PROPOSTA PARA A ELABORAÇÃO DE UM GLOSSÁRIO PORTUGUÊS - INGLÊS DA DISCIPLINA LINGUÍSTICA DE *CORPUS* DO LEA-MSI

Marcos Vinícius da Silva[1]

Karine Dourado Silva [2]

RESUMO: A disciplina Linguística de *Corpus*, situada no eixo de Terminologia do curso Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI), por lidar com uma linguagem de especialidade, é rica em termos técnicos, o que pode dificultar a compreensão do conteúdo e a leitura dos textos da disciplina. Sendo assim, a Terminografia, disciplina que se preocupa com a elaboração de glossários e dicionários especializados, é destacada neste trabalho por (i) ser uma potencial área para a minimização dessas problemáticas; (ii) ser um campo no qual o bacharel em LEA-MSI pode atuar. Desta forma, neste trabalho busca-se refletir sobre teorias que tangem à Terminologia, Terminografia e Linguística de *Corpus* e apresentar uma proposta de elaboração de um glossário da disciplina Linguística de *Corpus* do LEA-MSI, a fim de minimizar as possíveis dificuldades causadas pelo número de termos da disciplina.

PALAVRAS-CHAVE: Glossário. Terminologia. Terminografia. Linguística de *Corpus.*

ABSTRACT: The discipline of Corpus Linguistics, situated on the axis of Terminology of the course Applied Foreign Languages to Multilingualism and the Information Society (LEA-MSI), for dealing with a specialized language, is rich in technical terms, which may hinder the understanding of the content and the reading of the discipline texts. Moreover, Terminography, discipline that is concerned with the

[1] Trabalho de conclusão de curso apresentado junto ao curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação da Universidade de Brasília, na área de Línguas, Léxico e Terminologia como requisito parcial à obtenção do título de Bacharel, no 2º semestre de 2014.
[2] Professora do curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação e orientadora deste projeto.

development of glossaries and specialized dictionaries, is highlighted in this paper for (i) being a potential area to minimize these problems; (ii) being a field in which a professional who has a degree in LEA-MSI can work. Thereby, this work seeks to reflect on theories that concern Terminology, Terminography and Corpus Linguistics and present a proposal to develop a glossary of the Corpus Linguistics LEA-MSI's discipline, in order to minimize possible difficulties caused by the number of terms of the discipline.

KEYWORDS: Glossary. Terminology. Terminography. Corpus Linguistics.

"*Filho, desde a tua mocidade aplica-te à disciplina e até com cabelos brancos encontrarás a sabedoria. Como o lavrador e o semeador, cultiva-a, e espera pacientemente seus bons frutos, porque te cansarás um pouco em seu cultivo, mas em breve comerás de seus frutos*." – Eclesiástico 6, 18 – 20.

1. INTRODUÇÃO

A Linguística de *Corpus* é considerada essencial para o curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação, tanto é que o fluxo do curso oferta uma disciplina obrigatória, de mesmo nome, para os alunos do último semestre. Situada no eixo de Terminologia do curso, a Linguística de *Corpus*, além de seus aportes teóricos, dispõe de metodologias e ferramentas que o bacharel em LEA-MSI pode aplicar em estudos linguísticos que ele venha a trabalhar. Sendo uma disciplina de suma importância, percebeu-se a necessidade de uma atenção especial ao termos utilizados nessa disciplina que, por experiência pessoal, não são fáceis de compreender sem pesquisas prévias. De tal modo, surgiu a problemática: como potencializar o processo de aprendizagem dos estudantes facilitando a compreensão dos termos utilizados na disciplina Linguística de *Corpus* do LEA-MSI?

Através dos subsídios teóricos que o curso oferece aos estudantes, percebeu-se que, por meio da Terminologia e da Terminografia, essa a problemática poderia ser minimizada. Isto é, se os textos da disciplina fossem analisados, os termos fossem reconhecidos e coletados, seria possível propor a elaboração de um glossário especializado na disciplina Linguística de *Corpus*. Desta forma, o objetivo deste trabalho é apresentar uma proposta de elaboração de um glossário que contemple termos utilizados na disciplina Linguística de *Corpus* do curso Línguas Estrangeiras Aplicadas ao MSI.

Para levar tal projeto a cabo, este trabalho apresenta Referencial Teórico, Objetivos, Justificativa, Problematização e uma Metodologia detalhada que procura expor as áreas de Terminologia, Terminografia e Linguística de *Corpus*, relacionadas em vista da elaboração de glossários e dicionários especializados. Após a exposição da Metodologia, como Resultado Final, será apresentada a proposta de elaboração de um glossário (em português com equivalências em inglês) da disciplina Linguística de *Corpus* do LEA-MSI.

2. REFERENCIAL TEÓRICO

Pretende-se apresentar uma proposta para a elaboração de um glossário, em português com equivalências em inglês, da disciplina Linguística de Corpus do LEA-MSI por meio de aportes oriundos de três grandes áreas: a Terminologia, a Terminografia e a Linguística de Corpus, que serão apresentadas a seguir:

2.1 TERMINOLOGIA

A primeira compreensão de “Terminologia” a ser apresentada neste projeto é como disciplina, na qual se trata do estudo de termos técnico-científicos, podendo ser aplicada em diversos estudos da linguagem, como Tradução, ensino de línguas estrangeiras, compilação de termos e elaboração de dicionários especializados.

> “Com o objetivo prático de estudar, coletar e elucidar os usos dos termos das diversas áreas de conhecimento humano, a Terminologia é uma ciência que busca aprofundar o entendimento dos termos nas áreas de especialidade e precisar as relações comunicativas entre diferentes áreas de conhecimento a fim de evitar ambiguidades.” (MARINI, p. 37, 2013).

Sendo o termo o objeto de estudo da Terminologia, é necessário conceitua-lo: “A diferença entre termo e palavra se observa, fundamentalmente, na situação comunicativa” (ALMEIDA, p. 88, 2006.). Desta maneira, entende-se por termos, unidades lexicais pertencentes à língua geral (português, francês, inglês, alemão, esperanto...), mas que se diferenciam do léxico (acervo de palavras) comum porque ocorrem de forma natural e, em maior frequência, em específicos cenários comunicativos, podendo estes ser de caráter científico, técnico ou teórico. E, também, porque são realmente significativos quando inseridos no contexto específico a que pertencem, isto é, em sua linguagem de especialidade.

> “En efecto, la peculiaridad más notable de la terminología, en contraste con el léxico común, consiste en que sirve para designar los conceptos propios de las disciplinas y actividades de especialidad. En consecuencia, los términos son conocidos fundamentalmente por los especialistas de cada una de esas materias, y aparecen con una frecuencia muy elevada en los

> documentos especializados de cada disciplina.” (CABRÉ, p. 169, 1993.)

A linguagem de especialidade, ou também linguagem para fins específicos (*Language for Specific Purposes*), corresponde ao conceito de uma comunicação especializada constituída por termos técnicos que busca ser precisa e objetiva, sendo um “discurso funcional e um subsistema compreendido no sistema total da língua, como tal recorrendo apenas parcelarmente ao material lexical, sintático e semântico que a língua disponibiliza” (GIL, p. 115, 2003).

Em relação aos termos, temos a segunda acepção de “terminologia”, caracterizada por ser um conjunto de componentes lexicais de uma área específica. Em outras palavras, além de disciplina, a “terminologia” pode ser o vocabulário especializado de uma linguagem de especialidade (KRIEGER & FINATTO, p. 13, 2004). Para diferenciar as duas acepções apresentadas emprega-se, “Terminologia”, com T maiúsculo, em referência à disciplina, e “terminologia”, grafada com t minúsculo, para os termos presentes em distintos ofícios e ciências.

2.1.1 TEORIAS DA TERMINOLOGIA

Cada projeto terminológico possui objetivos distintos e, por isso, são exigidas metodologias apropriadas e específicas para alcança-los. A Terminologia possui duas grandes vertentes que, consequentemente, apresentam procedimentos metodológicos de pesquisa diferentes, sendo elas: a Teoria Geral da Terminologia (TGT) e a Teoria Comunicativa da Terminologia (TCT).

Formada pelas bases clássicas da disciplina, a Teoria Geral da Terminologia foi estabelecida por Eugen Wüster nos anos 60, e defende que a terminologia deve ser um instrumento padronizador da comunicação especializada. Isto é, por meio de princípios e métodos terminológicos normativos[3], deve-se eliminar a ambiguidade de linguagens técnicas e assim buscar um perfeito sistema de diálogo técnico-científico (CABRÉ, p. 27, 1993). O caráter normatizador da TGT inspirou, inclusive, a criação

[3] Aqui propõe-se uma distinção entre os termos “normatizar” e “normalizar”. Na visão de Marini (2013, p.59), o primeiro refere-se a um processo de prescrição de termos, e o segundo ao reconhecimento de termos como pertencentes a uma determinada linguagem de especialidade.

de instituições propagadoras de diretrizes com o intuito de unificar os métodos de trabalho da Terminologia, como a ISO (*International Standard Organization*). Ainda, na metodologia da TGT, segue-se o processo onomasiológico, no qual se prioriza o conceito, atuando na direção 'conceito → termo'. Isto se dá, pois a TGT compreende o termo como uma simples atribuição de um nome a um conceito, ou etiqueta denominativa, segundo Krieger e Finatto (2004).

Os fundamentos prescritivos, normatizadores e onomasiológicos da teoria inspirada por Wüster têm levado a TGT a ser contestada nas últimas décadas. A Teoria Comunicativa da Terminologia (TCT), idealizada por Maria Teresa Cabré, contrapõe a TGT, primeiramente, por ser uma teoria descritiva e não normativa. Diferenciando-se da TGT, a TCT reforça uma perspectiva linguística da Terminologia ao seguir o processo semasiológico, partindo do termo e o caracterizando funcional e semanticamente, de forma que o objeto central se torne o termo, e o conceito não seja desprezado (ALMEIDA, p. 86, 2006). Essa visão direcionada: 'termo → conceito' torna possível, segundo a TCT, a variação conceitual[4] e denominativa[5] nas linguagens de especialidade, indo contra o princípio prescritivo da TGT. Ainda, na TCT, a compreensão de termo é revista, pois acredita-se que "uma unidade lexical pode assumir o caráter de termo em função de seu uso em um contexto e situação determinados" (KRIEGER & FINATTO, p. 35, 2004). Em outras palavras, perante o princípio comunicativo, uma palavra pode se tornar parte de uma terminologia e ter seu conteúdo (significado) alterado conforme o cenário comunicativo em que se apresenta.

Considerando as duas teorias apresentadas, a proposta do presente TCC se adequa à Teoria Comunicativa da Terminologia pelo seu aspecto descritivo, e não de padronização dos termos. E, também, porque, neste trabalho, pretende-se observar os termos em seu ambiente natural de ocorrência (metodologia da TCT), isto é, nos próprios textos da disciplina, além de seguir a direção semasiológica de pesquisa.

[4] Chama-se essa variação conceitual de polissemia, na qual um termo pode ter dois ou mais significados diferentes (MARINI, p. 41, 2013).
[5] Chama-se essa variação denominativa de sinonímia, na qual dois ou mais termos podem denominar um determinado conceito (MARINI, p. 41, 2013).

2.2 TERMINOGRAFIA

Enquanto a Terminologia pode ser categorizada como uma matéria teórica e também metodológica, a Terminografia se apresenta como uma vertente aplicada da Terminologia (CABRÉ, p. 263, 1993). E, tendo a elaboração de dicionários técnico-científicos e glossários como um de seus objetivos específicos, a Terminografia consiste em "identificar a terminologia da área, redigir as definições dos termos e ainda organizar esses dados de acordo com o produto visado" (MÜLLER & RABELLO, p. 31, 2013). Portanto, ela se torna subsidiária deste trabalho por se tratar de uma proposta para a elaboração de um glossário da disciplina Linguística de *Corpus* do LEA-MSI.

Considerando esta proposta, antes de tratar sobre glossários, é necessário diferenciar Terminografia de Lexicografia, pois são áreas que podem ser confundidas por terem, inicialmente, objetivos semelhantes. A Lexicografia, entendida como arte ou técnica de produção de dicionários (KRIEGER & FINATTO, p. 50, 2004), busca registrar o léxico geral de um ou mais idiomas. E, embora os dicionários gerais possam apresentar léxico oriundo de linguagens de especialidade, os termos especializados não são o foco na aplicação lexicográfica. Já a Terminografia é uma disciplina que parte da Terminologia e, portanto, é aplicada na geração de glossários e dicionários especializados, possuindo então metodologia e objetivos específicos diferenciados da Lexicografia por serem voltados à "coleta, sistematização e apresentação dos termos de uma determinada área do saber ou atividade humana" (CABRÉ, p. 263, 1993). A partir desta diferenciação entre Lexicografia e Terminografia, segue-se para um dos produtos da Terminografia: o glossário.

2.2.1 GLOSSÁRIO

> "Ao lado de fundamentos teóricos, há também uma dimensão aplicada [da Terminologia], refletida na produção de glossários e dicionários técnicos, entre outros instrumentos de organização formal das terminologias." (KRIEGER & FINATTO, p. 13, 2004).

Para este projeto, propõe-se que um glossário seja entendido como

“repertório de unidades lexicais de uma especialidade com suas respectivas definições ou outras especificações sobre seus sentidos”. (KRIEGER & FINATTO, p. 51, 2004). Esse tipo de repertório terminográfico diferencia-se dos dicionários especializados principalmente pela não pretensão de abarcar todo o léxico de uma linguagem de especialidade, e sim os termos mais representativos[6] dela, possuindo assim uma grande distinção também pelo tamanho da obra.

Sendo uma obra terminográfica, um glossário de termos apresentará suas entradas, isto é, os termos, da mesma maneira em que aparecem na linguagem de especialidade a que pertencem. Desta forma as entradas podem ser formadas por termos que sejam substantivos, adjetivos, verbos, sintagmas terminológicos ou fraseologias (KRIEGER & FINATTO, p. 129, 2004).

Para levar a cabo a proposta deste trabalho, é necessário, também, apresentar dois termos muito recorrentes nos estudos terminográficos: a macroestrutura e a microestrutura. A orientação na qual são apresentadas as entradas no glossário pode ser chamada de macroestrutura, e, segundo Cabré (1993, p. 329), essa apresentação pode ser feita por ordem alfabética ou temática. Já a orientação que configura as informações dentro do verbete, segundo Marini (2013, p. 78), pode ser chamada de microestrutura. Tanto a macroestrutura quanto a microestrutura devem considerar os objetivos do trabalho e as necessidades dos usuários levando em conta o uso a que se propõe a obra terminográfica (KRIEGER & FINATTO, p. 130, 2004).

Portanto, com base na informatização que vêm se desenvolvendo nas últimas décadas, a Terminografia passa a contar com subsídios advindos de outras áreas para a obtenção de terminologias e dados para suas pesquisas. Nesse sentido, a Terminologia e a Terminografia recebem apoio da Linguística de *Corpus* e suas ferramentas, como concordanciadores, listadores de palavras e colocados, a fim de, em menos tempo, manipular e coletar dados de forma mais precisa.

[6] Representatividade, quando em relação a termos, é aplicada por Krieger e Finatto (2004, p. 129) a um termo que é, efetivamente, significativo de uma área do saber e “diz” algo para as pessoas que estão inseridas na linguagem de especialidade em questão.

2.3 LINGUÍSTICA DE *CORPUS*

Corpus é um termo oriundo do latim, e pode significar “conjunto de uma obra”. No âmbito da Linguística de *Corpus*, *corpus* é, na maioria das vezes, visto como um conjunto de textos que são armazenados e compilados considerando critérios pré-estabelecidos pelo pesquisador, visando uma pesquisa linguística. Em outras palavras:

> “As definições de *corpus*, via de regra, ressaltam que um *corpus* é uma coletânea de textos em linguagem natural, escritos ou falados, geralmente armazenados de forma organizada e informada, além de serem digitalizados a fim de que possam ser lidos por computador. Ainda que a definição de autores diversos ressalte uma ou outra característica, um *corpus* para a LC espelhará esses fatores principais” (SHEPHERD, p. 151, 2009).

Já a Linguística de *Corpus* (doravante LC) é definida por Tony McEnery e Andrew Wilson (2001, p. 1) como “estudo da linguagem baseado em exemplos de uso retirados da ‘vida real’”[7]. Nesta direção, a Linguística de *Corpus* tem como objetivo proporcionar dados de línguas naturais (linguagem concebida naturalmente pelo ser humano) em razão de um determinado estudo, seguindo a abordagem baseada em *corpus*, ou a abordagem direcionada pelo *corpus*.

A abordagem baseada em corpus seria uma pesquisa empírica na qual o corpus é utilizado para gerar exemplos capazes de ilustrar e experimentar teorias linguísticas já existentes. E a abordagem direcionada pelo corpus seria uma pesquisa na qual são geradas hipóteses no âmbito lexical e gramatical sobre o corpus analisado, conforme os dados sejam apresentados (SHEPHERD, p. 153 e 154, 2009). A abordagem direcionada (ou dirigida) pelo corpus, diferentemente da abordagem baseada em corpus, não visa fazer testes e exemplificar teorias já existentes, mas observar e analisar padrões e frequências lexicais. O que faz com que este presente trabalho se aproxime mais da abordagem direcionada pelo *corpus*.

A Linguística de *Corpus* conta com programas de computador para manipular

[7] Tradução de: “Study of language based on examples of ‘real life’ language use” (MCENERY & WILSON, p. 1, 2001).

os *corpora* e obter os mais diversos dados. Pode-se citar o *AntConc* e o *Wordsmith Tools* como exemplos dentre os muitos *softwares* desenvolvidos em vista à LC. Estes programas geralmente dispõem de ferramentas como: listador de palavras, concordanciador, lista de palavras-chave, colocados, etc. (i) O listador de palavras proporciona ao pesquisador uma listagem de todas as palavras do corpus, apresentando todas as palavras seguindo uma ordem de frequência, além de direcionar o pesquisador que clicar em uma palavra à sua respectiva ocorrência na ferramenta concordanciador; (ii) o concordanciador é uma ferramenta desenvolvida para que termos possam ser localizados em um *corpus* junto ao contexto em que estejam inseridos, gerando resultados chamados de “concordâncias”; (iii) a lista de palavras-chave faz um levantamento das palavras-chave mais frequentes nos *corpora*. Já a ferramenta colocados apresenta uma lista com a tendência de determinadas palavras ocorrerem juntas, em uma associação sintagmática de itens lexicais. Também existem outras ferramentas nos *softwares* voltados à Linguística de *Corpus* que não são apresentadas aqui porque não se fazem pertinentes ao presente trabalho.

Por fim, a Linguística de Corpus ainda é discutida na academia em respeito ao seu caráter de área de pesquisa ou de metodologia. Neste trabalho, acima dessa discussão, a LC se faz objeto e também parte da metodologia em apoio à Terminologia e a Terminografia. A categorização de objeto ocorre pela disciplina do LEA-MSI, Linguística de *Corpus*, ser o foco desta pesquisa, sendo, assim, o tema do glossário proposto. A LC também é determinada como apoio metodológico deste trabalho por auxiliar a pesquisa terminológica na obtenção de dados, e a pesquisa terminográfica, por tornar possível a manipulação dos dados e termos, auxiliando nas definições que constarão no glossário.

3. OBJETIVO GERAL

O objetivo geral deste trabalho é apresentar uma proposta de elaboração de um glossário em português com equivalências em inglês da disciplina Linguística de *Corpus* do curso Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da

Informação (LEA-MSI), de forma que o produto final possa destacar a Terminologia como eixo temático do curso supracitado, a Terminografia como área correlata a esse eixo, a Linguística de *Corpus* como disciplina e metodologia aplicada à Terminografia, e propor um instrumento de apoio didático aos estudantes na disciplina de LC.

3.1 OBJETIVOS ESPECÍFICOS

- Apresentar os procedimentos teórico-metodológicos escolhidos no desenvolvimento da proposta de elaboração do glossário;
- Montar dois *corpora*, um em português e outro em inglês, com os textos utilizados na disciplina Linguística de *Corpus* do LEA-MSI no primeiro semestre de 2014;
- Analisar as terminologias encontradas nos textos com o auxílio do *Freeware Concordance Program AntConc*;
- Selecionar os termos para o glossário considerando critérios de frequência e de indicação de professores do LEA-MSI;
- Criar fichas terminológicas para os termos então selecionados, separar as informações dos termos indispensáveis ao público-alvo e utilizá-las para constituir os verbetes do glossário;

4. JUSTIFICATIVA

Em função das grandes transformações verificadas no mundo nas últimas décadas, como os intercâmbios linguísticos, políticos, culturais e sociais, a Universidade de Brasília (UnB) sentiu a necessidade de implementar uma nova habilitação no Instituto de Letras, em 2010: o curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação. O LEA-MSI se tornou um

curso inovador e promissor pelos seus três eixos temáticos: Audiovisual, Multilinguismo e Terminologia.

Curiosamente, foi observado um grande interesse dos estudantes das primeiras turmas do LEA pelo eixo terminológico, dado verificado pelo considerável número percentual de propostas de TCC neste âmbito no ano de 2014[8]. Sendo assim, faz-se essencial a apresentação da Terminografia e da Linguística de *Corpus* para os estudantes do eixo de Terminologia. Ambas as ciências são interessantes, pois a primeira, Terminografia, torna-se relevante para quem pretende trabalhar com dicionários técnicos e glossários, podendo gerar, assim, novas perspectivas profissionais aos bacharéis em LEA-MSI; a segunda, Linguística de *Corpus,* foi selecionada como objeto deste trabalho por alguns motivos, a saber: (i) por ser uma disciplina que busca aprimorar os estudos de linguagem, tendo aplicação nas mais diversas áreas de pesquisa, como a tradução e ensino de línguas estrangeiras, por exemplo. (ii) porque além de ser objeto do trabalho, é, também, subsidiária da metodologia empregada, oferecendo aportes metodológicos a uma relação já comum aos profissionais da Terminologia, que já não imaginam uma pesquisa terminológica que não seja baseada em *corpus*. (ALMEIDA, & CORREIA, p. 91, 2008). (iii) por oferecer ferramentas computadorizadas, como os *softwares* concordanciadores, para alcançar os objetivos da Terminografia.

Voltando-se ao tema deste projeto, a proposta para a elaboração de um glossário surgiu pela dificuldade pessoal encontrada ao ler os primeiros textos da ementa da disciplina de Linguística de *Corpus*. Frequentemente, a leitura tinha que ser interrompida para que se fizesse uma busca aos termos específicos, necessitando, muitas vezes, do auxílio de textos externos à ementa para a compreensão dos termos.

Considerando todos esses levantamentos, este trabalho é impulsionado pelas seguintes justificativas: (i) é necessário facilitar o processo de aprendizagem do estudante da disciplina Linguística de *Corpus* de modo a evitar que a grande quantidade de termos técnicos que ela exige torne-se um obstáculo ao aluno; (ii) um glossário, quando finalizado, se torna uma ferramenta de busca rápida e prática, o

[8] No 1º semestre de 2014, considerando o total de duas propostas de TCC apresentadas ao LEA, uma era diretamente ligada à Terminologia; no 2º semestre de 2014, considerando o total de nove propostas de TCC apresentadas ao LEA, cinco eram diretamente ligadas à Terminologia.

que possibilita consultas mais direcionadas e objetivas; (iii) por fim, porque se pretende com esta proposta de elaboração de um glossário, impulsionar o interesse dos estudantes pela criação de projetos terminológicos e terminográficos no curso, para que no futuro tenhamos, talvez, um dicionário especializado no LEA-MSI, o que trará uma maior consolidação para a graduação e facilidade no processo de aprendizagem dos estudantes.

5. PROBLEMATIZAÇÃO

Considerando os apontamentos e as justificativas citadas na seção anterior, surge uma indagação: Como facilitar a compreensão dos termos utilizados na disciplina Linguística de *Corpus* aos estudantes do LEA-MSI, de modo a, também, apresentar a Terminologia e a Terminografia?

6. METODOLOGIA

A metodologia da produção de dicionários especializados e glossários, segundo Krieger e Finatto (2004), pode ser dividida em etapas, das quais, para a proposta de um glossário, adota-se as seguintes: Planejamento do Trabalho; Reconhecimento Terminológico e Preparação Inicial; Listagem de Termos; Registro de Dados; Fase Final.

6.1 PLANEJAMENTO DO TRABALHO

Nesta primeira etapa, foram esquematizadas questões pertinentes a todo o processo de elaboração do glossário, a saber:

Para levar a cabo a elaboração de um repertório terminográfico é necessário compor um *corpus* para análise terminológica, pois assim torna-se possível a observância de aspectos morfológicos, sintáticos e discursivos presentes no *corpus*, além de outros fatores que não são percebidos pela intuição (ALMEIDA, p. 88, 2006). Portanto, para compor o *corpus* deste trabalho, partiu-se da ementa disponibilizada no 1º semestre de 2014, na disciplina Linguística de *Corpus*, na qual se obteve a lista dos textos que deveriam ser compilados.

Como os textos da disciplina estavam divididos em duas línguas, inglesa e portuguesa, decidiu-se fazer dois *corpora*. Um para os textos em inglês e outro para os textos em português. Para analisar esses *corpora*, considerou-se utilizar o concordanciador *AntConc*.

O *AntConc*[9] foi o concordanciador selecionado para esta proposta por ser (i) disponibilizado gratuitamente na internet e ser executado nos principais sistemas operacionais (Microsoft, Linux...); (ii) ser fácil de ser manuseado, não exigindo o auxílio de um especialista para a utilização de suas ferramentas; (iii) apresentar configuração para UTF-8, que é um tipo de codificação que permite que o programa leia todos os caracteres presentes na língua portuguesa e inglesa; (iv) ser um software que revela eficiência e bom desempenho para o projeto proposto, podendo suprir todas as necessidades da metodologia de elaboração de um glossário.

Após planejar questões relativas ao *corpus*, partiu-se para as características de composição do glossário, elevando-se aspectos sobre a macroestrutura e microestrutura.

Segundo Almeida (2006, p. 98), ao pensar na macroestrutura de uma obra que segue a metodologia da Teoria Comunicativa da Terminologia, deve-se pensar no caráter comunicacional que a obra deve projetar. Desta forma, é necessário considerar a realidade dos futuros usuários do glossário. Visto que uma divisão temática do glossário, por exemplo, é indicada para um público que já conheça bem o tema do glossário, uma divisão sistemática pode não ser indicada para aprendizes, até porque “muitas pessoas que consultam dicionários sequer conhecem uma outra

[9] Segundo Anthony (2014), o *Freeware Concordance Program AntConc* é um software gratuito para a realização de pesquisas de Linguística de Corpus desenvolvido pelo Dr. Laurence Anthony, Professor da Faculdade de Ciência e Engenharia na Universidade de Waseda, no Japão.

ordem de organização que não seja a alfabética" (ALMEIDA, p. 98, 2006). Como o glossário é voltado aos estudantes da disciplina Linguística de *Corpus*, pode-se afirmar que o público-alvo do glossário não é especialista nessa área. Portanto, decidiu-se que a macroestrutura seria organizada em ordem alfabética para propiciar ao público-alvo maior praticidade nos momentos de pesquisa no glossário.

Considerando que se propõe nesse trabalho a elaboração de um glossário em português com equivalências em inglês, planejou-se elaborar um índice em ordem alfabética com todas as equivalências dos termos, pois assim o público-alvo pode encontrar os termos partindo de suas equivalências em inglês.

Optou-se, também, por adotar uma macroestrutura simples, sem sub-entradas. Isto é, para cada termo seria formulado um novo verbete. Por exemplo, sendo o termo "*corpus*" uma entrada, termos que poderiam ser sub-entradas, como "*corpus* de referência", configurarão outra entrada, a fim de deixar a pesquisa no glossário mais prática aos estudantes.

A microestrutura do glossário, sendo a estrutura interna dos verbetes, seria formada com informações que se julgaram pertinentes ao usuário final do glossário, como: definição do termo, seu equivalente em língua estrangeira, e a sigla do termo (se houver).

Ainda dentro da microestrutura, planejou-se a aplicação de remissivas. As remissivas são campos presentes nos verbetes que podem direcionar o leitor a outros verbetes, e "representam uma opção de ampliar o uso pragmático do instrumento uma vez que auxilia o leitor a recuperar, de forma rápida e objetiva outras informações sobre o tema tratado" (MARINI, p. 81, 2013). Desta forma, decidiu-se aplicar no glossário remissivas que fazem referência a termos que aparecem dentro da definição dos verbetes, e remissivas que fazem referências a termos conexos, possibilitando ao usuário ser direcionado a termos que possuem algum tipo de ligação ao termo pesquisado, de forma a complementar o conhecimento do usuário com outro verbete relacionado à definição que foi lida. Também se optou por elaborar remissivas para os termos que possuem sigla, de forma a corresponder por meio de remissivas tanto o termo escrito por extenso tanto a sigla do termo.

No modo de armazenagem dos dados, foi estabelecido que seriam criadas fichas terminológicas para cada termo que fosse selecionado. Essas fichas possuem informações imprescindíveis para o trabalho terminográfico, porque nelas se registram todos os dados necessários para a composição dos verbetes.

Também foram planejadas as características de definição dos termos. Conforme já explanado no Referencial Teórico, a Teoria Comunicativa da Terminologia afirma que o termo deve ser observado dentro do contexto temático em que ele é empregado (ALMEIDA, p. 99, 2006). Desta forma, as definições seriam obtidas utilizando os próprios textos dos *corpora*, garantindo que a definição fosse um reflexo do contexto em que o termo está inserido. Além disso, quando necessário, outras obras externas aos *corpora* seriam consultadas a fim de tornar a definição mais precisa.

Por fim, para assegurar que a microestrutura e a macroestrutura do glossário, bem como o índice de equivalências em inglês fossem entendidos por todos os usuários, planejou-se elaborar um texto introdutório ao glossário apresentando informações que colaborem com o público-alvo.

6.2 RECONHECIMENTO TERMINOLÓGICO E PREPARAÇÃO INICIAL

> “O reconhecimento terminológico, frisamos, está intrinsecamente relacionado ao reconhecimento de textos técnicos ou científicos e à identificação de tipos textuais, sejam eles mais ou menos “especializados” ou mais ou menos terminologicamente densos” (KRIEGER & FINATTO, p. 129, 2004).

Nesta etapa de reconhecimento terminológico e preparação inicial, alguns dos princípios básicos da Terminografia foram observados, como: atender às necessidades de um público-alvo (no caso, os estudantes da disciplina Linguística de *Corpus*) e a utilização de dados confiáveis.

Neste momento, foi fundamental analisar se os textos que seriam utilizados no *corpus* possuíam termos e contextos realmente representativos para a disciplina Linguística de *Corpus* do LEA-MSI, a fim de atender as necessidades dos usuários.

Por isso, sendo o glossário a respeito da disciplina, foram selecionados os textos listados na sua própria ementa, o que, no âmbito de utilização de dados confiáveis, atribuiu o caráter de confiabilidade ao *corpus*.

6.2.1 COMPILAÇÃO DO *CORPUS*

Para analisar um *corpus* utilizando o *AntConc* é necessário que os textos estejam no formato *.txt*. Portanto, para prosseguir com a compilação dos *corpora* deste trabalho, fez-se a conversão dos textos em inglês e português de *.pdf* para *.txt* por meio de um programa de OCR (*Optical Character Recognition*), responsável por reconhecer os caracteres de uma imagem e transformá-los em texto.

Após fazer a conversão dos textos, compilou-se primeiramente o *corpus* com textos em português. Assim, obteve-se um *corpus* com 6.396 palavras (*word types*) e 31.919 itens (*tokens*), como pode ser observado no Anexo 1, da seção *Anexos.* Depois, partiu-se para os textos em inglês e o *corpus* compilado gerou um resultado de 4.857 palavras (*word types*) e 31.592 itens (*tokens*), como pode ser observado no Anexo 2, da seção *Anexos*.

6.2.2 SELEÇÃO DOS TERMOS

Para selecionar os termos que fariam parte do glossário, critérios geralmente adotados em obras terminográficas foram observados, a saber: frequência de ocorrência no *corpus* e, também, sugestões de especialistas (alguns professores do eixo terminológico do LEA que já lecionaram a disciplina Linguística de *Corpus*).

Segundo Almeida (2006, p. 89), há três métodos de extração de termos em *corpus*: método estatístico, linguístico e híbrido. O método estatístico de seleção de termos trabalha com o critério de frequência. Esse método de análise das frequências de ocorrências no *corpus* elenca os candidatos a termos mais frequentes em um *corpus*, possibilitando que o profissional da Terminografia

selecione os termos que mais ocorrem para a formulação de um glossário ou dicionário especializado. Entretanto, esse método também é criticado por desconsiderar termos que podem ser importantes a uma linguagem de especialidade simplesmente por estes não terem grande frequência de ocorrência, podendo ser transpostos por palavras que não possuem o papel de termo dentro de um determinado *corpus*, mas que possuem elevada frequência de ocorrência (LAGUNA, p. 22, 2014).

A segunda abordagem de extração de termos é pelo método linguístico. Conforme apresentado por Laguna (2014, p. 27 e 28), o método linguístico procura extrair termos analisando a morfologia das palavras (a presença de morfemas greco-latinos nos termos, por exemplo)[10], mas também por meio de análises sintáticas (observando a estrutura das sentenças), semânticas, e morfossintáticas (no qual são observadas as categorias gramaticais dos candidatos a termos). Ainda segundo Laguna (2014), a abordagem linguística possui alguns problemas, a saber: (i) depender de conhecimento linguístico; (ii) depender de ferramentas etiquetadoras, que, frequentemente, geram erros, atrapalhando o processo de obtenção dos termos; (iii) se dispensadas as ferramentas etiquetadoras, o trabalho de obtenção dos termos se torna custoso e lento.

A terceira e última abordagem para a extração dos termos de um *corpus* é a pelo método híbrido. O método híbrido diz respeito à utilização tanto do método estatístico quanto do método linguístico, de forma que um complete o outro a fim de aprimorar a extração dos termos.

Neste trabalho, o método estatístico foi selecionado, pois se trata de uma proposta para a elaboração de um glossário. Desta forma, não há a necessidade de elencar toda a terminologia presente nos *corpora* analisados, e sim alguns poucos, porém importantes, termos à disciplina Linguística de *Corpus*. Assim sendo, os termos mais frequentes já tornam possível a apresentação da proposta para a elaboração de um glossário da disciplina de LC.

[10] Esse método de obtenção de termos se adapta à Teoria Geral da Terminologia, que defende uma terminologia normatizadora, e que acredita no emprego de morfemas greco-latinos para denominar conceitos e ferramentas.

Como este trabalho se trata da apresentação de uma proposta para a elaboração de um glossário, decidiu-se que no produto final seriam apresentados dez termos. Sendo assim, esses dez termos deviam ser os mais importantes para a disciplina Linguística de *Corpus* do LEA-MSI. Para escolhê-los, foram realizados os seguintes procedimentos:

Após a compilação do *corpus*, foi gerada uma lista de frequência utilizando a ferramenta "*Word List*" do *AntConc* para selecionar os termos. Partindo do *corpus* dos textos em português[11], buscou-se os quinze candidatos a termos mais frequentes no *corpus*[12]. Este número foi selecionado para que fosse possível, posteriormente, selecionar entre os 15 candidatos a termos apenas dez termos considerando a opinião de especialistas.

É importante informar que, os repertórios terminográficos (dicionários e glossários especializados) possuem grande aceitação a entradas com mais de uma palavra, ou seja, sintagmas, pois estes são frequentes nas linguagens de especialidade. Portanto, os quinze termos selecionados como mais frequentes no *corpus* de textos em português, foram analisados um por um a fim de verificar se estes faziam parte de relações sintagmáticas com outros itens lexicais. Essa análise foi feita por meio da ferramenta concordanciador (*concordance*)[13], na qual todas as concordâncias eram analisadas, e, assim que se detectava a presença de relações sintagmáticas, utilizava-se a ferramenta colocados (*collocates*)[14] a fim de exibir todas as ocorrências do sintagma pesquisado no *corpus*. De tal modo, termos como "abordagem baseada em *corpus*" foram selecionados no projeto deste glossário, pois se reconheceu, tanto pelo critério de frequência, quanto pela indicação de professores, a importância de alguns desses termos sintagmáticos para a disciplina de LC.

Considerando que a opinião de especialistas em uma linguagem de especialidade é de suma importância em pesquisas terminológicas, foi solicitado o

[11] Decidiu-se direcionar a pesquisa deste trabalho a partir do *corpus* dos textos em português por dois motivos: (i) ser o maior *corpus* em comparação ao formado com os textos em inglês; (ii) pelo fato do glossário ser em português, não em inglês.
[12] A lista com os quinze candidatos a termos pode ser vista no Anexo 3.
[13] Ver Anexo 4.
[14] Ver Anexo 5.

auxilio de três professores do curso Línguas Estrangeiras Aplicadas ao MSI[15] que já lecionaram a disciplina Linguística de *Corpus*. Foi-lhes solicitado que discriminassem, no mínimo, cinco termos representativos da disciplina do curso, e, após isso, todos os termos foram compilados em uma lista[16]. Como o glossário é voltado a uma disciplina que eles já lecionaram e por terem tomado parte até mesmo na seleção dos textos da ementa (que formam os *corpora* deste trabalho), pode-se afirmar que os termos elencados pelos professores certamente serão trabalhados na disciplina Linguística de *Corpus*.

De tal modo, uma nova lista de candidatos a termo foi gerada. Nela foram listados os 15 candidatos a termo mais frequentes no *corpus* de textos em português. Em frente foi colocado um campo que representava a presença ou não do candidato a termo na lista gerada pelas sugestões dos professores. Os termos que foram apresentados nesta lista, mas que não apareciam na lista dos professores foram descartados. Já os termos que constavam nas duas listas foram selecionados por ordem de frequência. Os dez termos mais frequentes foram escolhidos para compor a nomenclatura da proposta de glossário deste trabalho.[17]

A obtenção das equivalências em inglês dos termos foi efetuada através de[18]: (i) pesquisas no *corpus* dos termos em inglês, para aqueles termos que, por ventura, possuíssem o mesmo radical tanto na língua inglesa, quanto na língua portuguesa. (ii) uma lista dos vinte termos em inglês mais frequentes[19], na qual foram analisados os termos juntos ao contexto em que apareciam no *corpus*, a fim de descobrir os conceitos equivalentes aos dos termos em português previamente selecionados. (iii) através do dicionário em inglês especializado em Linguística de *Corpus* elaborado por Baker, Hardie & McEnery (2006).

[15] Professor doutorando Thiago Pires; Professor doutorando Marcos Carneiro; e Professora mestranda Clarissa Marini.
[16] Ver Anexo 6.
[17] Ver Anexo 7.
[18] É importante salientar que não foi possível fazer um alinhamento dos corpora para a obtenção das equivalências, pois os textos da ementa da disciplina de LC foram disponibilizados somente em suas línguas de partida, não possuindo versão traduzida para fazer alinhamento.
[19] Ver Anexo 8.

6.3 REGISTRO DE DADOS EM FICHAS TERMINOLÓGICAS

Nesta etapa, foram elaboradas fichas terminológicas para cada termo selecionado para o glossário. “A ficha terminológica [...] pode ser definida como um registro completo e organizado de informações referentes a um dado termo”. (KRIEGER & FINATTO, p. 136, 2004). Ou seja, na ficha terminológica devem constar as informações mais importantes de cada termo, pois essas fichas são utilizadas para elaborar as entradas e definições de cada verbete de um repertório terminográfico.

Segundo Cabré (1993, p. 282), em uma ficha terminológica comumente constam os seguintes campos: identificação do termo, termo de entrada, fonte do termo, categoria gramatical, área temática, definição, fonte da definição, contexto, fonte do contexto, remissão a termos sinônimos, conceito da remissão, outros tipos de remissão, conceito de cada tipo de remissão, autor e redator da ficha, notas para informações não previstas, equivalências em outras línguas (com indicação da língua), e fonte de cada equivalência. Mesmo assim, para Cabré (1993), alguns desses campos podem ser omitidos e outros campos podem ser adicionados considerando os objetivos do glossário a ser elaborado. Considerando o planejamento feito anteriormente sobre a microestrutura dos verbetes para este projeto, elaborou-se uma ficha terminológica que pudesse subsidiar todas as informações a serem apresentadas na microestrutura do glossário. Desta forma, as fichas terminológicas deste trabalho foram formadas com os seguintes campos: entrada, marca gramatical, equivalência em inglês, definição do termo, fonte da definição, contexto do termo no *corpus*, fonte do contexto, remissivas aos termos conexos, remissivas às siglas, e notas (para informações não previstas)[20].

Entrada		Marca gramatical	
Equivalência(s) em inglês			

[20] Alguns exemplos de Fichas Terminológicas elaboradas neste trabalho podem ser encontrados no Anexo 9.

Definição do termo			
Fonte da definição			
Contexto do Termo no *corpus*			
Fonte do contexto			
Remissiva(s): Termo(s) conexo(s)		Remissiva: sigla	
Notas			

Exemplo de Ficha Terminológica utilizada neste trabalho

Segue abaixo a explanação de cada um dos campos presentes na Ficha Terminológica elaborada para este trabalho:

- Entrada – O termo escrito por inteiro conforme encontrado junto ao seu contexto no *corpus*.
- Marca gramatical – Masculino, feminino ou neutro, representada pelas letras M, F ou N, respectivamente.
- Equivalência em inglês – As traduções dos termos.
- Definição do termo – Constituída por informações extraídas dos conceitos encontrados no *corpus*. Quando as informações do *corpus* foram insuficientes para constituir uma completa definição, outras fontes foram consultadas, como artigos científicos e dicionários especializados, por exemplo. Também buscou-se apresentar as características mais essenciais do objeto que se define e seguir um padrão de definição para todos os itens do glossário.

- Fonte da definição – Para utilização em pesquisas posteriores sobre o termo em questão.
- Contexto do termo no *corpus* – Exemplo real de aplicação do termo junto ao seu contexto em um dos textos do *corpus*.
- Fonte do contexto – Referência ao exemplo retirado do *corpus*.
- Remissivas aos termos conexos – (Se houver) Neste campo serão inseridos termos conexos ao conteúdo da ficha terminológica por terem sido mencionados na mesma ficha, por terem o mesmo radical que a entrada, ou por ser considerado um termo importante para elucidar ainda mais a definição do termo apresentado.
- Remissiva as siglas do termo – (Se houver) especialmente para termos que são sintagmas e possam ter uma representação por sigla.
- Notas – (Se houver) informações pertinentes ao termo.

6.4 FASE FINAL

Esta etapa constitui a fase de finalização da proposta para a elaboração do glossário. Neste momento, fez-se um recorte de informações das fichas terminológicas produzidas anteriormente, selecionando as informações mais relevantes para o público-alvo, estudantes da disciplina Linguística de *Corpus* do LEA-MSI. Informações como: marca gramatical, fonte da definição e exemplo retirado do *corpus*, foram omitidas nas entradas para facilitar a busca dos estudantes, visto que glossários são utilizados nos momentos de consulta rápida e uma entrada muito extensa pode dispersar o estudante do que ele estava estudando.

Após a seleção das informações que constariam nos verbetes, as entradas do glossário foram organizadas por ordem alfabética. A introdução do glossário, responsável por apresentar todas as informações pertinentes à obra ao leitor foi redigida neste momento. Além da introdução, também foram redigidas a tabela de

abreviações utilizadas no glossário e a lista de equivalências em inglês de todos os termos da nomenclatura para facilitar buscas a partir dos termos em língua estrangeira. Por fim, a diagramação, isto é, a forma com que um glossário é apresentado ao público, foi elaborada seguindo o modelo utilizado na obra dicionarística de Baker, Hardie & McEnery (2006)[21], na qual se observa que: (i) a entrada aparece com letras minúsculas, à esquerda e em negrito. (ii) logo após vem a definição e, os termos externos a essa entrada que aparecem na definição são marcados em negrito para remeter o leitor a outras entradas. (iii) ao final do verbete são colocadas as remissivas com a equivalência em língua estrangeira e termos conexos marcados também em negrito a fim de direcionar o leitor a outras pesquisas no glossário. Após essas considerações finais, a proposta para a elaboração de um glossário da disciplina Linguística de *Corpus* do curso Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação foi finalizada.

[21] Ver Anexo 10.

7. RESULTADO FINAL

APRESENTAÇÃO DO GLOSSÁRIO

Caro leitor / Cara leitora,

a presente obra se trata de uma proposta para a elaboração de um glossário da disciplina Linguística de *Corpus* do curso de graduação Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação, da Universidade de Brasília. Por ser uma proposta, este glossário não busca elencar todos os termos da disciplina, mas, pelo menos, os dez mais frequentes a fim de exemplificar e propor um modelo de glossário, de modo que, no futuro, talvez tenhamos um glossário mais completo e que atenda a totalidade das necessidades dos estudantes da disciplina de Linguística de *Corpus* da graduação.

LISTA DE ABREVIATURAS

- pl. – Apresentação da forma plural do termo.
- Ing. – Remissiva à equivalência do termo em língua inglesa.
- V. – Remissiva a um ou mais termos conexos à entrada pesquisada.

ALGUMAS CONSIDERAÇÕES

Este glossário está organizado em ordem alfabética para proporcionar a você, caro usuário, maior praticidade. As entradas podem ser identificadas em letras minúsculas, à esquerda e em negrito. Logo após a entrada, é disponibilizada a definição do termo, e, se por ventura houver termos do glossário presentes na definição, estes serão marcados em negrito. Ao final do verbete são colocados o equivalente do termo em língua inglesa e termos conexos marcados também em negrito a fim de aprimorar a pesquisa.

LISTA DE EQUIVALÊNCIAS EM INGLÊS DOS TERMOS DESTE GLOSSÁRIO

A

***alignment* –** alinhamento

***annotation* –** anotação

C

***compilation* –** compilação

***corpora* –** corpora

***corpus* –** corpus

***corpus-based* –** abordagem baseada em corpus

***corpus-driven* –** abordagem direcionada pelo corpus

P

***parallel corpus* –** corpus paralelo

***processing* –** processamento

R

***representativeness* –** representatividade

Proposta de glossário para a Disciplina Linguística de *Corpus* do LEA-MSI

Marcos Vinícius da Silva[22]

A

abordagem baseada em *corpus* Metodologia empírica na qual o ***corpus*** é utilizado para gerar exemplos capazes de ilustrar e experimentar teorias linguísticas já existentes. Ing. ***corpus-based***. V. **abordagem direcionada pelo *corpus*** *(corpus-driven)*.

abordagem direcionada pelo *corpus* Metodologia na qual são geradas hipóteses no âmbito lexical e gramatical sobre o ***corpus*** analisado, conforme os dados sejam apresentados. A abordagem direcionada pelo *corpus*, diferentemente da **abordagem baseada em *corpus***, não visa fazer testes e exemplificar teorias já existentes, mas observar e analisar padrões e frequências lexicais. Ing. ***corpus-driven***. V. **abordagem baseada em *corpus*** *(corpus-based).*

alinhamento Encaixe das versões de um mesmo texto que foi traduzido para línguas distintas. A intenção ao alinhar esses textos é relacionar partes de uma das versões com as partes equivalentes da versão em outra língua, de forma a garantir que os parágrafos e sentenças equivalentes das duas versões possam ser visualizados paralelamente. Ing. ***alignment****.* V. ***corpus*** **paralelo** *(parallel corpus)***.**

anotação Processo no qual são adicionadas informações linguísticas como o gênero dos textos, por exemplo, a um ***corpus*** com o propósito de aprimorar o desenvolvimento da pesquisa. Ing. ***annotation****.* V. **etiquetagem** (*tagging*).

C

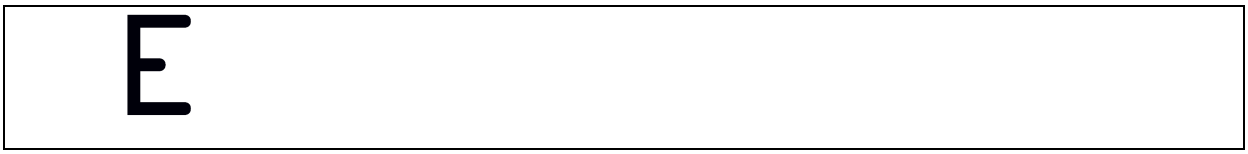
compilação Organização e armazenamento de textos em arquivos predeterminados a fim de elaborar um ***corpus*** ou ***corpora***. No processo de compilação, os textos são selecionados, limpos e convertidos para um formato adequado a ser utilizado em computador, para então ser armazenados segundo requisitos observados na Linguística de *Corpus*. Ing. ***compilation****.*

corpus pl. ***corpora*** Conjunto de textos em linguagem natural que podem ser convertidos para formato eletrônico para ser tratados por computador. São, então, armazenados e compilados considerando critérios pré-estabelecidos pelo pesquisador, visando à finalidade a que estão propostos. O termo vem do latim, e pode significar: “conjunto de uma obra”. No plural se torna *corpora*. Ing. ***corpus***. V. ***corpora****;* **abordagem**

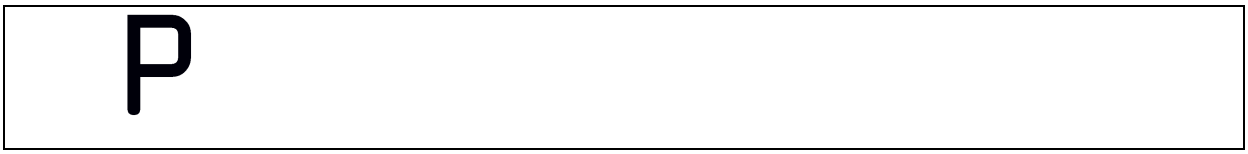
[22] Graduando em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação, pela Universidade de Brasília.

baseada em *corpus* *(corpus-based);* **abordagem direcionada pelo *corpus*** *(corpus-driven);* ***corpus* paralelo** *(parallel corpus).*

corpus* paralelo** Textos em determinada língua compilados junto a outras versões dos mesmos textos, isto é, textos iguais, porém traduzidos para outra língua diferente da primeira. Unidos, e por meio de um **alinhamento** dos textos, eles formam um ***corpus* paralelo** (ou ***corpora paralelos), que podem ser utilizados tanto para auxiliar tradutores no processo de tradução, quanto em pesquisas de linguística contrastiva, além de possuir outras aplicações. Ing. ***parallel corpus***. V. **alinhamento** (*alignment*).



etiquetagem Ato de marcar o ***corpus*** com etiquetas (códigos), que variam conforme o nível de informação, podendo indicar, por exemplo, as categorias gramaticais dos itens como verbo, advérbio, substantivo ou adjetivo, além de outras informações. Relaciona-se com o processo de **anotação** de *corpus*. Ing. ***tagging***. V. **anotação** (*annotation*).



processamento Processo no qual significativa quantidade de dados linguísticos, ***corpus***, ou ***corpora*** são computados a fim de tornar possível a automatização de determinados tipos de análises em favor de investigações linguísticas. Ing. ***processing***.

representatividade Conceito que, ao se referir a um ***corpus*** ou a ***corpora***, é utilizado para dizer que o *corpus* ou os *corpora* possuem volume de palavras e amostras suficientes e significativas para determinada investigação linguística, considerando o conteúdo do *corpus* e a sua relação com a temática da pesquisa. Ing. ***representativeness***.

REFERÊNCIAS BIBLIOGRÁFICAS DO GLOSSÁRIO

OBRAS EMPREGADAS PARA OBTENÇÃO DOS DADOS PARA A ELABORAÇÃO DO GLOSSÁRIO:

EVISON, J. . What are the basics of analysing a corpus? . In: *Routledge Handbook of Corpus Linguistics*. Oxford: Routledge, p. 122-135, 2010.

FERNÁNDEZ, P. M. . Aproximação à linguística de corpus como metodologia de base empírica. Compilação e anotação do Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial. In: *Revista AGÁLIA*, n. 89-90, Compostela: AGAL. 9-80, 2007.

GRANGER, S. . A bird's eye view of learner corpus research. In: Granger, S. ; Hung, J. ; Petch-Tyson, S. (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam e Philadelphia: Benjamins, p. 3-33, 2002.

KLÜBER, N. ; ASTON, G. . Using corpora in translation. In: Michael McCarthy; Anne O'Keeffe (eds). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, p. 501-516, 2010.

SARDINHA, T. B. . Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem – ReVEL*. v. 2, n. 3, ago. 2004.

SHEPHERD, T. M. G. . *O Estatuto da Linguistica de Corpus: Metodologia ou Área da Línguística*. Matraga (Rio de Janeiro), v. 16, p. 150-172, 2009.

SINCLAIR, J. . Developing Linguistic Corpora: a guide to good practice, Corpus and Text – basic principles [em linha]. 2004. Disponível em: <http://users.ox.ac.uk/~martinw/dlc/chapter1.htm>. Acesso em: 15 abr. 2014.

DICIONÁRIOS E GLOSSÁRIOS UTILIZADOS COMO APOIO NA ELABORAÇÃO DAS ENTRADAS:

BAKER, P. ; HARDIE, A. ; MCENERY, T. . *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press. 187 p. , 2006.

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

8. CRONOGRAMA DE ATIVIDADES

ATIVIDADES	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ
Escolha do tema e do orientador.	X							
Encontros com o orientador.	X	X	X	X	X	X	X	X
Pesquisa bibliográfica preliminar.	X	X	X					
Leituras e elaboração de fichamentos.	X	X	X					
Elaboração do projeto.	X	X						
Entrega do projeto de pesquisa.		X						
Revisão bibliográfica complementar.			X	X				
Coleta de dados e formulação de fichas terminológicas.			X	X	X			
Redação da metodologia do projeto.				X	X	X		
Distribuição das entradas em formato de glossário.				X	X	X		
Revisão e finalização do glossário.							X	X
Entrega oficial do								X

trabalho.								
Apresentação do trabalho em banca.								X

9. REFERÊNCIAS BIBLIOGRÁFICAS

9.1 OBRAS EMPREGADAS NOS *CORPORA* DESTE TRABALHO

EVISON, J. . What are the basics of analysing a corpus? . In: *Routledge Handbook of Corpus Linguistics*. Oxford: Routledge, p. 122-135, 2010.

FERNÁNDEZ, P. M. . Aproximação à linguística de corpus como metodologia de base empírica. Compilação e anotação do Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial. In: *Revista AGÁLIA*, n. 89-90, Compostela: AGAL. 9-80, 2007.

GRANGER, S. . A bird's eye view of learner corpus research. In: Granger, S. ; Hung, J. ; Petch-Tyson, S. (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam e Philadelphia: Benjamins, p. 3-33, 2002.

KLÜBER, N. ; ASTON, G. . Using corpora in translation. In: Michael McCarthy; Anne O'Keeffe (eds). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, p. 501-516, 2010.

SARDINHA, T. B. . Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem – ReVEL*. v. 2, n. 3, ago. 2004.

SHEPHERD, T. M. G. . *O Estatuto da Linguistica de Corpus: Metodologia ou Área da Línguística*. Matraga (Rio de Janeiro), v. 16, p. 150-172, 2009.

SINCLAIR, J. . Developing Linguistic Corpora: a guide to good practice, Corpus and Text – basic principles [em linha]. 2004. Disponível em: <http://users.ox.ac.uk/~martinw/dlc/chapter1.htm>. Acesso em: 15 abr. 2014.

9.2 DICIONÁRIOS E GLOSSÁRIOS UTILIZADOS COMO APOIO NA ELABORAÇÃO DAS ENTRADAS

BAKER, P. ; HARDIE, A. ; MCENERY, T. . *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press. 187 p. , 2006.

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

9.3 APOIO TEÓRICO GERAL

ALMEIDA, G. M. B. . A Teoria Comunicativa da Terminologia e a sua prática. São Paulo: *Alfa*, v. , p. 85-101, 2006.

_____. ; CORREIA, M. . Terminologia e corpus: relações, métodos e recursos. In: Stella E. O. Tagnin; Oto Araújo Vale. (Org.). *Avanços da Lingüística de Corpus no Brasil*. 1 ed. São Paulo: Humanitas/FFLCH/USP, v. 1, p. 63-93, 2008.

ANTHONY, L. . Help file version for AntConc 3.4.3. Tokyo: [s.n.], 2014. Disponível em: <http://www.laurenceanthony.net/software/antconc343/AntConc_readme.pdf>. Acesso em: 20 set. 2014.

BAKER, P. ; HARDIE, A. ; MCENERY, T. . *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press. 187 p. , 2006.

BOCORNY, A. E. P. ; VILLAVICENCIO, A. ; KILIAN, C. K. ; WILKENS, R. . Projeto glossário: a construção de um glossário online colaborativo com elementos multimeios para aprendizes da área de relações internacionais e seus resultados preliminares. *Revista Virtual de Estudos da Linguagem*, v. 9, p. 305-321, 2011.

CABRÉ, M. T. . *La terminología.* Teoría, metodología, aplicaciones. Barcelona: Editorial Antártida/Empúries, 469 p., 1993.

COSTA FILHO, J. E. . Elaboração de um glossário de termos utilizados na teoria da metáfora conceitual. In: XII Seminário de Teses em Andamento (SETA), 2007, Campinas. *Anais do SETA* (UNICAMP), v. 1. p. 327-332, 2007.

EVISON, J. . What are the basics of analysing a corpus? . In: *Routledge Handbook of Corpus Linguistics*. Oxford: Routledge, p. 122-135, 2010.

FERNÁNDEZ, P. M. . Aproximação à linguística de corpus como metodologia de base empírica. Compilação e anotação do Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial. In: *Revista AGÁLIA*, n. 89-90, Compostela: AGAL. 9-80, 2007.

GIL, I. T. M. . Algumas considerações sobre línguas de especialidade e seus processos lexicogênicos. *Máthesis*, v.12, p. 113-130, 2003.

GRANGER, S. . A bird's eye view of learner corpus research. In: Granger, S. ; Hung, J. ; Petch-Tyson, S. (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam e Philadelphia: Benjamins, p. 3-33, 2002.

KILIAN, C. K. ; BOCORNY, A. E. P. ; VILLAVICENCIO, A. ; WILKENS, R. . Critérios de seleção de termos utilizados na construção de glossários pedagógicos online baseados em corpus especializado. *Entrelinhas* (UNISINOS. Online), v. 6, p. 277-292, 2012.

KLÜBER, N. ; ASTON, G. . Using corpora in translation. In: Michael McCarthy; Anne O'Keeffe (eds). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, p. 501-516, 2010.

KRIEGER, M. G. . Sobre terminologia e seus objetos.. In: LIMA, M. dos Santos; RAMOS, P. C.. (Org.). *Terminologia e Ensino de Segunda Linha*. Porto Alegre: Gráfica UFRGS, v. , p. -.2001.

_____. ; FINATTO, M. J. B. . *Introdução à Terminologia*: teoria & prática. São Paulo: Contexto, v. 1. p.223, 2004.

LAGUNA, M. S. C. . *Extração automática de termos simples baseada em aprendizado de máquina*. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 159 p., 2014.

MARINI, S. . *Da tradução terminológica em glossário temático na área de saúde suplementar*. Brasília: Departamento de Línguas Estrangeiras e Tradução, Universidade de Brasília, p. 152, 2013.

MCENERY, T. ; WILSON, A. . The use of corpora in Language Studies. In: *Corpus Linguistics*: An introduction. 2.ed. Endiburgh: Endiburgh University Press, p.233, 2001.

MÜLLER, A. F. ; RABELLO, C. . A terminologia presente no interior das empresas: um estudo de caso sobre a variação terminológica em uma empresa de manutenção, reparo e revisão de aeronaves (MRO). *ReVEL*. v. 11, n. 21, 2013.

PERROTTI–GARCIA, A. J. . Glossários: Como elaborá-los? Como publicá-los?. [Em linha]. SBS - Livraria e Editora SBS - Ministrado em 27 de Outubro de 2003. Disponível em <http://www.sbs.com.br/e-talks/glossarios-como-elabora-los-como-publica-los/>. Acesso em: 17 ago. 2014.

_____. Reflexões sobre a qualidade de um bom glossário técnico: limites e limitações. *Confluências. Revista de tradução científica e técnica.* [S.l.] n. 01, p. 68-76, nov. 2004.

ROBERTS, J. Elaboration d'un dictionnaire. Association SIL. Cours Découvre Ta Langue, [S.l.], out. 1996.

ROCHA, C. F. . A elaboração de um glossário bilíngue da área de comércio tendo como subsídio a Linguística de Corpus. *Estudos Linguísticos* (São Paulo. 1978), v. 40, p. 1133-1144, 2011.

SARDINHA, T. B. . Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem – ReVEL*. v. 2, n. 3, ago. 2004.

SHEPHERD, T. M. G. . *O Estatuto da Linguistica de Corpus: Metodologia ou Área da Línguística*. Matraga (Rio de Janeiro), v. 16, p. 150-172, 2009.

_____. ; SARDINHA, T. B. . Panorama da Linguistica de Corpus. In: Tania M G Shepherd; Tony Berber Sardinha; Marcia Veirano. (Org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado das Letras, v. , p. -, 2012.

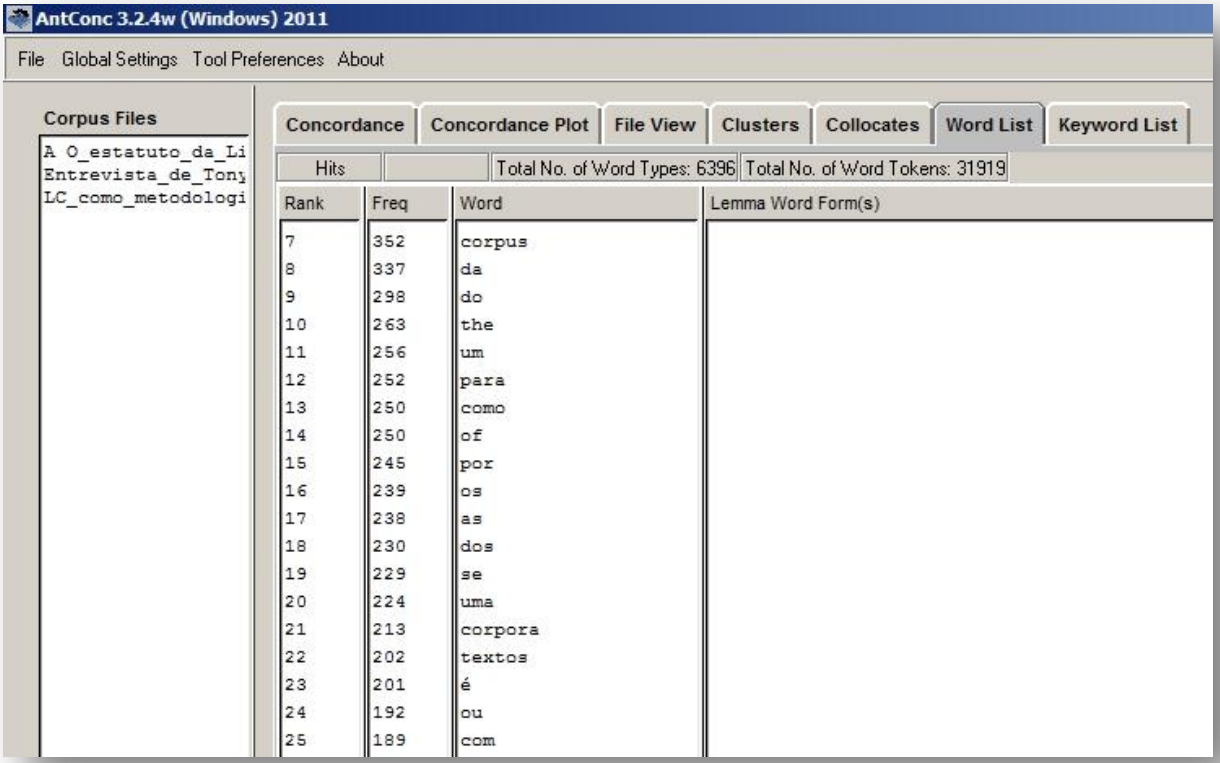
SILVEIRA, A. O. ; BERNARDON, M. . Construção de um glossário de termos técnicos para o curso de Engenharia Química. *Revista Expectativa* (Impresso), Toledo, v. 02, n.2, p. 62-67, 2003.

SINCLAIR, J. . Developing Linguistic Corpora: a guide to good practice, Corpus and Text – basic principles [em linha]. 2004. Disponível em: <http://users.ox.ac.uk/~martinw/dlc/chapter1.htm>. Acesso em: 15 abr. 2014.

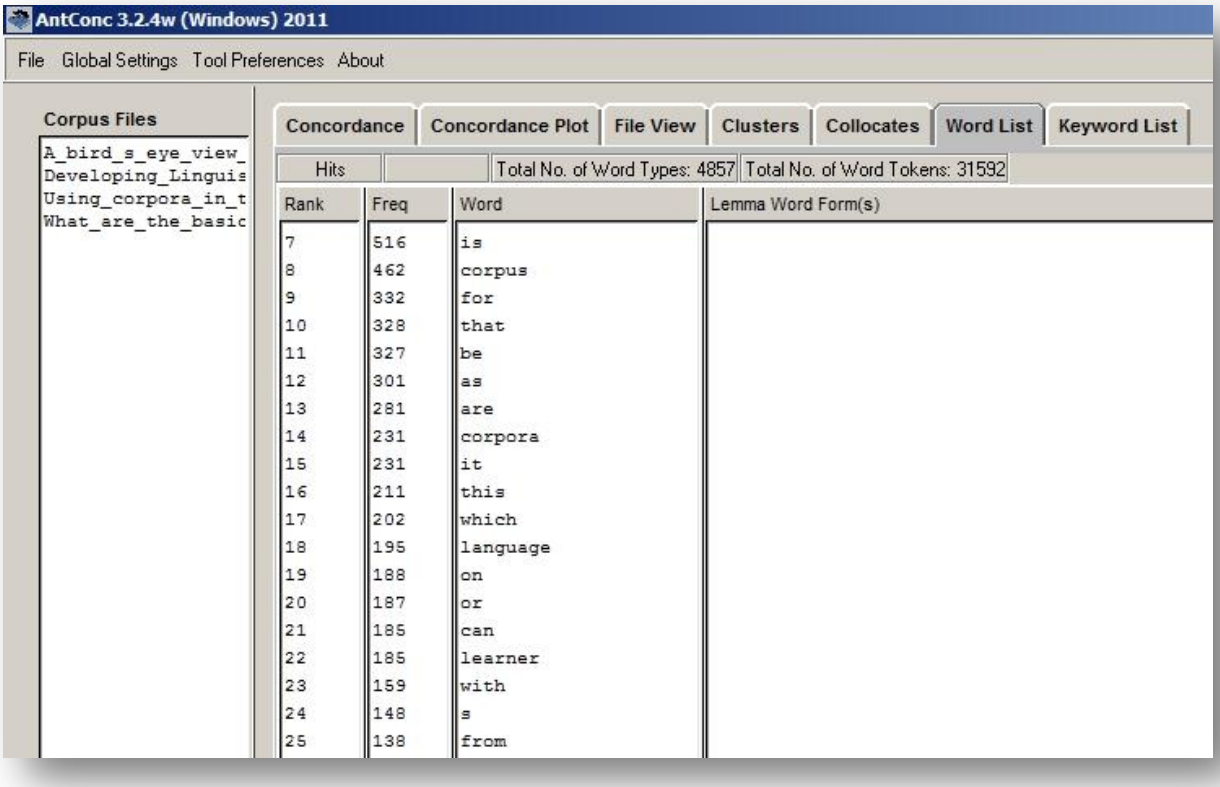
TRIBBLE, C. . Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In: Melia J & B Lewandowska-Tomaszczyk (eds). Proceedings of the First International Conference on Practical Applications in Language Corpora. Lodz: Lodz University Press, p. 106-117, 1997. Disponível em: <http://www.ctribble.co.uk/text/Palc.htm>. Acesso em: 18 ago. 2014.

10. ANEXOS

Anexo 1 - Lista de Palavras do *corpus* de textos em português.



Anexo 2 - Lista de Palavras do *corpus* de textos em inglês.

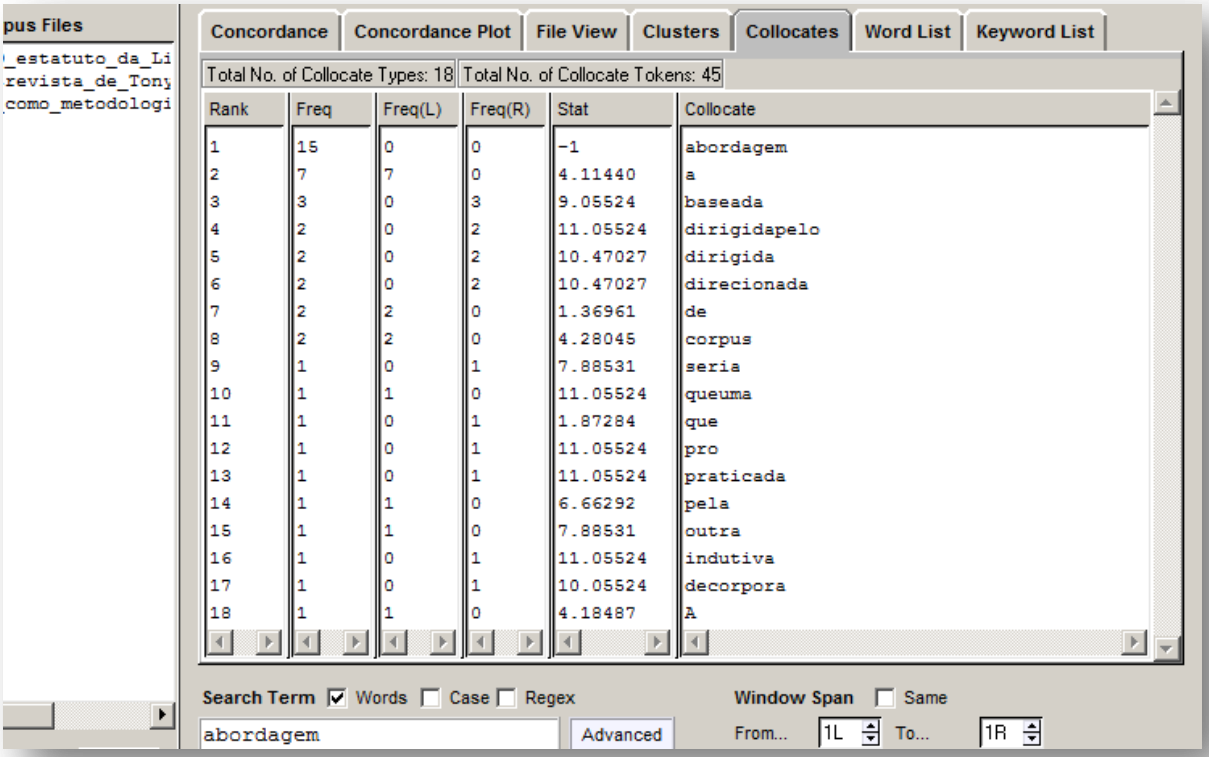


Anexo 3 - Lista com candidatos a termos obtidos por critério de maior frequência de ocorrência no *corpus* de textos em português.

CANDIDATO A TERMO	FREQUÊNCIA DE OCORRÊNCIA
corpus	352
corpora	213
linguística	179
dados	46
paralelo(s)	42
alinhamento(s)	38
processamento	24
anotação	21
abordagem(ns)	20
armazenagem	15
compilação	15
representatividade	15
etiquetagem	12
aplicabilidade	10
amostragem	7

Anexo 4 - Uso da ferramenta concordanciador para analisar possíveis ocorrências de termos sintagmáticos.

Anexo 5 - Uso da ferramenta colocados para listar as ocorrências de sintagmas no corpus.



Anexo 6 - Lista de candidatos a termos sugeridos pelos professores do LEA-MSI

Alinhamento;
Amostra;
Anotação;
Balanceamento;
Balizagem;
Colocado;
Compilação;
Concordância;
Concordanciador;
Corpora;
Corpus;
Corpus de referência;
Corpus paralelo;
Estudo baseado em corpus;
Estudo direcionado por corpus;
Etiquetador;
Etiquetagem;
Frequência;
Lematização;
Léxico;
Limpeza;
Manipulação;
Nódulo;
Ocorrência;
Processamento;
Representação;
Stop-list;

Anexo 7 – Lista utilizada para análise dos termos sugeridos pelos professores e dos termos obtidos pelo *corpus*.

CANDIDATO A TERMO	FOI SUGERIDO POR ALGUM PROFESSOR?	FREQUÊNCIA DE OCORRÊNCIA
corpus	SIM	352
corpora	SIM	213
linguística	NÃO	179
dados	NÃO	46
paralelo(s)	SIM	42
alinhamento(s)	SIM	38
processamento	SIM	24
anotação	SIM	21
abordagem(ns)	SIM	20
armazenagem	NÃO	15
compilação	SIM	15
representatividade	SIM	15
etiquetagem	SIM	12
aplicabilidade	NÃO	10
amostragem	SIM	7

Obs.: Os dez mais frequentes termos que constavam nas sugestões dos professores foram selecionados para compor a nomenclatura da proposta do glossário, sendo representados pela cor verde. Os termos representados pela cor vermelha foram descartados por não estarem entre os dez mais frequentes termos ou por não terem sido sugeridos pelos professores do LEA-MSI citados anteriormente.

Anexo 8 – Lista de candidatos a termos obtidos por critério de maior frequência de ocorrência no *corpus* (inglês), a fim de obter equivalências para os termos elencados em português.

CANDIDATO A TERMO	FREQUÊNCIA DE OCORRÊNCIA
corpus	462
corpora	231
data	123
frequency	90
linguistic	41
specialised	35
concordance	31
ocurrence(s)	29
tagging	27
parallel	24
reference	24
frequent	21
annotation	20
concordancing	15
tag	15
tagged	15
collocates	10
comparable	10
compiled	10
representativeness	10

Anexo 9 – Exemplos de Fichas Terminológicas Preenchidas neste trabalho

Entrada	*corpus*. pl. *corpora*	Marca gramatical	M
Equivalência(s) em inglês	*corpus*. pl. *corpora*		
Definição do termo	O termo vem do latim, e pode significar: "conjunto de uma obra". É caracterizado, por diversos autores, como um conjunto de textos em linguagem natural que podem ser convertidos para formato eletrônico para ser tratados por computador. São, então, armazenados e compilados considerando critérios pré-estabelecidos pelo pesquisador, visando à finalidade a que estão propostos.		
Fonte da definição	SHEPHERD M. G., Tania. *O estatuto da Linguística de Corpus: metodologia ou área da Linguística?*. Matraga, Rio de Janeiro, v. 16, n.24, p. 151, jan./jun. 2009.		
Contexto do Termo no *corpus*	"A Lingüística de ***corpus*** se ocupa de quase todas as áreas de investigação linguística. O léxico é a que mais recebe a atenção dos linguistas de ***corpus*** e é a que mais se projeta para o mundo, basta ver os dicionários de inglês atuais, que são produzidos com base em ***corpus***".		
Fonte do contexto	BERBER SARDINHA, Tony. Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem – ReVEL*. Vol. 2, n. 3, ago. 2004.		
Remissiva(s): Termo(s) conexo(s)	linguística de *corpus* (*corpus linguistics*); abordagem baseada em *corpus* (*corpus-based*); abordagem direcionada pelo *corpus* (*corpus-driven*); *corpus* de referência (*reference corpus*); *corpus* especializado (*specialised corpus*); *corpus* paralelo (*parallel corpus*).	Remissiva: sigla	
Notas			

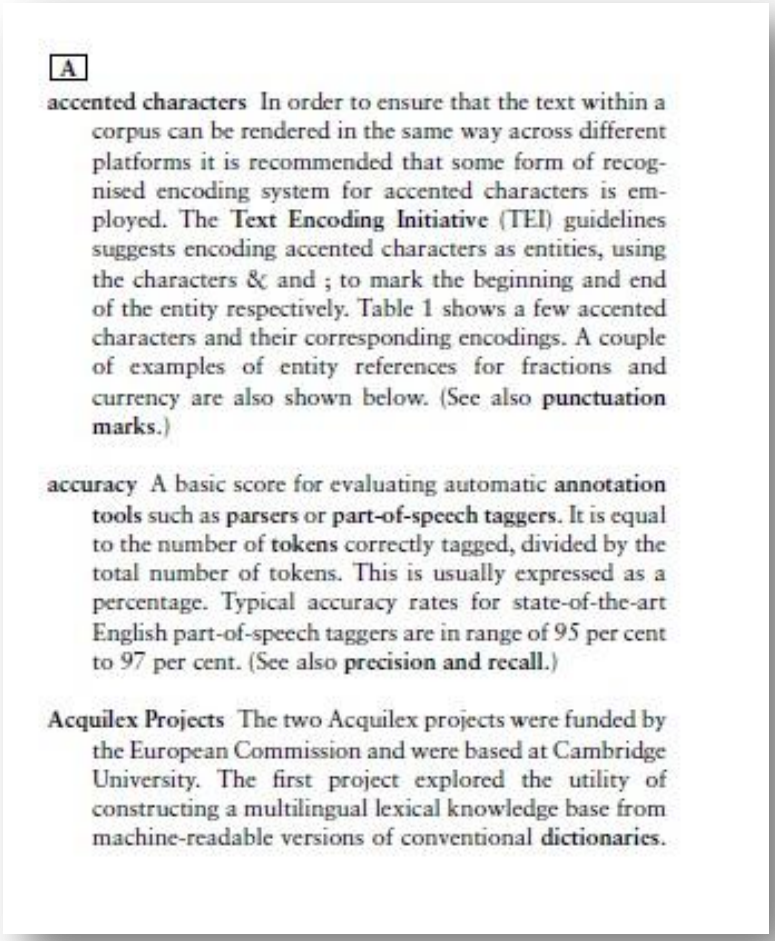
Entrada	abordagem direcionada pelo *corpus*.	Marca gramatical	F

Equivalência(s) em inglês	*corpus-driven.*		
Definição do termo	É uma metodologia na qual são geradas hipóteses no âmbito lexical e gramatical sobre o *corpus* analisado, conforme os dados sejam apresentados. A abordagem direcionada pelo corpus, diferentemente da abordagem baseada em *corpus*, não visa fazer testes e exemplificar teorias já existentes, mas observar e analisar padrões e frequências lexicais.		
Fonte da definição	SHEPHERD, T. M. G. . *O Estatuto da Linguistica de Corpus: Metodologia ou Área da Línguística*. Matraga (Rio de Janeiro), v. 16, p. 154, 2009.		
Contexto do Termo no *corpus*	"A **abordagem direcionada pelo *corpus*** se deve, segundo Sinclair (2004ª, p.xviii) à ausência de uma teoria que explicasse as relações lexicais na gênese dos trabalhos com corpora eletrônicos".		
Fonte do contexto	SHEPHERD, T. M. G. . *O Estatuto da Linguistica de Corpus: Metodologia ou Área da Línguística*. Matraga (Rio de Janeiro), v. 16, p. 150-172, 2009.		
Remissiva(s): Termo(s) conexo(s)	abordagem baseada em *corpus* (*corpus-based*).	Remissiva: sigla	
Notas			

Entrada	etiquetagem.	Marca gramatical	M
Equivalência(s) em inglês	*tagging.*		
Definição do termo	Relacionado à anotação de *corpus*, constitui-se no ato de marcar o *corpus* com etiquetas (códigos), que variam conforme o nível de informação. Ver também: etiquetagem morfossintática.		
Fonte da definição	BAKER, P. ; HARDIE, A. ; MCENERY, T. . *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press, p.154 , 2006.		

Contexto do Termo no *corpus*	[Tradução livre]: “Experimentos têm mostrado que, se os resultados de um aluno são bastante avançados, com uma baixa proporção de erros de ortografia e morfológicos, a taxa de sucesso do etiquetador é semelhante à obtida quando se faz a **etiquetagem** de dados de estudantes nativos. Entretanto, quanto mais desviantes são os dados, menos precisa será a **etiquetagem**, a ponto de tornar o uso do etiquetador impraticável”.		
Fonte do contexto	GRANGER, S. . A bird’s eye view of learner corpus research. In: Granger, S. ; Hung, J. ; Petch-Tyson, S. (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam e Philadelphia: Benjamins, p. 3-33, 2002.		
Remissiva(s): Termo(s) conexo(s)	etiquetagem morfossintática (*POS-tagging*); etiquetador (*tagger*); e etiqueta (*markup*).	Remissiva: sigla	
Notas			

Anexo 10 – Exemplo de diagramação selecionada para ser aplicada na proposta de glossário deste trabalho



A

accented characters In order to ensure that the text within a corpus can be rendered in the same way across different platforms it is recommended that some form of recognised encoding system for accented characters is employed. The **Text Encoding Initiative** (TEI) guidelines suggests encoding accented characters as entities, using the characters & and ; to mark the beginning and end of the entity respectively. Table 1 shows a few accented characters and their corresponding encodings. A couple of examples of entity references for fractions and currency are also shown below. (See also **punctuation marks**.)

accuracy A basic score for evaluating automatic **annotation tools** such as **parsers** or **part-of-speech taggers**. It is equal to the number of **tokens** correctly tagged, divided by the total number of tokens. This is usually expressed as a percentage. Typical accuracy rates for state-of-the-art English part-of-speech taggers are in range of 95 per cent to 97 per cent. (See also **precision and recall**.)

Acquilex Projects The two Acquilex projects were funded by the European Commission and were based at Cambridge University. The first project explored the utility of constructing a multilingual lexical knowledge base from machine-readable versions of conventional **dictionaries**.

Referência: BAKER, P. ; HARDIE, A. ; MCENERY, T. . *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press, p.7, 2006.