	Palavra	Quantidade absoluta	Frequência relativa
1	PEOPLE*	902.490	0,1736%	TIME*	7.828	0,1894%
2	TIME*	830.659	0,1597%	PEOPLE*	7.211	0,1745%
3	YEARS	606.024	0,1165%	WAY*	6.635	0,1605%
4	WAY*	532.713	0,1024%	BIT	6.444	0,1559%
5	YEAR	409.439	0,0787%	HAIR	6.068	0,1468%
6	WORLD*	382.149	0,0735%	THING*	5.245	0,1269%
7	DAY*	377.528	0,0726%	THINGS*	4.514	0,1092%
8	LIFE	365.621	0,0703%	VIDEO	4.492	0,1087%
9	MAN	343.992	0,0662%	LOT*	3.858	0,0933%
10	SCHOOL	337.163	0,0648%	GUYS	3.391	0,0820%
11	PRESIDENT	313.832	0,0604%	DAY*	3.013	0,0729%
12	MR	302.560	0,0582%	KIND	3.012	0,0729%
13	STUDENTS	299.880	0,0577%	TOP	2.964	0,0717%
14	STATE	295.855	0,0569%	WATER	2.908	0,0704%
15	CHILDREN	295.373	0,0568%	SIDE	2.906	0,0703%
16	THINGS*	293.999	0,0565%	LOOK	2.755	0,0667%
17	HOUSE	289.323	0,0556%	STRING	2.752	0,0666%
18	WOMEN	266.695	0,0513%	FINGER	2.745	0,0664%
19	PERCENT	255.983	0,0492%	WORLD*	2.708	0,0655%

20	FAMILY	253.530	0,0488%	MINUTES	2.655	0,0642%
21	WORK	246.161	0,0473%	MUSIC	2.555	0,0618%
22	THING*	244.737	0,0471%	CHORD	2.463	0,0596%
23	CITY	231.294	0,0445%	BACK	2.395	0,0579%
24	LOT*	230.403	0,0443%	FOOD	2.386	0,0577%
25	PART*	226.097	0,0435%	PART*	2.372	0,0574%

Fonte: elaborado com base na busca pela etiqueta [nn*] nos dois *corpora*.

Na tabela, as palavras marcadas com asteriscos aparecem nas listas dos dois *corpora*. Assim, dentre as 25 palavras mais frequentes no COCA e no CELV, 9 aparecem em ambos. Além disso, as palavras *people*, *time* e *way* aparecem em posições iniciais nos dois *corpora*. Apesar dessas semelhanças, nota-se que a lista do CELV possui algumas especificidades: primeiramente, apresenta palavras relacionadas ao contexto do YouTube, como *video* e *guys*, que aparece comumente no início dos vídeos como uma saudação ("*Hey, guys!*"); além disso, apresenta palavras específicas, relacionadas aos tópicos com maiores quantidades de *tokens*, como *hair* (do tema beleza e estilo), *string, finger, chord e music* (do tema música), e *water* e *food* (do tema culinária).

Apesar dessa limitação, creio que o objetivo de obter uma amostra ampla e variada foi alcançado, já que, apesar da discrepância, é possível encontrar vídeos de vários assuntos, e a amostra possui um tamanho considerável, com mais de 4 milhões de palavras. Dessa forma, um professor ou aluno que realize uma consulta ao *corpus* conseguirá encontrar uma boa quantidade de exemplos daquilo que estiver buscando, com relativa variedade de contextos. Futuramente, quando houver uma maior produção de legendas em língua inglesa para os vídeos do YouTube, será possível balancear e, até mesmo, aumentar a amostra, para prover uma maior quantidade de exemplos de língua sem tendências para determinado assunto.

4.2 A ferramenta e a página na *internet*

O segundo resultado do trabalho é a ferramenta disponibilizada *on-line* em www.celvonline.com, cujas funções principais foram explicadas ao longo deste texto. Os vídeos intitulados *Tutorial 1: Introducing main search functions*[50], cujo roteiro está no Apêndice 2, e *Tutorial 2: Using advanced search options*[51], cujo roteiro está no Apêndice 3, demonstram todas as funções de pesquisa, e estão disponíveis no YouTube.

[50] www.youtube.com/watch?v=N-yqGpV2ozs
[51] www.youtube.com/watch?v=-qqWFFKmqB0

Para ilustrar a página da ferramenta na *internet*, criei o seguinte logotipo:

Figura 19: logotipo do CELV.
Fonte: elaboração própria.

A imagem contém a sigla CELV, com a letra V em posição horizontal, com um círculo e triângulo vermelhos ao fundo, imitando o botão de reprodução (*play*) usado nos programas e aparelhos de exibição de vídeos. Esse logotipo é usado na interface principal de busca do site, e também a página de abertura, demonstrada a seguir:

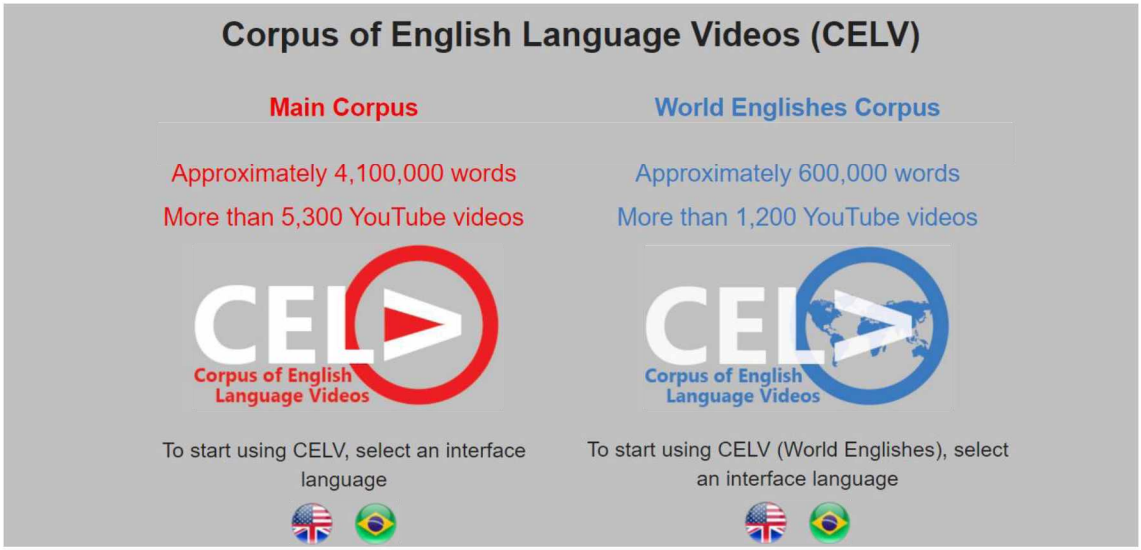


Figura 20: página de abertura do *site* do CELV.
Fonte: captura de tela de www.celvonline.com.

Essa página inicial serve para dar boas-vindas ao usuário e apresentar as informações quantitativas do *corpus*. Nessa página, é possível escolher entre duas amostras diferentes para consulta: o *Main Corpus* e o *World Englishes Corpus*. Todo o

texto desta dissertação se refere ao *Main Corpus*. O *World Englishes Corpus* é uma amostra de tamanho menor, mas que contém vídeos de uma grande quantidade de países, inclusive falantes não nativos do inglês. Essa amostra existe porque, durante a fase de coleta de legendas, deparei-me com alguns vídeos com variações interessantes do inglês, provindos de vários países do mundo, e percebi que poderiam ser uma ilustração da teoria de *World Englishes* de Kachru (1985). No entanto, a quantidade desses vídeos é pequena em relação aos da amostra principal, de forma que não foi possível encaixar essa amostra menor neste trabalho. Mesmo assim, mantive esse *corpus* no banco de dados do CELV. Ele pode ser consultado com todas as funções de pesquisa da ferramenta e poderá servir de inspiração e como amostra inicial para trabalhos futuros.

A página de abertura oferece duas opções de língua para acesso à interface do *corpus*: português e inglês. Ao clicar em uma das opções, o usuário é levado à página de busca, exibida na língua escolhida.

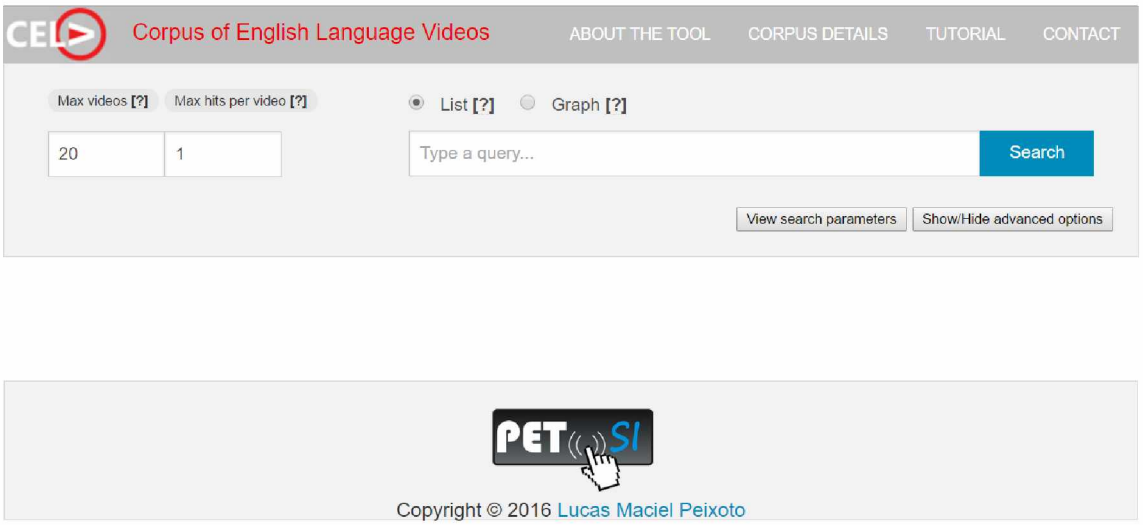


Figura 21: página de busca do *site* do CELV.
Fonte: captura de tela de www.celvonline.com/search.jsp?corpus=general&lang=en.

Observa-se que o CELV possui uma interface simples, buscando fornecer um mecanismo de pesquisa de fácil uso. Acredito que interfaces simples podem ser um fator de motivação para que mais linguistas, professores e alunos passem a explorar as possibilidades dos *corpora*, já que, como mencionado anteriormente, uma das barreiras encontradas para esse contato é o alto nível técnico das ferramentas de *corpus*.

Como demonstrado pela figura, além das funções previamente descritas neste texto e também nos vídeos, a página principal do CELV possui um menu superior com

links para as seguintes subpáginas: Sobre a Ferramenta (*About the Tool*), que exibe informações resumidas sobre a ferramenta e seus possíveis usos, além de especificações sobre as funções de pesquisa; Detalhes do *Corpus* (*Corpus Details*), que exibe as informações quantitativas do *corpus* e descreve suas características principais; Tutorial (*Tutorial*), que apresenta os três vídeos gravados e publicados no YouTube incorporados à página do CELV; e Contato (*Contact*), que indica o endereço de correio eletrônico criado para o CELV. Adicionalmente, é possível exibir uma mensagem informativa sobre cada função da interface por meio do posicionamento do ponteiro do mouse sobre os símbolos [?] que acompanham os elementos da página, como demonstrado a seguir.

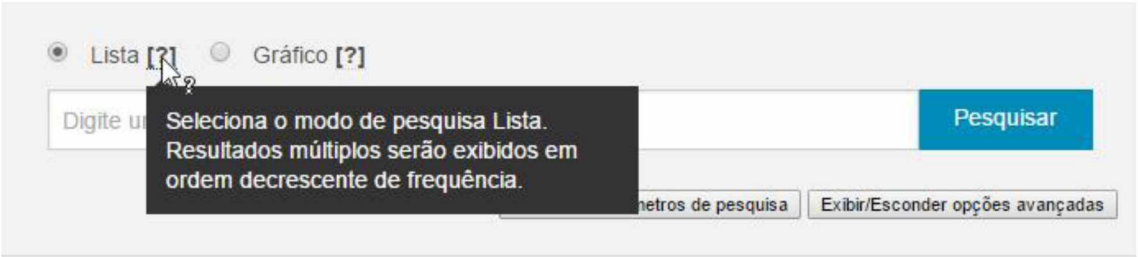


Figura 22: exemplo de mensagem informativa na interface do CELV.
Fonte: captura de tela do CELV.

Seguindo a linha de raciocínio da promoção de facilidade de acesso à ferramenta, as páginas Sobre a Ferramenta, Detalhes do Corpus e Tutorial, bem como a exibição de mensagens informativas na interface, têm o propósito de apresentar de maneira clara e detalhada as formas de se usar o CELV e as suas possibilidades de pesquisa, além de esclarecer as características da amostra. Como demonstrado no capítulo teórico deste trabalho, é necessário que o usuário de uma ferramenta como o CELV possua letramento de *corpus* para que possa tirar dela o máximo proveito, e o conteúdo informativo da página foi elaborado para ajudar a desenvolver esse letramento, servindo como um manual de instruções da ferramenta.

A página do CELV foi ao ar em outubro de 2014, embora nessa data nem o *corpus* e nem a ferramenta estivessem completos. Mesmo assim, a publicação da página foi útil para facilitar a comunicação com os programadores, uma vez que a visualização prévia do que estávamos produzindo permitiu o ajuste de detalhes e a identificação de problemas. A compilação do *corpus* e a implementação das funções de busca foram finalizados nos meses seguintes à publicação do *site*. No início de 2015, foi instalado o *script* do Google Analytics, mencionado anteriormente, para registrar o tráfego na página.

Esse registro de tráfego não estava previsto nos objetivos específicos deste trabalho, mas decidi implementa-lo para poder ter conhecimento sobre a quantidade de acessos à página e os lugares de origem desses acessos. Esses dados serão exibidos nas figuras seguintes.

	País	Sessões	Porcentagem do Sessões
1.	United States		
	Todos os usuários	1.303	30,55%
	Usuários que retornaram	17	5,54%
2.	(not set)		
	Todos os usuários	1.011	23,70%
	Usuários que retornaram	1	0,33%
3.	Brazil		
	Todos os usuários	335	7,85%
	Usuários que retornaram	156	50,81%
4.	China		
	Todos os usuários	227	5,32%
	Usuários que retornaram	6	1,95%
5.	United Kingdom		
	Todos os usuários	160	3,75%
	Usuários que retornaram	0	0,00%
6.	Russia		
	Todos os usuários	147	3,45%
	Usuários que retornaram	103	33,55%
7.	Japan		
	Todos os usuários	134	3,14%
	Usuários que retornaram	1	0,33%
8.	Germany		
	Todos os usuários	102	2,39%
	Usuários que retornaram	1	0,33%
9.	Netherlands		
	Todos os usuários	79	1,85%
	Usuários que retornaram	1	0,33%
10.	South Korea		
	Todos os usuários	76	1,78%
	Usuários que retornaram	1	0,33%

Figura 23: 10 países que mais acessaram o CELV entre janeiro de 2015 e junho de 2016.
Fonte: Google Analytics.

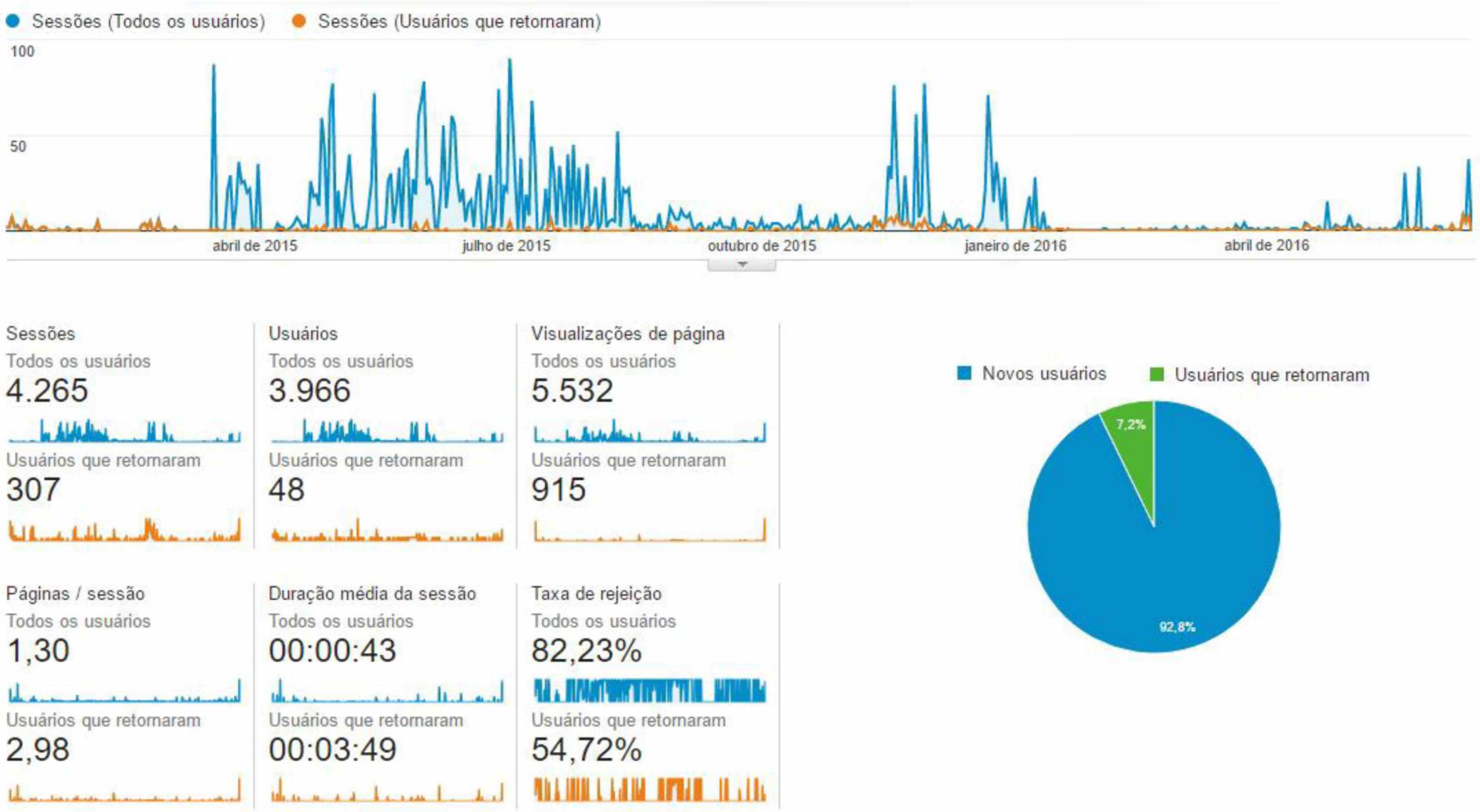


Figura 24: tráfego no CELV entre janeiro de 2015 e junho de 2016.
Fonte: Google Analytics.

A Figura 23 indica os dez países que acessaram o CELV com mais frequência entre janeiro de 2015 e junho de 2016. Nota-se que o país com o maior número total de acessos são os Estados Unidos, e outros países que acessaram a página com frequência total considerável foram o Brasil[52], a China, o Reino Unido e a Rússia. A categoria *not set* se refere a acessos cuja origem o Google Analytics não consegue registrar, e, portanto, não é possível retirar conclusões a respeito dessa categoria. Nas duas figuras, é importante destacar a diferença entre o número total de acessos e o número de acessos feitos por usuários que retornaram. Os usuários que retornaram são aqueles que acessaram a página mais de uma vez. Como o Google Analytics não fornece detalhes mais específicos sobre quem são esses usuários, as conclusões retiradas desses dados são apenas especulativas. Supõe-se que os usuários que retornaram são aqueles que consideraram a ferramenta interessante ou útil, e que os que não retornaram não consideraram a ferramenta útil, chegaram até o *site* acidentalmente, ou simplesmente não retornaram por outros motivos.

Assim, os dados sobre os usuários que retornaram são os mais relevantes para este trabalho. Os países com quantidades maiores de usuários que retornaram são: evidentemente, o Brasil, por ser o meu país e por terem ocorrido nele os acessos feitos por pessoas que conheceram o CELV durante apresentações que realizei em eventos acadêmicos ou para as quais demonstrei a ferramenta em outras ocasiões, além de outros possíveis acessos feitos por brasileiros; e, surpreendentemente, a Rússia, com um total de 147 acessos, dos quais 103 foram de usuários que retornaram. Com base apenas nos dados quantitativos oferecidos pelo Google Analytics, não é possível averiguar o propósito desses acessos, mas eles servem para fornecer uma noção sobre o alcance mundial da ferramenta e se pensar em, futuramente, desenvolver métodos para obter informações mais detalhadas sobre a forma como esses usuários estão usando a ferramenta.

A Figura 24 oferece dados mais específicos sobre os acessos: um gráfico de linhas, exibindo uma distribuição da quantidade de acessos ao longo do tempo; um gráfico circular, exibindo as quantidades de novos usuários e daqueles que retornaram; e outros dados quantitativos como o número de sessões (visitas ao CELV), o número de páginas visualizadas por visita, o número total de usuários, a duração média das visitas, as visualizações de página e a taxa de rejeição. Esses dados corroboram a suposição de que os usuários que retornaram são aqueles que consideraram a ferramenta útil, já que: (i) são

---

[52] O Google Analytics do CELV possui um filtro para que não registre os acessos feitos a partir do meu computador e nem a partir dos computadores dos programadores que trabalharam na ferramenta.

os usuários com maior número de páginas por visita (média de 2,98); (ii) são os usuários que gastaram mais tempo em suas visitas (duração média de 3 minutos e 49 segundos); e são os usuários com menor taxa de rejeição (média de 54,72%). A taxa de rejeição indica a porcentagem de usuários que encerraram a visita sem explorar a página, ou seja, abandonando a sessão na página de entrada.

Este subcapítulo apresentou as características da página do CELV e demonstrou os dados obtidos por meio da ferramenta de registro de tráfego Google Analytics. Em seguida, serão apresentados os resultados do teste da ferramenta com professores.

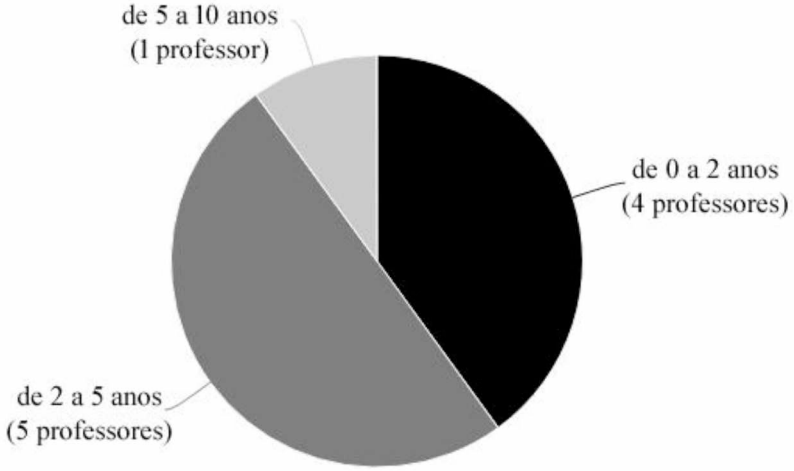
4.3 O teste da ferramenta com professores

Esta seção será dedicada a apresentar e discutir as respostas dadas pelos professores que participaram da pesquisa às perguntas contidas no questionário aplicado durante a oficina do CELV (Apêndice 5).

No momento de aplicação do questionário, todos os participantes da pesquisa eram alunos do Instituto de Letras e Linguística da Universidade Federal de Uberlândia, alguns em nível de graduação e outros em nível de pós-graduação. Vários deles eram professores em formação, e o trabalho no programa Inglês sem Fronteiras foi sua primeira experiência docente; outros, no entanto, já possuíam experiência com o ensino de inglês antes da participação no programa.

Como a amostra de participantes possui professores com diferentes níveis de experiência, a primeira pergunta do questionário foi "há quanto tempo você é professor de inglês"? As respostas para essa pergunta estão apresentadas no Gráfico 1, a seguir.

Gráfico 1: anos de experiência docente dos participantes da pesquisa.

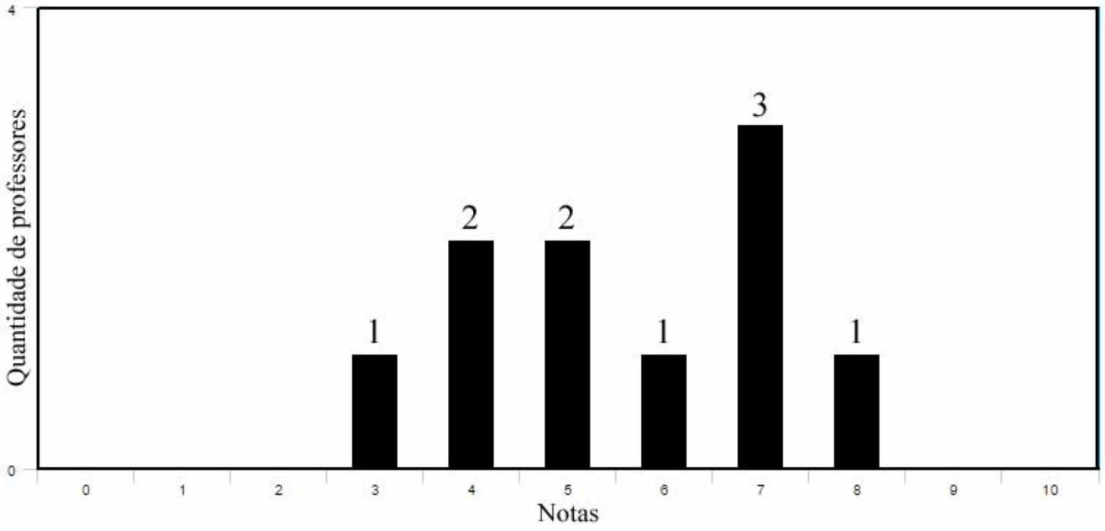


Fonte: elaborado com base nos dados de pesquisa.

A segunda pergunta do questionário está subdividida em duas partes. Na primeira parte, pede-se que os professores consideram a autenticidade da língua inglesa contida nos livros didáticos que já utilizaram em suas aulas, atribuindo a eles uma nota de 0 a 10, sendo 0 "totalmente artificial" e 10 "totalmente autêntico". Na segunda parte, deve-se justificar e/ou comentar a nota atribuída. O objetivo dessa pergunta foi averiguar a opinião dos professores sobre os exemplos de língua contidos em seus materiais didáticos, já que este trabalho parte da ideia de que materiais que apresentam usos reais da língua são vantajosos para a sua aprendizagem.

As notas atribuídas pelos professores aos seus livros didáticos quanto ao quesito autenticidade da língua usada foram distribuídas conforme o Gráfico 2, a seguir.

Gráfico 2: notas atribuídas pelos professores à autenticidade dos seus materiais didáticos.



Fonte: elaborado com base nos dados de pesquisa.

Conforme o gráfico, um professor atribuiu a nota 3, dois professores atribuíram a nota 4, dois professores atribuíram a nota 5, um professor atribuiu a nota 6, três professores atribuíram a nota 7, e um professor atribuiu a nota 8. Nenhum professor atribuiu as notas 0, 1, 2, 9 e 10. Os seis professores que atribuíram as notas mais baixas (entre 3 e 6) comentaram que: as faixas de áudio contidas nos livros são gravadas em estúdio de maneira artificial; os livros, às vezes, trazem textos autênticos, como de revistas, mas isso não é suficiente; os livros didáticos tendem a ser normativos, não apresentando a linguagem que os alunos poderão encontrar em situações reais; os livros não possuem diálogos próximos a situações reais; e que a autenticidade pode variar conforme o editor do livro. Os três professores que atribuíram nota 7 comentaram que:

algumas atividades propostas pelos livros são próximas de situações reais, mas frequentemente é necessário fazer adaptações; em alguns cursos ministrados, usaram material autêntico como filmes e seriados televisivos, mas, em outros cursos, usaram material não autêntico; e que os livros em geral trazem exemplos não autênticos, embora suas versões mais recentes busquem trazer textos de revistas e *sites*. O professor que atribuiu a nota 8 comentou que o material que estava usando no momento da pesquisa era baseado em pesquisas acadêmicas, trazendo conteúdo atual e relevante. As respostas dos professores a essa pergunta condizem com as informações mencionadas na introdução deste texto; os livros didáticos, de forma geral, trazem exemplos artificiais da língua, embora suas versões mais atuais estejam começando a trazer maior quantidade de textos autênticos.

A terceira pergunta do questionário pede que os professores afirmem se consideram a linguagem contida em vídeos do YouTube autêntica ou não, e comentem sobre como a linguagem dos vídeos pode ser comparada à contida nos seus livros didáticos no quesito autenticidade. Todos os professores afirmaram que consideram a linguagem dos vídeos autêntica, e comentaram que: os vídeos não são destinados especificamente ao ensino da língua; os vídeos são produzidos espontaneamente e, por isso, possuem uma linguagem mais natural; o conteúdo dos vídeos é a forma real como a língua é falada, sem facilitações; os vídeos são produzidos com propósitos reais de comunicação; e que os vídeos demonstram variações de pronúncia e de uso da língua. As respostas dos professores a essa pergunta estão de acordo com o que foi exposto no capítulo teórico deste texto sobre o uso de vídeos no ensino-aprendizagem de línguas: são recursos vantajosos por possuírem características típicas de contextos reais de comunicação.

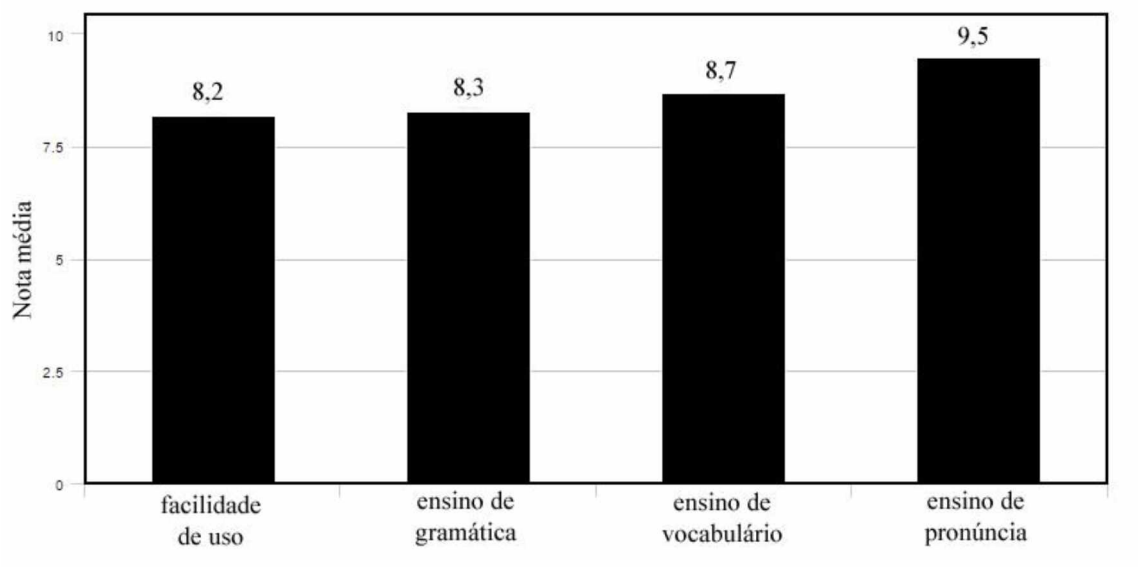
A pergunta quatro inquire aos professores se já haviam usado alguma ferramenta de *corpus* antes de aprender a respeito do assunto na oficina. Oito dos professores disseram que nunca haviam usado ferramentas de *corpus* anteriormente, e dois desses oito afirmaram que pretendem começar a usar *corpora* em suas aulas. Dois dos professores disseram que já haviam usado ferramentas de *corpus* anteriormente, para conferir o uso e padrões de colocação de determinadas palavras e para obter exemplos de frases para elaboração de exercícios. As respostas a essa pergunta demonstram que, nesta amostra de participantes, poucos já haviam buscado auxílio de *corpora* para suas atividades docentes, o que reflete as questões discutidas no capítulo teórico deste texto

sobre as dificuldades de incentivar o uso mais amplo de *corpora* nos contextos de ensino-aprendizagem.

A pergunta cinco inquire aos professores se já haviam ouvido falar da abordagem da ADD antes da oficina, e pede que indiquem pontos positivos e negativos dessa abordagem. Cinco dos professores nunca haviam ouvido falar da ADD, e os outros cinco já haviam ouvido a respeito em aulas da graduação ou da pós-graduação, mas nunca usaram a abordagem em sua prática docente. Os pontos positivos da ADD mencionados pelos professores foram: é uma maneira interessante de se trabalhar com vocabulário; fornece exemplos de língua mais próximos à realidade do aluno; coloca os alunos em contato com situações reais de comunicação a partir de uma grande variedade de exemplos, permitindo a aquisição de vocabulário e compreensão da estrutura e pronúncia; e possibilita que o aluno observe exemplos de variação linguística. Os pontos negativos foram: para implementação da ADD em sala de aula, é necessário usar instruções muito específicas, caso contrário é fácil se perder durante o uso das ferramentas; a abordagem é excessivamente acadêmica, o que pode não ser adequado para todos os tipos de público; a abordagem é mais voltada para alunos de nível avançado, que já possuem certo conhecimento do vocabulário e estrutura da língua; usar unicamente exemplos reais de comunicação, especialmente se houver muita variação, pode confundir o aluno; a implementação da abordagem depende muito de ferramentas computacionais como a *internet* e programas específicos. Tanto os pontos positivos quanto os pontos negativos mencionados pelos participantes da pesquisa corroboram os resultados obtidos nos outros trabalhos sobre a aplicação de *corpora* no ensino de línguas expostos no capítulo teórico desta dissertação, como as vantagens dessa abordagem para a aquisição de vocabulário e estrutura da língua e o contato com contextos reais de comunicação, mas também as dificuldades de implementação da abordagem fora do contexto acadêmico e as barreiras técnicas para o uso de *corpora*.

A pergunta de número seis está subdividida em quatro partes, e pede que os professores atribuam ao CELV notas de 0 a 10 nos seguintes quesitos: facilidade de uso, utilidade no ensino de gramática, utilidade no ensino de vocabulário e utilidade no ensino de pronúncia. O Gráfico 3, a seguir, apresenta a nota média recebida pelo CELV em cada um dos quatro quesitos, considerando todas as notas atribuídas pelos professores.

Gráfico 3: nota média recebida pelo CELV nos quatro quesitos da pergunta 6.



Fonte: elaborado com base nos dados de pesquisa.

A menor nota média recebida pelo CELV foi no quesito facilidade de uso, e a maior nota média foi no quesito ensino de pronúncia. Apesar de todas as notas terem sido altas, a diferença entre o quesito que recebeu a menor nota e o quesito que recebeu a maior nota é proporcionalmente grande, de maneira que é possível afirmar que os professores consideraram o CELV como um recurso especialmente útil no ensino de pronúncia. Além disso, não houve nenhuma nota abaixo de 7 dentre todos os professores e todos os quesitos, o que indica que, na opinião dos participantes da pesquisa, o CELV, de forma geral, é um recurso útil para o ensino-aprendizagem de línguas.

Diferentemente da questão seis, que pediu aos professores uma avaliação quantitativa do CELV, as questões sete, oito, nove e dez pediram uma avaliação qualitativa da ferramenta. A questão sete perguntou aos professores se usariam o CELV em suas aulas, e de que maneira. Todos os professores responderam que sim, das seguintes formas: para consultas de vocabulário; para demonstrar variações de pronúncia; para ensino de tópicos gramaticais; como complemento a outras atividades de fala ou de escuta; para demonstrar mudanças gramaticais e de sentido que determinadas palavras assumem em certos contextos; e para descobrimento de padrões de pronúncia de maneira indutiva. Os comentários dos professores nesta pergunta apontam para possíveis usos do CELV no ensino-aprendizagem de línguas, principalmente no que diz respeito ao estudo da pronúncia e suas variações, o que é, de fato, a principal vantagem de um *corpus* audiovisual.

A questão oito pede sugestões aos professores sobre recursos que podem ser

acrescentados ao CELV. Os recursos sugeridos foram: acréscimo de uma função que especifique o uso de uma palavra com uma determinada classe gramatical; formas de se realizar buscas sem que seja necessário conhecer as etiquetas morfossintáticas e os outros parâmetros de pesquisa; tornar a ferramenta mais compatível com dispositivos móveis; e maior quantidade e variedade de vídeos disponíveis na amostra. Todas essas sugestões apontam caminhos para o desenvolvimento futuro da ferramenta e ampliação da sua capacidade de contribuir com o ensino de inglês.

A questão nove pergunta aos participantes se, em sua opinião, o CELV seria bem aceito por professores de inglês. De forma geral, os participantes responderam que sim, já que a ferramenta oferece a oportunidade de elaborar atividades com base em situações comunicativas reais. No entanto, houve comentários sobre as dificuldades que podem aparecer nesse processo. Na opinião dos participantes, seria necessário que professores tivessem amplo conhecimento da ferramenta e dos princípios fundamentais da LC, por meio de cursos preparatórios. Os participantes também comentaram que, a princípio, esse tipo de ferramenta só seria aceito pelos professores mais abertos ao uso de novas tecnologias em sala de aula, e que atividades com uso do CELV necessitariam de um grande tempo para elaboração. Além disso, um dos participantes mencionou que a maior barreira para implementação do CELV em sala de aula seria a necessidade de um ambiente de laboratório.

A questão dez, por fim, pergunta aos participantes se, em sua opinião, o CELV seria bem aceito por alunos de inglês. Todos os participantes responderam que sim, pois os alunos seriam motivados a conferir o uso e pronúncia das palavras por meio dos vídeos, o que seria uma maneira dinâmica de aprendizagem. Além disso, os participantes mencionaram que, uma vez que o aluno tenha aprendido como usar o CELV, ele poderá acessar a ferramenta por conta própria, em casa. Alguns participantes mencionaram que para tirar total proveito da ferramenta, seria necessário que o aluno fosse autônomo e motivado a aprender.

A partir das respostas dos professores ao questionário, foi possível concluir que as características comuns referentes ao trabalho com *corpora* no ensino-aprendizagem de línguas que foram apresentadas anteriormente neste texto, tanto positivas quanto negativas, também apareceram nesta coleta de dados: a utilidade dos *corpora* na aprendizagem de padrões linguísticos e a vantagem do estudo de contextos reais de comunicação, e também as principais barreiras que costumam impedir o trabalho com

corpora nesse contexto. Mais especificamente, os dados demonstram as vantagens especiais de um *corpus* audiovisual no ensino da pronúncia. De maneira geral, os dados obtidos foram bastante elucidativos sobre o tipo de recepção que se pode esperar a uma ferramenta como o CELV no ensino-aprendizagem de línguas, e forneceram ideias norteadoras para a continuação do desenvolvimento da ferramenta em projetos futuros.

Além do teste formal da ferramenta em ambiente de laboratório com os professores participantes da oficina, tive a oportunidade de demonstrar a ferramenta de maneira informal durante a minha atuação como Professor Substituto no curso de graduação em Letras da Universidade Federal da Uberlândia. Esses momentos informais de demonstração, embora não registrados, ajudaram a encontrar indicações sobre como o CELV poderá ser usando no ensino-aprendizagem de línguas e quais são suas principais vantagens. Os comentários feitos nessas situações condisseram com os levantados formalmente durante a oficina, principalmente no que diz respeito às vantagens de uso de um *corpus* audiovisual.

4.4 Sugestões para uso do CELV no ensino-aprendizagem de inglês

Este trabalho sugere quatro possibilidades de uso do CELV no ensino-aprendizagem de língua inglesa. A primeira possibilidade é o uso do CELV para consultas de pronúncia. Todo *corpus* pode ser usado de maneira semelhante a um dicionário, no sentido que se está buscando por uma palavra específica para encontrar informações sobre ela. No caso de um dicionário, a busca é por uma definição, e no caso de um *corpus*, a busca é por um contexto de uso. Em um *corpus* audiovisual como o CELV, além do contexto de uso, é possível consultar a pronúncia das palavras ao se reproduzir os vídeos, encontrando exemplos de sotaque, entonação e contexto em um nível visual e auditivo.

A segunda possibilidade, também relacionada aos sons das palavras, é a audição de certos fenômenos de pronúncia como *Connected Speech*, ou fala conectada. Quando falamos, é comum unirmos o som final de uma palavra com o som inicial da próxima. Se a pronúncia das palavras for estudada isoladamente, esse fenômeno pode passar desapercebido. No entanto, como o CELV é uma amostra de linguagem autêntica, é possível ouvir exemplos de frases com fala conectada com frequência.

A terceira possibilidade é para a consulta de padrões lexicogramaticais em geral. Qualquer *corpus on-line* pode prover exemplos úteis para o estudo dos padrões de colocação das palavras, e o CELV possui a vantagem de exibir vídeos além das linhas de

concordância usuais. Um exemplo de padrão linguístico que pode ser estudado em um corpus on-line são os *Verb Patterns* do inglês, ou padrões verbais, como *stop to do*, *stop doing*, *like to do* e *like doing*. O recurso de busca por classes gramaticais permite a pesquisa por verbos que sigam determinados padrões de uso e a observação suas colocações com outras palavras, e as possíveis alterações de sentido decorrentes dessas combinações. Por exemplo: para se encontrar frases com uso do padrão *stop to do*, pode ser feita a pesquisa por *stop to [vvi]*, onde a etiqueta inserida se refere a qualquer verbo no infinitivo; e para se encontrar frases com o padrão *stop doing*, pode ser feita a pesquisa por *stop [vvg]*, onde a etiqueta especifica qualquer verbo no gerúndio.

Uma outra possibilidade de uso do CELV no ensino-aprendizagem de inglês é o estudo de *Stress Derivation*, ou derivação da sílaba tônica, fenômeno existente na língua inglesa no qual palavras homógrafas podem sofrer alterações de pronúncia da sílaba tônica, o que muda sua classe gramatical e, consequentemente, seu sentido. Exemplos de palavras que ilustram esse fenômeno são *record*, *permit* e *conflict*. Esse tipo de fenômeno não pode ser observado em *corpora* que contém linguagem somente escrita, já que as palavras possuem a mesma ortografia. Em um *corpus* audiovisual, no entanto, é possível consultar vários exemplos de cada uma dessas palavras e observar as alterações de pronúncia e de uso.

Essas sugestões de uso do CELV também estão registradas em forma de vídeo, no tutorial intitulado *Tutorial 3: Applications of CELV in English learning*[53], que foi publicado no canal no CELV no YouTube e também incorporado na página de tutoriais do *site* da ferramenta, e cujo roteiro se encontra no Apêndice 4.

Essas considerações encerram o capítulo de exposição dos resultados da pesquisa. Em seguida, serão apresentadas as conclusões e considerações finais.

[53] www.youtube.com/watch?v=DrSr1mmXfQw

5. CONSIDERAÇÕES FINAIS

No capítulo introdutório desta dissertação, foi apresentada a seguinte hipótese: o uso de uma ferramenta de *corpus on-line* com acesso a vídeos, ao possibilitar um contato empírico com a língua escrita e também oral, oferecerá ao aluno oportunidades de enriquecimento da sua aprendizagem do inglês. Com base nos resultados aqui apresentados, é possível chegar a uma confirmação parcial dessa hipótese, já que o teste da ferramenta foi feito com professores, mas não com alunos. Os dados obtidos por meio dos comentários dos professores, como demonstrado anteriormente, indicam que o CELV poderá ser um recurso útil na aprendizagem de inglês, especialmente pelo fato de ser um *corpus* audiovisual facilmente acessível por meio da sua página na *internet* e com uso de sua interface simples de busca. Para expandir essa conclusão, seria necessário testar a ferramenta em uma quantidade maior de contextos, com professores de vários perfis, e também com alunos. O teste com alunos não foi possível dentro das restrições de tempo da pesquisa, já que a maior parte do tempo dedicado ao trabalho foi gasta na compilação do *corpus* e desenvolvimento da ferramenta.

Quanto aos objetivos propostos, julgo que foram alcançados, com algumas limitações. O primeiro objetivo proposto foi compilar o *corpus* a partir da seleção de legendas de vídeos do YouTube. Como demonstrado, o *corpus* foi compilado e alcançou um número de mais de 4 milhões de palavras, embora não balanceadas. O objetivo implícito era que o CELV se aproximasse de uma amostra de língua inglesa geral, o que não pode se afirmar deste *corpus* em sua forma atual, se comparado a um *corpus* monitor como o COCA. Mesmo assim, o *corpus* serve o propósito de suprir exemplos da língua inglesa usada nos vídeos do YouTube, dentro dos gêneros, temas e países contemplados, e esses exemplos são úteis para a aprendizagem do inglês por meio das técnicas da ADD.

O segundo objetivo proposto foi desenvolver a ferramenta que permite consultas ao *corpus* e publica-la na *internet*. A função mais importante da ferramenta, que é a associação das linhas de concordância aos vídeos originais, permitindo sua reprodução a partir de momentos específicos, foi desenvolvida com sucesso. As outras funções descritas para os programadores também foram implementadas conforme as instruções, e funcionam corretamente. As limitações da ferramenta estão relacionadas a outras funções que não foram implementadas, mas já existem em outros *corpora on-line* como o COCA. Alguns exemplos de funções que não existem no CELV são: pesquisa por uma palavra usada em uma classe gramatical específica (por exemplo: busca por *water* sendo

usada, especificamente, como verbo) e pesquisa por palavras contendo parâmetros coringas no meio da palavra (por exemplo: busca por *dis*ed*, encontrando palavras como *discovered*, *disconnected* e *disoriented*).

O terceiro objetivo proposto foi levantar opiniões de professores de inglês sobre o funcionamento e utilidade do CELV. Esse objetivo foi alcançado por meio da aplicação do questionário e análise das respostas dos professores, obtendo dados que poderão nortear a continuação do desenvolvimento da ferramenta. Para encontrar mais formas de aprimorar a ferramenta e averiguar a sua utilidade, será necessário realizar o teste, também, com alunos de inglês de vários perfis e contextos.

Considerando as limitações aqui expostas e a reflexão sobre os dados obtidos durante o teste da ferramenta, alguns caminhos para desenvolvimento futuro do CELV são: expansão e balanceamento da amostra, para que ofereça uma maior variedade de exemplos de língua inglesa em uso, sem tendências sobre determinados temas; inclusão de novas funções de pesquisa na ferramenta, para aumentar as possibilidades de uso e as suas aplicações no ensino-aprendizagem de inglês; e testar a ferramenta em uma quantidade maior de contextos do seu público-alvo de usuários, com professores de perfis diferentes dos que participaram desta pesquisa e também com os próprios alunos.

A realização deste trabalho teve muito a acrescentar à minha formação acadêmica. Entrei em contato com os procedimentos para a elaboração de um *corpus* de grande tamanho, aprendendo muito sobre os critérios de compilação de *corpora* e os diversos fatores que devem ser considerados para a obtenção de uma boa amostra. Participei, dentro das minhas limitações, do processo computacional de desenvolvimento de um sistema linguístico de busca textual, aprendendo conceitos importantes e úteis para o trabalho com a Linguística Computacional, como a indexação e recuperação de dados e o uso de expressões regulares. Pude observar e formar opiniões sobre a aplicação de *corpora* ao ensino-aprendizagem de línguas, aprendendo sobre as vantagens e desvantagens desse tipo de abordagem e aprendendo, também, com as opiniões de outros professores sobre o mesmo tema. Em conclusão, espero que o meu trabalho com o desenvolvimento do CELV possa ser útil para pesquisadores, professores e alunos interessados no ensino-aprendizagem de língua inglesa, e represente uma nova e interessante possibilidade de aplicação de *corpora* nesse contexto.

REFERÊNCIAS BIBLIOGRÁFICAS

ACUNZO, C. M. **Uso de *corpora* para ensino de Língua Inglesa para profissionais de publicidade**. 151 f. Dissertação de Mestrado. São Paulo: Pontifícia Universidade Católica de São Paulo, 2012. Disponível em: <www.sapientia.pucsp.br//tde_busca/arquivo.php?codArquivo=14316>. Acesso em 20 de abril de 2015.

ALMEIDA, V. C. **Investigando colocações em um *corpus* de aprendiz**. 155 f. Tese de Doutorado. Belo Horizonte: Universidade Federal de Belo Horizonte, 2014. Disponível em: <www.bibliotecadigital.ufmg.br/dspace/handle/1843/MGSS-9PHMJ6?show=full>. Acesso em 20 de abril de 2015.

ALMEIDA FILHO, J. C. P. A Importância do Artigo de Edward M. Anthony (1963) e da sua Tradução Hoje. **HELB**, v. 1, n. 5, 2011.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para a pesquisa linguística. **Calidoscópio**, v. 4, n. 3, p. 156-178, 2006. Disponível em: <revistas.unisinos.br/index.php/calidoscopio/article/view/6002/3178>. Acesso em 01 de maio de 2016.

ANTHONY, L. **AntConc 3.4.3**. Tokyo. Waseda University, 2014. Disponível em: <www.laurenceanthony.net/software/antconc>. Acesso em 21 de fevereiro de 2015.

ANTHONY, E.M. Approach, Method and Technique. **English Language Teaching**, v. 17 (p. 63-67), 1963.

AZEVÊDO, C. **Movie Segments to Assess Grammar Goals** [*Internet*]. Brasília: Claudio Azevêdo. 2008. Disponível em: <moviesegmentstoassessgrammargoals.blogspot.com.br/>. Acesso: 10 de junho de 2016.

_____. **Movie Segments For Warm-Ups and Follow-Ups** [*Internet*]. Brasília: Claudio Azevêdo. 2009 – Disponível em: <warmupsfollowups.blogspot.com.br>. Acesso em 10 de junho de 2016.

BEILKE, N. S. V. Ach Jå! Fraseologismos em pomerano e em alemão. **Domínios de Lingu@gem**, v. 8, n. 2, p. 178-201, 2014. Disponível em: <www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/27732/15770>. Acesso em 15 de abril de 2015.

BANG, M.; FROMM, G. Terminologia em série: House M. D. In: **EntreLetras**, v. 4, n. 2. Araguaína: UFT, 2013. Disponível em: <www.revista.uft.edu.br/index.php/entreletras/article/download/995/533>. Acesso em 20 de abril de 2015.

BERBER SARDINHA, T. Como usar a Linguística de *Corpus* no ensino de língua estrangeira: por uma Linguística de *Corpus* educacional brasileira. In: VIANA, V.; TAGNIN, S. E. O. (Orgs.) **Corpora no ensino de línguas estrangeiras**. São Paulo, SP: HUB Editorial, 2010.

_____. **Linguística de *Corpus*.** 1ª Ed. Barueri: Manole, 2004.

BIEMANN, C. **Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm**. 199f. Tese (Doutorado em Ciência da Computação). Faculdade de Matemática e Ciência da Computação, Universidade de Leipzig, Leipzig, 2007. Disponível em: <wortschatz.uni-leipzig.de/~cbiemann/pub/2007/Biemann07diss_Structure-Discovery-final.pdf>. Acesso em 3 de setembro de 2014.

BORDAG, S. **Elements of Knowledge-free and Unsupervised Lexical Acquisition**. 263f. Tese de Doutorado em Ciência da Computação. Leipzig, Universidade de Leipzig, Faculdade de Matemática e Ciência da Computação, 2007. Disponível em: <wortschatz.uni-leipzig.de/~sbordag/BordagDiss.pdf>. Acesso em 3 de junho de 2014.

BOULTON, A. Data-Driven Learning: The Perpectual Enigma. In: GOZDZ-ROSZKOWSKI, S. (Org.), **Explorations Across Languages and Corpora**. Frankfurt: Peter Lang, 2011, p. 563-580.

BRAUN, S. ELISA - a pedagogically enriched corpus for language learning purposes. In: BRAUN, S., KOHN, K., MUKHERJEE, J. (Orgs.), **Corpus Technology and Language Pedagogy:** New Resources, New Tools, New Methods. Frankfurt, 2006, p. 25-47.

CARNEIRO, R. M. O. **Fraseologia e Padronização Linguística**: colocação, coligação, preferência semântica e prosódia semântica. Uberlândia, 2013. Trabalho não publicado.

ÇELİK, S. Develloping Collocational Competence Through Web Based Concordance Activities. In: **Novitas-ROYAL** (Research on Youth and Language), n. 5, v. 2, 2011, p. 273-286.

CHOMSKY, N. **Aspects of the Theory of Syntax**. Cambridge, Massachusetts: The MIT Press, 1965.

COBB, T. Is there Any Measurable Learning from Hands-on Concordancing In: **System**. v. 25, n. 3, 1997, p. 301-315.

CONTRERA, S. **Autenticidade em livros didáticos para o ensino de inglês como língua estrangeira**: um estudo diacrônico sob a perspectiva da Linguística de *Corpus*. 141 f. Dissertação de Mestrado. São Paulo: Pontifícia Universidade Católica de São Paulo, 2010. Disponível em: <www.sapientia.pucsp.br/tde_busca/arquivo.php?codArquivo=10984>. Acesso em 20 de abril de 2015.

DAVIES,M. **The Corpus of Contemporary American English**: 520 million words, 1990-present. 2008-. Disponível em: <corpus.byu.edu/coca>. Acesso em 21 de maio de 2016.

DUARTE, V. R. **Ensino e produção de material de inglês instrumental para a área de Tecnologia Ambiental com base na Linguística de *Corpus***: uma interface com a Linguística Cognitiva. 229 f. Dissertação de Mestrado. Santa Cruz do Sul: Universidade de Santa Cruz do Sul, 2011. Disponível em:

<btd.unisc.br/Dissertacoes/VitorDuarte.pdf >. Acesso em 20 de abril de 2015.

ERMAN, B.; WARREN, B. The idiom principle and the open choice principle. In: **Interdisciplinary Journal for the Study of Discourse**. v. 20, n. 1, p. 29–62, 2009.

FERREIRA, E. **Palavra frequente, pronúncia diferente**: A linguística de *corpus* auxiliando o ensino da pronúncia do inglês como língua estrangeira. 151 f. Dissertação de Mestrado. São Paulo: Pontifícia Universidade Católica de São Paulo, 2006. Disponível em: <www4.pucsp.br/pos/lael/lael-inf/teses/elias_ferreira.pdf>. Acesso em 20 de Abril de 2015.

FROMM, G. Lingüística Computacional: uma intersecção de áreas. In: **Revista Factus**. nº 5. Taboão da Serra: FTS, 2006. p. 135-140.

GABRIELATOS, C. Corpora and language teaching: Just a fling, or wedding bells? **TESL-EJ** 8(4), p. 1-37, 2005.

GARSIDE, R. The robust tagging of unrestricted text: the BNC experience. In: THOMAS, J., SHORT, M. (Orgs.) **Using corpora for language research**: Studies in the Honour of Geoffrey Leech. Londres, Inglaterra: Longman, 1996, p. 167 – 180.

GILMORE, A. Authentic materials and authenticity in foreign language learning. **Language teaching**, v. 40, n.2, 97-118, 2007.

_____. A comparison of textbook and authentic materials. **ELT Journal**, v. 58, n. 4, p. 263-274, 2004.

Halliday, M. A. K. Corpus Studies and Probabilistic Grammar. In: AIJIMER, K., ALTENBERG, B. (orgs.) **English Corpus Linguistics:** Studies in Honour of Jan Svartvik. Harlow: Longman. 1991. pgs. 30-43.

HALLIDAY, M. A. K. **Computational and Quantitative Studies**. v. 6. Londres, Inglaterra: Continuum, 2005.

HEATHER, J., HELT, M. Evaluating Corpus Literacy Training for Pre-Service Language Teachers: Six Case Studies**. Journal of Technology and Teacher Education**, 20(4), 415-440. Chesapeake, VA: Society for Information Technology & Teacher Education. Disponível em: <http://www.editlib.org/p/39324>. Acesso em 20 de Agosto de 2015.

JOHNS, T. F. Should you be persuaded: Two samples of data-driven learning materials. In: JOHNS, T.F.; KING, P. (org.). ***Classroom Concordancing***. Birmingham: ELR Journal, v.4, 1991, p. 1-16.

KACHRU, B. Standards, codification and sociolinguistic realism: the English language in the outer circle. In: QUIRK, Randolph; WIDDOWSON, Henry G. (Orgs.). **English in the world:** Teaching and learning the language and literatures. London e New York: Cambridge University Press, 1985. p. 11-30.

KENNEDY, C., MICELI, T. Corpus-Assisted Creative Writing: Introducing Intermediate Italian Learners to a Corpus as a Reference Resource. In: Language Learning & Technology. v. 14, n. 1, 2010, p 28-44.

KINDERMAN, C. A. E. **Linguística de Corpus e Multiletramentos**: uma nova interface pedagógica para a formação do professor de Língua Inglesa. 133 f. Tese de Doutorado. São Paulo: Universidade de São Paulo, 2011. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8147/tde-26052011-134021/pt-br.php> Acesso em 20 de abril de 2015.

KREYER, R. Corpora in the classroom and beyond. In: ZHANG, F.; BARBER, B. (Org.). **Handbook of research on computer-enhanced language acquisition and learning**. London: IGI Global, 2008, p. 422- 437.

MA, K. C. Learning strategies in ESP classroom concordancing: An initial investigation into data-driven learning. In FLOWERDEW, J., TONG, A (Orgs.), **Entering texts.** Hong Kong: Language Centre, The Hong Kong University of Science and Technology. 1994. p. 197–214.

McENERY, T.; WILSON, A. **Corpus linguistics**. Edinburgh, Edinburgh University Press, 1996.

MEUNIER, F. Corpus linguistics and second/foreign language learning: exploring multiple paths. **Revista Brasileira de Linguística Aplicada**, Belo Horizonte, v. 11, n. 2, p. 459-477, 2011.

MOREIRA FILHO, J. L. **Desenvolvimento de um *software* para preparação de aulas de inglês com *corpora***. 172 f. Dissertação de Mestrado. São Paulo: Pontifícia Universidade Católica de São Paulo, 2007. Disponível em: <www4.pucsp.br/pos/lael/lael-inf/teses/jose_lopes_moreira_filho.pdf>. Acesso em 20 de abril de 2015.

PEACOCK, M. The effect of authentic materials on the motivation of EFL Learners. **ELT Journal**, v. 51, n. 2, p. 144-53, 1997.

PEIXOTO, L. M. Identificação de unidades fraseológicas no vocabulário de Star Trek: abordagens *corpus-driven* e *corpus-based*. **Domínios de Lingu@gem**, v. 8, n. 2, p. 139-163, 2014. Disponível em: <www.seer.ufu.br/index.php/dominiosdelinguagem/article/viewFile/27630/15768>. Acesso em 15 de fevereiro de 2015.

PEIXOTO, L. M., AFRA BRITO, L. F Procedimentos para compilação de um *corpus* composto por legendas e construção de uma ferramenta de *corpus on-line*: o *Corpus of English Language Videos*. **Domínios de Lingu@gem**, v. 9, n. 3, p. 275-299, 2015. Disponível em: <http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/29266/16982>. Acesso em 18 de abril de 2016.

PULLUM, G. K. Computational Linguistics and Generative Linguistics: The Triumph of Hope over Experience. **EACL 2009 Workshop.** Atenas, Grécia, 2009. Disponível em: <aclweb.org/anthology-new/W/W09/W09-0104.pdf>. Acesso em 3 de maio de 2015.

QUEVEDO, A. G. Video use possibilities in autonomous learning. In: LEFFA, V. J. (ed.). **Autonomy in Language Learning**. Porto Alegre: Ed. Universidade/ UFRGS, 1994, p. 89-94.

RÖMER, U. 7. Corpora and language teaching. In: Lüdeling, Anke & Merja Kytö (eds.). **Corpus Linguistics**: An International Handbook (volume 1). [HSK series] Berlin: Mouton de Gruyter, 2008, p. 112-130.

_____. Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. In : GAST, V. (Org.). **The Scope and Limits of Corpus Linguistics** – Empiricism in the Description and Analysis of English. Special Issue: Zeitschrift für Anglistik und Amerikanistik, v. 54, n. 2, p. 121-134, 2006.

SAUSSURE, F. de. **Curso de linguística geral**. Tradução de Antônio Chelini et al. 28. ed. São Paulo: Cultrix, 2012.

SCOTT, M. Aprendizagem Direcionada por Dados: uma homenagem a Tim Johns (1936-2009). In: VIANA, V.; TAGNIN. S. E. O. (Orgs.) **Corpora no Ensino de Línguas Estrangeiras**. São Paulo: Hub Editorial, 2010, p. 7-11.

_____. **WordSmith Tools**. Versão 7. Stroud: Lexical Analysis Software, 2016a.

_____. **WordSmith Tools Help**. Liverpool: Lexical Analysis Software, 2016b.

SILERO, R. W. P. **Os quantificadores *a few* e *few***: questões de interlíngua e prosódia semântica em *corpus* de aprendizes. 107 f. Dissertação de Mestrado. Belo Horizonte: Universidade Federal de Belo Horizonte, 2014. Disponível em: <www.bibliotecadigital.ufmg.br/dspace/handle/1843/MGSS-9MQQ4B> Acesso em 20 de abril de 2015.

SINCLAIR, J. McH. Introduction. In: SINCLAIR, J. McH. (ed.). **How to Use Corpora in Language Teaching**. Amsterdam/Philadelphia: John Benjamins, v. 12., 2004, p. 1-10

_____. **Corpus, Concordance, Collocation**. 1ª Ed. Oxford: Oxford University Press, 1991.

_____. (org.) **Looking Up**. An Account of the COBUILD Project in Lexical Computing. London: HarperCollins, 1987.

TAGNIN, S. E. O. **O jeito que a gente diz**. São Paulo: Disal, 2005.

TARTONI, M. R. R**. A Linguística de *Corpus* no ensino do inglês**: Um estudo empírico exploratório com atividades com *to* e *for*. 164 f. Dissertação de Mestrado. Belo Horizonte: Universidade Federal de Belo Horizonte, 2012. Disponível em: <www.letras.ufmg.br/poslin/defesas/1497m.pdf> Acesso em 20 de abril de 2015.

VIANA, V. Linguística de Corpus: Conceitos, Técnicas & Análises. In: VIANA, V.; TAGNIN, S. E. O. (Orgs.) **Corpora no ensino de línguas estrangeiras**. São Paulo, SP: HUB Editorial, 2010.

APÊNDICES

Apêndice 1: Instruções aos programadores para desenvolvimento do CELV

1) O que é o CELV?

CELV é um *corpus*. Um *corpus* é uma coleção de textos sobre um determinado tema ou tópico.

Atualmente, os textos de um *corpus* sempre estão em forma de arquivos de texto (geralmente, o simples .txt) em computador, para poderem ser analisados por meio de programas de análise lexical.

Os programas de análise lexical mais usados são:

- Wordsmith Tools (http://www.lexically.net/wordsmith/)
- AntConc (http://www.antlab.sci.waseda.ac.jp/antconc_index.html)

Também existem *corpora* (plural de *corpus*) disponibilizados na *internet*, com ferramentas de busca e análise integradas no próprio *website*, como, por exemplo:

- COCA (Corpus of Contemporary American English): http://corpus.byu.edu/coca/
- BNC (British National Corpus): http://www.natcorp.ox.ac.uk/

Assim como o COCA e o BNC, o CELV será disponibilizado online e seu sistema deverá ter uma ferramenta de busca e algumas outras ferramentas parecidas com as desses dois *corpora*.

Os textos do CELV serão **legendas de vídeos do YouTube, em inglês**. Alguns produtores de vídeos no YouTube transcrevem sua fala, e disponibilizam essas legendas em seus vídeos. O YouTube chama essas legendas de *Closed-Captions*.

Para coletar essas legendas, está sendo usado o *software* **Google2SRT** (http://google2srt.sourceforge.net/en/).

Esse *software* extrai um arquivo .srt (o formato mais usado para legendas) a partir dos links dos vídeos. É um arquivo de texto simples, igual ao .txt, e pode ser lido pelo Bloco de Notas.

Somente são usados vídeos que já tenham legendas produzidas manualmente por seus autores. Não são baixadas as legendas produzidas pelo ***automatic captioning*** do YouTube, pois esse tipo de transcrição produz legendas com muitos erros.

Uma vez extraídos, os arquivos .srt são renomeados seguindo uma nomenclatura que poderá facilitar o desenvolvimento do sistema, possibilitando que as informações sobre cada arquivo possam ser resgatadas a partir do nome do arquivo.

As informações importantes são: país, ano, gênero, nome do canal e link para o vídeo.

2) Sobre a interface

A proposta inicial para a interface do site do CELV é usar uma organização parecida com a do COCA (http://corpus2.byu.edu/coca/):

- Menu com a ferramenta de busca do lado esquerdo da tela.
- Interface para visualização de informações sobre ano, gênero e país dos termos pesquisados, do lado direito e em cima.
- Interface para visualização de **Linhas de Concordância**, no lado direito e embaixo.

Detalhes sobre as opções na interface de busca:

- **Format: List / Graph**
- A opção **List** deverá exibir os resultados da busca como uma lista simples, ordenada conforme a frequência de ocorrências;
- A opção **Chart** deverá exibir os resultados em gráficos de barras, demonstrando o número de ocorrências dos termos buscados distribuídos de acordo com os anos, países e gêneros.
- Ambas as opções devem permitir que o usuário clique em um resultado (seja na lista ou nas barras dos gráficos) para visualizar esse resultado em linhas de concordância, na interface debaixo.

- **Filters: country, genre**
- Filtram a busca para exibir resultados somente de determinado país ou gênero

- **Maximum results**
- Aqui, o usuário poderá digitar um número, e então a busca encontrará, no máximo, aquela quantidade de ocorrências. Talvez programar para que esse número não passe de, por exemplo, 100 ou 1000, para não travar o sistema em buscas muito amplas.

- **As Linhas de Concordância**
- As linhas de concordância devem exibir os resultados da busca com o termo procurado no centro, ladeado pelo restante da frase em que se encontra, tanto na esquerda quanto na direita, até um determinado número de caracteres.
- O usuário deverá ter a opção de ordenar alfabeticamente as linhas de concordância conforme as palavras à esquerda ou à direita da palavra central (termo buscado).

3) Suporte ao uso de TAGS em pesquisas

Um usuário do site deverá ter a opção de pesquisar, por exemplo:

[VVI] the door

Onde "[VVI]" é uma TAG (etiqueta) que simboliza qualquer verbo no infinitivo.

Para isso, o *corpus* será previamente etiquetado, especificando a classificação (substantivos, verbos, etc.) de todas as palavras contidas nas legendas.

O formato de etiquetagem usado já existe e será o sistema CLAWS - *Constituent Likelihood Automatic Word-tagging System* (http://ucrel.lancs.ac.uk/claws/trial.html).

O sistema do site deverá ser capaz de "entender" as tags usadas pelo CLAWS (http://ucrel.lancs.ac.uk/claws7tags.html), e interpretá-las nas pesquisas feitas pelos usuários, retornando os resultados correspondentes.

Outra tag que deve poder ser usada em buscas é o símbolo *, que significa "qualquer palavra". Isso deve possibilitar buscas como:

play the *

Resultando em qualquer frase (dentro do *corpus*) que comece com "play the" e seja seguida de mais uma palavra qualquer, como por exemplo:

play the game, play the guitar, play the song, etc…

4) Lematização

Um usuário deverá ter a opção de pesquisar, por exemplo:

[open] the door

O uso dos colchetes na palavra "open" significa que os resultados devem exibir qualquer forma da palavra open: open, opens, opened, opening

O sistema poderá fazer isso a partir de uma **lista de lemas**, já previamente disponível.

5) Links para os vídeos no YouTube

Os links para os vídeos no YouTube não deverão apenas reproduzir o vídeo a partir do tempo "00:00", mas sim a partir da marcação de tempo correspondente ao resultado de busca. Por exemplo, na legenda abaixo:

00:01:35,670 --> 00:01:38,950
I have a standard kitchen knife, just a butter knife.

Nota-se, pela marcação de tempo, que essa frase é dita no tempo 01:35. Portanto, a ferramenta deverá produzir um link para o YouTube de forma que o vídeo seja reproduzido a partir desse momento, e não do início. O próprio YouTube já disponibiliza uma opção para incorporar vídeos e reproduzi-los a partir de determinado momento.

Apêndice 2: Roteiro do vídeo *Introducing main search functions*

Hello. In this video, I'm going to introduce the Corpus of English Language Videos and show you some of the things you can do with this corpus.

The Corpus of English Language Videos, or CELV, is an online corpus tool for linguistic queries. It is similar to other online corpus tools you may know. Currently, the sample has over 4 million words, and many functionalities found in this kind of tool.

What is special about CELV is it was compiled using YouTube video subtitles. This means that all texts contained in the sample come from YouTube videos, which can be accessed from the corpus. Only subtitles written manually by YouTube channel owners were used, not the ones produced automatically by YouTube. To ensure that the sample has some variety of texts, videos from different themes were selected, including cooking tutorials, Ted Talks, vlogs and many others.

To start using CELV, navigate to celvonline.com.

Currently, there is the Main Corpus and the World Englishes corpus. The Main Corpus contains the main sample of texts and includes video subtitles from Australia, Canada, the UK and the USA. The World Englishes corpus contains a smaller but more varied sample from many different countries of the world, including non-native speakers of English. At the moment, this is a small sample, which was compiled for future researches.

Let's start exploring CELV using the Main Corpus, selecting the English language interface.

Here, you see a simple search interface, similar to the one you would see in Google, for example. You can type any text in English in the search box to check it in the corpus. Let's look for the word 'language'. Type 'language' in the search box and click on Search, or press Enter.

You will see the word 'language' listed. The number in parenthesis shows the frequency of this word in CELV. In this case, the word 'language' has 356 occurrences in the corpus.

Click on the word 'language' to open a list of concordance lines, showing examples of that word.

In this screen, you see information about each concordance line: the country where it came from, the genre of the video, the theme of the video, the YouTube channel who produced it, and the text itself, centered on the search word.

Notice that you can double click on any concordance line to watch the original video. For instance, let's click on line number 7, which reads 'Who invented the German language anyway'?

As you can see, the video is incorporated in a new window and starts playing a few seconds before the moment when the speaker says the sentence we were interested in.

You can control this delay by going back to the search interface and clicking on 'Show/Hide advanced options'. You will then see the option 'Start videos 2 seconds earlier', and you can adjust the amount of seconds, which is 2 by default. To demonstrate, let's change this value to 8 seconds, and open that same video.

This time, the video started 8 seconds before the sentence 'Who invented the German language anyway'? Usually, it is not necessary to adjust this option, but it is there in case you are interested in seeing more context from the video.

You can click on 'Show/Hide Advanced options' again to hide this part of the interface if you wish.

You will notice that this concordance list currently goes up to 20 concordance lines. You can change the amount of concordance lines to view more examples, by adjusting the value in 'Max videos'. Let's change this value to 50 and generate the concordance list again by clicking on the word 'language' one more time.

We now have the same 20 examples from before, with 30 more, for a total of 50.

By default, this list is configured so you don't see more than one example from each video. For instance, in this list with the word 'language', all 50 lines come from different videos – you don't see repetitions on the word 'language' in the same video. You can change this by adjusting the value in 'Max hits per video'. Let's change this value to 3 and click on 'language' again to generate a new list.

This new list has up to 3 instances of the word 'language' in each video, so we have new examples to work with. You will also notice that the number of examples is now 103. This means that, from the number of 50 videos which we configured previously, 103 total occurrences of the word 'language' were found.

Notice that the number in 'Max videos', which we set to 50, multiplied by the number in 'Max hits per video', which we set to 3, equals 150. However, this doesn't mean that you should expect to find 150 concordance lines, because some videos may contain less than 3 occurrences of the search word. In this case, as we saw previously, we have a total of 103 lines.

To clean the screen and start a new search, you can click on the CELV logo, on the top left of the screen. As you can see, resetting the screen changes all options back to their default values, so we are back to 20 'Max videos' and 1 'Max hit per video'.

Now let's explore the filters we can add to our search. To do this, click on 'Show/Hide advanced options'. You will see that your searches can be refined by a number of filters.

The current sample contains 3 genres of videos you can choose from: How To, Talks, and Vlogs. You can search in all 3 genres or choose a specific one.

The sample also contains videos from 4 countries: Australia, Canada, United Kingdom and United States.

Finally, the sample contains a number of themes you can choose from. Notice that each theme is associated with a genre. There are How To videos of 3 different themes: Beauty and Style, Cooking, and Music. There are Talks of 3 different themes: Environment and Sustainability, Politics and Society, and Science and Technology. And there are Vlogs of 3 different themes as well: General Topics, Scientific and Educational, and Travel.

If you filter your search by a specific genre, you will see that selecting a genre limits the amount of themes available, because each theme is associated to a genre. If you click again on 'Search all', then all options of themes become available again.

The buttons 'Mark/Unmark all' and 'Invert Selection' can help you quickly select the filters you are interested in.

To demonstrate these filters, let's suppose we want to look at the pronunciation of a certain word in different countries. A common example of difference in pronunciation in the English language is the word 'water'. Let's type 'water' in the search box.

Now, let's watch and listen to examples of the pronunciation of this word from the United States.

Let's compare this pronunciation with the United Kingdom, by changing the country filter and repeating the search.

As you can see, the ability to watch and listen to each video from the corpus allows interesting comparisons and other types of linguistic studies.

Now, let's look at the channel filter. Our current concordance list contains a variety of channels to see examples from, but suppose we only want to see examples from the channel 'This Is Genius'.

Type the name of the channel in the channel filter box. It must be typed exactly as it appears in the concordance list, including capitalization. You will notice that the filter will help you complete the name of the channel by showing it to you.

Click on search again, and you will notice we have a much smaller number of occurrences of the word 'water' now. That is because this search is limited to only one channel. By clicking on the word 'water', we can see examples from the channel 'This Is Genius'.

One more thing you can do to help organize your concordance lines is to sort them. Let's sort our current concordance list by the first word to the right on the search word. To do so, click the first 'Sort' option and select R1, which means the first word to the right. Click on the word 'water' again to generate a new list. This time, the list is sorted alphabetically by the first word to the right of 'water', starting with punctuation marks.

You can add up to three sort options to further organize your concordances. Let's sort first by R1, then by R2, and then by R3. You'll notice that each sort option corresponds to a different color in the concordance lines. This can be helpful for analyzing long lists of concordances in search for linguistic patterns.

That's all for this video. We explored the main search functions of the CELV interface with simple searches containing only one word. To learn about how to do more complex searches, including the use of part-of-speech tags, lemmatization and more, watch the other videos in this channel or navigate to the tutorial page in the CELV website. Thank you for watching.

Apêndice 3: Roteiro do vídeo *Using advanced search options*

Hello. In this video, we will learn how to use the advanced search options in CELV, to enable complex searches in the corpus. Before watching this video, it is recommended that you watch the first video tutorial available in this channel, which introduces CELV and teaches how to use its basic functions.

To start using advanced search options in CELV, notice that, in the website, you can click the button 'View search parameters' to look at some of the input formats available.

As you can see, the first parameter we can use is the asterisk symbol. It symbolizes any word, and can be used when you want to leave a space open for any word in a given sentence. Here, we see an example with the phrase 'open the' and then the asterisk symbol. Let's type this example in the search box to see what happens.

As you can see, many phrases starting with 'open the' were found, followed by any word in the position where the asterisk was, such as 'open the eye', 'open the door', 'open the eyes', and so on. Notice that this list is displayed from top to bottom in descending order of frequency, which is the number between parenthesis.

The asterisk can be used in any position in a phrase: not only at the end, but also in the middle or at the start.

The next parameter we can use is to type a word inside curly brackets. This specifies that any inflection of the word inside curly brackets can be found. In this example, we see that the word 'break' inside curly brackets indicates that this word and all its lemmas, such as 'breaks' or 'breaking', can be found. Let's try this search using the search box.

Notice that all lemmas or inflections of the word were found in the corpus and displayed, again, in descending order of frequency.

Another parameter that can be used is the vertical bar. A vertical bar can be inserted, without spaces, between two or more search words, to specify that any of those words can be found in a search. In the example, we see the word 'do' and then the words 'a', 'an' and 'the', the articles of the English language, separated by vertical bars without spaces. This search will find results for 'do a', 'do an', and 'do the'. Let's try it out.

Again, we see the search results displayed in descending order of frequency. Remember that you can click on any of these results to see concordance lines with examples of it. You can also apply any of the advanced options you learned in the first video, such as filters and the sort function, together with any of these advanced search parameters to further refine your searches.

To illustrate how search parameters can be combined in the same search, let's type the following phrase in the search box: 'take' inside curly brackets, followed by the prepositions 'off', 'from' and 'in', separated by vertical bars, and then an asterisk. This search should find any inflection of the word 'take', followed by either 'off', 'from' or 'in', followed by any word.

The search results are now a combination of all the variable search parameters we specified, displayed in descending order of frequency.

Finally, the last search parameter available in CELV are the part-of-speech tags. Here we see some examples of general tags which can be used in searches to specify grammatical categories, such as [nn*] for any noun, [v*] for any verb, or [i*] for any preposition. All these tags must be typed inside square brackets. As an example, let's see which prepositions are more frequently used with the phrase 'the *internet*'. For that, let's type [i*] inside square brackets in the search box followed by the phrase 'the *internet*'.

The search results show that in the CELV corpus, the collocation 'on the *internet*' is the most frequent with 97 occurrences, followed by 'of the *internet*' with 30 occurrences, and so on. This type of search using part-of-speech tags can be very useful for finding collocation patterns. These tags can also be used in combination with any of the other search parameters and options available in CELV.

To exemplify, let's use CELV to study a common question among English language students. What's the difference between 'do' and 'make'?

There are many ways to approach this question and many different searches we can do in CELV to help answer it. Let's suppose we are interested in collocations between either do or make and any noun. To translate that into search parameters for CELV, we type 'do', followed by a vertical bar, followed by 'make', and then we type the tag for any noun, which is [nn*] inside square brackets.

The results show, for example, that both verbs appear frequently with the word 'things', although 'do things' is almost twice as frequent as 'make things' in this corpus. We can always click on these results to see examples of the collocations.

It's also important to notice that in this type of corpus search, the results are not always clear at first glance. For example, one of the results in this list is 'make vegan', with 20 occurrences. When analyzed by itself, this phrase doesn't seem to make sense, so we must look at some concordance lines to understand it better.

We can now conclude that the word 'vegan' in this collocation is usually part of a noun phrase, such as vegan cornbread, vegan sugar cookies, vegan recipes, and so on.

For even more refined searches, specific word tags can be used. CELV was tagged using the CLAWS part-of-speech tagger, which means that any tag from its tagset can be applied in searches. Not only the general tags shown in this window can be used, but also a wide range of very specific tags, which can be seen by clicking on the link for the C7 tagset.

This will direct you to the CLAWS website where you can see the available tags in this tagset, all of which can be used in CELV if typed inside square brackets in the search box.

This list shows, for example, that the specific tag for comparative adjectives is JJR, and the specific tag for superlative adjectives is JJT. These tags can be used in CELV if inserted inside square brackets. Searching for [jjr] inside square brackets will produce a list of all comparative adjectives found in the corpus, in descending order of frequency. Searching for [jjt] inside square brackets will produce a similar list, this time with superlative adjectives. Again, these specific tags can be combined with other search parameters. And as usual, we can click on any result in the list to see examples. Also, any of the examples can be accessed in video format.

That's all for this video. To learn about other functions and applications of CELV, refer to the other videos in this channel. Thank you for watching.

Apêndice 4: Roteiro do vídeo *Applications of CELV in English learning*

Hello. In this video, you will see some possibilities for using CELV in English language teaching and learning. This video will not explain how to use the basic and advanced search functions of the corpus. If you'd like to know more about how to make searches using CELV, watch the other videos in this channel.

We are going explore four ways in which you can use CELV as a student or teacher.

The first application we are going to explore is pronunciation checking. One way to quickly check the pronunciation of a word in English is to use an online dictionary. Besides showing definitions of words, this type of dictionary also has audio clips with their pronunciation. In this dictionary, for example, you can check the American and the British pronunciation of the word water.

CELV can work in a similar way, with some advantages, because it presents the word in a contextualized way in a real communicative situation. You won't see a definition of the word water, but you will see it inside a sentence such as 'This is where I get water from'. In addition, you can hear the pronunciation of the word and also see the water itself in video format.

Another valuable feature of the naturally occurring language found in CELV is that you can often find peculiar characteristics of spoken language, such as connected speech. For instance, if we search for the phrase 'get out' in an online dictionary, we can hear examples of its isolated pronunciation. However, in natural spoken language, it is common to connect the two words in this expression and pronounce 'gerout'. We can find examples of this type of connected speech in CELV. The ability to check the pronunciation of words in their real context can be very helpful to practice listening and speaking skills.

The third application we will look at in this video is verb patterns. In English, when using two verbs in sequence, we can create verb patterns that work differently depending on the verbs used. For example, some verbs can be followed either by an -ing verb or by an infinitive verb with 'to', such as 'start', which accepts both patterns: 'start doing' and 'start to do'. Other verbs can only be followed by an –ing verb, such as 'keep' as in 'keep doing'. The pattern 'keep to do' is not used. And other verbs can only be followed by an infinitive verb with 'to', such as the phrase 'would like' in 'would like to do'. The pattern 'would like doing' is not used.

This can be observed and studied in CELV using searches with part-of-speech tags. The tagset shows that the tag VVG corresponds to any verb in the –ing participle form, and the tag VVI corresponds to any verb in the infinitive form. Let's use these tags to find examples of verb patterns in CELV.

Searching for 'start' and the tag [vvg] produces a list with examples of this pattern, which shows that it's very common in the English language with many different verbs. Searching for 'start to' and the tag [vvi] also shows many results. This demonstrates that both patterns exist in the language. They can always be further analyzed by clicking in one of the results to see concordance lines.

Now, let's search for 'keep' with the tag [vvg]. The list demonstrates that it is also a common verb pattern in English. But let's see what happens if we search for 'keep to' and then [vvi]. No results were found in the corpus, which is a strong indication that this pattern is not used in the language.

The same process can be done with the phrase 'would like'. Searching for 'would like to' [vvi] will show many different results. However, the results for 'would like' and

the tag [vvg] will probably be different. Again, finding no results is a strong indication that this pattern is not used.

This type of search can be used by both teachers and students as a way of proving that some patterns exist but others don't. Furthermore, examples of each pattern can be analyzed in more detail by looking at concordance lines, to find semantic differences between various patterns.

The fourth and final application of CELV that we will explore in this video is stress derivation, a phonological process of the English language. Some words can be pronounced with different stressed syllables depending on their part-of-speech inside a sentence.

Take, for example, the word 'record'. When this word is a noun, it is pronounced 'REcord', but when it is a verb, it is pronounced 'reCORD'. Let's look at how we can use CELV to find examples of this stress derivation.

Here we see that the word has 217 occurrences in this corpus. Let's increase the number of examples we can see to 100, and then click on the result to look at some concordance lines.

Next, we can double click on some concordance lines to find examples of pronunciation. In this example, the pronunciation was 'reCORD', and the word was a verb. In this example, the pronunciation was 'REcord', and the word was a noun.

We can facilitate the process of finding different stress derivations if we use the sort function, by clicking on 'Show/Hide advanced options', and then specifying one sort filter by the first word to the left of the search word, represented here by L1. If we click again on the word on the list, we generate the concordances again, but this time with the sort filter.

One way to use this filter to find what we are looking for is to use our linguistic knowledge to our advantage. For example, we know that a word that appears after an article, such as 'a', 'an', or 'the', is usually a noun. So let's look for examples of 'the record' in our sorted list and check its pronunciation.

In this example, the pronunciation was 'REcord', because the word was, indeed, a noun, as we correctly predicted from the article before it.

Now, let's use our sorted list again to look for occurrences of our search word preceded by the preposition 'to', because this preposition usually appears before verbs.

In this example, the pronunciation was 'reCORD', because the word was a verb, as we predicted.

This type of search is useful for clarifying variations in pronunciation, and the steps taken for the word 'record' can be repeated for any word that has stress derivation and for other types of phonological phenomena as well.

That's all for this video. Feel free to find more applications of CELV for your classroom or to study at home. If you'd like to learn how to use the search functions of this tool, watch the other videos in this channel. Thank you for watching.

Apêndice 5: Questionário para professores sobre o CELV

1. Há quanto tempo você é professor de inglês?
   ( ) de 0 a 2 anos
   ( ) de 2 a 5 anos
   ( ) de 5 a 10 anos
   ( ) mais de 10 anos

2. Sobre a autenticidade do conteúdo de um ou mais livros didáticos com os quais você já trabalhou: em uma escala de 0 a 10, sendo 0 "totalmente artificial" e 10 "totalmente autêntico", como você o(s) classificaria? Porquê?

3. Os exemplos encontrados no CELV são retirados de vídeos do YouTube. Você considera esse tipo de linguagem autêntico? Como ela se compara com a linguagem usada nos livros didáticos com os quais você trabalha ou trabalhou?

4. Antes de conhecer o CELV, você já havia trabalhado com *corpora* para auxiliar sua prática docente no ensino de língua inglesa? Se sim, de que maneira?

5. Antes desta oficina, você já havia ouvido falar da abordagem da Aprendizagem Direcionada por Dados (*Data-Driven Learning*)? Quais pontos positivos e negativos você observa nessa abordagem?

6. Como você avalia o CELV, em notas de 0 a 10, nos seguintes quesitos?

6.1. Facilidade de uso ____
6.2. Utilidade no ensino de gramática ____
6.3. Utilidade no ensino de vocabulário ____
6.4. Utilidade no ensino de pronúncia ____

7. Você usaria o CELV diretamente com seus alunos, em uma de suas aulas? Por quê? Em que tipo de atividade?

8. Você tem sugestões de recursos que poderiam ser acrescentados para aprimorar o CELV, tornando-o mais útil ou facilitando seu uso? Se sim, quais?

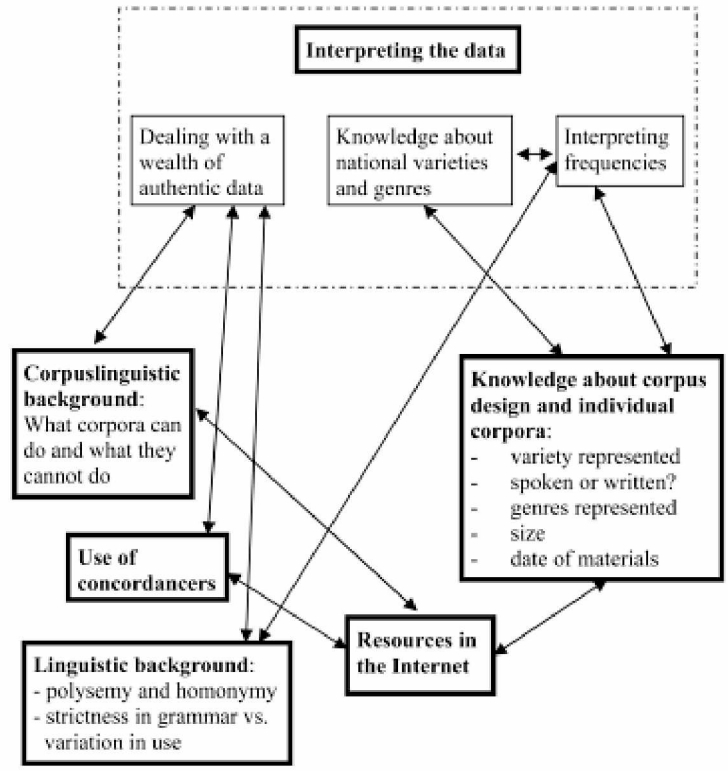
9. Na sua opinião, o CELV seria bem aceito como recurso pedagógico por professores de inglês? Por quê?

10. Na sua opinião, o CELV seria bem aceito como recurso pedagógico por alunos de inglês? Por quê?

ANEXOS

Anexo 1: Aspectos da competência de *corpus*.

(KREYER, 2008, p. 433)



Anexo 2: Tipos de aplicação pedagógica de *corpus*.

(RÖMER, 2008, p. 113)