seminário de iniciação científica
ISSN 2558-6052

CALIEMT: CORPUS DE APRENDIZAGEM DA LÍNGUA INGLESA DO ENSINO MÉDIO TÉCNICO

Matheus Emmanuel Ferreira da Silva[1]; Pedro Aquiles Mazon Araujo[2]; Shirlene Bemfica de Oliveira[3]

1 Matheus Emmanuel Ferreira da Silva, Bolsista (IFMG), Ensino Médio Técnico Integrado em Edificações, IFMG Campus Ouro Preto, Ouro Preto - MG; matheusemmanuelsilva@gmail.com
2 Pedro Aquiles Mazon Araujo, Ensino Médio Técnico Integrado em Automação Industrial, IFMG Campus Ouro Preto, Ouro Preto – MG; aquilesaquiles710@gmail.com
3 Orientador: Pesquisadora do IFMG, Campus Ouro Preto; shirlene.o@ifmg.edu.br

RESUMO

Este projeto de pesquisa tem como objetivo organizar o Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico (CALIEMT). O CALIEMT é um corpus de amostras escritas e orais construído por alunos de língua inglesa do Ensino Médio Técnico. A proposta traz uma abordagem heurística para o ensino de Inglês no âmbito da escola técnica integral onde os alunos discutem temas transversais e escrevem textos na língua inglesa em uma atmosfera animada de discussão e descoberta acadêmica genuína, confirmando a velha máxima latina '*Docendo discimus*' (Ensinando aos outros, aprendemos). Esta abordagem de Multiletramentos é promovida por meio de aulas que focam no desenvolvimento das habilidades integradas (*reading, writing, listening, speaking*), e na análise e produção de textos escritos de diversos gêneros que demonstram o posicionamento crítico dos alunos sobre temas transversais. A pesquisa visa tornar o ensino de idiomas mais orientado pelos dados (*data-driven*) e o processo de aprendizagem mais centrado no aluno que pode aprender por descoberta de padrões linguísticos, pelo uso do computador e suas ferramentas (*Computer-Assisted Language Learning - CALL*), enfatizando o pensamento crítico e a metacognição dos alunos. Os textos são produzidos individualmente, em pares e em grupos em tarefas de sala de aula, e são disponibilizados a comunidade científica e aos alunos no formato eletrônico. Os dados para a pesquisa são coletados presencialmente em salas de aula ou no Laboratório de Ensino, Pesquisa e Extensão e à distância por meio de atividades propostas na plataforma *Moodle* e com o uso da ferramenta *Google Forms*. A compilação de um corpus de aprendizes do ensino médio técnico se justifica pela carência de estudos dessa natureza e pela raridade de amostras de escrita de alunos deste nível. O CALIEMT pode preencher a lacuna de estudos com ênfase na interlíngua de aprendizes, de identificação de padrões de vocabulário, sintaxe, gramática e estilo na escrita dos alunos adolescentes dos níveis de proficiência básico, intermediário e avançado. A importância dos estudos com corpora de aprendizes é enfatizada por diversos pesquisadores (NESSELHAUF, 2004; GRANGER, 2002), principalmente pelo fato de revelarem as dificuldades dos aprendizes de uma determinada língua, sendo possível, assim, analisar a sua produção linguística. Além disso, as amostras de escrita de alunos podem ser utilizadas em sala de aula em tarefas de conscientização linguística, análise de erros, correção em pares, etc. Dessa forma, o valor do corpus está vinculado tanto ao processo de construção textual dos alunos em sala de aula quanto ao produto educacional final que é o próprio banco de dados para consulta.

INTRODUÇÃO:

> "A linguagem deve ser estudada em instâncias reais, atestadas e autênticas de uso, não como frases intuitivas, inventadas, isoladas" (STUBBS, 1993, p. 2).

Pesquisas com foco em corpora de aprendizes são fontes ricas que podem propiciar diversas investigações nas áreas de Educação, Linguística Aplicada, Linguística de Corpus, uma vez que os dados advindos desses estudos propiciam a compreensão de como alunos empregam a linguagem nos gêneros e registros textuais, como entendem os conceitos estudados em sala de aula, contribuindo para sua formação cidadã, e na promoção de ações de justiça social. Além disso, esse processo de produção textual sistematizado se

ISSN 2558-6052

constitui em novas metodologias de investigação, de ensino e de aprendizagem de línguas (BERBER SARDINHA, 2004; ALUISIO et al., 2006; GONZALES, 2007). Especificamente, no contexto da escola regular de Educação Profissional e Tecnológica, as metodologias de criação, compilação e divulgação de corpora eletrônicos de aprendizes dessa faixa etária são raras e, geralmente os corpora são construídos com foco de investigação em aspectos formais da língua e/ou para avaliar os níveis de proficiência dos alunos.

Esta pesquisa objetiva a organização e compilação do CALIEMT (Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico), que tem potencial inovador por não existir no Brasil um corpus dessa natureza. Os resultados da compilação e organização do banco de dados eletrônico é um corpus com amostras de textos escritos e transcrições do discurso oral de alunos dos níveis de proficiência básico, intermediário e avançado que estudam em um Instituto Federal.

> O corpus de aprendiz constitui de "uma compilação de textos escritos, não publicados, produzidos num ambiente de ensino ou treinamento, geralmente para serem avaliados" (SCOTT & TRIBBLE, 2006, p.133) que possibilita a investigação de fenômenos da língua, tendo como base a frequência e os padrões de uso.

No que tange as abordagens de ensino e aprendizagem de línguas, este projeto parte da leitura dos termos de assentimento para que os pais possam autorizar o uso dos textos dos alunos no CALIEMT[1], testagem dos alunos que fazem um teste de proficiência para a conscientização de seu nível linguístico. Em seguida, há o planejamento e aplicação de aulas de língua inglesa em grupos heterogêneos, que promovem a ativação dos conhecimentos prévios dos alunos sobre temas transversais e Objetivos de Desenvolvimento Sustentáveis (ODSs), que promovem a interação e debates entre os discentes sobre as temáticas e a construção e reconstrução de conhecimentos que podem se materializar em uma diversidade de gêneros textuais orais e escritos. Esse processo de construção textual visa à socialização dos alunos, o fomento ao debate, a construção do discurso, a promoção de andaimes coletivos e a Aprendizagem Significativa (STORCH, 2005; VYGOSTSKY, 1987; AUSUBEL et al., 1983; OLIVEIRA et al., 2018).

Após a coleta, com financiamento específico (interno da instituição ou do pesquisador responsável), esse corpus será divulgado em servidor próprio por meio do Software CqpWEB (gratuito na Web) a ser divulgado a comunidade científica interna e externa, após autorização, podendo ser utilizado em pesquisas de salas de aula, publicações sobre as nuances da produção em língua inglesa de alunos dessa faixa etária e sobre os conteúdos a serem discutidos e produzidos pelos alunos. Este projeto propõe uma interface teórico-metodológica entre as áreas de Linguística Aplicada (LA) e a Linguística de Corpus (LC) para a compilação e construção de um Corpus com amostras textos em inglês produzidos por alunos do ensino médio técnico, o CALIEMT.

Linguística de Corpus

A Linguística de Corpus (LC) é a área do conhecimento que estuda a linguagem por meio da utilização do computador (GONZÁLES, 2007, p. 8). Ela é definida como uma maneira de se chegar à linguagem por meio da análise dos padrões probabilísticos que se constroem nos contextos em que os falantes os empregam (BERBER-SARDINHA, 2004). A principal característica da LC é, segundo Gonzáles (2007, p. 8), a observação de dados empíricos armazenados em bancos de dados que compõem um corpus, com a utilização de ferramentas eletrônicas que auxiliam na análise de dados verificando os fenômenos da língua em uso. Para este tipo de análise, recorremos a um corpus que é entendido como

> um conjunto de dados linguísticos (pertencentes ao uso escrito da língua), sistematizado segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso de algum de seus âmbitos, dispostos de tal modo que possam se processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SANCHEZ, 1995, p.8-9).

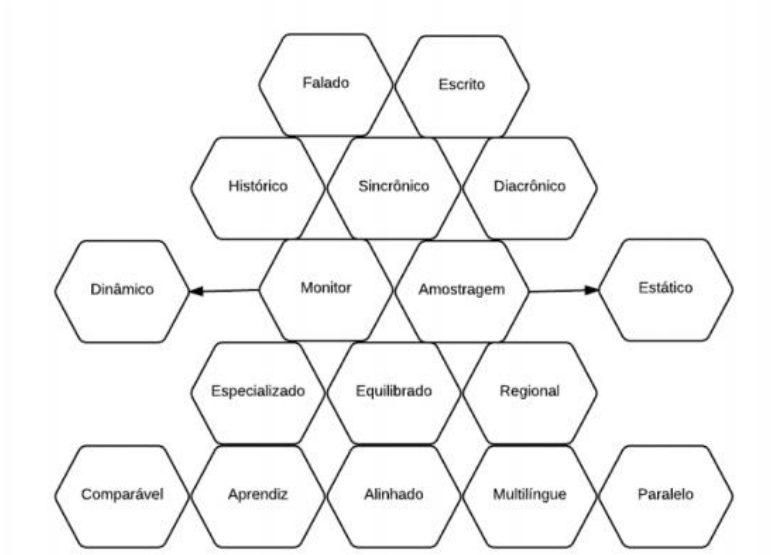
De acordo com Resende e Maverick (2016, p. 8), o corpus pode ser falado ou escrito. O corpus falado é aquele composto por porções de falas que foram transcritas e preparadas para serem armazenas em um banco de dados, no caso desse estudo, as transcrições de falas dos alunos e da professora produzidas durante as interações para as produções textuais gravadas presencialmente com o uso de filmadoras. Os dados escritos do CALIEMT serão textos escritos individualmente, em pares ou grupos na língua inglesa por

[1] CAAE: 60114922.4.0000.8507

ISSN 2558-6052

alunos do ensino médio técnico. Esses materiais serão digitalizados e como todos os corpora, serão armazenados em um banco de dados. Abaixo seguem outras tipologias dos corpora:

Figura 1: Tipologia dos Corpora



Fonte: RESENDE; MAVERICK (2016, p. 8)

Além de serem falados ou escritos, os corpora também podem ser considerados sincrônicos, diacrônicos ou históricos. Os corpora sincrônicos são formados por textos com uma linguagem contemporânea, os corpora diacrônicos retratam diferentes períodos de tempo e os corpora históricos retratam um tempo passado (RESENDE; MAVERICK, 2016, p. 8). Segundo os autores, há ainda os corpora de amostragem e os monitores. O CALIEMT será um corpus sincrônico composto por textos de temáticas transversais contemporâneas e de cunho social. Segundo os autores, há ainda os corpora de amostragem e os monitores. Os corpora de amostragem são compostos por porções de textos ou variedades textuais finitas e são planejados para uma amostra finita da linguagem como um todo. Eles são sempre estáticos, ou seja, depois de compilados, não é mais possível acrescentar e nem diminuir dados.

Os corpora monitores são compostos por dados constantemente reciclados de forma que possam refletir o estado atual de uma língua e por essa característica são também chamados de corpora dinâmicos ou corpora orgânicos, como é o caso do CALIEMT (RESENDE; MAVERICK, 2016, p. 8). Quando os corpora são compostos por um número semelhante de textos ou de dados, ou seja, se possuem uma quantidade de dados equilibrada, como por exemplo, o mesmo número de textos ou o mesmo número de gêneros ou registros, eles recebem o nome de corpora equilibrados (RESENDE; MAVERICK, 2016, p. 9). O CALIEMT é um corpus especializado, pois é específico de uma área: de língua inglesa. Ele será composto por dados de alunos das disciplinas de língua inglesa do ensino médio técnico da região sudeste e por isso é denominado corpus de aprendiz. Ele será construído a partir de textos de diversos registros em um idioma: o inglês.

Os corpora paralelos são também compostos por textos em dois idiomas diferentes, mas, diferente dos comparáveis, os textos são formados por traduções e podem ou não ser alinhadas. Quando os corpora são paralelos e alinhados isto quer dizer que uma linha do texto original fica alinhada com a linha do texto traduzido e pode-se ler o texto fonte e o texto traduzido um ao lado do outro ou um acima do outro, não é o caso do CALIEMT. Por último, temos os corpora de treinamento que também são chamados de corpora de testes, eles são desenhados para permitir o desenvolvimento de aplicações e ferramentas de análise linguística (RESENDE; MAVERICK, 2016, p. 10).

Os dados produzidos para o CALIEMT serão coletados anualmente, o que demonstra uma constante dinâmica e reflete algumas características da interlíngua dos aprendizes. Ele não será equilibrado, pois está sendo construído com diversos gêneros e registros por aprendizes de níveis diferentes. O corpus será especializado com foco no emprego da língua inglesa por alunos da disciplina de língua inglesa do ensino médio técnico, conforme mostra a tabela abaixo:

ISSN 2558-6052

Tabela 1: Composição do CALIEMT

CRITÉRIOS	CORPUS DE LINGUA ESTRANGEIRA	
Abrangência	Regional: região sudeste do Brasil	
Meio	Escrito e oral	
Tempo	Contemporâneo	
Renovação	Dinâmico Anual	
Línguas	Inglês	
Produtores	Aprendizes de língua inglesa do Ensino Médio Técnico de níveis de proficiência A1, A2, B1, e B2	
Emprego	Estudo	
Intervenções	Sem intervenção ou correção dos professores	
Estratégia	2011: Texto produzido a partir de duas versões não editadas pelo professor 2012-2014: Texto produzido a partir de mapas conceituais e produção escrita colaborativa 2015-2017: Escrita colaborativa e gravação das interações 2017-2023: Textos escritos em duas versões	
Gêneros Registros	Texto expositivo de definição (2011) Mapas Conceituais (2012-2014) Textos argumentativos (2015-2023) Poemas, Narrativas, Depoimentos, gravações de interações (2015-2023)	- Distúrbios alimentares - Bullying - Violência na sociedade moderna - Arte e sociedade - Racismo - ODSs - Temas propostos pelo International Corpus of Learner English (ICLE)

Do autor

METODOLOGIA:

As pesquisas empíricas sobre corpora de aprendizes são muito recentes e o caráter inovador deste estudo se deve a “uma grande carência de estudos sobre a interlíngua de aprendizes brasileiros” e de “compilações de corpus de aprendizes no Estado de Minas Gerais” (DUTRA, 2010, p. 03).

Esta pesquisa é uma pesquisa de sala de aula de natureza empírica, com análises quantitativas e qualitativas. Ela é desenvolvida em salas de aula do ensino médio regular e no Laboratório de Ensino, Pesquisa e Extensão em Línguas Estrangeiras de um Instituto Federal com a participação da pesquisadora e de 2 bolsistas do ensino médio técnico. Nos institutos federais, em geral, os alunos são matriculados em seus cursos técnicos e as turmas de língua inglesa são muito heterogêneas e categorizadas como Língua Inglesa I, II e III.

No contexto desta pesquisa, os alunos permanecem em seus grupos de origem (Administração, Automação Industrial, Edificações, Mineração e Metalurgia), fazem um teste de proficiência apenas para conscientização de seu processo de aprendizagem e participam das dinâmicas de sala de aula para o debate dos temas e para o processo de ativação de conhecimentos prévios e escrita das primeiras versões. A pesquisadora conduz as atividades em sala, com aproximadamente 300 alunos das disciplinas de língua inglesa distribuídos em 9 turmas. A maioria dos alunos participantes são do nível A1 e por isso as discussões são feitas na língua inglesa e na língua portuguesa. Os alunos com níveis de proficiência B1 e B2 são convidados a participar da produção de textos em pares e as discussões são gravadas em vídeo e áudio. Eles discutem a proposta e

ISSN 2558-6052

escrevem o texto em coautoria na língua inglesa. Os textos produzidos pelos participantes em sala de aula, no laboratório, bem como as transcrições das gravações em áudio fazem parte do CALIEMT.

O processo de compilação e organização do CALIEMT está ancorado pela literatura da área de Linguística de Corpus que sugere que para a compilação de um corpus, os pesquisadores devem seguir alguns procedimentos: primeiramente, a elaboração do projeto do corpus, que inclui a seleção ou organização dos textos considerando a autenticidade, representatividade, balanceamento e diversidade (BIBER et al., 1998; BERBER SARDINHA, 2004; SINCLAIR, 1991). O processo de compilação está sendo feito com base em (RESENDE; MAVERICK, 2016) que consiste na:

> 1. seleção de textos e pedidos de permissão de uso;
> 2. manipulação e limpeza do corpus;
> 3. etiquetagem (anotação estrutural, cabeçalhos) e nomeação dos arquivos;
> 4. exclusão das etiquetas XML2 para processamento do corpus e contagem de palavras (RESENDE; MAVERICK, 2016, p. 11).

No que diz respeito à metodologia de compilação e organização do CALIEMT- Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico, a base teórica e metodológica advém da Linguística de Corpus (LC) que é a área do conhecimento que estuda a linguagem por meio da utilização do computador, conforme discutido anteriormente (GONZÁLES, 2007, p. 8). O CALIEMT está em processo de construção e será um corpus sincrônico composto por textos de temáticas transversais contemporâneas e de cunho social. Os dados escritos do CALIEMT são textos escritos individualmente, em pares ou grupos na língua inglesa por alunos do ensino médio técnico, além da transcrição da discussão oral para a produção de textos de alunos mais proficientes. Esses materiais são digitados e como todos os corpora, armazenados em um banco de dados. São feitas duas coletas por semestre, justificando o caráter dinâmico do corpus. A alimentação do corpus eletrônico é anual e, o emprego dos textos, inicialmente é feito para fins de estudo dos alunos, para fins acadêmicos de análise e publicações de pesquisadores acerca do processo de ensino e aprendizagem das línguas estrangeiras. A professora pesquisadora utiliza os textos em tarefas de conscientização das estruturas da língua, e para fomentar as discussões temáticas.

As estratégias de ensino envolvem a discussão das temáticas transversais e os Objetivos de Desenvolvimento Sustentáveis que já fazem parte dos temas curriculares discutidos em sala de aula e a produção de textos serão feitas na língua inglesa por meio das seguintes ferramentas e gêneros textuais:

Estratégias

- Textos produzidos a partir de versões (listas feitas a partir de tempestade de ideias, rascunhos e versão final).
- Textos produzidos a partir de mapas conceituais (mapas conceituais a partir de discussão das temáticas, produção de texto a partir dos mapas).
- Textos produção em coautoria ou escrita colaborativa (alunos proficientes no laboratório gravam a discussão e escrevem uma versão em coautoria).

Gêneros ou registros

- Textos expositivos de definição / opinião.
- Textos argumentativos ou dissertativos.
- Mapas Conceituais.
- Poemas.
- Narrativas.
- Depoimentos.

RESULTADOS E DISCUSSÕES:

Em sala de aula, os alunos têm encontros semanais de 1h e 40 min. e utilizam os materiais didáticos aprovados pelo PNLD. Durante as aulas são desenvolvidas atividades que contemplam as habilidades de compreensão e produção oral e escrita (*listening, speaking, reading and writing*), além de pronúncia, gramática e vocabulário. Além disso, a língua é ensinada como uma língua global (CRISTAL, 2003) com base na valorização dos processos de mullticulturalidade e ênfase nos conhecimentos e experiências prévias dos participantes. Nessas aulas, o professor pesquisador também tem a preocupação em promover o

ISSN 2558-6052

empoderamento dos jovens por meio das interações, do ativismo e do engajamento de causas sociais para promover um processo de reconstrução de suas identidades e de formação cidadã, por meio das discussões dos Objetivos de Desenvolvimento Sustentáveis para a reflexão e ação para a justiça social (RAJAGOPALAN, 2003).

Além das coletas em sala de aula, por meio das gravações das aulas, os alunos usam o laboratório de línguas para a gravação de trabalhos em grupos menores. Nesses grupos menores é estimulada a livre expressão, a imaginação e da criatividade, para a discussão dos temas e produções escritas individuais e colaborativas. O foco dessas atividades é nos Letramentos Sociais, com propostas de tarefas em que os participantes são desafiados a trabalhar em equipe, a questionar, resolver problemas e produzir textos de forma colaborativa (STREET, 2014), conforme mostra a imagem de um desses encontros:

Imagem 1: Alunos em pares discutindo uma das temáticas propostas

Fonte: do autor

Esses dados orais e escritos são anteriormente analisados com foco na produção textual e no processo de construção colaborativa. Estas análises são feitas com o auxílio de concordanciadores eletrônicos apresentando as construções lexicais mais frequentes e seus colocados.

Desde o início do projeto, há um foco na construção dos dados para o corpus, para sistematizar o processo de compilação, manipulação, nomeação dos arquivos de textos e arquivamento digital dos pedidos de permissão de uso das produções. Os arquivos selecionados são salvos em um servidor local já criado pela TI no Laboratório de Ensino, Pesquisa e Extensão em Línguas Estrangeiras[2] com a Terminologia (CALIEMT). As imagens dos mapas conceituais serão digitalizadas utilizando o Cmaptools para serem integradas e linkadas aos textos escritos. Para a limpeza e manipulação do corpus e para que os textos, em seus formatos originais de disponibilização, possam ser corretamente processados por ferramentas computacionais, será necessário converter manualmente todos os arquivos de texto de seus formatos originais (Microsoft Word de extensão “.doc”, para um único formato padrão, no caso o Bloco de Notas de extensão “.txt”, pois este não possui códigos de formatação, mas apenas caracteres do teclado (letras, números e símbolos ortográficos). Além disso, serão excluídas possíveis tabelas, gráficos, fórmulas, cálculos, fotos e toda informação que não esteja em forma de texto para a limpeza do corpus.

O trabalho posterior seguindo a proposta de Resende e Maverick (2016), é a anotação estrutural, geração de cabeçalhos e nomeação de arquivos. Por meio da versão adaptada do Editor de Cabeçalhos12 do Projeto Lacio-Web13 ou pela Plataforma TEITOK, será possível realizar a anotação estrutural dos textos, adicionando-se em cada texto um cabeçalho e uma nomeação específica e padronizada.

A anotação estrutural, de acordo com Aluísio e Almeida (2006), compreende a marcação de dados externos e internos dos textos. Os dados externos são formados pela documentação do corpus na forma de um cabeçalho que inclui os metadados textuais, isto é, tamanho do arquivo em palavras, autoria, a tipologia

[2] O laboratório de Ensino, Pesquisa e Extensão de Línguas Estrangeiras conta com 6 computadores com acesso a internet e webcams acopladas, 4 gravadores digitais e uma câmera filmadora com tripé advindos de um edital de fomento interno.

seminário de iniciação científica
ISSN 2558-6052

textual e informação sobre a distribuição do corpus. Os dados internos é a anotação de segmentação do texto cru, que envolve:

a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos;
b) marcação da estrutura de subparágrafos – elementos que são de interesse linguístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc (ALUÍSIO; ALMEIDA, 2006, p. 14).

De acordo com os autores, a anotação estrutural do texto é um processo pelo qual parcelas de texto que constituem informações diferenciadas por sua relevância (ou pela falta de relevância), e são marcadas por etiquetas por meio do Editor de Cabeçalhos.

Aluísio e Almeida (2006) instruem que são necessárias para a geração dos cabeçalhos: as informações: título, subtítulo, ano e local da produção, comentários (informações adicionais sobre o processo de produção), autoria (Individual, par, grupo, ano de instrução, sexo do autor), Gênero Textual, Tipo de Texto. Essas informações mostram-se relevantes para o CALIEMT bem como para outros projetos que pretendam reutilizar o corpus. Reproduzem-se, abaixo, a título de exemplo, todas as informações que podem ser recuperadas do texto:

**** separa os textos
* separa as variáveis
sexo 1 masculino, 2 feminino
ins ano no ensino médio técnico / nível de proficiência (1, 2, 3 ou 4)
cur curso no ensino médio técnico MIN Mineração

Exemplo:
**** *ano_2011 *n_001 *sex_11 *ins_1 *cur_MIN

Com a finalização do corpus, será preciso contar o número de palavras que ele contém. Para isso, segundo Aluisio e Almeida (2006) não é recomendável considerar as etiquetas XML inseridas pelo Editor de Cabeçalhos, pois o programa contaria todas as palavras que estão entre as etiquetas como sendo palavras do corpus. Dessa forma, é preciso remover tais etiquetas a fim de deixar o corpus enxuto. Uma outra razão para a exclusão das etiquetas é o fato de que muitos programas de PLN não processam com facilidade (e alguns simplesmente não processam) corpus com etiquetas XML.

Um gesto de interpretação de textos sobre bullying no contexto escolar

O recorte apresentado a seguir é da análise de uma parte do corpus escrito sobre *bullying* no sistema escolar. O foco é dado no uso de concordanciadores para comparar as diferentes perspectivas dos alunos sobre o mesmo assunto. A análise quantitativa foi feita com base nas produções de textos de oito alunos que escreveram quatro textos em pares.

Com o auxílio do Iramuteq, o corpus foi subdividido em 44 segmentos de texto sendo 1656 palavras, sendo 524 formas e 310 hapax, 59,16% das formas e 18,72% das ocorrências. O Iramuteq criou as listas de frequência em tabelas apresentando o corpus subdividido em 465 lemas , 306 formas ativas, 159 formas suplementares, formas ativas com frequência menor ou igual 3:63. O entrecruzamento feito com as formas ativas ou palavras de conteúdo, mais frequentes do corpus são *bully* (35), *school* (27), *problem* (19), *kid* (16), *violence* (13), *case* (11), *parent* (11), *student* (11), *person* (9), *victim* (8), *happen* (7), *teenager* (7), *family* (6), *society* (6). Elas indicam que o *bullying* é um problema social principalmente relacionado às crianças e adolescentes no contexto escolar. A lista apresentada a seguir apresenta as 10 primeiras ocorrências do corpus discutido:

Quadro 1: Formas Ativas e Suplementares do Corpus

Formas Ativas				**Formas Suplementares**			
n.	**Item**	**Frequência**	**Classe predominante**	**n.**	**Item**	**Frequência**	**Classe predominante**
1	*bully*	35	Adj	1	the	157	sw

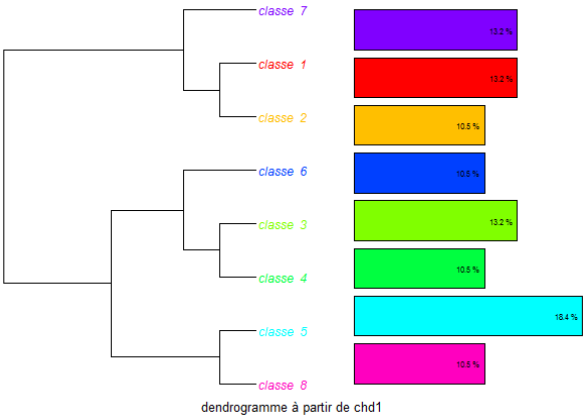
ISSN 2558-6052

2	*school*	27	Nom	2	and	74	sw
3	*problem*	19	Nom	3	to	71	sw
4	*kid*	16	Nom	4	of	48	sw
5	*violence*	13	Nom	5	a	37	sw
6	*student*	11	Nom	6	is	30	sw
7	*parente*	11	Nom	7	in	27	sw
8	*case*	11	Nom	8	that	24	sw
9	*case*	11	Nom	9	it	23	sw
10	*person*	9	Nom	10	with	21	sw

Fonte: Dados da pesquisa

Além das tabelas em excel, o Iramuteq oferece a análise das Estatísticas textuais: o número de textos e segmentos de textos, ocorrências, frequência média das palavras, bem como a frequência total de cada forma; e sua classificação gramatical. Nesta etapa, Nascimento e Meandro (2006) afirmam que o programa seleciona as formas reduzidas com frequência igual ou superior a 04 e define as Unidades de Contexto Elementar a partir da pontuação. O Iramuteq agrupou o corpus dos aprendizes em 44 segmentos de texto (UCEs). O dendograma apresentado pode ser visualizado por meio de uma imagem da Análise Fatorial de Correspondência (AFC) que segundo Nascimento e Meandro (2006) permite verificar as relações entre as classes num plano gráfico, o que permite a visualização geral articulada dos agrupamentos das palavras presentes no discurso dos autores. No corpus dos aprendizes, quanto mais próximos os elementos dispostos no plano, mais eles falam das mesmas coisas, por exemplo, no quadrante em alaranjado que representa a classe 2, o substantivo teenager que é o mais frequente e, portanto, maior, está mais relacionado aos vocábulos *agression, form, research, physical, attack*. O mesmo substantivo está distante dos vocábulos de outros quadrantes, por exemplo, com vocábulos tais como: *consequence, parente*, ou todos os itens da classe 5. Isto demonstra que foram menos relacionados nos textos do grupo mesmo pertencendo a mesma classe.

Figura1: Dendograma das palavras mais frequentes do corpus



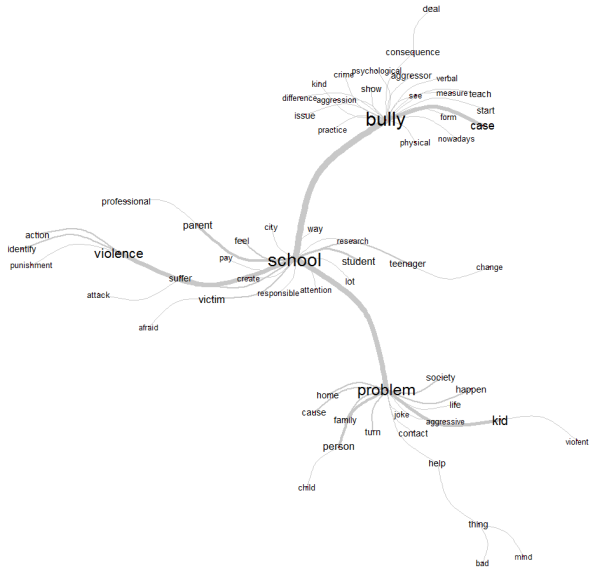
Fonte: extraído pelos autores com o auxílio do Iramuteq

Outra possibilidade de análise com o auxílio do Iramuteq, feita automaticamente é a análise da semelhança semântica das palavras e a representação em forma de árvore. Ela é baseada na teoria dos grafos, e mostra as co-ocorrências e as conexões entre as palavras do corpus. Na teoria dos grafos, de acordo com Lucchesi (1979), é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Segundo o autor, para comprovação, são empregadas estruturas chamadas de grafos, G (V,E), onde V é um conjunto não vazio de objetos denominados vértices e E é um subconjunto de pares não ordenados de V, chamados arestas. Lucchesi argumenta que dependendo da aplicação, arestas podem ou não ter direção, pode ser permitido ou não arestas ligarem um vértice a ele próprio e vértices e/ou arestas podem ter um peso (numérico) associado. Segundo ele, se as arestas têm uma direção associada (indicada por uma seta na representação gráfica) temos um dígrafo (grafo orientado). Um grafo com um único vértice e sem arestas é conhecido como grafo trivial. Estruturas que podem ser representadas por grafos estão em toda parte e muitos problemas de interesse prático podem ser formulados como questões sobre certos grafos. Por exemplo, no

ISSN 2558-6052

corpus de aprendizes apresentados neste artigo, a estrutura de ligações dos itens lexicais pode ser representada por um dígrafo: os vértices são os textos do corpus e existe uma aresta do texto 1 para o texto 2 se e somente se o texto 1 contém um link para 2 e assim por diante com os outros textos. Dígrafos são também usados para representar máquinas de estado finito. O Iramuteq gerou a árvore semântica da relação entre os quatro textos como mostra a figura 3 abaixo:

Figura 4: Análise de semelhança semântica em árvore



Fonte: extraído pela autora com o auxílio do Iramuteq

Partindo-se de agrupamentos de itens lexicais obtidos com o auxílio do software, foram encontrados três grandes grupos. O primeiro deles, correspondente ao item lexical *school* e itens associados, reflete um dos focos da proposta para a produção textual: o bullying no ambiente escolar. Itens como *parent* e professional, que refletem as opiniões dos alunos em relação àqueles responsáveis por tomar medidas sobre as ocorrências de bullying. O item lexical *violence* aparece em destaque dentro desse grupo, indicando o bullying como um causador e potencializador da violência no ambiente escolar. O segundo grupo tem como centro o item lexical *bully* e estão associados itens lexicais referentes a formas reconhecidas como *bullying* – verbal e *physical* -, bem como os termos *aggression*, agressor e crime, que refletem uma relação negativa que os autores dos textos estabelecem com a temática.

O terceiro grupo também reflete o ponto de vista negativa sobre o *bullying* – visto que é centrado no item lexical *problem* – de forma que a prática do *bullying* é colocada como causadora de problemas para os envolvidos, em seu ambiente familiar e escolar. Os alunos não utilizaram itens lexicais ou construções, como a primeira pessoa do singular, que representam relação direta pessoal com a temática, estando ausentes relatos pessoais de experiência e memória. A posição geral nos textos foi contrária à prática do bullying, refletindo as categorias estética negativa, funcionalidade negativa e valor mercantil negativo, de acordo com os critérios de descrição do objeto propostos por Bardin (2009, p. 68) e o sentimento de recusa, também citado por Bardin (2009, p.69) em relação à atitude perante o objeto.

A palavra central foi *school*, está com a interligação mais forte com a palavra bully logo com as palavras *problem* e *violence*. Após observar as palavras em maior destaque podemos concluir que o principal problema, o qual esta ocorrendo nas escolas é a violência por parte dos “valentões”. O problema ocorre principalmente com as crianças e jovens, estes por serem mais indefesos, tudo isto dentro de uma sociedade. O agressor pode agir de maneiras diferentes desde o modo verbal ao modo físico, ambas as agressões causam consequências psicológicas. Trata-se de um contato agressivo, uma piada que prejudicam modo de viver. Comunicar com a autoridade pode ser uma boa solução, porém antes disso o problema deve ser comunicado a uma pessoa da família, por exemplo, alguém em quem se confia para que assim exista apoio por parte destes, o qual é necessário. Deve-se identificar o tipo de violência cometida para que assim se encontre a melhor forma de punição. Infelizmente o bullying também pode ocorrer por parte dos membros da própria casa ou *família*. Todavia o local mais comum onde isto ocorre é nas escolas, local de plena

ISSN 2558-6052

responsabilidade, onde crianças e jovens, no caso, devem se sentir-se seguros assim como os pais por deixaremos lá. Os estudantes sentem-se atormentados por um novo ataque que pode estar por vim. Os pais e profissionais do mesmo devem estar atentos quanto as vítimas e o bullying, para que assim este seja evitado. A palavra *life* também é destacada, seguida pelas palavras *crime, violence, persons* e *prison*. Após observar estas palavras podemos concluir que a violência cometida pelas pessoas da sociedade é o problema central, este consiste em um crime que prejudica a vida de cada um, faz com que a vítima mude e transforme, mesmo que aos poucos, em um "prisioneiro" de si mesmo. A solução seria buscar a autoridade, como consequência a prisão.

CONCLUSÕES:

Este artigo teve como objetivo apresentar uma proposta de criação, compilação e análise do corpus de aprendizes da língua inglesa do Ensino Médio Técnico – o CALIEMT e demonstrar um recorte de análise de dados com foco na frequência das palavras usadas por um determinado grupo com o uso do concordanciador *Iramuteq*.

A pesquisa de organização e compilação do CALIEMT está em andamento, tem potencial inovador por não existir no Brasil um corpus dessa natureza. Os resultados da compilação e organização do banco de dados podem contribuir fortemente para as pesquisas nas áreas de Educação, Linguística de Corpus e Linguística Aplicada. O CALIEMT tem grande potencial social e impacto tecnológico: para os bolsistas, a participação traz o desenvolvimento de habilidades no processo de criação de bancos de dados e no uso de softwares. O docente enquanto pesquisador de sua própria prática pode refletir e tomar consciência do nível de abstração e generalidade presentes na profissão e das possibilidades de inclusão de novos saberes, metodologias e construções discursivas que beneficiem o aprendizado dos alunos e sua própria práxis. Aos alunos e bolsistas é dada a oportunidade de se tornarem mais ativos e conscientes de seu papel enquanto cidadãos e de serem pessoas capazes de transformar a educação por meio da produção de conhecimento científico. Além disso, o CALIEMT tem impacto econômico na medida em que pode contribuir com outros pesquisadores que poderão utilizar o banco de dados para outras pesquisas otimizando o processo científico.

Em relação ao recorte proposto com foco no bullying escolar analisou quantitativamente os dados das produções de textos escritos em coautoria sobre um dos problemas que eles vivenciam no contexto escolar. A proposta de ensino demonstra que alunos em pares, em tarefas de produção escrita colaborativa, escreveram textos inteligíveis, organizados e, em certa medida, coesos, além de promoverem a autonomia dos alunos. A natureza do conhecimento que os alunos constroem varia de conhecimento pessoal, científico, linguístico a conhecimentos originários de dentro e de fora do contexto escolar que funcionam como ferramentas simbólicas para o processo de aprendizagem mais consciente.

Este estudo é uma pequena amostra do potencial do *Iramuteq*, que viabiliza o mapeamento dos itens lexicais e permite a interpretação de corpora maiores, sem perder o contexto de uso dos itens. Este estudo pode ser replicado em comparação a corpus de referência, possibilitando a integração das análises quantitativas e qualitativas trazendo objetividade às interpretações adas ao conteúdo dos textos.

REFERÊNCIAS BIBLIOGRÁFICAS:

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 (UNISINOS), v. 4, p. 156-178, 2006. Disponível em: http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf.

AUSUBEL, D. P.; NOVAK, J. D.; HANESIAN, H. Psicología Educativa: Un punto de vista cognoscitivo. 2° Ed. TRILLAS México. 1983.

BERBER SARDINHA, A. P. Linguística de Corpus: Histórico e Problemática. In: D.E.L.T.A. v.16, n. 2, 2000, p. 323-367.

BERBER SARDINHA, A. P. Linguística de Corpus. Barueri-SP. Manole, 2004.

BIBER, D. S.; CONRAD; REPPEN, R. Corpus Linguistics: Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

ISSN 2558-6052
BRASIL. Parâmetros Curriculares Nacionais: Língua Inglesa. Secretaria de Educação Fundamental. Brasília: MEC/SEF, 1997.

BRASIL. Resolução Nº 466, DE 12 DE DEZEMBRO DE 2012. Disponível em https://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf Acesso em 28/04/2022.

CPQWeb. Disponível em: https://cqpweb.lancs.ac.uk/ Acesso em 28/04/2022.

CPqweb Portal. Disponível em: https://www.research.lancs.ac.uk/portal/en/upmprojects/cqpweb(855e22e7-97fb-4863-9101-9b245729ccbd).html Acesso em 28/04/2022.

CRISTAL, D. English as a global language. Cambridge: Cambridge University Press, 2003.

DUTRA, D. P. Agrupamentos lexicais na escrita de aprendizes brasileiros de inglês: um estudo baseado em corpus. Plano de trabalho apresentado ao Programa Pesquisador Mineiro. Edital FAPEMIG 03/2010.

GONZÁLES, Z. M. G. Linguística de Corpus na análise do Internetês. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Faculdade de Letras. Pontifícia Universidade Católica de São Paulo, São Paulo, 2007.

NESSELHAUF, N. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In: ASTON, G.; BERNARDINI, S. & STEWART, D. (Eds.). Amsterdam and Philadelphia: Benjamins, 2004, p.109-124.

OLIVEIRA, S. B.; ANICETO, A. B. V.; ALCANTARA, B. G.; SILVA, C. A. L.; AZEVEDO, L. T. ; MELO, L. O. ; FONSECA, M. M. S. Escrita colaborativa no ensino e aprendizagem da Língua Inglesa. In: Gislayne Elisana Gonçalves; Shirlene Bemfica de Oliveira. (Org.). Pesquisa e extensão na escola pública: possibilidades e Desafios. 1ed. Goiânia: Editora Espaço Acadêmico, 2018, v. 1, p. 140-154.

OLIVEIRA, E. B. SOUZA, W. F. Educação Profissional Tecnológica: implicações do movimento CTSA para o currículo. In: OLIVEIRA, A. R.; XAVIER, H. C. SILVA, J. F.; OLIVERIA S. B. (Orgs.). Educação Profissional no Brasil: da história à teoria, da teoria à práxis. Coleção Educação Profissional no Brasil. v. 1. Curitiba: Editora CRV, 2020, p. 137-152.

RAJAGOPALAN, K. Por uma Linguística crítica: linguagem, identidade e a questão ética. São Paulo: Parábola Editorial, 2003.

RESENDE, S. V.; MAVERICK, R. Planejamento, Compilação e Organização de Corpora. Blucher Social Sciences Proceedings, v.2, n. 3, Março de 2016. Disponível em: http://pdf.blucher.com.br.s3-sa-east-1.amazonaws.com/socialsciencesproceedings/viii-eblc-xiii-elc/06_artigo_03.pdf Acesso em 28/01/2018.

SINCLAIR, J. Corpus, concordance and collocation. Oxford: Oxford University Press, 1991.

SCOTT, M.; TRIBBLE, C. (2006). Textual Patterns: keywords and corpus analysis in language education. Amsterdam: John Benjamins, 2006.

STORCH, N. Collaborative writing: Product, process, and students' reflections. Journal of Second Language Writing. v. 14, 2005, p. 153–173. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1060374305000172 Acesso em 28/01/2018.

STREET, B. Letramentos Sociais. Abordagens críticas do letramento no desenvolvimento, na etnografia e na educação. Trad. Marcos Bagno. São Paulo: Parábola Editorial, 2014.

VYGOTSKY, Lev Semenovick. A formação social da mente. São Paulo: Martins Fontes. 1987.