Procedimentos para construção do *Corpus* da Computação da Língua Inglesa (CoCLI) e cálculo do esforço na construção manual de *corpora*

Procedures for Corpus of Computing in English (CoCLI) construction and effort calculation in manual construction of corpora

Fernando Paulino de Oliveira
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil
fernandooliveira@ufu.br
http://orcid.org/0000-0002-7002-9664

Resumo: O presente trabalho tem como objetivo descrever os procedimentos metodológicos da pesquisa intitulada "*ToGatherUp*: um protótipo de ferramenta para a construção de corpora" que verificou o efeito da incorporação da ferramenta *ToGatherUp* no tempo e no esforço necessários para a construção manual de um *corpus* que elaboramos: o *Corpus* da Computação da Língua Inglesa (CoCLI). Para tanto, discorremos sobre como os autores da pesquisa desenvolveram um conjunto de métricas de medição de esforço – Esforço da Atividade (EA), Esforço Total de Coleta do Texto (ETCT) e Esforço Total do Projeto (ETP) – que serviram de base para a realização de um experimento estatístico comparativo entre os projetos de elaboração manual de duas versões idênticas do CoCLI que se diferenciam por em um deles utilizar o *ToGatherUp* e o outro não. O resultado do experimento demonstrou uma redução média de 7,47% no ETP do projeto em que o *ToGatherUp* foi incorporado em relação ao ETP do projeto em que a ferramenta não foi utilizada, o que corroborou a hipótese de que ela reduz o tempo e o esforço despendidos pelo pesquisador em projetos de elaboração manual de *corpora*.

Palavras-chave: Linguística de *Corpus*; construção manual de *corpus*; métricas de medição de esforço; *ToGatherUp.*

Abstract: The present work aims to describe the methodological procedures of the research entitled "*ToGatherUp*: a prototype of a tool for corpora construction" that verified the effect of incorporating *ToGatherUp* in necessary time and effort invested

in manual construction of *Corpus* of Computing in English (CoCLI). To this end, we discuss how the research authors developed a set of metrics for measuring effort – Activity Effort (EA), Total Effort for Text Collection (ETCT) and Total Project Effort (ETP) – which served as the basis for conducting a comparative statistical experiment between the manual elaboration of two identical versions of the CoCLI: which differ from each other by one of them using the *ToGatherUp* and the other one not using it. The experiment shows an average reduction of 7.47% in the ETP when using *ToGatherUp* compared to the ETP when not using the tool. This result corroborates the hypothesis that the tool reduces the time and effort spent by the researcher on manual elaboration projects of *corpora*.

Keywords: *Corpus* Linguistics; manual construction of *corpus*; effort measurement metrics; *ToGatherUp*.

Submetido em 25 de agosto de 2020

Aceito em 09 de novembro de 2020

1 Introdução

O desenvolvimento de pesquisas com base na observação empírica de dados da língua favoreceu o surgimento e o crescimento da Linguística de *Corpus*, doravante LC, que é "uma nova metodologia (que utiliza textos naturais e ferramentas informáticas para descrever a língua) e uma nova disciplina (no sentido de uma nova abordagem à descrição linguística)" (FRANKENBERG-GARCIA, 2012, p. 12). Conforme esclarece Berber Sardinha (2004), para que seja possível o uso prático da LC, o interessado precisa de "um ingrediente essencial: o *corpus*" (BERBER SARDINHA, 2004, p. 45).

A construção de *corpora* de pequenas extensões[1] pode não representar um desafio complicado, mas a de *corpora* compostos por grande volume de dados tem sido reportada como uma das partes mais difíceis do desenvolvimento de uma pesquisa (cf. KÜBLER; ASTON, 2010; ATKINS; CLEAR; OSTLER, 1992; BAKER, 2010; BIANCHI, 2012; EDWARD, 2015; MACMULLEN, 2003; MCENERY; HARDIE, 2011; MCENERY; XIAO; TONO, 2006; MINSHALL, 2013;

[1] A extensão ou o tamanho de um *corpus* representa o volume de dados linguísticos disponíveis para análise. Na seção Fundamentação teórica, discutimos sobre a extensão de *corpora*.

RENOUF, 2007; SEMINO; SHORT, 2004; VOORMANN; GUT, 2008; ZANETTIN, 2014). A principal reclamação dos linguistas refere-se à quantidade enorme de tempo e esforço necessária para a realização das atividades relativas à construção de um *corpus*. Além do tempo e esforço, Edward (2015) e Garretson (2008) afirmam que, ao começar a construção de um *corpus*, uma das primeiras barreiras enfrentadas pelos pesquisadores é encontrar ferramentas computacionais capazes de dar suporte[2] especializado às atividades do projeto.

Diante desses desafios, a proposta deste trabalho é contribuir com a prática de linguistas e pesquisadores de áreas afins por meio da apresentação dos procedimentos metodológicos adotados na pesquisa intitulada "*ToGatherUp*: um protótipo de ferramenta para a construção de *corpora*"[3] (OLIVEIRA, 2019). O objetivo dessa pesquisa - determinar os efeitos da incorporação do *ToGatherUp* no esforço necessário para a construção manual de *corpora* – levou seus autores a percorrerem um interessante e produtivo caminho que gerou uma proposta de sistematização do trabalho de construção manual de *corpora*, a criação de métricas de aferição de esforço e culminou na realização de um experimento que revelou a eficácia do *ToGatherUp* na redução do tempo e esforço investido na criação de *corpora*. Para alcançarmos nosso objetivo, nas próximas seções deste artigo, apresentaremos a fundamentação teórica, a metodologia, os resultados alcançados na pesquisa e nossas considerações finais sobre ela.

2 Fundamentação teórica

A Linguística é a área em que se desenvolve o estudo científico da linguagem humana com base em fatos linguísticos (MARTINET, 1978). De acordo com Widdowson (1996), de modo geral, os fatos linguísticos podem ser inferidos por meio da introspecção, da elicitação e da observação de dados provenientes do uso real da língua pelos seus usuários. Widdowson (1996) esclarece que os fatos linguísticos apreendidos por meio da introspecção e da elicitação não revelam o uso

[2] Do ponto de vista dos autores da pesquisa retratada por nós, as ferramentas que oferecem suporte à construção manual de *corpora* são aquelas que oferecem recursos que facilitam as atividades e o gerenciamento do projeto de construção manual de *corpora*.
[3] Disponível em: www.ileel.ufu.br/togatherup. Acesso em: 1 mar. 2019.

efetivo da língua, pois partem das intuições que os seus usuários têm sobre ela. Já a observação de dados linguísticos decorrentes do uso real da língua e que refletem o comportamento linguístico de seus usuários constitui-se como uma forma mais segura para a realização de inferências sobre a língua. Nesse sentido, as análises linguísticas com base na LC podem ser consideradas altamente confiáveis, uma vez que partem da observação de *corpora* compostos por dados linguísticos reais.

Sinclair (2005) afirma que a construção de um *corpus* deve ser realizada de acordo com critérios bem definidos e eficientes o bastante para que o seu delineamento final possa garantir que o conjunto de textos seja representativo. O conceito de representatividade na LC está associado à capacidade que um *corpus* tem de representar uma língua ou uma variedade dela e ao modo como foi construído. Podemos dizer que um *corpus* é representativo quando, a partir da análise do conjunto de textos provenientes das várias situações comunicativas reais de uma comunidade linguística, é possível obter conclusões, a respeito de suas propriedades, que permitam generalizações sobre a língua ou sobre a variedade de língua em estudo.

A fase de construção de um *corpus* em que são definidos os seus critérios tem sido referenciada pelos autores da LC como o "desenho do *corpus*".[4] Firmar o desenho de um *corpus* não é uma tarefa simples, pois, conforme Berber Sardinha (2004), não existem critérios objetivos para isso. Segundo Blecha (2012), a delimitação do desenho de um *corpus* deve ser orientada em consonância com os objetivos da pesquisa. Tagnin (2010) coaduna com Blecha (2012) e afirma que cabe ao criador do *corpus* a responsabilidade de definir os critérios que possam garantir sua representatividade. Dentre os critérios para a construção de *corpora*, na pesquisa aqui relatada, os fundamentos e implicações referentes à extensão do *corpus* ganham importância.

A extensão do *corpus* representa o volume de dados linguísticos que ele dispõe para análise. Na literatura da LC, não encontramos a definição exata do tamanho necessário para que um *corpus* seja representativo. No entanto, para estudos que consideram a chavicidade[5]

[4] Na literatura da LC, em língua inglesa, encontramos o termo *corpus design*.

[5] De acordo com Fromm (2007), a chavicidade (*keyness*) informa o quanto uma palavra se destaca na relação entre a sua frequência no *corpus* de estudo e no *corpus* de referência.

de palavras, encontramos recomendações e estimativas, como a de Berber Sardinha (2004), que afirma que a relação de tamanho entre os *corpora* de estudo e os *corpora* de referência influencia a quantidade de palavras-chave obtidas. O autor recomenda que um *corpus* deve ser o mais extenso possível.

Mesmo com a recomendação de Berber Sardinha (2004), cabe pontuar que a extensão de um *corpus* está sujeita à disponibilidade de dados que atendam às especificidades do desenho dele. A obtenção de dados suficientes para cada campo semântico de uma Árvore de Domínio, no caso das pesquisas terminológicas, ou para cada gênero textual que compõe um *corpus* de estudo do léxico, de modo que seja garantido o balanceamento[6] do *corpus*, é um exemplo dessa situação. Ademais, Fromm (2003) chama a atenção para o fato de que o desenvolvimento de um *corpus* extenso requer a participação de vários pesquisadores e auxiliares; caso contrário, a construção dele pode demorar anos para ser concluída. Nessa situação, há a questão do tempo que o pesquisador (ou a equipe de pesquisadores) tem para dedicar à obtenção de dados.

Berber Sardinha (2005) observa que, na prática, "o pesquisador coleta uma certa quantidade de dados de acordo com suas possibilidades, efetua a análise, mas não sabe se sua coleta foi além ou aquém do que seria teoricamente mais adequado" (BERBER SARDINHA, 2005, p. 188). Por essa razão, Nelson (2010) afirma que a criação de um *corpus* é "uma aceitação entre o que é o esperado e o que é possível" (NELSON, 2010, p. 30)[7] e Meyer (2004) explica que as mudanças no desenho inicial do *corpus* são naturais e inevitáveis (desde que não comprometam a integridade do *corpus*) diante dos obstáculos e complicações que podem surgir durante a sua compilação.

2.1 A organização do trabalho em projetos de construção manual de um *corpus*

A LC não oferece padrões pré-estabelecidos ou modelos sistematizados para a construção manual de um *corpus*. O que encontramos na sua literatura são abordagens que, embora obedeçam às

[6] Aluísio e Almeida (2006) definem o balanceamento como o equilíbrio entre as categorias atribuídas aos textos que compõem um *corpus*.

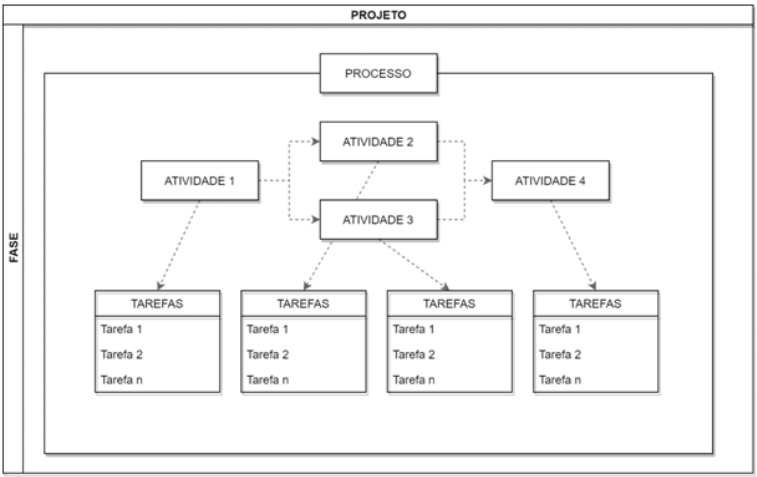
[7] Original: "any attempt at corpus creation is therefore a compromise between the hoped for and the achievable".

diretrizes criadas por Sinclair (2005), diferenciam-se entre si em aspectos de organização, uso de ferramentas e técnicas. Cabe destacar que, ao mesmo tempo em que a inexistência de um padrão promove a flexibilidade das práticas de elaboração de *corpora*, ela também gera problemas como a variação significativa dos nomes que são atribuídos às ações que envolvem a construção de um *corpus*. Nesse sentido, há autores que se referem ao trabalho de criação de *corpus* como um processo dividido em estágios (cf. ATKINS; CLEAR; OSTLER, 1992; ESCARTÍN, 2012; KENNEDY, 1998), em ciclos (cf. BIBER, 1993) ou em passos (cf. SANTOS, 2011). Além dessas denominações, é comum encontrarmos palavras, tais como: "tarefas", "atividades" e "procedimentos", sendo utilizadas com o mesmo sentido, isto é, remetendo-se às mesmas ações.

Por essa razão, os autores da pesquisa aqui retratada decidiram adotar os termos "fases", "processos", "atividades" e "tarefas" utilizados na área de Gerenciamento de Projetos para designar as partes do "ciclo de vida" de um projeto. A adoção dessa nomenclatura pelos autores foi feita por considerarem que a construção manual de um *corpus* equivale à realização de um projeto em conformidade com o conceito de projeto e, em partes, nos princípios da área de Gerenciamento de Projetos, expostos no guia *Project Management Body of Knowledge* (PMBOK), publicado em 2013 e considerado como a principal referência da área de Gerenciamento de Projetos. De acordo com o PMBOK, um projeto é "um esforço temporário empreendido para criar um produto, serviço ou resultado exclusivo" dentro de um "ciclo de vida" (PMBOK, 2013, p. 3). O ciclo de vida de um projeto corresponde à sequência de fases pelas quais ele passa ao longo do seu desenvolvimento.

Durante o ciclo de vida do projeto, cada fase pode comportar um ou mais processos. Estes, por sua vez, podem admitir uma ou mais atividades. Uma atividade pode relacionar-se com outra(s), de maneira lógica, de modo que seu início ou sua continuidade somente seja possível após a geração de um ou mais resultados (entregas) de outra(s) atividade(s). No nível mais baixo do ciclo de vida de um projeto, encontram-se as tarefas, que são as menores unidades de trabalho possíveis pertencentes ao escopo de uma atividade. A Figura 1 ilustra as relações dos componentes do trabalho de um projeto apresentadas neste parágrafo.

FIGURA 1 – Organização do trabalho de um projeto



Fonte: Oliveira (2019, p. 42.)

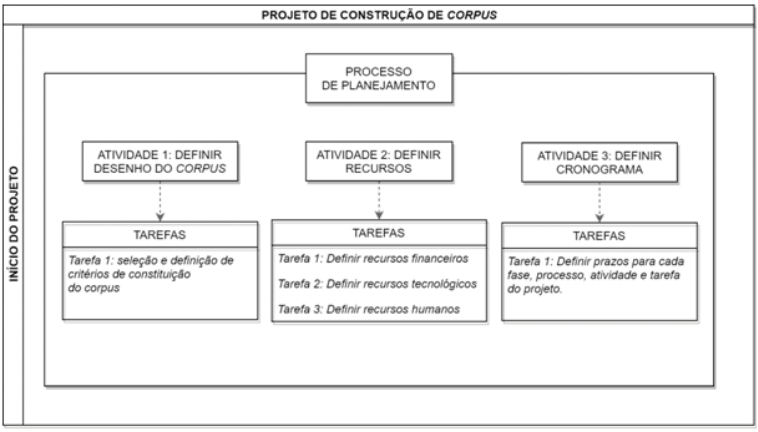
Além de resolver o problema da nomenclatura utilizada na construção de *corpora*, os autores da pesquisa perceberam que era possível transpor o modelo de organização do trabalho dos projetos da área de Gerenciamento de Projetos para os projetos de construção manual de *corpora*. A partir dessa percepção, eles procuraram sistematizar o trabalho relativo à construção manual de *corpora* de com base nesse modelo. Deste modo, propuseram que o projeto de construção manual de um *corpus*, de modo geral, pode ser dividido em três fases distintas: a) a inicial, em que há o planejamento do *corpus*; b) a intermediária, caracterizada pela obtenção, preparação e armazenamento dos dados do *corpus* e c) a de encerramento, na qual ocorre a distribuição dos dados do *corpus*. Nos próximos parágrafos, explicitamos a sistematização concebida pelos autores da pesquisa, situando-a com contribuições teóricas dos autores da LC.

A fase inicial do projeto de construção manual de um *corpus* é caracterizada pela execução das atividades de planejamento do *corpus*, de definição dos recursos necessários para elaborá-lo e de esquematização do cronograma de execução do projeto. De acordo com Nelson (2010, p. 53), há uma série de variáveis que precisam ser consideradas antes do início da compilação dos dados, a saber: o tamanho do *corpus*, o balanceamento dele, a estrutura conceitual em que os textos serão

organizados, o formato de armazenamento dos textos, a maneira como será feita a coleta dos textos, o padrão que será usado para a nomeação dos arquivos e o controle em relação à coleta e ao gerenciamento dos textos. Algumas dessas questões são analisadas durante o desenho do *corpus*, que é a primeira atividade do processo de planejamento do *corpus*. Para o estabelecimento do desenho do *corpus*, o seu criador precisa executar as tarefas de seleção e definição dos critérios que nortearão a constituição do *corpus*.

Ademais, Atkins, Clear e Ostler (1992, p. 3) mencionam o fato de que o planejamento do *corpus* deve prever o uso de recursos financeiros, tecnológicos e humanos necessários para garantir a conclusão do projeto. Santos (2011) complementa as exigências do planejamento do *corpus* ao afirmar que é necessário estabelecer o cronograma para a execução do projeto, pois várias decisões que precisam ser tomadas durante a elaboração do *corpus* estão vinculadas às restrições de tempo para a sua realização. A Figura 2 ilustra o processo de planejamento e as atividades da fase inicial do projeto de construção de um *corpus*.

FIGURA 2 – Processo de planejamento da construção de um *corpus*

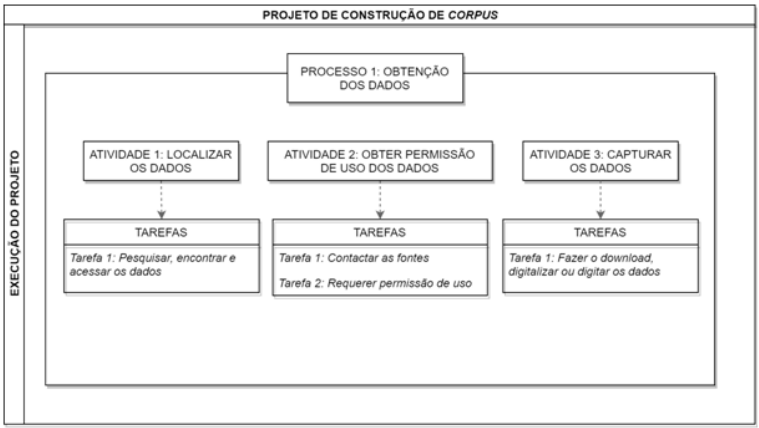


Fonte: Oliveira (2019, p. 44.)

Após planejar a construção do *corpus*, o pesquisador tem em mãos os parâmetros que guiarão a obtenção de dados linguísticos que irão compor o *corpus* e pode iniciar a fase de execução do projeto que consiste na realização das atividades relativas aos processos de obtenção,

preparação e armazenamento de dados do *corpus*. A Figura 3 exibe o processo de obtenção de dados para composição de um *corpus*.

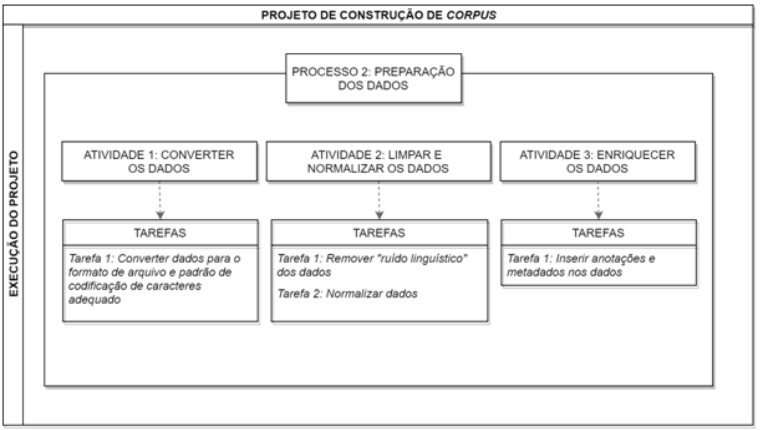
FIGURA 3 – O processo de obtenção de dados para composição de um *corpus*



Fonte: Oliveira (2019, p. 45.)

A primeira atividade desse processo consiste na pesquisa, localização e acesso aos materiais que comportam os dados desejados. Para Sinclair (1991), os dados podem ser encontrados, em suas versões originais, na forma eletrônica, impressa ou escrita à mão. Esse autor salienta que a obtenção de textos no formato eletrônico é a mais fácil e desejável comparada às demais, visto que exige um menor esforço do pesquisador no momento de adaptá-los para posterior processamento feito pelas ferramentas computacionais. Após a obtenção dos textos, o pesquisador precisa certificar-se de que eles são "úteis" para a inclusão em um *corpus*. A utilidade de um texto, para as pesquisas da LC, está associada, obrigatoriamente, à condição favorável dele para o processamento através de ferramentas computacionais e, opcionalmente, à integridade e ao enriquecimento dele. Esses aspectos configuram o processo de preparação dos dados do *corpus* e estão contemplados na Figura 4.

FIGURA 4 – O processo de preparação dos dados do *corpus*



Fonte: Oliveira (2019, p. 48.)

O processamento de um texto por meio de uma ferramenta computacional exige que ele esteja em um formato "compreensível pelos computadores" (*machine-readable*). Porém, os recursos que tornam um texto propício à interpretação humana podem ser prejudiciais ao processamento feito pelos computadores, já que estes ainda não possuem as mesmas capacidades de decodificação que os homens. Em virtude disso, na metodologia da LC, é indispensável a conversão das versões originais de textos para versões apropriadas ao trabalho que a máquina executa.

Após a realização da conversão, segundo Santos (2011), os arquivos podem apresentar resíduos como, por exemplo, números de páginas, informações de cabeçalhos e rodapés de páginas, anotações sobre a divisão das seções do texto, conteúdos de tabelas (que perdem o sentido ao serem desprovidos da estrutura da tabela) e erros de codificação de caracteres (resultantes da transposição de um padrão de codificação para outro). No contexto dos dados linguísticos, os resíduos presentes nos textos podem ser considerados como ruído linguístico e podem gerar problemas no que diz respeito à análise da frequência dos elementos linguísticos um *corpus*.

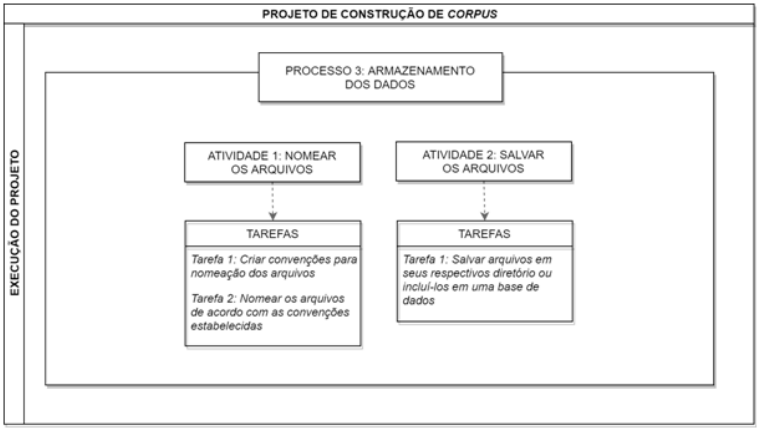
Conforme Gries (2009), a frequência de um elemento linguístico é base para a formação de listas de palavras que são utilizadas para a construção de listas de palavras-chave. As listas de palavras-chave, de

acordo com Tagnin (2015) e Edward (2015), apresentam os elementos linguísticos cujas frequências são estatisticamente relevantes a partir do resultado da comparação entre listas de palavras de um *corpus* de estudo e um *corpus* de referência. Pensando nessas relações, para que uma ferramenta computacional possa gerar listas com a frequência das palavras e, a partir disso, possa executar os cálculos que determinam a relevância dos elementos linguísticos, é necessário, em um primeiro momento, que ela identifique cada um dos elementos linguísticos presentes num texto. Essa identificação é realizada através do processo computacional conhecido na área de Processamento de Linguagem Natural como tokenização que consiste na segmentação das sentenças de um texto em elementos significativos, chamados *tokens*.

A tokenização de um texto que apresenta ruído linguístico pode gerar *tokens* sem nenhuma relação com qualquer elemento significativo da língua (por exemplo: *tokens* formados por partes de palavras que foram separadas incorretamente) e, consequentemente, gerar cálculos imprecisos sobre a frequência de um elemento, comprometendo a qualidade das listas provenientes da análise realizada por ferramentas computacionais. Por isso, uma das formas de reduzir ou de eliminar erros de tokenização, é a realização da limpeza e da normalização dos textos de um *corpus*. A limpeza de um texto, de acordo com Aluísio e Almeida (2006), consiste na remoção dos ruídos linguísticos. Já a normalização consiste na uniformização de palavras (em termos ortográficos), de siglas e de abreviaturas que possuem variações de escrita, a remoção de espaçamentos e de quebras de linhas desnecessários e a homogeneização de caracteres de pontuação do texto, como hifens, traços, aspas e apóstrofos.

Após a obtenção e preparação do *corpus*, os arquivos de texto que o constituem precisam ser nomeados e armazenados de modo que possam ser recuperados facilmente (NELSON, 2010). As atividades relacionadas à nomeação, ao armazenamento e à disponibilização dos dados de um *corpus* são realizadas no processo de armazenamento de dados, ilustrado na Figura 5.

FIGURA 5 – O processo de armazenamento dos dados do *corpus*

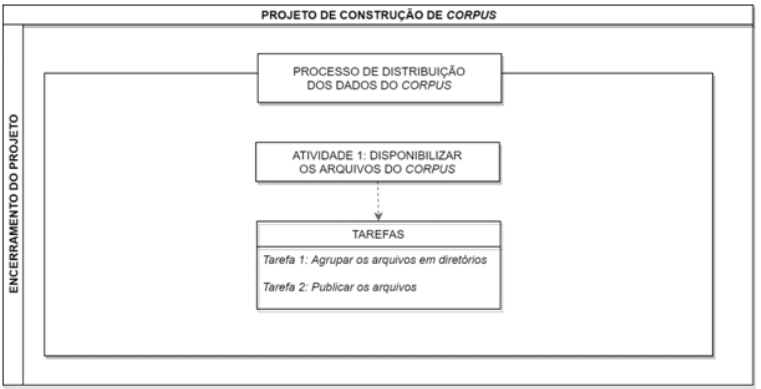


Fonte: Oliveira (2019, p. 59.)

A nomeação dos arquivos de texto para o armazenamento consiste na atribuição de um nome ao arquivo (*filename*) a fim de que ele possa ser identificado no sistema de arquivos do sistema operacional. Para facilitar a associação do nome de um arquivo ao seu conteúdo o pesquisador pode fazer uso de uma convenção de nomeação de arquivos, no inglês, *File Naming Conventions* (FNC). A convenção de nomeação de arquivos pode ser definida como um conjunto de regras que determina a estrutura da nomeação – constituída por diferentes segmentos que abrigam elementos informativos, ou seja, aqueles que fazem referência ao conteúdo, à descrição, ao contexto ou ao propósito dos arquivos. A ideia central de convencionar a nomeação é combinar informações suficientes para que a identificação do conteúdo do arquivo seja feita a partir de seu nome. Uma vez nomeado o arquivo de texto, comumente, é armazenado em um diretório organizado de acordo com a estrutura hierárquica adotada no projeto de construção do *corpus*, de modo que seja possível realizar uma associação significativa entre seu conteúdo e a categoria do diretório.

Depois das fases de planejamento e execução, a construção de um *corpus* segue para a fase de encerramento, que compreende o processo de distribuição dos dados do *corpus*, ilustrado na Figura 6, que trata da disponibilização dos dados do *corpus* com o propósito de serem processados pelas ferramentas computacionais, nas quais os métodos de análises e recuperação de informações são aplicados.

FIGURA 6 – O processo de distribuição dos dados do *corpus*



Fonte: Oliveira (2019, p. 64.)

A partir da sistematização apresentada, os autores da pesquisa debruçaram-se sobre a questão do tempo e do esforço na construção de *corpora*, procurando descortinar as principais dificuldades percebidas pela comunidade linguística quanto à execução das atividades de construção de *corpora*. No próximo tópico, encontramos as reflexões teóricas feita por eles sobre essas questões.

2.2 O tempo e o esforço na construção de um *corpus*

A introdução dos computadores no fazer linguístico e a Internet facilitaram a obtenção de dados linguísticos e o trabalho de construção de *corpora*. Porém, conforme mencionamos na introdução deste artigo, não é raro encontrarmos reclamações relacionadas à quantidade de tempo e esforço para se construir um *corpus*. De fato, a construção manual de *corpora* pode exigir bastante esforço e tempo do pesquisador, uma vez a sua intervenção, praticamente, é necessária em todas as fases do projeto (BAKER, 2010, p. 109), como podemos observar na descrição das atividades enumeradas a seguir:

1) **Definir o desenho do *corpus*, os recursos utilizados para lidar com ele e o cronograma referente ao projeto de construção do *corpus*:** exige que o pesquisador, basicamente, tome decisões teóricas e gerenciais;

2) **Localizar dados linguísticos:** a coleta manual de textos, em oposição à automática exige que o pesquisador busque fontes e verifique o acesso ao material (eletrônico, impresso ou em outro estado) que será utilizado em sua seleção. Os dados em formato eletrônico, por exemplo, são frequentemente localizados por meio de pesquisas feitas em sistemas de busca como o *Google*. Mesmo com a facilidade oferecida por esse tipo de sistema, a filtragem dos materiais encontrados (feita com base nos critérios do desenho do *corpus*) pode ser árdua devido ao grande volume e à qualidade das informações retornadas por *sites* como *Google*;
3) **Obter permissão de uso dos dados linguísticos:** para Santos (2011), essa atividade implica que o pesquisador precisará identificar a pessoa ou a entidade detentora dos direitos autorais de um texto, solicitar-lhe consentimento para usá-lo e, em seguida, aguardar retorno. O autor alerta que o consumo de tempo é ampliado nos casos em que o pesquisador precisa realizar várias tentativas de contato com o detentor para ter sucesso ou nas situações em que a solicitação de autorização tenha de partir de níveis hierárquicos superiores de uma instituição para que se obtenha uma resposta;
4) **Capturar os dados:** demanda a intervenção humana em uma escala que varia de acordo com o formato em que os dados estão quando são encontrados. Se estiverem no eletrônico, o pesquisador necessitará de intervir menos. Se os dados estiverem em materiais impressos ou escritos à mão, o pesquisador terá de convertê-los para o formato eletrônico, de preferência, por meio da digitalização com o auxílio de *scanners* e *softwares* de OCR[8] – uma das atividades que mais demandam esforço e tempo. Para Simske (2006), a precisão oferecida atualmente pelos OCRs na conversão de textos ainda é limitada e pode gerar erros referentes à troca, à inserção e à exclusão de caracteres, principalmente, quando a qualidade dos documentos originais é ruim. Isso faz com que autores como Nelson (2010), Kübler e Aston (2010), Santos (2011) e Bianchi (2012) preconizem a revisão manual cuidadosa dos dados resultantes da digitalização de textos por intermédio de *scanners* e OCRs para que se tenha certeza de que eles correspondem às suas versões originais;

[8] OCR é um *software* de reconhecimento ótico de caracteres. A sigla OCR vem do inglês *Optical Character Recognition*.

5) **Converter os dados:** exige pouco do pesquisador, apenas que ele manipule ferramentas computacionais que convertam os arquivos para o formato TXT e para o padrão *Unicode UFT-8*. Transformar um texto escrito em PDF para TXT, por exemplo, pode ser feito de forma gratuita por meio de serviços *on-line* de conversão, como o *Lightpdf* (lightpdf.com), entre outros. E a mudança para o padrão *Unicode UTF-8* pode ser efetuada por ferramentas como o *EncodeAnt* (ANTHONY, 2016);
6) **Limpar e normalizar os dados:** requer que o pesquisador proceda como auditor no que diz respeito aos dados do *corpus* para identificação e posterior eliminação ou correção das anomalias (ruído linguístico). A limpeza e a normalização estão diretamente relacionadas a algumas variáveis: volume e qualidade dos dados (resultante dos métodos de captura, conversão e codificação dos textos), finalidade (necessidades) da pesquisa e, por fim, métodos escolhidos para a execução da limpeza e da normalização. Vale lembrar que as tarefas em questão podem ganhar proporções gigantescas e, portanto, serem difíceis no caso de *corpora* compostos por grandes volumes de informação;
7) **Enriquecer os dados:** pressupõe que o pesquisador realize uma conferência no que alude à etiquetagem automática de *corpora*. Conforme Neumann e Hansen-Schirra (2012), o enriquecimento de *corpora* grandes depende da etiquetagem automática, pois o processamento manual de grandes volumes de dados é praticamente inviável. Semino e Short (2004) reforçam essa ideia ao afirmarem que até mesmo a etiquetagem manual de *corpora* pequenos é extremamente demorada. Contudo, a utilização de ferramentas computacionais para a etiquetagem não dispensa a intervenção manual do pesquisador (MEYER, 2004), uma vez que *taggers* e *parsers* não conseguem alcançar uma precisão total no processamento dos dados. Para Meyer (2004), a precisão dos etiquetadores, geralmente, é comprometida pela inconsistência dos dados (dados não limpos ou normalizados) e pela dificuldade que apresentam para lidar com as características idiossincráticas (GARSIDE; SMITH, 1997 *apud* MEYER, 2004) da linguagem humana. Em decorrência dos possíveis erros, o resultado da etiquetagem automática precisa ser conferido pelo pesquisador (*post-editing*) com o objetivo de corrigir e resolver possíveis

ambiguidades nas etiquetas (LEECH, 2005). Segundo Bianchi (2012), essa atividade é desenvolvida manualmente e exige muito tempo e esforço.

8) **Nomear arquivos:** prevê que o pesquisador atribua nomes aos arquivos do *corpus*, de preferência, após estabelecer uma convenção. Nessa atividade, o pesquisador poderá ter de checar como foi definida a estrutura da convenção quando for nomear cada arquivo do *corpus* caso não consiga memorizá-la. Ademais, ele precisará selecionar a informação mais adequada para compor cada segmento da estrutura do nome do arquivo;

9) **Salvar arquivos:** requer pouco do pesquisador quando é feito por meio da alocação dos arquivos em diretórios de um sistema de arquivos, de acordo com a hierarquia estabelecida em um projeto. Entretanto, segundo Sedlar (2005), em situações em que o pesquisador decida salvar os arquivos em uma base de dados, a execução da tarefa dependerá do uso de uma ferramenta computacional que ofereça a interface necessária para a inclusão dos arquivos no banco de dados. O pesquisador poderá optar pelo uso de uma ferramenta já existente ou pela criação de uma ferramenta customizada para o seu projeto. No primeiro caso, ele precisará de um esforço adicional para a escolha de uma ferramenta e para a aprendizagem do seu uso. No segundo, além do esforço para aprender a usar a ferramenta, ele investirá recursos financeiros, tempo e esforço por ter de contratar um profissional para desenvolver a aplicação ou por desenvolvê-la por conta própria;

10) **Disponibilizar arquivos:** demanda pouco esforço e tempo do pesquisador quando ele opta por apenas copiar os arquivos em dispositivos de armazenamento de dados. Já a publicação *on-line* do *corpus* pode requerer recursos financeiros (por exemplo, para a contratação de serviços de hospedagem ou de armazenamento de dados na nuvem) e, ainda, mais esforço e tempo do pesquisador, pois ele deverá se preocupar com questões: como a escolha de um local para a publicação, a compactação dos arquivos, a disponibilização de documentação sobre o *corpus* com informações suficientes para que a sua utilização seja feita por outros pesquisadores e a explicitação de uma licença de uso dos dados do *corpus*.

A fim de contornar as dificuldades relacionadas à coleta de dados, os pesquisadores podem adotar os *web corpora* ou *corpora ad-hoc*, que são *corpora* compostos por dados coletados da Internet de forma automática. Nesse caso, eles precisam lançar mão de ferramentas computacionais, como o *WebBootCat* (BARONI *et al*., 2006), o *WebCorp Linguist's Search Engine* (KEHOE; GEE, 2007) e o *Bootcat* (BARONI; BERNARDINI, 2004), que, segundo Aluísio e Almeida (2006, p. 168), utilizam motores de busca (*Google*, por exemplo) e um "pequeno conjunto de itens léxicos, denominados sementes (*seeds*)" para efetuarem a compilação.

Schäfer e Bildhauer (2013) consideram que a realização de inferências estatísticas a partir de *corpora* construídos com base em resultados de pesquisas de motores de busca não é uma boa prática de pesquisa, pois os buscadores privilegiam a precisão (*precision*) em detrimento da revocação (*recall*),[9] podem ser influenciados por fatores econômicos,[10] usam variáveis como a língua e a localização de quem fez a pesquisa e realizam alterações automáticas nas expressões fornecidas para a pesquisa (otimizam as expressões por meio de reduções ou expansões). Além disso, as buscas não podem ser reproduzidas devido à constante entrada e saída de conteúdos (indexação) na Internet.

Mais do que as questões relacionadas aos critérios de recuperação de informações dos motores de busca, Schäfer e Bildhauer (2013) acreditam que a opção pelo uso de *corpora* provenientes de métodos automáticos de coleta requer precaução extra do pesquisador no que diz respeito a aspectos, tais como: remoção do *boilerplate* (quais partes do documento foram removidas) e do ruído linguístico dos documentos (quais os tipos de ruídos existentes e qual a precisão da remoção deles); introdução de ruído linguístico (quais ruídos foram introduzidos após o processamento dos documentos); remoção de arquivos duplicados (*deduplication*) (quais documentos foram removidos e quais foram os

[9] Consoante Rubi (2009), a revocação "pode ser mensurada por meio da relação entre o número de documentos relevantes sobre determinado tema, recuperados pelo sistema de busca, e o número total de documentos sobre o tema, existentes nos registros do mesmo sistema" (RUBI, 2009, p. 85). A precisão "pode ser mensurada por meio da relação entre os documentos relevantes recuperados e número total de documentos recuperados" (RUBI, 2009, p. 85-86).

[10] Por exemplo, os conteúdos patrocinados.

critérios de remoção) e a forma pela qual a amostragem dos dados foi criada. Além das precauções de Schäfer e Bildhauer (2013), na literatura da LC, identificamos outros problemas que podem surgir numa situação em que se opta por automatizar a coleta de um *corpus*:

1) **Problema da replicabilidade:** está relacionado à mutabilidade dos dados na Internet. Para Mcenery e Hardie (2011), os estudos com *corpora* coletados de forma automática na Internet são difíceis de serem replicados com o passar do tempo em virtude de haver constante mudança de dados na rede;
2) **Problema dos falso-positivos:** os falso-positivos são *tokens* e *types* que não possuem relação com qualquer elemento significativo de uma língua alvo de pesquisa, provenientes de erros de tokenização provocados pelo ruído linguístico de um *corpus*. Conforme Schäfer e Bildhauer (2013), os *web corpora* tendem a conter um alto nível de ruído linguístico;
3) **Problema da amostragem:** refere-se à incontrolabilidade e arbitrariedade da escolha dos dados dos *web corpora* no universo heterogêneo (RENOUF, 2007) dos dados disponíveis na Internet. De acordo com Schäfer e Bildhauer (2013), a coleta automática de documentos, geralmente, não segue um esquema amostral preestabelecido. Para esses autores, na melhor das hipóteses, *web corpora* são uma amostra randômica de dados da Internet, cuja composição exata é desconhecida e precisará ser estabelecida após a sua compilação. Para Mcenery e Hardie (2011), um dos aspectos do desconhecimento do conteúdo de *web corpora* é a dificuldade em determinar o gênero textual dos documentos coletados sem tê-los lido;
4) **Problema legal:** consiste no *download* e uso de textos de *sites* da Internet e na sua distribuição como parte de um *corpus* sem o consentimento dos autores (MCENERY; HARDIE, 2011, p. 58). Segundo Mcenery e Hardie (2011), as leis de direito autoral aplicam-se aos textos coletados automaticamente da Internet do mesmo modo que se aplicam aos materiais impressos e, por isso, podem gerar as mesmas implicações legais que outras formas de construção de *corpora*;

5) **Problema da violação da integridade dos textos:** Schäfer e Bildhauer (2013) argumentam que o processamento automático de coleta dos textos introduz erros nos dados originais deles que podem reduzir a qualidade dos dados. Para exemplificar, os autores mencionam a remoção automática de dados duplicados que, se for feita no interior dos textos, no nível dos parágrafos, pode implicar a inclusão de materiais incompletos em um *corpus*;
6) **Problema das consultas em massa *(batch or bulk requests*):** a obtenção dos dados dos *web corpora* depende do envio das sementes (ALUÍSIO; ALMEIDA, 2006, p. 168) que serão utilizadas como parâmetro de consulta pelos motores de busca. Schäfer e Bildhauer (2013) explicam que, no intuito de evitar abusos, os motores de busca apresentam restrições para o processamento gratuito de grandes volumes de consultas automáticas. Portanto, a construção de *web corpora* de grandes proporções por meio da coleta automática de dados demandará do pesquisador o pagamento pelo serviço de busca que ultrapassa os limites de consultas dos motores até que o volume de dados necessário para os *corpora* seja atingido.

A compilação automática de *corpora* pode ser "adequada para uma grande variedade de propósitos" (MCENERY; HARDIE, 2011, p. 8)[11] e os *web corpora* são extremamente valiosos para as pesquisas que demandam a análise de grandes volumes de informação (BERGH; ZANCHETTA, 2008, p. 320) e em que "o valor do volume dos dados se sobrepõe à qualidade proporcionada pela sua limpeza" (SCHÄFER; BILDHAUER, 2013, p. 126),[12] ainda que apresentem problemas e possam ter sua utilidade considerada limitada por grupos de linguistas (RUNDELL; KILGARRIFF, 2011, p. 262). Para Rundell e Kilgarriff (2011), os *web corpora* adequam-se, por exemplo, às pesquisas lexicográficas para a criação de dicionários gerais em que "os benefícios da abundância de dados superam os problemas dos *web corpora*" (RUNDELL; KILGARRIFF, 2011, p. 262).[13]

---

[11] Original: "suitable for a wide variety of purposes".
[12] Original: "values the amount of available data more highly than the cleanliness of a corpus".
[13] Original: "the benefits of abundant data outweigh most of the perceived disadvantages of web corpora".

3 Metodologia

Como explicado na introdução deste artigo, o objetivo principal da pesquisa nele descrita é determinar os efeitos da incorporação do *ToGatherUp* no esforço necessário para a construção manual de *corpora*. A forma encontrada pelos autores para atingirem esse propósito foi a realização de um experimento de comparação entre os esforços necessários para a construção de duas versões idênticas do CoCLI, sendo que o projeto de elaboração de uma delas contou com a incorporação do *ToGatherUp* e o outro não. Para que a confrontação fosse possível, em um primeiro momento, eles estabeleceram um critério objetivo e um método para a medição do esforço das atividades de cada um dos projetos de construção de *corpora*. Na sequência, à medida que executaram a construção dos *corpora*, tabularam os esforços necessários para a realização de cada uma das atividades dos projetos. Por fim, realizaram o experimento por meio de um teste estatístico para a comparação dos dados tabulados. Nos tópicos desta seção, explanamos cada um desses passos.

3.1 Como mensurar o esforço?

Apesar de o esforço ser um tema recorrente entre os autores da LC, na revisão da literatura da área, os autores não identificaram trabalhos que tenham se debruçado sobre a sua investigação. Por essa razão, tendo em vista o resultado dessa averiguação e o objetivo da pesquisa, eles precisaram formular métricas e um método de mensuração do esforço dos projetos de construção de *corpora*. A criação da métrica feita por eles baseou-se no conceito de medição proposto por Fenton e Bieman (2014), no livro "*Software Metrics: A Rigorous and Practical Approach*". Conforme esses dois autores, a "medição é o processo pelo qual números ou símbolos são associados aos atributos[14] de uma entidade[15] do mundo real, de modo que seja possível descrevê-los de acordo com um conjunto

[14] Os atributos são as características ou propriedades das entidades.

[15] As entidades são representações de objetos e eventos do mundo real. Por exemplo: uma pessoa, um lugar, um objeto, uma ideia, um produto, um processo ou uma atividade. Do mesmo modo que uma pessoa (entidade) pode ser descrita a partir de suas características (por exemplo: altura, sexo e idade), as atividades podem ser descritas a partir de seus atributos (por exemplo: duração, *inputs* e *outputs*).

de regras[16] bem definidas"[17] (FENTON; BIEMAN, 2014, p. 5) e compará-los com atributos semelhantes de outras entidades.

A partir do conceito de métrica citado, os pesquisadores assumiram as atividades dos projetos de construção de *corpora* como entidades e estabeleceram que o *input*,[18] o *output*[19] e o tempo de duração seriam os seus atributos. Deste modo, eles passaram a dispor de elementos para mensurar o esforço das atividades e foram capazes de compor uma métrica,[20] o Esforço da Atividade (EA), que quantifica, em segundos, o esforço despendido para a completude de uma atividade de um projeto de construção de *corpus*. O Esforço da Atividade (EA) estabelece que o esforço de uma atividade é igual ao quociente entre o intervalo de tempo decorrido entre o início e o fim da atividade (a duração da atividade) e o resultado da atividade, ou seja, a sua completude. Nessa relação, o tempo do pesquisador é o *input*[21] que equivale ao tempo de duração da atividade e o resultado da atividade é o *output* que corresponde ao número 1 (um)[22] – forma que os autores estabeleceram para quantificar e denotar a completude da atividade.[23] A Figura 7 ilustra os 3 atributos de uma atividade no escopo do EA.

---

[16] As regras ditam como a medição deve ser realizada.

[17] Original: "Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way so as to describe them according to clearly defined rules".

[18] Os *inputs* são as entradas necessárias para a realização de uma atividade. No caso, realizamos um recorte nas entradas que considerou apenas o tempo despendido pelo criador do *corpus* na execução da atividade.

[19] Os *outputs* são os produtos ou entregas (resultados) de uma atividade.

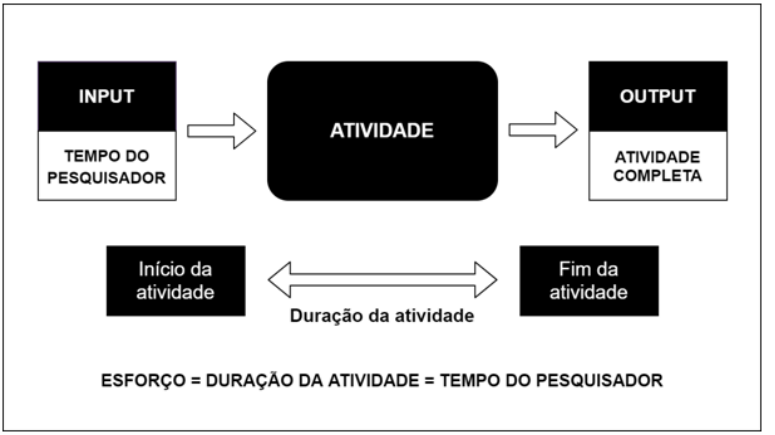
[20] Métricas são unidades de medidas criadas a partir de medições.

[21] Apesar de utilizarem somente o tempo do pesquisador como *input* da atividade, os autores estão cientes da existência de outros *inputs* necessários para a realização de uma tarefa, como o conhecimento do pesquisador. A decisão pelo uso do tempo do pesquisador justifica-se pelo fato de o tempo ser, geralmente, reportado como o recurso primário para a execução de uma atividade. Ademais, o tempo do pesquisador apresenta-se como um *input* quantificável e de fácil mensuração em relação aos *inputs* mais abstratos, como o conhecimento.

[22] De acordo com o raciocínio aplicado, a não completude da atividade corresponderia ao número 0 (zero).

[23] A completude de uma atividade pode ser compreendida como a finalização de 100% de suas tarefas.

FIGURA 7 – Atributos do modelo de regras do EA



Fonte: Oliveira (2019, p. 77.)

Ao criarem o EA, os autores passaram a dispor da variável básica para o cálculo do esforço investido em um projeto de construção de *corpora*. A forma de calcular esse esforço foi a criação de uma outra métrica, o Esforço Total do Projeto (ETP). O ETP é uma métrica que quantifica, em segundos, o esforço despendido para a completude[24] de um projeto de construção de um *corpus* e corresponde à soma de todos os EAs das atividades do projeto. Assim, o ETP pode ser expresso da seguinte maneira: ETP = EA 1 + EA 2 + EA 3 + EA 4 + EA *n*. Em suma, a comparação entre os esforços empregados na construção de cada versão do CoCLI deu-se pela comparação entre os ETP de cada um dos projetos. Para encontrar o ETP de cada projeto, em um primeiro momento, foi calculado o EA de cada atividade dos projetos. O método para o cálculo do ETP referente à construção da versão do CoCLI que não passou pela intervenção do *ToGatherUp* consistiu nos passos a seguir:

1. Identificação das atividades realizadas para a construção do projeto de acordo com a sistemática de organização de projetos proposta pelos autores da pesquisa. As atividades identificadas foram: a) localização dos dados; b) permissão de uso dos dados; c) captura dos dados; d) conversão dos dados; e) limpeza e normalização dos

[24] A completude de um projeto pode ser compreendida como a finalização de 100% das suas atividades.

dados; f) salvamento de arquivos; g) enriquecimento dos dados; h) nomeação dos arquivos;

2. Cálculo do EA das atividades identificadas. Para melhor compreensão, os autores atribuíram uma sigla para cada EA calculado. Desse modo, obtiveram a seguinte lista: a) Esforço da Atividade da localização dos dados (EALD); b) Esforço da Atividade da obtenção de permissão de uso dos dados (EAOPD); c) Esforço da Atividade de captura dos dados (EACD); d) Esforço da Atividade de conversão dos dados (EACVD); e) Esforço da Atividade de limpeza e normalização dos dados (EALND); f) Esforço da Atividade de salvamento de arquivos (EASA); g) Esforço da Atividade de enriquecimento dos dados (EAED); h) Esforço da Atividade de nomeação dos arquivos (EANA).
3. Aplicação do modelo de cálculo do ETP, expresso por: ETP1[25] = EALD + EAOPD + EACD + EACVD + EALND + EASA + EAED + EANA.

No que tange ao método para o cálculo do ETP concernente à elaboração da versão do CoCLI que contou com a incorporação do *ToGatherUp*, os passos foram:

1. Identificação das atividades realizadas para a construção do projeto de acordo com a sistemática de organização de projetos proposta pelos autores da pesquisa. As atividades identificadas[26] foram: a) localização dos dados; b) permissão de uso dos dados; c) captura dos dados; d) conversão dos dados; e) limpeza e normalização dos dados; f) cadastramento de textos;[27]
2. Cálculo do EA das atividades identificadas. De forma análoga ao passo dois do método citado anteriormente, os autores atribuíram siglas para cada EA calculado. Logo, obtiveram a seguinte lista:

[25] O ETP1 diz respeito ao projeto não intervencionado pelo *ToGatherUp*.

[26] As atividades a, b, c e d são comuns aos dois projetos. As atividades de salvamento, nomeação de arquivos e enriquecimento dos dados foram automatizadas pelos recursos do *ToGatherUp* e, por isso, não geraram seus respectivos EAs. Portanto, não as incluímos no cálculo do projeto intervencionado pelo *ToGatherUp*.

[27] O cadastramento de texto é uma atividade específica da construção de *corpora* no *ToGatherUp*.

a) EALD; b) EAOPD; c) EACD; d) EACVD; e) EALND; f) Esforço da Atividade de cadastramento de textos (EACT).

3. Aplicação do modelo de cálculo do ETP, expresso por: ETP2[28] = EALD + EAOPD + EACD + EACVD + EALND + EACT.

Em ambos os projetos, os pesquisadores fizeram a medição da duração das atividades, necessária para a obtenção do EA de cada uma das atividades, com o uso de um cronômetro disponível na interface do *ToGatherUp*. Para a obtenção da quantidade de segundos relativa à duração de uma atividade, o cronômetro foi acionado assim que a atividade foi iniciada e paralisado logo após a conclusão dela. A informação[29] fornecida pelo cronômetro (Instrumento 1) foi tabulada em uma planilha do *Google* (Instrumento 2), que serviu para a extração do conjunto de dados (*dataset*) analisado no experimento da pesquisa.

Além do EA e do ETP, os pesquisadores criaram a métrica Esforço Total de Coleta do Texto (ETCT) para determinar o esforço despendido para a inclusão de uma única unidade de texto em um *corpus.* O ETCT pode ser expresso, de forma semelhante ao ETP, pela fórmula ETCT = EA 1 + EA 2 + EA 3 + EA 4 + EA *n.* Porém, o contexto de aplicação da expressão é limitado somente aos EAs de uma única unidade textual.

3.2 O *ToGatherUp*

O *ToGatherUp*[30] é uma ferramenta *on-line* (http://www.ileel.ufu.br/togatherup) desenvolvida pelos autores da pesquisa aqui retratada que oferece suporte a projetos de construção manual de *corpora*. As principais funcionalidades da ferramenta são a inserção automática de cabeçalho de metadados nos arquivos do *corpus*, a nomeação do

[28] O ETP2 alude ao projeto intervencionado pelo *ToGatherUp*.

[29] O tempo decorrido entre o início e o fim da atividade. Ou seja, a duração da atividade.

[30] O nome *ToGatherUp* surgiu da associação entre o ato de construir um corpus e o verbo frasal *gather up*, da língua inglesa, que, de acordo com o *Macmillan Dictionary* significa "pegar coisas de lugares diferentes e colocá-las juntas", no original "to pick up things from several different places and put them together" (MACMILLAN DICTIONARY, 2018). Para reforçar a associação, no *design* da logomarca da ferramenta, os autores incluíram o símbolo 輯, um ideograma da língua japonesa que, conforme Jisho (http://jisho.org), um dicionário japonês *on-line*, pode ser traduzido para as seguintes palavras da língua inglesa: a) *gather*; b) *collect;* c) *compile*.

arquivo de acordo com uma convenção preestabelecida pelo criador do *corpus* e o armazenamento do arquivo no diretório correspondente ao seu posicionamento na estrutura hierárquica do projeto. Além dessas funcionalidades, o *ToGatherUp* exibe ao pesquisador uma interface em que é possível visualizar estatísticas sobre a quantidade de textos e palavras coletadas, conferindo a ele maior controle em relação ao andamento de um projeto. O *ToGatherUp* possui os seguintes recursos:

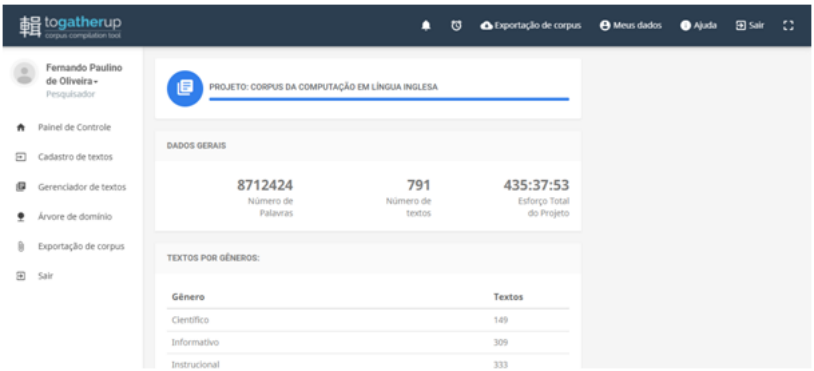
1) Painel de Controle (*Data Overview*): permite a visualização da quantidade total de palavras e de textos de um *corpus*, do ETP, das quantidades de palavras e textos para cada um dos gêneros, tipos textuais, meios de distribuição, áreas e subáreas de um *corpus* e facilita o acompanhamento visual da evolução da coleta de textos de um *corpus*;
2) Cadastro de Textos (*Data Entry*): apresenta um formulário com os campos para a entrada dos dados do texto. Os campos são: Subárea, Título, Língua, Fonte, Gênero textual, Tipos textuais, Meio de Distribuição e ETCT;
3) Gerenciador de Textos (*Data Manager*): interface que exibe uma lista com os textos de um *corpus*, em forma de tabela, e possibilita a localização (pesquisa) de um texto (ou textos) a partir dos metadados dele;
4) Árvore de Domínio (*Domain Tree*): interface para a visualização da organização hierárquica adotada no projeto de um *corpus*;
5) Exportação de *Corpus* (*Data Exporter*): funcionalidade que exporta os arquivos de um *corpus* em diretórios organizados de acordo com a hierarquia do projeto.

Na sequência, abordamos cada um dos referidos recursos do *ToGatherUp*, ilustrando-os com exemplos extraídos do projeto de construção do CoCLI em que ocorreu a incorporação do *ToGatherUp*.

3.2.1 Painel de Controle

O Painel de Controle é a interface principal do *ToGatherUp* e exibe as informações gerais do projeto de construção de um *corpus*. O objetivo dele é oferecer ao pesquisador uma visão geral da evolução da coleta de textos do *corpus*. A Figura 8 ilustra parcialmente o Painel de Controle do *ToGatherUp* com informações do projeto do CoCLI.

FIGURA 8 – Painel de Controle com informações do CoCLI



Fonte: *ToGatherUp*.

O Painel de Controle contém outros cinco painéis: a) Dados gerais, que apresenta o número total de palavras, a quantidade total de textos e o ETP de um *corpus*; b) Textos por Gêneros, que apresenta as quantidades totais de textos para cada gênero textual que compõe um *corpus*; c) Textos por Tipos Textuais, que apresenta o número total de textos para cada tipo textual de um *corpus*; d) Textos por Meios de Distribuição, que apresenta, de maneira discriminada, os meios de comunicação em que os textos de um *corpus* foram obtidos durante sua coleta, quantificando-os; e) Textos por Áreas e Subáreas, que fornece a visão da quantidade de textos e de palavras para cada item da hierarquia adotada num projeto de construção de *corpus*.

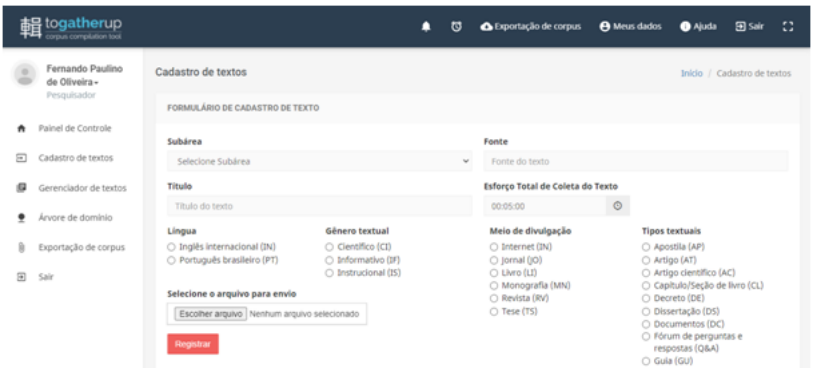
3.2.2 Cadastro de Textos

No *ToGatherUp*, a inclusão de um texto em um *corpus* é realizada através do recurso denominado Cadastro de Textos, presente na Figura 9. O Cadastro de Textos é uma interface *web* que apresenta um formulário, a ser preenchido pelo pesquisador de forma manual, composto pelos seguintes campos:[31] (a) Subárea; (b) Título; (c) Língua; (d) Fonte; (e) Gênero Textual; (f) Tipos Textuais; (g) Meio de Distribuição; (h) ETCT.

[31] Os campos citados foram estabelecidos para o projeto do CoCLI. Os campos do Cadastro de Textos podem ser definidos pelo pesquisador no momento da configuração do projeto no *ToGatherUp*. A data de publicação do texto e a sua autoria são exemplos de informações que podem ser incluídas durante a configuração do projeto.

Além desses campos, o formulário apresenta, ainda, a opção para que o pesquisador possa anexar o arquivo do texto.

FIGURA 9 – Formulário de Cadastro de Textos do *ToGatherUp*



Fonte: *ToGatherUp*.

Ao incluirmos um texto no *corpus* por meio do Cadastro de Textos, o *ToGatherUp* desencadeia, de forma automática, as atividades: a) Atividade 1: Registro dos metadados do texto no banco de dados; b) Atividade 2: Nomeação[32] do arquivo do texto; c) Atividade 3: Inserção de cabeçalho no arquivo do texto; d) Atividade 4: Armazenamento do arquivo do texto. Na sequência, descrevemos cada uma dessas atividades e como elas foram configuradas no projeto de construção do CoCLI.

a) Atividade 1: Registro dos metadados do texto no banco de dados

Ao armazenar os textos de um *corpus*, o pesquisador precisa estabelecer padrões descritivos que otimizem o acesso a eles, a recuperação e o reuso deles. Para atender essa necessidade, o *ToGatherUp* faz uso de metadados para a catalogação dos textos dos *corpora*. A utilização de metadados surgiu no âmbito das Ciências da Informação

[32] Na realidade, o que ocorre é uma renomeação, porque, para que seja possível a sua submissão no *ToGatherUp*, o arquivo precisa ter sido previamente salvo pelo pesquisador. O *ToGatherUp* desconsidera qualquer que seja o nome dado a um arquivo submetido a ele e procede com a sua renomeação em conformidade com os metadados do texto e com a convenção de nomeação de arquivos do projeto.

como uma solução para a organização de dados. Para Alves (2010), os metadados podem ser definidos como:

> [...] atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação (ALVES, 2010, p. 47).

Para o projeto do CoCLI, cada campo do formulário de Cadastro de Textos correspondeu à um metadado estabelecido de acordo com os critérios de desenho. Desse modo, os metadados do CoCLI apresentam-se conforme o Quadro 1.

QUADRO 1 – Metadados do CoCLI

Metadados	Descrição
(a) Subárea	Informa a subárea do texto.
(b) Título	Informa o nome dado para o texto.
(c) Língua	Informa o idioma em que o texto foi escrito.
(d) Fonte	Informa a origem do texto.
(e) Gênero textual	Informa o gênero textual do texto.
(f) Tipos textuais	Informa o tipo textual do texto.
(g) Meio de distribuição	Informa o meio em que o texto foi divulgado.
(h) ETCT	Informa o esforço total referente à soma de todos os EAs realizados para a inclusão de uma unidade textual no *corpus*.[33]

Fonte: Oliveira (2019, p. 92.)

[33] A obtenção do ETCT depende do registro do EA de cada uma das atividades necessárias para a coleta do texto. É importante lembrar que o ToGatherUp não apresenta uma forma de registro para cada um dos EAs. O software tem somente um cronômetro que pode ser utilizado para a captura da duração de cada atividade, que pode ser registrada em um tipo de controle escolhido pelo pesquisador.

Além dos metadados do Quadro 1, o *ToGatherUp* registra, de forma automática, o um conjunto de metadados sem que ocorra a intervenção do pesquisador. O Quadro 2 apresenta esses metadados.

QUADRO 2 – Metadados gerados de forma automática pelo *ToGatherUp*

Metadados	**Descrição**
(a) Domínio	Informa o domínio do texto (área do conhecimento/ especialidade a qual pertence).[34]
(b) Número de palavras	Informa o número de palavras do texto.[35]
(c) Data da inclusão	Informa a data e a hora em que o texto foi incluído no *corpus*.[36]
(l) Identificador do arquivo (ID)	Informa o número de identificação do texto no banco de dados do *ToGatherUp*.[37]

Fonte: Oliveira (2019, p. 92.)

b) Atividade 2: Nomeação dos arquivos dos textos

O *ToGatherUp* faz a nomeação automática dos textos de um *corpus* durante a submissão deles pelo Cadastro de Textos de acordo com os metadados do texto e com uma convenção de nomeação de arquivos definida durante a configuração do projeto no sistema. Com base na convenção estabelecida para a nomeação dos textos do CoCLI, um dos textos desse *corpus* foi nomeado, por exemplo, desta forma: IN-CO-IF-AT-IN-25Sep2017-797.txt. O nome do arquivo é constituído por sete partes distintas, separadas por hífen, e finalizado com a extensão correspondente ao formato dele (.txt). Cada uma das partes é formada por uma abreviação que se associa a um metadado do texto:

a) a primeira (IN) informa a língua do texto. Para a língua inglesa, foi utilizada a abreviação IN;

[34] O domínio do texto é estabelecido durante as configurações do projeto no *ToGatherUp*. Por essa razão, o *ToGatherUp* é capaz de incluí-lo, automaticamente, como um metadado.
[35] O *ToGatherUp* possui um algoritmo que contabiliza a quantidade de palavras do texto.
[36] O *ToGatherUp* considera a data e a hora do servidor em que o sistema está instalado. Por isso, o pesquisador não precisa informar esses dados.
[37] O ID é gerado de forma incremental e automática pelo *ToGatherUp*.

b) a segunda (CO) diz respeito ao domínio (área do conhecimento/especialidade a que pertence o texto). Como o CoCLI é do domínio da Computação, foi utilizada CO para abreviá-lo;

c) a terceira (IF) refere-se ao gênero do texto. A abreviação CI foi utilizada para o gênero científico, a IF para o gênero informativo e a IS para o gênero instrucional;

d) a quarta (AT) alude ao tipo do texto. As abreviações referentes aos tipos textuais dos textos do CoCLI foram estabelecidas da seguinte maneira:
   - Apostila (AP);
   - Artigo (AT);
   - Artigo científico (AC);
   - Capítulo/Seção de livro (CL);
   - Decreto (DE);
   - Dissertação (DS);
   - Documentos (DC);
   - Fórum de perguntas e respostas (Q&A);
   - Guia (GU);
   - Livro (LV);
   - Manual (MA);
   - Monografia (MN);
   - Norma técnica (NR);
   - Nota técnica (NT);
   - Notícia (NO);
   - Portaria (PA);
   - Relatório (RL);
   - Reportagem (RP);
   - Tese (TS);
   - Transcrição (TR);
   - Tutorial (TT).

e) a quinta parte (IN) é relativa ao meio de divulgação do texto. Como todos os textos do CoCLI são provenientes da Internet, foi utilizada a abreviação IN para representá-la;

f) a sexta parte (25Sep2017) informa a data de coleta do texto;

g) a sétima parte (797) indica o identificador (ID) do texto no banco de dados do *ToGatherUp*. Cada texto recebe um ID único ao ser registrado no banco de dados do sistema, o que evita a possibilidade de que textos com metadados idênticos recebam um mesmo nome.

c) Atividade 3: Inserção de cabeçalho nos arquivos de texto

Os metadados dos textos do CoCLI foram usados pelo *ToGatherUp* para a criação e inserção automática de cabeçalho nos arquivos dos textos. Para que isso fosse possível, nas configurações da ferramenta, foi necessário estabelecer a estrutura do cabeçalho a ser utilizada que, na pesquisa, conteve apenas a origem do texto e a sua data de inclusão no *corpus*. Com base nisso, o *ToGatherUp* procedeu com a inserção do cabeçalho nos textos, alimentando-os com os metadados fornecidos no Cadastro de Textos da ferramenta.

d) Atividade 4: Armazenamento do arquivo do texto

O armazenamento de arquivos através de métodos tradicionais comuns no cotidiano das organizações e no gerenciamento de informações pessoais é natural. No entanto, Dourish (2003, p. 4) aponta que estudos realizados por Barreau e Nardi (1995) e por Kaptelinin (1996) revelam que essa prática é problemática, pois dificulta a reorganização das informações quando elas assumem funções diferentes das originais ou quando elas não se adequam a somente um dos *loci* de armazenamento.

Considerando essa problemática, o *ToGatherUp* foi desenvolvido de modo que ele fosse capaz tanto de armazenar os textos do CoCLI de acordo com a Árvore de Domínio da Computação (uma estrutura hierárquica fixa) como de reorganizá-los seguindo outras configurações hierárquicas. A capacidade do *ToGatherUp* de reorganizar o armazenamento dos textos deve-se à incorporação, em seu desenvolvimento, de um modelo conceitual de gerenciamento de arquivos chamado *Placeless Documents.* O *Placeless Documents* foi criado por Paul Dourish (2003), pesquisador do *Xerox Palo Alto Research Center*, localizado em Palo Alto, na Califórnia, nos Estados Unidos, e propõe a organização de documentos a partir das suas propriedades, conforme as diferentes necessidades de seus usuários.

Nesse modelo, a associação das propriedades dos documentos (informações sobre os próprios documentos), chamadas de *active properties*, aos documentos permite que eles sejam organizados de acordo com essas propriedades ao invés de obedecerem a uma estrutura hierárquica predeterminada. Ao oferecer essa nova forma de organização baseada em propriedades, o *Placeless Documents* possibilita o agrupamento, de diferentes maneiras, de um conjunto de documentos,

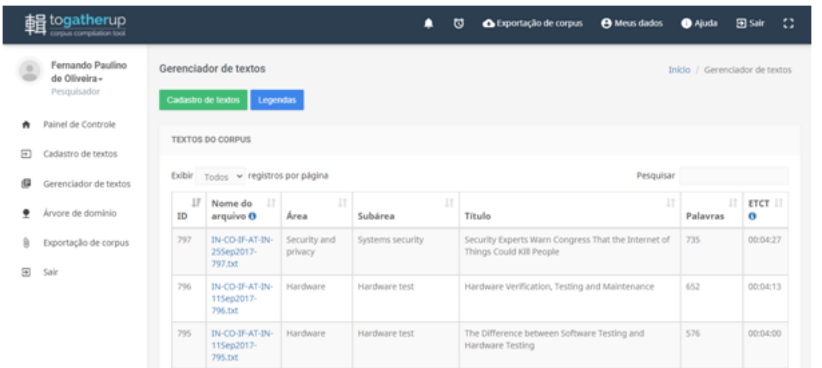
o que soluciona o problema da reorganização de arquivos de acordo com suas funções. A flexibilidade proporcionada pelo *Placeless Documents* foi a principal razão para a sua incorporação ao *ToGatherUp*. No entanto, o modelo implementado no *ToGatherUp* baseou-se na associação dos metadados dos textos do CoCLI ao invés das propriedades dos seus arquivos.

Ao submetermos um texto por meio do formulário do Cadastro de Textos do *ToGatherUp*, seus metadados são registrados no banco de dados do sistema e seu arquivo é armazenado em um diretório comum do servidor *web* em que o sistema está instalado. Por seguir o modelo *Placeless Documents*, o local de armazenamento dos arquivos dentro da infraestrutura do *ToGatherUp* é irrelevante, uma vez que será a necessidade do pesquisador que irá determinar seu posicionamento na estrutura de diretórios, que é gerada no momento da sua exportação para o processamento em outras ferramentas computacionais.

3.2.3 O Gerenciador de Textos

Além das informações quantitativas disponíveis no Painel de Controle, o *ToGatherUp* apresenta uma interface, nomeada como Gerenciador de Textos, que permite ao pesquisador a visualização dos textos de um *corpus*, em forma de tabela, e a pesquisa por um ou mais textos do *corpus* com base em suas informações. A Figura 10 mostra a interface do Gerenciador de Textos.

FIGURA 10 – Gerenciador de Textos do *ToGatherUp*



Fonte: *ToGatherUp*.

A tabela do Gerenciador de Textos possui as colunas: a) ID; b) Nome do arquivo; c) Área; d) Subárea; e) Título; f) Palavras (número de palavras); g) ETCT. O clique sobre o título de cada coluna faz com que suas informações sejam visualizadas em ordem crescente ou decrescente, no caso dos dados numéricos, ou em ordem alfabética, no caso dos dados alfabéticos ou alfanuméricos.

3.2.4 A Exportação de *Corpus*

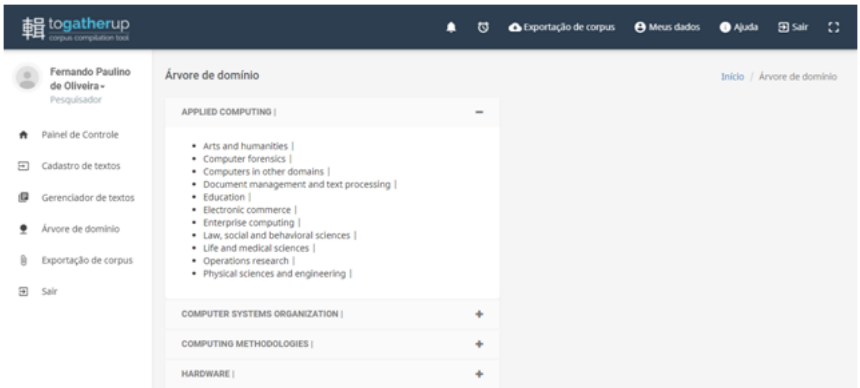
Ao término do projeto de construção de um *corpus*, os seus dados precisam ser disponibilizados para o processamento em ferramentas computacionais. Pensando nisso, o *ToGatherUp* possui o recurso Exportação de *Corpus*, que operacionaliza a exportação dos arquivos do *corpus* de modo que eles possam ser manipulados em outras ferramentas computacionais. Para realizar a exportação, o pesquisador deve utilizar a opção Exportação de *Corpus*, disponível na barra superior e no menu principal do sistema, a qualquer momento que julgar necessário. Ao acioná-la, o *ToGatherUp* cria, de forma automática, um arquivo compactado contendo os textos do *corpus* organizados em diretórios e subdiretórios, conforme a estrutura estabelecida para o projeto, nomeados de forma padronizada e com os seus cabeçalhos já inseridos no início do conteúdo de cada arquivo.

No caso da exportação dos dados do CoCLI, o *ToGatherUp* criou, automaticamente, um arquivo compactado com a extensão .zip, contendo os textos do *corpus* alocados em diretórios correspondentes aos seus campos semânticos na Árvore de Domínios da Computação. No entanto, é importante salientarmos que a flexibilidade oferecida pelo modelo *Placeless documents* e pelo uso dos metadados permite que a exportação seja feita de acordo com outros esquemas. Para exemplificarmos, poderíamos gerar, a partir do conjunto de textos do CoCLI, um *subcorpus* composto somente por textos do gênero científico, caso as configurações de exportação do *ToGatherUp* fossem definidas para esse novo esquema.

3.2.5 A Árvore de Domínio

A Árvore de Domínio é a interface do *ToGatherUp* que exibe a organização hierárquica adotada no projeto de construção de um *corpus*. A Figura 11 exibe a Árvore de Domínio da Computação, com suas áreas e subáreas, adotada no projeto de construção do CoCLI.

FIGURA 11 – Árvore de Domínio do CoCLI



Fonte: *ToGatherUp*.

3.3 A construção do CoCLI

Conforme explicamos anteriormente, a pesquisa aqui retratada comparou os esforços investidos na construção de duas versões idênticas do *Corpus* da Computação da Língua Inglesa (CoCLI), sendo que o projeto de uma delas contou com a incorporação do *ToGatherUp* e o outro não. Ou seja, os projetos foram executados por meio de métodos distintos. Com o intuito de facilitar a compreensão do texto, utilizamos a expressão "Método 1" para referenciar o método que não envolveu a incorporação do *ToGatherUp* e "Método 2" para o que adotou a ferramenta. Apesar da distinção, os dois métodos apresentam um conjunto de atividades em comum e um conjunto de atividades próprias de cada um deles. A seguir, descrevemos a parte comum entre os métodos e, na sequência, tratamos da parte em que eles se distinguem um do outro.

3.3.1 Atividades comuns dos métodos 1 e 2

A parte comum entre os métodos 1 e 2 compreende atividades das fases inicial e de execução dos projetos de construção de *corpora*. A primeira atividade realizada foi a definição do desenho do *corpus*. Após a seleção e definição dos critérios, o desenho do CoCLI apresentou a configuração do Quadro 3.

QUADRO 3 – Desenho do CoCLI

Critério	Definição
Objetivo	Recuperar informações, extrair termos, definir termos e identificar exemplos de uso de termos.
Domínio:[38]	Textos restritos às áreas e subáreas da Computação.
Tipo	Especializado (composto por textos das áreas e subáreas da Computação).
Tempo	Sincrônico (contempla textos publicados no período de 2000 a 2018).
Língua	Monolíngue (apenas textos escritos na língua inglesa).
Gênero e tipo textual	Textos científicos (artigos científicos, capítulos/seções de livro, teses, dissertações, monografias e livros), informativos (artigos, notícias, relatórios e reportagens) e instrucionais ou normativos (apostilas, perguntas e respostas de fóruns, guias, manuais, decretos, normas técnicas, notas técnicas, portarias, tutoriais e documentos).
Tamanho	Cada campo nocional da CSS deverá contar com, no mínimo, 100 mil palavras.[39]
Modalidade	Escrita.
Público-alvo	Pesquisadores, aprendizes e profissionais da Computação.
Estado natural dos textos	Formato eletrônico e sem a necessidade de reconhecimento de seus caracteres.[40]

Fonte: Oliveira (2019, p. 111)

Após o estabelecimento do *design* do CoCLI, foram definidos os recursos financeiros, tecnológicos, materiais e humanos que seriam despendidos para a execução dos projetos e o cronograma das suas realizações. Como todo o trabalho da pesquisa foi realizado pelos seus autores e a hospedagem do *ToGatherUp* foi feita, de forma gratuita, no

[38] Assunto do *corpus*.

[39] Os autores da pesquisa não identificaram, na literatura da LC, um número padrão estabelecido para um *corpus* ou para as ramificações de uma Árvore de Domínio. Por essa razão, estabeleceram o número de 100 mil palavras como padrão para a pesquisa, partindo do pressuposto de que esse valor é suficiente para a recuperação de informações em uma pesquisa terminológica

[40] Essa condição dos textos facilita a captura deles.

servidor *web* do ILEEL[41] da UFU,[42] como parte dos projetos do GPELC,[43] sob o domínio www.togatherup.ileel.ufu.br, não foi necessário investir recursos financeiros para a sua realização. Com essas definições, foi encerrada a fase inicial dos projetos de construção do CoCLI e passou-se à fase de execução deles, na qual iniciou-se o processo de obtenção dos dados dos *corpora*.

A primeira atividade desse processo foi a localização de textos que pudessem ser incluídos nos *corpora*. A escolha dos textos foi feita com base nas referências dos currículos dos cursos da área da Computação de centros acadêmicos de referência. Os textos foram obtidos na Internet e a permissão para o seu uso não foi solicitada devido à dificuldade de obtenção de autorização para a grande quantidade de materiais necessária para os *corpora*. Além desse motivo, a obtenção da permissão de uso foi desconsiderada pela falta de intenção de publicização do CoCLI ao término de sua construção.

Logo após, iniciou-se o processo de preparação dos textos para incluí-los nos *corpora.* Os textos dos *corpora* foram obtidos em suas fontes originais nos formatos PDF, DOC e HTML, convertidos para o formato TXT, limpos por meio da remoção dos elementos citados no Quadro 4 e normalizados por meio dos procedimentos descritos no Quadro 5.

QUADRO 4 – Procedimentos de limpeza de *corpus*

Procedimentos
(a) Remoção de cabeçalhos e rodapés de páginas.
(b) Remoção de elementos gráficos (figuras, imagens e gráficos).
(c) Remoção de imagens.
(d) Remoção de notas de rodapé e fim.[44]
(e) Remoção de números de página.
(f) Remoção de referências bibliográficas.

[41] Instituto de Letras e Linguística.

[42] Universidade Federal de Uberlândia.

[43] Grupo de Pesquisa e Estudos em Linguística de *Corpus.*

[44] Optamos por excluir esses elementos dos textos por julgarmos o restante das informações das produções escritas suficiente para os objetivos da pesquisa

(g) Remoção de listas (sumários, tabelas, figuras, abreviações e gráficos).
(h) Remoção de tabelas e quadros.
(h) Remoção de títulos e subtítulos.
(i) Remoção de legendas de tabelas, figuras e quadros.

Fonte: Oliveira (2019, p. 114.)

QUADRO 5 – Procedimentos de normalização textual

Procedimentos
(a) Remoção de hifens no final de linha.
(b) Remoção de quebras de linhas/parágrafos/páginas/seções.
(c) Remoção de espaços em branco duplicados.
(d) Remoção de marcas de parágrafos/recuos.
(e) Remoção de linhas em branco.
(f) Padronização de hifens, apóstrofos, traços e aspas.

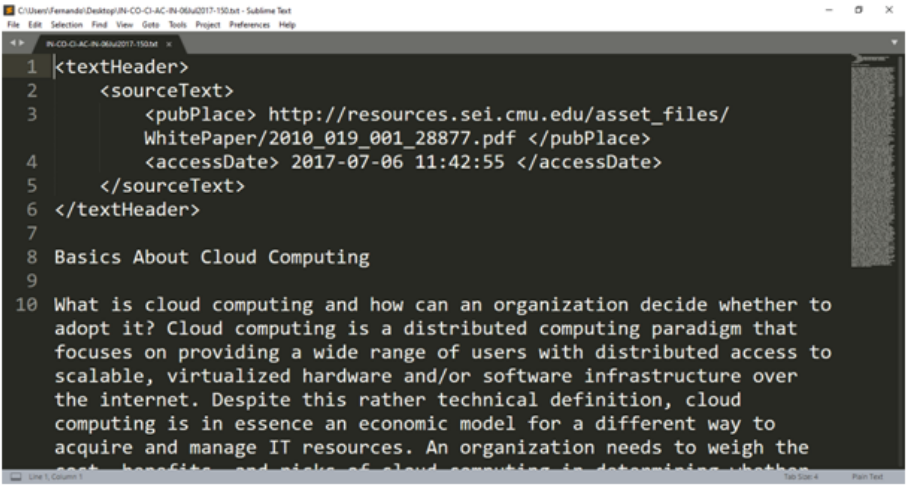
Fonte: Oliveira (2019, p. 114.)

Após a conclusão da conversão, limpeza e normalização dos textos, iniciou-se a execução da última atividade do processo de preparação dos *corpora* – o enriquecimento dos dados.

3.3.2 Atividades distintas dos métodos 1 e 2

As atividades de enriquecimento e relativas ao armazenamento dos dados compõem o conjunto de atividades que tiveram a forma de execução completamente alterada com a incorporação do *ToGatherUp*. O enriquecimento dos dados dos *corpora* da pesquisa consistiu na inserção de cabeçalhos construídos a partir de metadados dos textos. A Figura 12 ilustra um exemplo de cabeçalho inserido pelo *ToGatherUp* em um dos textos do CoCLI.

FIGURA 12 – Cabeçalho do texto *Basics about Cloud Computing*



Fonte: *ToGatherUp*.

No Método 1, o cabeçalho da Figura 12 foi construído e o inserido no arquivo do texto de forma manual. Já no Método 2, o *ToGatherUp* foi programado para construir e inserir, automaticamente, o cabeçalho no texto, de acordo com a estrutura XML definida em sua programação e os metadados do texto. A inclusão dos cabeçalhos encerrou o processo de preparação dos textos, que foi sucedido pelas atividades de armazenamento dos dados dos *corpora*. Para a nomeação dos arquivos dos textos, foi utilizada a convenção de nomeação de arquivos apresentada no subtópico Cadastro de Textos. No Método 1, os textos dos *corpora* foram nomeados manualmente e, no Método 2, todo o trabalho foi executado de maneira automática pelo *ToGatherUp* durante o registro do texto no Cadastro de Textos da ferramenta. Para que isso fosse possível, o *ToGatherUp* foi programado para nomear os arquivos de acordo com as regras da convenção de nomeação de arquivos adotada no projeto e com os metadados dos textos.

A última atividade dos projetos de construção do CoCLI foi o salvamento (arquivamento) dos textos. No Método 1, os arquivos dos *corpora* foram salvos manualmente nos diretórios, criados com o *Windows Explorer* do *Windows*, correspondentes às áreas e subáreas presentes na Árvore de Domínio da Computação. No Método 2, o *ToGatherUp* armazenou automaticamente os arquivos em consonância com os princípios

e a funcionalidade de armazenamento da ferramenta apresentados no item d) Atividade 4: Armazenamento do arquivo do texto do subtópico 3.2.2 Cadastro de Textos. No contexto do *ToGatherUp*, o local dos arquivos é indiferente, uma vez que a ferramenta é capaz de organizar os arquivos de acordo com a consulta estabelecida pelo usuário do sistema. No caso do CoCLI, o *ToGatherUp* foi programado para exportar os arquivos conforme a estrutura da Árvore de Domínio da Computação.

3.4 O experimento

O experimento realizado na pesquisa consistiu na realização de um teste estatístico que comparou os ETP de cada um dos projetos de construção do CoCLI. O objetivo do experimento foi testar a hipótese de que a incorporação do *ToGatherUp* em projetos de construção manual de *corpora* poupa o tempo e minimiza o esforço do pesquisador dispensados à execução das atividades de elaboração de *corpora*, de modo semelhante ao que ocorre com as atividades de análise de *corpora* mediadas pelo uso de computadores (criação automática de listas de palavras e linhas de concordância, evidenciação de padrões linguísticos e etiquetagem de *corpora*). Para a realização do experimento foi necessário realizar a tabulação manual dos EAs, fornecidos pelo cronômetro do *ToGatherUp* (Instrumento 1), em uma planilha do *Google* (Instrumento 2), para cada uma das atividades de cada um dos projetos de construção do CoCLI. Desses conjuntos de dados (*dataset*) foram extraídas amostras aleatórias referentes aos mesmos 50 textos de cada *corpus*. Os dados das amostras foram submetidos a um teste estatístico que permitiu determinar o efeito da incorporação do *ToGatherUp* na construção manual das duas versões do CoCLI.

De acordo com Rumsey (2010), testar uma hipótese é uma tentativa de se "confirmar ou negar uma declaração sobre uma população[45] a partir dos dados de sua amostra"[46] (RUMSEY, 2010, p. 87).[47] Para

[45] Para Correia (2003), população é "uma coleção completa de todos os elementos a serem estudados" (CORREIA, 2003, p. 9).

[46] Consoante Correia (2003), uma subcoleção de elementos extraídos de uma população" (CORREIA, 2003, p. 9).amostra é "uma subcoleção de elementos extraídos de uma população" (CORREIA, 2003, p. 9).

[47] Original: "trying to confirm or deny a claim about a population using data from a sample".

a autora, quando um teste de hipóteses[48] envolve a comparação entre parâmetros numéricos, o objeto de interesse é a diferença entre as médias[49] (*means*) desses parâmetros. Como a análise envolveu a comparação entre o ETP dos diferentes projetos de construção do CoCLI (dois parâmetros numéricos), foi utilizado um teste de hipóteses conhecido como *T-Test* que, segundo Dodge (2008), é apropriado para testar hipóteses a partir da comparação entre as médias de duas populações em que os elementos de uma delas possuem uma relação com os elementos da outra.

Para a realização do T-Test, os dados tabulados foram importados no *software Statistics Statistical Package for the Social Sciences* (SSPS),[50] uma ferramenta de análise estatísticas, desenvolvida pela IBM, amplamente usada em pesquisas acadêmicas que envolvem a realização de testes estatísticos. Na sequência, foi utilizada uma função do SSPS para criar uma amostra aleatória[51] de cinquenta textos do CoCLI. Por fim, o SSPS comparou o ETCT resultante da aplicação do Método 1 (que abreviamos como ETCT – Método 1) e o ETCT resultante da aplicação do Método 2 (que passamos a chamar de ETCT – Método 2) dos 50 textos selecionados. Os dados referentes ao ETCT – Método 1 constituíram o Grupo de Controle[52] (*control group*) do experimento e os dados relativos ao ETCT – Método 2 formaram o Grupo Experimental (*treatment group*). O tratamento que diferenciou o Grupo de Controle do Grupo Experimental foi a manipulação dos EAs automatizados pelo *ToGatherUp* no Método 2.

---

[48] Segundo Correia (2003), um teste de hipóteses é "técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses realizado com os dados amostrais, pode-se fazer inferências sobre a população" (CORREIA, 2003, p. 100).

[49] Segundo Correia (2003), um teste de hipóteses é "técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses realizado com os dados amostrais, pode-se fazer inferências sobre a população" (CORREIA, 2003, p. 100).

[50] O SSPS foi escolhido por realizar os cálculos estatísticos de forma automática. Disponível em: https://www.ibm.com/br-pt/products/spss-statistics. Acesso em: 23 fev. 2019.

[51] Os registros que compuseram o conjunto de dados analisados foram criados de forma automática e aleatória pelo SSPS.

[52] De acordo com Rumsey (2010), as amostras que são expostas a condições normais (não recebem tratamento ou recebem um tratamento falso, também chamado de placebo) denominam-se Grupo de Controle. Já as amostras sujeitas a tratamento que afeta seus atributos são chamadas de Grupo Experimental.

A análise dos resultados fornecidos pelo SSPS considerou os conceitos estatísticos de hipótese nula[53] (*null hypothesis*) e hipótese alternativa (*alternate hypothesis*). Segundo Charles Brase e Corrine Brase (2011, p. 411), a hipótese nula ou "hipótese estatística"[54] é a declaração que está sob teste e, geralmente, associa-se a resultados como "não houve efeito", "não houve diferença" ou "nada foi alterado" entre a média calculada para o Grupo de Controle e a média calculada para o Grupo Experimental. A hipótese alternativa[55] é definida pelos autores como qualquer declaração diferente da hipótese nula. De acordo com os conceitos de hipótese nula e alternativa, a nossa hipótese da pesquisa pode ser feita da seguinte maneira: a hipótese deve ser rejeitada caso o resultado do *T-Test* revele que o ETCT do método que utiliza o *ToGatherUp* é igual ou maior do que o ETCT do método que não utiliza a ferramenta. Se a hipótese nula for rejeitada, ou seja, se o *T-Test* mostrar que o ETCT do método que utiliza o *ToGatherUp* é menor do que o ETCT do método que não utiliza a ferramenta, a hipótese alternativa deve ser aceita e a hipótese confirmada.

O resultado de um teste de hipótese é estatisticamente significativo quando a probabilidade de que ele tenha ocorrido por acaso seja muito improvável. Por essa razão, os autores da pesquisa preocuparam-se em determinar o nível de significância que foi considerado no teste. Para Rumsey (2010), o nível de significância de um teste de hipótese, também conhecido como *alpha level* (α), é dado pelo *p-value* (*probability value*) que, geralmente, é definido em 0.05[56] ou 0.01. Caso o *p-value* é maior ou igual a α, a hipótese nula deve ser aceita e, caso o *p-value* é menor que α, a hipótese nula deve ser rejeitada. Em outras palavras, o resultado de um teste de hipótese é estatisticamente significativo quando, a partir do seu

[53] Na estatística, a hipótese nula é representada por H0 e a hipótese alternativa, por H1.

[54] Para Correia (2003), a hipótese estatística "trata-se de [i.e. trata de] uma suposição quanto ao valor de um parâmetro populacional, ou quanto à natureza da distribuição de probabilidade de uma variável populacional" (CORREIA, 2003, p. 100).

[55] Autores como Rumsey (2010) também usam a expressão "hipótese de pesquisa" para referenciar a hipótese alternativa.

[56] De acordo com Rumsey (2010), um *p-value* de 0.05 e um *p-value* de 0.01 indicam, respectivamente, que em 95% e 99% das vezes os resultados da amostra poderão se repetir caso o experimento seja realizado novamente com outras amostras aleatórias da mesma população sob as mesmas condições. Para Rumsey (2010), outros valores podem ser assumidos para o *p-value* e essa determinação depende de cada pesquisador.

p-value, é possível rejeitar a hipótese nula devido à improbabilidade de que ela ocorra. A rejeição da hipótese nula, consequentemente, leva-nos a acreditar que a hipótese alternativa pode ser verdadeira. Sendo assim, os autores definiram o *p-value* a ser considerado no teste em 0.05 por julgarem esse nível de significância bastante aceitável para o propósito da pesquisa.

4 Resultados

A execução do *T-Test* no SSPS gerou o resultado apresentado na Tabela 1.

TABELA 1 – Resultado do *T-Test*

		Paired Differences					*t*	*df*	*Sig. (2-tailed)*
	Mean	*Std. Deviation*	*Std. Error Mean*	*95% Confidence Interval of the Difference*					
				Lower	*Upper*				
Pair 1	ETCT 1[57] & ETCT 2[58]	131,140	4,333	613	129,909	132,371	214,003	49	0,000

Fonte: Oliveira (2019, p. 141.)

O ETCT da amostra construída com Método 1, indicado na coluna *Mean* da Tabela 1, foi, em média, 131 segundos maior do que o ETCT da amostra elaborada com o Método 2. Esse valor pode ser considerado estatisticamente significativo, uma vez que o *p-value* do teste foi igual a 0,000, conforme indica a coluna *Sig. (2-tailed) da* Tabela 4, um valor bem inferior ao *p-value* (0.05) estabelecido para a garantia da significância estatística do *T-Test*. Portanto, com base no resultado do *T-Test*, a hipótese nula da pesquisa (Hipótese nula (H_0): ETCT – Método 2 = ou > ETCT – Método 1) pode ser rejeitada pelos pesquisadores e eles puderam afirmar, por inferência, que os resultados encontrados sugerem que a incorporação do *ToGatherUp* reduz o ETP de construção manual de *corpora*.

[57] Referente ao Método 1.
[58] Referente ao Método 2.

6 Considerações finais

A pesquisa retratada neste artigo é o resultado de um trabalho sistemático para a determinação do efeito da incorporação do *ToGatherUp* em projetos de construção manual de *corpora* e, até onde pudemos verificar por meio da revisão bibliográfica da LC, consiste em um dos primeiros trabalhos a propor uma forma de mensurar o esforço necessário para a realização de projetos de elaboração manual de *corpora* e a propor uma sistematização do trabalho de criação manual de *corpora*, respeitando princípios e métodos da LC e da área de Gerenciamento de Projetos.

O uso do *ToGatherUp* que, no momento da redação deste artigo, está passando por ajustes para que possa ser disponibilizado em 2021 e utilizado, gratuitamente, em outras pesquisas é outro ponto de destaque da pesquisa. Acreditamos que a disponibilização da ferramenta irá contribuir para o preenchimento da lacuna[59] existente na LC referente à carência de ferramentas voltadas para o suporte das atividades de construção manual de *corpora* compostos por grande volume de dados.

Além dessas contribuições, a pesquisa traz uma importante discussão sobre possíveis complicações do uso de *web corpora* nas pesquisas em que existe a preocupação quanto à precisão de análises, visto que as ferramentas de coleta automática de textos, no estágio atual da tecnologia, não conseguem lidar com os problemas apontados na fundamentação teórica deste artigo. Essa discussão ganha mais relevância ao considerarmos o fato identificado na pesquisa de que o percentual do EALND, em ambos os métodos de construção do CoCLI, foi maior do que todos os demais esforços somados juntos, atingindo 83,29% no Método 1 e 90,02% no Método 2, corroborando a ideia de Dasu e Johnson (2003) de que a limpeza e a normalização podem ocupar cerca de 80% do tempo compreendido entre a obtenção de um texto e sua análise. Se o maior esforço de um projeto de construção de *corpora* está nas atividades de limpeza e normalização e as ferramentas de coleta automática de textos negligenciam essas atividades, as análises feitas a partir de *corpora* coletados automaticamente correm o risco de serem postas em xeque.

[59] A referida lacuna foi identificada por meio de um levantamento realizado durante a pesquisa em que foram analisadas dez ferramentas da LC apontadas para a criação de *corpora* pelo projeto *Corpus Analysis* (KLEIBER; BERBERICH, 2018), desenvolvido por Ingo Kleiber e Kristin Berberich, da Universidade de Heidelberg, na Alemanha.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Leopoldo, v. 4, n. 3, p. 156-178, 2006. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 2 abr. 2019.

ALVES, R. C. V. *Metadados como elementos do processo de catalogação.* 2010. 132f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2010. Disponível em: https://repositorio.unesp.br/handle/11449/103361. Acesso em: 2 abr. 2019.

ANTHONY, L. *EncodeAnt*. Version 1.2.0. [Computer Software]. Tokyo: Waseda University, 2016. Disponível em: http://www.laurenceanthony.net. Acesso em: 2 abr. 2019.

ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Literary and Linguistic Computing,* Oxford, v. 7, n. 1, p. 1-16, 1992. DOI: https://doi.org/10.1093/llc/7.1.1. Disponível em: https://academic.oup.com/dsh/article-abstract/7/1/1/1028498?redirectedFrom=fulltext. Acesso em: 17 abr. 2019.

BAKER, P. Corpus Methods in Linguistics. *In*: LITOSSELITI, L. (ed.). *Research Methods in Linguistics.* New York: Continnum International Publishing Group, 2010. p. 93-113.

BARONI, M.; BERNARDINI, S. *BootCaT*. Version 1.08. [Computer Software]. Trento/Forlì: Universities of Bologna, 2004. Disponível em: http://bootcat.dipintra.it. Acesso em: 2 abr. 2019.

BARONI, M. *et al*. WebBootCaT: A Web Tool for Instant Corpora. *In*: EURALEX INTERNATIONAL CONGRESS, 12th., 2006, Torino. *Proceedings* […]. Torino: Edizioni dell'Orso s.r.l., 2006. p. 123-131. Disponível em: https://euralex.org/publications/webbootcat-a-web-tool-for-instant-corpora/. Acesso em: 2 abr. 2019.

BARREAU, D.; NARDI, B. Finding and Reminding: File Organization from the Desktop. *ACM SIGCHI Bulletin,* New York, v. 27, n. 3, p. 39-43, 1995. DOI: https://doi.org/10.1145/221296.221307. Disponível em: https://dl.acm.org/citation.cfm?id=221307. Acesso em: 17 abr. 2019.

BERBER SARDINHA, T. A influência do tamanho do corpus de referência da obtenção de palavras-chave usando o Programa Computacional Wordsmith Tools. *The ESPecialist*, São Paulo, v. 26, n. 2, p. 188, 2005. Disponível em: https://revistas.pucsp.br/esp/article/view/9290. Acesso em: 27 nov. 2020.

BERBER SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

BERGH, G.; ZANCHETTA, E. Web linguistics. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics:* An International Handbook. Berlin: Mouton de Gruyter, 2008. p. 309-327.

BIANCHI, F. *Culture, Corpora and Semantics*: Methodological Issues in Using Elicited and Corpus Data for Cultural Comparison. Lecce: ESE Salento University Publishing, 2012. Disponível em: http://siba-ese.unisalento.it/index.php/culturecorpora/article/viewFile/12427/11066. Acesso em: 10 jan. 2019.

BIBER, D. Representativeness in Corpus Design. *Literary and Linguistic Computing,* Oxford, v. 8, n. 4, p. 223-257, 1993. DOI: https://doi.org/10.1093/llc/8.4.243. Disponível em: http://otipl.philol.msu.ru/media/biber930.pdf. Acesso em: 2 abr. 2019.

BLECHA, J. *Building Specialized Corpora.* 2012. 159f. Thesis (Master in English Language and Literature) – Faculty of Arts, Department of English and American Studies, Masaryk University, Brno, República Tcheca, 2012. Disponível em: https://is.muni.cz/th/aki90/179991_Building_Specialized_Corpora.pdf. Acesso em: 2 abr. 2019.

BRASE, C. H.; BRASE, C. P. *Understandable Statistics*: Concepts and Methods, 10. ed. Boston: Cengage Learning, 2011.

CORREIA, M. S. B. B. *Probabilidade e estatística*. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003. Disponível em: http://estpoli.pbworks.com/f/livro_probabilidade_estatistica_2a_ed.pdf. Acesso em: 25 fev. 2019.

DASU, T.; JOHNSON, T. *Exploratory Data Mining and Data Cleaning*. Hoboken: John Wiley & Sons, 2003. DOI: https://doi.org/10.1002/0471448354.

DODGE, Y. *The Concise Encyclopedia of Statistics*. New York: Springer-Verlag, 2008.

DOURISH, P. The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents. *Computer Supported Cooperative Work (CSCW),* Dordrecht, v. 12, n. 4, p. 465-490, 2003. DOI: https://doi.org/10.1023/A:1026149119426. Disponível em: https://link.springer.com/article/10.1023/A:1026149119426. Acesso em: 17 abr. 2019.

EDWARD, R. P. Computational Tools and Methods for Corpus Compilation and Analysis. *In*: BIBER, D; REPPEN, R. (ed.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015. p. 32-49.

ESCARTÍN, C. P. Design and compilation of a specialized Spanish-German parallel corpus. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC), 2012, Istanbul. *Proceedings* […]. Istanbul: European Language Resources Association (ELRA), 2012. p. 2199-2206. Disponível em: http://www.lrec-conf.org/proceedings/lrec2012/pdf/577_Paper.pdf. Acesso em: 2 abr. 2019.

FENTON, N.; BIEMAN, J. *Software Metrics*: A Rigorous and Practical Approach. 3. ed. Boca Raton: CRC Press, 2014. DOI: https://doi.org/10.1201/b17461.

FRANKENBERG-GARCIA, A. Prefácio. *In*: SHEPHERD, T. M. G.; BERBER SARDINHA, T.; PINTO, M. V. (org.). *Caminhos da linguística de corpus*. São Paulo: Mercado de Letras, 2012. p. 11-14.

FROMM, G. O uso de *corpora* na análise linguística. *Revista Factus,* São Paulo, v. 1, n. 1, p. 69-76, 2003. Disponível em: http://www.ileel.ufu.br/guifromm/upload/ousodecorporanaproducaolinguistica.pdf. Acesso em: 17 abr. 2019.

FROMM, G. *VoTec*: a construção de vocabulários eletrônicos para aprendizes de tradução. 2007. 214 f. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2007. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/pt-br.php. Acesso em: 2 abr. 2019.

GARRETSON, G. Desiderata for Linguistic Software Design. *Internatinal Journal of English Studies (IJES),* Murcia, v. 8, n. 1, 67-94, 2008. Disponível em: http://revistas.um.es/ijes/article/view/49101. Acesso em: 2 abr. 2019.

GARSIDE, R.; SMITH, N. A Hybrid Grammatical Tagger: CLAWS 4. *In*: GARSIDE, R.; LEECH, G.; MCENERY, T. (eds.). *Corpus annotation*: Linguistic Information from Computer Text Corpora. London: Routledge; Taylor & Francis, 1997. p. 102-121. DOI: https://doi.org/10.4324/9781315841366

GOOGLE. *Refinar pesquisas na Web*, 2019. Disponível em: https://support.google.com/websearch/answer/2466433?hl=pt-BR. Acesso em: 1 abr. 2019.

GRIES, S. T. What is Corpus Linguistics? *Language and Linguistics Compass*, Hoboken, v. 3, n. 5, p. 1225-1241, 2009. DOI: https://doi.org/10.1111/j.1749-818X.2009.00149.x. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1749-818X.2009.00149.x. Acesso em: 2 abr. 2019.

KAPTELININ, V. Creating Computer-Based Work Environments: An Empirical Study of Macintosh Users. *In*: ACM SIGCPR/SIGMIS CONFERENCE ON COMPUTER PERSONNEL RESEARC, 1996, Denver. *Proceedings* […]. Denver: ACM, 1996. p. 360-366. DOI: https://doi.org/10.1145/238857.238921. Disponível em: https://dl.acm.org/citation.cfm?id=238921. Acesso em: 17 abr. 2019.

KEHOE, A.; GEE, M. New Corpora from the Web: Making Web Text More 'Text-Like'. *Studies in Variation, Contacts and Change in English,* Helsinki, v. 2, [s.p.], 2007. Disponível em: http://www.helsinki.fi/varieng/series/volumes/02/kehoe_gee/. Acesso em: 18 jan. 2019.

KENNEDY, G. *An Introduction to Corpus Linguistics*. New York: Longman, 1998.

KLEIBER, I.; BERBERICH, K. *Corpus Analysis*. Heidelberg: Heidelberg University, 2018. Disponível em: https://corpus-analysis.com/. Acesso em: 25 jan. 2019.

KÜBLER, N.; ASTON, G. Using Corpora in Translation. *In*: O'KEEFFE, A.; MCCARTHY, M. J. (org.). *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 2010. p. 501-515. DOI: https://doi.org/10.4324/9780203856949-36

LEECH, G. Adding Linguistic Annotation. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford:

Oxbow Books, 2005. p. 17-29. Disponível em: http://ota.ox.ac.uk/documents/creating/dlc/. Acesso em: 2 abr. 2019.

MACMILLAN DICTIONARY. *Gather Up*. 2018. Disponível em: https://www.macmillandictionary.com/dictionary/british/gather-up. Acesso em: 21 jun. 2018.

MACMULLEN, W. J. Requirements Definition and Design Criteria for Test Corpora in Information Science. *SILS Technical Report 2003-03*. School of Information and Library Science: University of North Carolina at Chapel Hill. p. 3-21, 2003. Disponível em: https://sils.unc.edu/sites/default/files/general/research/TR-2003-03.pdf. Acesso em: 10 jan. 2019.

MARTINET, A. *Elementos de linguística geral.* 8. ed. Lisboa: Martins Fontes, 1978.

MCENERY, T.; HARDIE, A. *Corpus Linguistics*: Method, Theory and Practice. Cambridge: Cambridge University Press, 2011. DOI: https://doi.org/10.1017/CBO9780511981395.

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-Based Language Studies*: An Advanced Resource Book. London; New York: Routledge, 2006. Disponível em: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A10.pdf. Acesso em: 10 jan. 2019.

MEYER, C. F. *English Corpus Linguistics*: An Introduction. Cambridge: Cambridge University Press, 2004.

MINSHALL, D. E. *A Computer Science Word List.* 2013. 98f. Dissertation (Master of Arts - MA TEFL) – University of Swansea, Swansea, UK, 2013. Disponível em: https://www.baleap.org/wp-content/uploads/2016/03/Daniel-Minshall.pdf. Acesso em: 10 jan. 2019.

NELSON, M. Building a Written Corpus: What Are the Basics? *In*: O'KEEFFE, A.; MCCARTHY, M. J. (org.). *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 2010. p. 53-65. DOI: https://doi.org/10.4324/9780203856949-5

NEUMANN, S.; HANSEN-SCHIRRA, S. Corpus Methodology and Design. *In*: HANSEN-SCHIRRA, S.; NEUMANN, S.; STEINER, E. (org.). *Cross-Linguistic Corpora for the Study of Translations:* Insights from the Language Pair English-German. Berlin: De Gruyter Mouton, 2012. p. 21-34. DOI: https://doi.org/10.1515/9783110260328.

OLIVEIRA, F. P. *ToGatherUp:* um protótipo de ferramenta para a construção de *corpora* a produção de vocabulários bilíngues direcionada por corpus. 2019. 219f. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, 2019. Disponível em: https://repositorio.ufu.br/bitstream/123456789/25433/1/ToGatherUpProtótipoFerramenta. Acesso em: 2 abr. 2019.

PROJECT MANAGEMENT INSTITUTE. *Um Guia do Conhecimento em Gerenciamento de Projetos (Guia PMBOK)*. 5. ed. Newtown Square: Project Management Institute, 2013.

RENOUF, A. Corpus Development 25 Years on: From Super-Corpus to Cybercorpus. *Language and Computers: Studies in Practical Linguistics*, [*S.l.*], v. 62, n. 1, p. 27-49, 2007. DOI: https://doi.org/10.1163/9789401204347_004.

RUBI, M. P. Os princípios da política de indexação na análise de assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários. *In*: FUJITA, M. S. L. *et al*. (org.). *A indexação de livros:* a percepção de catalogadores e usuários de bibliotecas universitárias: um estudo de observação do contexto sociocognitivo com protocolos verbais. São Paulo: Cultura Acadêmica, 2009. p. 81-93.

RUMSEY, D. *Statistics Essentials for Dummies*. Hoboken: John Wiley & Sons, 2010.

RUNDELL, M.; KILGARRIFF, A. Automating the Creation of Dictionaries: Where Will It All End? *In*: MEUNIER, F. *et al*. (ed.). *A Taste for Corpora:* A Tribute to Professor Sylviane Granger. Amsterdam: Benjamins, 2011. p. 257-281. DOI: https://doi.org/10.1075/scl.45.15run.

SANTOS, A. *Contributions for Building a Corpora-Flow System.* 2011. 100f. Dissertação (Master in Informatics Engineering) – Escola de Engenharia, Universidade do Minho, Guimarães, PT, 2011. Disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/28122/1/eeum_di_dissertacao_pg15973.pdf. Acesso em: 17 abr. 2019.

SCHÄFER, R.; BILDHAUER, F. *Web Corpus Construction*. Toronto: University of Toronto, 2013. DOI: https://doi.org/10.2200/S00508ED1V01Y201305HLT022.

SEDLAR, E. *Database-Managed File System*. US Pat. US20050091287A1. Redwood Shores, CA: Oracle International Corporation, 2005.

SEMINO, E.; SHORT, M. *Corpus Stylistics*: Speech, Writing and Thought Presentation in a Corpus of English Writing. London: Routledge, 2004. DOI: https://doi.org/10.4324/9780203494073.

SIMSKE, S. J. *Systems and Methods for Processing Text-Based Electronic Documents*. U.S. Patent n. 7,106, 905. [*S.l.*]: Hewlett-Packard Development Company, 2006.

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

SINCLAIR, J. McH. Corpus and Text – Basic Principles. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford: Oxbow Books, 2005. [s.p.]. Disponível em: https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm. Acesso em: 2 abr. 2019.

TAGNIN, S. E. O. Glossário de linguística de *corpus*. *In*: VIANA, V.; TAGNIN, S. E. O. (org.). *Corpora no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2010. p. 349-353.

TAGNIN, S. E. O. *Corpora na tradução*. São Paulo: Hub Editorial, 2015.

VOORMANN, H.; GUT, U. Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory,* Berlin, v. 4, n. 2, p. 235-251, 2008. DOI: https://doi.org/10.1515/CLLT.2008.010.

WIDDOWSON, H.G. *Linguistics*. Oxford: Oxford University Press, 1996.

ZANETTIN, F. *Translation-driven corpora*: Corpus resources for descriptive and applied translation studies. London: Routledge, 2014. DOI: https://doi.org/10.4324/9781315759661.