PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS – MESTRADO
ÁREA DE CONCENTRAÇÃO EM LEITURA E COGNIÇÃO

Vitor Ricardo Duarte

ENSINO E PRODUÇÃO DE MATERIAL DE INGLÊS INSTRUMENTAL PARA A ÁREA DE TECNOLOGIA AMBIENTAL COM BASE NA LINGUÍSTICA DE *CORPUS*: UMA INTERFACE COM A LINGUÍSTICA COGNITIVA

Santa Cruz do Sul
2011

Vitor Ricardo Duarte

ENSINO E PRODUÇÃO DE MATERIAL DE INGLÊS INSTRUMENTAL PARA A ÁREA DE TECNOLOGIA AMBIENTAL COM BASE NA LINGUÍSTICA DE *CORPUS*: UMA INTERFACE COM A LINGUÍSTICA COGNITIVA

Dissertação apresentada ao Programa de Pós-Graduação em Letras – Mestrado, Área de Concentração em Leitura e Cognição, Linha de Pesquisa em Processos cognitivos e textualidade, Universidade de Santa Cruz do Sul – UNISC, como requisito parcial para obtenção do título de Mestre em Letras.

Orientadora: Prof.ª Dr. Onici Claro Flôres
Co-orientador: Prof. Dr. Antônio Paulo Berber Sardinha

Santa Cruz do Sul
2011

Vitor Ricardo Duarte

ENSINO E PRODUÇÃO DE MATERIAL DE INGLÊS INSTRUMENTAL PARA A ÁREA DE TECNOLOGIA AMBIENTAL COM BASE NA LINGUÍSTICA DE *CORPUS*: UMA INTERFACE COM A LINGUÍSTICA COGNITIVA

Esta Dissertação foi submetida ao Programa de Pós-Graduação em Letras – Mestrado; Área de Concentração em Leitura e Cognição; Linha de Pesquisa em Processos cognitivos e textualidade, Universidade de Santa Cruz do Sul – UNISC, como requisito parcial para obtenção do título de Mestre em Letras.

Prof.ª Dr. Onici Claro Flôres
Orientadora - UNISC

Prof. Dr. Antônio Paulo Berber Sardinha
Co-Orientador – PUC/SP

Prof.ª Dr. Alessandra Dahmer
Professora examinadora - UNISC

Prof.ª Dr. Simone Sarmento
Professora examinadora - UFRGS

AGRADECIMENTOS

À CAPES, pelo apoio financeiro.

Agradeço ao Dr. Tony Berber Sardinha, pela generosidade, paciência, sabedoria, *expertise* e grande conhecimento sempre disponíveis. Tony, sem teu apoio este trabalho jamais teria acontecido, sou muito grato a ti.

A Dra. Onici Claro Flôres pela leitura minuciosa e sempre atenta dos textos produzidos ao longo do curso.

À coordenadora do Mestrado em Tecnologia Ambiental da UNISC/RS, Dra. Adriane Lawisch Rodríguez pelo apoio para a realização desta pesquisa. Aos demais professores do curso, pela disponibilidade e auxílio na seleção de artigos do *corpus* aqui desenvolvido.

Aos pesquisadores do LAEL/PUCSP, Cristina Mayer Acunzo, Marcia Veirano Pinto, Renata Condi de Souza, Denise Delegá-Lúcio, Maria Cecília Lopes, Wendel Mendes Dantas, Eduardo de Carvalho Cassimiro por saber que pude contar com a ajuda de vocês. Agradecimento muito especial à Márcia, Cristina e Wendel pelas muitas sessões online de ajuda e suporte em momentos de dúvida.

Ao professor Elvio Funck. Esta pesquisa teve sua gênese em suas aulas de inglês instrumental, aulas que jamais esqueci.

À Profa. Dulci Boettcher pelo carinho, confiança e grande apoio na realização do estágio acadêmico.

À minha mãe, Gisella Maria Duarte, por tudo. A ela, por ter me contado histórias quando criança. Por continuar sendo uma grande contadora de histórias: da vida, dos livros, das vivências, das memórias.

Ao Vincent, pelo *nonsense*.

(...)bem antes de servir para comunicar, a linguagem serve para viver.

(Émile Benveniste)

RESUMO

O trabalho aqui apresentado teve como objetivo a produção de material de ensino para um curso de inglês instrumental para acadêmicos do curso de Tecnologia Ambiental (também denominado ESP – English for Specific Purposes), visando à preparação dos alunos para a leitura de artigos acadêmicos desta área. Para tanto, o percurso investigativo partiu do referencial metodológico e teórico da Linguística de *Corpus* com o aporte de alguns fundamentos da Linguística Cognitiva. A Linguística de *Corpus* é uma área de pesquisa que se baseia na visão probabilística da linguagem, considerando, para tanto, o registro de dados provenientes da linguagem em uso, a partir da elaboração de *corpora* de estudos que são analisados por computador. Sua metodologia contempla a produção de *corpus*, sua análise e o reconhecimento de padrões linguísticos, entre eles os pacotes lexicais e as sequências formulaicas. Além do estudo de um *corpus*, é possível, também, utilizar o instrumental da Linguística de *Corpus* para o ensino de línguas. Em vista disso, o presente estudo descreve o processo de elaboração e análise de um *corpus* composto por artigos acadêmicos da área de Tecnologia Ambiental e a subsequente utilização desses dados na elaboração de atividades de ensino.

Palavras-chave: Linguística de *Corpus*. Inglês Instrumental. Padrões linguísticos. Ensino de língua inglesa. Produção de material de ensino.

ABSTRACT

The present work was aimed at producing teaching material for an ESP (English for specific purposes) course which targeted Environmental Technology academic students. This course purpose is very focused on preparing students to read academic papers in this area. To this end, the investigative path was suported by the methodological and theoretical framework from *Corpus* Linguistics, which interfaced with some principles from Cognitive Linguistics. *Corpus* Linguistics is a study area that takes an empirical approach to language and sees it as a probabilistic system. It is based on the collection and analysis of *corpus* data criteriously selected from real language use, which could be read by computers with the aim of linguistic research. Its methodology also involves the analysis, recognition and extraction of language patterns from a *corpus*, such as the lexical bundles and the formulaic sequences. Besides the study of the *corpus* itself, *Corpus* Linguistics also offers ways of using its methodology and tools towards language teaching. Therefore, this study describes the process of elaboration and analysis of a *corpus* of academic papers in the area of environmental technology and subsequent use of such data in developing teaching activities.

Keywords: *Corpus* Linguistics. ESP (English for Specific Purposes). Language Patterns. English Language teaching. Course material design.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Não é incomum alunos de nível avançado de cursos de língua estrangeira, ao tomarem contato com textos autênticos, isto é, textos não adaptados para estudantes da língua inglesa, não os compreenderem. Estudos (SINCLAIR: 1991, FLOWERDEW, 2002b) têm evidenciado que parte dessa problemática pode estar relacionada com o material (não autêntico) utilizado para o ensino da L2. Esses materiais são, muitas vezes, produzidos para o contexto da sala de aula, adaptados às necessidades linguísticas do aprendiz de acordo com o seu nível de conhecimento. Como resultado dessa pasteurização linguística, tornam-se materiais artificiais, não representando a língua tal como é empregada em situações de comunicação real. Materiais adaptados produzem, portanto, um repertório linguístico diferente daquele da língua falada ou escrita por usuários nativos ou usuários com comunicação autêntica, num determinado contexto linguístico.

Embora os textos adaptados da língua original sejam produzidos para facilitar o aprendizado da segunda língua, os resultados que produzem, podem ser desastrosos. O processo de 'adequação linguística' faz com que a língua alvo passe por um processo de simplificação, eliminando assim algumas de suas marcas linguísticas peculiares. Os textos tornam-se sim, mais fáceis para a leitura do aluno. No entanto, o estudante da L2 acaba se habituando a essa textualidade artificial. Ao se deparar com textos autênticos, esses estudantes podem não conseguir lê-los. A língua da sala de aula, isto é, a língua idealizada e propagada por tais materiais de ensino, infelizmente, não é a mesma utilizada pelos produtores dos textos que circulam na sociedade.

Problemática semelhante atinge o ensino de língua inglesa para acadêmicos. A falta de material adequado às necessidades dos alunos, que variam, de curso para curso, faz com que professores utilizem o material que tenham em mãos, geralmente composto de textos adaptados e direcionados para um público geral, provenientes, muitas vezes, de livros de cursos gerais de língua inglesa. A distância entre os

textos autênticos versados em língua inglesa que circulam no espaço acadêmico – textos esses que os acadêmicos precisarão ler ao longo de seu percurso de estudo - se comparados com aqueles do ambiente de ensino da língua inglesa (cursos gerais) é gigantesca, pois pertencem a domínios textual, vocabular e sintático, completamente, diferentes.

O presente estudo insere-se justamente no espaço que diz respeito ao ensino de língua inglesa para acadêmicos, através da proposição de um instrumental e metodologia, para a produção de material de ensino, baseado em textos autênticos que são utilizados no meio acadêmico e científico. O ensino da língua inglesa para acadêmicos tem propósitos e objetivos diferenciados dos cursos gerais de língua inglesa. Uma das diferenças diz respeito à especialização da língua, ou seja, o ensino da língua estrangeira é atrelado a campos do conhecimento muito específicos. Tais cursos parecem estar diretamente associados a cursos de formação profissional de nível universitário ou a disciplinas muito específicas de um curso acadêmico. Além disso, cursos de língua inglesa para acadêmicos têm objetivos linguísticos muito específicos, isto é, podem ter como foco o desenvolvimento de uma determinada competência, como por exemplo, a escrita de relatórios de pesquisa, o aprendizado da linguagem utilizada para apresentações, a leitura de textos científicos, entre outras possibilidades.

O estudo aqui relatado apresenta uma proposta para a produção de material de ensino de língua inglesa para o desenvolvimento da competência leitora de acadêmicos da área de Tecnologia Ambiental. Isto é, esta proposta é duplamente específica por determinar uma única área acadêmica e, mais ainda, por limitar-se ao desenvolvimento e ensino de uma única habilidade linguística: a leitura de artigos acadêmicos. As razões dessas escolhas são detalhadas no capítulo três.

A proposta de produção de material de ensino de língua inglesa para alunos da pós-graduação em Tecnologia Ambiental, isto é, de um curso de inglês instrumental, encontrou suporte teórico e metodológico na Linguística de *Corpus*, área da Linguística Aplicada que utiliza uma abordagem empírica e vê a linguagem como um sistema probabilístico com base em análises de dados linguísticos reais, da língua em uso, através da utilização de recursos informáticos (SARDINHA, 2004). O capítulo um introduz o escopo e metodologia da Linguística de *Corpus*.

Os estudos da linguagem através da metodologia e instrumental da Linguística de *Corpus* (LC) vêm alterando a concepção da maneira como se estuda e ensina uma língua. A facilidade em coletar, armazenar e analisar quantidades massivas de textos (originais, usados em situação real de comunicação) para a formação de *corpora*, através dos dispositivos tecnológicos, permitiu que linguistas estudassem a língua por um viés empírico e probabilístico, quantificando o fenômeno linguístico, verificando frequências e recorrências de termos, entre outras possibilidades. Resultados desses estudos têm também evidenciado que a língua segue padrões que se repetem e são frequentes no uso. Os linguistas de *corpus* provam que as palavras se combinam de forma padronizada, formando agrupamentos lexicais que ocorrem numa dada configuração, com frequências significativas dentro de um *corpus* linguístico. Um dos conhecimentos mais significativos produzidos pela Linguística de *Corpus* diz respeito ao mapeamento desses agrupamentos linguísticos, conhecidos por diferentes denominações, entre elas, colocações, pacotes lexicais, sequências formulaicas, clusters, chunks. Mais ainda, a Linguística de *Corpus* demonstrou, através da análise de dados estatísticos advindos da análise de *corpora*, que os padrões seguidos pela língua são altamente recorrentes, que utilizamos unidades pré-fabricadas e padrões lexicogramaticais na comunicação, como não se imaginara até então.

Por outro lado, a Linguística Cognitiva trouxe evidências de que a fluência em uma língua pode estar diretamente ligada ao uso dos padrões linguísticos recorrentes, tais como sequências formulaicas e pacotes lexicais. A preferência de falantes nativos pelo uso das sequências formulaicas relaciona-se ao funcionamento da memória, isto é, sequências pré-fabricadas são memorizadas e articuladas, holisticamente, garantindo um processamento cognitivo mais eficiente e, em consequência, resultando na fluência linguística de toda uma comunidade. O capítulo dois discutirá algumas pesquisas da Linguística Cognitiva que apontam evidências de que a mente humana processa a linguagem a partir dos agrupamentos de palavras, contribuindo, assim, para a fluência comunicativa.

Assim, sequências formulaicas e pacotes lexicais, manifestações da lexicogramática, são objetos de estudo desta pesquisa a qual pretende, a partir da formação de um *corpus* de estudo, composto por artigos acadêmicos da área de

Tecnologia Ambiental, com o instrumental metodológico da Linguística de *Corpus*, mapear e quantificar os pacotes lexicais desse *corpus* para posterior utilização na produção de material de ensino. Resumindo, o presente estudo percorre duas jornadas: a primeira delas produz e analisa um *corpus* de textos da área de Tecnologia Ambiental; a segunda, utiliza os dados coletados e analisados desse *corpus* para a produção de material de ensino. Aspectos atinentes à metodologia utilizada para a produção e análise do *corpus* de Tecnologia Ambiental estão detalhados no capítulo quatro. Já os procedimentos envolvidos na produção do material de ensino encontram-se no capítulo cinco.

Assim, a prática de ensino de inglês instrumental aqui articulada utiliza material autêntico e busca subsídios para o ensino dos padrões linguísticos recorrentes, relevantes, extraídos de um *corpus* de estudo. Embora as fórmulas linguísticas e as estruturas pré-fabricadas ocorram simultânea e reiteradamente na comunicação, não são percebidas pelo estudante de idiomas estrangeiros, ou são completamente ignoradas por ele. Pressupõe-se que conhecê-las e reconhecê-las na leitura favoreça o desenvolvimento da fluência leitora na L2 (LANGACKER (2008), WRAY(2002), SINCLAIR(1991)), nesse caso, da língua inglesa. O ensino desses grupos lexicais, a partir da sua identificação em textos originais, pode contribuir para o aprendizado da leitura, segundo resultados de estudos psicolinguísticos, os quais são analisados no capítulo dois.

A importância do léxico para o aprendizado de uma língua estrangeira é destacado e o capítulo três discute metodologias de ensino que enfatizam o vocabulário e pensam um programa de ensino estruturado a partir do léxico. O ensino das sequências formulaicas, pacotes lexicais e outros elementos da lexicogramática, levando em conta o conhecimento advindo dos estudos da Linguística de *Corpus*, bem como da Linguística Cognitiva, são conjuntamente articulados na proposição de um programa de ensino.

Da Linguística de *Corpus* deriva uma proposta para ensino da língua, o DDL (Data Driven Learning/Aprendizagem Movida por Dados), que propõe a utilização do instrumental tecnológico e metodológico da Linguística de *Corpus* diretamente com alunos. O capítulo três diz respeito aos encaminhamentos pedagógicos e discute possibilidades de levar para a sala de aula o conhecimento produzido a partir dos

dados do *corpus* aqui elaborado, apresentando propostas condizentes para a produção de material de ensino, objetivo último deste estudo.

Propõe-se aqui pesquisar textos acadêmicos pertinentes ao curso de pós-graduação em Tecnologia Ambiental. Também se opta em investigar as questões relacionadas à leitura e vocabulário desse campo científico devido à relevância das pesquisas e tecnologias pertinentes ao meio ambiente para a sustentabilidade da vida e do mundo. Devido ao rápido avanço das tecnologias, entre outros fatores, muitos textos não chegam a ser traduzidos, sendo rapidamente atualizados ou mesmo substituídos por novos artigos que apresentam pesquisas ainda mais recentes. A atualização e renovação, proveniente de pesquisas mais e mais recentes, são constantes e parecem andar *ad-infinitum. Logo*, ler na língua inglesa é condição necessária para pesquisadores desse campo.

Embora existam diversas pesquisas e trabalhos desenvolvidos na área de Linguística Aplicada tendo como referencial a Linguística de *Corpus*, parece haver um grande espaço para pesquisas que estabeleçam uma interface entre Linguística de *Corpus* e ensino de uma segunda língua (L2) no Brasil, seja no que se refere à produção de materiais de ensino, seja à reflexão sobre as práticas pedagógicas. Até onde o sabemos, encontramos poucos exemplos de trabalhos desenvolvidos no cenário nacional que procurem articular questões sobre a cognição, o aprendizado da leitura com os pressupostos da LC, o aprendizado dos padrões formulaicos e o desenvolvimento da competência linguística. Além do mais, é visível a escassez de materiais específicos para o ensino de língua inglesa para acadêmicos, isto é, material que diga respeito ao curso ao qual o aprendiz está vinculado. Procurando sanar parte desta lacuna, este estudo é desenvolvido.

OBJETIVOS E QUESTÕES DE PESQUISA

Objetivo geral:

Utilizar referencial teórico e técnico da Linguística de *Corpus* em consonância com princípios da Linguística Cognitiva para a produção de material e propostas didáticas para o ensino de inglês instrumental para a área de Tecnologia Ambiental.

Objetivos específicos

Produzir um *corpus* pequeno (*small corpus*) composto de textos acadêmicos utilizados no programa de mestrado em Tecnologia Ambiental da UNISC a partir de metodologia proposta pela Linguística de *Corpus*.

Mapear e descrever o vocabulário e padrões linguísticos/formulaicos mais recorrentes, analisando sua frequência no *corpus*, para posterior produção de material e programa didático.

Desenhar atividades e tarefas pedagógicas, apropriando-se de recursos tecnológicos para o ensino (computadores, internet, banco de dados, *corpus*), enfatizando o reconhecimento dos pacotes lexicais mais frequentes e importantes do *corpus*.

Questões da pesquisa

A fim de atingir o objetivo proposto, foram formuladas as seguintes questões de pesquisa.

Quais são as palavras-chave do *corpus*?

Quais são os pacotes lexicais (lexical bundles) que contém palavras-chave?

Que pacotes lexicais, de maneira frequente no *corpus*, apresentam frequência mínima de 5 ocorrências e em 3 textos diferentes?

Qual o índice de cobertura dos *bundles* selecionados em relação ao *corpus*?

Qual o texto-chave deste *corpus*?

A partir do estudo linguístico dos textos que formam o *small corpus* desenvolvido nesta pesquisa, os dados são interpretados e utilizados para o

desenho de atividades e tarefas pedagógicas. Para isso, foram estabelecidas as diretrizes :

- Como explorar didaticamente os padrões lexicais mais frequentes, nas atividades propostas, em consonância com os fundamentos da Linguística Cognitiva?
- Que estratégias podem ser desenvolvidas para o ensino do vocabulário técnico para acadêmicos?
- Como utilizar e explorar os recursos tecnológicos associados à LC para oportunizar aos aprendizes maior probabilidade de exposição à língua(input)?

Respostas para as questões que nortearam esta pesquisa estão distribuídas entre os capítulos quatro e cinco. O capítulo quatro apresenta em detalhes a metodologia empregada para a elaboração do *corpus* de estudo, composto por artigos acadêmicos da área de Tecnologia Ambiental, bem como a descrição e apresentação dos dados coletados deste *corpus*. O capítulo cinco utiliza os dados obtidos do *corpus* para o desenvolvimento de tarefas e para a produção de material de ensino, buscando aplicar as questões teóricas apresentadas e discutidas.

As questões investigadas neste estudo foram articuladas em consonância com a área de concentração “Leitura e Cognição” e vincula-se à linha de pesquisa “Processos cognitivos e textualidade” do mestrado em letras desta universidade.

1 NOÇÕES SOBRE A LINGUÍSTICA DE *CORPUS* PARA A INSTRUMENTALIZAÇÃO DO ENSINO DE LÍNGUAS ESTRANGEIRAS

Neste capítulo, apresenta-se uma parte da fundamentação teórica que norteia o presente estudo. De início, procura-se esclarecer alguns princípios atinentes à Linguística de *Corpus*, seu escopo, áreas preferenciais de investigação e, principalmente, sua relação com o trabalho aqui descrito, que em conjunto com o instrumental e a metodologia aportados por essa vertente teórica são centrais para o seu desenvolvimento. Para tanto, faz-se uma compilação de questões relativas à padronização da língua, ao conhecimento específico trazido pela Linguística de *Corpus*, verificando-se, ainda, a possibilidade de utilização de um formato específico de *corpus*, o *small corpus*, para o ensino da língua inglesa.

Em prosseguimento, busca-se esclarecer como os procedimentos dessa área de investigação linguística favorecem a identificação e o estudo das unidades convencionais da língua a partir de um viés probabilístico e quantitativo. Também se abordam alguns desdobramentos práticos da LC, com ênfase na possibilidade de sua exploração no ensino de uma língua estrangeira.

1.1 Definição da Linguística de *Corpus*

Durante muitos anos os estudos línguísticos foram influenciados pelos pressupostos teóricos apregoados pela vertente racionalista, que teve como seu principal mentor Noam Chomsky. Na visão racionalista, o conhecimento tem origem em princípios estabelecidos *a priori* e “se fundamenta no estudo da linguagem por meio da introspecção, como forma de verificar modelos de funcionamento estrutural e processamento cognitivo da linguagem” (BERBER SARDINHA, 2004 p. 30). A Linguística de *Corpus*, por assumir um percurso de investigação de base empírica – empírico, significando a primazia conferida aos dados provenientes da observação da linguagem -, desde seus primórdios, contrapôs-se à visão racionalista da língua.

Nesse sentido, a língua, como objeto de análise da Linguística de *Corpus*, é estudada a partir de exemplos coletados de situações de uso real, seja da língua escrita ou da falada. O conjunto de exemplos desses textos ou seus fragmentos constitui o *corpus*.

O estatuto da Linguística de *Corpus*, aparentemente, não encontrou ainda uma definição comum compartilhada pelos agentes envolvidos em seu estudo. Segundo Berber Sardinha (2004, p. 35) a "LC não é uma disciplina tal qual a psicolinguística, sociolinguística ou semântica, pois seu projeto de pesquisa não é delimitado como em outras áreas." De fato, mostra-se mais do que uma metodologia, pois traz em seu bojo algo além do instrumental computacional, ou seja, produz um corpo de conhecimento novo. Segundo Hoey (apud BERBER SARDINHA, 2004) a "Linguística de *Corpus* não é um ramo da linguística, mas rota para a linguística". Conforme apontado por Berber Sardinha (2004), a preferência de alguns linguistas do *corpus*, entre eles, Douglas Biber, tem sido pelo termo *abordagem baseada em corpus (corpus-based approach).* Nesse sentido, a proposta de pesquisa aqui apresentada configura-se como uma *abordagem baseada em corpus.*

Em primeiro lugar, cabe ressaltar que a Linguística de *Corpus* (LC) estabelece como pré-requisito operacional a construção de um *corpus* representativo de uma determinada língua em estudo. "Um *corpus* é uma coleção de fragmentos de textos, no formato eletrônico, selecionados de acordo com um critério externo para representar, o mais amplamente possível, uma língua ou variedade linguística, formando assim uma base de dados para a pesquisa linguística[1]" (SINCLAIR, 2004[2]). A seleção de textos é feita de acordo com critérios definidos pelo pesquisador. Uma coleção de citações literárias, por exemplo, não seria necessariamente um *corpus* linguístico, a não ser que as citações fossem selecionadas e organizadas para fazer determinado estudo. Um *corpus* é considerado uma amostragem da linguagem, pois é impossível produzir um, que englobe a totalidade de variações de uma língua. Assim, o *corpus* será significativo para o propósito para o qual ele foi produzido, de vez que somente poderá conter

[1] "A *corpus* is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (SINCLAIR, 2004).
[2] Cf. o site http://ahds.ac.uk/creating/guides/linguistic-*corpora*/chapter1.htm, acesso em 06.01.10.

parte da língua em estudo e esta fração representa sua amostragem. Assim: "*Corpus* é uma coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios lingüísticos explícitos, a fim de serem usadas como uma amostra da linguagem" (PERCY, C. E; MEYER, C. F.; LANCASHIRE, I. (Orgs.) apud BERBER SARDINHA, 2004, p. 17).

Os *corpora* vinculam-se ao uso de computadores, softwares e outros dispositivos tecnológicos. De fato, o desenvolvimento da LC está diretamente relacionado ao acesso à tecnologia informática. Somente com esse desenvolvimento é que se viabilizou a construção de grandes bancos de dados, tornando possível a compilação de quantidades maiores de textos, como até agora ainda não fora possível fazer. Através da Tecnologia da Informação é possível fazer a análise e o estudo de volumes monumentais de textos. No presente, há *corpora* linguísticos do idioma inglês formados por mais de cem milhões de palavras como o British National *Corpus* (BNC)[3] e o Bank of English[4], este último contendo 450 milhões de palavras, estando, além disso, em processo de atualização e reformatação para atingir a marca de 524 milhões de palavras. O COSMAS[5], um *corpus* ainda maior, representativo da língua alemã, é composto por 1,7 bilhões de vocábulos. O Banco de Português[6] é o maior *corpus* representativo da língua portuguesa do Brasil, sendo formado por um bilhão de palavras (*tokens*) e o Banco de Português versão 2 contém 700 milhões de palavras (BERBER SARDINHA, 2004). Tais quantidades, sem o uso dos dispositivos tecnológicos, jamais poderiam ser organizadas, manuseadas, analisadas e estudadas. O computador permite que diversos tipos de análises sejam conduzidos, desde a contagem de palavras e frequência, o reconhecimento de agrupamentos de palavras recorrentes (padrões linguísticos) e as comparações entre *corpus* de diferentes línguas ou variações linguísticas, entre outras possibilidades. Os estudos realizados pela Linguística de *Corpus* a partir da interface tecnológica são preponderantemente orientados por dados (*data oriented*).

[3] http://www.natcorp.ox.ac.uk/
[4] Dados acessados em 20/04/2010 diretamente no site do Bank of English http://www.titania.bham.ac.uk/
[5] http://www.ids-mannheim.de/kl/projekte/cosmas_I/
[6] http://www2.lael.pucsp.br/*corpora*/bp/conc/index.html

Se por um lado a Tecnologia da Informação possibilitou a formação de *corpus* de proporções gigantescas, por outro lado, a democratização do computador e a criação de softwares, de vários tipos para vários fins, tornaram possível que um simples *desktop* ou *laptop* fosse suficiente para produzir e utilizar *corpus* de estudos. Dessa forma, a Linguística de *Corpus* pôde alcançar outros domínios, entre eles, a educação. O fácil acesso a softwares específicos da Linguística Computacional, muitos deles com distribuição gratuita, possibilita a educadores consultar os grandes *corpora* de instituições de pesquisas ou até mesmo produzir *corpus* linguístico para determinados objetivos. Uma das modalidades de *corpus* produzidas no campo educacional, pelos próprios professores, é a do *small corpus* (*corpus* pequeno).

1.2 *Small corpus*

Um dos critérios classificatórios de um *corpus* linguístico como pequeno (*Small corpus*), segundo Sinclair (2001), é o fato de, para determinados tipos de pesquisa, ser pequeno o suficiente para ser analisado manualmente, isto é, sem usar algoritmos computacionais, diferentemente, de *corpora* grandes, compostos por milhões ou bilhões de palavras, cuja análise os exige. Os *corpora* pequenos permitem que o próprio pesquisador analise diretamente os dados, além de não demandar uma ampla estrutura tecnológica: basta um único computador doméstico e alguns softwares. “Assim, *corpora* ‘pequeno’ e ‘grande’ são assim entendidos, enquanto contrastantes um com o outro. A diferença existente é de ordem metodológica, pois não se pode diferenciá-los considerando apenas o tamanho, seja relativo ou absoluto[7]” (SINCLAIR, 2001, p. vii). Em outras palavras, a dimensão do *corpus* está atrelada ao que é possível com ele fazer, aos procedimentos metodológicos envolvidos na sua confecção e utilização. Sobre a diferença metodológica, pontua Sinclair:

[7] (…)now “small” and “large” *corpora* are seen as in contrast with each other. So the difference must be methodological, because it cannot be just size, whether relative or absolute size. (SINCLAIR:2001 p. xi)

> Há assim um contraste claro na metodologia; *corpus* conhecidos como *Small Corpora* são aqueles projetados para receber, desde o início, a intervenção humana (early human intervention (EHI)), enquanto que *Large Corpus* (*Corpus* Grandes) são aqueles destinados à intervenção humana posterior (delayed human intervention (DHI)). Claro que no processo DHI o homem está indiretamente controlando o processo e este processo foi desenvolvido e construído através de muitas sessões de EHI (com intervenção humana) e que também o agente humano, no final, participa na interpretação dos dados[8](SINCLAIR, 2001, p. xi).

A principal técnica investigativa utilizada na maioria dos estudos com intervenção humana (EHI) e utilização de *Small corpus* é a da comparação (SINCLAIR, 2001). A comparação de diferentes gêneros textuais, após a análise dos dados realizada pelo computador, é um bom exemplo de análise manual. Essa abordagem possibilita a descoberta de diferenças, através da interpretação dos dados obtidos. A comparação e análise manual, ou seja, a interpretação de dados também pode ser utilizada no estudo de dois *corpora* paralelos. *Corpora* paralelos (FRANKENBERG-GARCIA, 2008) contêm o mesmo material linguístico em, pelo menos, duas línguas diferentes, possibilitando que o usuário verifique como a língua alvo foi traduzida, ou vice-versa. Essa modalidade de *corpus*, entre outras existentes, pode ser explorada pelo professor de idiomas.

Além disso, o *small corpus* apresenta um diferencial importante no que se refere a sua possibilidade de uso no ensino: possibilita a compilação e organização de *corpus* formado por textos específicos de uma determinada área do conhecimento. A montagem e a organização do *small corpus* pode atingir o grau de especificidade que for necessário, sendo possível, por exemplo, elaborar um *corpus* de textos de uma determinada disciplina, ou de disciplinas de uma mesma linha de pesquisa, de um mesmo autor, de um mesmo período literário ou histórico, entre outros tantos critérios. “Normalmente, *corpora* compilados em pequena escala por pesquisadores individuais acabam sendo mais representativos do que os respectivos sub*corpora* dos *corpora* gerais” (BERBER SARDINHA, 2004, p. 28).

[8] There is thus a fairly sharp contrast in method; the so-called Small *Corpora* are those designed for early human intervention (EHI) while the Large *Corpora* are designed for late or delayed human intervention (DHI). (Of course in DHI the human being is indirectly controlling the process, and the process has probably been built up over many EHI sessions, and the human being must eventually participate in order do interpret the results.)(SINCLAIR:2001 p. xi)

Adicionalmente, Ghadessy, Henry e Roseberry (2001) fazem um comentário sobre o uso de *small corpus* no ensino:

> (...) ao focar em determinados gêneros textuais e os propósitos inerentes a eles, nos é oferecida uma janela para investigar os mecanismos da língua. Com ajuda de computadores podemos observar por nós mesmos o vocabulário, funções e padrões dos textos produzidos para propósitos específicos. Ao envolver os estudantes nesse processo, podemos ajudá-los a se tornarem mais conscientes dos padrões linguísticos que estão tentando aprender[9] (GHADESSY, HENRY & ROSEBERRY, 2001, p. xxii-xxiii).

O ensino da leitura instrumental da língua inglesa no meio acadêmico muitas vezes vincula-se a disciplinas específicas da grade curricular do curso ao qual o aluno está vinculado. Nesse sentido, a organização de um *small corpus* formado por uma coletânea de textos específicos de uma área do conhecimento, como o de Tecnologia Ambiental que aqui será apresentada, possibilita que no processo de aprendizagem, a atenção seja focalizada em aspectos da língua pertinentes ao campo conceitual em que os aprendizes estão interessados. Assim, o estudo do *small corpus* montado para este fim possibilita que o aprendiz verifique *in loco* e por si mesmo padrões recorrentes da língua, vocabulário específico, estruturas ligadas à movimentação do texto, isto para citar algumas possibilidades. Em outras palavras, *o small corpus* viabiliza que os estudantes sejam expostos à linguagem que provavelmente utilizarão em sua área científica de estudos, de trabalho ou de pesquisa, no próprio ambiente de estudo (FLOWERDEW, 2001).

A utilização de textos autênticos, não-adaptados, é um dos pressupostos gerais da LC de extrema relevância para o ensino de *English for Specific Purposes (ESP).* A seleção de exemplos reais possibilita a apresentação acurada da linguagem daquele nicho. Flowerdew (2001) afirma que a utilização de material linguístico produzido, artificialmente, para ensino representa um perigo para o aprendiz, pois apresenta aos alunos um modelo distorcido da linguagem em uso. Sobre isso, de forma muito incisiva Flowerdew (2001) retoma um ponto já destacado por Sinclair,

[9] "(…) by focusing on types of texts and the purposes they are intended to achieve, offers us a window to understanding into the works of language. In part with the aid of computers, we can observe for ourselves the vocabulary, functions and discourse patterns of texts created for specific purposes. By involving our students in this process, we can help them to raise their awareness of the language patterns they are trying to learn" GHADESSY, Mohsen; HENRY, Alex & ROSEBERRY, Robert L (2001), p. xxii-xxiii)

asseverando que “claramente, os materiais didáticos disponíveis no mercado, ao apresentarem modelos linguísticos não-autênticos, estão fazendo um desserviço aos aprendizes e precisam ser substituídos por materiais que apresentem um modelo mais preciso de como as definições são expressas[10]” (FLOWERDEW, 2001, p. 83). Nesse âmbito e para evitar situações semelhantes, a produção e a organização de um *small corpus* constituído por textos da área acadêmica em foco, criteriosamente selecionados, pode ser uma alternativa de ensino valiosa. Outro pesquisador, Robert de Beaugrande (2001), reconhece o problema gerado pelo uso de materiais que não apresentam a língua adequadamente, alertando que “nosso maior problema não é somente o ‘inglês ruim ou incorreto’ como frequentemente se lamenta, mas muito mais, a quantidade insuficiente de inglês”. Neste mesmo excerto salienta que “os usos de *corpora* são, com certeza, mais necessários e urgentes para falantes não-nativos que não tiveram extensiva exposição ao inglês fluente[11]” (DE BEAUGRANDE, 2011, p. 11). Logo, o uso de *corpus* não só pode como deve ser feito no ensino de inglês instrumental, inclusive, como uma alternativa de o aprendiz ter acesso ao input linguístico.

Assim, o *small corpus* produzido para a educação seria a base de referência da língua a ser utilizada pelos aprendizes e professor(es) como fonte de consulta e pesquisa. A partir desse *corpus*, o professor teria condições de investigar o comportamento da língua da área delimitada, planejar atividades didáticas derivadas dos dados ali contidos. Os alunos, por sua vez, poderiam recorrer ao *small corpus* para tomar contato com exemplos autênticos de uso da língua.

A proposta de elaboração de um *small corpus* para esta pesquisa teve razões diversas. Primeiramente, esse formato de *corpus* é viável de ser produzido pelo professor-pesquisador utilizando-se de computadores comuns, como desktops e notebooks de utilização doméstica, sem exigência de configurações especiais, além da possibilidade de utilização de softwares com ampla distribuição. Segundo, a dimensão de um *small corpus* permite o fácil manuseio dos dados, tanto pelo

[10] “Clearly, the commercially available materials, in presenting an inauthentic formula are doing a disservice to learners and need to be replaced by materials presenting a more accurate model of how definitions are expressed”. (John Flowerdew:2001, p. 83)

[11] “The uses of *corpora* are surely most urgent for non-native speakers who have not had extensive exposure to fluent English. Our major problem is so much not *bad English* or *incorrect English*, as is often lamented, but rather *insufficient English.* (**BEAUGRANDE, Robert de.** p. 11)

professor, quanto pelos alunos quando for explorado para fins pedagógicos, podendo ser também facilmente distribuído entre alunos para o utilizarem quando necessitarem. Além disso, um conjunto pequeno de textos especializados - tal como a compilação aqui realizada contendo 86 artigos científicos - e com temática muito específica não requer a elaboração de um grande *corpus*. Esse tamanho de *corpus* possibilita que as perguntas desta pesquisa sejam respondidas. Assim, a escolha do formato *small corpus* para este estudo ocorreu pela fácil operacionalização do mesmo e pelas possibilidades de uso pedagógico.

Adiante, em capítulo específico desta dissertação, serão apresentadas possibilidades de uso de *corpus* linguístico em educação. Considera-se que a utilização da Linguística de *Corpus* no meio educacional, seja para conhecer e empregar seu instrumental e modelo teórico na produção e análise linguística de um *corpus,* seja para usar as bases de dados como material de referência, pode trazer contribuições significativas para o ensino de segunda língua. Sinclair (1991) exemplifica a dificuldade de muitos aprendizes em utilizar adequadamente as expressões comuns de uma língua, tais como expressões idiomáticas, *phrasal verbs* e outras fórmulas linguísticas. Segundo ele, os aprendizes evitam utilizá-las e muitas vezes as substituem por perífrases confusas, muito grandes, que soam estranhas e rompem com a naturalidade da língua. Conforme pondera, isso "não é certamente responsabilidade deles (dos alunos, grifo meu), também não é responsabilidade do professor, que somente pode trabalhar e acessar o modelo de descrição da língua que está disponível[12]" (SINCLAR, 1991, p. 79). Nesse sentido, produzir um *small corpus* e utilizá-lo para o ensino de idiomas pode ser uma forma de agregar qualidade e de tornar mais significativo o ensino, bem como de fornecer, aos alunos, material didático mais coerente e focado em suas necessidades.

1.3 Trust the text: Probabilidade, estatística e representatividade

Por ser orientado pelos dados, o uso de computadores permitiu que os fenômenos linguísticos fossem observados de forma muito mais objetiva, sendo

[12] This is certainly not their fault, nor is the fault of the teacher, who can only work within the kind of language descriptions that are available" (SINCLAIR, 1991, P. 79).

esse uso favorecido pela utilização dos algoritmos computacionais. Embora já houvesse pesquisas focadas na análise de *corpus* linguísticos sem a utilização da tecnologia da informação, tal como o *corpus* de Thorndike composto por 18 milhões de palavras, estudos pré-computacionais de *corpora* gigantescos recebiam críticas contundentes por não serem confiáveis, “pois o ser humano não é talhado para tarefas desse tipo” (BERBER SARDINHA, 2004, p. 4). De fato, a expansão das pesquisas em Linguística de *Corpus* disparou nos anos 1980, com os microcomputadores. A pesquisa sobre a língua, com um viés empírico, tal como enfatizado pela LC, passou a ter maior credibilidade a partir das evidências aportadas via estabelecimento de relações entre probabilidade de ocorrência e representatividade. Conforme refere Berber Sardinha:

> A representatividade está ligada à questão da probabilidade. A linguagem tem caráter probabilístico, conforme dito, havendo a possibilidade de estabelecer uma relação entre traços que são mais e menos comuns em determinado contexto. O conhecimento da probabilidade de ocorrência de traços lexicais, estruturais, pragmáticos e discursivos está no cerne da Linguística de *Corpus* e, portanto, o conhecimento acerca da probabilidade de ocorrência ainda está sendo adquirido. (BERBER SARDINHA, 2004, p. 23-24)

Ademais, a utilização de textos naturais ou autênticos para estudo linguístico, como visto, é outro dos pressupostos da Linguística de *Corpus*, assim como de toda a linguística que concebe a linguagem como vinculada à práxis, ao seu uso em situações reais. Entende-se aqui como sendo textos naturais/autênticos todos aqueles que não foram produzidos artificialmente para figurarem no *corpus* (BERBER SARDINHA, 2004). O *corpus*, segundo o entendimento da LC, seria composto por gêneros textuais utilizados na sociedade em situações reais de comunicação, coletados e organizados de acordo com o propósito do estudo que motivou a sua criação. Há, no entanto, uma exceção: os *corpora* linguísticos formados pela produção textual de estudantes de uma segunda língua, não sendo, pois, compostos por textos provenientes de exemplos de língua nativa (GRANGER, 1998). No entanto, essa modalidade é construída, a fim de possibilitar pesquisas pertinentes ao desenvolvimento linguístico de estudantes, conhecer pontos da língua

onde encontram maior dificuldade, identificar diferenças entre usos da língua materna em comparação à segunda língua, entre outros fenômenos linguísticos.

A opção pela seleção e análise de textos autênticos em contraposição à utilização de exemplos inventados para o estudo da língua e a utilização da tecnologia computacional como catalisadora, para a análise probabilística da língua, colaboraram para que a linguística se distanciasse de um percurso de investigação baseado na intuição. Conforme destaca Sinclair, "exemplos inventados iriam buscar a validação da sua autenticidade em um contexto que simplesmente não existe e que, no final das contas, seriam avaliados pela intuição de alguém acompanhada de todos os equívocos provenientes de tal ato[13]" (SINCLAIR, 1991, p. 5). Sinclair cunhou a expressão "trust the text/confie no texto" como um mote dessa visão empírica de análise da língua. Segundo ele, falhas podem ocorrer no processo de investigação científica baseado na intuição e o caminho reverso é percorrido pela LC: "sem abandonar nossas intuições, tentamos encontrar explicações que se encaixam à evidência em vez de ajustá-la para se encaixar em uma explicação prévia elaborada[14]"(SINCLAIR, 1991, p. 36). Ou seja, a partir da análise do texto natural - a fonte da evidência linguística -, juntamente com a utilização do instrumental tecnológico, são feitas algumas constatações com base nos dados, as quais depois de detectadas, são postuladas. Na mesma obra, Sinclair sintetiza seu ponto de vista e conclui ressaltando o papel negativo da intuição na pesquisa linguística, ao afirmar que "o problema relativo a todos os tipos de introspecção é que nenhum deles fornece evidência sobre o uso[15]" (SINCLAIR, 1991, p. 39).

A análise de *corpora* compostos de textos autênticos possibilita a identificação de padrões recorrentes da língua, trazendo à tona, conhecimentos inovadores sobre o comportamento da linguagem. A partir da abordagem da LC é possível realizar mapeamentos, evidenciando a existência de grupos lexicais com funções específicas, os quais até então eram ignorados e mesmo desconhecidos pelo o

[13] Invented examples would, therefore, appeal for their authenticity to a non-existent context, which would eventually be evaluated by someone´s intuition, with all misleading consequences of that (SINCLAIR, 1991, p. 5).
[14] "without relinquishing our intuitions, of course, we try to find explanations that fit the evidence, rather than adjusting the evidence to fit a pre-explanation" (SINCLAIR, 1991, p. 36).
[15] The problem about all kinds of introspection is that it does not give evidence about usage"(SINCLAIR, 1991, p. 39).

senso e observação comuns. Além disso, a LC também vem possibilitando diversos questionamentos sobre pressupostos anteriormente estabelecidos acerca do funcionamento linguístico. Assim, conforme Kennedy (1998) "a análise de um *corpus* pode revelar, e frequentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar[16]" (Kennedy, 1998, p. 9). A comprovação, a partir da análise dos dados do *corpus*, do alto grau de recorrência de certos padrões dentro de um sistema linguístico tem sido um dos resultados advindos de pesquisas conduzidas pela LC. Assim, a língua passou a ser vista e entendida como um fenômeno probabilístico, recorrente e não-aleatório. Conforme o afirma Berber Sardinha:

> Dizer que a variação não é aleatória, na verdade, é afirmar que a linguagem é padronizada (patterned). A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura que se repete significativamente mostra sinais de ser, na verdade, um padrão lexical ou léxico-gramatical. A linguagem forma padrões que apresentam regularidade (estáveis em momentos distintos, isto é, têm frequência comparável em *corpora* distintos) e variação sistemática (correlacionam-se com variedades textuais, genéricas, dialetais etc.) (BERBER SARDINHA, 2004, p. 31)

1.4 Ferramentas Computacionais do pacote Wordsmith Tools

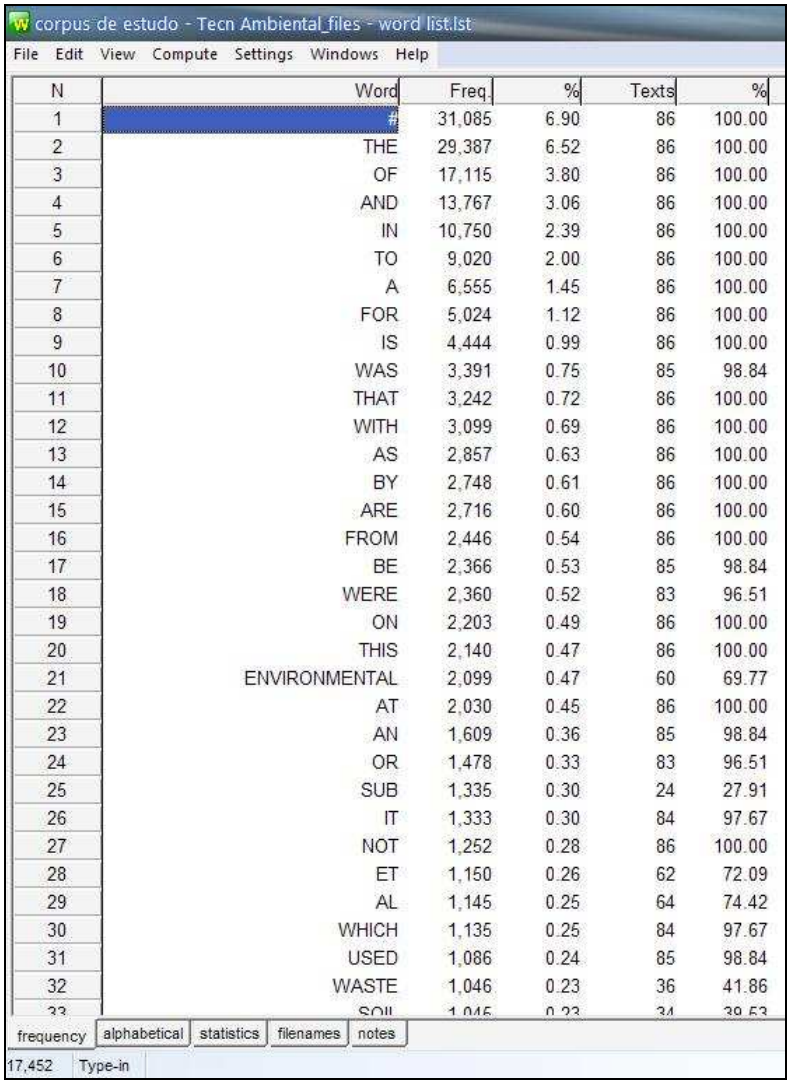
A Linguística de *Corpus* faz uso de uma série de softwares com funções e propósitos diferenciados. O Wordsmith Tools 5.0[17] (Scott, 2010), software a ser utilizado na condução deste estudo, é uma suíte de programas para a realização da análise linguística de *corpora.* "Esse software permite fazer análises baseadas na frequência e na co-ocorrência de palavras em *corpora*" (BERBER SARDINHA, 2009, p. 8). É composto por três ferramentas: Wordlist, Concord e KeyWords.

O Wordlist possibilita que sejam elaboradas listas de frequência das palavras do *corpus* selecionado. Assim, é possível listar todas as palavras e verificar suas frequências absolutas e percentuais, podendo-se, ainda, determinar quais são mais importantes, mais recorrentes, quais tendem a aparecer em determinado texto e não em outro, entre outras possibilidades de análise. Também é possível ordenar a lista

[16] Tradução de Tony Berber Sardinha (BERBER SARDINHA, 2004, p. 37).
[17] http://www.lexically.net/wordsmith/

alfabeticamente. A figura 1 mostra o exemplo de uma lista de palavras, elaborada com o Wordlist e ordenada segundo as frequências.

Figura 1 – Lista de Frequências da ferramenta Wordlist



corpus de estudo - Tecn Ambiental_files - word list.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	#	31,085	6.90	86	100.00
2	THE	29,387	6.52	86	100.00
3	OF	17,115	3.80	86	100.00
4	AND	13,767	3.06	86	100.00
5	IN	10,750	2.39	86	100.00
6	TO	9,020	2.00	86	100.00
7	A	6,555	1.45	86	100.00
8	FOR	5,024	1.12	86	100.00
9	IS	4,444	0.99	86	100.00
10	WAS	3,391	0.75	85	98.84
11	THAT	3,242	0.72	86	100.00
12	WITH	3,099	0.69	86	100.00
13	AS	2,857	0.63	86	100.00
14	BY	2,748	0.61	86	100.00
15	ARE	2,716	0.60	86	100.00
16	FROM	2,446	0.54	86	100.00
17	BE	2,366	0.53	85	98.84
18	WERE	2,360	0.52	83	96.51
19	ON	2,203	0.49	86	100.00
20	THIS	2,140	0.47	86	100.00
21	ENVIRONMENTAL	2,099	0.47	60	69.77
22	AT	2,030	0.45	86	100.00
23	AN	1,609	0.36	85	98.84
24	OR	1,478	0.33	83	96.51
25	SUB	1,335	0.30	24	27.91
26	IT	1,333	0.30	84	97.67
27	NOT	1,252	0.28	86	100.00
28	ET	1,150	0.26	62	72.09
29	AL	1,145	0.25	64	74.42
30	WHICH	1,135	0.25	84	97.67
31	USED	1,086	0.24	85	98.84
32	WASTE	1,046	0.23	36	41.86
33	SOIL	1,046	0.23	34	39.53

frequency | alphabetical | statistics | filenames | notes

17,452 Type-in

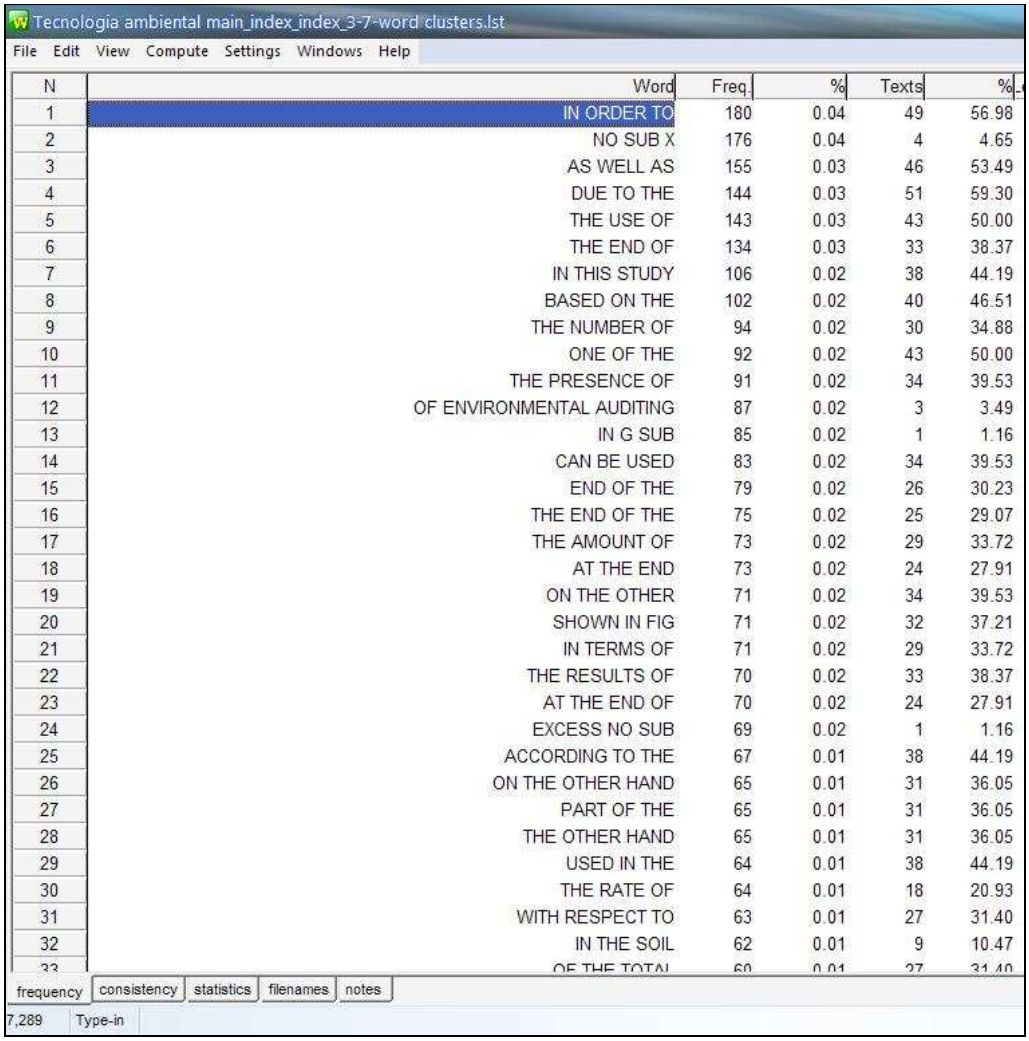
O Wordlist também gera listas de *clusters,* ou seja, "um grupo de palavras consecutivas em um texto[18]" (Scott, 2010, p. 307), a partir de critérios estabelecidos pelo pesquisador. Podem-se localizar clusters compostos de duas a oito palavras que co-ocorrem e o software informa a quantidade de ocorrências, percentagem no *corpus*, em quantos e quais textos do *corpus* a multipalavra está presente, entre

[18] "A cluster is a group of words which follow each other in a text. (Scott, 2010, p. 307)

outras informações. A figura 2 apresenta uma tela do Wordlist com a seleção dos primeiros e mais frequentes *clusters* detectados pelo software. Na pesquisa deste exemplo, foi solicitado ao software que localizasse *clusters* compostos de três a sete palavras. A lista da figura 2 apresenta os 32 clusters mais frequentes de um total de 7289.

O termo *cluster* é utilizado amplamente na literatura sobre Linguística de *Corpus* e parece ter como principal divulgador Mike Scott. Douglas Biber, para referir-se à co-ocorrência de termos na língua utiliza o termo *lexical bundle,* o qual tem sido traduzido para a língua portuguesa como pacote lexical. Há diferenças sutis no significado entre um e outro termo, no entanto, ambos referem-se ao mesmo fenômeno linguístico: da co-ocorrência de termos na língua. Outras denominações são utilizadas por diferentes autores para referir-se a esse fenômeno linguístico. Quando houver necessidade da utilização, ao longo deste texto, de outro termo além da trinca *cluster*, pacote lexical (*lexical bundles*) e sequência formulaica (*formulaic sequence*), esses serão introduzidos e acompanhados de devida explicação, se for o caso. No entanto, quando determinado termo for parte do repertório de um autor citado, será respeitada a terminologia por ele adotada. Por questões metodológicas, a serem discutidas em capítulo específico, adota-se, ao longo desta dissertação, a utilização da expressão pacote lexical e sequência formulaica. O próximo capítulo explicitará as diferenças entre as duas denominações escolhidas e razões que levaram a sua escolha.

Figura 2 – Lista de *clusters* da ferramenta Wordlist



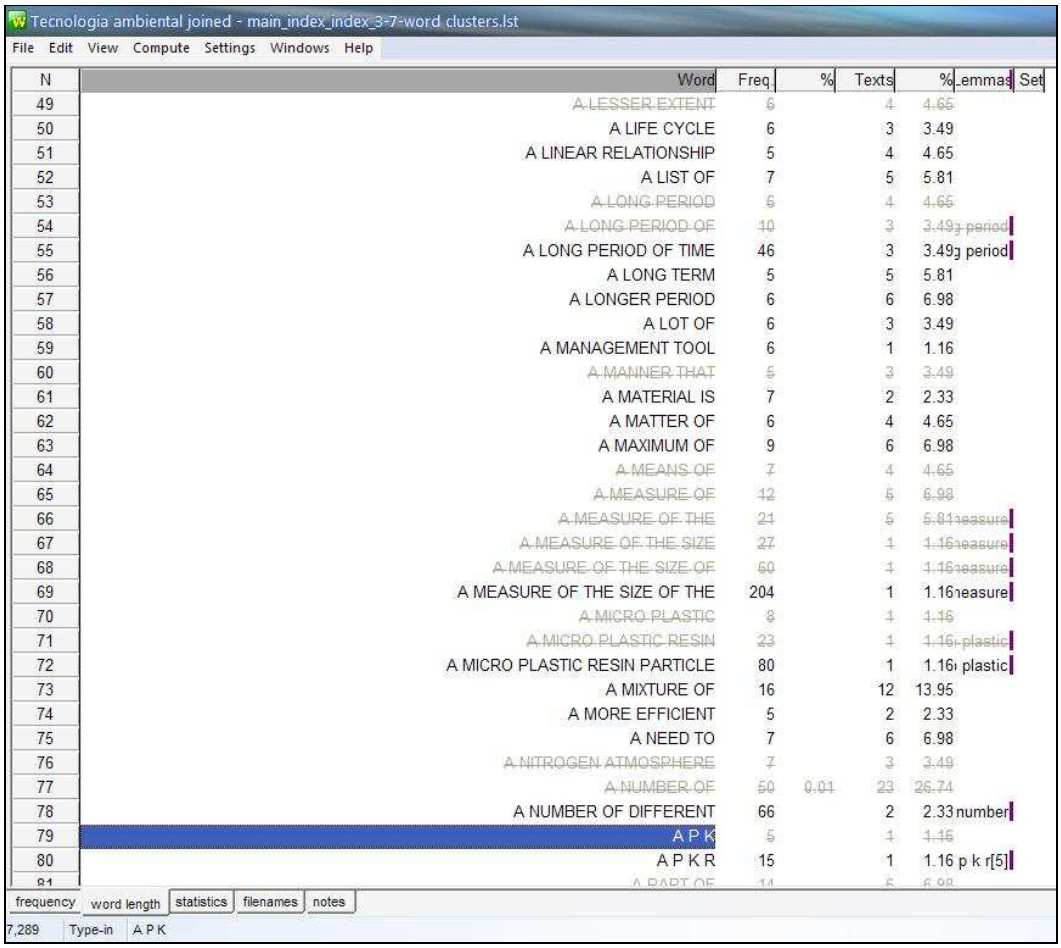
N	Word	Freq.	%	Texts	%
1	IN ORDER TO	180	0.04	49	56.98
2	NO SUB X	176	0.04	4	4.65
3	AS WELL AS	155	0.03	46	53.49
4	DUE TO THE	144	0.03	51	59.30
5	THE USE OF	143	0.03	43	50.00
6	THE END OF	134	0.03	33	38.37
7	IN THIS STUDY	106	0.02	38	44.19
8	BASED ON THE	102	0.02	40	46.51
9	THE NUMBER OF	94	0.02	30	34.88
10	ONE OF THE	92	0.02	43	50.00
11	THE PRESENCE OF	91	0.02	34	39.53
12	OF ENVIRONMENTAL AUDITING	87	0.02	3	3.49
13	IN G SUB	85	0.02	1	1.16
14	CAN BE USED	83	0.02	34	39.53
15	END OF THE	79	0.02	26	30.23
16	THE END OF THE	75	0.02	25	29.07
17	THE AMOUNT OF	73	0.02	29	33.72
18	AT THE END	73	0.02	24	27.91
19	ON THE OTHER	71	0.02	34	39.53
20	SHOWN IN FIG	71	0.02	32	37.21
21	IN TERMS OF	71	0.02	29	33.72
22	THE RESULTS OF	70	0.02	33	38.37
23	AT THE END OF	70	0.02	24	27.91
24	EXCESS NO SUB	69	0.02	1	1.16
25	ACCORDING TO THE	67	0.01	38	44.19
26	ON THE OTHER HAND	65	0.01	31	36.05
27	PART OF THE	65	0.01	31	36.05
28	THE OTHER HAND	65	0.01	31	36.05
29	USED IN THE	64	0.01	38	44.19
30	THE RATE OF	64	0.01	18	20.93
31	WITH RESPECT TO	63	0.01	27	31.40
32	IN THE SOIL	62	0.01	9	10.47
33	OF THE TOTAL	60	0.01	27	31.40

Convém salientar que nem todo *cluster* é necessariamente uma expressão recorrente da língua. Ou seja, “o que é importante no contexto de análise de *corpus*, usando o Wordsmith Tools, é que os *clusters* têm sua gênese como um fenômeno puramente distributivo” (SCOTT E TRIBBLE, 2006, p. 131). Por isso, no momento de sua detecção pelo software, eles não passam de uma sequência de palavras que pode, ou não, ter sentido na língua. Como exemplo, na figura 3, o cluster *‘a material is’* não é uma sequência formulaica da língua, pois não é um padrão de linguagem recorrente e também parece não fazer sentido. Ao contrário, o *cluster ‘a long period of time’* apresenta grande probabilidade de ser uma fórmula linguística reconhecida, pois possui alta frequência no *corpus* analisado, podendo ser um padrão linguístico

consagrado, pelo menos, pela comunidade acadêmica leitora dos artigos que compõem o *corpus* de estudo.

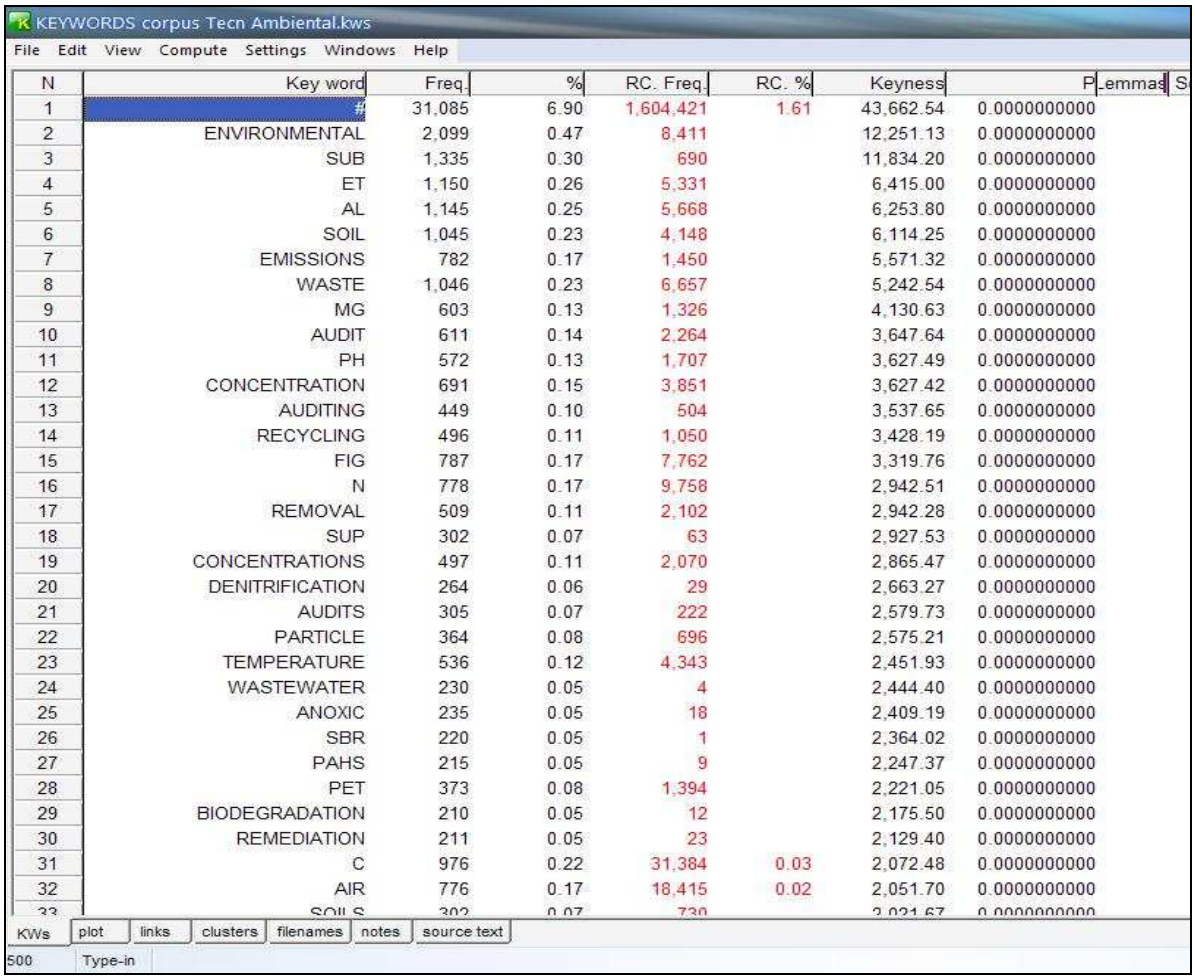
A detecção final das sequências dotadas de sentido é um trabalho realizado a partir da análise dos dados apresentados pelo software, a que se agrega, posteriormente, o exercício crítico e interpretativo do linguista. Outro recurso da ferramenta Wordsmith Tools, denominado *joining clusters,* permite que *clusters,* com sequências repetidas e semelhantes, sejam aglutinados, para que mais facilmente seja detectado o padrão recorrente. A figura 3 representa o processo denominado de *joining clusters* (aglutinação de clusters). Observe que as sequências de palavras que aglutinam o maior número de termos estão em destaque e podem ter maior probabilidade de serem padrões linguísticos recorrentes utilizados na língua. Os clusters eliminados, no processo de detecção e aglutinação dos *"joining clusters",* estão riscados.

Figura 3 – Lista de *clusters* após processamento da função 'joined clusters' da ferramenta Wordlist.



Assim, o *KeyWords* possibilita que sejam extraídas e organizadas, em uma lista, as palavras-chave de um *corpus*, sendo a extração feita a partir da comparação com um *corpus* de referência. O *KeyWords* faz a listagem das palavras-chave a partir da extração das "palavras de uma lista cujas frequências são estatisticamente diferentes (maiores ou menores) do que as frequências das mesmas palavras num outro *corpus* (de referência)" (BERBER SARDINHA, 2009, p. 9). A partir dessa comparação é possível estabelecer que termos têm maior destaque para um determinado contexto ou área do conhecimento, por exemplo. A figura 4 ilustra um exemplo de uma lista de palavras-chave da área de Tecnologia Ambiental elaborada com o *KeyWords*.

Figura 4 – Lista de palavras-chave na ferramenta KeyWords



N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P_emmas	S
1	#	31,085	6.90	1,604,421	1.61	43,662.54	0.0000000000	
2	ENVIRONMENTAL	2,099	0.47	8,411		12,251.13	0.0000000000	
3	SUB	1,335	0.30	690		11,834.20	0.0000000000	
4	ET	1,150	0.26	5,331		6,415.00	0.0000000000	
5	AL	1,145	0.25	5,668		6,253.80	0.0000000000	
6	SOIL	1,045	0.23	4,148		6,114.25	0.0000000000	
7	EMISSIONS	782	0.17	1,450		5,571.32	0.0000000000	
8	WASTE	1,046	0.23	6,657		5,242.54	0.0000000000	
9	MG	603	0.13	1,326		4,130.63	0.0000000000	
10	AUDIT	611	0.14	2,264		3,647.64	0.0000000000	
11	PH	572	0.13	1,707		3,627.49	0.0000000000	
12	CONCENTRATION	691	0.15	3,851		3,627.42	0.0000000000	
13	AUDITING	449	0.10	504		3,537.65	0.0000000000	
14	RECYCLING	496	0.11	1,050		3,428.19	0.0000000000	
15	FIG	787	0.17	7,762		3,319.76	0.0000000000	
16	N	778	0.17	9,758		2,942.51	0.0000000000	
17	REMOVAL	509	0.11	2,102		2,942.28	0.0000000000	
18	SUP	302	0.07	63		2,927.53	0.0000000000	
19	CONCENTRATIONS	497	0.11	2,070		2,865.47	0.0000000000	
20	DENITRIFICATION	264	0.06	29		2,663.27	0.0000000000	
21	AUDITS	305	0.07	222		2,579.73	0.0000000000	
22	PARTICLE	364	0.08	696		2,575.21	0.0000000000	
23	TEMPERATURE	536	0.12	4,343		2,451.93	0.0000000000	
24	WASTEWATER	230	0.05	4		2,444.40	0.0000000000	
25	ANOXIC	235	0.05	18		2,409.19	0.0000000000	
26	SBR	220	0.05	1		2,364.02	0.0000000000	
27	PAHS	215	0.05	9		2,247.37	0.0000000000	
28	PET	373	0.08	1,394		2,221.05	0.0000000000	
29	BIODEGRADATION	210	0.05	12		2,175.50	0.0000000000	
30	REMEDIATION	211	0.05	23		2,129.40	0.0000000000	
31	C	976	0.22	31,384	0.03	2,072.48	0.0000000000	
32	AIR	776	0.17	18,415	0.02	2,051.70	0.0000000000	
33	SOILS	302	0.07	730		2,021.67	0.0000000000	

Por sua vez, os concordanciadores são programas que possibilitam a extração e visualização das ocorrências de uma palavra ou de um *cluster* em um *corpus. O concord* é um software que permite que palavras ou *clusters* selecionados, provenientes dos textos que compõem o *corpus*, sejam visualizados, em destaque, no contexto em que ocorrem, isto é, ladeados por alguns dos termos que sequencialmente os antecedem ou sucedem.

A palavra ou expressão pesquisada aparece centralizada, em destaque, acompanhada de co-textos. Se, por exemplo, no *corpus*, a palavra pesquisada foi *environmental,* são apresentadas as concordâncias, cada uma tendo a palavra *environmental* centralizada e cercada pelos co-textos relativos a cada uma das ocorrências, em quantidade e disposição (à direita e à esquerda da palavra destacada) definidas pelo usuário do programa. A palavra em destaque é denominada KWIC (Key Word in context/ palavra-chave em contexto). Dessa forma, todas as palavras-chave ou KWIC são alinhadas. Esta é a convenção para a apresentação da palavra-chave utilizada pela maioria dos softwares concordanciadores, pois segundo Tribble & Jones (1990), ela facilita o estudo do contexto imediato da palavra-chave em destaque. A figura 5 mostra um exemplo da palavra *environmental* pesquisada através do uso do concordanceador.

Assim sendo, é o concordanciador a ferramenta computacional utilizada para estudar as fraseologias e padrões de uma dada língua, porque ele permite que os textos sejam rearranjados de maneira a possibilitar a percepção e a comparação dos padrões linguísticos encontrados em um *corpus* de estudo, os quais, de outra maneira, não seriam visíveis. No exemplo ilustrado na figura 5, cópia da tela de apresentação de uma pesquisa realizada no Concord, a análise das primeiras linhas possibilita constatar que a palavra *environmental* ocorre com grande frequência com os termos *auditing* e *clearance*, por exemplo. Tal evidência enseja acrescentar que:

> Um software concordanciador permite que se descubram os padrões existentes na língua natural pelo agrupamento dos textos, de maneira tal, que tais padrões tornam-se visíveis. Esses padrões são características importantes da língua, mas sempre foi um problema muito difícil conseguir isolá-los. O verdadeiro valor do concordanceador reside nas possibilidades

de visualização (dos padrões linguísticos, grifo meu) apresentadas[19] (Tribble e Jones, 1990, p. 9).

Figura 5 – Linhas de concordância na ferramenta Concord

[19] Concordancing software enables you to discover patterns that exist in natural language by grouping text in such a way they are clearly visible. These patterns are important feature of language but the problem has always been that it is extremely difficult to isolate them. The real value of the concordancer lies in this question of visibility (Tribble e Jones, 1990, p. 9).

A partir da utilização do conjunto de softwares que compõe o *Wordsmith Tools 5.0* (SCOTT, 2010), é viável realizar o estudo de uma dada língua ou variedade linguística, gênero textual, registro, etc., com base na análise das frequências, recorrências, identificação de palavras-chave, reconhecimento de padrões linguísticos e de expressões usuais. Pela utilização dessas ferramentas, novas formas de conhecimento a respeito da dinâmica e das construções da língua podem ser identificadas e mapeadas.

1.5 Padrões de Linguagem

Prosseguindo com as ponderações de Berber Sardinha (2004), a busca de evidências de que o léxico é padronizado, da regularidade existente entre determinados agrupamentos de vocábulos e das associações que se estabelecem entre eles, é foco de interesse dos linguistas de *corpus*. E é justamente a investigação que advém da atenção dada aos padrões da língua, parte da peculiar contribuição da LC ao conhecimento sobre o funcionamento e dinâmica de uma língua. “O aspecto mais interessante do processamento de dados de longos textos não é, no entanto, o espelhamento das categorias intuitivas de descrição. É a possibilidade de novas abordagens, novas formas de evidências, e novas possibilidades de descrição[20]” (SINCLAIR, 1991, p. 36). As *colocações* e *coligações* - duas formas de padronização lexical - são exemplos de construções estáveis, identificadas a partir do instrumental de investigação da Linguística de *Corpus* e, devido a sua relevância, são em seguida detalhadas.

As colocações referem-se à co-ocorrência de palavras de significado, de combinações lexicais consagradas. “O termo *collocation* foi introduzido pelo linguista britânico J. R. Firth para designar casos de co-ocorrência léxico-sintática, ou seja, palavras que usualmente “andam juntas”. “Há casos em que essa co-ocorrência é extremamente restritiva, ou seja, tem alto grau de fixidez”(TAGNIN, 2005, p. 37). Por exemplo, não há regra que defina por que a expressão *cão e gato* seja assim

[20] “The most exciting aspect of long-text data-processing, however, is not the mirroring of intuitive categories of description. It is the possibility of new approaches, new kinds of evidence, and new kinds of description”(SINCLAIR, 1991, p. 36).

apresentada, ao invés de *gato e cão.* É o uso que consagra muitas colocações. Há várias classificações de colocações, entre elas: colocações adjetivas (main course/prato principal), colocações nominais (baking powder/fermento em pó), colocações verbais (make trouble /criar problema), colocações adverbiais em que o advérbio modifica o adjetivo (deeply offended / profundamente ofendido) e aquelas em que o advérbio modifica o verbo (love blindly/amar cegamente). Expressões especificadoras de unidade (a block of ice/um cubo de gelo), coletivos (a flight of birds/ um bando de pássaros) e binômios (profit and loss / lucros e perdas; bit by bit /pouco a pouco) são também exemplos de colocações. "O fenômeno da colocação é o mais tradicionalmente enfocado no estudo de *corpus*. Os pacotes lexicais (bundles/clusters) são também padrões colocacionais e, devido sua centralidade neste estudo, serão discutidos detalhadamente e com aprofundamento no decorrer desta dissertação.

As coligações, de sua parte, referem-se a uma combinação gramatical, em que o colocado é necessariamente uma expressão gramatical. Segundo Berber Sardinha (2004) é uma associação que ocorre entre itens lexicais e gramaticais. Por exemplo, expressões em português como "confiar *em*" e na língua inglesa "to be good *at*" (ser bom em) são exemplos de coligações nesses dois idiomas. As coligações de regência incluem todos os tipos de regência, ou seja, termos necessariamente seguidos de preposição, podendo ser substantivos (obedience *to*/ obediência *a*), adjetivos (good *at*/bom *em*), verbos (depend *on*/confiar *em*) e advérbios (because of/por causa de). Phrasal verbs (verbos frasais) ou *two-word* verbs são formados por um verbo seguido de uma partícula adverbial, formando uma única unidade linguística como em "switch on" e "break in", nas quais o significado está atrelado à co-ocorrência desses dois termos. Todas essas expressões – colocações, coligações e outros tipos de padrões lexicais ou léxico-gramaticais - devem ser entendidas em sua totalidade, não fazendo sentido decodificá-las através da soma de suas partes.

A prosódia semântica é outro exemplo de estudo dos padrões da linguagem possibilitado pela análise de *corpus*, a partir do instrumental da LC. Refere-se à associação entre itens lexicais e determinada conotação, podendo ser positiva, negativa ou neutra (BERBER SARDINHA, 2004). Isto é, determinadas palavras

tendem a atrair termos com determinada carga semântica. Conforme exemplifica (BERBER SARDINHA, 2004), o verbo *cause* tem uma prosódia semântica negativa pois tende a vir acompanhado de palavras desfavoráveis, tais como *problem(s), damage, disease, death, cancer.* Assim, o estudos acerca dos padrões de uma língua estão no cerne das pesquisas da LC, sendo uma de suas principais contribuições ao estudo da linguagem. Na esteira do estudo dos padrões recorrentes da língua Hunston afirma que:

> Os padrões de uma palavra podem ser definidos como todas as palavras e estruturas com as quais são regularmente associados, contribuindo para seu significado. Um padrão pode ser identificado se uma combinação de palavras ocorre com relativa frequência, se é dependente de uma palavra específica, e se há um significado claro associado[21] (HUNSTON E FRANCIS, 1999, p. 37)

O estudo da padronização da língua tem produzido diferentes taxonomias dos grupos lexicais, originando diferentes terminologias de acordo com a sua funcionalidade, ou com o nível linguístico considerado: sintático, semântico ou pragmático. Alison Wray (1999) fez um extenso levantamento bibliográfico em que contabilizou mais de cinquenta termos utilizados para nomear as diferentes formas de padronização da linguagem. Nessa lista estão incluídos, além dos termos colocações/coligações já citados outros tantos tais como, chunks, linguagem formulaica, padrões pré-fabricados, itens lexicais, expressões fixas, lexical phrases, para citar apenas os mais usuais. Embora muitos desses termos sejam sinônimos, alguns deles apresentam diferenças na composição, função ou utilização. Por exemplo, Nattinger e DeCarrico (1992) utilizam o termo *lexical phrase* como um subtipo de colocação, associado diretamente a funções pragmáticas. No entanto, essa variação terminológica tem em comum o fato de referir o fenômeno linguístico da co-ocorrência de termos e expressões, tendo em seu bojo o "*idiom principle*", conforme definido por Sinclair (1991, p. 110): "O princípio idiomático significa que o usuário da língua tem disponível para si um grande número de estruturas linguísticas

[21] Tradução realizada por Tony Berber Sardinha (BERBER SARDINHA, 2004, p. 39-40).

semi ou pré-construídas que se constituem de fato como unidades, embora possam parecer analisáveis em segmentos[22]".

A constatação da existência de padrões recorrentes da língua põe em evidência a não separação entre léxico e gramática, isto é, salienta que ambos estão quase sempre amalgamados na produção do sentido. Sinclair (1991) destaca que o fato da forma estar alinhada com o significado, constitui o cerne das questões apontadas pela LC. Assim:

> O reconhecimento de que a forma está com frequência alinhada com o significado foi um importante passo que afetou a concepção ortodoxa de significado predominante. Então, evidenciou-se que a forma, de fato, poderia ser um determinante do significado e uma conexão causal foi postulada, fornecendo argumentos a favor da forma para a produção do significado. Assim, um ajustamento conceitual ocorreu a partir da observação de que a escolha de um significado, onde quer que esteja localizado no texto, deverá ter um profundo efeito nas escolhas do entorno. Seria fútil imaginar o contrário. Afinal, não existe nenhuma distinção entre forma e significado[23]". (SINCLAIR, 1991, p. 7)

De fato, tal constatação rompe com a pressuposição racionalista de que a gramática/sintaxe constitui-se em um sistema independente, onde o usuário poderia, à sua escolha, encaixar palavras numa determinada estrutura preexistente, abstrata e pronta a receber as palavras provenientes do léxico do falante. O aprendizado da língua seria o domínio das estruturas sintáticas que deveriam ser combinadas com todo e qualquer vocabulário que fizesse sentido no contexto. A partir dessa concepção, o usuário teria toda a liberdade criativa para produzir a língua.

Por outro lado, ao utilizar padrões linguísticos, o usuário da língua mostra que lança mão de conjuntos linguísticos pré-constituídos pela mesclagem entre uma estrutura linguística e vocábulos, formando assim uma unidade. Por exemplo, a

---

[22] "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110).

[23] The recognition that form is often in alignment with meaning was an important step, and one that cut across the received orthodoxy of the explanation of meaning. Soon it was realized that form could actually be a determiner of meaning, and a causal connection was postulated, inviting arguments from form to meaning. Then a conceptual adjustment was made, with the realization that the choice of a meaning, anywhere in a text, must have a profound effect on the surrounding choices. It would be futile to imagine otherwise. There is ultimately no distinction between form and meaning. (SINCLAIR, 1991, p. 7)

expressão “Tenha um bom final de semana” comumente utilizada, surge na comunicação nesta estrutura e forma, sem alterações. Alguém poderia dizer “Te desejo um bom final de semana”, ou “espero que tenhas um final de semana bom” e seria, com certeza, entendido, porém não teria a mesma naturalidade e eficiência que a estrutura pré-fabricada. Esse fato parece colocar por terra o entendimento da língua como um fenômeno preponderantemente abstrato e criativo. Segundo Berber Sardinha (2004):

> Os modelos estruturais da linguagem em geral (incluindo os gerativistas de Chomsky) a descrevem por meio de esquemas de *slot and filler*, nos quais as lacunas (slots) sintáticas podem ser preenchidas lexicalmente de qualquer modo, desde que o conjunto de lacunas seja estruturalmente plausível. Essa visão tem críticos ferozes dentro da Linguística de *Corpus*, dentre os quais se destaca John Sinclair. O programa da pesquisa de Sinclair tem-se pautado pela descrição da linguagem do ponto de vista lexical, cuja perspectiva é a descrição de quais agrupamentos lexicais são realmente empregados pelos falantes, isto é, atestados pelo uso. Essa perspectiva se concretizou em um princípio de entendimento da linguagem chamado de idiomático (*idiom principle)*, explicado como o fato de o usuário de uma língua ter à sua disposição “um grande número de frases pré ou semiconstruídas, que se constituem em escolhas únicas, muito embora pareçam analisáveis em segmentos (SINCLAIR, 1987, p. 230, tradução do autor)”. (BERBER SARDINHA, 2004, p. 33).

Desse modo, a Linguística de *Corpus* evidenciou as estreitas relações entre léxico e estrutura, a inseparabilidade de ambos na produção do sentido, e o alto grau de frequência e uso de padrões léxico-gramaticais determinados culturalmente. A partir desses dados foi possível postular que o fluxo linguístico dá-se muito mais pela reutilização de estruturas prontas do que pela invenção de novas combinações linguísticas durante o ato comunicativo, tal como o pressupõe a gramática universal. De fato, pesquisas mostram que as fórmulas linguísticas possuem alta frequência de uso, em índices até então não imaginados. De acordo com estudos realizados desde meados dos anos 80, a recorrência de fórmulas linguísticas chega a constituir 70% da linguagem de um falante nativo adulto. Convém, por isso mesmo, atentar para os dados coletados por Wray and Perkins (2001):

> Se considerarmos, como alguns o fazem, que os padrões formulaicos incluem o enorme conjunto de simples colocações lexicais, cujos padrões do ponto de vista da gramática formal são tão notáveis e intrigantes (e.g. Sinclair, 1991), então provavelmente algo em torno de 70% de nossas

> línguas nativas constitui-se de padrões formulaicos têm mostrado que o padrão seguido pelas palavras e expressões na linguagem comum manifesta muito menos variação do que poderia ser previsto se tivesse por base a gramática e o léxico, separadamente. De fato, a maior parte da linguagem natural, escrita ou falada, parece constituir-se em grande parte de conjuntos ou estruturas formados por colocações[24] (WRAY e PERKINS, 2000, p. 1-2).

Essa constatação coloca na berlinda os pressupostos da Gramática Universal, ou seja, a língua entendida como um modelo altamente criativo de produção de sentido. Como se sabe, tal modelo predomina no sistema de ensino, existindo ainda pouca difusão do conhecimento existente sobre os padrões lexicais. Ressalta-se, no entanto, que embora se entenda que exista predominância dos padrões e fórmulas linguísticas, isso não significa dizer que não há momentos em que o falante utilize a língua, criativamente, talvez em situações em que tenha de pensar para expressar algo que ainda não está definido ou posto, ou quando utiliza a língua de forma expressiva, como na literatura ou poesia. Nesses momentos, é provável que o indivíduo faça o uso de seu conhecimento sintático da língua, estabelecendo uma combinação inusitada, com determinado vocabulário, na tentativa de produzir sentido. Sinclair (1991), pesquisador que está no centro desta discussão, pondera que:

> A percepção do significado é muito mais explícita do se poderia crer pela forma sugerida por gramáticos abstracionistas. O modelo formal de uma sintaxe altamente generalizável, com lacunas para serem preenchidas com palavras retiradas de uma lista organizada, funciona apenas em casos raros ou em textos especializados. De longe, a maioria dos textos é constituída por palavras comuns em padrões corriqueiros, ou em sutis variações desses padrões. A maior parte das palavras de uso diário não possui um significado ou significados independentes, sendo elas componentes de um rico repertório de padrões de multipalavras que criam o texto. Isso é totalmente obscurecido pelos procedimentos da gramática convencional[25] (SINCLAR, 1991, P.108).

---

[24] "If we take formulaicity to encompass, as some do, also the enormous set of 'simple' lexical collocations, whose patterns are both remarkable and puzzling from a formal grammatical point of view (e.g. Sinclair, 1991), then possibly as much as 70% of our adult native languages may be formulaic (Altenberg, 1990). A range of *corpus* studies (e.g. Kjellmer, 1984; Baayen and Lieber, 1991; Altenberg, 1993; Barkema, 1993) have shown that the patterning of words and phrases in ordinary language manifests far less variability than could be predicted on the basis of grammar and lexicon alone, and in fact most natural language, written or spoken, appears to consist largely of collocations 'sets' or 'frameworks' (Renouf and Sinclair, 1991; Renouf, 1992)." (WRAY & PERKINS :2001, p. 1-2)

[25] "The realization of meaning is much more explicit than is suggested by abstract grammars. The model of a highly generalized formal syntax, with slots into which fall neat lists of words, is suitable only in rare uses and specialized texts. By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday

Um dos aspectos mais interessantes sobre a língua é a possibilidade de se poder dizer e criar, pelo menos em tese, o que se bem entende. Pode-se brincar com a língua e colocá-la ao serviço de nossas necessidades de comunicação e, talvez, justamente este discurso tenha estado no centro das discussões da GU. No entanto, o que as pesquisas em *corpus* têm evidenciado é que a língua segue padrões de repetição e recorrência como até então não se conhecera, ou seja, a língua que é utilizada em situações corriqueiras e normais de comunicação é "a língua que não está tentando produzir efeitos, a língua que é usada de forma semelhante por um grande número de pessoas[26]" (FOX, 1998, p. 31).

1.6 Linguística de *Corpus* e ensino de línguas

O conhecimento trazido pela LC tem produzido efeitos em diversas instâncias relacionadas ao uso, estudo e divulgação da linguagem falada ou escrita: produção de material educativo, publicação de dicionários e gramáticas e ensino, propriamente. O campo da educação e ensino de línguas é diretamente afetado por essas novas concepções acerca da linguagem. Contudo, a quase totalidade dos materiais didáticos até o momento produzidos é elaborada a partir da criação de exemplos inventados, baseados num registro subjetivo da linguagem. O resultado disso são materiais que apresentam para o estudante de uma segunda língua modelos de uso da língua que em situações reais de comunicação muitas vezes não funcionam. Fazer uma revisão de práticas de ensino, incluindo aqui a produção de materiais a serem utilizados por alunos e professores parece ser oneroso. Sinclair, em sua obra seminal, já antevia essa dificuldade ao afirmar que "para muitos linguistas aplicados, deixar de lado a prática da invenção ou adaptação de exemplos significaria uma grande ruptura: o abandono de métodos considerados importantes e

---

words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text. This is totally obscured by the procedures of conventional grammar". (SINCLAR, 1991, P.108)

[26] [...] language which is not striving for effect, language which is used in a similar way by a large number of people (FOX, 1998, p. 31).

a completa revisão de muitas publicações[27]" (SINCLAIR, 1991, p. 5). No presente, já existem alguns dicionários e gramáticas da língua inglesa (BIBER, 1999; CARTER e MCCARTHY, 2006; COLINS COBUILD DICTIONARY, LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH) baseados no estudo de *corpus* linguístico, os quais apresentam exemplos autênticos, registrando o uso real da língua, seja escrita ou falada. No entanto, pouco material didático foi produzido, mas parece que a indústria pedagógica começa a prestar atenção aos resultados trazidos pelas pesquisas da LC.

Publicações originadas no estudo de um *corpus* linguístico procuram apresentar exemplos de uso da língua a partir do mapeamento realizado com o instrumental tecnológico utilizado pela LC. Nessa perspectiva, um dicionário baseado em pesquisa de LC apresenta as palavras e expressões organizadas pela frequência de uso, buscando facilitar o acesso do usuário, pois há maior probabilidade de algum interessado buscar o significado daquela palavra que é mais frequente na língua do que aquela de frequência menor, por exemplo. Esses dicionários também apresentam vários exemplos e anotações relacionadas ao contexto de uso, fornecendo informações pertinentes à pragmática. Mais ainda, dicionários elaborados a partir de uma abordagem da Linguística de *Corpus* listam outras palavras que com frequência ocorrem juntamente com a palavra listada, e que pela forma colocada ou combinada, produzem outras significações. Fórmulas linguísticas, em suas diferentes convenções, são claramente demarcadas nos dicionários produzidos a partir de uma análise de *corpus* e, quase sempre, acompanhadas de exemplos de uso coletados de textos autênticos.

Além de tudo mais, as tecnologias informáticas têm permitido que se registre sonoramente a língua tal como é falada em determinadas situações e contextos. Como resultado, após a compilação e estudo das gravações, descobertas importantes estão sendo feitas acerca de como a língua realmente se comporta, sobre seus padrões, muitas vezes diferindo do modelo desenhado nas gramáticas tradicionais. Resultados desses estudos já constam da primeira gramática descritiva da Língua Inglesa, produzida totalmente a partir da descrição das gravações

[27] "for many applied linguists, to abandon the practice of inventing or adapting examples would mean a big change; the demise of cherished methods and the wholesale revision of many publications" (SINCLAIR, 1991, p. 5).

registradas da língua falada (CARTER e MCCARTHY, 2006). Tudo indica, que materiais didáticos para o ensino de línguas serão também produzidos a partir desses referenciais.

A utilização de um *corpus* linguístico, como fonte de pesquisa - tanto para o professor, para o pesquisador como para os estudantes - pode servir como fonte de referência de modelos linguísticos utilizados pelos falantes nativos. O uso de um *corpus* linguístico produzido por instituições de pesquisa, tal como o BNC (British National *Corpus*), pode vir a ser utilizado pelo educador, como uma base de dados para o ensino de uma segunda língua. Assim, por exemplo, o estudante pode verificar como determinada expressão linguística foi utilizada por vários autores em situações semelhantes àquela que investiga. Por esse meio, uma das grandes dificuldades do ensino de uma língua no contexto escolar pode ser neutralizada: a escassez de materiais autênticos. Através da tecnologia utilizada pela LC, o input linguístico necessário para o aprendizado deixa de ser escasso e torna-se abundante na forma de bancos de dados que podem conter desde dezenas até milhões de exemplos de uso da língua.

A popularização da tecnologia computacional torna possível a educadores interessados produzirem seus próprios *corpora* de estudos. Um *corpus* produzido pelo professor de um idioma, bem como *corpora* públicos disponibilizados por universidades podem ser utilizados diretamente com os alunos para a exploração da língua, além de servir de referência para o desenho e criação de cursos ou materiais de ensino, por parte do professor ou organizadores de um currículo. O terceiro capítulo desta dissertação apresentará propostas para utilizar o conhecimento produzido pela LC para a produção de material de ensino de língua inglesa, focado no desenvolvimento da competência leitora de acadêmicos do curso de Tecnologia Ambiental.

Conforme o que foi até aqui descrito, parece claro que conhecer os padrões linguísticos de uma língua deve ser ponto a ser contemplado no planejamento de ensino de uma segunda língua (e provavelmente, também, no ensino da língua materna). O próximo capítulo aprofunda aspectos das contribuições da Linguística de *Corpus* para o ensino de uma língua estrangeira, com ênfase no ensino do vocabulário, trazendo, além disso, evidências da Linguística Cognitiva sobre o

processamento mental dos padrões linguísticos, além de outros pontos destacados pela LC (frequência, recorrência, exposição, uso de textos naturais, fluência, idiomaticidade e uso de sequências formulaicas) que parecem ser significativos para o aprendizado de uma segunda língua.

2 CONTRIBUIÇÕES DA LINGUÍSTICA COGNITIVA PARA O ESTUDO DOS PADRÕES LINGUÍSTICOS: PONTOS DE ENCONTRO COM A LINGUÍSTICA DE *CORPUS*

Este capítulo inter-relaciona teoricamente conceitos provenientes da Linguística de *Corpus* e da Linguística Cognitiva. De início, faz-se um levantamento das contribuições teóricas da psicolinguística sobre processamento de uma segunda língua, com ênfase na análise do papel da frequência e da recorrência de termos e sua relação com o aprendizado, na quantidade de acesso e de exposição à língua alvo, na qualidade do input disponível (autenticidade), no processamento psicolinguístico dos grupos de palavras e sua relação com o funcionamento da memória, assim como sua possível relação com a fluência verbal. Entende-se que resultados aportados por esses estudos e, também, por aqueles realizados pela Linguística de *Corpus* apresentam aspectos investigativos que se complementam para se entender os processos envolvidos na aquisição de uma segunda língua. Conhecer tais pesquisas pode trazer subsídios valiosos para o ensino da língua inglesa, em especial, e, sobretudo, para a produção de material de ensino, objetivo último desta dissertação de mestrado.

Assim, pretende-se descortinar algumas das possíveis intersecções entre as duas vertentes de investigação linguística citadas. Schmitt, Grandage e Adolphs (2004) apresentam uma síntese bastante precisa acerca das contribuições advindas dessas duas áreas de estudo.

> As abordagens sobre aquisição da linguagem com base em *corpus* e psicolinguística se intercomplementam e, de fato, há vários links entre esses dois modos de pesquisa. Em particular, os estudos psicolinguísticos, com frequência, recorrem a dados de *corpus* para selecionar e controlar itens lexicais a serem tratados como objeto de estudo. Logo não é estranho afirmar que os dados provenientes da análise de um *corpus* refletem a realidade psicolinguística, de como a língua é processada e usada. Além do mais, quase todos os *corpora* são compilados de vários tipos de linguagem autêntica produzida por pessoas reais. Em alguns casos, as evidências de um *corpus* podem ser diretamente interpretadas como um reflexo do

> verdadeiro estado mental subjacente das pessoas que contribuíram com o *corpus*[28]" (SCHMITT, GRANDAGE e ADOLPHS, 2004, p. 128).

O propósito é, pois, articular os achados teóricos da Linguística Cognitiva e da Linguística de *Corpus*, a partir da revisão de alguns pressupostos básicos, os quais dizem respeito ao estudo do vocabulário para o ensino da língua inglesa, destacando-se a relevância do *input* e do conhecimento das frequências dos termos recorrentes no *corpus* de uma dada língua, a fim de definir seu programa de ensino. Ensino e instrução diretos serão considerados à luz de pesquisas calcadas no uso linguístico (*usage based approach*), em especial, para o aprendizado de uma segunda língua. Várias explicitações teóricas serão mobilizadas para fundamentar a utilização dos conceitos de *pacotes lexicais* e de *sequências formulaicas*, pontos centrais desta investigação.

Num segundo momento, serão definidos os conceitos de pacotes lexicais (*lexical bundles*) e de sequências formulaicas (*formulaic sequences*), esclarecendo-se qual a nomenclatura adicional a ser utilizada para referir as unidades lexicais compostas por uma sequência de termos. No processo, será dado especial destaque à função e à relevância desses pacotes e dessas sequências no processamento linguístico e no entendimento de uma língua, sobretudo, no aprendizado de LE e, ainda, sua relevância no desenvolvimento da fluência leitora. Para tanto serão aportados resultados de pesquisas da psicolinguística que apresentam evidências de que a mente processa a linguagem a partir de grupos de palavras, ou seja, utilizando preponderantemente as sequências formulaicas. Também serão consideradas as pesquisas sobre o processamento da memória e os estudos empíricos sobre movimento dos olhos durante a leitura, sobre a função da pausa no processamento das sequências formulaicas, sobre a velocidade de leitura entre falantes nativos e estudantes da língua inglesa, todas elas pontuando

---

[28] "The *corpus* and psycholinguistics/acquisition approaches complement each other, and indeed there are clear links between the two modes of research. In particular, psycholinguistics studies often draw upon *corpus* data to select and control target lexical items. It is not unnatural them to assume that the data drawn from *corpus* analyses reflects the psycholinguistic reality of how language is processed and produced. After all, nearly all *corpora* are compiled from authentic language of various types, which real people have produced. In some cases, *corpus* evidence can be directly interpreted as reflecting the true underlying mental state of the people contributing to the *corpus*. " (SCHMITT, GRANDAGE & ADOLPHS, 2004, p. 128)

evidências e trazendo novos conhecimentos sobre como a mente humana processa a linguagem.

2.1 Vocabulário, ensino e aprendizado

2.1.1 A relevância do vocabulário

Embora o estudo do vocabulário seja de extrema importância para o aprendizado de uma língua, a ênfase dada ao estudo da sintaxe relegou-o a segundo plano, por bom tempo, dada a influência decisiva do gerativismo. De acordo com os pressupostos chomskyanos acerca do conhecimento linguístico, o aprendiz é visto como criador de sistemas de linguagem (GASS e SELLINKER, 2008). Nessa concepção, a língua utilizada por um indivíduo seria por ele criada (criatividade linguística) através de um dispositivo mental que Chomsky denominou Gramática Universal. A GU considera a sintaxe como a área de estudo privilegiada da linguagem articulada e, em vista disso, aprender uma língua passa necessariamente pelo aprendizado das regras gramaticais sintagmáticas que são semelhantes em todas as línguas, já que determinadas exclusivamente pelos mecanismos internos da mente humana. Assim, em consonância com os pressupostos do gerativismo, usar a língua significa mobilizar a estrutura geral, inata, e nela “encaixar” o vocabulário. A sintaxe, então, teria precedência sobre os demais sistemas linguísticos - fonologia e semântica -, ou seja, a sintaxe consistiria do conhecimento das estruturas gramaticais e do seu equacionamento em diferentes situações. Ao indivíduo caberia completar, com itens lexicais, as lacunas da estrutura. Atrelada a essa concepção reside a ideia de o fenômeno linguístico ser um ato plenamente criativo, sendo cada sentença proferida pelo indivíduo uma manifestação de sua criatividade linguística.

Nos últimos vinte anos, outra abordagem desenvolveu-se, paralelamente, contrapondo-se ao gerativismo. Essa proposta teórica - a Linguística Cognitiva - ganhou espaço teórico e acabou se impondo, enquanto teoria de estudo alternativa. Seu fundamento teórico parte da premissa de que as línguas diferem muito entre si “e que cada uma delas contém centenas de construções, na forma de expressões

idiomáticas ou padrões linguísticos, que têm de ser aprendidos na base do *input* junto com o desenvolvimento de processos cognitivos gerais" (GOLDBERG e CASENHISER, 2008, p. 197). Segundo O´Keeffe et al. (2007), o "desenvolvimento em Linguística de *Corpus* tem convencido muitos linguistas de que o vocabulário é muito mais do que aquilo que Chomsky chamou de 'uma lista desordenada de todas a formações lexicais.'" Assim, tendo em vista as posições de teóricos da L. Cognitiva e da L. *Corpus*, o léxico começa a ganhar destaque em pesquisas relacionadas ao aprendizado de uma segunda língua, tendo em vista o aumento do número de evidências da centralidade de sua função na organização da língua e na criação de significado. Tom Cobb (2007) destaca que "após décadas de trabalho de adivinhação, agora há um entendimento compartilhado por muitos pesquisadores de que a compreensão de um texto depende em muito de um conhecimento detalhado da maioria das palavras de um texto" (p. 38).

Considera-se, então, que é a partir do conhecimento do vocabulário de um idioma que a comunicação se efetiva, de vez que o léxico é o principal veiculador do significado no sistema linguístico. "Quando as pessoas falam e ouvem, não é pelo puro prazer em manipular formas sintáticas – elas estão preocupadas com o significado que expressam" (LANGACKER, 2008). Segundo esse estudioso, destacar o vocabulário no sistema linguístico, não significa dizer que a gramática não tem importância para a língua ou para o seu ensino, mas sim que "é útil dar-se conta de que a gramática orienta a significação muito mais do que ser um fim em si mesma" (LANGACKER, 2008, p. 67). Este estudo, em consonância com as teorias referidas, com os resultados das pesquisas e com os princípios da Gramática Cognitiva (LANGACKER, BYBEE, 2008), reconhece a importância do léxico no sistema linguístico, conferindo-lhe especial atenção.

2.1.2 O que é vocabulário?

Ao se mencionar o termo 'vocabulário', de imediato, a ele atrela-se o termo 'palavra', como se ambos formassem um par indissociável vocabulário/palavra, sendo considerados sinônimos. Contudo, embora "as palavras tenham sido desde sempre entendidas como unidades independentes" (Lewis, 1997, p. 256), como

signos autônomos que identificam o vocabulário de uma língua, tal concepção vem sendo contestada já há algum tempo por diversos linguistas (Sinclair (1991), Nation (2003), Wray (2000), Nattinger & DeCarrico (1992), entre outros). Do ponto de vista morfológico, no entanto, o termo vocabulário tanto pode ser uma única palavra, como um conjunto delas que se aproximam, formam e adquirem significado diferente do sentido individual de cada uma delas. Assim, o termo vocabulário refere-se ao signo linguístico independente da quantidade de termos que o compõem. "Logo, é por isso necessário entender vocabulário como sendo constituído também por termos compostos ou 'multipalavras' (Nation, 2003, p. 131)".

Lewis, 1997, compartilhando entendimento semelhante, utiliza a expressão *itens lexicais* para se referir a toda a variedade de significantes linguísticos compostos por uma ou mais palavras, contribuindo assim para a extensão do conceito de vocabulário. Para ele, *itens lexicais* "são socialmente sancionados como unidades independentes. Podem ser palavras individuais, ou frases completas – elocuções institucionalizadas – que carregam significados sociais fixos ou pragmáticos peculiares de uma determinada comunidade" (Lewis, 1997, p. 255). Tal definição parece ter sua gênese em definição muito semelhante já proposta por Nattinger e DeCarrico (1992):

> Anteriormente definimos "*lexical phrases*" como compostos de forma e função, unidades léxico-gramaticais que ocupam uma posição em algum lugar entre os polos tradicionais do léxico e da sintaxe: são semelhantes ao léxico ao serem tratadas como unidades, embora em sua maioria consistam de mais de uma palavra e ao mesmo tempo possam ser derivadas de regras sintáticas regulares, tais como frases e orações. Seu uso é regido pelos princípios da competência pragmática que também seleciona e determina funções para as unidades lexicais (*phrase units*). [...] Os itens lexicais (*lexical phrases*) por isso diferem de outras formas convencionalizadas ou congeladas, tais como expressões idiomáticas ou clichês, principalmente por serem utilizados para realizar determinadas funções[29] (Nattinger e DeCarrico, 1992, p. 36).

[29] "Previosuly, we defined lexical phrases as form/function composites, lexico-grammatical units that occupy a position somewhere between the traditional poles of lexicon and syntax: they are similar to lexicon in being treated as units, yet most of them consist of more than one word, and many of them can, at the same time, be derived from the regular rules of syntax, just like other sentences. Their use is governed by principles of pragmatic competence, which also select and assign particular functions to lexical phrases units. (…)Lexical phrases , then, differ from other conventionalized or frozen forms such as idioms or clichés mainly in that they are used to perform certain functions. (Nattinger e DeCarrico, 1992, p. 36)

Norbert Schmitt (2008) acrescenta que “para lidar com essas multipalavras, foi cunhado o termo *lexema (unidade lexical* ou *item lexical)*. Esses três modos de designar são intercambiáveis sendo definidos como ‘um item que funciona como uma única unidade de significado, independentemente do número de palavras que contém” (SCHMITT, 2008, p. 2). Essa concepção do conceito de vocabulário como unidade de significado e sentido, podendo ser constituída tanto por um único termo linguístico quanto por um agrupamento de palavras percorre todo este estudo, independentemente da designação utilizada (chunks, lexical items, lexical bundles, patterns, formulaic language, formulaic sequence, constructor, pré-fabricados, etc.) por diferentes linguistas, para se referir ao mesmo fenômeno linguístico. O´Keeffe; McCarthy e Carter (2007) também chamam a atenção para a relevância dos agrupamentos de palavras na formação do vocabulário e sua relação com o ensino da língua:

> É necessário chamar a atenção para o fato de que muitos *chunks* são tão frequentes quanto as palavras individuais que compõem o vocabulário principal [...]. O que é sugerido é que o programa de vocabulário para ensino do nível básico seria incompleto sem ser dada a devida atenção para os *chunks* mais frequentes, pois muitos deles são tão ou mesmo mais recorrentes do que as palavras individuais que todos concordam que deveriam ser ensinadas[30]”. (O´KEEFFE; MCCARTHY E CARTER 2007, p. 46).

Devido à importância inquestionável dos itens lexicais na constituição das redes conceituais de uma dada língua, bem como no ensino e aprendizado de uma segunda língua, além de ser objeto de análise deste estudo, sua discussão será realizada com maior profundidade logo adiante.

[30] “It needs to be pointed out that many chunks are as frequent that the single-word item which appear in the core vocabulary […]. What it suggests is that the vocabulary syllabus for the basic level is incomplete without due attention being paid to the most frequent chunks, since many of them are as frequent as or more frequent than single item which everyone would agree must be taught” (O´KEEFFE; MCCARTHY E CARTER 2007, P. 46).

2.1.3 Input e vocabulário

Como o conhecimento linguístico (tal como outras formas de conhecimento) ancora-se no contexto social, torna-se necessário rever o conceito de *input e* sua relação com o aprendizado e ensino de uma segunda língua. Segundo Gass e Selinker (2008), "*input* refere-se ao que está disponível para o aprendiz", ou seja, "(...) a língua, tanto na forma escrita quanto falada à qual o aprendiz está exposto" (p. 305). Ou seja, *input* no contexto de aprendizagem de um idioma seria toda a informação linguística disponível para o aprendiz, podendo esse *input* estar atrelado ao plano fonológico, ortográfico, semântico, sintático, etc.

Nessa ótica, o léxico deixa o papel de coadjuvante e passa a ser o componente central para o aprendizado de uma segunda língua, sendo ainda a forma de *input* mais importante, tanto para a produção quanto para a recepção. Gass 1988 (apud GASS e SELLINKER, p. 449, 2008) comenta que "erros gramaticais geralmente resultam em estruturas que são compreendidas, enquanto erros relacionados ao léxico podem interferir na comunicação". Ou seja, o significado é mais facilmente inferido quando os erros são produzidos por desconhecimento da estrutura gramatical do que quando advindos do uso de vocabulário inadequado. Logo, práticas de ensino de uma segunda língua deveriam ter seu foco no desenvolvimento do vocabulário de seus aprendizes.

2.1.4 Vocabulário: o mínimo e o necessário

Dentre as pesquisas acerca do vocabulário, destacam-se os estudos realizados por Paul Nation (2001, 2003), Batia Laufer(1997), Tom Cobb (1997, 2007), Coxhead (1998, 2000) que discutem a quantidade mínima de vocábulos necessária para o aprendizado de uma língua, no caso, o inglês como LE. Esses autores investigam exatamente o número de palavras que um aprendiz precisa conhecer para conseguir comunicar-se na segunda língua. A pergunta básica é: Qual o conhecimento semântico mínimo necessário para poder ler um texto com fluência numa língua estrangeira? Nation (2001) afirma que o conhecimento das 2000 famílias de palavras mais recorrentemente utilizadas na produção escrita seria o mínimo

necessário para iniciar a compreensão de qualquer texto em língua inglesa, atingindo um cobertura de aproximadamente 80% do vocabulário, ou seja, uma palavra a cada cinco seria desconhecida. Laufer, de sua parte, acrescenta que:

> Há evidências que aprendizes - independentemente de sua alta qualificação - que não dominavam as 3000 palavras mais frequentes (5000 itens lexicais) da língua apresentaram fraco desempenho em testes de leitura. Em outras palavras, mesmo os estudantes mais capazes e inteligentes, que são bons leitores em suas línguas nativas, não conseguem ler com fluência na L2 se o conhecimento do vocabulário está abaixo desse limite mínimo (LAUFER, 1997, p. 24).

Paul Nation e colegas “analisaram textos acadêmicos e determinaram que, dentro desse domínio, existem certas palavras que embora não sejam frequentes na língua como um todo, são muito frequentes na produção escrita acadêmica” (COBB, 1997). Esses autores utilizaram uma lista contendo as palavras mais recorrentes na produção textual acadêmica, a AWL (Academic Word List definida por Coxhead, 1998). A AWL é composta de aproximadamente 570 famílias da palavras que não estão incluídas entre as 2000 mais frequentes da lista geral. São palavras que possuem frequência atestada em textos acadêmicos provenientes de diferentes áreas do conhecimento.

Nation (2001) apresenta dados comparativos (Tabela 1) indicando a frequência em porcentagem das primeiras 1000 palavras da lista ordenada, incluindo as demais 1000 palavras que totalizam o grupo das 2000 mais frequentes. A tabela contém ainda a contagem de vocabulário acadêmico nos quatro tipos de textos (conversação, ficção, jornais e textos acadêmicos) analisados. Conforme consta na tabela, as primeiras 1000 palavras cobrem entre 73% e 84% dos textos analisados, sendo que as palavras classificadas entre as posições 1001 e 2000 cobrem aproximadamente 5-6% do *corpus*. Além disso , os textos acadêmicos possuem uma frequência de vocabulário acadêmico específico – conforme a *Academic Wordlist,* elaborada por Coxhead (1998) - que totaliza 8,5%, chegando a ser quatro vezes mais frequente nos textos acadêmicos do que nos demais.

Tabela 1. Gêneros textuais e abrangência das 2000 mais frequentes palavras da Língua Inglesa e uma lista de vocabulário acadêmico em quatro diferentes gêneros textuais.

Níveis	Conversação	Ficção	Jornais	Textos Acadêmicos
1st 1000 (1 a 1000)	84,3%	82,3%	75,6%	73,5%
2nd 1000 (1001 a 2000)	6%	5,1%	4,7%	4,6%
Acadêmico	1,9%	1,7%	3,9%	8,5%
Outros	7,8%	10,9%	15,7%	13,3%

Fonte: (NATION, 2001, p. 17)

No caso do leitor acadêmico em formação que temos em mente, caso possua o domínio das 2000 palavras mais frequentes da língua inglesa, terá segundo esse entendimento, conhecimento de aproximadamente 80% do vocabulário desse idioma. A partir desse número mínimo de vocabulário, acredita-se que o aprendiz teria condições de autônomamente expandir seu vocabulário, bem como, a começar, de fato, a realizar inferências mais precisas acerca do tema do texto lido. Segundo Cobb (1997), no entanto, compreender 80% do vocabulário de um texto ainda não é suficiente para a compreensão de um texto acadêmico e nem para a aquisição independente de vocabulário a partir da leitura e do aprendizado incidental. A soma dos vocábulos listados nos textos acadêmicos (conforme *Academic Wordlist, Coxhead (1998))com a lista geral* que indica as 2000 palavras mais frequentes, totalizando 2570 termos, aproxima-se do mínimo de 90% dos vocábulos necessários para o entendimento da produção escrita acadêmica.

Logo, seria estratégico focar o ensino dos vocábulos altamente recorrentes na língua para o desenvolvimento da competência leitora do grupo composto por acadêmicos. Ainda segundo Cobb (1997), “para o restante da jornada (90% a 95%)”, tendo esse domínio de 90% do vocabulário, o estudante estaria muito bem, pois teria uma base adequada para realizar inferências ou pesquisas necessárias para a compreensão leitora (dicionários, livros de referência, etc.). Ou seja, somente a partir desse estágio, o leitor teria condições de ler com autonomia e continuar seu aprendizado com independência. Além do mais, o falante de língua portuguesa, no

processo de aprendizagem da leitura em Língua Inglesa, pode ainda beneficiar-se do fácil reconhecimento de cognatos latinos. Essa semelhança linguística pode ser benéfica ao leitor culto da Língua Portuguesa, de vez que iniciará o aprendizado da Língua Inglesa, podendo de imediato utilizar um domínio vocabular herdado de sua língua materna. Tal transferência positiva da língua materna, por exemplo, não teria um falante de uma língua oriental, como o chinês ou o japonês.

Laufer (1997) atribui uma denominação especial ao vocabulário mínimo, essencial, para a evolução do conhecimento de uma segunda língua - *sight vocabulary* (vocabulário fácil e rapidamente reconhecível). De acordo com a autora, essa seria a base vocabular que permitiria ao leitor o rápido reconhecimento dos vocábulos constituintes de um texto durante o ato de leitura. Nesse caso, seus olhos *escaneariam* o texto e, quase que simultânea e automaticamente, reconheceriam o sentido atrelado aos itens lexicais já conhecidos e registrados em seu léxico mental. Segundo Laufer, o *sight vocabulary* é o nível de vocabulário necessário para poder ocorrer a transferência de estratégias de leitura desenvolvidas na L1 para a L2. Mais ainda, o *sight vocabulary* seria composto por “palavras cuja forma e significado são reconhecidos de modo automático, independentemente do contexto” (LAUFER, 1997, p. 22). O desenvolvimento desse tipo de vocabulário (*sight vocabulary)* resulta de treinamento e de prática do aprendiz e pode ser alcançado em programas de aprendizagem.

Tal domínio vocabular vincula-se a funções cognitivas importantes e ao funcionamento da memória do leitor. O domínio pleno do *sight vocabulary*, ou seja, o reconhecimento automático das 3000 palavras mais frequentes da língua inglesa, permitiria ao leitor liberar o funcionamento da memória para realizar outras atividades pertinentes ao ato da leitura. Parece evidente, então, que somente a partir da automatização do reconhecimento instantâneo dessa base lexical, a memória estaria liberada para realizar a leitura em outros níveis mais complexos. Isto é, apenas depois de atingir esse limiar mínimo seria possível interpretar, argumentar e contra-argumentar, firmar posicionamentos frente ao texto, concordar e discordar, comparar, analisar, inter-relacionar textos e conhecimentos, entre outras funções do intelecto realizadas no ato da leitura. Ou seja, só após a automatização de

procedimentos *bottom-up,* a mente do leitor fica liberada, tendo em vista o seu domínio do vocabulário.

> Uma vez que a quantidade de informação que pode ser cognitivamente manipulada, por processos controlados em um determinado período de tempo, é limitada; focar constantemente em palavras parcial ou totalmente desconhecidas sobrecarregaria o uso de capacidades cognitivas que, do contrário, poderiam estar sendo utilizadas em níveis mais altos de processamento do texto. O reconhecimento automático de uma grande quantidade de vocabulário – grande quantidade de *sight vocabulary* – por outro lado, libera os dispositivos cognitivos do leitor para (1) compreender o sentido de termos desconhecidos/novos ou não lembrados e (2) interpretar o significado global do texto. [...] Por exemplo, se o esforço cognitivo de um leitor é direcionado e sobrecarregado para a compreensão num nível frasal ou no reconhecimento de expressões idiomáticas, isto é, ao tentar frequentemente decifrar palavras desconhecidas, o leitor terá dificuldades em perceber as relações e conexões entre parágrafos. Até finalizar a leitura do parágrafo seguinte, poderá ter esquecido sobre o que tratava o parágrafo anterior e por isso não terá condições de estabelecer as conexões entre os dois parágrafos" (LAUFER, 1997, p. 22-23).

2.1.5 Listas de palavras

O instrumental tecnológico utilizado pela Linguística de *Corpus*, combinação de softwares e equipamentos específicos, conforme já referido, permite a compilação de *corpora* de diferentes dimensões. Um dos resultados obtidos pela análise de *corpus*, a partir de um algoritmo computacional em interface com um software específico, é a elaboração de listas de frequência das palavras que compõem o *corpus*. Tais listas podem ser apresentadas e organizadas em ordem de frequência possibilitando que se conheça qual o termo mais frequente e qual o menos recorrente do *corpus*.

A aplicabilidade dessas listas pode ser de grande valia tanto para o ensino de língua, como para a elaboração e definição de programas de ensino de língua estrangeira. Conforme O'Keeffe et al. os "*corpora* têm confirmado nossas intuições sobre língua e com frequência têm mostrado que essas podem ser enganosas quando se referem a questões pertinentes à semática ou à gramática" (2007, p. 21). As listagens possibilitam, por exemplo, conhecer e elencar, a partir de análise empírica, o vocabulário que o aluno terá maior probabilidade de encontrar em sua área de estudos. Segundo Nation (2001, p. 16):

> As palavras de alta frequência da língua são visivelmente tão importantes que tempo considerável deveria ser dispensado por alunos e professores a seu estudo. Essas palavras compõem um conjunto razoavelmente pequeno que torna possível, durante um programa de ensino de inglês, em longo prazo, analisar e focar a atenção na grande maioria delas. Essa atenção pode ser dada na forma de ensino direto, aprendizagem direta, aprendizagem incidental, além de um planejamento preciso para abordá-las no programa de ensino. O tempo dispensado a este vocabulário é muito bem justificado em função de sua frequência, abrangência e alcance (NATION, 2001a, P. 16).

No que se refere ao ensino de inglês instrumental para acadêmicos, a elaboração e utilização de listas das palavras mais frequentes do vocabulário é de grande utilidade no sentido de orientar o educador a respeito do que terá maior relevância para o ensino. Tanto termos de alta frequência como termos que têm baixa probabilidade de ocorrência merecem ser analisados, pois itens lexicais desses dois grupos têm função e importância diferenciadas. "Quando falamos que o vocabulário de alta frequência é importante, sua importância vem da alta probabilidade deste vocabulário ser encontrado em uma ampla variedade de usos da língua" (Nation, 2001b, p. 33). Ou seja, o conhecimento deste vocabulário recorrente pode consolidar o domínio linguístico do aprendiz em diversas instâncias de uso linguístico.

De modo geral, os termos de baixa frequência têm relevância para o ensino de áreas técnicas, já que, comumente, se associam ao vocabulário específico do campo conceitual em análise. Por exemplo, em relação a esta pesquisa, termos com menor frequência e recorrência no *corpus* têm grande probabilidade de constituir o vocabulário técnico da área de Tecnologia Ambiental. Assim, pelo visto, termos de baixa frequência não são irrelevantes, de vez que tendem a integrar o grupo das palavras-chave de um dado *corpus (keywords),* constituindo o rol dos itens lexicais formadores do vocabulário técnico e especializado da área estudada.

Por isso mesmo, em projetos de ensino da língua estrangeira, neste caso, ensino de inglês, no meio acadêmico, os quais comumente associam-se a alguma disciplina técnica ou específica dos cursos, a utilização das listas de frequência produzidas a partir de um *corpus* de estudo pode auxiliar o educador a detectar a ênfase e o direcionamento a ser dado ao ensino, de acordo com o nível de

conhecimento dos alunos. Para alunos com pouco conhecimento da língua inglesa, por exemplo, a proposição inicial, poderia enfocar o estudo dos termos recorrentes apontados na análise do *corpus* e, posterior e gradativamente, seriam feitas aproximações daqueles termos não tão frequentes, porém relevantes e significativos para o contexto de estudo do aprendiz. Nation (2001b) indica algumas possibilidades de interpretação e utilização dos dados apresentados nas listagens de frequência, os quais poderiam ser utilizados pelos educadores, com especial atenção aos itens de baixa frequência:

> Uma das formas de tornar o vocabulário de baixa frequência mais controlável é aumentar o número de palavras de alta frequência. Isso pode ser feito através de uma abordagem de ensino por áreas específicas. Ou seja, os objetivos de uso da língua dos aprendizes são examinados e, em seguida, uma pesquisa é feita para ver se existe vocabulário especializado que não está presente entre as 2000 palavras mais frequentes da língua, mas é frequente e de grande utilização dentro da área especializada em que os alunos estão interessados (Nation, 2001b, p. 33).

O conjunto de softwares utilizados pela L. *Corpus*, destacando-se dentre eles o Word Smith Tools, permite tal análise. O presente estudo, por exemplo, foi realizado a partir da produção e análise de um pequeno *corpus* (*small corpus*) e fez um levantamento dos termos e da linguagem pertinentes ao campo da Tecnologia Ambiental. De início, fez-se a análise tanto dos vocábulos altamente frequentes, ou seja, daqueles termos constantes dentre os 2000 mais recorrentes do inglês, como do vocabulário de baixa frequência o qual, provavelmente, se relacione às palavras-chave e ao vocabulário especializado da área de Tecnologia Ambiental. O estudo das frequências pode ser indicador da densidade das palavras para o contexto em análise e, segundo Nation (2001b):

> A principal ideia subjacente ao uso de alguns desses programas é que as palavras do vocabulário de uma língua não têm igual importância. Algumas palavras são muito mais importantes do que outras. A forma comum de determinar a importância de uma palavra é pela análise de sua frequência e pela faixa de ocorrência. Isto é, palavras que ocorrem numa larga faixa de uso da língua são geralmente muito mais úteis para um usuário da língua, do que palavras que raramente ocorrem sendo utilizadas numa área limitada, particularmente em áreas como biologia, informática, geografia, etc., as quais talvez não sejam de interesse imediato para o usuário (não especializado) da língua (Nation, 2001b, p. 32).

Há indícios que o aprendizado de uma língua parece ser favorecido pela relação entre o termo e sua frequência dentro de um determinado contexto, ou seja, segundo esta ótica, os termos mais recorrentes seriam assimilados pelo aprendiz com maior facilidade. Nation (2001b, p. 33), mencionando Read, afirma que "há evidência suficiente de que os aprendizes assimilam as palavras de alta frequência antes de aprenderem os termos de baixa frequência (Read, 1988)." Segundo o autor, isso não é novidade alguma e esses dados já vêm sendo utilizados por cursos de línguas para focar o ensino nos termos mais recorrentes (embora essa pareça não ser, ainda, uma prática amplamente divulgada no sistema educacional brasileiro). De acordo com Nation (2001b), o fluxo de termos menos frequentes pode ser um dificultador do aprendizado:

> Para os aprendizes que se enquadram no nível intermediário ou superior, o fluxo de termos de baixa frequência será, geralmente, o que lhes causa dificuldades. É por isso importante ser capaz de analisar textos rapidamente e quantificar o vocabulário de baixa frequência e o que essa baixa frequência significa. Isso pode auxiliar um professor a decidir se o texto precisa ser simplificado, se precisa ser trocado por um texto mais fácil ou se é viável alguma pré-atividade acerca do vocabulário (Nation, 2001b, p. 33).

Logo, conhecer o vocabulário mais frequente é condição suficiente e necessária para atingir um nível superior de conhecimento do idioma, ou seja, a fluência em uma ou mais habilidades da língua. No caso do ensino de inglês instrumental, se o leitor não dispuser do conhecimento mínimo das 3000 palavras mais frequentes que compõem a língua inglesa, poderá não ser capaz de compreender e inferir. Como já comentado anteriormente, Laufer (1997, p. 24) afirma que bons leitores da L1 somente poderão transferir as estratégias de leitura da língua materna para a língua estrangeira após atingirem o conhecimento mínimo das 3000 famílias de palavras da língua alvo. Sem esse domínio vocabular, segundo a pesquisadora citada, a transferência seria afetada pelo conhecimento insuficiente do vocabulário.

Em outras palavras, a modalidade de leitura *bottom-up* é condição *sine qua non* para alcançar um outro estágio do ato da leitura: a interpretação. O que implica

dizer que para poder interpretar é fundamental ter domínio do vocabulário básico da língua alvo. Conforme coloca Flôres (2008, p. 44) “para interpretar é preciso ir além da materialidade significante do texto”. Ao destacarmos a importância dos processos *bottom-up,* de forma alguma se está negando as estratégias *top-down,* mas sim reiterando que elas apenas entrarão em cena após o domínio pleno dos processos cognitivos mais elementares da leitura. Num determinado estágio da leitura, entende-se que os dois processos se complementarão na complexa tarefa de interpretação de um texto. Laufer e Sim (1985 apud LAUFER, 1997, p. 21) acrescentam ademais que “descobriram que estudantes, ao interpretar textos, tendem a considerar as palavras do texto como o principal ponto de referência do significado”. Nesse mesmo estudo encontraram evidências de que “o conhecimento de mundo é considerado em menor medida e que a sintaxe é praticamente ignorada” (LAUFER e SIM, 1985 apud LAUFER, 1997, p. 21). Outros pesquisadores citados por Laufer (1997) chegaram a conclusões semelhantes:

> Haynes and Baker (1993) também chegaram à conclusão de que a maior desvantagem para leitores da L2 não é ausência de estratégias de leitura, mas vocabulário insuficiente na língua inglesa. O que esses estudos indicam é que o limiar para a compreensão leitora é, em grande medida, lexical. Logo, problemas relacionados ao conhecimento do léxico limitarão a leitura e compreensão fluentes (LAUFER, 1997, p. 21).

Tais dados remetem diretamente à modalidade de ensino da leitura numa segunda língua que preconiza o desenvolvimento de estratégias de leitura, desconsiderando o nível de conhecimento lexical do aprendiz. Essa abordagem, para o ensino da leitura, estimula a dedução de termos desconhecidos a partir do contexto e parece favorecer o processo de adivinhação de termos não conhecidos. Conforme anteriormente afirmado, o leitor somente terá condições de começar a realizar inferências e deduções a partir do conhecimento de pelo menos 80% dos vocábulos que compõem a trama textual. Laufer (1997) critica a abordagem em questão e “acha difícil aceitar a afirmação de que adivinhar o significado em uma L2 é de fato possível com a maioria das palavras desconhecidas e que o seu sucesso depende exclusivamente das estratégias de adivinhação do aprendiz” (LAUFER, 1997, p. 27-28). A autora embasa sua crítica em outros estudos seus e de colegas e

apresenta algumas informações complementares com base em Bensoussan e Laufer (1984) que

> descobriram em um estudo, utilizando uma passagem de um texto acadêmico, que das 70 palavras que os estudantes foram solicitados a adivinhar/deduzir, pistas contextuais claras e precisas puderam ser utilizadas para apenas 13 palavras. Talvez isso não seja o caso com outros gêneros textuais. Alguns talvez possam fornecer mais pistas, outros até menos. Mas ter como certo que um texto fornecerá pistas para as palavras desconhecidas é exageradamente otimista (LAUFER, 1997, p. 28).

Em suma, entende-se que somente após a decifração e compreensão do vocabulário o leitor poderá chegar ao estágio de realizar interpretações. Logo, o entendimento é que focalizar o aprendizado de estratégias *bottom-up,* associando-o ao estudo minucioso do vocabulário, é imprescindível para o aprendiz tornar-se fluente na leitura de uma segunda língua.

Conforme até aqui exposto, é clara a relação entre vocabulário e input. Procurou-se trazer dados provenientes de estudos que evidenciam o valor do vocabulário e sua função na compreensão leitora. Nessa ótica, o vocabulário é a forma de input a ser enfatizada durante o ensino e aprendizagem de uma segunda língua. “Uma função significativa na instrução de línguas está relacionada à manipulação do input. O que significa que o professor pode possibilitar diversos graus de explicitude ao input” (GASS e SELLINKER, p. 387). Assim, a Linguística de *Corpus*, a partir da possibilidade de análise e estudo de *corpus* e do mapeamento do vocabulário, oferece instrumental para o educador direcionar, de forma mais objetiva, o ensino da língua. Na seção a seguir, serão discutidos alguns estudos que enfatizam o ensino do vocabulário como base e fundamento para a aprendizagem de uma segunda língua.

2.2 Contribuições teóricas da Linguística Cognitiva para ensino de língua estrangeira

2.2.1 Instrução explícita, ensino e léxico

Abordar a instrução direta significa enfocar o aprendizado específico que ocorre em ambiente de ensino. Uma das maiores diferenças entre um ambiente de aprendizagem e um ambiente natural relaciona-se à quantidade e qualidade do input (GASS e SELLINKER, 2008). Na sala de aula, o input disponível é proveniente de seleções determinadas e planejadas pelo professor, como também de conhecimentos trazidos pelos próprios aprendizes, embora o conhecimento específico da língua em estudo possa ser limitado. Muitas vezes, o único contato que os alunos têm com a língua estrangeira é aquele baseado na experiência em sala de aula. Conforme sintetizam Gass e Sellinker (2008), no espaço escolar, o input é disponibilizado, sobretudo, de três maneiras: (1) através das interações com o professor, (2) através do material de ensino e (3) através do conhecimento socializado entre os alunos.

Estudantes que tomam contato com a língua inglesa em contexto onde essa não seja falada, como língua nativa ou como língua de comunicação, geralmente, não têm interlocutores com os quais praticar os conhecimentos linguísticos desenvolvidos em classe. Essa dificuldade é amenizada se os estudantes estiverem focados no aprendizado da habilidade leitora. Até porque, o indivíduo que tiver acesso a um computador conectado à internet, tem acesso a uma infinidade de textos em língua inglesa. A internet permite a circulação dos mais diferentes gêneros textuais, possibilitando que os interessados acessem publicações para aprendizado na língua alvo com a mesma facilidade com que as acessam em sua língua nativa. No entanto, ter acesso ao texto em si não é garantia por si só de desenvolvimento do aprendizado da leitura em uma segunda língua.

Krashen defende (apud Cobb 2007) e “acredita que todo o léxico necessário para a leitura pode ser adquirido naturalmente através da própria leitura, tanto na segunda língua como na primeira”. Porém, se assim fosse, todo leitor fluente na sua língua materna seria um leitor em potencial na segunda língua, tendo todas as

condições de desenvolver a leitura de forma autônoma. Os fatos têm demonstrado que não é tão simples assim. Conforme Cobb (2007), com o auxílio de softwares que permitem a análise textual e sua aplicação ao aprendizado de línguas, a quantificação dessa análise demonstra que é extremamente improvável o desenvolvimento de um léxico adequado em L2 somente através da leitura, mesmo nas condições mais favoráveis. Por outro lado, ficam mais e mais numerosos os resultados positivos advindos da instrução explícita do vocabulário da língua alvo.

Além disso, Gass e Sellinker (2008) comentam que os aprendizes não absorvem os erros cometidos pelos colegas, mas que a instrução e a intervenção do professor é essencial para que os envolvidos saiam da sala de aula esclarecidos sobre impasses que possam ter ocorrido na interação com colegas. Embora tal comentário refira-se à abordagem comunicativa, acredita-se que o papel desempenhado pelo professor no ensino da leitura é crucial. A partir da leitura dos alunos, acompanhada (ou não) de interpretação, o professor pode intervir, evitando equívocos e sanando dúvidas relacionadas à compreensão, esclarecendo, por exemplo, a função de determinadas estruturas e vocábulos na produção do sentido do texto. De fato, a intervenção de um terceito pode proporcionar ao aprendiz uma oportunidade para se dedicar à resolução de um problema, com a colaboração de um agente mais capaz. Este agente normalmente é o professor, mas também pode ser um colega que esteja em um nível de conhecimento superior ao do estudante, podendo intervir beneficamente para que o colega avance, através da interação/socialização, no aprendizado da língua. Materiais pedagógicos e tecnológicos utilizados para o ensino, os quais registram e referenciam a língua, tais como dicionários, sites de pesquisa, tarefas e exercícios, concordanciadores, podem ser entendidos também como agentes mediadores do processo de aprendizagem, de pleno direito.

2.2.2 O processamento da instrução: instrução direta e intervenção

O processamento da instrução, segundo GASS e SELLINKER (2008), refere-se ao tipo de instrução que toma como base a maneira como os aprendizes processam o input. Esse modelo instrucional é baseado na proposta feita por VanPatten e

colegas (apud GASS e SELLINKER, 2008) e se sustenta na atenção à forma e ao seu papel no processo de aprendizagem, vinculando a passagem do *input* para o *intake* e, consecutivamente, para o *output*. No contexto de ensino da leitura, considera-se o vocabulário e as expressões constituintes do texto a ser lido pelo aprendiz como o principal *input*; o *output* seria a evidência da compreensão/interpretação por parte do aprendiz. Nosso interesse recai no modelo de processamento da instrução proposto por VanPatten através do qual a mediação do professor se dá no sentido de elucidar, de alguma forma, a maneira como o input é percebido e processado pelo aprendiz.

Segundo os princípios instrucionais propostos por VanPatten (2008) há três premissas básicas presentes no processo instrucional, as quais foram adaptadas, pensando-se no ensino de leitura: (1) aprendizes precisam de input para aquisição; (2) dificuldades na aquisição podem estar relacionadas ao modo que os aprendizes processam o input e (3) se o educador entender como o input é processado, então terá condições de eleger procedimentos para destacar o *input* (*input enhancement*), ou a estrutura linguística (*focus on form*) responsável por determinada significação presente no texto. Conforme aqui se propõe, este estudo dá saliência ao reconhecimento dos padrões linguísticos e sua focalização visa a auxiliar o aprendiz no processo de decodificação de determinadas características ou dispositivos de linguagem presentes no texto. Conhecer os pressupostos teóricos envolvidos no processamento cognitivo da linguagem durante a leitura, ou seja, durante o processamento do input do aprendiz de uma segunda língua, pode favorecer o professor no sentido de fazer intervenções mais adequadas e de utilizar procedimentos de intervenção eficazes.

A mediação proposta pode ser feita de diversas maneiras, seja pela correção de uma interpretação, pela antecipação da explicação de uma estrutura semântica complexa, pela exemplificação pragmática do uso uma expressão, pelo oferecimento de diversos exemplos de uso de um determinado termo ou estrutura, pelo destaque dado à maneira como determinados termos linguísticos agrupam-se formando uma estrutura lexical única, entre tantos outros possíveis encaminhamentos a serem realizados pelo professor de línguas. VanPatten sintetiza tais procedimentos, agrupando-os em três diretrizes: (1) fornecer informações sobre a estrutura ou

forma, destacando a relação da estrutura com o significado da expressão ou termo linguístico em jogo; (2) informar os aprendizes sobre processos cognitivos / estratégias que possam ser utilizadas na seleção de determinada estrutura; (3) estruturar o input a fim de que os estudantes possam confiar na estrutura linguística para chegar ao significado. Gass e Sellinker (2008) concluem que esse modelo de processamento da instrução lida não somente com a dificuldade linguística, mas com estratégias de processamento problemáticas, procurando interrompê-las no ato de sua formação, através da instrução direta e prática. Dentre essas estratégias, pode-se incluir a interferência da L1 do estudante em relação à L2, podendo esta interferência tanto ser positiva, quanto negativa. Pela instrução direta, ao observar os movimentos cognitivos do aprendiz, o tutor poderá auxiliá-lo a perceber que interferências colaboram com o aprendizado da segunda língua e quais devem ser desconsideradas.

Tomasello e Herron (apud GASS e SELLINKER, 2008) confirmam alguns dos pontos do modelo proposto por VanPatten, ao mostrarem que a correção (*corrective feedback*) costuma ser mais significativa após os aprendizes terem sido induzidos a correr o risco, na tentativa de utilizarem determinado recurso linguístico, ao invés de serem prevenidos com uma correção antecipada para que evitem o erro. O mesmo procedimento é válido no aprendizado da leitura de uma segunda língua e foi utilizado no sentido de os alunos correrem o risco de expor suas tentativas de compreensão e interpretação da leitura do texto proposto. Assim, ao refletir sobre as tentativas feitas talvez possam por si mesmos comprovar a sua validade e, então, caso se sintam motivados, produzir os ajustes antes da intervenção. No entanto, reitera-se que é fundamental, quando possível, a intervenção cooperativa do leitor mais experiente, seja para elucidar o significado de algum vocábulo ou de estrutura linguística. A sala de aula é o local privilegiado para o exercício dessa prática, acompanhada, quando necessário, de observações e orientações de outros leitores mais experientes. A experiência trazida pelo aprendizado da língua e pela prática continuada de leitura possibilitará a fluência, a segurança e a competência leitora almejada, através da busca e repetição consciente de algumas informações e do uso recorrente da língua em estudo.

2.2.3 Exposição, frequência, recorrência

Até este ponto fez-se a análise da relação entre a qualidade e a quantidade do input para o aprendizado da leitura, dando destaque ao ensino do vocabulário. No entanto, há outros fatores que precisam ser articulados ao cenário de aprendizagem de uma língua, seja ao ambiente formal de ensino, seja à sua extensão, seja à vida do indivíduo. A prática, o treino, o uso frequente do conhecimento aprendido são, como se sabe, pelo senso comum, determinantes do aprendizado. Tais variáveis têm sido equacionadas a partir de um quadro teórico que vem ganhando espaço nas pesquisas acerca da linguagem nas últimas décadas: *a usage based approach* (abordagem baseada no uso) (ELLIS e ROBINSON, 2008). Verifica-se, ademais, que alguns de seus pressupostos relacionam-se com o ensino e aprendizado da leitura, em segunda língua, estabelecendo relações com algumas das proposições constantes no escopo da Linguística de *Corpus*.

A abordagem baseada no uso (*a usage based approach*) insere-se no quadro teórico da linguística cognitiva. Conforme Robinson e Ellis (2008) o sintetizam na introdução do *Handbook of Cognitive Linguistics and Second Language Acquisition*, a linguagem é adquirida, enquanto os sujeitos estão envolvidos e engajados na comunicação, ou seja, segundo essa visão teórica, a língua é aprendida a partir do seu uso. Tal concepção, conforme já sinalizado anteriormente, contrapõe-se à visão da Gramática Gerativa que entende a linguagem humana como inata. Conforme o esclarece Langacker:

> a ênfase da abordagem baseada no uso (*usage based approach*) é focada propriamente no aprendizado da linguagem. Seja lá qual for sua base inata, dominar uma linguagem requer seu uso específico e o aprendizado de uma vasta variedade de unidades linguísticas convencionais. Isto por si só tem suas implicações pedagógicas (o que pode parecer óbvio não o é para todas as teorias linguísticas). Sugere-se a importância de possibilitar ao aprendiz exposição suficiente a usos representativos de uma dada unidade linguística. De forma ideal, essa exposição deveria ocorrer no contexto em que ocorrem trocas significativas, aproximando-se do uso de situações normais inscritas na sociedade e na cultura (LANGACKER, 2008, p. 81).

Pelo visto, há relação direta, até mesmo interdependência, entre o uso das construções linguísticas e o *input* disponível. O aprendizado é dependente do *input* e

da exposição frequente a ele. Em tais circunstâncias, a frequência e a recorrência dos termos linguísticos, o seu uso em situações reais, ou o mais próximo possível da realidade linguística são imprescindíveis para que ocorra o aprendizado. Hudson, de forma muito clara e direta resume a pertinência dessas variáveis para o processo de aprendizado de uma segunda língua: “O mote para o aprendiz de uma segunda língua é conhecido: use-a ou esqueça-a” (HUDSON, 2008, p. 105).

As pesquisas baseadas no *usage based approach* têm defendido, ainda, que a questão da frequência e do *input* estão relacionadas de forma muito profunda com o aprendizado de um idioma, a ponto de estarem diretamente conectadas com o aprendizado da gramática. Embora esse tema escape ao escopo deste estudo, mesmo assim, faz-se relevante mencionar o assunto até por ser uma questão central e estar implícita naquilo que aqui se propõe. Assim, essa inter-relação é enfatizada por Bybee, nos seguintes termos:

> Quando as construções são combinadas com um modelo baseado no uso o resultado é uma teoria que propõe que as estruturas gramaticais são desenvolvidas através da experiência com exemplos específicos de construções, os quais estão categorizados na memória através de um processo de mapeamento que combina sequências de construções buscando similaridades ou diferenças. As representações cognitivas resultantes são abstrações sobre a experiência do usuário com a língua. (...) Nessa visão da gramática, o uso e a frequência desempenham um papel importante ao determinar estruturas cognitivas [31]” (BYBEE, p. 217-218).

Bybee (2008), a partir de seus estudos sobre a frequência de palavras e expressões linguísticas utilizadas numa língua, demonstrou a relação entre elas e o aprendizado tanto da língua materna, como de uma segunda língua. Segundo ela, as diferenças relacionadas à frequência dos termos e expressões na língua manifestam-se nas representações cognitivas dos falantes. Formas linguísticas com alto índice de recorrência têm representações relativamente fortes, sendo de fácil acesso; enquanto termos de baixa frequência no léxico, são representações menos acessíveis. Aprendizes de uma segunda língua, segundo a autora, também são

[31] Tradução livre de Vitor Duarte. “When constructions are combined with a Usage-Based model the result is a theory that proposes that grammatical structures are built up through experience with specific examples of constructions which are categorized in memory by a mapping process that matches strings for similarity and difference. The resulting cognitive representations are abstractions over one´s cumulative experience with language. (…) In this view of grammar, then, frequency of use plays an important role in determining cognitive structures”. (BYEBEE, p. 217-218)

favorecidos pelos termos mais recorrentes na língua, da mesma forma que os falantes nativos. Mais ainda, as formas menos frequentes num idioma podem ser aprendidas como derivações das formas mais frequentes. Inclusive, pode-se afirmar que a própria língua, por seus mecanismos, vai constituindo a cognição do aprendiz; estruturas mais complexas produzem as menos complexas, a partir do uso, frequência e interação.

2.2.4 Usage-based e ensino direto

Nick Ellis (2008b) afirma que embora o aprendizado baseado no uso tenha se mostrado eficiente para o ensino e aprendizado da primeira língua, quando se trata de aprendizado de L2, um breve período de prática intensiva, por si só, não produz efeitos tal como na primeira língua. Segundo ele, a interferência de um mediador não pode ser dispensada e tal intervenção continuará a ser importante e necessária, na forma de instrução explícita. Segundo ele:

> Esses processos (a diferença na descontinuidade da abordagem baseada no uso, grifo meu) também explicam porque a instrução focada na forma (form-focused) é um componente necessário da L2A e também porque o sucesso na aquisição de uma segunda língua necessita de um maior nível de atenção explícita às construções da L2, uma tensão dialética tensa entre as forças conflitantes do estágio atual da interlinguagem do aprendiz e a evidência de um feedback explícito tanto linguístico, quanto pragmático ou metalinguístico, o qual possibilitará um suporte ao desenvolvimento do aprendiz[32](ELLIS, 2008b, p. 372).

A intervenção e o ensino explícito, na perspectiva de Ellis, atuariam como conciliadores frente à tensão estabelecida entre o conhecimento da interlíngua (bem como da primeira língua) e a base linguística da L2. Assim, a instrução funcionaria como balizadora, indicando ao aprendiz o funcionamento mais provável de determinado recurso linguístico, auxiliando-o a buscar referências ou modelos na língua utilizada por uma comunidade fluente, ou oferecendo-lhe o modelo que lhe

[32] These processes also explain why form-focused instruction is a necessary component of L2A, and why successful L2A necessitates a greater level of explicit awareness of the L2 constructions, a dialetic tension between the conflicting forces of the learner´s current stable states of interlanguage and the evidence of explicit form-focused feedback either linguistic, pragmatic, or metalinguistic, that allows socially scaffolded development" (ELLIS, p. 372)

faz falta naquele momento. A intervenção é benéfica, nessas circunstâncias, porque a língua é um código compartilhado e aprendido a partir de uma comunidade, logo, a ação do mediador poderia ser uma forma de complementar a experiência do aprendiz com a língua alvo através de uma informação linguística que talvez lhe falte, ou de um esclarecimento suplementar a respeito de algum tipo de interferência de sua primeira língua no processamento daquela que está sendo aprendida. Também poderia ser um auxílio para focar sua atenção em determinados dispositivos linguísticos que não tivessem sido espontaneamente percebidos. Parece, então, que o modelo proposto por VanPatten, acerca da instrução explícita, anteriormente detalhada, afina-se com a interpretação fornecida por Ellis.

Em outro artigo, Ellis (2008a) explica a função cognitiva da aprendizagem explícita, complementando postulações anteriores e afirmando que:

> Há alguns aspectos sutis numa segunda língua, os quais aprendizes propriamente não assimilam, apesar de serem de alta frequência no ambiente de uso da língua: é onde o *input* falha e não se torna *intake*. Tais situações ocorrem porque aprendizes não reconhecem pistas linguísticas não salientadas, destacadas e redundantes ao processar o significado, ou por haver interferência de determinados elementos linguísticos da primeira língua, quando estes devem ser processados diferentemente na segunda língua[33] (Ellis, 2008a, p. 93)

Esses comentários explicativos levam a algumas reflexões acerca do aprendizado da L2 no contexto da sala de aula. Nesse ambiente a exposição à L2 alvo é dramaticamente prejudicada, pois nesse caso os aprendizes têm muito menos contato com a língua do que falantes que a estão aprendendo entre nativos. O *input*, obviamente, é muito menos frequente no contexto da sala de aula do que numa situação de uso real do idioma. Faz-se necessário, então, pensar em uma forma de contemplar, no planejamento de ensino, o equilíbrio entre o uso de termos mais frequentes e menos frequentes de uma dada língua. Ou mesmo, quando necessário, de criar estratégias de ensino para que aqueles termos que naturalmente ocorrem

[33] Tradução livre de Vitor Duarte. “There are ‘fragile’ aspects of second languages which learners fail to acquire despite high frequency in the ambient language: where input fails to become intake. Such situations arise because learners fail to notice cues which are lacking in salience and redundant in cuing meaning, or because of interference where the features need to be processed in a different way from that usual in their L1” (Ellis:2008, p. 93)

com menor frequência no contexto de uso, possam ser mais facilmente acessados pelos estudantes. Conforme já mencionado, os itens mais recorrentes são mais facilmente lembrados e acessados pelo usuário de uma língua. Dessa forma, uma exposição planejada ao vocabulário necessário poderia contribuir para que termos importantes, porém menos recorrentes, fossem, com maior facilidade, aprendidos pelos alunos. O professor, a partir de instrumentalização específica, teria condições de promover essa facilitação, pois a Linguística de *Corpus* apresenta algumas propostas para tal fim, entre elas, o uso de concordanciadores, listas de frequências e facilidade de acesso ao *corpus* digitalizado. Esses recursos podem contribuir para que o professor contemple em seu planejamento oportunidades de interação e utilização da linguagem mais significativa.

Na pesquisa realizada por Bybee (2008), consta uma indagação sobre até onde a exposição a uma segunda língua, no contexto da sala de aula, deve espelhar situações de uso da língua em situações naturais. Ellis e Robinson (2008, p. 14) concluem, dizendo que “a resposta sugerida é que um exato paralelo entre a sala de aula e as situações naturais não é necessário”. Porém, enfatizam que a “atenção à questão da frequência de *types* e *tokens* continua importante, havendo ali muitas oportunidades para comunicação e uso autêntico da língua, o que reflete as distribuições naturais de frequência [...]”. Pode-se aceitar, é claro, que seria mesmo impossível reproduzir com exatidão, na sala de aula, uma situação de comunicação da vida cotidiana, no entanto, aproximações verossímeis são possíveis, não há dúvida. Uma exposição planejada, por exemplo, pode direcionar o uso adequando-o através do emprego da linguagem mais significativa ao contexto de aprendizagem, a partir da análise dos índices de frequência apresentados em *corpus* de estudo composto por textos autênticos. Tal iniciativa pode tornar-se uma experiência pedagógica valiosa para o professor e para os alunos.

É importante rever o conceito de prática e sua relação com o de aprendizado de uma língua, principalmente ao se fazer a apologia do ensino dos termos mais recorrentes de uma dada língua, à repetição, à prática, à exposição frequente à língua alvo e ao ensino direto como modos para a aprendizagem. Segundo Gass e Sellinker (2008), “em anos anteriores, prática significou pouco mais do que mera decoreba ou exercícios de substituição. Em uma abordagem cognitiva do

aprendizado da língua, prática assume outras formas, porém o ingrediente comum é que o aprendiz interaja com a língua de maneira significativa (não somente saber de cor)[34]" (GASS e SELLINKER, 2008, p. 387). Em suma, a tecnologia da informação apresenta subsídios para qualificar a experiência do aprendiz de uma segunda língua, possibilitando acesso fácil e contato frequente com a língua "real" utilizada por uma determinada comunidade.

O destaque ao ensino do vocabulário, conforme a posição aqui assumida, implica o ensino de multipalavras, a utilização de textos autênticos, a seleção do vocabulário segundo a frequência e a probabilidade de ocorrência dos vocábulos num dado contexto, a relevância da instrução explícita a partir da intervenção do educador e da utilização de materiais adequados, a prática e o uso reiterado da língua. Em conjunto e, se harmonizados, dada a sua importância para o processamento psicolinguístico da língua estudada, são fatores imprescindíveis ao se estruturar um curso ou um programa de ensino de um idioma estrangeiro.

Além disso, os *lexical bundles* (pacotes lexicais) e as sequências formulaicas merecem destaque especial no ensino da leitura em língua inglesa, devido a sua alta frequência, o que vai ser tratado a seguir. Para atender a esse objetivo, a próxima seção deste estudo analisa evidências de que a mente humana processa a linguagem a partir da memorização dos itens lexicais e/ou sequências formulaicas.

2.3 Sequências formulaicas e processamento cognitivo

2.3.1 Sequências formulaicas e pacotes lexicais

Entender uma língua, segundo o quadro conceitual apresentado pela Linguística Cognitiva, significa também conhecer os padrões da língua em suas diversas manifestações: sequências formulaicas, lexical bundles, chunks, clusters, itens lexicais, entre outras denominações utilizadas. Atualmente, cada vez mais se destacam a importância e a centralidade do estudo das multipalavras na constituição

[34] "In earlier years, practice meant little more than rote repetition and/or substitution drills. In cognitive accounts of language learning, practice takes on a number of forms, but the common ingredient is that the learner interacts with the language in some meaningful (no solely rote) manner". (GASS e SELLINKER, 2008, p. 387).

da língua e no processamento linguístico. Este capítulo pretende apresentar resultados de investigações que trazem evidências tanto da existência das multipalavras quanto da importância desses agrupamentos lexicais no processamento mental. Ou seja, busca-se aqui estabelecer pontos de encontro entre as evidências já comprovadas pela Linguística de *Corpus* acerca da existência e recorrência dos agrupamentos lexicais, na fala e na escrita, com estudos que procuram analisar, por um viés psicolinguístico, a presença destas multipalavras no léxico mental.

Vale salientar que a variedade de vocábulos utilizados para definir o fenômeno da co-ocorrência de termos em uma dada língua poderia ser razão para confusão conceitual. Alison Wray (2000) realizou um levantamento bibliográfico e detectou o uso de mais de 50 expressões diferentes para nomear esse mesmo fenômeno. Todos os diferentes termos por ela arrolados, embora com oscilações de sentido, de uma forma ou outra, procuram definir a ocorrência das multipalavras. Conforme coloca Wood (2002), "há um consenso geral sobre a definição básica do que se constitui uma sequência formulaica e sobre quais características são compartilhadas e quais as distintivas". O ponto em comum compartilhado entre linguistas e pesquisadores de áreas correlatas está relacionado à suposição acerca da forma como as sequências formulaicas são cognitivamente processadas. "O consenso parece ser que as unidades da língua constituídas por multipalavras são armazenadas na memória de longa duração como se fossem um único item do léxico[35]" (Wood 2002, p. 2)

Douglas Biber et al. (1999) cunhou o termo *lexical bundles,* – pacotes lexicais, conforme tradução adotada[36] – para referir-se à recorrência de termos, tanto da língua falada quanto da produção escrita e assim o define:

[35] "There is general agreement on basic definition of what constitutes formulaic sequence and what characteristics such sequences share that make them distinct. The consensus seems to be that they are multiword units of language that are stored in long-term memory as if they were single lexical units." (Wood 2002, p. 2)

[36] Tradução utilizada por Tony Berber Sardinha e demais pesquisadores vinculados ao LAEL/GELC/PUCSP ao longo do IX ELC (Encontro de Linguística de *Corpus*, ocorrido em Porto Alegre, em outubro de 2010.

> Os pacotes lexicais são expressões recorrentes, independentemente de sua idiomaticidade ou de sua estrutura linguística. Isto é, os pacotes lexicais simplesmente são sequências de palavras que comumente ocorrem no discurso natural (BIBER et al., 1999, p. 990)[37].

Logo, a co-ocorrência de dois termos poderia ser critério para a definição dos pacotes lexicais. No entanto, devido à alta frequência de ocorrências de pares de palavras numa língua e, muitas vezes por tais pareamentos linguísticos não constituírem expressões de significado, bem como para delimitar sua pesquisa, Biber e colegas, estabeleceram como critério para a identificação de um pacote lexical a co-ocorrência mínima de três termos. "Para tornar o escopo de nossas investigações manuseável, um pacote lexical é definido aqui como uma sequência recorrente de três ou mais palavras[38]" (BIBER et al., 1999, p. 990). Biber argumenta a razão desta escolha, afirmando que

> Os pacotes lexicais são empiricamente identificados com as combinações de palavras que, de fato e mais usualmente, recorrem em um determinado gênero. Pacotes lexicais de três palavras podem ser considerados como uma espécie de extensão de uma colocação e por isso são extremamente comuns. Por outro lado, pacotes lexicais constituídos por quatro, cinco ou seis palavras são mais frasais e, correspondentemente, menos frequentes. Na conversação, há também pacotes lexicais formados pela contração de três termos que, na forma falada e contraída, são formados por duas palavras (exemplo: *she didn´t → she did not*). Na produção escrita, seriam expressos como três palavras separadas e, assim, essas duas palavras sequenciais contraídas na conversação podem ser comparadas aos pacotes lexicais compostos por três termos presentes na produção escrita acadêmica[39] (BIBER et al., 1999, p. 992).

A definição de pacote lexical, cunhada e explicitada por Biber, será adotada ao longo deste estudo por apresentar um critério objetivo para a definição de um padrão de co-ocorrências a ser investigado. No capítulo que diz respeito à metodologia,

---

[37] Tradução livre de Vitor Duarte. "Lexical Bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse" (BIBER et al., 1999, p. 990).

[38] "To make the scope of our investigations more manageable, a lexical bundle is defined here as a recurring sequence of three or more words" (BIBER et al., 1999, p. 990).

[39] "Lexical Bundles are identified empirically, as the combinations of words that in fact recur most commonly in a given register. Three-word bundles can be considered as a kind of extended collocational association, and they are thus extremely common. On the other hand, four-word, five-word, and six-word bundles are more phrasal in nature and correspondingly less common. In conversation, there are also recurrent two-word contracted bundles, which are composed of three grammatical word forms (*e.g. she didn´t → she did not*). In typical written prose, these would be expressed as three separate words; thus, these two-word contracted sequences in conversation might also be compared to three-word bundles in academic prose." (BIBER et al., 1999, p. 992)

serão detalhados os procedimentos e critérios propostos por esse linguista para formular a definição, bem como para localizar e selecionar os pacotes lexicais presentes em um dado *corpus*.

No entanto, na definição de pacote lexical acima proposta não há menção a aspectos cognitivos ou psicolinguísticos relacionados ao processamento mental, por parte do falante ou escritor, ao se utilizar dessas "palavras que parecem estar coladas entre si", de acordo com Ellis (1996). Simplesmente, parece não fazer parte do escopo do estudo de Biber questionamentos dessa ordem. Na tentativa de preencher a lacuna teórica existente e estabelecer uma ponte entre questões psicolinguísticas presentes no processamento da língua e sua relação com a alta recorrência dos pacotes lexicais na língua, tornou-se necessário adotar, paralelamente, uma definição que desse conta dos aspectos cognitivos. Para isso, será utilizada a definição "sequência formulaica" proposta por Wray e Perkins (2000):

> A minha definição de sequência formulaica é a seguinte: uma sequência, contínua ou descontínua de palavras ou outros elementos de significado, os quais são ou parecem ser pré-fabricados: isto é, memorizados e recuperados como uma unidade/totalidade da memória no momento em que são utilizados, ao invés de estarem sujeitos à produção ou análise pela linguagem gramatical[40] (WRAY e Perkins, 2000, p. 3).

Wray inclui na definição de sequência formulaica termos que, não necessariamente, aparecem lado a lado, e assim mesmo a constituem. Biber, por outro lado, considera apenas as co-ocorrências contínuas, que estão lado a lado no fluxo de um texto. Desse modo, optou-se por considerar o conceito apresentado por Wray como um hiperônimo que faz referência a todas as possibilidades de co-ocorrência, sejam contínuas ou descontínuas, e por enfatizar a possibilidade de a mente humana registrar e processar a sequência formulaica como uma totalidade. Ademais, Schmitt e Carter(2004) identificam características semelhantes às aqui apresentadas, ao reconhecerem o valor da definição de sequência formulaica e, segundo ele:

---

[40] "My definition of the formulaic sequence is as follows: a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p. 3).

> Este termo abrange uma grande variedade de linguagem formulaica e toca em dois critérios-chave pertinentes à ênfase deste livro: a) estamos preocupados com as sequências lexicais b) a mente lida, ou pelo menos parece lidar, com essas sequências como se fosse uma unidade lexical, em algum nível da representação[41]. (Schmitt e Carter, 2004, p. 3-4).

Pacote lexical e sequência formulaica são, pelas razões apresentadas, as duas expressões referenciais adotadas neste estudo. Ao se utilizar a expressão pacote lexical se estará referindo uma dada co-ocorrência detectada na análise de um *corpus*. Já a expressão sequência formulaica será utilizada como indicador de algum item lexical, ou seja, de determinado pacote lexical poder também fazer parte do léxico mental do indivíduo. Na definição de Wray e Perkins (2000), por ser mais abrangente, pode-se incluir o conceito de pacote lexical. Biber (1999), por outro lado, não estabelece relação entre a co-ocorrência de termos e o processamento psicolinguístico. Assim, justifica-se essa dupla escolha conceitual que, conforme se entende, parece ser complementar.

A escolha das duas denominações adotadas aqui– sequência formulaica e pacote lexical - também ocorreu, em parte, por ainda existirem limitações tanto na análise de um *corpus* linguístico quanto no conhecimento sobre o processamento linguístico. Ainda não há como saber com precisão, se os pacotes lexicais mapeados em um *corpus* são necessariamente sequências formulaicas, e se assim são processados pela mente. Um limitador da tecnologia é ainda não possibilitar esta certeza e sobre esse ponto Schmitt e Carter (2004) elucidam que "os modernos concordanciadores são bons em identificar sequências contíguas, mas ainda não temos um software que possa identificar sequências formulaicas flexíveis a partir da análise de *corpora*[42]". Mesmo com a grande eficiência com que são realizados os mapeamentos das sequências de um *corpus*, não há ainda condições de afirmar com plena certeza que elas são assim processadas pela mente. Por outro lado, a

[41] This term covers a wide range of formulaic language, and touches on two keys criteria of the emphasis in this book: a) we are concerned with sequences of lexis and b) the mind handles, or appears to handle, these sequences at some level of representation as wholes. (Schmitt e Carter, 2004, p. 3-4).

[42] "Modern concordancers are good at identifying contiguous sequences, but we do not yet have software which can identify flexible formulaic sequences automatically from *corpora*. Once this software is developed, we may find that flexible formulaic sequences are even more prevalent than totally fixed ones" (Schmitt & Carter., 2004, p. 7).

tecnologia ainda não possibilita o mapeamento exato e preciso das áreas da linguagem no cérebro, a fim de entender como ocorre o processamento mental das sequências formulaicas. Conforme é muito bem ponderado por Schmitt, Grandage e Adolphs (2004) é necessário ainda ser muito cauteloso a respeito da questão, não formulando quaisquer conclusões apressadas:

> Os dados de um *corpus* são extremamente úteis para identificar *clusters* recorrentes na linguagem. Isso continuará a ser de grande utilidade para aplicações no campo da linguística aplicada. No entanto, este estudo sugere que os dados de um *corpus* por si próprios não são bons indicadores se, realmente, os *cluster*s encontrados em um *corpus* são representados como unidades lexicais na mente humana. Parece haver um consenso tácito de que os dados de um *corpus* são de alguma forma também válidos por um viés psicolinguístico, e em muitos sentidos isto deve ser verdadeiro pois a língua representada em *corpora* foi produzida por usuários da língua e isso reflete, em certa medida, a sua competência linguística. Entretanto, este estudo sugere que é imprudente considerar a recorrência dos *clusters* de um *corpus* como evidência de que tais *clusters* estão representados na mente como sequências formulaicas. Abordagens baseadas em *corpus* e estudos psicolinguísticos são intercomplementares, e parece não ser novidade que precisamos de ambas as abordagens a fim de explicar como a língua é processada e usada[43]" (SCHMITT, GRANDAGE e ADOLPHS, 2004, p. 146-147).

Mesmo assim, há vários indícios de que é provável que pelo menos uma parte dos pacotes lexicais mapeados em um *corpus* sejam também sequências formulaicas processadas pela mente humana. Afinal, a sua presença e alta recorrência em um *corpus* linguístico não é produto do acaso. Neste sentido, pesquisadores têm buscado instrumentos e metodologias de pesquisa diversos, tanto da Linguística de *Corpus* quanto da Psicolinguística, com o objetivo de buscar maiores evidências a esse respeito. Mais ainda, tais estudos têm sido realizados, intercomplementando o conhecimento produzido pelas duas metodologias. Logo

[43] "*Corpus* data is very useful in identifying recurrent clusters in language production. This will continue to be of considerable use in applied linguistic applications. However, this study suggests that *corpus* data on its own is a poor indicator of whether those clusters are actually stored in the mind as wholes. There seems to have been an unspoken assumption that *corpus* data is somehow psycholinguistically valid, and in many senses this must be true because the language in *corpora* has been produced by people using language and so must reflect language competence to some extent. However, this study suggests that it is unwise to take recurrence of clusters in a *corpus* as evidence that those clusters are also stored as formulaic sequences in the mind. *Corpus* and psycholinguistic approaches complement each other, and unsurprisingly it seems we need both in order to explain how language is processed and used." ." (SCHMITT, GRANDAGE & ADOLPHS, 2004, p. 146-147

adiante, serão apresentadas algumas pesquisas que têm contribuído para o entendimento do processamento mental das sequências formulaicas.

2.3.2 Memória, pausas e sequência formulaica: processamento mental de sequências formulaicas

Estudos diversos (Wray, 2002; Wray e Perkins, 2000; Schmitt, 2004; Erman, 2007, Schmitt (2004), etc) apontam que o uso das sequências formulaicas pode estar relacionado ao processamento humano da memória. Ao utilizar esses padrões fraseológicos, a memória de trabalho é liberada para outras funções do processo de compreensão e interpretação. Tal pressuposição torna-se um forte argumento na defesa da tese de que o uso predominante das fórmulas linguísticas na comunicação está atrelado à capacidade de processamento da memória humana. Segundo Perkins (1999, p. 56) parece ser razoável que "a principal razão para a predominância das sequências formulaicas, no sistema de linguagem do adulto, parece ser o princípio da economia de esforço no processamento[44]" (Perkins, 1999, p. 56 apud Wray, 2002, p. 15-16). Para Becker (apud Wray, 2002, p. 16), essa economia ocorre por permitir o uso de estruturas prontas, as quais permitem que as ideias possam ser rapidamente processadas, sem a necessidade de estar sempre iniciando, do zero, o trabalho de formular e gerar novos enunciados.

Pressuposições teóricas relacionadas à base biológica e neurológica da atividade mental humana são apontadas por Wray ao afirmar que "uma das razões do déficit entre a capacidade gramatical e a capacidade de processamento on-line está relacionada às limitações da memória de curto prazo" (Wray, 2002, p. 16). De sua parte, Miller (1956), Bower (1969) and Simon (1974) mostraram "como a organização da informação em uma única e complexa unidade aumenta a quantidade total de material que pode ser armazenado na memória de curto prazo

[44] "In this light, it seems reasonable that "the mainsreason for the prevalence of formulaicity in the adult language system appears to be the simple processing principle of economy of effort! (Perkins, 1999:56). This economy occurs because it gives us access to "ready-made frameworks on which to hang the expression of our ideas, so that we do not have to go through the labor of generating an utterance all the way out from 'S' every time we want to say something" (Becker 1975:17) (Wray, 2002, p. 15-16)

ou na memória de trabalho[45]" (apud Wray, 2002, p. 16). Code (1994, 139-140) de forma incisiva argumenta que "seria fisiologicamente impossível falar, de forma proficiente e rápida, da maneira como o fazemos, se tivéssemos que planejar e executar cada segmento individualmente[46]" (apud Wray, 2002, p. 16). Ou seja, a fluência do ato de comunicação parece vincular-se tanto ao uso das sequências formulaicas como à capacidade de processamento da memória.

Parte-se do entendimento de que a memória constitui-se de três subsistemas – memória de trabalho, memória declarativa e memória de produção (working memory, declarative memory e production memory (1983, apud Erman, 2007, p. 29). Ao considerar a proposta de Anderson, Erman considera que as estruturas pré-fabricadas são armazenadas, como uma única unidade lexical, na memória declarativa (ou de longo prazo). Ela explica que, de acordo com a teoria proposta por Anderson, há três rotas para ativar os conceitos na memória declarativa: percepção, ativação distribuída e focalização (*perception, spreading activation* e *focusing*), porém apenas a ativação distribuída e a focalização dizem respeito a seu foco de estudo. Conforme ela alega, as estruturas pré-fabricadas são, talvez, ativadas de maneira distribuída: um conceito conecta outro conceito associado, de forma contínua.

Antes de prosseguir, convém atentar para o conceito de pausa, o qual é basilar para a compreensão da tese proposta. Segundo Erman (2007) e outros pesquisadores:

> O pressuposto básico de que as pausas, durante a produção, são involuntárias, tem como base o fato de que elas interrompem o fluxo da fala e aparecem em unidades sintáticas elementares. As pausas relacionadas à produção, que indicam o esforço cognitivo durante o processamento, foram identificadas com base na sua posição em relação ao contexto de uso, ocorrendo tipicamente no limiar de uma expressão ou frase, depois de um termo gramatical ou funcional (function Word) tais como um determinante,

[45] "One of the explanations for the shortfall between grammatical capability and on-line processing capability is limitations in short-term memory. […]Miller (1956), Bower (1969) and Simon (1974) have shown how chunking information into single complex units increases the overall quantity of material that can be stored in short-term or working memory. Ellis and Sinclair (1996) note that a person's phonological working memory span correlates with his or her language learning capacity." (Wray, 2002, p. 16).

[46] "It would be physiologically impossible for us to produce speech with the rapidity and proficiency that we are able to if we had to plan and perform each segment individually" (Code 1994:139-140 apud Wray, 2002, p. 16)

> um auxiliar, uma preposição ou antes de uma palavra de significado. Estudos têm mostrado que as pausas ocorrem com muito maior probabilidade antes de palavras de significado do que de termos gramaticais ou funcionais (function Word), porque quando "o falante chega ao ponto em que uma palavra de significado é requerida, ele naturalmente tem uma opção de escolha muito maior do que se fosse requerido um termo gramatical (function Word) (Butterworth 1980:167). Maclay e Osgood (1959:31) constataram que falantes têm a tendência a hesitar antes de começarem a falar uma expressão e que eles tendem a relutar novamente antes de proferir a palavra de significado nessa frase ou expressão. Boomer, de forma semelhante, enfatiza que "hesitações em orações fonêmicas[47] são muito mais prováveis de ocorrer **depois** de pelo menos uma decisão preliminar ter sido feita no que diz respeito à própria estrutura e **antes** de palavras de significado terem sido finalmente emitidas[48]" (Boomer 1965:152-3) (ERMAN, 2007, p. 32).

A ativação distribuída é um processo inerentemente automático e sem a manifestação de pausas. Ou seja, o usuário da língua emite uma sequência formulaica 'automaticamente' de forma contínua e fluida, sem intervalos ou pausas (silêncios, respirações, emissões sonoras glotais ou guturais, etc.) entre os termos que a compõem, independentemente de esta estrutura ser composta por dois, três ou mesmo quantidade superior de palavras. "A focalização, por contraste, demanda esforço e ocorre a partir da inserção de pausas" (Shilperhood, 1996 apud Erman, 2007, p. 29). Durante a focalização, a fala é produzida de forma menos fluida e precisa ser articulada, pausas são inseridas entre termos que vão sendo formados à medida que o falante compõe o enunciado, buscando em sua memória vocábulos que possam se encaixar na estrutura linguística à medida que esta vai sendo elaborada.

---

[47] "A expressão oração fonêmica é uma unidade que foi usada originalmente na PSICOLINGUÍSTICA, nas pesquisas sobre a distribuiçãio e a função das PAUSAS: refere-se à estrutura GRAMATICAL produzida dentro de um único CONTORNO DE ENTONAÇÃO e limitado por JUNTURAS". (Dicionário de LINGUÍSTICA E FONÉTICA. David Crystal Rio de Janeiro: Jorge Zahar Ed., 2000.

[48] "The basic assumption that production pauses are involuntary is supported by the fact that they interrupt the flow of speech and appear in lower-level syntactic units. Production pauses reflecting cognitive processing effort were identified on the basis of position and surrounding context, typically occurring within the boundaries of the phrase, after a function word, viz., a determiner, an auxiliaury, a preposition, or before a content word. Studies have shown that pauses are much more likely to occur before content words than function words, because "when the speaker comes to a point where a content word is required he naturally has a much wider choice than where a function word is required" (Butterworth 1980:167). Maclay and Osgood (1959:31) found that speakers tend to hesitate before they begin a phrase, and then they tend to hesitate again before uttering the lexical word in that phrase. Boomer likewise emphasizes that "hesitations in phonemic clauses are most likely to occur **after** at least a premilinary decision has been made concerning its structure and **before** the lexical choices have been finally made" (Boomer 1965:152-3) (ERMAN, 2007, p. 32)

Erman alega que as duas rotas que ativam os conceitos na memória declarativa – a ativação distribuída e a focalização – relacionam-se diretamente a dois princípios propostos por Sinclair (1991): o princípio idiomático (*idiom principle*) e o princípio da livre escolha (*open choice principle)* "e de fato poderia se dizer que lhes agregam força explanatória" (ERMAN, 2007, p. 29). Segundo Erman (2007) "o princípio idiomático pode estar relacionado à ativação distribuída e o princípio da livre escolha pode envolver um estado de focalização, embora, claro, nem sempre assim ocorra[49]" (ERMAN, 2007, p. 29).

A partir de experimento realizado por Erman (2007), "os resultados claramente indicam que é raro ocorrerem pausas entre os componentes dos *prefabs"* (ERMAN, 2007, p. 47). Por outro lado, a pesquisadora afirma que os resultados desse estudo "mostram que a grande maioria de pausas de longa duração ocorrem quando um falante procura uma palavra para preencher um espaço aberto na estrutura e isso acontece precisamente quando se espera que a focalização aconteça[50]" (ERMAN, 2007, p. 47). Segundo ela, no estudo comentado, as pausas ocorreram quase que exclusivamente em situações de escolha, seguindo o princípio da livre escolha, fornecendo manifestações claras de processos cognitivos ocorridos em função das pausas. Ao comparar dois grupos etários diferentes, Erman (2007) verificou que 90,6% das pausas produzidas por adultos e 84,6% das pausas produzidas por adolescentes ocorreram quando o usuário analisava a língua, ou seja, quando o *open choice principle*, tal como descrito por Sinclair, se manifestava.

Schmitt e outros pesquisadores (2004) apontam diferenças significativas entre a quantidade de pausas produzidas na fala de um nativo de língua inglesa e falantes aprendizes da língua inglesa, na utilização de sequências formulaicas. "Considerando como uma característica a habilidade dos candidatos para reproduzir as sequências formulaicas fluentemente, é importante notar que os falantes não-nativos produziram hesitações, gagueiras e falsos inícios duas vezes mais que os

[49] The idiom principle is assumed to be at work in connection with spreading activation, and the open choice principle could be said to involve a state of focusing, although of course far from always (ERMAN, 2007, p. 29).

[50] "The results of the present study shows that the vast majority of pauses of longer duration occur when a speaker does word search to fill an open slot and this is precisely where we expect focusing to be at work" (ERMAN, 2007, p. 47).

falantes nativos participantes do estudo[51]"(SCHMITT, GRANDAGE e ADOLPHS, 2004, p. 143). Tais dados corroboram os resultados das demais pesquisas, indicando a pausa como uma evidência do processamento da fala de sequências formulaicas.

Essas conclusões dão suporte ao princípio idiomático proposto por Sinclair, segundo o qual "a pouca frequência de pausas entre as multipalavras parece indicar que essas unidade lexicais não são analisadas durante a sua produção[52]" (ERMAN, 2007, p. 48). A autora chega à conclusão que a combinação de vários itens lexicais constituindo uma única multipalavra – ou uma sequência formulaica – está diretamente associada à ideia de que a ativação distribuída facilita a produção de associações automáticas, em consonância com a teoria sobre a memória, tal como entendida por Anderson (1983, apud ERMAN, 2007, p. 48).

2.3.3 Mais evidências da Linguística Cognitiva

Outros estudos, além das pesquisas pertinentes ao uso das pausas na fala, têm encontrado evidências de que as sequências formulaicas são processadas, holisticamente, pela mente. Há dados empíricos provenientes de estudos fonológicos, de estudos de uso da língua em situações que exigem habilidade comunicativa do falante, bem como de pesquisas acerca do processamento de leitura que contribuem favoravelmente para a consolidação dessa hipótese.

David Wood (2002) chama a atenção para um aspecto intrigante relacionado à memorização e à recuperação das sequências formulaicas: além de serem memorizadas de forma holística, são também articuladas como unidades, o que "viabiliza a coerência fonológica que é uma característica de sua produção". Wood destaca, com base em Bolander (1989), que a qualidade fonológica alcançada na produção das sequências formulaicas pode decorrer justamente do fato de seu processamento ocorrer dessa forma – unitariamente. Conforme Ladefoged:

---

[51] "As a feature of the candidate's ability to reproduce the strings fluently, it is worth noting that the non-native speakers displayed hesitations, stutters and false starts in twice as many strings as the native speaker participants." (SCHMITT, GRANDAGE & ADOLPHS, 2004, p. 143)

[52] "[...]the low frequency of pauses in multi-word units seems to indicate that they remain unanalyzed in production.

> Há várias evidências de que os movimentos musculares são organizados em função dos complexos e fixos *chunks* (sequências formulaicas, grifo meu) de pelo menos um quarto de segundo de duração (e frequentemente mais longos) e nada indica que segmentos pequenos e simultâneos , que necessitam ser processados a partir de regras rígidas contextuais, organizem-se assim. (Ladefoged, p. 85 apud Wood, 2002, p. 8).

Schmitt, Grandage e Adolphs (2004) também apresentam dados indicadores advindos de outros estudos comprovando que as sequências formulaicas são articuladas de maneira fluente, como se fossem uma única palavra, ou seja, "com uma entonação normal, quer dizer, uso natural de acentuação tônica, tom e encadeamento sonoro dos termos" (van Lancker, Canter, e Terbeek, 1981 apud Wood, 2002, p. 131). E, qualquer desvio deste padrão, "por exemplo, uma hesitação entre as palavras de um cluster, tal como a pausa de um segundo de tempo, sugere que o *cluster* não foi holisticamente armazenado na memória[53]" (SCHMITT, GRANDAGE e ADOLPHS, 2004, p. 131)

Kuiper (2004) relaciona o uso das sequências formulaicas com situações de comunicação onde, segundo ele afirma, pode haver uma grande demanda no uso da memória de trabalho, condição em que a memória encontra-se sob grande pressão. Narradores de corridas de cavalo, como de outras atividades esportivas, leiloeiros, locutores de rádio são alguns dos exemplos de nichos específicos – ou tradições orais específicas – nas quais é comum o ato comunicativo acontecer sob muita pressão. Tais situações demandam, por parte do comunicador, rapidez e agilidade na utilização da língua para que possa se comunicar adequadamente com sua audiência.

Referindo-se a situações desse tipo ou a outras assemelhadas, Kuiper (2004) afirma que "a memória, claramente, está sob grande pressão; porém, o comentário

[53] More importantly, it has been noted that formulaic sequences are typically articulated in a fluent manner (e.g. van Lancker, Canter, and Terbeek, 1981), with a 'normal' intonation contour, that is, with a natural pitch, stress and juncture profile. This has been accepted as one of the criteria of formulaicity (e.g. Pawley and Syder, 1983; Peters, 1983) and any deviation from this profile (e.g. a hesitation between words with a cluster: *as a matter* (1 second pause) *of fact* suggests that the cluster is not stored holistically (although note that other explanations are possible: see Rosenberg, 1977). Thus, although is admittedly not a direct measure of holistic storage, in this study we take fluently-articulated reproduction of the recurrent clusters embedded in the dictation contexts as evidence that they are likely to be holistically-stored formulaic sequences". (SCHMITT, GRANDAGE & ADOLPHS, 2004, p. 131)

fluente é mantido totalmente sob controle através da utilização de sequências formulaicas tradicionais da fala[54]" (KUIPER, 2004, p. 40). O conhecimento da tradição cultural do nicho desempenha um papel importante: por um lado permite que o narrador ou comunicador, em situação que demandam alta pressão e raciocínio rápidos, possa efetivamente transmitir sua mensagem a partir de uma mensagem altamente convencionalizada – as sequências formulaicas. "O resultado é um comentário completamente fluente, sem nenhum tipo de hesitação, falsos inícios, pausas expressas ou mudas[55]" (KUIPER, 2004, p. 42). Por conseguinte, toda uma comunidade – muitas vezes, também pressionada emocionalmente pela situação – rapidamente compreende os comandos ou informações transmitidas pelo narrador ou comunicador. Segundo Kuiper (2004) este aprendizado leva muitos anos e

> é também circunscrito pelas restrições de nossa capacidade para produzir e entender linguagem, causados pelas limitações dos recursos da nossa memória. Nós temos uma grande capacidade para relembrar e rapidamente resgatar da memória o que precisamos. Temos, porém, uma capacidade de processamento relativamente limitada pelo fato de nossa memória de trabalho ser muito pequena. As sequências formulaicas viabilizam-nos controlar esses recursos de uma maneira eficiente, desde que o que precise ser dito não necessite ser feito de uma forma radicalmente original. A maior parte daquilo que falamos no curso normal dos eventos de nossas vidas não o é[56]. (KUIPER, 2004, p. 51-52)

2.3.4 Sequência formulaica: função social e fluência comunicativa

As sequências formulaicas, conforme anteriormente comentado, possuem alta frequência tanto no discurso oral como na produção escrita, e há razões tanto de ordem cognitiva como de ordem social que contribuem para isso. De um lado, o uso

---

[54] Clearly there is significant pressure on memory, but fluent commentary is maintained through the utilization of a totally formulaic speech tradition. (KUIPER, 2004, p. 40)

[55] The result is a highly fluent commentary without any hesitation phenomena such as false starts or pauses voiced or unvoiced." (KUIPER, 2004, p. 42)

[56] It is also circumscribed by the constraints on our ability to produce and understand language caused by the limitations of our memory resources. We have an immense capacity to remember, and to retrieve very quickly from memory what we need. But we have relatively restricted processing capacities because our working memory is quite small. Formulaic speech enables us to harness theses resources in an efficient way so long as what we wish to say does not need to be radically novel. Much of what we say in the normal course of social events is not." (KUIPER, 2004, p. 51-52)

das sequências formulaicas significa um auxílio no processamento cognitivo, liberando a carga e o esforço da memória de trabalho. Por outro lado, o uso das fórmulas tem alta relevância na interação social, sendo, muitas vezes a moeda de troca que garante a comunicação fluente entre os membros de uma comunidade. Wray e Perkins (2000) propõem uma maneira de acomodar essas duas funções em um único modelo. "Quando um indivíduo escolhe uma sequência pré-fabricada da língua a fim de reduzir a pressão durante o processamento linguístico, o objetivo é ser fluente e proceder, sem interrupções, na produção da mensagem completa ou, ainda, para garantir que a informação esteja facilmente acessível quando necessária". Há uma forte inter-relação entre a redução da sobrecarga da memória pela utilização das sequências formulaicas e a fluência linguística. "[...]Em outras palavras, a seleção de uma sequência formulaica no contexto sócio-interacional objetiva que se maximizem as chances de compreensão efetiva[57]" (Wray, 2000, p. 477).

A noção de fluência em um idioma passa necessariamente pelo uso contínuo e predominante dos pacotes lexicais. Ao fazer uso deles, a atenção é liberada para a interpretação dos enunciados, não exigindo que o falante focalize sua habilidade analítica na produção de novas expressões e preste atenção às regras que compõem o pacote linguístico. Com a liberação da memória, é possível concentrar-se na avaliação das proposições emitidas pelo interlocutor, na atualização das informações contextuais do ato comunicativo e também na realização de predições sobre o que poderá em seguida acontecer no processo comunicativo (Wray, 1992 apud Wray e Perkins, 2000, p. 19). Em outras palavras, ao utilizar as fórmulas linguísticas recorrentes no contexto em que está inserido, o indivíduo libera sua memória e atenção para a interpretação, para a interação comunicativa. Wray e Perkins contextualizam informações acerca do processamento cognitivo, a partir de suas análises, bem como de dados aportados por outros pesquisadores:

[57] When na individual chooses a prefabricated stretch of language in order to reduce the pressure on processing, the aim is to be fluent and to suceed in producing the entire message without interruption, or to ensure that information is reliably to hand when needed. (…) "In other words, the selection of a formulaic sequence in the socio-interactional context aims to achieve the maximum chance of efficient comprehension." (Wray 2000:477)

> Parece que empregamos sequências pré-fabricadas como uma forma de minimizar os efeitos do descompasso entre nossas potencialidades linguísticas e a real capacidade de nossa memória de curto prazo. Conforme aponta Becker (1979), faz pouco sentido produzir do nada sequências linguísticas usadas com frequência. Parece que utilizamos as sequências formulaicas para reduzir a quantidade de novo processamento, permitindo que somente aquilo que realmente tem que ser novo seja realmente processado como tal. Pesquisa recente (Raichle, 1998; Mc Crone, 1999) mostra que uma vez que o cérebro se familiariza com uma tarefa linguística, é capaz de ignorar a rota processual utilizada para aprendê-la. (Wray & Perkins, 2001, p. 15-16). (...) As palavras podem se dispor (agrupar) de modo a formar expressões que poderiam, em princípio, significar diferentes coisas, mas geralmente são interpretadas de uma única forma convencionalizada (e.g. bullet point); em alguns casos podem até se tornar clichês (e.g. the current economic climate). Por isso, os benefícios da linguagem pré-fabricada, ao reduzir o esforço no processamento, são considerados uma explicação para o fato de um indivíduo, ou mesmo toda uma comunidade, preferir certas colocações e expressões para comunicar uma ideia a outras que poderiam igualmente ser mobilizadas[58] (Pawley & Syder, 1983). (Wray & Perkins, 2000, p. 15-16).

Munido desse conhecimento linguístico, o aprendiz usaria a língua de forma mais efetiva, pois a fluência em uma dada língua nativa associa-se ao uso da linguagem formulaica. Pawley and Syder (1983), afirmam que um aprendiz de nível avançado de um idioma diferencia-se de um falante nativo justamente por não dominar as fórmulas linguísticas por ele utilizadas. Wray (2000) diz que tornar-se fluente em uma nova língua "requer que o aprendiz torne-se sensível às preferências dos falantes nativos por certas sequências de palavras em detrimento de outras que poderiam parecer tão possíveis quanto as usadas[59]" (Wray, 2000, p 463). Wood (2002), da mesma forma, relaciona a fluência linguística ao uso das sequências formulaicas, pois "uma grande proporção dos conceitos e dos atos de fala mais

[58] It seems that we use prefabricated sequences as a way of minimizing the effects of a mismatch between our potential linguistic capabilities and our actual short term memory capacity. As Becker (1979) points out, it makes little sense to produce from scratch word strings which we use many times, and we appear to use formulaic sequences to reduce the amount of new processing to only that which has to be new. Recent research (Raichle, 1998; Mc Crone, 1999) shows that once the brain is familiar with a linguistic task, it is able to by-pass the processing route that was used to learn it. (Wray & Perkins, 2001, p. 15-16) (…)Words may collocate to form phrases which could, in principle, mean several different things, but which are only normally interpreted in one agreed way (e.g. bullet pointi); in some cases there may even become clichés (e.g. the current economic climate). Thus, the benefits of prefabricated language in reducing processing effort can account for why an individual or indeed a whole community comes to prefer certain collocations and expressions of an idea over other equally permissible ones (Pawley & Syder, 1983)" (Wray & Perkins, 2001, p. 15-16).

[59] "Gaining a full command of a new language requires the learner to become sensitive to the native speakers' preference for certain sequences of words over others that might appear just as possible" (Wray, 2000, p. 463).

conhecidos são expressos formulaicamente e, se o falante for capaz de resgatá-los da memória, como unidades, a fluência estará reforçada[60]"

2.3.5 Leitura e sequências formulaicas

Embora até este ponto tenha se discutido e mostrado evidências da função das sequências formulaicas, preponderantemente, na comunicação oral, a seguir serão analisados resultados trazidos por outras pesquisas no que tange à produção escrita em língua inglesa. Tais pesquisas apresentam resultados semelhantes no que se refere ao desempenho cognitivo no processamento da leitura, corroborando os dados já apontados por Sinclair (1997), Wray e Perkins (2000), Erman e Warren (2000), Wood (2002). "Desde que praticamente todos os textos contêm estruturas pré-fabricadas e que pré-fabricados, compostos pela colocação de dois ou mais termos, podem ser considerados como uma única palavra - uma multipalavra - acessada do léxico mental, a quantidade de acessos tem que ser menor do que o número de palavras num texto de qualquer tamanho[61]" (ERMANN e WARREN, 2000, p. 48). Infere-se a partir daí que o aprendiz que conhecer as fórmulas linguísticas recorrentes, constantes em textos escritos, será beneficiado. Entende-se que o benefício se daria por razões semelhantes às apontadas por Wray e Perkins, nos estudos acerca da comunicação oral: liberação da memória para o processamento da informação e a possibilidade de ter sua fluência desenvolvida. Um aspecto sociológico, frente ao uso das sequências formulaicas, também adquire destaque: ao dominar o linguajar, o jargão ou a linguagem de um nicho específico, o aprendiz terá maior probabilidade de ser considerado um membro daquele grupo, como alguém capaz de partilhar o mesmo sistema simbólico.

Erman e Warren (2000) demonstram a frequência dos elementos pré-fabricados na comunicação escrita: "Nós descobrimos que em média algo um pouco acima da metade (em torno de 55%) de um texto é constituído por linguagem pré-

[60] "A great proportion of the most familiar concepts and speech acts can be expressed formulaically, and if a speaker can pull these readily from memory, as wholes, fluency is enhanced" (Wood, 2002).
[61] "Since practically all texts contain prefabs and since prefabs can be assumed to constitute single multi-word retrievals from our mental store of words, the number of retrievals must be fewer than the number of words in a text of some size" (ERMANN & WARREN, 2000, p. 48).

fabricada[62] [63]" (ERMAN e WARREN, 2000, p. 50). Segundo as pesquisadoras, esses dados "dão forte sustentação ao *idiom principle* (princípio idiomático), como formulado por Sinclair, e reiteram que a proporção de pré-fabricados na língua, de forma geral, tem sido subestimada"(ERMAN e WARREN, 2000, p. 50).

Nattinger e DeCarrico (1992), ao refletirem sobre as estratégias de leitura utilizadas por aprendizes de uma segunda língua, retomam um ponto com o qual provavelmente todo professor de língua inglesa já se defrontou: aprendizes iniciantes, principalmente, durante a leitura, tendem a fixar-se muito mais no texto em si, ou seja, utilizam muito mais estratégias locais e dão muito mais atenção à superfície do texto do que um leitor nativo. Em outras palavras, tendem a ler o texto palavra por palavra na tentativa de vasculhar sua memória em busca de um sentido para cada palavra ali impressa. Ou seja, "leitores de uma L2, com frequência, utilizam uma estratégia *bottom-up* para processar o texto, focando nas características estruturais presentes na superfície textual e elaborando a compreensão através de análise e síntese desta informação visual[64] (*visual input*)" (Nattinger e DeCarrico, 1992, p. 159).

Muitos aprendizes, por não dominarem a quantidade mínima indispensável de conhecimento linguístico da língua alvo e talvez também por não conhecerem outros procedimentos e estratégias de leitura levam muito tempo, tentando deduzir (ou mesmo adivinhar) o significado de um vocábulo ou, mesmo, utilizam em demasia o dicionário, recorrendo a ele a cada vez que se deparam com uma palavra. Com a superutilização do dicionário, ou outro referencial externo ao texto, para a identificação do vocabulário, o fluxo da leitura é interrompido em muitos pontos, dificultando o entendimento da temática textual. Em outros termos, nesses casos a memória de trabalho está sobrecarregada com a atividade de decodificação mais

[62] As autoras utilizam o termo pré-fabricado como um quase-sinônimo para pacote lexical (cluster/bundle). Segundo elas: Um pré-fabricado é uma combinação de pelo menos duas palavras selecionadas por uma falante nativo, em preferência a alguma combinação alternativa que poderia ser equivalente, caso não existisse nenhuma convencionalização. "A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have benn equivalent had there been no conventionalization". (BITT ERMAN; BEATRICE WARREN, 2000)

[63] "We have found that on average somewhat more than half (around 55 percent) of a text will consist of prefabricated language" ERMAN & WARREN, 2000, p. 50).

[64] L2 readers often process text in a bottom-up manner, focusing on surface structure features and building comprehension through analysis and synthesis of this visual input (Nattinger & DeCarrico, 1992, p. 159).

elementar, na tentativa de reconhecer os vocábulos e a eles associar um sentido, com isso não conseguindo o leitor avançar na leitura: a compreensão global do texto, o estabelecimento de relações entre partes do texto e outros textos, entre outras atividade cognitivas comuns na leitura fluente. Com a sobrecarga da atividade cognitiva concentrada no decifrado de elementos mínimos, esses leitores não chegam a perceber e nem a identificar os pacotes lexicais que, na maior parte das vezes, têm significados diferentes daquele da soma do significado de que cada um de seus termos individuais e, lamentavelmente, acabam não sendo capazes de realizar uma interpretação coerente do texto. Nattinger e DeCarrico (1992) explicitam os movimentos cognitivos presentes durante a leitura, quando os pacotes lexicais não são processados pelo leitor, reiterando que:

> Quando a leitura das sequências formulaicas não é realizada, menos informação pode ser armazenada, de uma única vez, na memória de curto prazo. Tal redução na capacidade de armazenamento significa que uma quantidade menor de dados linguísticos poderá ser analisada simultaneamente, o que implica um uso ineficiente da redundância e das pistas contextuais. Por causa das limitações da atenção humana e do processamento da capacidade da memória, essas demandas cognitivas adicionais podem ser responsáveis pelo fato de que bons leitores de L1, com frequência, não são capazes de aplicar as habilidades de leitura de sua primeira língua na leitura de textos na L2[65] (Kern, 1989) (Nattinger & DeCarrico, 1992, p. 159-160).

A partir da revisão dos estudos e das pesquisas aqui comentados, pode-se presumir, com relativa segurança, que tanto na fluência leitora, como na fluência oral da língua inglesa, pode haver forte correlação entre capacidade de processamento da memória de trabalho e conhecimento de sequências formulaicas. Estudos realizados pelo grupo de pesquisadores da Universidade de Nottingham, vinculados a Norbert Schmitt, têm procurado evidências a respeito. Esses estudos envolvem comparação entre tempo de leitura e processamento da leitura realizada por leitores de L1 comparados com leitores de L2 e, também, estudos sobre a movimentação ocular durante a leitura.

[65] When chunking is impeded, less information can be stored at one time in short-term memory. Such a reduction in storage capacity means that less linguistic data can be analyzed simultaneously, which results in inefficient use of redundancy and contextual cues. Because of limitations in human attention and memory processing capacity, these additional cognitive demand may account for the observation that good L1 readers are often not able to apply their reading skills to L2 texts (Kern, 1989). (Nattinger & DeCarrico, 1992, p. 159-160)

Parte das evidências de que a mente organiza e processa a língua por meio de agrupamentos lexicais decorre de estudos sobre o movimento dos olhos. Segundo Underwood, Schmitt e Galpin (2004, p. 154), “o apelo em medir o movimento ocular está no fato de que os olhos dão uma indicação sobre que processos estão ocorrendo na mente do leitor”. Em decorrência, as informações existentes se apóiam em relatos de estudo que mostram que o número de regressões (*regressions*) e fixações (*forward fixations*) aumentam em quantidade e tempo de duração, de acordo com o nível de dificuldade do texto. “Mais ainda, maus leitores tendem a realizar mais fixações regressivas (*regressive fixations*) na leitura de um texto do que bons leitores” (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154).

Assim, como afirmado no parágrafo anterior, durante a leitura, os olhos realizam dois tipos básicos de movimento: fixações e sacadas. As fixações são as paradas – imperceptíveis, sem o uso de um instrumento preciso de registro – que os olhos realizam em determinadas partes de um texto. Assim, através das sacadas ocorrem os movimentos de avanço ou recuo dos olhos no texto. Esses pequenos saltos que os olhos dão, de uma palavra a outra, ou mesmo de um intervalo maior de palavras, traduzem o ritmo evolutivo da leitura. As sacadas avançam, quando a leitura segue seu fluxo contínuo; recuam, quando o leitor precisa reler algo; podem ser intencionais ou, ao contrário, movimentos mecânicos e inconscientes desempenhados pelo aparato biológico da leitura. Há evidências que sustentam que “as fixações são indicadores ‘*on-line’* da dificuldade de leitura”, pois quando os olhos se fixam “em uma parte relativamente importante do campo, nossos olhos permanecerão estáticos, durante determinado tempo, o que é indicativo do aumento da quantidade de processamento que está ocorrendo” (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154). A versão extrema dessa visão teórica sustenta que as palavras que não são fixadas, não são processadas. Conforme o afirmam Underwood, Schmitt e Galpin, (2004, p. 154), a teoria de Just e Carpenter (1980), abaixo referida, baseia-se em duas premissas: a pressuposição imediata *(immediacy assumption*) e a suposição nomeada olho-mente (*eye-mind assumption*). Essa teoria propõe que

> A pressuposição imediata indica que “o leitor tenta interpretar o significado de cada palavra de um texto à medida que a encontra” e a pressuposição olho-mente assegura que “o olho permanece fixado numa palavra pelo tempo que a palavra estiver sendo processada. Por isso, o tempo necessário para processar um termo recém-fixado é indicado de modo direto pela duração da mirada[66] (Just and Carpenter, 1980, p. 330 apud UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154).

O tempo da fixação, segundo os autores referidos, está também relacionado com a frequência dos termos na língua. “À medida que o índice de frequência da palavra reduz, a quantidade de tempo requerida para extrair as informações necessárias dessa palavra, aumenta[67]" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154-155). Essas constatações da psicolinguística vão ao encontro dos estudos sobre a relevância da frequência das palavras no *corpus* de uma língua. Sobre esse aspecto, os autores acima citados concluem dizendo que “palavras que demandam maior processamento visual recebem fixações mais longas e, explicar o efeito da frequência é um dos objetivos primários dos modelos teóricos sobre o controle dos movimentos dos olhos na leitura[68]” (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154-155).

Os dados acerca da frequência de termos e sua relação com o tempo de fixação desses termos durante a leitura comprovam haver diferenças no comportamento e na qualidade da leitura realizada por falantes nativos de língua inglesa, quando comparados com estudantes do idioma (pós-graduandos e alguns graduandos sujeitos do estudo reportado). De fato, falantes nativos são mais fluentes do que aprendizes, o que não é nenhuma novidade. Porém, o estudo em tela ajuda a entender um dos mecanismos que parece justificar o porquê de a fluência leitora ser superior entre os falantes nativos: esses sujeitos de forma constante fazem poucas fixações do olhar e, na sua maioria, essas fixações duram pouco tempo, em todos os diferentes contextos de leitura analisados. “A explicação

---

[66] Just and Carpenter’s theory is in fact based on two assumptions, the immediacy assumption which states that “the reader tries to interpret each content word of a text as it is encountered”, and the eye-mind assumption that “the eye remains fixated on a word as long as the word is being processed. So the time it takes to process a newly fixated word is directly indicated by the gaze duration” (Just and Carpenter, 1980: 330).

[67] “As frequency decreases, so the amount of time required to extract the necessary information from the Word increases”. (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 155)

[68] “Words that need more visual processing receive longer fixations, and explaining the frequency effect is a primary goal of theoretical models of eye movement control in reading.” (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 154-155)

indica que os não-nativos fixaram o olhar em cada palavra, em média 1,4 vezes mais do que os nativos, o que é particularmente óbvio quando se observa o percurso (ocular) percorrido durante a leitura". Os pesquisadores também observaram que os "não-nativos tinham a tendência a fazer mais regressões, além de a maioria das palavras serem fixadas, de maneira usual, mais de uma vez[69]" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 161). Conforme apontam os autores

> Os não-nativos tiveram, de forma geral, o tempo de fixação mais longo quando da leitura das passagens dos textos. Isso sugere que as frequências pessoais das palavras mostradas não eram tão altas como aquelas dos falantes nativos. Isto é um produto direto do tempo de exposição a tais palavras, ao longo de uma vida[70]. (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 166)

O estudo do movimento dos olhos trouxe várias evidências sobre o modo de processamento intralinguístico das sequências formulaicas. Ehrlich e Rayner's (1981, apud UNDERWOOD, SCHMITT e GALPIN, 2004, p. 162), por sua vez, mostraram, por exemplo, que, durante a leitura, as palavras previstas pelo contexto precedente eram mais rapidamente fixadas do que palavras em contextos neutros e "parece que o contexto fornecido por uma sequência formulaica, por si mesmo, é suficiente para facilitar o processamento da palavra final da sequência" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 162). Tais dados, de acordo com Underwood e colegas, "vão plenamente ao encontro da visão de que as sequências formulaicas são armazenadas e processadas como unidades. Uma vez que uma sequência formulaica é reconhecida, deve haver menor necessidade de verificar o final da sequência, pois, simplesmente, a pessoa já sabe o que consta no final[71]" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 162). Em outras palavras, uma vez

---

[69] […]This is indicated by the fact that nonnatives fixated on each word 1.4 times on avegare, and is particularly obvious when we observe the actual tracking during reading. The nonnatives tended to have many regressions and most of the words were fixated, often more than once." (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 161)

[70] "The non-native speakers had longer fixation durations overall when reading the passages, suggesting that their personal frequencies of the words being shown were not as high as those of the native speakers. This is a product of their lifetime's exposure to these words. (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 166)

[71] This is largely consistent with the view that such sequences are stored and processed as wholes. Once a sequence is recognized, there should be less need to sample the end of the sequence, simply because the person already knows what that ending is." (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 162)

percebido que uma ou mais palavras indicam o início de uma sequência formulaica, os olhos dão uma sacada, não se fixando na última, ou últimas palavras da sequência. Por isso mesmo, o leitor capaz de automatizar esse procedimento ganhará velocidade na leitura, processará mais rapidamente a informação e, por conseguinte, sua leitura será mais fluente. Os autores citados descrevem em maiores detalhes o processo, alegando, ainda, que:

> Quando um falante nativo lê uma sequência formulaica de palavras tal como *I can see what you mean* (Eu sei o que você quer dizer) ou *the black sheep of the family* (a ovelha negra da famíia), as palavras se tornam mais previsíveis à medida que ele avança na leitura da sequência, e a última palavra da sequência (WordN) torna-se quase redundante. [...] A previsibilidade tem duas consequências. Pelo fato de a última palavra da sequência ser reconhecida mais rápido, o sinal para iniciar a sacada é deflagrado antes e, assim, os olhos do leitor, se fixam na palavra alvo por menos tempo. O fato se manifesta no tempo de fixação reduzido das palavras finais das sequências formulaicas, se comparado com as mesmas palavras, quando não presentes em tais sequências (0.71 de fixação por palavra vs. 0.86 por palavra). Ao olhar para a penúltima palavra de uma sequência (WordN-1), a constatação da familiaridade permitirá que o leitor certifique-se de que a sequência é previsível e que as palavras são familiares, e, então, a atenção será deslocada para a última palavra[72] (WordN). (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 165)

Segundo Underwood, Schmitt e Galpin (2004), também foram observadas diferenças positivas entre as fixações realizadas na leitura de sequências formulaicas por não-nativos. Os não-nativos, como os nativos, fixaram menos vezes o olhar quando se deparavam com sequências formulaicas. Isto é, davam sacadas, mais frequentemente, quando percebiam uma sequência lexical como formulaica. Mesmo dando menos sacadas do que os nativos, os aprendizes demonstraram, também, vantagem no processamento das sequências formulaicas, em comparação com a leitura de outros fragmentos (não formulaicos) do texto. “Agora temos

[72] When a native speaker reads a formulaic sequence of words such *as I can see what you mean* or *the black sheep of the family*, the words become more predictable as they progress through the sequence, and the final word (WordN) is almost redundant. The Familiarity Check would be completed earlier than for the equivalent terminal word placed in a non-formulaic text, thereby allowing faster word recognition overall. This has two consequences. Because the final word in the sequence is recognised early, the signal to begin the saccadic programme is started early, and so the reader’s eyes fixate the target word for less time than otherwise. This is manifest in the reduced fixation duration on final words in formulaic sentences compared to the same words in non-formulaic sequences (0.71 fixations per word vs. 0.86 fixation per word). When looking at the penultimate word in a sequence (WordN−1) the Familiarity Check would allow the reader to ascertain that the sequence is predictive and the words familiar, and attention would move to the final word (WordN).” (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 165)

evidência de que as últimas palavras de uma sequência formulaica são processadas mais rapidamente do que as mesmas palavras, quando apresentadas em contextos não-formulaicos. O ocorrido é uma evidência a favor do posicionamento que defende que as sequências formulaicas são armazenadas e processadas holisticamente[73]" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 166). Porém, em que pese o tom enfático da conclusão, os autores são cautelosos ao advertirem sobre a necessidade de outras pesquisas aprofundando, a partir de diversos pontos, o estudo das sequências formulaicas. "Dado o amplo reconhecimento da importância das sequências formulaicas, agora é tempo de fazer uso de todas as ferramentas disponíveis no kit de ferramentas da psicolinguística para investigar esses itens[74]" (UNDERWOOD, SCHMITT e GALPIN, 2004, p. 166).

Outro estudo de pesquisadores da Universidade de Nottingham analisou o tempo de reconhecimento das sequências formulaicas, comparando o tempo despendido por nativos e por estudantes estrangeiros. O estudo buscou verificar se realmente as sequências formulaicas são processadas holisticamente, tal como o apregoam vários linguistas. Schmitt e Underwood (2004) comentam que "talvez a pessoa precise ver a sequência formulaica ou partes da sequência conjuntamente a fim de reconhecer a sequência completa, ao invés de palavra por palavra" (SCHMITT e UNDERWOOD, 2004, p. 180). No caso da leitura das sequências formulaicas ser realizada termo a termo, conforme o afirmam alguns autores, seria difícil reconhecê-las como uma unidade. Em outras palavras, sem ter conhecimento prévio delas, dificilmente haveria o seu reconhecimento num texto (ou outras práticas discursivas).

Conforme esses autores, a partir dos dados encontrados nesse estudo, o conhecimento prévio das sequências formulaicas passou a ser considerado determinante para o estabelecimento da diferença de resultados entre falantes nativos e não nativos. Mais ainda, sequências formulaicas pequenas foram

[73] We now have evidence that the terminal words in formulaic sequences are processed more quickly than the same words when in nonformulaic contexts. This provides evidence for the position that formulaic sequences are stored and processed holistically. (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 166).

[74] Given the widely recognized importance of formulaic sequences, it is now time to use all of the tools available in the psycholinguistic toolkit to investigate these items." (UNDERWOOD, SCHMITT & GALPIN, 2004, p. 166)

reconhecidas mais rapidamente pelos aprendizes do que sequências longas. Para falantes proficientes, por outro lado, o tamanho das sequências formulaicas não produziu nenhuma diferença no tempo de reconhecimento (tempo dispensado a cada palavra). As diferenças entre os resultados apresentados pelos dois grupos de pesquisados, segundo os autores, não estão necessariamente atreladas somente ao tamanho das sequências formulaicas, mas ao conhecimento e à experiência prévia com tais sequências. Conforme concluem, "o menor tempo de reconhecimento é, pelo menos, parcialmente causado pelo fato dos não-nativos conhecerem melhor as sequências menores do que as maiores" (SCHMITT e UNDERWOOD, 2004, p. 182). Nas palavras dos autores:

> Quando foi comparado o tempo de reconhecimento de palavras em sequências formulaicas conhecidas com o tempo de reconhecimento de sequências desconhecidas, descobrimos que o tempo de reconhecimento era significativamente mais lento para as sequências desconhecidas (458 msec) do que para as sequências já conhecidas (419 msec). A maioria das palavras componentes das sequências pesquisadas possuía frequência relativamente alta e os não-nativos tinham grande possibilidade de já conhecê-las. Mas as (presumivelmente conhecidas) palavras individuais foram reconhecidas mais rapidamente pelos não-nativos, quando faziam parte de uma sequência formulaica. Isso pode indicar certo tipo de facilidade (na compreensão, grifo nosso) pelo fato dessas palavras fazerem parte de um pacote lexical[75]" (SCHMITT e UNDERWOOD, 2004, p. 181).

Embora a utilização das sequências formulaicas possa ser entendida como a de um agente catalisador da velocidade da leitura, sua principal função vincula-se à promoção da compreensão. Ou seja, as "sequências formulaicas não existem na língua porque estimulam e tornam o processamento e reconhecimento das palavras mais rápido, mas porque facilitam um melhor entendimento da mensagem[76]" (SCHMITT e UNDERWOOD, 2004, p. 180). No entanto, para que cumpram tal

[75] "When we compared recognition times for words embedded in formulaic sequences which were known against words in unknown sequences, we found that those times were significantly slower for the unknown sequences (458 msec) than for known sequences (419 msec). Most of the component words in the target sequences were of relatively high frequency, and the nonnatives were likely to know them. But these (presumably known) individual words were recognized faster by the nonnatives when they were part of a formulaic sequence than when they were not. This could indicate some facilitation as a result of these words being 'packaged' in unitary sequences." (SCHMITT & UNDERWOOD, 2004, p. 181)

[76] […]formulaic sequences do not exist in language because they provide a benefit in terms of faster recognition and processing, but because they facilitate better understanding of the message." ." (SCHMITT & UNDERWOOD, 2004, p. 180)

função é necessário que o aprendiz da língua tenha consciência delas, saiba de sua presença e função no discurso. “De forma muito clara, um fator importante, é a consciência que uma série de palavras em um texto é uma sequência formulaica[…]” (SCHMITT e UNDERWOOD, 2004, p. 174). O aprendiz precisa saber vê-las e sobre este tema versará o próximo subitem deste capítulo.

2.3.6 Consciência das sequências formulaicas

Aprendizes de uma segunda língua têm, muitas vezes e com frequência, conforme o atestam diversos autores, dificuldades em reconhecer as sequências formulaicas presentes no discurso. “Há um consenso geral de que as sequências formulaicas são extremamente difíceis para o aprendiz de uma segunda língua dominar[77]” (WRAY 2000, p. 468). Wray aponta como uma das causas prováveis dessa dificuldade a pobreza de experiências linguísticas do aprendiz. Segundo ela e outros pesquisadores, as sequências formulaicas são “frequentemente omitidas na fala endereçada a aprendizes de uma segunda língua”, embora tais expressões sejam comuns na televisão e em filmes[78]” (Irujo, 1986, p. 236-237 apud Wray 2000, p 468). A omissão das fórmulas linguísticas em programas de ensino de idiomas estrangeiros pode estar entre as causas dessa dificuldade; livros didáticos raramente desenvolvem propostas para o seu ensino. “Mais ainda, elas não são bem ensinadas além de ser fácil de serem ensinadas de modo inadequado[75]” (Wray 2000, p. 468) Parece, então, que se poderia até dizer que alunos, e mesmo muitos professores desconhecem ou mesmo ignoram completamente a existência deste fenômeno linguístico.

---

[77] “There is a general consensus that formulaic sequences are extremely difficult for the L2 learner to master”. (Wray 2000, p.468)

[78] “If formulaic sequences are so difficult to learn, then unless we understand why, we are unlikely to hit on a successful way of teaching them. One possible cause is the poverty of the learner experience. Irujo (1986) points out that formulaic sequences are ‘frequently omitted in the speech addressed to second-language learners’ and although they are common in television and movies’ (p. 236-237). Furthermore, they are not taught very well (Granger, 1998, Irujo 1986:237) and it is easy to the wrong ones to be taught (Williams: 1988:51). (Irujo, 1986, p. 236-237 apud Wray 2000, p 468).

Talvez fosse até mais sensato afirmar que os aprendizes não as reconhecem como tais, pois não as conhecem como uma unidade lexical. Howarth (1996, p. 186 apud BISHOP, 2004, p. 227-228) argumenta que os problemas apresentados pelos aprendizes de uma L2 com as sequências formulaicas são "atribuídos à falta de consciência do fenômeno.[79]" É preciso, inicialmente, aprender a vê-las e reconhecê-las em um texto até porque as palavras de um texto, por sua estrutura gráfica, possuem uma delimitação, fronteiras. O leitor reconhece de imediato o começo e o fim de uma palavra pelos espaços que se estabelecem entre elas. No entanto, as sequências formulaicas não possuem essa transparência. O falante nativo as reconhece através da experiência linguística construída ao longo de muitos anos, sabendo os marcadores iniciais e finais, bem como suas possibilidades de variação. O aprendiz, por outro lado, precisa de orientação que lhe indique, primeiramente, a configuração das sequências formulaicas, pois sem experiência prévia sobre essas formações lexicais dificilmente terá condições de inferir a sua presença no discurso. A instrução explícita, agregada a outras estratégias de ensino, desempenha um papel extremamente importante para o aprendiz de uma L2 no sentido de instrumentalizá-lo, para que possa perceber e identificar tais agrupamentos semânticos, para posteriormente considerá-los na leitura e interpretação.

O conhecimento prévio das sequências formulaicas e a consciência da presença delas em um texto tem implicações cognitivas profundas no processo da leitura, isto é, no processo de entendimento da língua. Conforme citação de Bishop (2004):

> Schmidt (1990, 1992) alega que a percepção consciente ao nível da observação/*noticing* (definida como disponibilidade de realizar um relato verbal) é uma condição suficiente e necessária para converter *input* em *intake*, e que o requerimento da observação/*noticing* é aplicado tanto para sintaxe, quanto para fonologia e pragmática. A maioria dos estudos sobre *noticing en*focaram seu aspecto gramatical (e.g., Schmidt 1995; Doughty and Williams, 1998). No entanto, se *noticing* se aplica ao vocabulário, deveria, por consequência, aplicar-se às sequências formulaicas. Não notar a forma holística de uma sequência formulaica interfere em seu próprio processamento e em sua subsequente aprendizagem como uma unidade linguística[80] (BISHOP, 2004, p. 228).

[79] Howarth (1996: 186) argues L2 learners' problems with formulaic sequences are attributable to "a lack of awareness of the phenomenon" (BISHOP, 2004, p. 227-228).

[80] Schmidt (1990, 1992) claims that conscious awareness at the level of noticing (defined as 'availability for verbal report') is a necessary and sufficient condition for converting input to intake, and

A partir de uma adaptação do modelo de processamento lexical de Level (1989, 1993), em adaptação proposta por De Bot, Paribakht, e Wesches's (1997), Bishop (2004) a utiliza para explicar os possíveis mecanismos envolvidos na leitura e compreensão de uma sequência formulaica. O modelo proposto por Levelt (1989) alega que as informações sobre uma palavra são representadas no léxico mental em dois níveis: lexema e lema. O lexema representa as informações fonológicas e ortográficas de uma palavra, ou seja, é a palavra como um signo gráfico ou sonoro. Por outro lado, o lema representa as informações sintáticas e semânticas de uma língua, as propriedades de sentido atreladas ao signo linguístico.

A leitura inicia com o reconhecimento da palavra impressa no texto. Havendo seu reconhecimento, segundo o modelo proposto, a mente realizaria o pareamento entre o lexema (a forma gráfica presente no texto e conhecida pelo leitor) com o lema (o sentido da palavra representado em sua memória). Pelas palavras de Bishop (2004), a partir de Bot et al, "quando uma palavra conhecida é processada durante a leitura, o seu padrão ortográfico precisa, primeiramente, ser reconhecido e combinado com um lexema, o que ativará o lema e só então serão acessadas as propriedades sintáticas e semânticas da palavra[44]" (BISHOP, 2004, p. 228).

No caso de não conhecer ou reconhecer uma palavra presente em um dado texto, o leitor faz um percurso cognitivo distinto daquele que faz quando a conhece. "Com uma palavra desconhecida, o leitor de uma segunda língua precisa primeiramente focar na forma desconhecida e, então, verificar se há necessidade de realizar uma tentativa para encontrar seu significado[81]"(BISHOP, 2004, p. 228). No contexto da leitura, primeiro terá que decidir se é relevante interromper a leitura para encontrar o significado da palavra, atrelando seu lexema desconhecido a um dado lema, para só depois realizar o pareamento entre lema e lexema. O pareamento

---

that the requirement of noticing applies to vocabulary as well as syntax, phonology, and pragmatics. Most studies of noticing have focused on grammar (e.g., Schmidt 1995; Doughty and Williams, 1998). However, if noticing applies to vocabulary, it should therefore apply to formulaic sequences. Not noticing the holistic form of a formulaic sequence should interfere with its processing and subsequent learning as a unitary whole (BISHOP, 2004, p. 228).

[81] According to de Bot et al, when a known word is processed during reading, the orthographic pattern must first be recognized and matched with a lexeme, which activates the lemma thus accessing the syntactic and semantic properties of the word. With an unknown word, however, the L2 reader must first focus on the unknown form and attend to it as being of sufficient interest to attempt to find the meaning. (BISHOP, 2004, p. 228)

pode ser efetivado a partir da consulta a uma referência externa ao texto, ou mesmo a partir da inferenciação ou dedução, entre outras possibilidades. “O processo de preenchimento de um lema vazio implica compreensão e essa é uma condição necessária para o aprendizado[82]” (BISHOP, 2004, p. 228). Somente a partir desse processo, a palavra desconhecida para o leitor terá sentido.

As sequências formulaicas desconhecidas do leitor não se constituem nem em lexemas, nem em lemas. Por não apreendê-las e visualizá-las como tal, ou por não reconhecê-las no texto, já que elas não se configuram nem como um lexema, nem como objetos sígnicos, logo, não possuem nenhum significado para esse leitor. Por outro lado, conhecendo a possibilidade da formação de sequências formulaicas, o leitor atento e treinado poderá, na tentativa de entender um texto, considerar a possibilidade de um vocábulo não entendido fazer parte de uma sequência formulaica. Para isso acontecer, conforme diz Bishop (2004) “as sequências de palavras precisam primeiramente ser reconhecidas como lexemas, isto é, como totalidades portadoras de significado” (BISHOP, 2004, p. 229).

Nesse sentido, fica evidente a relevância que o ensino e o treinamento do leitor podem ter no processo de formação de leitores proficientes. É preciso aprender a ver e a reconhecer as sequências formulaicas, o que pode configurar-se como um facilitador do processo de aprendizagem da língua. “Fica claro que o reconhecimento da forma das sequências formulaicas é um passo essencial em direção ao seu aprendizado[83]” (BISHOP, 2004, p. 229). A Linguística de *Corpus* permite que os pacotes lexicais sejam mapeados e localizados em um *corpus* de estudo, elencando os mais recorrentes e relevantes no contexto de aprendizagem do estudante. Com tais dados em mãos, o educador poderá estabelecer estratégias para seu ensino. O próximo capítulo procurará dar conta de questões relacionadas ao ensino das sequências formulaicas e dos pacotes lexicais, levando em conta o conhecimento trazido pela Linguística de *Corpus* em consonância com os pressupostos da Linguística Cognitiva.

---

[82] “Filling the empty lemma implies comprehension, and this is a necessary condition for learning.” (BISHOP, 2004, p. 228)

[83] According to this account, it is clear that recognition of the form of the formulaic sequences is an essential step towards its being learned.” (BISHOP, 2004, p. 229)

No capítulo que aqui se encerra procurou-se elencar pontos em comum entre a Linguística de *Corpus* e a Linguística Cognitiva para o ensino e aprendizado de uma língua. Vocabulário e léxico, frequência, exposição à língua, ensino e instrução direta, fluência linguística e, principalmente, o aprendizado das sequências formulaicas são aspectos contemplados tanto pelo projeto de pesquisa da Linguística de *Corpus* como pelo projeto da Linguística Cognitiva e se integram, formando um amálgama. O cruzamento de tais dados e informações parecem poder, quando nas mãos do professor ou do produtor de materiais de ensino, contribuir para a melhoria do ensino da língua inglesa. No próximo capítulo, será feito uma proposta de como esse conhecimento trazido pela Linguística de *Corpus* e pela Linguística Cognitiva pode ser utilizado para a produção de material de ensino de língua inglesa.

3 ENSINO DE LEITURA COM ÊNFASE NA AQUISIÇÃO DE VOCABULÁRIO

Este capítulo apresenta uma integração do conhecimento teórico constante nos capítulos anteriores para o desenvolvimento de uma prática de ensino de inglês instrumental. A proposta de ensino aqui articulada insere-se no contexto de ensino acadêmico conhecido como EAP (English for Academic Purposes) ou ESP (English for Specific Purposes). Reitera-se a relevância das sequências formulaicas para o ensino de língua inglesa, destacando-se sua função na linguagem acadêmica, além de se fazer uma proposta de elaboração de um programa de ensino, organizado a partir do léxico. Ao longo deste capítulo, questões pertinentes ao ensino dos pacotes lexicais serão aprofundadas com o propósito de desenvolver uma tarefa de ensino específica – incluindo-se a produção de material de ensino, objetivo último desta pesquisa. A base teórica da Linguística de *Corpus* referente ao ensino da língua inglesa será utilizada tanto para a elaboração do programa de ensino como para o desenvolvimento de tarefas e materiais, a partir de propostas de aplicação do instrumental da LC no contexto da sala de aula, utilizando-se os dados do *corpus* desenvolvido neste estudo,

3.1 Ensino de língua inglesa para acadêmicos

3.1.1 EAP, ESP e Inglês Instrumental

O ensino da língua inglesa para acadêmicos tem propósitos e objetivos diferentes dos cursos gerais de língua inglesa (EGP – English for General Purposes). Uma dessas diferenças diz respeito à especialização da língua, ou seja, ao ensino da língua estrangeira atrelado a campos do conhecimento muito específicos. Gavioli (2005, p. 5) diz que “tradicionalmente, ESP tem sido definido como o estudo da língua inglesa em contextos e áreas de conhecimento especializadas, tais como medicina, engenharia, negócios, entre outras”. Cursos de

ESP parecem estar diretamente associados a cursos de formação profissional ou a discipllinas muito específicas de um curso acadêmico. Em continuidade, Gavioli afirma que

> o conceito de LSP (Language for Specific Purposes- grifo meu) tem se tornado cada vez mais padronizado ao longo dos anos, tornando-se restrito a cenários envolvendo profissionais em formação que necessitam aprender uma língua estrangeira para lidar com questões relevantes em sua futura profissão. (GAVIOLI, 2005, p. 5).

Nesse sentido, a proposta de ensino de língua inglesa para acadêmicos da área de Tecnologia Ambiental coaduna-se às definições apresentadas por Gavioli (2005) tanto por vincular-se ao ensino da língua para acadêmicos de uma área muito específica quanto por levar em conta habilidades necessárias para o desenvolvimento profissional dos estudantes. Neste caso, dá-se destaque ao desenvolvimento da competência leitora de artigos científicos versados em língua inglesa, habilidade que os pesquisadores em formação deverão desenvolver para o pleno desempenho de suas atividades acadêmicas e profissionais, tanto do presente, quanto do futuro.

O alto grau de especificidade alcançado em cursos de ESP dá-se não somente no recorte conceitual, mas também na seleção de variedades de discursos e de registros priorizadas e desenvolvidas ao longo de um curso nesse formato. Gavioli (2005) destaca, ainda, que, provavelmente, a marca distintiva de um curso geral de inglês (EGP) de um curso de ESP

> é que o curso de ESP focaliza um número menor de variedades de gêneros, tipos textuais e situações, frequentemente desenvolvido um de cada vez. Neste sentido, uma diferença saliente de ESP, quando comparado com um curso de inglês geral (EGP), é aquela decorrente de sua abordagem do estudo da língua[84] (GAVIOLI, 2005, p. 6).

O comentário [apresentado] de Gavioli, mais uma vez, contempla a proposta de ensino de inglês intrumental aqui desenvolvida, pois esta enfoca o ensino da língua

[84] Both EGP and ESP are concerned with a variety of registers and sub-languages and speakers of General English do speak for a variety of specific purposes every day. What distinguishes the two is probabloy that ESP focuses on a smaller number of varieties, text-types and situations, often one at a time. In this sense, then, a distinguisihing feature of ESP as compared to EGP is a difference in their approach towards the study of language. (GAVIOLI, 2005, p. 6)

inglesa tal como utilizada em textos científicos, propondo-se a desenvolver a competência leitora do acadêmico na leitura de um gênero textual muito específico: artigos científicos da área de Tecnologia Ambiental.

Duas outras definições de modalidades de curso de língua inglesa recorrentemente utilizadas para um curso ESP são: Inglês Instrumental e EAP (English for Academic Purposes). Cursos de *English for Specific Purposes,* no Brasil, são predominantemente referidos como "Inglês Instrumental", às vezes com especificações como "Inglês Instrumental para Engenharia", "Inglês Intrumental para Informática", "Inglês Instrumental para Enfermagem", etc.. Contudo, um curso de Inglês Instrumental, majoritariamente, refere-se ao ensino da habilidade de leitura em detrimento das outras habilidades linguísticas. No entanto, um curso instrumental não precisa ter como escopo exclusivo o desenvolvimento da habilidade leitora, podendo ser estruturado a partir de outros objetivos de ensino. Por outro lado, um curso de *English for Academic Purposes* (EAP) diz respeito ao desenvolvimento de uma série de habilidades linguísticas, comunicativas e cognitivas todas elas necessárias para o estudante poder acompanhar um curso acadêmico em língua inglesa.

Segundo Hyland (2006, p. 1) o "EAP é geralmente definido como o ensino da língua inglesa que tem o objetivo precípuo de auxiliar o estudo ou pesquisa do acadêmico naquela língua[85]", sendo tais cursos oferecidos por universidades que recebem estudantes estrangeiros, podendo haver, entretanto, cursos de EAP para falantes nativos, como aqueles voltados ao desenvolvimento da habilidade de escrita acadêmica. Assim, áreas relacionadas ao desenvolvimento de habilidades comunicativas pertinentes ao meio acadêmico têm chances de inclusão em programas de ensino de um curso de EAP. Práticas de ensino que proponham o desenvolvimento de competências diversas, como leitura e escrita de diferentes gêneros textuais acadêmicos, preparação de apresentações, linguagem para interação em sala de aula, processo de pesquisa, procedimento para produção de resumos e anotação durante palestras e aulas (note taking), entre outras competências, também, fazem parte do rol de possibilidades. Hyland (2006), citando

[85] EAP is usually defined as teaching English with the aim of assisting learners' study or research in that language (e.g. Flowerdew and Peacock, 2001: 8; Jordan, 1997: 1).

Dudley-Evans (2001), observa que o EAP tende a lidar com questões práticas, “em termos que se referem aos contextos locais e às necessidades específicas de determinados estudantes[86]” (HYLAND, 2006, p. 1), objetivo comum a todo curso ESP, conforme já referido.

De modo geral, os cursos de EAP não se vinculam a uma área específica do conhecimento, destinando-se ao desenvolvimento de uma determinada competência comum a diversos cursos acadêmicos, como um curso de *note-taking* (habilidade de sintetizar e tomar notas de palestras e aulas), comumente, oferecido em universidades inglesas, por exemplo. Hyland (2006) afirma que em sua origem os cursos de EAP[87] foram concebidos como uma das duas divisões dos cursos ESP, juntamente com cursos relacionados ao uso da linguagem em ambientes profissionais (EOP – English for Occupational Purposes). Embora as duas definições sejam muito semelhantes, neste estudo adota-se a sigla ESP.

3.1.2 Sequências formulaicas e EAP

O capítulo anterior apresentou e discutiu a importância das sequências formulaicas no processamento da língua, destacando a função facilitadora que elas desempenham com relação à memória e seu forte vínculo com a fluência verbal. Também foi esclarecida a importância que o conhecimento dessas sequências detém numa comunidade linguística sendo elas, muitas vezes, o fator que diferencia um falante fluente de um aprendiz da língua; um estrangeiro, de um membro da comunidade.

De modo similar, as sequências formulaicas desempenham papel de destaque e função de extrema relevância no discurso acadêmico, sendo amplamente utilizadas na produção textual de diversos gêneros de textos que circulam entre

[86] As Dudley-Evans (2001: ix) notes, EAP often tends to be a practical affair, and these areas are typically understood in terms of local contexts and the needs of particular students. (HYLAND, 2006, p. 1)

[87] O termo EAP, segundo Hyland (2006, p. 2) parece ter sido cunhado por Tim Johns, em 1974, tendo aparecido pela primeira vez em uma compilação de artigos editados por Cowie e Heaton, em 1977[87]”. Tim Johns foi desenvolvedor do DDL (Data Driven Learning), proposta para utilização de um *corpus* linguístico, diretamente com alunos, que tem central importância neste estudo e será oportunamente detalhada.

pesquisadores e estudantes universitários. Wray (2002, p. 95) sugere que a questão mais importante no uso das sequências formulaicas "é a promoção dos interesses do usuário[88]". Jones e Haywood (2004) alegam que, talvez, um dos maiores interesses dos estudantes de língua inglesa para fins acadêmicos seja ser bem sucedido na sua formação. A busca de melhor formação e de conhecimentos linguísticos atrelados à área de formação específica dos estudantes, pode ser um indicador do interesse em conhecer e dominar o código do nicho acadêmico do qual os estudantes querem fazer parte. Por isso, "focar no ensino das sequências formulaicas, durante aulas de produção textual, nos cursos de EAP, parece plenamente justificável, pois pode auxiliar os estudantes a atingir seus objetivos acadêmicos[89]" (JONES & HAYWOOD, 2004, p. 273). Desse modo, em consonância com o que até aqui se argumentou a respeito das sequências formulaicas, entende-se que trabalhar com elas é válido e favorável, não somente para a melhoria da produção textual, conforme referido pelos autores, mas também por propiciar o desenvolvimento de outras habilidades linguísticas, incluindo-se e destacando-se a habilidade leitora. Adicionalmente, Jones e Haywood (2004) explicitam a importância das sequências formulaicas, no contexto do ensino acadêmico, frisando que

> Tanto graduandos quanto alunos da pós-graduação seguem uma espécie de iniciação nas disciplinas escolhidas, gradualmente se familiarizando, não somente com o conhecimento e habilidades de suas áreas, mas também com a linguagem daquele campo, para se tornarem capazes de expressar suas ideias de uma forma aceita e reconhecida. À medida que o fazem, o manejo das sequências formulaicas os torna capazes de, por exemplo, expressarem ideias técnicas de forma concisa, sinalizarem estágios em seus discursos, além de saberem usar o nível necessário de formalidade. A ausência destas características (marcas linguísticas- grifo meu) talvez resulte em a escrita de um estudante ser julgada inadequada. [...] Por outro lado, familiaridade com e domínio da língua de suas áreas técnicas indica que "pertencem ao clube", neste caso, a comunidade e área acadêmica por eles escolhida. Mais ainda, quando o estilo de sua produção escrita for convencional, atrai pouca atenção, pois (as sequências formulaicas da área – grifo meu) não sobrecarrega o processamento da informação durante a leitura e permite que a mensagem do escritor seja mais facilmente apreendida[90]. (JONES & HAYWOOD, 2004, p. 273).

[88] Wray's model suggests that the overriding purpose of the use of formulaic sequences is "the promotion of the [user's] interests" (2002: 95). (JONES & HAYWOOD, 2004, p. 273)

[89] Thus a focus on formulaic sequences in academic writing in the EAP classroom seems fully justified as it can help the students reach their academic goals. (JONES & HAYWOOD, 2004, p. 273)

[90] Both undergraduates and postgraduates serve a kind of apprenticeship in their chosen discipline, gradually familiarising themselves not only with the knowledge and skills of their field, but also with the language of that field, so that they become capable of expressing their ideas in the form that is

Em vista disso, o ensino das sequências formulaicas e dos *lexical bundles*, conforme várias vezes já pontuado, é basilar para o aprendizado da língua inglesa, com especial destaque para o ensino da língua no *métier* acadêmico. A seguir serão apresentadas e discutidas propostas de ensino que utilizam o léxico como elemento organizador e orientador de uma proposta de ensino, dando-se especial destaque ao ensino dos padrões lexicais, tais como sequências formulaicas e pacotes lexicais.

3.2 Léxico e currículo

3.2.1 A centralidade do léxico no ensino de língua inglesa

Tradicionalmente, o programa de ensino de um curso de língua estrangeira é organizado a partir de estruturas da língua elencadas de acordo com algum critério estabelecido pelo produtor do material. Nesses currículos, as estruturas gramaticais estão no centro do programa, sendo a partir delas organizado o currículo de ensino que parte do pressuposto de que a aquisição da língua é linear, decorrente do aprendizado de uma sucessão de estruturas gramaticais. Esse pressuposto é compartilhado por diferentes autores. Wilkins (1976, apud WILLIS, 1990, p. 42) refere essa abordagem como "estratégia sintética" (synthetic strategy), definindo-a como uma estratégia onde "as diferentes partes da língua são ensinadas separadamente e passo-a-passo" e, a aquisição da língua, resultante desse encadeamento, é entendida como "um processo acumulativo gradual de partes até a estrutura completa da língua estar desenvolvida[91]". Lewis (1993) afirma que "[...] a

expected. As they do this, their use of formulaic sequences enables them, for example, to express technical ideas economically, to signal stages in their discourse and to display the necessary level of formality. The absence of such features may result in a student's writing being judged as inadequate. […]On the other hand, familiarity with and control of the language of their field indicates their membership of the group, in this case, the community of their chosen academic discipline. In addition, when the writing style is conventional, it attracts little attention. This lightens the processing load for the reader and allows the writer's message to be more easily perceived. (JONES & HAYWOOD, 2004, p. 273)

[91] A synthetic strategy is is one in which the different parts of the language are taught separately and step by step so thatacquisition is a process of gradual accumulation of the parts until the whole structure of the language has been built up. (Wilkins, 1976, apud Willis, 1990, p. 42).

maioria dos currículos especifica uma dada sequência. Em decorrência, ficam intrinsicamente amarrados ao conceito de nível. A natureza multi-dimensional do nível, necessariamente, implica multi-dimensionalidade do currículo[92]" (LEWIS, 1993, p.47). Segundo Willis (1990, p. 42), outro problema relacionado à estratégia sintética está na presunção de que os "itens gramaticais podem ser ordenados de maneira lógica, não somente a partir do ponto de vista do escritor do material, mas também do aprendiz".

Com a democratização do acesso aos computadores, porém, ocorreu uma mudança significativa no modo de conceber o ensino, e novas propostas de programas e de reformulações entraram em cena, dentre elas, aquelas advindas do conhecimento produzido pela Linguística de *Corpus*. Esta, além de estar produzindo conhecimento sobre a língua (colocações, padrões linguísticos, etc), ainda tem sido utilizada com propósitos pedagógicos na organização e planejamento de programas de ensino de língua estrangeira. Além disso, está servindo como instrumental de ensino e aprendizagem para a produção de materiais e fornecendo referencial teórico para repensar práticas de ensino de línguas. A seguir são apresentadas três propostas de ensino, diretamente, relacionadas ao conhecimento gerado pela Linguística de *Corpus*: o *Lexical Syllabus*, proposto por Dave Willis; o *Lexical Approach*, de Michael Lewis; e o DDL (Data Driven Learning) desenvolvido por Tim Johns.

A discussão apresentada neste capítulo focalizará alguns aspectos de cada uma dessas três propostas, os quais podem ser utilizados no desenvolvimento de um programa de ensino de língua inglesa para acadêmicos e, ainda, na proposição de tarefas e na produção de material didático para o segmento acadêmico em pauta. Não se objetiva, pois, detalhar diferenças ou semelhanças entre as propostas, nem fazer críticas a uma ou outra, mas sim apropriar-se de uma base de conhecimento já consolidada como referencial para a reflexão e proposição de uma dada prática de ensino.

[92] Typically, in addition to content, most syllabuses specify a sequence. As such, they are intrinsically linked to the concept of level. The multi-dimensional nature of level, necessarily imples a multi-dimensionality of syllabus. (LEWIS, 1993, p.47)

3.2.2 Centralidade do significado

Embora essas três abordagens apresentem divergências entre si, todas elas consideram o léxico como central para a organização de um programa de ensino. Também enfatizam o uso de material autêntico, de exemplos de usos da língua produzidos por usuários genuínos do idioma, em situação real de comunicação. O aporte em pauta desloca o foco do ensino da estrutura frasal para o discurso, tanto na produção como na recepção. Ou seja,

> o uso da língua não é uma questão de conformá-la a um conjunto de regras restritivas. É uma questão de explorar o sistema linguístico para alcançar propósitos comunicativos. A língua usada é moldada pelo propósito para o qual está sendo usada. A língua usada para simplesmente ilustrar um sistema gramatical abstrato não possui nenhum propósito e por isso não oferece nenhuma base para a sua escolha[93] (WILLIS, 1990, p. 126).

O entendimento de que o que tem importância para um sistema linguístico é a capacidade de produzir sentido constitui-se no ponto fulcral compartilhado por essas abordagens. Willis (1990), referenciando Halliday, afirma que o que o estudante precisa desenvolver é a habilidade de interagir com os outros de tal forma que atinja os resultados por ele almejados. Pode-se mesmo afirmar que hoje o ensino da língua é entendido por um viés pragmático, centrando-se a maior preocupação dos teóricos na obtenção de resultados, sejam eles comunicativos ou operacionais. Lewis (1993, p. vi), proponente do *Lexical Approach*, parece compartilhar dessa visão ao afirmar que "a língua é reconhecida como um recurso pessoal e não como uma idealização abstrata[94]". Willis (1990, p. iii) consegue ser mais assertivo, ainda, ao generalizar que "no presente, há um consenso geral de que aprendemos uma língua a partir de seu uso, para alcançar resultados[95]." Para esses autores, o léxico, seja na forma de uma única palavra, seja numa sequência formulaica ou na forma

---

[93] Language use is not a matter of conforming to a set of restrictive rules. It is a matter of exploiting the language system to achieve communicative intentions. The language used is shaped by the purpose for which it is being used. Language which is being used simply to illustrate an abstract grammatical system has no purpose and therefore offers no basis for choice. (WILLIS, 1990, p. 126)

[94] Language is recognised as a personal resource, not an abstract idealisation. (LEWIS, 1993, p. vi)

[95] "There is a general agreement nowadays that we learn a language by using it to do things, to achieve outomes." (WILLIS, 1990, p. iii

de um pacote lexical, é o grande propulsor do processo de produção de sentido, sendo priorizado em suas propostas pedagógicas.

3.2.3 Um currículo Lexical

Dave Willis (1990) elaborou um programa de ensino, cujo ponto de partida é a palavra e o denominou *Lexical Syllabus* (currículo lexical). Sua proposta foi fortemente influenciada pelo *Collins COBUILD Project* desenvolvido por John Sinclair na pesquisa sobre padrões de linguagem. Willis (1990, p. 91), citando alguns passos iniciais da elaboração da proposta relata que "acreditávamos que os padrões e significados associados com as palavras mais comuns da língua inglesa poderiam constituir uma base para a especificação de um currículo, o qual forneceria aos aprendizes uma boa cobertura[96]." Tal como a pesquisa do *COBUILD project*, o *Lexical Syllabus* iniciou pelo estudo de um *corpus* que foi analisado e mapeado a partir do referencial teórico e do instrumental metodológico da Linguística de *Corpus*. A organização de um currículo, tendo a palavra por base, segundo Willis pautou-se pela presunção de "que essa abordagem, para a especificação e desenho de um currículo, nos daria uma melhor cobertura (da língua, grifo meu) do que os currículos tradicionais organizados a partir do inventário de padrões gramaticais e/ou funções linguísticas[97]". Willis indica outros princípios orientadores que estariam presentes na organização do léxico no currículo, afirmando que "esse currículo especificaria as palavras, seus significados, e os padrões fraseológicos nos quais eram utilizadas[98]. (WILLIS, 1990, p. 15)

Seu ideador, pelo visto, opunha-se à exposição aleatória, devendo-se, segundo ele, evitá-la. Em sua ótica, a língua a que os aprendizes serão expostos deve ser criteriosamente selecionada, promovendo o encontro do aprendiz com aqueles

[96] We believed that the patterns and meanings associated with the commonest words of English would afford a basis for syllabus specification which would provide learners with good coverage - and would provide that coverage economically. (WILLIS, 1990, p. 91)
[97] We believed that this approach to syllabus specification and design would give us better coverage than the more traditional syllabus based on an inventory of grammatical patterns and/or functional realisations. (WILLIS, 1990, p. 91)
[98] This would specify words, their meanings, and the common phrases in which they were used. (WILLIS, 1990, p. 15)

vocábulos que tenham maior probabilidade de ocorrência quando a língua for utilizada no cotidiano. “As palavras mais comuns e mais importantes, os significados mais básicos da língua inglesa são aqueles expressos pelos seus termos mais frequentes[99]” (WILLIS, 1990, p. 46). Essa seleção pode ser tarefa muito complexa, porém a partir de instrumental apropriado para a seleção dos termos e significados mais frequentes, tal como o recomenda a Linguística de *Corpus*, pode tornar-se muito simples. Em termos históricos, o *Lexical Syllabus* foi o primeiro currículo a ser organizado a partir da frequência e relevância dos termos de uma língua.

A sistematização do ensino do vocabulário é também preocupação do *Lexical Approach;* Lewis (1993, p. 117) afirma que “o ensino do vocabulário raramente apresentou-se sistematizado”. Para Lewis, em conformidade com o pensamento de Willis, o *Lexical Approach* requer “um sistema organizado por princípios para a introdução e exploração do léxico, mesmo para o vocabulário simples, na sala de aula[100]” (LEWIS, 1993, p. 117).

Desse modo, a organização de um programa de ensino a partir do léxico está no cerne das preocupações do presente estudo, o que se coaduna perfeitamente com as proposições de Willis e Lewis, anteriormente referidos. A tarefa e o material de ensino produzidos (e, posteriormente, apresentados) foram organizados a partir da enumeração dos termos mais recorrentes, mais frequentes e mais importantes, sendo selecionados a partir dos textos que compõem o *corpus* da Tecnologia Ambiental. Em vista disso, o *corpus* de artigos científicos, analisado com as ferramentas do Wordsmith Tools 5.0, foi o ponto de partida para a elaboração do programa de ensino de inglês instrumental para a Tecnologia Ambiental. Berber Sardinha (2004, p. 284) assegura que os proponentes do *Lexical Syllabus* tinham como preocupação central “evitar a intuição na especificação do curso. Esse é mais um dos pilares da Linguística de *Corpus* segundo John Sinclair, e a intuição era evitada justamente pelo atrelamento do curso ao *corpus*”. Por isso mesmo, a sequência das palavras desta investigação define a sequencialidade da tarefa

[99] The commonest and most important, most basic meanings in English are those meanings expressed by the most frequent words in English. (WILLIS, 1990, p. 46)

[100] Vocabulary teaching has rarely been systematic. [...]A *Lexical Approach* requires a much more principled system of introducing and exploiting lexis, and even simple vocabulary, in the classroom. (LEWIS, 1993, p. 117)

elaborada, que se pauta pelas listagens de palavras, palavras-chave e *lexical-bundles* extraídos do *corpus*. Em decorrência, obtiveram-se guias mais precisos e pontuais para o direcionamento do ensino.

Lewis (1993) critica o *Lexical Syllabus,* proposto por Willis, dizendo ter sido atribuída ênfase demasiada ao ensino de vocábulos isolados, omitindo, a proposta feita, o ensino dos padrões lexicais. Berber Sardinha (2004) sintetiza essa divergência ao destacar a relevância da frequência das palavras para a sequencialidade do Currículo Lexical, apesar de a situação gerar alguns conflitos como os decorrentes do aparecimento de grande número de palavras funcionais. Contudo, inter-relacionar termos e padrões parece ser bem produtivo como o postula Sardinha, em seguida:

> Entretanto, as palavras mais comuns são as mais vazias lexicalmentee, tais como *have* e *with* e também as mais difíceis de ensinar. Dessa forma, ao seguir estritamente o critério de quanto maior a frequência, mais cedo elas devem ser ensinadas, o Currículo Lexical coloca em risco o sucesso de sua própria implementação. Os vários sentidos de uma mesma palavra seriam ensinados juntos, no Currículo Lexical, pois a unidade da atividade é a palavra. Como decorrência do foco na palavra isolada, os itens lexicais compostos foram preteridos. O resultado é que muitos itens lexicais realizados por múltiplas palavras (*multiwords)* não eram ensinados, ou o eram de modo inadequado. Os alunos acabavam, consequentemente, tendo uma visão distorcida do léxico da língua. (BERBER SARDINHA, 2004, p. 288-289).

No *Lexical Approach*, a ênfase é dada ao ensino dos padrões colocacionais da língua, isto é, a abordagem confere especial destaque ao caráter composto do léxico, como o evidenciam as sequências formulaicas, as colocações e os *lexical bundles* e, além do mais, utiliza apenas textos autênticos. Resta esclarecer, que as contribuições teóricas in*corpora*das ao presente estudo originaram-se das proposições do *Lexical Approach e* dizem respeito, sobretudo, ao ensino dos padrões lexicais. Busca-se, além disso, estabelecer um equilíbrio, entre as propostas formuladas por Willis e Lewis, para, a partir daí, elaborar um programa de ensino, destinado a acadêmicos da área de Tecnologia Ambiental. Em vista disso, a tarefa proposta considera a frequência e sequencialidade dos termos, tal como Willlis, mas também investiga possibilidades e estratégias para o ensino dos pacotes lexicais, conforme o propõe Lewis.

3.2.4 Lexicogramática: por uma gramática da palavra

As combinações linguísticas acontecem, também, por meio da combinação de termos lexicais, sintaticamente, e seu resultado é a formação de uma nova unidade de sentido, que não diferencia estrutura e vocabulário, formando uma unidade da língua que representa um amálgama entre léxico e sintaxe. Esse fenômeno linguístico tem sido denominado lexicogramática. "Há muitos termos diferentes na Linguística de *Corpus* para caracterizar lexicogramática, incluindo, colocações, coligação, padrão lexical, chunk (porção), pacote lexical, linguagem formulaica e multipalavras, entre outros[101]" (BERBER SARDINHA, p. 3, no prelo). O conceito de lexicogramática, segundo Berber Sardinha (no prelo), é o que se segue:

> lexicogramática (ou léxico-gramática) é um nível da estrutura linguística onde o léxico ou vocabulário, e gramática ou sintaxe, combinam-se em uma unidade. Nesse nível, palavras e estruturas gramaticais não são vistas como independentes, mas mutuamente dependentes, onde um nível faz interface com o outro. (BERBER SARDINHA, p. 1, no prelo).

Ute Römer (2009, p. 148), ao analisar a inseparabilidade entre gramática e léxico em diferentes abordagens teóricas (*idiom principle, lexical bundles, pattern grammar, lexical priming e collostructional analysis*), conclui que "todas elas alegam que forma e significado são inseparáveis e que a unidade de sentido na língua não é a palavra isolada mas a construção ou a unidade frasal (em diferentes níveis de complexidade)[102]".

Em prosseguimento, Römer (2009, p. 146) destaca um aspecto da lexicogramática saliente nos pacotes lexicais, referindo que, com relativa frequência, estes "atravessam as fronteiras das categorias da gramática tradicional tais como

[101] There are many different terms in CL to characterize lexicogrammar, including, collocations, colligation, phraseology, lexical pattern, chunk, lexical bundle, formulaic language, and multi-word units among others. . (BERBER SARDINHA, p. 3, no prelo)

[102] They all find that form and meaning are inseparable and that the unit of meaning in language is not the word in isolation but a construction or phrasal unit (at different levels of complexity).

sintagmas nominais (*noun phrases*) e sintagmas verbais (*verbal phrases*)[103]". Biber, Conrad and Cortes (2004, p. 377), pesquisadores engajados na invesigação detalhada e aprofundada dos *lexical bundles* afirmam que "a maioria dos pacotes lexicais não representam uma unidade estrutural completa" registrando que menos de cinco por cento dos pacotes lexicais presentes na prosa acadêmica constituem uma unidade estrutural completa. Esses pesquisadores demonstram, a partir dos dados coletados em seus estudos, como ocorre a combinação de léxico e gramática, em uma unidade (lexicogramatical), afirmando que

> De fato, a maioria dos pacotes lexicais liga duas unidades estruturais: eles iniciam no limiar de uma oração ou estrutura fraseológica, mas as últimas palavras do pacote lexical são os primeiros elementos de uma segunda unidade estrutural. A maioria dos pacotes lexicais em conversações conecta duas orações (por exemplo, *I want to know* – eu quero saber; *Well, that's what I* – bem, é o que eu), enquato pacotes lexicais na prosa acadêmica geralmente conecta duas estruturas fraseológicas (por exemplo, *in the case of* – no caso de*; the base of* – a base de)[104]. (BIBER, CONRAD and CORTES, 2004, p. 377).

Römer (2009), no artigo já mencionado, destaca a importância da inclusão da lexicogramática no ensino de língua inglesa, reiterando a relevância das estruturas lexicogramaticais para a comunidade acadêmica e sugere

> que respondêssemos às observadas divergências da regra especialista nos cursos de língua inglesa (ELT) e também nas aulas de EAP (English for Academic Purposes), tanto para níveis introdutórios quanto para avançados, e destacássemos para nossos estudantes e calouros na universidade os padrões léxico-gramaticais, as colocações, os pacotes lexicais e os construtos, para ajudá-los a se tornarem membros aceitos da específica comunidade de prática à qual almejam pertencer[105]. (RÖMER, 2009, p. 160)

---

[103] A particularly interesting aspect of lexical bundles is that they, more often than not, cross the boundaries of traditional grammatical categories such as noun phrases or prepositional phrases. (RÖMER, 2009, 146)

[104] Instead, most lexical bundles bridge two structural units: they begin at a clause or phrase boundary, but the last words of the bundle are the first elements of a second structural unit. Most of the bundles in conversation bridge two clauses (e.g. *I want to know, well that's what I)*, while bundles in academic prose usually bridge two phrases (e.g. *in the case of, the base of)*. (BIBER, CONRAD AND CORTES, 2004, p. 377)

[105] Based on the observations made in this paper, I would suggest that we respond to the observed deviations from the expert norm in general ELT and in EAP classes, both on introductory and advanced levels, and highlight for ours students and novices in academia that lexical-grammatical patterns, collocations, lexical bundles, and construction matter, thus helping them become accepted members of the specific community of practice they aim to belong to. (RÖMER, 2009, p. 160)

Embora alguns autores ressaltem que o aprendizado dos padrões lexicais possa não ser fácil para os alunos, conforme já comentado antes, existe a possibilidade de a dificuldade estar relacionada às práticas de ensino, pois os alunos raramente têm contato com conceitos relacionados à lexicogramática. A sugestão é que os padrões lexicogramaticais sejam introduzidos desde os primeiros contatos do estudante com o ensino da língua estrangeira. Isso porque talvez um dos legados mais importantes do *Lexical Approach* esteja, justamente, no destaque dado ao ensino dos padrões lexicais, em programas de língua estrangeira. Lewis (1993, p. 100) adverte que ao se adotar o *Lexical Approach* deve ficar bem claro que os padrões lexicais "têm que ser introduzidos relativamente cedo no programa de ensino[106]". O mesmo argumento é apresentado por Willis (2009, p. 166), ao afirmar que deveriam ser dadas oportunidades aos alunos para desde cedo reconhecerem os usos de padrões lexicais, assim "pavimentando o caminho para a assimilação e reconhecimento dos padrões da língua em estágios futuro" do seu percurso de aprendizagem.

3.3 A Linguística de Corpus na sala de aula

Laura Gavioli (2005) comenta que o uso do *corpus* no ensino de uma língua estrangeira pode acontecer de duas formas. Na primeira, o *corpus* é utilizado como uma ferramenta para o professor, para a própria pesquisa do educador com o propósito de conhecer e analisar a língua e para a composição e elaboração de um programa de ensino. O *corpus* como uma "ferramenta para o professor" aparece, claramente, na proposta do *Lexical Syllabus* de Willis (1990) que se volta à elaboração de um programa de ensino a partir da seleção dos termos mais relevantes da língua. Essa possibilidade de uso do *corpus*, no estudo aqui relatado, desdobra-se no processo de criação e análise do *corpus* de Tecnologia Ambiental para fazer o reconhecimento do léxico específico da área analisada, de sua lexicogramática e das possibilidades de exploração do conhecimento trazido pelo

[106] If we are to take a lexical approach to language teaching it is clear that some of these sentences, grouped and choses for archetypicality, must be introduced relatively early in the learning programme. (LEWIS, 1993, p. 100)

corpus no estabelecimento de diretrizes e prioridades sobre o que ensinar. A propósito, conforme o atesta Hunston (2002, p. 198), "para o professor de Inglês para fins acadêmicos (EAP), a questão "o que ensinar" (em oposição a "como ensinar") tem significância especial[107]".

Ainda mais, Gavioli (2005) amplia a dimensão do conhecimento produzido pela utilização de um *corpus*, na definição de um programa de ensino, entendendo que não somente o aluno de um curso especialista é beneficiado, mas o próprio professor de língua que muitas vezes (senão, sempre) precisa lecionar para alunos de áreas em que não possui domínio técnico. A autora sugere

> que nas aulas de ESP, "o que" é ensinado (quando comparado a "como" é ensinado) é um problema crucial para os professores de idiomas, que com frequência não são *experts* na disciplina em questão e precisam encontrar um "caminho linguístico" para poder acessar tanto as convenções da língua, como conceitos presentes na disciplina especializada, para assim também poderem elaborar seu próprio entendimento. Eu penso que o trabalho com *corpus* fornece indícios que podem ser organizados e utilizados pelos alunos (e professores) e isto os capacita a se infiltrarem nas práticas discursivas da comunidade na qual atuam[108]. (GAVIOLI, 2005, p. 14-15)

3.3.1 Aprendizagem Movida a Dados (Data Driven Learning – DDL)

A segunda possibilidade de uso do *corpus* no ensino de língua estrangeira, segundo Gavioli (2005), seria "como uma ferramenta para os aprendizes", ou seja, o conhecimento e metodologia da Linguística de *Corpus* são utilizados como um instrumental dos alunos, a ser usado por eles mesmos em seu próprio aprendizado. O *Data Driven Learning* (DDL), traduzido para a língua portuguesa como "Aprendizagem Movida por Dados", foi um estudo pioneiro na utilização dos dados de um *corpus* linguístico, impressos ou interfaceados por algum software específico da Linguística de *Corpus*, diretamente pelos alunos no contexto da sala de aula.

[107] For the teacher of English for academic purposes, the issue of 'what to teach' (as opposed to 'how to teach') is of particular significance. (Hunston, 2002, p. 198)

[108] I suggest that in ESP "what" is taught (as compared to "how" it is taught) is a very crucial problem for language teachers who are often not experts in the discipline in question and need to find a "linguistic path" to gain access to both language conventions and concepts that are entailed in the specialized discipline and construct its own meanings. I suggest that *corpus* work provides clues that can be put together by learners (and teachers) performing it and that this enables them to infiltrate the discourse community's communicative activity. (GAVIOLI, 2005, p. 14-15)

Segundo Berber Sardinha (2004, p. 290) o DDL é “uma das propostas mais sólidas para a utilização de material de *corpus* na sala de aula”. Esta proposta foi defendida por Tim Johns e continua central e relevante para o ensino.

Entre os pontos relevantes da proposta do DDL está a possibilidade de o próprio aprendiz utilizar o *corpus* de textos, sem qualquer agente intermediário e, a partir da leitura dos dados apresentados e organizados na interface do conconcordanciador, chegar às suas próprias conclusões. De acordo com Johns:

> O que distingue a abordagem DDL é a tentativa de eliminar o máximo possível o agente intermediário, possibilitando acesso direto aos dados, a fim de que o aprendiz tome parte na construção de seu próprio repertório de significados e usos da língua. A suposição subjacente a essa abordagem é que o aprendizado efetivo da língua é por si mesmo uma forma de pesquisa linguística, e que os modelos apresentados pelo concordanciador oferecem uma fonte genuína para a estimulação de estratégias indutivas de aprendizado – em particular as estratégias para perceber similiaridades e diferenças e para formar e testar hipóteses[109] (Johns, 1988 apud Johns, 1994).

A partir dessa linha de raciocínio, ao utilizar um *corpus* especializado e formular suas próprias indagações sobre o funcionamento da língua, sobre o significado de um determinado termo ou expressão, procurando respostas a essas perguntas (entre tantas outras possibilidades de investigação), os aprendizes terão a oportunidade de “ter sua própria participação no discurso de sua comunidade” acadêmica (Gavioli, 2005, p. 99). Um *corpus* composto de material autêntico proveniente do discurso acadêmico, da área especialista do aluno de um curso ESP, tal como aqui proposto, pode ser uma fonte abundante de *input* e fornecer estímulos variados para que o estudante esmiuce os meandros da linguagem utilizada pelos pares da comunidade acadêmica em que pretende se inserir.

---

[109] “What distinguishes the DDL approach is the attempt to cut out the middleman as far as possible and to give direct access to the data so that the learner can take part in building up his or her own profiles of meaning and uses. The assumption that underlies this approach is that effective language learning is itself a form of linguistic research, and that the concordance printout offers a unique resource for the stimulation of inductive learning strategies – in particular the strategies of perceiving similarities and differences and of hypothesis formation and testing (Johns 1988)”

3.3.2 O pesquisador, o estudante e o lexicógrafo

Alguns autores comparam o processo de investigação crítica, protagonizado pelo aluno na manipulação dos dados de *corpora*, ao trabalho do pesquisador e do lexicógrafo. Um dos motes presentes na proposta do DDL é dar ao aprendiz a oportunidade de pesquisar, de exercer o papel de investigador frente ao aprendizado da língua. Segundo Johns (1991, p. 2), o que é inovador sobre a proposta do DDL está atrelado à percepção de que "a pesquisa é uma ferramenta valiosa demais para ficar na mão dos pesquisadores[110]", que o aprendiz de uma língua é também um pesquisador, cujo aprendizado precisa ser dirigido para o acesso a dados linguísticos – por essa razão a expressão 'aprendizagem movida por dados' (DDL – data driven learning) é utilizada para descrever essa abordagem (Johns, 1991, p. 1). Willis (1990, p. 68) sugere "que esse processo é análogo àquele realizado pelo lexicógrafo[111]". Esse argumento é assim desenvolvido considerando-se que os aprendizes terão a oportunidade de utilizar técnicas semelhantes às utilizadas por lexicógrafos e gramáticos, principalmente no que diz respeito à formulação e testagem de hipóteses acerca da utilização da língua.

No entanto, Gavioli (2005) é reticente em relação a essa analogia e tece alguns comentários que merecem ser considerados. O primeiro deles refere-se ao fato dos aprendizes estarem em posição muito diferenciada da posição de um profissional da língua: não são falantes da língua, não necessariamente são estudantes de nível avançado e não dominam um conhecimento teórico tão específico como o do linguista ou lexicógrafo. Segundo, o propósito do aprendiz na utilização do concordanciador e do *corpus* é aprender e entender a língua. "No contexto de ESP, com frequência, o que leva os estudantes a investigar *corpora* de linguagem especializada é a tentativa de entender aspectos característicos de tal linguagem[112]" (GAVIOLI, 2005, p. 88). Um aprendiz da língua inglesa, provavelmente, fará

[110] Conforme tradução de Berber Sardinha (2004)
[111] What is novel about the work reported in this paper is the perception that 'research is too serious to be left to resarchers": that the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data – hence the term "data driven learning" (DDL) to describe the approach. (Johns, 1991, p. 1).
[112] In the ESP environment, what often leads students to investigate *corpora* of specialized language is the attempt to work out characteristic aspects of such language. (GAVIOLI, 2005, p. 88).

perguntas muito diferentes das do pesquisador e, consequentemente, encontrará respostas também diferenciadas em sua consulta ao *corpus* linguístico. A autora aponta essas diferenças, fazendo referência à Bernardini e salientando "que o propósito do aprendiz não é fornecer uma descrição sistemática das características da língua, mas "entender" a linguagem especializada, buscando pistas sobre valores, ideias, conceitos e convenções que estão implícitos no texto[113]" (GAVIOLI, 2005, p. 88). Ainda segundo Gavioli:

> Enquanto a "força" das concordâncias na pesquisa da língua é descrever características interessantes ou desconhecidas do seu uso, a "força" das concordâncias, a partir de um ponto de vista pedagógico, é estimular e aprimorar a percepção[114] linguística dos aprendizes. A primeira abordagem está relacionada com o produto da análise; a segunda, com o processo. Enquanto nenhuma exclui a outra, e é de fato possível que as análises produzidas por estudantes são também "cientificamente interessantes" do ponto de vista da pesquisa linguística; no entanto, a partir de uma perspectiva pedagógica, deveríamos primeiramente estar preocupados com o que é alcançado através do processo, independentemente de ser uma "descoberta científica"[115]. (GAVIOLI, 2005, p. 31-32)

Entende-se aqui, em relação à comparação feita por Johns (1991) entre o aprendiz e o pesquisador, que a semelhança está relacionada às possibilidades oferecidas pelo *corpus* na obtenção de respostas às perguntas da pesquisa. O movimento cognitivo do aprendiz, no processo de investigação e utilização das concordâncias, ou seja, dos padrões linguísticos, poderia mesmo ser semelhante ao do pesquisador profissional, como o do gramático, do linguista ou do lexicógrafo. No entanto, os usos na utilização do instrumental da Linguística de *Corpus* diferem entre estudantes e linguistas. Portanto, entende-se essa analogia como um contraponto a metodologias tradicionais de ensino que centralizam no professor o

---

[113] As Bernardini observes, however, the learners' aim is not to provide systematic descriptions of linguistic characteristics, but rather to "understand" specialized language by collecting clues about values, ideas, concepts and conventions that are implicit in the text. (GAVIOLI, 2005, p. 88)

[114] A autora, orignalmente utilizou o temo intuição. Porém, optou-se traduzir por percepção pois torna-se mais coerente com o conjunto do discurso da autora, bem como o referencial teórico a qual ela se vincula.

[115] While the "power" of concordances in language research is that of describing interesting or previously unknown features of language use, the "power" of concordances from a pedagogic point of view is that of stimulating and enhancing the linguistic intuition of the learners. The former is related to the product of the analysis, the latter to the process. While neither excludes the other, and it is indeed possible that students' analyses are also "scientifically interesting" from the point of view of language research, from a language learning perspective, we should primarily be concerned with what is achieved through the process, independently of the "scientific discovery". (GAVIOLI, 2005, p. 31-32)

papel de transmissor do conhecimento. No entanto, é preciso ter muito claro que os propósitos da busca (além da metodologia de pesquisa) são muito diferenciados. O aprendiz utilizará o DDL para aprender a língua, para apropriar-se de conhecimento linguístico a fim de melhorar seu desempenho na utilização do idioma em sua comunidade ou área de estudo. Esta é a função, neste estudo aceita e replicada, que o DDL desempenhará no ensino da língua estrangeira no contexto de ESP.

A respeito do papel do agente intermediário (*middleman*), Gavioli (2005) pondera que, poderia haver sérios problemas se o intermediário fosse totalmente eliminado, conforme o preconizou Johns. Para operarem efetiva e autonomamente frente ao concordanciador e para fazerem consulta aos dados do *corpus*, os estudantes necessitam de orientação sistemática. Para a autora, essa orientação inicia com o processo de aprender a ler e interpretar os dados proveninentes de um *corpus*. Na ótica de Gavioli, (2005, p. 29) "uma implicação pedagógica que emerge aqui, não é se o 'agente intermediário' (por exemplo, professores) deve ser eliminado ou não, mas muito mais, qual tipo de 'filtro' deveria ser exercido e de que maneira[116]". Embora Johns (1991) proponha que os aprendizes analisem os dados diretamente, Gavioli (2005) considera que ele não trata da problemática relacionada à identificação de questões (problemas) de linguagem a serem mapeadas e nem do modo de ajudar a desenvolver uma metodologia analítica apropriada às necessidades dos alunos.

Feitas essas ressalvas ao DDL, argumenta-se favoravelmente à utilização dessa proposta para o ensino de língua inglesa. Em consonância com as críticas de Gavioli, entende-se o papel e a presença do mediador/professor como necessários, porém relativiza-se sua função e exclui-se a necessidade de sua onipresença, em contraste com o modelo tradicional de ensino. Também fica estabelecido que as questões de pesquisa de um aprendiz na certa produzirão conhecimento e informação pertinentes e necessários ao aprendizado da língua, não, necessariamente, resultando em descoberta científica.

As sugestões de Gavioli (2005) para iniciar os aprendizes na utilização do concordanciador e no uso das concordâncias dependem da mediação do professor

[116] A pedagogic implication emerging here, then, is not whether "the middlemen" (i.e. the teachers) should be cut out or not, but rather which type of "filter" they should exercise and in what way. . (GAVIOLI, 2005, p. 29)

que os auxilia no aprendizado do software (o que é bastante simples), no aprendizado e na leitura de uma concordância que se apresenta distribuída em um layout não-convencional. Além disso, a autora acentua que "[...] o principal ponto consiste em orientar os aprendizes, capacitando-os a fazerem as perguntas apropriadas e a "ler" e interpretar os dados para obterem respostas que façam sentido[117]" (GAVIOLI, 2005, p. 71). A autora também destaca que os aprendizes

> precisam familiarizar-se com diferentes visões da língua, e em particular, com o fato de as combinações lógicas serem apenas uma parte do uso de uma língua. O uso da língua é de forma massiva baseado em combinações, as quais são muito mais idiomáticas do que de natureza lógica e, como tais, não podem ser explicadas racionalmente[118]." (GAVIOLI, 2005, p. 40-41)

3.3.3 Peculiariedade das concordâncias ou os dados brutos na sala de aula

Antes de proseguir, convém retomar o conceito de concordância, o qual será reiteradamente utilizado neste capítulo e nos próximos. Um software concordanciador, como o Concord (do pacote Wordsmith Tools) e o Antconc[119], produz concordâncias. Isto é, o software apresenta a listagem de um item específico, podendo ser formado por uma ou mais palavras, acompanhado do seu co-texto, ou seja, uma certa quantidade de palavras que o antecedem ou precedem (BERBER SARDINHA, 2009). O termo pesquisado é apresentado centralizado, destacado com alguma saliência tipográfica, como cor, itálico, negrito ou sublinhado, variando isso de acordo com configurações dos softwares. Para referir-se a esse termo em destaque na tela do concordanciador também é utilizado a definição "KWC – Key Word in Context".

A figura 6 mostra a tela do Concord a qual apresenta o resultado da pesquisa para o termo "waste". A figura 7 mostra o resultado da consulta para o pacote lexical

---

[117] the main issue consists in enabling students to ask appropriate questions and to "read" and interpret the data to get sensible answers. (GAVIOLI, 2005, p. 71).

[118] In order to enable students to appreciate the type of information they may get from a concordance, they need to be familiarized with a different view of language and in particular with the fact that logical combinations account only in part for language use. Language use is also massively based on combinations which are of an idiomatic rather than a logical nature and as such cannot be explained rationally.

[119] http://www.antlab.sci.waseda.ac.jp/software.html

“due to the”, também realizada com o concord. Cada uma das linhas visualizadas na tela do software concordanciador são as concordâncias.

Figura 6 - Concordâncias do termo "waste"

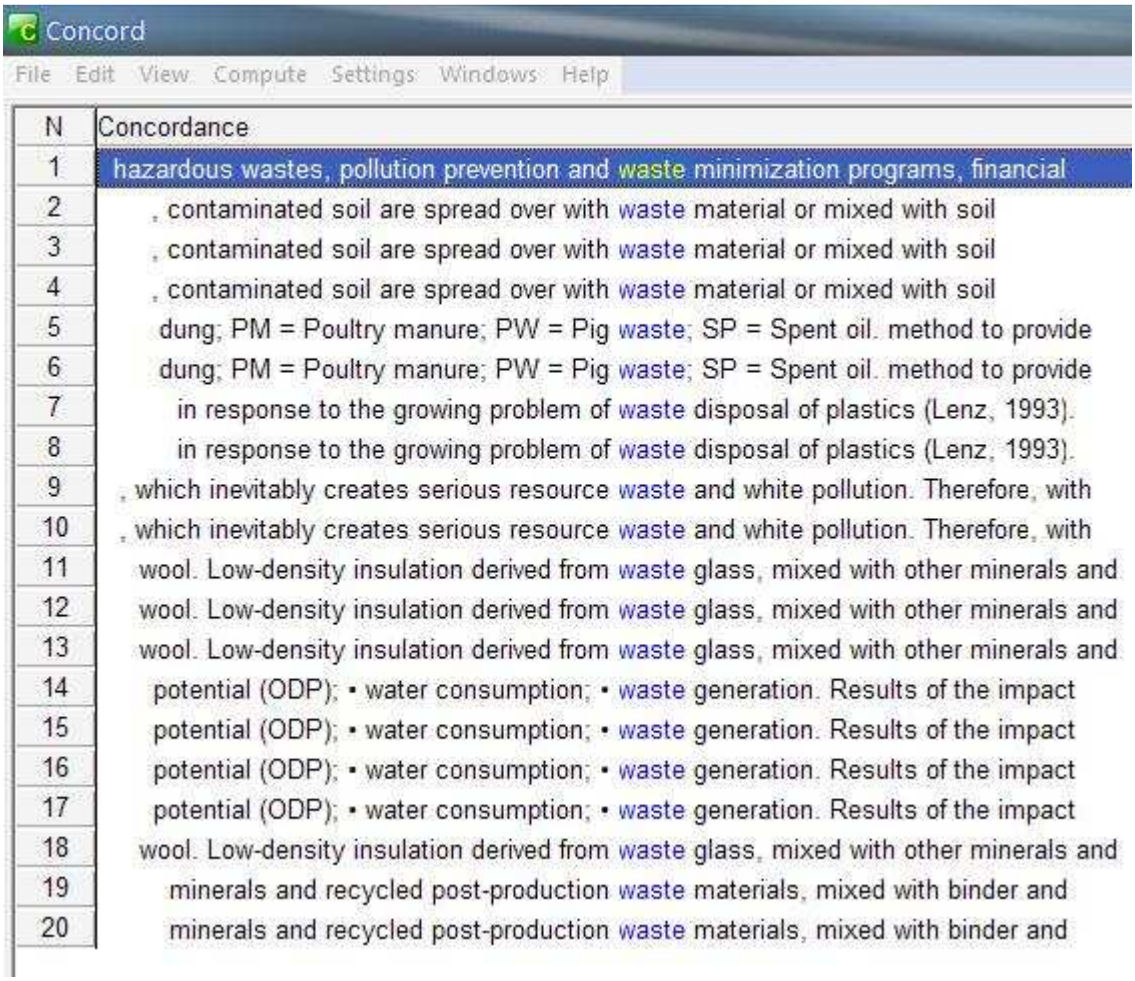
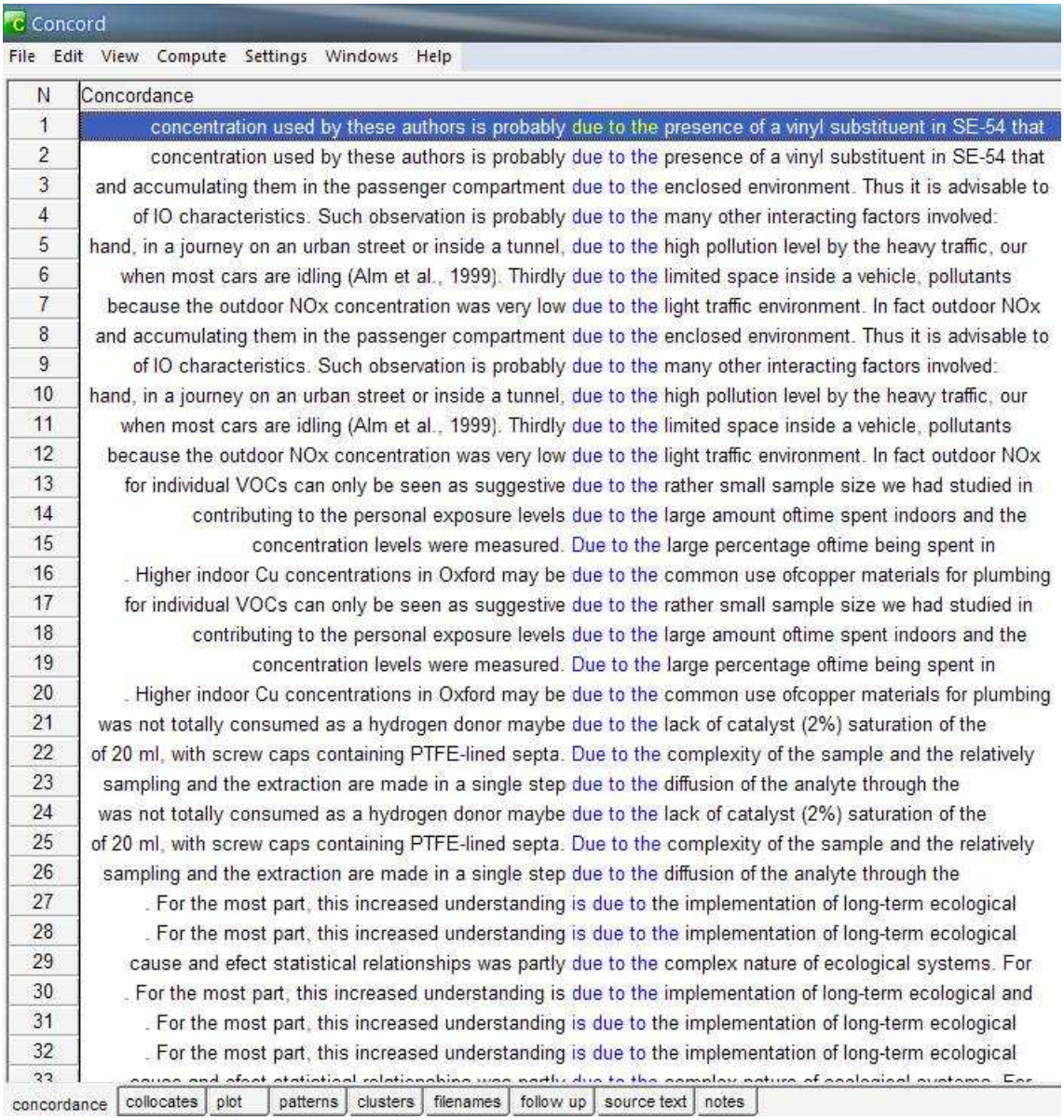


Figura 7 - Concordância do pacote lexical "due to the"



Uma das grandes diferenças na utilização dos dados linguísticos provenientes de um *corpus* composto por textos autênticos, utilizado como material pedagógico para o ensino da língua estrangeira, está no fato de que respostas às perguntas e dúvidas podem não serem conhecidas de antemão. Conforme o afirma Berber Sardinha (2004, p. 292-293) “o DDL é uma abordagem de cunho essencialmente indutivo, ou seja, os alunos produzem conhecimento de modo ascendente (bottom-

up) a partir da observação de dados". No uso do DDL, há certo grau de imprevisibilidade nas buscas das respostas no *corpus*, pois, não se sabe de antemão o que ali será apresentado, ou seja, que dados constam no *corpus*, pois cada pesquisa, de acordo com os termos ou combinações de termos pesquisados, gera um resultado que não é previamente conhecido. A consulta ao *corpus*, tal como o uso da língua em situações de comunicação real, sempre guarda uma margem de incerteza; em ambas, não se sabe de antemão o que será encontrado. Há sempre algo que escapa ao controle pedagógico, quando se utiliza um *corpus*, já que não há como antever com precisão que pesquisa será realizada pelo aluno e qual resultado poderá aparecer, porque as possibilidades são muitas; além do mais, as interpretações produzidas também são variadas. Talvez essa margem de imprevisibilidade possa ser motivo de desconforto para o educador, de vez que o coloca numa posição em que não tem controle absoluto sobre o que poderá surgir da pesquisa feita pelo estudante. Essa circunstância exigirá dele uma postura mais flexível frente ao conhecimento linguístico, ao seu próprio saber e ao ensino. Berber Sardinha, neste mesmo artigo sintetiza a questão ao afirmar que "a diferença no DDL é *que o professor não sabe necessariamente a resposta de antemão*. (BERBER SARDINHA, 2004, p. 292-293). A esse respeito, Tim Johns (1991) destaca a margem de imprevisibilidade como caracterizadora do DDL, ao comentar que:

> O que é distintivo à abordagem DDL em relação ao ensino indutivo de uma língua, é o princípio de que o dado é primário e o professor não sabe com antecedência exatamente que regras ou padrões os alunos irão descobrir: de fato, eles com frequência perceberão coisas desconhecidas, não somente para o professor, mas também para o padrão de materiais de referência sobre a língua. É esse elemento desafiador e estimulador de descobertas que confere ao DDL caráter estimulante e um sabor especial[120]. (Johns, 1991, p. 3).

[120] What is distinctive about the DDL approach to inductive language teaching is the principle that the data is primary, and the teacher does not know in advance exactly what rules or patterns the learners will discover: indeed, they will often notice things that are unknown no only to the teacher, but also to the standard works of reference on the language. It is this element of challenge and of discovery that gives DDL its special flavour and stimulys. (Johns, 1991, p. 3).

Prosseguindo com Laura Gavioli (2005), que preconiza o uso de *corpora* linguístico para o ensino de língua inglesa, para áreas técnicas, a autora destaca, também, que um dos pontos mais interessantes dessa abordagem está na possibilidade de "os alunos observarem coisas que com frequência passam despercebidas, não captando a atenção do professor", e acrescenta que fatos como esse "promovem discussão sobre características e convenções da língua durante as aulas, colocando estudantes e professores "em pé de igualdade" (GAVIOLI, 2005, p. 99). Segundo ela, ao analisar, pedagogicamente, a utilização do *corpus* no ensino, esse uso é interessante não apenas por fornecer uma base de textos, "mas por fornecer uma metodologia para investigar esses textos" (GAVIOLI, 2005, p. 69). O vínculo entre textos autênticos e uso das concordâncias em sala de aula, pode provocar indagações, questionamentos e discussões sobre a linguagem, ou seja, o trabalho com *corpus* e concordâncias fornece dados lexicogramaticais que estão à espera de interpretação. Sobre isso, Gavioli assegura que:

> Concordâncias originadas de um *corpus* fornecem material ainda não interpretado a que é necessário atribuir algum sentido. A tentativa de dar sentido aos dados provoca o que Widdowson (1998, p. 713) chama de "reação pragmática", que ocorre quando os analistas examinam os dados em busca de resposta às suas indagações. Por essa razão, o que torna as concordâncias materiais interessantes está relacionado ao seu limite em produzir material ainda não interpretado/analisado. Do mesmo modo, *corpora* especializados fornecem materiais pouco generalizáveis em comparação com o material derivado de outros tipos de *corpus* (por exemplo, *corpora* gerais) sendo necessário compará-los dentre si. A necessidade de comparar o material, novamente, produz uma abordagem investigativa, que é exatamente o que faz do trabalho interessante[121] (GAVIOLI, 2005, p. 69)

O processo de investigação dos dados coletados de um *corpus* pode continuar a ser enriquecedor para os aprendizes, que após adquirirem autonomia no uso do instrumental e da metodologia, poderão estender as pesquisas a outros locais que não a sala de aula, podendo incorporá-los a seu aprendizado continuado da língua.

[121] Concordances derived from *corpora* provide uninterpreted material which needs to be given a sense. The attempt to give a sense to the data provokes what Widdowson (1998a: 713) calls a "pragmatic reaction" by the analysts who examine their data to see if it answers their questions. So the limit of producing uninterpreted material is what makes concordance material interesting. Likewise, specialised *corpora* provide material which is hardly generalizable and comparison with material derived from other types of *corpora* (e.g. general ones) is necessary. The necessity to compare the material,again, produces an investigative approach that is exactly what makes the work interesting. (GAVIOLI, 2005, p. 69)

Mais do que isso, pode ocorrer que análises e observações produzidas pelos alunos contribuam para investigações mais amplas, tornando as explicações obtidas mais claras e precisas, relacionando-as às tentativas de explicação do professor, agregando assim, valor ao conhecimento e a todos os agentes envolvidos no processo de sua construção.

3.3.4 Amostras X exemplos

Os dados de um *corpus* linguístico, ao serem analisados no concordanciador, devem ser vistos como um tipo "particular" de dado. Gavioli (2005) argumenta que a razão dessa particularidade está, primeiramente, no fato de esses dados serem provenientes de *amostras (samples)* ao invés de *exemplos,* pois *amostras* precisam ser analisadas e interpretadas, em relação ao seu *corpus* originário (e tal relação deve ser muito clara para o usuário), através de uma metodologia apropriada, devido à configuração com que os dados são apresentados no concordanciador. Assim posto, é importante ter em mente que o concordanciador apresenta *amostras de linguagem*, pois, a partir de uma operação realizada por um algoritmo computacional, os dados apresentados, pela lógica computacional, não passam de uma combinação de caracteres. De acordo com Gavioli (2005, p. 73), somente a partir da intervenção humana, ou seja, a partir do processo de leitura e interpretação, os dados poderão ser entendidos e então classificados como exemplos. Conforme suas próprias palavras "enquanto amostras podem, com certeza, ser usadas como fonte de exemplos, essa transformação requer a intervenção humana através seleção e análise[122]" (GAVIOLI, 2005, p. 73).

Os dados coletados de um *corpus* linguístico e visualizados em um concordanciador, na perspectiva de Gavioli (2005, p. 73), tornam-se "um tipo de material didático completamente diferente daqueles tradicionalmente utilizados em sala de aula". Ao contrário de dicionários, gramáticas e livros didáticos, "não oferecem exemplos, [...], pois simplesmente fornecem dados, sendo

[122] The concordance, by contrast, provides samples, selected for the sole reason that they contain a particular combination of characters. While samples can of course be used as a source of examples, this transformation requires human intervention via analysis and selection. (GAVIOLI, 2005, p. 73)

responsabilidade e tarefa dos usuários explicá-los[123]", ou melhor, construir suas próprias explicações. A autora aprofunda suas reflexões a esse respeito, dizendo que mesmo quando uma pesquisa não gera uma quantidade suficiente de concordâncias para poder ser propriamente analisada, mesmo essa falta, é capaz de disparar um movimento cognitivo importante nos alunos, que os auxilia no aprendizado da segunda língua. Conforme o atesta a autora:

> Eu já chamei a atenção para o fato de as concordâncias não fornecerem "exemplificações", o que é, pelo meu entendimento, o que provoca uma reação pragmática nos estudantes, levando-os a produzir sentido a partir de um material aparentemente sem sentido. Mesmo naqueles casos em que uma pesquisa não produz resultado ou produz resultado insuficiente para levantar uma hipótese, essa escassez de material provoca uma reação nos estudantes, fazendo-os se questionarem sobre o porquê disso. Esse "limite" do material, proveniente das concordâncias, leva os alunos a interagir com outros materiais, tanto proveniente de um *corpus* diferente (por exemplo, um *corpus* maior) ou de *corpora* em uma língua diferente (por exemplo, uma língua estrangeira ou sua própria), a fim de realizarem uma análise comparativa[124]. (GAVIOLI, 2005, p. 99)

Quanto à utilização dos dados linguísticos advindos de *corpora* em sala de aula, Gavioli (2005) ressalta a importância de os alunos perceberem a natureza dos dados contidos no *corpus*, isto é, a não tratá-los como exemplos da língua, mas sim como uma amostra. Não havendo essa clareza, a potencialidade do *corpus* seria subutilizada ou se correria o risco de a proposta não ter sido plenamente compreendida.

---

[123] This makes concordances a very different type of material from those traditionally used in the classroom. Unlike dictionaries, grammars and textbooks, a concordance does not offer explanations; as noted above (Section 2.3 in particular), it merely provides data which it is the user's task or responsibility to explain. (GAVIOLI, 2005, p. 72)

[124] Indeed, they may provide no occurrence at all or not enough of them. In Chapter 3, I noted that the very fact that concordances do not provide "exemplifications" is what, in my view, provokes a pragmatic reaction from the students who are encouraged to make sense out of seemingly senseless material. Even in those cases when a search does not produce any occurrence or enough evidence to build up hypotheses, such lack of material provokes a reaction from the students who generally ask why this is so. This "limit" of concordance material, then, prompts the students to interact with more materials, either from different *corpora* (e.g. larger *corpora*) or from *corpora* in ifferent languages (e.g. the foreign language and their own), thus carrying out comparative work. (GAVIOLI, 2005, p. 99)

3.3.5 Concordâncias e texto na sala de aula

Berber Sardinha (2009, p. 20) além de salientar a necessidade da intervenção do professor, principalmente, durante o processo de familiarização com as concordâncias, destaca que "[...] é fundamental ganhar experiência na análise de concordâncias, 'treinando os olhos' para perceber padrões que sejam 'interessantes'" (BERBER SARDINHA, 2009, p. 20). Ou seja, o treino aliado à experiência na análise das concordâncias, pode levar os alunos a "uma apreciação diferente da língua, baseada na lexicogramática" (BERBER SARDINHA, 2009, p. 20). Esse processo de conscientização do aluno sobre a lexicogramática pode ser catalisado pelo uso do concordanciador, pois

> Normalmente os aprendizes de língua estrangeira tendem a ver a linguagem como um conjunto de palavras individuais sustentadas por regras gramaticais; desse modo, a compreensão e a produção se dão, em muitos casos, com base na junção de palavras individuais. O resultado disso é que, por exemplo, a fluência, um aspecto natural da fala na língua materna, é conseguida somente a duras penas na língua estrangeira. (BERBER SARDINHA, 2009, p. 13)

O autor insiste na questão explicitando porque é tão importante o aluno desenvolver sua capacidade de percepção e consciência linguística, como bem o atesta a sequência da citação, logo a seguir:

> O trabalho com *corpora* em sala de aula pode ajudar a melhorar essa situação, pois ele naturalmente faz saltar aos olhos essas 'unidades pré-fabricadas', que são os padrões léxico-gramaticais. Ou seja, começando por um trabalho de conscientização, o aluno passa a ter uma visão diferente do que é o vocabulário de uma língua, deixando de lado a ideia de que se trata de um conjunto de palavras isoladas que se juntam por meio de regras gramaticais. Ele passa a perceber que as palavras se juntam por meio de atração mútua, via de regra explicada somente pelo uso e não por regras de gramática, e que essa atração cria os agrupamentos, pacotes e 'chunks', que por sua vez se juntam e formam o tecido da linguagem. (BERBER SARDINHA, 2009, p. 13)

O uso de um *corpus* linguístico na sala de aula, seja através das concordâncias, seja através de outras possibilidades de utilização, deveria estar atrelado ao uso do texto, preferencialmente, originário dessas concordâncias utilizadas. Nessa situação, no dizer de Sardinha: "Temos, desse modo, duas

grandes fontes de input de língua na sala de aula: o texto e o *corpus*" (BERBER SARDINHA, 2009, p. 12). Além disso, com a utilização integrada do material proveniente do *corpus* e dos textos, o aluno poderá, sempre que sentir necessidade, estabelecer relações entre o gênero textual utilizado e o discurso organizado – neste caso, na forma de artigos científicos. Isso pode ser provindenciado pelo professor, tanto pela disponibilização de textos impressos que acompanhem páginas (também impressas) de linhas de concordância (quando, por exemplo, for utilizado em um local sem acesso a computador) ou mesmo, através de processos informáticos automatizados, no caso da utilização de algum software concordanciador como o *Wordsmith Tools* 5.0 ou o *AntConc* 3.2.1[125]. Ao utilizar alguns desses softwares (entre outros existentes), o usuário pode, ao acessar uma concordância, se assim o preferir, imediatamente localizar o arquivo completo contendo todo o texto utilizado para a composição do *corpus*. Ainda, dependendo do software ou das configurações, poderá acessar frações menores do texto, como o parágrafo, ou mesmo, apenas a frase contendo o termo de busca investigado. Conforme pontua Berber Sardinha:

> [...] texto e *corpus* são duas unidades da língua que se completam naturalmente, até porque a segunda é uma coletânea da primeira; em outras palavras, o trabalho com *corpus* na sala de aula *pede* um texto. Sem um texto em que se orientar, o aluno por se sentir perdido, sem um apoio em uma unidade concreta de comunicação humana. (BERBER SARDINHA, 2009, p. 12)

Além disso, ao desenvolver uma prática de sala de aula movida por dados (DDL), orientada por um princípio que procura articulá-la ao uso dos textos (incluindo textos analisados e pesquisados no *corpus*) conjugadamente à proposição de atividades que intercomplementem a pesquisa no *corpus* com a utilização do concordanciador e com a leitura dos textos originais, acredita-se estar evitando a repetição das práticas tradicionais de uma gramática da sentença, completamente dissociada de seu discurso, conforme já referido. Mais ainda, o uso de textos autênticos e de *corpora* na sala de aula de língua estrangeira coloca os alunos em contato com linguagem autêntica e com uma grande quantidade de *input,* facilmente

[125] Software distribuído gratuitamente, acessado em: http://www.antlab.sci.waseda.ac.jp/software.html

disponibilizada pela tecnologia. Uma prática de ensino assim articulada “pode trazer benefícios para os alunos, pois eles entram em contato com muitas ocorrências de padrões que se repetem em outros textos, permitindo assim desenvolver a consciência a respeito da natureza probabilística e associativa da língua, visando ao desenvolvimento da fluência e ao enriquecimento do vocabulário” (BERBER SARDINHA, 2009, p. 20). Objetivos estes, prioritários na proposta aqui desenvolvida.

3.4 Produção de materiais e tarefas

O propósito desta pesquisa é a utilização do *corpus* linguístico aqui apresentado para a produção de material didático e para o desenvolvimento de uma tarefa. Conjuntamente, prevê-se o ensino das sequências formulaicas e dos pacotes lexicais. Nesse sentido, são apresentados alguns estudos sobre o ensino de sequências formulaicas, os quais forneceram subsídios para a investigação e podem fornecê-los aos educadores. Além disso, dada a importância das tarefas para o ensino de língua inglesa, seu conceito será discutido, sempre que possível, relacionadamente, à produção de material didático para o ensino de ESP.

O DDL (Data driven learning) introduziu a possibilidade de utilizar diretamente com os alunos os dados linguísticos, com grande ênfase na utilização das concordâncias. A prática de sala de aula que se fundamenta no uso das concordâncias é também conhecida por *classroom concordancing* (BERBER SARDINHA, 2011). No entanto, outras possibilidades de uso de recursos advindos da Linguística de *Corpus*, com aprendizes, coexistem, ou seja, existem “alternativas que in*corpora*m instrumentos além das concordâncias, como listas de palavras, palavras-chave e pacotes lexicais/cluster, e não se restringem à concordância como o ‘texto de trabalho’ da atividade” (BERBER SARDINHA, 2011, p. 2).

Berber Sardinha (2011) apresenta três tipos de atividades de ensino que podem ser produzidas ou executadas com *corpora*:

Atividades centradas na concordância

Atividades centradas no texto e em gênero

Atividades Multimídia/Multigênero

A atenção será focada nas duas primeiras atividades (centradas na concordância e centradas no texto) por serem pertinentes ao ensino da leitura no contexto de ESP.

3.4.1 Atividades centradas na concordância

Atividades centradas na concordância, segundo Berber Sardinha (2011, p. 3) são os materias de ensino que têm como referência a utilização da concordância, isto é, são "'centrados na concordância', visto que ela é a peça central, senão a única, da atividade" (BERBER SARDINHA, 2011, p. 2). Nesse tipo de atividade a Linguística de *Corpus* é "constitutiva da atividade, ou seja, sem ela, a modalidade não existiria." Nessa modalidade encaixam-se todas as propostas e concepções acerca do DDL, discutidas anteriormente.

3.4.2 Atividades centradas no texto

Nas atividades centradas no texto, as concordâncias tornam-se "coadjuvantes" e o texto tranforma-se no "ator principal" da cena. "O foco passa a ser o texto de apoio" e a concordância, seja ela impressa ou apresentada diretamente na tela do computador, "transforma-se em mais um elemento da atividade" (BERBER SARDINHA, 2011, p. 14).

Nesse âmbito, atividades voltadas ao texto serão realizadas em paralelo com atividades que fazem uso das concordâncias, procurando estabelecer intercomplementação mútua entre os dados do *corpus* e o texto fonte. As atividades centradas no texto, segundo Berber Sardinha (2011), são uma resposta a algumas desvantagens apresentadas pelo uso das concordâncias, entre elas, a sua desvinculação dos textos que as originaram. Os aprendizes, muitas vezes, encontram dificuldade na utilização dos dados brutos, sem a possibilidade de consulta aos textos originais, onde a lexicogramática é representada. Enfim, "um texto contextualiza o uso linguístico em uma dada situação social, histórica, de produção, de reprodução, em um gênero (mais ou menos) determinado,

enfim, com as características que tornam a linguagem discurso" (BERBER SARDINHA, 2011, p. 14).

3.4.3 Seleção de concordâncias: critérios e cuidados

Dependendo do nível de conhecimento dos alunos, sobretudo, se o trabalho for desenvolvido com alunos pouco afeitos ao uso de concordanciadores, pode ser necessário selecionar as linhas de concordância. Para alunos avançados ou para alunos "com bastante exposição a concordâncias, a seleção intencional de linhas pode ser desnecessária" (Berber Sardinha, 2006, p. 154). É preciso ter em mente que o aluno, ao utilizar um concordanciador, poderá acessar uma grande quantidade de dados "brutos" para serem analisados e interpretados, o que exige grande esforço cognitivo. Dependendo do termo pesquisado no *corpus*, o concordanciador poderá apresentar centenas de linhas contendo a palavra ou expressão pesquisada (KWIC - key word in context). Por essa razão, Berber Sardinha (2006. P. 154) recomenda que, "como regra geral, creio ser desejável evitar concordâncias com mais de 30 linhas ou uma página de papel impresso de extensão, devido ao esforço cognitivo que sua análise demanda" (SARDINHA, 2006, p. 154).

No evento de a seleção tornar-se nececessária para realizar uma tarefa ou para produzir material pedagógico, ao fazê-la, é preciso ter o cuidado de não impor critérios que possam alterar o grau de representatividade do *corpus*. Tim Johns chama a atenção para essa possibilidade, afirmando que:

> O princípio mais importante a ser levado em conta ao realizar essa tarefa (a seleção, grifo meu) é que o inevitável processo de seleção não deve distorcer a evidência – isto é, os extratos de concordâncias selecionados devem representar o mais fidedignamente possível toda a gama de características linguísticas e comunicativas contidas nos dados brutos. Há duas principais formas de distorção. A primeira pode ocorrer se a seleção for realizada a partir de um critério linguístico imposto externamente – por exemplo, os preconceitos do professor sobre o que deveria estar selecionado nos dados ao invés do que realmente está registrado nos dados do *corpus*. O segundo pode ocorrer no evento de a seleção ser feita partindo de um critério pedagógico, o que poderia ser, por si mesmo, perfeitamente justificável (por exemplo, que preferência deveria ser dada a citações que são relativamente autônomas ou autoexplicativas), mas as quais podem influenciar tendenciosamente as amostras em termos dos

> significados representados pelo que aquelas formas expressam[126]. (JOHNS, 1994, p. 298)

3.5 Ensino das sequências formulaicas e dos pacotes lexicais

Já foi salientado anteriormente que, para o aprendiz conseguir apreender as sequências formulaicas e os pacotes lexicais, ele deve, primeiramente, estar ciente da sua função e de sua existência na língua. Os softwares concordanciadores, pela forma com que apresentam uma palavra em contexto (centralizada, destacada e acompanhada de cotexto), em consequência desse layout, acabam destacando os padrões lexicogramaticais, o que pode ser um facilitador para o seu reconhecimento, embora isso não seja necessariamente uma garantia de aprendizado. Em um texto, por outro lado, não há marcadores que indiquem onde inicia ou termina um pacote lexical ou uma sequência formulaica. Por isso mesmo, é importante, nessas circunstâncias, a intervenção do professor ou de outro mediador (incluindo materiais didáticos de qualidade) com mais experiênica no uso da língua, principalmente no início do trabalho com a lexicogramática. Essa intervenção será fundamental no sentido de familiarizar os aprendizes, orientando-os quanto ao que se entende por padrões linguísticos. Ou seja, os alunos precisam aprender a vê-los e reconhecê-los, tanto nos textos, quanto no concordanciador.

A utilização de estratégias que dão destaque às sequências formulaicas e à lexicogramática presente em um texto, é um item a ser contemplado na produção de materiais. Willis (2009, p. 185) enfatiza que os professores "podem auxiliar os aprendizes a desenvolver sistemas apropriados, destacando (*highlight*) termos no texto e mostrando as regularidades presentes." O processo de ensino da lexicogramática, segundo Willis, deve ser iniciado de imediato como parte integrante

[126] The most important principle that has to be borne in mind in carrying out this task is that the inevitable process of selection should not distort the evidence – that is to say, the concordance extracts chosen should represent as far as possible the full range of linguistics and communicative features of the raw data. There are two main sources of distortion. The first can occur if selection is made on linguistic criteria that are impose externally – for example, the teacher's preconceptions of what *ought to be* in the data rather than what is in the data. The second can occur if selection is made on pedagogic criteria that may in themselves be perfectly justifiable (for examples, that preference should be given to citations that are relatively self contained and self-explanatory) but which have the unforeseen effect of biasing the sample in terms of the forms represented of the meaninings that those forms convey. (JOHNS, 1994, p. 298)

do ensino do léxico. Conforme o autor: "À medida que o léxico é desenvolvido é possível expor os alunos a uma quantidade maior de texto, oportunizando-lhes mais e mais oportunidades para exploração (da língua, grifo meu)[127]" (WILLIS, 2009, p. 185). O destaque aos pacotes lexicais, bem como a utilização de um caderno de vocabulário, são estratégias que podem favorecer o aprendizado, pois

> Sublinhar ou marcar com cores (highlighting) os padrões que se repetem com frequência em textos e diálogos pode ser uma maneira de chamar a atenção para eles (raising awareness); estimular estudantes a registrar os itens lexicais com todos os seus termos, em seus cadernos de vocabulário, poderá sensibilizá-los (raise awareness) sobre a integridade desses padrões que poderão ser utilizados em uma grande variedade de situações comunicativas[128]" ." (O'KEEFFE, MCCARTHY e CARTER, 2007, p. 3)

Bishop (2004) relata que em um estudo, no qual determinadas sequências formulaicas desconhecidas tiveram algum tipo de "destaque/*highlight*" nos textos, elas foram com mais frequência consultadas em glossários ou em outros materiais de referência, quando comparadas às palavras, do grupo controle, sem nenhum destaque. Para o autor, a estratégia é uma evidência que suporta a alegação de que as sequências formulaicas previamente desconhecidas não são notadas durante a leitura, pois sequências formulaicas e pacotes lexicais não possuem marcadores que limitem seu iníco e fim, tal como palavras, que são demarcadas por espaços. Logo, destacar "tipograficamente as sequências formulaicas é, de certa forma, uma maneira de tornar o lexema (multipalavra/multiword) visível para o leitor. A cor e o sublinhado, argumenta-se, podem sinalizar a natureza holística de uma sequência formulaica[129]" (BISHOP, 2004, p. 239). Assim, é necessário pensar formas de fazer com que o aluno note as sequências formulaicas e os pacotes lexicais no texto.

---

[127] Teachers can help learners to develop appropriate systems by highlighting them in text (recognition) and by pointing to regularities in the way they are organised (system building), but the whole process must be kick-started by the acquisition of lexis. As lexis is acquired, so it is possible to expose learners to more and more texts, and provide more and more opportunities for exploration. (WILLIS, 2009, p. 185)

[128] "Underlining or colour-highliglhting patterns which are frequently repeated in texts and dialogues may be one of raising awareness of useful chunks, and encouraging students to record whole chunks in their vocabulary notebooks may raise awareness of their fullness as frames that can be used with a potentially large number of utterances." (O'KEEFFE, MCCARTHY e CARTER, 2007, p. 3)

[129] Adding typographical salience to a formulaic sequence is in a sense making the (multi word) lexeme visible to the reader. The color and the underlining, it is argued, signal the holistic nature of the FS. (BISHOP, 2004, p. 239)

A atenção dispensada a um pacote lexical ou a uma expressão formulaica não precisa se restringir à saliência tipográfica usada no texto impresso, podendo-se empregar outras estratégias, dentre elas comentários do professor, observações sobre determinado padrão e mesmo explicações mais detalhadas sobre o uso de uma expressão. No caso de serem utilizados textos sonoros como a gravação de uma palestra ou aula, o professor pode solicitar que os alunos observem determinadas características fonológicas de uma sequência formulaica. Em suma, atividades pedagógicas devem ser pensadas no sentido de direcionar a atenção dos alunos para os padrões da lexicogramática. Mais ainda, é necessário planejar atividades criativas que, de alguma forma, chamem a atenção dos aprendizes para esse fenômeno linguístico, para que eles o percebam e reflitam sobre a língua alvo, a fim de se tornarem usuários proficientes.

A introjeção do conceito de sequência formulaica e de pacote lexical, por parte do aprendiz, pressupõe-se, virá com o tempo, com a experiência adquirida na sala de aula por meio de práticas que o levem a perceber e reconhecer os padrões que compõem a trama textual, e com a sua experiência individual de leitura. Aos poucos, durante a atividade leitora, o leitor treinado começa a “desconfiar” que determinado termo que não entende pode fazer parte de um pacote lexical. Assim, no caso de buscar suporte em alguma obra de referência poderá também verificar se aquele termo e seu cotexto compõem uma estrutura lexicogramatical.

Outra estratégia a ser utilizada, na produção de materiais e tarefas, diz respeito à repetição de termos. Conforme já dito antes, aprende-se com mais facilidade termos recorrentes, como bem o enfatiza Wood: “As atividades de sala de aula poderiam consistir de exposição a grandes quantidades de inputs, com especial atenção ao uso das sequências formulaicas” (WOOD 2002, p. 10). A consulta ao *corpus*, a partir dos concordanciadores, geralmente apresenta vários exemplos da palavra ou expressão consultada, fornecendo grandes quantidades de *input*. A prática da consulta às concordâncias, pressupõe-se, favorece o contato com as ocorrências da língua. Mais ainda, é também uma forma de acessar a linguagem natural produzida pelos vários usuários dos gêneros em questão. No entanto, conforme já postulado com relação à utilização dos textos, a prática deverá ser realizada conjunta e integradamente com a exploração de textos (que podem conter

lexical bundles destacados), empregando estratégias que, de alguma forma, salientem as sequências formulaicas, a fim de sensibilizar os aprendizes acerca da sua função e importância.

O conceito de sensibilização (*awareness raising*) é central para o encaminhamento das práticas de ensino, sendo um princípio orientador das intervenções do professor, bem como da produção de materiais de ensino de língua estrangeira. Conforme definido por Lewis (1993),

> A sensibilização (awareness raising) é um termo que, recentemente, se tornou usual na terminologia relacionada ao ensino de línguas. A característica comum, que está por trás de todos os comentários, é a assertiva de que a habilidade do estudante para observar, de forma acurada, percebendo similiaridades e diferenças nos dados da língua alvo, provavelmente contribua para a aquisição do sistema gramatical. Nesse quadro teórico, a gramática, como uma habilidade receptiva, desempenha um papel importante[130]. (LEWIS, 1993, p. 154)

Atividades que despertem a atenção ou a consciência dos aprendizes (*awareness raising activities*) são de grande importância para o ensino das sequências formulaicas. A proposta dessas atividades é encorajar os aprendizes a observarem atentamente a língua estudada, prestando atenção às formas utilizadas em determinado discurso e/ou gênero textual, refletindo sobre a relação entre esses padrões e estruturas, o léxico e o sentido produzido. No desenvolvimento das tarefas, tanto de práticas de aula, como de exercícios constantes de materiais didáticos, a meta é utilizar estratégias que sensibilizem os alunos levando-os a pensar sobre a constituição da língua e sobre a trama textual produtora de sentido. Willis (1990) comenta a maneira pela qual se apropriou do princípio da sensibilização (*awareness raising*) na elaboração dos exercícios que compuseram o curso por ele elaborado. “Todos os exercícios reforçaram a mesma abordagem metodológica. Eles estimulavam os alunos a observar criticamente o *corpus* e a fazer generalizações sobre a língua a qual foram expostos. Também desafiamos os

[130] ‘Awareness raising’ is a term which has recently acquired currency in language teaching terminology. The unifyin feature behind all the commentator is the assertion that it is the students’ ability to observe accurately, and perceive similarity and difference within target language data which is most likely to aid the acquisition of the grammatical system. Withis this theoretical framework, grammar as a receptive skill has an important role to play.

alunos a referenciarem (a consultar novamente, grifo meu) os dados linguísticos anteriormente vistos" (WILLIS, 1990, p. 84).

Lewis (1993), em seu *Lexical Approach*, enfatiza que os alunos devem primeiramente aprender a observar a língua e que o aprendizado da "gramática" é consequência desse olhar atento ao fenômeno linguístico. Para ele, a gramática é um aprendizado receptivo, que é por ele chamado de "*receptive grammar*/gramática receptiva". Ou seja, o aprendizado da gramática, e também, destaca-se, o aprendizado da lexicogramática, é resultante das experiências do usuário com a língua. "O reconhecimento de que a gramática e o aprendizado dos itens lexicais (lexical chunking ou lexicogramática) como uma habilidade receptiva é questão central para o *Lexical Approach*[131]" (LEWIS, 1993, 149). Claro, que no contexto de ensino e aprendizagem, conforme já comentado, tanto a experiência do aluno, quanto a intervenção de seus pares e do professor, são catalisadores extremamentes importantes do processo.

A gramática receptiva proposta por Lewis é defendida por ele como um contraponto a cursos gerais de língua inglesa que, num primeiro momento, sobrecarregam os aprendizes com atividades de produção linguística, com grande ênfase na "frase perfeita", idealizada. Para ele, alunos iniciantes deveriam ser encorajados a observar e refletir sobre a língua alvo, e, somente após terem um bom repertório, então, sim, poderiam começar a correr riscos, a se expor com atividades de produção. Dada a sua importância para o ensino, em especial, a ênfase conferida à inclusão da lexicogramática como base de um programa de ensino, faz-se pertinente apresentar sua definição de gramática como uma habilidade receptiva. Segundo Lewis (1993):

> O *Lexical Approach* propõe uma redução dramática do papel daquilo geralmenre entendido como "ensino de gramática". Igualmente, há destaque especial ao trabalho da gramática que é aqui entendido de uma forma radicalmente diferente. Este novo estilo de gramática é essencialmente receptivo, e por ser baseado no desenvolvimento da sensibilização/atenção do aluno (student´s awareness), é enfaticamente centrado no aluno, ao invés de no professor. [...] Os princípios informativos são um entendimento da gramática como uma habilidade linguística receptiva e a relação simbiótica entre explicação e prática. Explicação e prática estão

[131] The recognition of grammar and lexical chunking as a receptive skill is central to the Lexical Approach. (LEWIS, 1993, p. 149)

> inextrincavelmente emaranhadas de um forma que tradicionalmente não foi o caso[132] (LEWIS, 1993, p. 149).

Por conhecimento da lexicogramática, entende-se o aprendizado das sequências formulaicas e dos pacotes lexicais que de início são observados e, quando for o caso, incluídos em construções maiores. Prevê-se, após, seu reconhecimento, que terá, em grande parte, uma base receptiva. Ao propor que os alunos devam ter grande oportunidade de exposição à língua, e, no caso de um curso de ensino instrumental da leitura, conforme aqui desenhado, no qual se privilegia o acesso a uma grande quantidade de *input* disponibilizada em um *corpus* e em textos, propõe-se que a ela se vincule, implicitamente, a noção de "gramática receptiva", tal como definida por Lewis. Sua definição de *gramática receptiva* é um conceito pertinente, indo ao encontro de outras tantas pesquisas da área de psicolinguística. Michael Hoey, por exemplo, considera que a gramática é "o produto da acumulação de todos os *primiming* lexicais (assimilação do léxico, grifo meu) de um indivíduo ao longo de sua vida[133]" (Hoey, 2005, p. 159 apud RÖMER, 2009, p. 145). Römer detalha o processo, acrescentando que:

> Central para a teoria de Hoey é a observação de que "cada palavra é por nós assimilada (*primed for us*, grifo meu) discursivamente como um resultado dos efeitos cumulativos dos encontros (prévios, grifo meu) de um indivíduo com as palavras" (Hoey, 2005, p. 13; see also Hoey, 2004, p. 386; and Hoey no prelo). Em outras palavras, à medida que nos deparamos com palavras, tanto no discurso oral como no escrito, e as usamos por nós mesmos, automaticamente, abstraímos seu padrão de uso (léxicogramática, grifo meu) e aprendemos em qual estrutura, posição textual ou em que tipo de texto elas tipicamente ocorrem. Nesse processo, termos já assimilados (*existing priming, grifo meu*) podem ser reforçados ou enfraquecidos (veja Hoey, 2005, p. 9). Como resultado, nosso conhecimento sobre uma palavra é, de acordo com Hoey, inteiramente dependente de nossas experiências com ela (por exemplo, como a vimos/escutamos sendo usada e como a utilizamos nós mesmos)[134]". (RÖMER, 2009, p. 145)

---

[132] The Lexical Approach proposes a gretlay diminished role for what is usually understood by 'grammar teaching'. Equally, there is an enhanced role for the grammar work which is radically different. This new style of grammar is primarily receptive, and because it is based on raising student's awareness, is powerfully student – rather than teacher-centred. Teacher will need to show students used to more authoritarian grammar teaching that this new approach is not only valid, but more helpful. The informing principles are an understanding of grammar as a receptive skill, and the symbiotic relationship between explanation and practice. Explanation and practice are inextricably intertwined in a way which has not traditionally been the case. (LEWIS, 1993, p. 149)

[133] Grammar then is "the product of the accumulation of all the lexical primings of an individual's lifetime" (Hoey, 2005, p. 159 apud RÖMER, 2009, p. 145)

[134] Central to Hoey's theory is the observations that "[e]very word is primed for us in discourse as a result of the cumulative effects of and individual's encounters with the words". (Hoey, 2005, p. 13; see

Embora, para um curso de língua inglesa instrumental, essencialmente, focado no desenvolvimento da capacidade leitora, possa parecer óbvio demais fazer apologia do ensino das habilidades receptivas, já que a leitura é entendida como tal, pretende-se deixar muito claro que a ênfase na *gramática receptiva* é uma escolha. Pressupõe-se que o desenvolvimento de uma grande quantidade de vocabulário receptivo, leve ao entendimento do texto e de seu sentido, e esse possa levar ao desenvolvimento do conhecimento gramatical do texto, conforme proposto por Hoey (entre outros autores). Em um período muito curto, como normalmente ocorre para o desenvolvimento de um curso de inglês instrumental (raramente maior que um semestre letivo) faz muito mais sentido desenvolver bem uma única habilidade do que medianamente todas as demais. Além disso, entende-se que o aprendiz que detiver conhecimento de uma quantidade razoável de vocabulário (conforme aqui proposto) e, ainda, conhecimento adicional da lexicogramática do texto, poderá obter melhores resultados no desenvolvidmento das outras habilidades. Essa pressuposição, no entanto, merece investigação e estudo aprofundados.

3.5.1 O todo e as partes

Um dos pontos controversos acerca do ensino das sequências formulaicas diz respeito ao modo como o seu ensino deva ocorrer, havendo duas possibilidades: a forma holística ou a analítica. De modo muito breve, serão apontadas algumas posições que defendem um ou outro estilo.

Alison Wray (1999) destaca a natureza não-analítica das sequências formulaicas para o desenvolvimento da competência e da fluência linguística de falantes nativos. A tentativa de professores e escritores de livros didáticos para aprendizes da língua inglesa, bem como de outros profissionais que podem ter influência no planejamento de programas e materiais de ensino, é estimularem a

---

also Hoey, 2004, p. 386; and Hoey forthcoming). In other words, as we encounter words in spoken and written discourse and use them ourselves, we automatically pick up their usage patterns and learn in which language structures, textual positions, or text types they typically appear. In this process, existing primings can either be reinforced or weakened (see Hoey, 2005, p. 9). As a result, our knowledge about word is, according to Hoey, entirely dependent on our experiences with it (i.e. on how we have seen/heard it being used and how we have used it ourselves). (RÖMER, 2009, p. 145)

análise das porções de linguagem (chunks of language) como uma tentativa de 'facilitar o ensino'. Conforme as próprias palavras de Wray (1999), tais tentativas estarão "fomentando o uso de uma competência linguística de falantes nativos através da promoção de um comportamento de processamento não-nativo[135] (Wray, 1999, p. 463). Wray expõe seu argumento defendendo que as sequências formulaicas devem ser ensinadas, enfatizando-se a sua integridade, justamente por ser essa a marca da idiomaticidade. A autora explica seu ponto de vista:

> Nos programas de ensino, o resultado esperado ao introduzir as sequências formulaicas ao aluno é levá-lo(a) a aprender a usar a língua de uma maneira similar àquela utilizada pelo falante nativo, porém, a forma pela qual ele (a) é estimulado a fazer isso, é através do fracionamento e análise das partes que compõem a sequência, justamente o que os falantes nativos parecem não fazer, embora serem capazes disso. Para resumir, não é característico dos falantes nativos voluntariamente prestarem atenção na estrutura interna de uma sequência formulaica[136] [...].(Wray, 1999, p. 480).

Outros pesquisadores, porém, discordam e têm entendimento diferente sobre a questão relacionada à análise das sequências formulaicas, no contexto do ensino de L2. David Willis, conforme citado por Wray (1999, p. 463), acredita que no trabalho de sala de aula, algum tipo de análise é inevitável e que deve ser estimulado porque os ganhos no manejo da língua ultrapassam em muito as perdas em termos da aquisição de idiomaticidade semelhante àquela do falante nativo.

Wray (1999, p. 484) explica que a posição de Willis embasa-se na ideia de que a solução seria abandonar o pressuposto de que os alunos aprendem pelo mesmo processo vivenciado pelo falante nativo. Ao legitimar o interesse inerente dos alunos e sua habilidade em analisar a língua, Willis faz disso o princípio basilar para a organização de seu material de ensino. Ao invés de negar a artificialidade da sala de aula, ele aceita essa característica, embora dessa forma, e como resultado dessa

---

[135] 'pursuing native-like linguistic usage by promoting entirely unnative-like processing behaviour' (Wray, 1999, p. 463).

[136] In the teaching syllabuses, the intended outcome of presenting formulaic sequences to the learner is to make hin/her behave in a linguistically more native-like way, but the process by which this is encouraged to occur is the breakdown of the sequences into their consituent components, the very thing that native speakers appear no to do, even thou they are capable of it. It is, in short, not native-like voluntarily to activate awareness of the internal structure of a formulaic sequence[...].(Wray, 1999, p. 480).

escolha, segundo Wray, os alunos não possam obter resultados comparáveis aos de falantes nativos.

O'Keeffe, McCarthy e Carter (2007) apresentam argumentos de outros pesquisadores que defendem uma abordagem analítica no ensino das sequências formulaicas. Há, segundo Spöttl e McCarthy (2003), (apud O'KEEFFE, MCCARTHY e CARTER 2007), evidências psicolinguísticas indicando que, mesmo entre falantes nativos, pelo menos um mínimo de literalidade é mantido no processamento de expressões figurativas, sugerindo que mesmo as expressões linguísticas mais rígidas que, em geral, não sofrem alterações em seu uso, preservam algo do significado individual de seus itens, o qual é facilmente disponível para o usuário. "Os aprendizes talvez realmente se inclinem mais a analisar as porções de linguagem (chunks of language) do que os falantes nativos e, talvez, entendam isso como uma parte importante do processo de aprendizagem da língua. Talvez mesmo o domínio receptivo da língua seja resultado de uma abordagem analítica ocasional[137]" (O'KEEFFE, MCCARTHY E CARTER, 2007, p. 79-80)

Em resumo, neste trabalho considera-se relevante que os alunos entendam as sequências formulaicas e os pacotes lexicais em sua totalidade, evitando-se proceder a análise individual de cada um de seus constituintes. Em decorrência, preconiza-se a prática de ensino que destaca a característica holística dessas estruturas linguísticas. Lewis (1993, p. 121) indica ser partidário de ponto de vista semelhante ao afirmar que "os estudantes precisam estar cientes de que devem compreender o item lexical em sua totalidade, sem necessariamente analisá-lo gramaticalmente. Precisam ser treinados para fazê-lo de maneira acurada". Neste mesmo excerto, Lewis também chama a atenção para a função maior do material de ensino que é a de auxiliar os aprendizes a assimilar a função dos pacotes lexicais e das sequências formulaicas da língua. Ou seja, em sua perspectiva os "[...] estudantes precisam ser treinados para reconhecer a importância do invólucro – as

[137] Learners may be even more inclined to analyse chunks than native speakers, and may see it as an important part of the learning process. Receptive mastery may indeed gain from an occasional analytical approach. (O'KEEFFE, MCCARTHY E CARTER, 2007, p. 79-80)

palavras exatas utilizadas para expressar uma determinada função pragmática[138]„ (LEWIS, 1993, p. 121).

A questão do treino parece ser extremamente relevante para o aprendizado das sequências formulaicas, pois os anos de ensino com ênfase total na análise da língua de modo fragmentado e descontextualizado deixaram como herança, para várias gerações, um olhar segmentado sobre o funcionamento da língua. Assim, a intervenção do professor ou de outro mediador torna-se crucial, no sentido de estimular os alunos a pensarem a língua a partir de outro viés, ou seja, de seu aspecto holístico, auxiliando-os também a abandonarem a sobrecarga da herança de sua formação linguística tradicional.

No entanto, embora se tenha em mente, como proposta de ensino, enfatizar o aprendizado dos padrões linguísticos, não se descarta o ensino e aprendizado de itens isolados. Nesse sentido, respeita-se e entende-se a legitimidade do aprendiz de uma segunda língua querer entender o significado de uma palavra específica. Talvez a atitude ideal frente a indagações de alunos sobre o significado de termos presentes em expressões formulaicas e pacotes lexicais, seja orientá-los a observar como o termo é utilizado, quando isolado e quando dentro da sequência, comparando os dois usos. A intervenção do educador seria valiosa no sentido de ajudar o aluno a perceber as nuanças de significado das palavras, conforme elas se relacionam com outros termos.

Uma consulta ao *corpus* poderá dar respostas ao aluno. Ao verificar o termo nas linhas de concordâncias, ele poderá analisar, por si mesmo, qual o significado que aquela palavra adquire, quando utilizada, isoladamente, ou quando presente em diferentes sequências formulaicas e pacotes lexicais. A prática, a experiência e a exposição continuada ao idioma estudado, as intervenções do professor, a qualidade do material de ensino que destaca essas prioridades, o estímulo e a oportunidade de refletir sobre o uso da língua, mais o fator tempo, são fatores que, conjuntamente, auxiliarão o aprendiz a consolidar sua compreensão dessas sutilezas da língua.

---

[138] [...] With language material intended to present lexical phrases for learning, students must be trained to recognise the importance of the wrapper – the **precise** words used to express a particular pragmatic function. (Lewis, 1993, p. 121)

3.6 Aprendizagem Baseada em Tarefas

No contexto de ensino e aprendizagem de ESP, como também nesta proposta de ensino de inglês instrumental para a área de Tecnologia Ambiental “o que” ensinar é fundamental. De outra parte, já foi devidamente enfatizada a contribuição metodológica da Linguística de *Corpus*, no sentido de permitir a seleção de aspectos linguísticos que mereçam ser destacados em *corpora* de quaisquer dimensões. No entanto, a questão do “como” colocar em prática, na sala de aula, o ensino da língua de uma forma coerente com todos os preceitos até aqui destacados é outra questão. Em primeiro lugar, é preciso atentar para que a proposta esteja focada nas necessidades dos alunos e que os conteúdos e encaminhamentos didáticos vinculem-se estreitamente às práticas comunicativas utilizadas no mundo profissional dos acadêmicos. Enfim, é necessário formular “uma proposta coerente de preparação de material didático, a fim de que as atividades não sejam apenas exercícios soltos de manipulação da língua como um fim em si mesmo” (BERBER SARDINHA, 2006, p. 149). Presumivelmente, a Aprendizagem Baseada em Tarefas (Task Based Learning), conforme discutida a seguir, permite que as questões levantadas, tanto a partir do viés da Linguística de *Corpus* quanto da Linguística Cognitiva para o ensino da leitura de artigos acadêmicos em língua inglesa, sejam consideradas através da articulação dessas duas vertentes teóricas.

Embora existam pequenas variações na definição da “Aprendizagem Baseda em Tarefas”, todas elas “enfatizam o fato de que as tarefas pedagógicas envolvem uso da língua com propósitos comunicativos e que a atenção do usuário é focada no significado/sentido ao invés de focar nas formas gramaticais[139]” (NUNAN, 2004, p. 4). A construção do sentido, por parte do aluno, é um dos requisitos dessa abordagem, o que se coaduna plenamente com os objetivos e propósitos do estudo aqui desenvolvido. De fato, o significado é o ponto de partida para o desenvolvimento da língua desde o momento do planejamento de atividades. Ou

[139] While these definitions vary somewhat, they all emphasize the fact that pedagogical tasks involve communicative language use in which the user’s attention is focused on meaning rather than grammatical form. (NUNAN, 2004, p. 4)

seja, “criamos tarefas para facilitar atividades significativas na sala de aula [140]” (WILLIS e WILLIS, 2007, p. 9).

Nunan (2004, p. 1) afirma que a abordagem pedagógica “Aprendizagem Baseada em Tarefas”, independentemente das diferentes definições encontradas na bibliografia especializada, fortaleceu os seguintes princípios:

- Seleção de conteúdo baseada nas necessidades dos alunos;
- Ênfase na aprendizagem de habilidades comunicativas através de interação na língua alvo;
- Introdução de textos autênticos no contexto de aprendizagem;
- Criação de oportunidades para os aprendizes focarem não apenas na língua, mas também no próprio processo da aprendizagem;
- Valoração das experiências pessoais dos alunos como um importante elemento para o aprendizado;
- Ligação entre a língua empregada na sala de aula e a língua usada, na sociedade, pelos falantes nativos.

Essa listagem é uma síntese de diversos pontos teóricos discutidos ao longo desta dissertação, sejam eles advindos do referencial da Linguística Cognitiva ou da Linguísitica de *Corpus*, os quais convergem para uma proposta de trabalho e de ensino.

No entanto, por se tratar da criação de um programa de desenvolvimento exclusivo da habilidade da leitura de artigos acadêmicos, há um ponto listado acima que merece breve reflexão: “Ênfase na aprendizagem de habilidades comunicativas através de interação na língua alvo”. Todos os autores que aqui foram citados, no que se refere à “Aprendizagem Baseda em Tarefas”, direta ou indiretamente, destacam a importânica do desenvolvimento da habilidade comunicativa. No que diz respeito a um curso de ESP focado no desenvolvimento de uma única habilidade linguística receptiva – a leitura – este ponto pode até parecer deslocado ou incoerente; no entanto, não o está. Entende-se que a capacidade de entender e

[140] So we create tasks to facilitate meaningful activities in the classroom. (WILLIS e WILLIS, 2007, p. 9).

interpretar o texto, ou seja, a construção ou reconstrução do sentido do texto, por parte do aluno, como a habilidade comunicativa que diz respeito à leitura. No momento em que o aprendiz de ESP conseguir operacionalizar a linguagem utilizada pelo autor do texto e conseguir dela apropriar-se; entender as proposições apresentadas no texto versado na língua alvo para condução de seus estudos e pesquisas, para aprofundar seu conhecimento sobre algum aspecto teórico; e, além disso, conseguir assumir uma posição frente ao enunciado do texto (a favor, contra, neutra, etc.) estará assim manifestando a habilidade comunicativa pertinente à leitura.

Dentre as diversas definições encontradas para “Aprendizagem Baseada em Tarefas” (Willis, 1996; Ellis (2003); Nunan (2004); Willis e Willis (2007), entre outras, a definição de Estaire e Zanon (1994, p. 14 apud BERBER SARDINHA, 2006, p. 150) será aqui adotada, pois é condizente com os propósitos e com o referencial teórico deste estudo. Segundo Estaire e Zanon (1994), uma tarefa comunicativa é:

> uma atividade de trabalho de sala de aula durante a qual a atenção dos aprendizes é focada no significado e não na forma, isto é, no que está sendo expressado ao invés do que nas formas linguísticas utilizadas para expressá-lo. (...) É uma atividade de sala de aula na qual, até onde for possível, assemelha-se com atividades que os estudantes ou outras pessoas desempenham em suas vidas cotidianas, reproduzindo, em vista disso, processos da comunicação do dia a dia (Estaire e Zanon, 1994, p. 14 apud BERBER SARDINHA, 2006, p. 150).

A aplicação de uma tarefa no contexto de ensino deve pelo menos dizer respeito a práticas sociais utilizadas pelos membros da comunidade envolvida. A leitura é uma atividade essencialmente cognitiva e metacognitiva, pois leva à realização de outros fazeres e à produção de outros saberes, tal como acontece de forma muito acentuada no meio acadêmico. No entanto, atividades cognitivas, não são necessariamente “tarefas”, embora façam parte delas.

Berber Sardinha (2006, p. 149) comenta que uma das vertentes da Aprendizagem Baseada em Tarefas entende “tarefa como uma prática social, uma atividade que desempenhamos genuinamente como parte de nossos afazeres cotidianos, de modo profissional ou não.” No contexto de ensino e ESP para alunos de Tecnologia Ambiental, uma pergunta que pede resposta é: de que forma a prática

da leitura seria uma tarefa? Talvez a resposta mais óbvia, seja a mais vaga: ler e entender os textos científicos. No entanto, essa resposta não basta. Berber Sardinha (2006, p. 149), nesse mesmo artigo, acrescenta que as "tarefas se concretizam discursivamente, de modo mais ou menos rotinizado, e recebem nomes socialmente aceitos". De fato, no meio acadêmico, a leitura dos artigos científicos é uma prática social muito específica da comunidade estudantil e científica, mas é uma prática que se vincula à realização de outras funções e atividades acadêmicas; tem ao mesmo tempo um fim em si mesma, porém também desencadeia muitas outras ações.

No meio acadêmico (e fora da academia), a leitura é condutora, ou é base para o desenvolvimento de tarefas bastante específicas, sejam elas de cunho prático, ou de cunho cognitivo, ou as duas, e geralmente são interdependentes. As tarefas aqui referidas como sendo de cunho prático "com nomes socialmente aceitos", ou seja, os gêneros textuais, são facilmente reconhecidas e resultam em algum produto ou prática cultural (abstract, apresentação, lista, resumo, tabela, etc.) reconhecida pela sociedade. Berber Sardinha (2006, p. 149) afirma que as tarefas são "do ponto de vista linguístico-discursivo gêneros: eventos comunicativos institucionalizados em determinados grupos sociais". Por outro lado, as atividades ditas cognitivas são inerentes às atividades de produção e sem elas nenhum produto cultural ou tarefa poderia ser executado, sendo essencialmente de ordem mental. Mesmo assim, são também produtos de uma cultura, "tecnologias intelectuais", que de alguma forma foram aprendidas.

Durante o ato da leitura, diversas operações mentais e cognitivas são realizadas pelo leitor. Algumas das atividades cognitivas presentes no ato de ler podem ser elencadas: compreender, classificar, predizer, induzir, inferir, discriminar, sumarizar, resumir, sintetizar, avaliar, ponderar, comparar, concordar, discordar, refutar, localizar uma informação específica, fazer uma leitura rápida (passar os olhos pelo texto), relacionar informações, interligar textos e contextos, localizar palavras-chave e "tirar uma ideia do texto", deduzir significados pelo contexto ou cotexto, estabelecer relações entre conceitos, relacionar parágrafos e sentenças, prever, predizer, etc.. Todas essas atividades, entre muitas outras, não são necessariamente "produtos culturais" tal como um texto impresso, um artigo

científico ou um gráfico, mas atividades que o leitor, em algum momento ou outro da leitura, realiza ou precisa realizar. No entanto, no que diz respeito à leitura, todas essas atividades mentais são produzidas na interação com o código linguístico, ou seja, através da língua. O domínio dessas atividades cognitivas presentes e necessárias na leitura levam à realização de muitas outras atividades práticas, de cunho acadêmico (ou não) e permitem a produção de diversos artefatos culturais. Assim, a resposta à pergunta acima feita, é que a leitura será considerada uma tarefa no contexto de ESP- conforme o conceito de tarefa aqui proposto - quando possibilitar a produção ou recepção de artefatos culturais acadêmicos, ou seja, dos gêneros do discurso acadêmico.

Para citar algumas tarefas de cunho prático resultantes da leitura e pertinentes ao meio acadêmico podem-se citar: escrever uma síntese, organizar um sumário, escrever uma resenha, escrever um *abstract*, produzir um esquema visual, elaborar uma apresentação, esquematizar uma fórmula a partir de uma descrição, traduzir um extrato de texto, citar uma parte da leitura, parafrasear um conceito do autor, realizar o fichamento de um artigo, referenciar um autor, referenciar obras lidas, localizar um artigo tratando de um tema X, localizar artigos de autores relacionados ao tema Y, organizar uma coleção de artigos científicos por tema ou assunto, organizar os dados de um artigo em uma tabela, selecionar artigos a serem utilizados em uma pesquisa; a lista poderia se estender por vários parágrafos. O que aqui se afirma, incisivamente, é que sem a realização de uma leitura na qual o leitor se aproprie, com profundidade, do conhecimento veículado no texto lido, a partir da mobilização de uma vasta gama de movimentos cognitivos, provavelmente, nenhuma dessas atividades poderia ser plenamente realizada. Enfim, por ser a leitura uma atividade essencialmente cognitiva (e metacognitiva) e levar à realização de outros fazeres e à produção de outros saberes, tal como acontece no meio acadêmico, o trabalho de sala de aula precisa contemplar tanto o desenvolvimento das habilidades cognitivas presentes na leitura, como a realização de tarefas que façam parte da vida acadêmica originadas ou atreladas à leitura.

A consolidação da proposta de “Aprendizagem Baseada em Tarefas” não elimina nem descarta a necessidade de atividades de ensino que digam respeito ao funcionamento da língua, como o entendimento da gramática, da lexicogramática,

das funções, etc. No entanto, o foco é diferenciado, pois estudar os padrões linguísticos diz respeito a entender como a língua funciona para produzir sentido e como se relaciona à função de sua lexicogramática dentro do discurso, dentro de um gênero textual. Assim, evitam-se práticas tais como "aprender o *past continuous*" e produzir uma série de exercícios para fixar a sua estrutura; ou realizar exercícios para praticar o comparativo, ou exercícios para preencher os espaços (fill the gaps) com os *phrasal verbs* que estão faltando. Estaire e Zanon (1994, apud BERBER SARDINHA, 2006, p. 150) definem tarefas de apoio (enabling tasks) como aquelas atividades que possibilitam ao aprendiz conhecer e refletir sobre os meandros da língua. Desse modo, as tarefas de apoio

> são como um suporte para as tarefas comunicativas. Seu propósito é fornecer aos estudantes o suporte linguístico necessário para desempenhar uma atividade comunicativa. Embora elas possam relacionar-se à produção de sentido, seu principal foco diz respeito aos aspectos linguísticos (gramática, vocabulário, pronúncia, funções, discurso) mais do que ao significado. Elas são de forma muito clara experiências que produzem sentido, cujo objetivo é possibilitar que os alunos se comuniquem da forma mais eficaz e harmoniosa possível (Estaire e Zanon, 1994, p. 15 apud BERBER SARDINHA, 2006, p. 150).

Logo, as tarefas de apoio também podem fazer parte das práticas de ensino desta proposta, desde que associadas ao discurso, a um dado texto, evitando-se exercícios de fixação, esvaziados de propósito. O entendimento do modo de funcionamento da linguagem no texto e a necessidade de abstrair o seu sentido serão a mola propulsora para aplicar tarefas de apoio e não o oposto.

Então, com base nos pressupostos teóricos discutidos neste capítulo serão apresentados nos capítulos a seguir, o plano de realização de uma tarefa pedagógica, para o ensino de leitura para acadêmicos do curso de Tecnologia Ambiental, utilizando-se os dados coletados no *corpus* de Tecnologia Ambiental, e a aplicação do conhecimento teórico desenvolvido ao longo dos três capítulos teóricos. Busca-se, também, a partir da proposta de Estaire e Zanon (1994, apud BERBER SARDINHA, 2006, p. 150), sugerir um modelo para orientação e produção de uma unidade didática.

4 METODOLOGIA DE PESQUISA DA ELABORAÇÃO DO *CORPUS* DE TECNOLOGIA AMBIENTAL

Este capítulo descreve a metodologia de investigação utilizada no desenvolvimento do *corpus* de estudo, sua análise e obtenção de dados relacionados às perguntas desta pesquisa. Também é apresentado a metodologia e procedimentos utilizados para a obtenção do texto-chave do *corpus*.

4.1 *Corpus* e metodologia

O processo de elaboração do *corpus* teve início com a definição de critérios para a seleção dos textos do *corpus* de estudo a ser constituído. A utilização de textos autênticos, isto é, não adaptados, respeitando um dos pressupostos da Linguística de *Corpus*, foi o primeiro critério a ser definido. Em seguida, propôs-se a seleção de textos utilizados ao longo do curso, para compor a amostra textual constitutiva do *corpus* do mestrado em Tecnologia Ambiental da UNISC. Tais opções derivaram-se da intenção de difundir e aplicar, posteriormente, o conhecimento produzido nesta investigação, seja no ensino de inglês instrumental para alunos vinculados ao curso em questão (possivelmente, também, em cursos de áreas afins), seja na socialização do *corpus* linguístico aqui compilado e apresentado para a comunidade acadêmica, de modo geral.

Após contatos com alguns professores e com a coordenação do curso de PPGTA (Programa de Pós-graduação em Tecnologia Ambiental – mestrado) da UNISC (Universidade de Santa Cruz do Sul/RS), visando a conhecer as necessidades dos alunos, em relação ao uso e aprendizado da língua inglesa, constatou-se que evidenciavam dificuldades para ler artigos da área de Tecnologia Ambiental versados em língua inglesa. Decidiu-se, então, que o *corpus* de textos da área de Tecnologia Ambiental seria formado exclusivamente por artigos acadêmicos, selecionados pelos professores que se voluntariaram a auxiliar, cooperando com o andamento da pesquisa. O critério fornecido aos professores para melhor se

orientarem em relação à seleção dos artigos foi o de que os textos indicados estivessem escritos em língua inglesa e que fizessem parte do currículo de suas disciplinas, ou que pudessem vir a ser lidos pelos alunos ao longo do curso, por se constituírem em referência complementar ao programa das disciplinas.

Os professores titulares das disciplinas do programa do PPGTA abaixo listadas, foram responsáveis pela seleção dos artigos científicos que compõem o *corpus* de estudo. Considerando-se a possibilidade de estudos futuros, o presente *corpus* foi subdvidido em quatro *subcorpora*, compostos por artigos pertinentes a cada uma das seguintes disciplinas:

- Tratamento e Reciclagem de Materiais
- Recuperação de Áreas Degradadas
- Gestão e Tecnologia Ambiental
- Controle e Poluição Atmosférica

Os artigos indicados pelos professores foram entregues em formato digital, tanto arquivos em PDF como arquivos em formato do Word, isto é, formato DOC. Para a construção do *corpus* e posterior análise pelo programa Wordsmith Tools 5.0, entretanto, todos os arquivos textuais precisam estar no formato TXT. A conversão de formatos foi realizada com a utilização de um software de OCR (Optical Caracter Recognition), que permite que arquivos textuais gerados como imagem – tal como o PDF –, bem como outros formatos de arquivos (como arquivos DOC do Word; JPG, de imagens) sejam convertidos em um formato textual final, escolhido pelo usuário. O software OCR utilizado para o processo de conversão dos arquivos em formato TXT foi o Abbyy Fine Reader 9.0. Referências bibliográficas, dados de identificação dos autores e instituição foram deletados, evitando assim, que esses termos fossem contabilizados pelo software. Gráficos e imagens foram automaticamente eliminados na conversão para o formato TXT. No entanto, optou-se pela permanência de legendas de imagens ou gráficos por entender que essas apresentam linguagem significativa para fazer parte do *corpus*.

As palavras de um *corpus*, quando analisadas e quantificadas pelo software, são contabilizadas de duas maneiras. Assim, os termos *type* e *tokens* são denominações utilizadas para referir-se às palavras de um texto, indicando diferenças no procedimento de contagem dos itens do *corpus*. O termo *type* idenfica

cada palavra do *corpus* analisado, não contabiliza repetições, apenas unidades diferentes; ou seja, a quantidade de *types* de um texto pode ser indicador de sua riqueza/pobreza de vocabulário. O termo *tokens* indica a quantidade de palavras totais, incluindo todas repetições de cada termo.

Com os textos convertidos e aptos a serem analisados pelo Wordsmith Tools, realizou-se a primeira análise quantitativa do *corpus*. Houve disparidade muito grande entre a quantidade de artigos (e consequente quantidade de *tokens*) de cada uma das disciplinas que compunham o *corpus* de estudo. Foi necessário, então, realizar um rebalanceamento do *corpus*, a partir da inclusão ou exclusão de alguns textos, para que o *corpus* de cada disciplina apresentasse, de forma equilibrada, tamanho e quantidade semelhantes. Para o rebalanceamento do *corpus*, estabeleceu-se como parâmetro a quantidade de aproximadamente 110.000 palavras para cada um dos sub*corpus*. O *corpus* final de Tecnologia Ambiental apresentou a distribuição final, conforme consta na tabela 2, logo abaixo.

Tabela 2 - Dados do *corpus* de Tecnologia Ambiental

Disciplinas Itens analisados	Tratamento e Reciclagem de Materiais	Recuperação de áreas degradadas	Gestão e Tecnologia Ambiental	Controle da poluição atmosférica	*Corpus* de Tecnologia Ambiental (total)
Quantidade de artigos	19	24	15	28	86
file size	713.135	672.109	792.746	702.100	2.880.090
tokens (running words) in text	112.316	107.485	117.996	112.768	450.565
tokens used for word list	104.436	98.892	112.376	103.779	419.483
types (distinct words)	8.120	6.604	7.920	7.625	17.452
type/token ratio (TTR)	7.78	6.68	7.05	7.35	4.16

4.1.1 Análise dos dados do *corpus* de Tecnologia Ambiental

Pela análise da Tabela 2 pode-se ver que o equilíbrio total do *corpus*, tendo como critério a quantidade de palavras/*tokens* em cada uma das disciplinas, foi alcançado, mesmo havendo diferença de até 8500 *tokens* entre cada uma delas. O *corpus* total - *Corpus* de Tecnologia Ambiental - composto pela análise global de todos os textos provenientes das quatro disciplinas acima listadas, alcançou o total de 450.565 palavras/tokens. Conforme é de conhecimento dos pesquisadores da área, não há um critério rígido que estabeleça o limite preciso entre um valor/quantidade que seja parâmetro para classificar um *corpus* como pequeno *(small corpus*). Inclusive, Gavioli alerta para a questão (2005, p. 7) dizendo que "as categorias 'pequeno' e 'grande' não são precisas e provavelmente, no presente, enganosas[141]". Como refere essa autora, surgiram discordâncias entre os investigadores e novos questionamentos acerca da utilização das definições *corpus especializado (specialized corpus)* e *corpus geral (general corpus):*

> *Corpora* especializados são *corpora* projetados com o propósito de apresentar uma amostra de linguagem especializada, tanto por coletar textos de conteúdo similar (por exemplo, física, medicina, negócios, filosofia) ou de gêneros textuais similares (por exemplo, artigos científicos, artigos de divulgação científica, capítulos de livros) ou ambos (por exemplo, artigos científicos da medicina ou palestras da área de ciências), ou até mesmo entre textos provenientes de outras categorias, tais como de linguagem jornalística ou acadêmica. "*Corpora* gerais" são, geralmente, os *corpora* que têm como propósito serem representativos de uma língua como um todo (por exemplo, a língua inglesa) ou uma variedade linguística regional (por exemplo, inglês britânico)[142] (GAVIOLI, 2005, p. 7).

A partir da classificação fornecida por Gavioli, o *corpus* aqui apresentado pode ser enquadrado como *corpus* especializado, tanto por ser constituído por textos

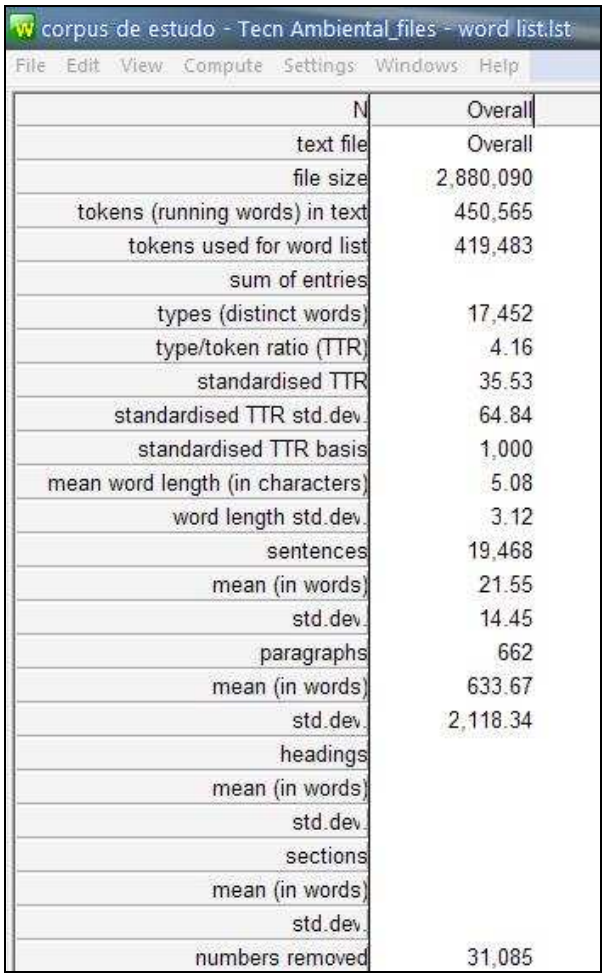
[141] [...]the 'small' and 'large' categories are very fuzzy ones and probably even misleading nowadays. (GAVIOLI, 2005, p. 7)

[142] "Specialized" *corpora* are *corpora* designed for the purpose of creating a sample of specialized language either by collecting texts of similar content (e.g. science, medicine, business, philosophy) or of similar text-type or genre (e.g. research papers, letters, book chapters) or both (e.g. medical research articles or science lectures), or even texts from other types of specialized categories, such as newspaper language or academic language. "General" *corpora* are normally *corpora* designed with the aim of representing a "whole" language (e.g. English) or a geographical variety (e.g. British English). (GAVIOLI, 2005, p. 7)

provenientes de uma área especializada (Tecnologia Ambiental), como por apresentar uma seleção de textos de um único gênero acadêmico (artigos científicos). Logo, o *corpus* de Tecnologia Ambiental é um pequeno *corpus* em razão de sua dimensão e, ainda, um *corpus* especializado em razão da dupla natureza dos textos que o compõem.

Mesmo tendo realizado a partição do *corpus*, para o balanceamento (e possível aproveitamento em outros estudos ou pesquisas), a análise realizada neste estudo referir-se-á sempre à totalidade do *corpus*. A figura 8 mostra os dados estatísticos coletados com a utilização do software Wordsmith Tools (WST, doravante), a partir da análise de todos os arquivos (86 arquivos, cada um referente a um artigo acadêmico) que compõem o *corpus* de Tecnologia Ambiental. Esta imagem é um fragmento capturado da tela do software WST, em que foram selecionados os dados referentes à totalidade do *corpus*.

Figura 8 - Dados estatísticos do *corpus* de Tecnologia Ambiental



N	Overall
text file	Overall
file size	2,880,090
tokens (running words) in text	450,565
tokens used for word list	419,483
sum of entries	
types (distinct words)	17,452
type/token ratio (TTR)	4.16
standardised TTR	35.53
standardised TTR std.dev.	64.84
standardised TTR basis	1,000
mean word length (in characters)	5.08
word length std.dev.	3.12
sentences	19,468
mean (in words)	21.55
std.dev.	14.45
paragraphs	662
mean (in words)	633.67
std.dev.	2,118.34
headings	
mean (in words)	
std.dev.	
sections	
mean (in words)	
std.dev.	
numbers removed	31,085

A partir da análise dos dados estatíticos do *corpus* de Tecnologia Ambiental, presente na figura 8, as seguintes informações do *corpus* são detectáveis, entre elas:

Tamanho do arquivo (file size): 2.880.090

Total de palavras do texto(tokens/running words): 450.565

Palavras usadas para a elaboração da lista (tokens used for list): 419.483

Palavras distintas (types, distinct words): 17.452

Abaixo segue a tabela 2 apresentando as 200 primeiras palavras do *corpus* de Tecnologia Ambiental.

Tabela 3 – 200 palavras mais frequentes do *Corpus* de Tecnologia Ambiental

Lista de palavras do *Corpus* deTECNOLOGIA AMBIENTAL		
N	Word	Freq.
1	#	31,085
2	THE	29,387
3	OF	17,115
4	AND	13,767
5	IN	10,75
6	TO	9,02
7	A	6,555
8	FOR	5,024
9	IS	4,444
10	WAS	3,391
11	THAT	3,242
12	WITH	3,099
13	AS	2,857
14	BY	2,748
15	ARE	2,716
16	FROM	2,446
17	BE	2,366
18	WERE	2,360
19	ON	2,203
20	THIS	2,140
21	ENVIRONMENTAL	2,099
22	AT	2,030
23	AN	1,609
24	OR	1,478
25	SUB	1,335
26	IT	1,333
27	NOT	1,252
28	ET	1,150
29	AL	1,145
30	WHICH	1,135
31	USED	1,086
32	WASTE	1,046
33	SOIL	1,045
34	CAN	1,012
35	C	976
36	HAVE	926
37	HAS	880
38	ALSO	858
39	NO	846
40	THESE	843
41	THAN	818
42	FIG	787
43	PROCESS	783
44	EMISSIONS	782
45	N	778
46	AIR	776
47	BEEN	765
48	OTHER	753
49	G	720
50	TABLE	697
51	CONCENTRATION	691
52	TIME	691
53	ALL	685
54	MORE	650
55	WATER	650
56	RESULTS	642
57	SYSTEM	639
58	DURING	636

59	HIGH	635
60	STUDY	615
61	AUDIT	611
62	MANAGEMENT	611
63	SUCH	608
64	USING	606
65	MG	603
66	INTO	582
67	PH	572
68	BETWEEN	552
69	CONTROL	552
70	TEMPERATURE	536
71	EACH	534
72	THEIR	534
73	USE	529
74	TOTAL	523
75	WHEN	521
76	TWO	520
77	MODEL	513
78	REMOVAL	509
79	B	507
80	HOWEVER	502
81	RATE	500
82	MATERIALS	499
83	CONCENTRATIONS	497
84	CONDITIONS	497
85	DIFFERENT	496
86	RECYCLING	496
87	MAY	495
88	THERE	489
89	UNDER	489
90	INCREASE	488
91	ONE	476
92	WILL	475
93	L	472
94	PRODUCTION	469
95	BOTH	468
96	E	465
97	DATA	464
98	ANALYSIS	463
99	ITS	463
100	POLLUTION	459
101	AFTER	457
102	BUT	457
103	LEVELS	450
104	AUDITING	449
105	ONLY	448
106	I	447
107	OIL	446
108	ENERGY	445
109	MATERIAL	444
110	DO	441
111	LOW	427
112	BECAUSE	416
113	MOST	416
114	S	412
115	THEY	404
116	GAS	402
117	M	398
118	TREATMENT	398
119	WE	394
120	H	392

121	IF	390
122	ENVIRONMENT	385
123	LEVEL	385
124	BASED	384
125	HIGHER	382
126	EFFICIENCY	378
127	PERIOD	378
128	DUE	375
129	P	375
130	VALUES	375
131	PET	373
132	CONTENT	368
133	COULD	368
134	SOME	368
135	PARTICLE	364
136	CO	363
137	THROUGH	358
138	X	358
139	CARBON	355
140	ORGANIC	347
141	ABOUT	343
142	NUMBER	343
143	OXYGEN	343
144	PHASE	343
145	FOUND	342
146	OBSERVED	339
147	OZONE	336
148	WHERE	333
149	REDUCTION	329
150	INCREASED	328
151	WOULD	327
152	PRODUCTS	326
153	VALUE	321
154	FIRST	320
155	CHEMICAL	309
156	QUALITY	309
157	AUDITS	305
158	THREE	303
159	AREA	302
160	SOILS	302
161	SUP	302
162	OBTAINED	295
163	SHOWN	295
164	PLANT	292
165	SOLID	291
166	EMISSION	290
167	ADDITION	288
168	ORDER	288
169	OUT	288
170	EXPOSURE	287
171	ACID	284
172	THEREFORE	283
173	SHOULD	282
174	SOURCE	281
175	SYSTEMS	281
176	ALPHA	280
177	D	276
178	POTENTIAL	275
179	WELL	274
180	COMPANIES	273
181	KG	273
182	AVERAGE	272

183	METHODS	272
184	COST	271
185	STUDIES	271
186	CONTAMINATED	270
187	PAPER	269
188	RESEARCH	269
189	HAD	266
190	PLASTIC	266
191	DENITRIFICATION	264
192	INDUSTRY	263
193	INFORMATION	263
194	OVER	263
195	RESPECTIVELY	263
196	LOWER	262
197	REACTION	261
198	EFFECT	260
199	IMPACT	260
200	YEAR	259

4.1.2 As palavras mais frequentes do *Corpus* de Tecnologia Ambiental

A tabela 3 apresenta as 200 palavras mais frequentes do *Corpus* de Tecnologia Ambiental. Nessa listagem que informa apenas a quantidade dos 200 primeiros termos do *corpus*, já é evidente a presença significativa de vocábulos específicos da área de Tecnologia Ambiental (destacados com sombreado cinza), bem como símbolos de notação científica (N, C, L, G, entre outros) e alguns vocábulos recorrentes na produção escrita acadêmica. Tal presença ultrapassa 50% dos 200 primeiros termos, tendo totalizado 105 termos atinentes à área conceitual do *corpus*. A tendência é aumentar a proporção de termos específicos à medida que mais itens, de acordo com a ordem de frequência, são apresentados.

O primeiro item da lista, representado pelo sinal "#", indica a quantidade de algarismos utilizados dentro do *corpus*. O software contabiliza todas as recorrências numéricas e as apresenta como um único termo da listagem. A alta incidência de algarismos no *corpus* de Tecnologia Ambiental pode ser interpretada como uma das marcas dos textos acadêmicos, principalmente das áreas técnicas, por estes corriqueiramente apresentarem diversos dados comparativos e estatísticos, fórmulas científicas, resultados de estudos quantitativos, entre outras notações numéricas pertinentes às áreas técnicas.

BNC – British National *Corpus*

Ao compararmos a lista de palavras do *corpus* de Tecnologia Ambiental com as 200 primeiras palavras do *corpus* BNC (British National *Corpus*, um *corpus* geral da língua inglesa) é visível a não existência de terminologia específica de nenhuma área acadêmica ou de algum campo conceitual, havendo alta predominância e recorrência de palavras funcionais da língua inglesa. Entre as 200 primeiras palavras do *corpus* BNC apresentadas na tabela 4, escassas palavras de significado são encontradas, tendência oposta a ocorrida no *corpus* de Tecnologia Ambiental.

Tabela 4 – 200 palavras mais frequentes do *Corpus* BNC (British National*Corpus*)

Lista de palavras do *Corpus* BNC	
N	Word
1	THE
2	OF
3	AND
4	TO
5	A
6	IN
7	#
8	THAT
9	IS
10	IT
11	FOR
12	WAS
13	I
14	ON
15	WITH
16	AS
17	BE
18	HE
19	YOU
20	AT
21	BY
22	ARE
23	THIS
24	HAVE
25	BUT
26	NOT
27	FROM
28	HAD
29	HIS
30	THEY
31	OR
32	WHICH
33	AN
34	SHE
35	WERE
36	HER
37	WE
38	ONE
39	THERE
40	ALL
41	BEEN
42	THEIR
43	IF
44	HAS
45	WILL
46	SO
47	WOULD
48	NO
49	WHAT
50	CAN
51	WHEN
52	MORE
53	UP
54	OUT
55	SAID
56	WHO
57	ABOUT
58	DO
59	SOME
60	THEM
61	ITS
62	INTO
63	THEN
64	TIME
65	HIM
66	OTHER

67	TWO
68	ONLY
69	LIKE
70	MY
71	THAN
72	COULD
73	WELL
74	QUOT
75	NOW
76	YOUR
77	ME
78	OVER
79	MAY
80	IT'S
81	JUST
82	NEW
83	THESE
84	ALSO
85	ANY
86	FIRST
87	VERY
88	KNOW
89	PEOPLE
90	SEE
91	AFTER
92	SUCH
93	SHOULD
94	WHERE
95	BECAUSE
96	MOST
97	HOW
98	BACK
99	GET
100	WAY
101	DON'T
102	OUR
103	DOWN
104	DID
105	MADE
106	RIGHT
107	BETWEEN
108	ER
109	MUCH
110	WORK
111	YEARS
112	THINK
113	MANY
114	BEING
115	EVEN
116	GO
117	THOSE
118	GOT
119	BEFORE
120	YEAH
121	THROUGH
122	GOOD
123	US
124	THREE
125	POUND
126	MAKE
127	HELLIP
128	LAST

129	STILL
130	TAKE
131	MUST
132	OWN
133	YEAR
134	OFF
135	BOTH
136	SAY
137	TOO
138	OH
139	COME
140	THAT'S
141	HERE
142	MR
143	USED
144	GOING
145	ERM
146	LITTLE
147	I'M
148	USE
149	SAME
150	UNDER
151	EACH
152	HOWEVER
153	MIGHT
154	DAY
155	ANOTHER
156	YES
157	PUT
158	AGAIN
159	GOVERNMENT
160	LONG
161	AGAINST
162	WANT
163	MAN
164	NEED
165	WHILE
166	LIFE
167	WORLD
168	THOUGHT
169	PER
170	PART
171	NEVER
172	OLD
173	LOOK
174	HOME
175	DOES
176	SOMETHING
177	HOUSE
178	COURSE
179	SINCE
180	NUMBER
181	END
182	WHY
183	AWAY
184	PLACE
185	DIFFERENT
186	FOUND
187	GREAT
188	REALLY
189	LOCAL

190	WENT
191	WITHIN
192	FOUR
193	CASE
194	LEFT
195	NEXT
196	CAME
197	WITHOUT
198	ALWAYS
199	SET
200	SYSTEM

4.1.3 Palavras-chave do *Corpus* de Tecnologia Ambiental

As palavras-chave (keywords) do *corpus* de Tecnologia Ambiental foram selecionadas com a utilização da ferramenta Keywords. O software para elaborar a listagem de palavras-chave necessita de uma listagem de palavras proveniente de um *corpus* de referência, em vista disso foi utilizado o *corpus* BNC (British National Corpus) como referência e, como resultado, foram encontradas 3.040 palavras-chave.

A tabela 5 apresenta a listagem das 100 primeiras palavras-chave, as mais frequentes do total de 3.040 palavras-chave do *corpus* de Tecnologia Ambiental. Pela análise da lista é possível, mesmo através de uma rápida leitura, perceber o alto grau de relevância destes termos para a área de Tecnologia Ambiental. Com isso, torna-se evidente a eficácia da ferramenta *Keywords* em detectar os termos significativos para um determinado campo teórico.

Tabela 5 – Lista parcial de Palavras-chave do *Corpus* de Tecnologia Ambiental

Lista parcial de Palavras-chave do *Corpus* de Tecnologia Ambiental		
N	**Keyword**	**Freq.**
1	#	31,085
2	OF	17,115
3	AND	13,767
4	IN	10,750
5	WERE	2,360
6	ENVIRONMENTAL	2,099
7	SUB	1,335
8	ET	1,150
9	AL	1,145
10	USED	1,086
11	WASTE	1,046
12	SOIL	1,045
13	C	976
14	FIG	787
15	PROCESS	783
16	EMISSIONS	782
17	N	778
18	AIR	776
19	G	720
20	TABLE	697
21	CONCENTRATION	691
22	WATER	650
23	RESULTS	642
24	SYSTEM	639
25	DURING	636
26	HIGH	635
27	STUDY	615

28	AUDIT	611
29	MANAGEMENT	611
30	USING	606
31	MG	603
32	PH	572
33	CONTROL	552
34	TEMPERATURE	536
35	TOTAL	523
36	MODEL	513
37	REMOVAL	509
38	B	507
39	RATE	500
40	MATERIALS	499
41	CONCENTRATIONS	497
42	CONDITIONS	497
43	RECYCLING	496
44	DIFFERENT	496
45	INCREASE	488
46	L	472
47	PRODUCTION	469
48	E	465
49	DATA	464
50	ANALYSIS	463
51	POLLUTION	459
52	LEVELS	450
53	AUDITING	449
54	OIL	446
55	ENERGY	445
56	MATERIAL	444
57	LOW	427
58	S	412
59	GAS	402
60	TREATMENT	398
61	M	398
62	H	392
63	ENVIRONMENT	385
64	LEVEL	385

65	BASED	384
66	HIGHER	382
67	EFFICIENCY	378
68	PERIOD	378
69	VALUES	375
70	DUE	375
71	PET	373
72	CONTENT	368
73	PARTICLE	364
74	CO	363
75	X	358
76	CARBON	355
77	ORGANIC	347
78	OXYGEN	343
79	PHASE	343
80	OBSERVED	339
81	OZONE	336
82	REDUCTION	329
83	INCREASED	328
84	PRODUCTS	326
85	VALUE	321
86	CHEMICAL	309
87	QUALITY	309
88	AUDITS	305
89	SUP	302
90	SOILS	302
91	OBTAINED	295
92	SHOWN	295
93	PLANT	292
94	SOLID	291
95	EMISSION	290
96	ADDITION	288
97	EXPOSURE	287
98	ACID	284
99	SOURCE	281
100	SYSTEMS	281

4.1.4 Clusters e pacotes Lexicais no *corpus* de Tecnologia Ambiental

Um dos propósitos deste estudo foi detectar, identificar e quantificar os pacotes lexicais (*bundles/clusters*) presentes no *corpus* de Tecnologia Ambiental, verificando sua frequência e relevância para o ensino. A ferramenta Wordlist foi utilizada para a extração dos clusters.

A extração dos pacotes lexicais demanda que o linguista estabeleça alguns parâmetros de antemão, definindo a quantidade mínima e máxima de palavras de cada *bundle* e a quantidade mínima de repetições no *corpus*. O Wordlist permite que sejam detectados *clusters* de tamanhos variados, contendo de 2 a 8 palavras. Após um primeiro experimento verificou-se que havia quantidade significativa de

clusters com mais de quatro palavras contendo termos específicos da área de Tecnologia Ambiental e, assim, optou-se por extrair todos os clusters contendo de 3 a 8 palavras.

Douglas Biber apresenta uma metodologia específica para a extração de bundles de um *corpus*. Segundo Biber (BIBER et al., 1999):

> Para ser qualificada como um pacote lexical, uma combinação de palavras precisa necessariamente recorrer com frequência em um dado registro. Nos seguintes resultados, as sequências lexicais são contadas como pacotes lexicais 'recorrentes' somente se ocorrerem pelo menos dez vezes por milhão de palavras em um registro. Essas ocorrências precisam estar distribuídas em pelo menos cinco textos diferentes (para excluir idiossincrasias individuais do falante/escritor). Pelo fato de pacotes lexicais que contenham cinco ou seis palavras serem geralmente pouco comuns, um ponto de corte mais baixo de pelo menos cinco vezes por milhão de palavras é usado para esses tipos[143]. (BIBER et al., 1999, p. 992-993)

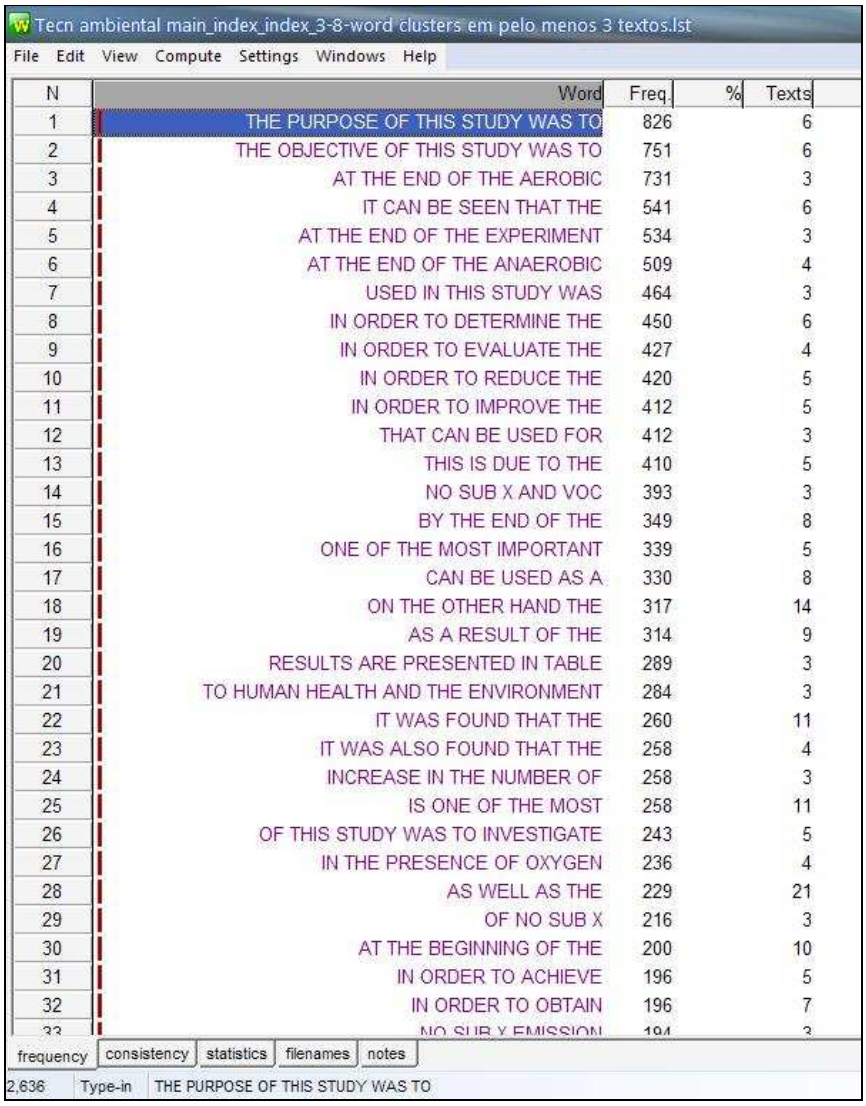
Em razão da dimensão do *corpus* criado para este estudo ter proporções e propósitos diferenciados - 450.000 palavras e textos muito específicos de uma determinada área acadêmica – foram necessárias algumas adaptações aos parâmetros estabelecidos por Biber. Por ser um *corpus* de dimensões pequenas optou-se por selecionar os *bundles* repetidos pelo menos 5 vezes e distribuídos em pelo menos três textos diferentes. A escolha desses parâmetros deu-se com o receio de omitir, na extração, *bundles* representativos e significativos, embora não apresentando um alto índice de recorrência no *corpus*. Por outro lado, percebeu-se, em uma primeira análise dos dados, que havia alta recorrência de determinados bundles em poucos textos (em apenas um, dois ou três textos) e, assim, o ajuste do ponto de corte, de acordo com o formato do *corpus* aqui apresentado e aliado às questões desta pesquisa, foi adaptado. Adaptação semelhante foi realizada, também, por Biber (1999), conforme citação anterior, em razão da extração dos pacotes lexicais de cinco ou seis palavras não serem tão recorrentes no *corpus* por ele analisado. Em outro artigo, Biber, Conrad e Cortes (2004, p. 376) esclarecem que o ponto de corte, em relação à frequência, "é de alguma forma arbitrário", o que parece legitimar as escolhas aqui realizadas.

[143] To qualify as a lexical bundle, a word combination must frequently recur in a register. In the following findings, lexical sequences are counted as 'recurrent' lexical bundles only if they occur at least ten times per million words in a register. These occurrences must be spread accross at least tive different texts in the register (to exclude individual speaker/writer idiosyncrasies). Because five-word and six-word bundles are generally less common, a lower cut-off of at least five times per million words is used for those types. (DOUGLAS BIBER ET AL., 1999, p. 992-993)

A figura 9 mostra a primeira parte da tela capturada do software Wordlist, contendo os primeiros 32 clusters de um total de 2.636, organizados pela ordem de frequência no *corpus*, extraídos de acordo com os critérios e ponto de corte acima definidos, ou seja, *bundles* de 3 a 8 palavras, com recorrência mínima de 5 vezes e distribuição em pelo menos 3 textos diferentes. Dentre estes primeiros *bundles* apresentados já é visível a recorrência de clusters de 6 e 7 palavras, evidenciando-se assim a relevância dos clusters compostos por mais de 4 palavras dentro do *corpus*. Biber (2009) salienta que ainda se sabe muito pouco sobre padrões linguísticos compostos por sequências de palavras mais longas que as usualmente pesquisadas (2, 3 ou 4 palavras). “Por isso, investigação adicional é requerida para documentar os tipos de padrões encontrados em sequências mais longas e desenvolver um quadro conceitual que capture as relações entre padrões formulaicos de qualquer tamanho”[144] (BIBER, 2009, p. 301). Assim, parece que um movimento nessa direção é aqui realizado procurando detectar a relevância dos bundles extensos no *corpus* de Tecnologia Ambiental.

[144] [...] we know much less about the formulaic patterns represented by longer sequences of words. Thus, additional reserach is requires to document the kinds of patterns found in longer sequences, and to develop a framework that captures the relationships among formulais patterns of any lenght. (BIBER, 2009, p. 301)

Figura 9 - Clusters extraídos com a ferramenta Wordlist



Tecn ambiental main_index_index_3-8-word clusters em pelo menos 3 textos.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts
1	THE PURPOSE OF THIS STUDY WAS TO	826		6
2	THE OBJECTIVE OF THIS STUDY WAS TO	751		6
3	AT THE END OF THE AEROBIC	731		3
4	IT CAN BE SEEN THAT THE	541		6
5	AT THE END OF THE EXPERIMENT	534		3
6	AT THE END OF THE ANAEROBIC	509		4
7	USED IN THIS STUDY WAS	464		3
8	IN ORDER TO DETERMINE THE	450		6
9	IN ORDER TO EVALUATE THE	427		4
10	IN ORDER TO REDUCE THE	420		5
11	IN ORDER TO IMPROVE THE	412		5
12	THAT CAN BE USED FOR	412		3
13	THIS IS DUE TO THE	410		5
14	NO SUB X AND VOC	393		3
15	BY THE END OF THE	349		8
16	ONE OF THE MOST IMPORTANT	339		5
17	CAN BE USED AS A	330		8
18	ON THE OTHER HAND THE	317		14
19	AS A RESULT OF THE	314		9
20	RESULTS ARE PRESENTED IN TABLE	289		3
21	TO HUMAN HEALTH AND THE ENVIRONMENT	284		3
22	IT WAS FOUND THAT THE	260		11
23	IT WAS ALSO FOUND THAT THE	258		4
24	INCREASE IN THE NUMBER OF	258		3
25	IS ONE OF THE MOST	258		11
26	OF THIS STUDY WAS TO INVESTIGATE	243		5
27	IN THE PRESENCE OF OXYGEN	236		4
28	AS WELL AS THE	229		21
29	OF NO SUB X	216		3
30	AT THE BEGINNING OF THE	200		10
31	IN ORDER TO ACHIEVE	196		5
32	IN ORDER TO OBTAIN	196		7
33	NO SUB X EMISSION	194		3

frequency consistency statistics filenames notes

2,636 Type-in THE PURPOSE OF THIS STUDY WAS TO

A tabela 6 apresenta os 200 pacotes lexicais mais frequentes e recorrentes no *corpus* de Tecnologia Ambiental, tornando evidente, a partir da apresentação de uma quantidade maior de dados, a importância dos bundles compostos por mais de quatro palavras dentro do *corpus* desta pesquisa.

Tabela 6 – Lista parcial de pacotes lexicais do *Corpus* de Tecnologia Ambiental

Lista parcial de pacotes lexicais do *Corpus* de Tecnologia Ambiental		
N	Word	Freq.
1	THE PURPOSE OF THIS STUDY WAS TO	826
2	THE OBJECTIVE OF THIS STUDY WAS TO	751
3	AT THE END OF THE AEROBIC	731
4	IT CAN BE SEEN THAT THE	541
5	AT THE END OF THE EXPERIMENT	534
6	AT THE END OF THE ANAEROBIC	509
7	USED IN THIS STUDY WAS	464
8	IN ORDER TO DETERMINE THE	450
9	IN ORDER TO EVALUATE THE	427
10	IN ORDER TO REDUCE THE	420
11	IN ORDER TO IMPROVE THE	412
12	THAT CAN BE USED FOR	412
13	THIS IS DUE TO THE	410
14	NO SUB X AND VOC	393
15	BY THE END OF THE	349
16	ONE OF THE MOST IMPORTANT	339
17	CAN BE USED AS A	330
18	ON THE OTHER HAND THE	317
19	AS A RESULT OF THE	314
20	RESULTS ARE PRESENTED IN TABLE	289
21	TO HUMAN HEALTH AND THE ENVIRONMENT	284
22	IT WAS FOUND THAT THE	260
23	IS ONE OF THE MOST	258
24	IT WAS ALSO FOUND THAT THE	258
25	INCREASE IN THE NUMBER OF	258
26	OF THIS STUDY WAS TO INVESTIGATE	243
27	IN THE PRESENCE OF OXYGEN	236
28	AS WELL AS THE	229
29	OF NO SUB X	216
30	AT THE BEGINNING OF THE	200
31	IN ORDER TO OBTAIN	196
32	IN ORDER TO ACHIEVE	196
33	NO SUB X EMISSION	194
34	THE USE OF THE	182
35	BE DUE TO THE	175
36	AS CAN BE SEEN IN FIG	171
37	AT THE SAME TIME THE	165
38	AS WELL AS IN	165
39	DUE TO THE HIGH	165
40	AS WELL AS FOR	165
41	THE MECHANICAL PROPERTIES OF THE	165
42	IT IS FOUND THAT THE	164
43	WAS DUE TO THE	163
44	THE AIR QUALITY IN THE	161
45	PROBABLY DUE TO THE	160
46	DUE TO THE HIGHER	157
47	THE USE OF A	156
48	CAN BE USED TO	155
49	TO THE USE OF	155
50	ON THE USE OF	155
51	THE END OF NITRIFICATION	154
52	THAT THE USE OF	153
53	WITH THE USE OF	153
54	BY THE USE OF	153
55	IN THE CASE OF	146
56	IT SHOULD BE NOTED THAT	146
57	TO THE END OF	146
58	THE END OF DENITRIFICATION	146
59	AS SHOWN IN FIG	145
60	IS BASED ON THE	144
61	BEFORE THE END OF	144
62	THE RESULTS OF THE	142
63	ON THE BASIS OF THE	142
64	ARE SHOWN IN FIG	141
65	IT WAS OBSERVED THAT THE	137
66	IS SHOWN IN FIG	135
67	ARE BASED ON THE	133
68	AS SHOWN IN TABLE	131
69	AN INCREASE IN THE	128
70	IN THIS STUDY THE	126
71	WAS BASED ON THE	125
72	IN THIS STUDY WE	124
73	ARE SHOWN IN TABLE	119
74	THE LIFE CYCLE ASSESSMENT LCA	119
75	BASED ON THE RESULTS	118
76	WAS FOUND TO BE	115
77	IT HAS BEEN ESTIMATED THAT	115
78	IS SHOWN IN TABLE	110
79	IN THE RANGE OF	109
80	THE PRESENCE OF A	109
81	LESS THAN OR EQUAL TO	109
82	AS ONE OF THE	108
83	ONE OF THE MAJOR	108
84	CAN BE USED IN	107
85	FROM THE POINT OF VIEW OF	106
86	ONE OF THE MAIN	106
87	BY THE NUMBER OF	106
88	AT A FLOW RATE OF	105
89	IN THE FORM OF	104
90	LIFE CYCLE ASSESSMENT LCA IS	104
91	NONE OF THE	103
92	TO THE PRESENCE OF	102
93	AND THE PRESENCE OF	101
94	THE PRESENCE OF ORGANIC	101
95	THE INCREASE IN THE	99
96	UNITED STATES ENVIRONMENTAL PROTECTION AGENCY	98
97	IT IS OBSERVED THAT THE	98
98	THE PULP AND PAPER INDUSTRY	97
99	AS PART OF THE	96
100	FOR THE PRODUCTION OF	95
101	COULD BE EXPLAINED BY THE	92
102	IN THE ANOXIC PHASE	90
103	AT THE BOTTOM OF THE	89

104	IN THE SOIL AND	89
105	IN TERMS OF THE	88
106	IN A SEQUENCING BATCH REACTOR	88
107	THE RESULTS OF THIS	87
108	BE USED IN THE	87
109	IT IS POSSIBLE TO	86
110	WITH RESPECT TO THE	85
111	IN THE AMOUNT OF	85
112	THE RATE OF THE	83
113	ARE PRESENTED IN FIG	81
114	FOR THE CASE OF	81
115	WERE CARRIED OUT IN	80
116	ARE SHOWN IN FIGURE	80
117	THE EFFECT OF THE	79
118	IN THE PRODUCTION OF	79
119	ARE SHOWN IN FIGS	78
120	IN THE PRESENT STUDY	76
121	THAT THERE IS A	76
122	BECAUSE OF THE	75
123	WAS CARRIED OUT IN	75
124	IN AN INCREASE IN	75
125	IS SHOWN IN FIGURE	74
126	IN THE UNITED STATES	73
127	THE EFFICIENCY OF THE	73
128	BE USED FOR THE	73
129	THE PERFORMANCE OF THE	72
130	THAT THERE IS NO	72
131	USED TO DETERMINE THE	72
132	WERE FOUND TO BE	71
133	WAS USED AS A	71
134	WAS USED AS THE	70
135	THE CONCENTRATION OF THE	70
136	OF THE ENVIRONMENTAL IMPACT	70
137	IN ADDITION TO THE	69
138	FOR THE PURPOSE OF	69
139	THE EFFECTS OF THE	68
140	ACCORDING TO THE	67
141	IN ACCORDANCE WITH THE	66
142	THE FACT THAT THE	66
143	THE PRODUCTION OF THE	66
144	OF EACH OF THE	65
145	MAY BE USED TO	65
146	WAS ADDED TO THE	64
147	THE DEVELOPMENT OF THE	64
148	IN A NUMBER OF	64
149	RESULTS ARE SHOWN IN	64
150	IT WAS SHOWN THAT THE	64
151	THE COMPOSITION OF THE	63
152	THE C N RATIO	63
153	WERE CARRIED OUT AT	63
154	ARE PRESENTED IN FIGS	63
155	WAS USED FOR THE	62
156	A DECREASE IN THE	62
157	IT IS WELL KNOWN THAT	62
158	THAT MOST OF THE	62
159	THAT THE ADDITION OF	61
160	THE PH OF THE	61
161	OF THE TOTAL	60
162	THAT SOME OF THE	60
163	WERE CARRIED OUT TO	60
164	IN SOME OF THE	60
165	OF TOTAL PETROLEUM HYDROCARBONS TPH	60
166	TO THE DEVELOPMENT OF	60
167	IN THE ANAEROBIC PHASE	60
168	AS A FUNCTION OF	59
169	WITH THE INCREASE IN	59
170	WITH THE ADDITION OF	59
171	CONSIDERED TO BE A	58
172	OF THE ANOXIC PHASE	58
173	THE DEVELOPMENT OF A	57
174	WAS CARRIED OUT BY	57
175	USED TO EVALUATE THE	57
176	WAS USED TO	56
177	IT IS IMPORTANT TO	56
178	DESPITE THE FACT THAT	56
179	IN THE DEVELOPMENT OF	56
180	IS BASED ON A	55
181	C N RATIO WAS	55
182	THE PURPOSE OF THE	55
183	WITH THE PURPOSE OF	55
184	HYDROGENATION OF CASTOR OIL	55
185	IS PRESENTED IN TABLE	54
186	THERE IS A NEED	54
187	FOR THE DEVELOPMENT OF	54
188	OF THE AEROBIC PHASE	54
189	WITH A FLAME IONIZATION DETECTOR	54
190	EXPERIMENTS WERE CARRIED OUT	53
191	MEASUREMENTS WERE CARRIED OUT	53
192	IN THE ABSENCE OF	52
193	RESULTS AND DISCUSSION THE	52
194	WAS OBSERVED IN THE	52
195	SEQUENCING BATCH REACTOR SBR	52
196	IN THE PRESENT WORK	52
197	AND THERE IS NO	52
198	BE ADDED TO THE	52
199	ECO MANAGEMENT AND AUDIT SCHEME	52
200	THE LENGTH OF THE	52

4.1.5 Lexical bundles com palavras-chave do *corpus* de TA

A investigação dos *lexical bundles* objetivou ser ainda mais específica e procurou detectar *lexical bundles* que contêm palavras-chave. A pressuposição era que, em relação ao *corpus* de textos de Tecnologia Ambiental, poderiam existir diversos pacotes lexicais contendo palavras-chave e que estes poderiam ser termos ou expressões peculiares da área de Tecnologia Ambiental.

A detecção dos bundles contendo as palavras-chave foi realizado a partir de scripts computacionais especializados desenvolvidos pelo professor Dr. Tony Berber Sardinha, co-orientador desta pesquisa. Esses scripts fizeram a comparação das listas de palavras do *corpus* com a lista dos clusters detectados. O resultado da comparação, ou seja, a extração de todos os clusters contendo pelo menos uma palavra-chave resultou na listagem final, contendo 2114 clusters com no mínimo uma palavra-chave. Na tabela 7 são apresentados os 100 primeiros *lexical bundles* compostos por pelo menos uma palavra-chave, específicos do *corpus* de artigos de Tecnologia Ambiental. Também foram identificados *lexical bundles* pertinentes à linguagem acadêmica geral, isto é, de termos e expressões presentes em diferentes gêneros textuais que circulam no meio acadêmico.

Tabela 7 – Lexical bundles com palavras-chave do *corpus* de TA

Lexical bundles com palavras-chave do *corpus* de TA			
	Cluster	Nr. KW	Keywords presentes no bundle
1.	THE PURPOSE OF THIS STUDY WAS TO	2	<OF, STUDY>
2.	THE OBJECTIVE OF THIS STUDY WAS TO	2	<OF, STUDY>
3.	RESULTS ARE PRESENTED IN TABLE	4	<IN, PRESENTED, RESULTS, TABLE>
4.	TO HUMAN HEALTH AND THE ENVIRONMENT	2	<AND, ENVIRONMENT>
5.	OF THIS STUDY WAS TO INVESTIGATE	2	<OF, STUDY>
6.	THE MECHANICAL PROPERTIES OF THE	3	<MECHANICAL, OF, PROPERTIES>
7.	THE LIFE CYCLE ASSESSMENT LCA	2	<CYCLE, LCA>
8.	UNITED STATES ENVIRONMENTAL PROTECTION AGENCY	1	<ENVIRONMENTAL>
9.	IN A SEQUENCING BATCH REACTOR	3	<BATCH, IN, REACTOR>
10.	OF ENVIRONMENTAL AUDITING	3	<AUDITING, ENVIRONMENTAL, OF>
11.	OF TOTAL PETROLEUM HYDROCARBONS TPH	5	<HYDROCARBONS, OF, PETROLEUM, TOTAL, TPH>
12.	OF THE TOTAL	2	<OF, TOTAL>
13.	WAS USED TO	1	<USED>
14.	WITH A FLAME IONIZATION DETECTOR	0	<>
15.	MEASUREMENTS WERE CARRIED OUT	2	<MEASUREMENTS, WERE>
16.	ECO MANAGEMENT AND AUDIT SCHEME	4	<AND, AUDIT, ECO, MANAGEMENT>
17.	THE INTERNATIONAL AGENCY FOR	0	<>

	RESEARCH		
18.	REAL TIME CONTROL	1	<CONTROL>
19.	THE ENVIRONMENTAL PROTECTION AGENCY	1	<ENVIRONMENTAL>
20.	FUSED SILICA OPEN TUBULAR COLUMN	1	<COLUMN>
21.	ENVIRONMENTAL PROTECTION AGENCY EPA	2	<ENVIRONMENTAL, EPA>
22.	SUCH AS THE	0	<>
23.	THE NITRIFICATION AND DENITRIFICATION	3	<AND, DENITRIFICATION, NITRIFICATION>
24.	IN THE FIRST	1	<IN>
25.	AN ENVIRONMENTAL MANAGEMENT SYSTEM	3	<ENVIRONMENTAL, MANAGEMENT, SYSTEM>
26.	OF ENVIRONMENTAL MANAGEMENT	3	<ENVIRONMENTAL, MANAGEMENT, OF>
27.	PHOSPHORUS AND NITROGEN REMOVAL	4	<AND, NITROGEN, PHOSPHORUS, REMOVAL>
28.	ENVIRONMENTAL LAWS AND REGULATIONS	3	<AND, ENVIRONMENTAL, REGULATIONS>
29.	OF ENVIRONMENTAL AUDITS	3	<AUDITS, ENVIRONMENTAL, OF>
30.	NITRIFICATION AND DENITRIFICATION PROCESSES	4	<AND, DENITRIFICATION, NITRIFICATION, PROCESSES>
31.	POLYCYCLIC AROMATIC HYDROCARBONS PAHS	3	<AROMATIC, HYDROCARBONS, PAHS>
32.	OF THE SOIL	2	<OF, SOIL>
33.	WERE USED TO	2	<USED, WERE>
34.	ENVIRONMENTAL MANAGEMENT SYSTEMS	3	<ENVIRONMENTAL, MANAGEMENT, SYSTEMS>
35.	IN THE FOLLOWING	1	<IN>
36.	AND IN THE	2	<AND, IN>
37.	DIFFERENT TYPES OF	3	<DIFFERENT, OF, TYPES>
38.	SCALE SEQUENCING BATCH REACTOR	3	<BATCH, REACTOR, SCALE>
39.	IN THE REACTOR	2	<IN, REACTOR>
40.	OF SOLID WASTE	3	<OF, SOLID, WASTE>
41.	OF THE STUDY	2	<OF, STUDY>
42.	THE LEVEL OF	2	<LEVEL, OF>
43.	G M SUP	3	<G, M, SUP>
44.	UNDER AEROBIC CONDITIONS	2	<AEROBIC, CONDITIONS>
45.	THE FORMATION OF	1	<OF>
46.	BIOLOGICAL NUTRIENT REMOVAL BNR	3	<BIOLOGICAL, NUTRIENT, REMOVAL>
47.	OF VOLATILE ORGANIC COMPOUNDS	3	<COMPOUNDS, OF, ORGANIC>
48.	HAZARDOUS WASTE MANAGEMENT	3	<HAZARDOUS, MANAGEMENT, WASTE>
49.	IN THE TARGET	1	<IN>
50.	IN ACTIVATED SLUDGE	3	<ACTIVATED, IN, SLUDGE>
51.	TO INCREASE THE	1	<INCREASE>
52.	ENVIRONMENTAL IMPACT ASSESSMENT EIA	3	<EIA, ENVIRONMENTAL, IMPACT>
53.	EXPERIMENTS WERE CONDUCTED TO	3	<CONDUCTED, EXPERIMENTS, WERE>
54.	OF ENVIRONMENTAL PROTECTION	2	<ENVIRONMENTAL, OF>
55.	THE RECYCLING OF	2	<OF, RECYCLING>
56.	IN ADDITION THE	2	<ADDITION, IN>
57.	THE CONCEPT OF	1	<OF>
58.	THE EXTENT OF	1	<OF>
59.	THE VALUE OF	2	<OF, VALUE>
60.	THE CONCENTRATIONS OF	2	<CONCENTRATIONS, OF>
61.	OF THE ENVIRONMENT	2	<ENVIRONMENT, OF>
62.	THE REMOVAL OF	2	<OF, REMOVAL>
63.	IT IS ALSO	0	<>
64.	OF ENVIRONMENTAL ISSUES	2	<ENVIRONMENTAL, OF>
65.	ALL RIGHTS RESERVED	0	<>
66.	BETWEEN THE TWO	0	<>
67.	OF THE SYSTEM	2	<OF, SYSTEM>
68.	HEALTH AND SAFETY	1	<AND>
69.	OF THE INTERNAL	2	<INTERNAL, OF>
70.	OF THE PROCESS	2	<OF, PROCESS>

71.	THE AUDIT TEAM	1	<AUDIT>
72.	CAN ALSO BE	0	<>
73.	NEED TO BE	0	<>
74.	EMISSIONS FROM THE	1	<EMISSIONS>
75.	THE APPLICATION OF	1	<OF>
76.	THE BEHAVIOR OF	2	<BEHAVIOR, OF>
77.	EFFECT ON THE	0	<>
78.	FOUND IN THE	1	<IN>
79.	THE NEED FOR	0	<>
80.	MIXED LIQUOR SUSPENDED SOLIDS	1	<MIXED>
81.	ASSOCIATED WITH THE	0	<>
82.	THE POSSIBILITY OF	1	<OF>
83.	SHOWS THAT THE	0	<>
84.	NEEDS TO BE	0	<>
85.	ALL OF THE	1	<OF>
86.	COMPUTATIONAL FLUID DYNAMICS CFD	0	<>
87.	BIOLOGICAL PHOSPHORUS REMOVAL	3	<BIOLOGICAL, PHOSPHORUS, REMOVAL>
88.	INFORMATION ON THE	0	<>
89.	THE PERCENTAGE OF	1	<OF>
90.	THERE ARE SEVERAL	0	<>
91.	IN THIS RESEARCH	1	<IN>
92.	IN THE EFFLUENT	2	<EFFLUENT, IN>
93.	OF THE PARTICLE	2	<OF, PARTICLE>
94.	RESULTED IN A	1	<IN>
95.	IN THE EARLY	1	<IN>
96.	OF THE MODEL	2	<MODEL, OF>
97.	THE TYPE OF	1	<OF>
98.	DISSOLVED OXYGEN CONCENTRATION	3	<CONCENTRATION, DISSOLVED, OXYGEN>
99.	EMISSIONS IN THE	2	<EMISSIONS, IN>
100.	COULD NOT BE	0	<>

4.2 Texto-chave

Uma análise possível e útil a ser realizada pelo professor de língua para o desenvolvimento de atividades é a seleção do texto-chave, também denominado de texto-focal. Um texto-chave é um texto do *corpus* que contém o maior número de palavras-chave. Já foi vista a importância das palavras-chave como indicadoras do campo semântico/conceitual do qual o texto é originário. Logo, conforme coloca Tribble, "uma análise de palavras-chave também oferece um meio poderoso de se determinar que palavras (e expressões) importam em uma coleção de exemplos de um gênero[145]" (TRIBBLE, 2000, p. 90). Assim, detectar o texto-chave significa encontrar o texto que é o mais representativo da linguagem analisada, ou seja, no caso desta pesquisa, o que mais se destaca do conjunto de 86 textos que compõem o *corpus* de Tecnologia Ambiental por conter o maior número de palavras-

[145] Tradução de Maurício Einstoss de Castro Barbosa utilizada em sua dissertação de mestrado, citada na bibliografia.

chave, as quais estão diretamente associadas à terminologia da área, conforme se pode constatar na descrição dos dados do *corpus*.

O texto chave foi selecionado a partir da utilização de outro *script* especializado desenvolvido pelo Dr. Tony Berber Sardinha. O script fez a comparação entre a lista das 500 palavras-chave mais frequentes do *corpus* (*standart top 500 keywords*) e a lista de palavras-chave de cada um dos textos que compõem o *corpus*. Assim, dos 86 textos analisados pelo script, foi possível detectar o texto-chave do *corpus*, segundo esse critério. A tabela 8 apresenta uma síntese de alguns detalhes dos 15 textos do *corpus* de Tecnologia Ambiental, organizados de acordo com a quantidade de palavras-chave, identificadas na coluna *Nr. KW* (número de palavras-chave).

Tabela 8 - Lista de chavicidade de textos – organizada de acordo com a quantidade de KEYWORDS

Lista de chavicidade de textos – organizada de acordo com a quantidade de KEYWORDS						
Ordem Classif.	Arquivo original do texto	Disciplina	Nr KW	Nr. Pág	Tokens/tokens usados para a lista	Proporção keywords/ tokens
1	wastefoodindustry.txt	Trat. E rec.	340	105	33859/32409	0.00100416
2	Emerging Opportunities for Environmental Auditing.txt	Gest. e Tecn. Amb.	279	183	45,118/42,761	0.0061837
3	surfactante02.txt	Recup. Areas. Degrad.	265	18	9,270/8,687	0.0285868
4	surfactante03.txt	Recup. Areas. Degrad.	218	25	11,650/11,153	0.0187124
5	5994 Characterization of compost-like outputs from mechanical biological treatment of municipal solid waste.txt	Trat. E rec.	198	16	6,277 / 5,668	0.0315437
6	sdarticle-06.txt	Recup. Areas. Degrad.	195	10	4,257 / 3,907	0.0458069
7	4402 Electron beam technology for multipollutant emissions control from heavy fuel oil-fired boiler.txt	Contr. Poluição Atmosférica	194	12	4,809 / 4,501	0.040341
8	ACVCOMPOSITOKENAF.txt	Trat. E rec.	190	10	6,010 / 5,679	0.0316139
9	04.txt	Recup. Areas. Degrad.	187	11	3,412 / 3,065	0.0548065
10	18.txt	Recup. Areas. Degrad.	185	10	4,297 / 3,910	0.0430532
11	10.txt	Recup. Areas. Degrad.	178	6	3,516 / 3,182	0.0506257
12	8153 Potential ozone impacts of excess [NO.sub.txt	Contr. Poluição Atmosférica	177	28	9,030 / 8,166	0.0196013
13	20.txt	Recup. Areas. Degrad.	176	11	4,337 / 4,084	0.040581
14	9199 Recycling hazardous wastes.txt	Trat. E rec.	176	23	9,526 / 8,813	0.0184757
15	09.txt	Recup. Areas. Degrad.	172	12	4,089 / 3,808	0.042064

A detecção do texto-chave utilizando como único critério a quantidade de palavras-chave, conforme demonstrado na tabela 8, pode gerar resultados não tão fidedignos. Até porque a quantidade de palavras-chave de um *corpus*, geralmente, está relacionada com o tamanho do arquivo, isto é, o maior texto do *corpus*, contendo o maior número de palavras, terá probabilidade maior de conter a maior quantidade de palavras-chave e, assim, ser selecionado como o texto-chave de um *corpus*. Conforme matematicamente previsto, tal fato ocorreu nesta análise e o texto contendo o maior número de palavras-chave foi o texto contido no arquivo *wastefoodintrustry.txt,* conforme dados da tabela 8. Neste caso, para se assegurar da chavicidade do texto foi necessário aplicar um algoritmo para calcular a proporção entre o número de palavras-chave e o número de palavras (tokens) de cada um dos textos, isto é, para saber o grau de chavicidade de cada texto, em relação ao *corpus* ao qual pertencia. O algoritmo da chavicidade do texto pode ser representado pela seguinte fórmula:

Texto-chave = número de palavras-chave / número total de palavras (tokens) do texto

Essa segunda análise (tabela 9), mais depurada que a anterior, apresentou resultados bastante diferentes. Para facilitar ao leitor a comparação entre os dados, incluiu-se na tabela 9, abaixo, na primeira coluna, a ordem de classificação dos dados a partir do cálculo da proporção entre palavras-chave e o número de palavras do texto (tokens). A segunda coluna indica a chavicidade dos textos a partir do critério único da contagem das palavras-chave (conforme já detalhado na tabela 8). A partir dessa comparação é perceptível que, por essa análise, os textos com maior grau de chavicidade são outros, diferindo em tamanho e mesmo em quantidade de palavras-chave.

Tabela 9 - Lista de chavicidade de textos – organizada pela média keyword/tokens (em comparação à quantidade de keywords)

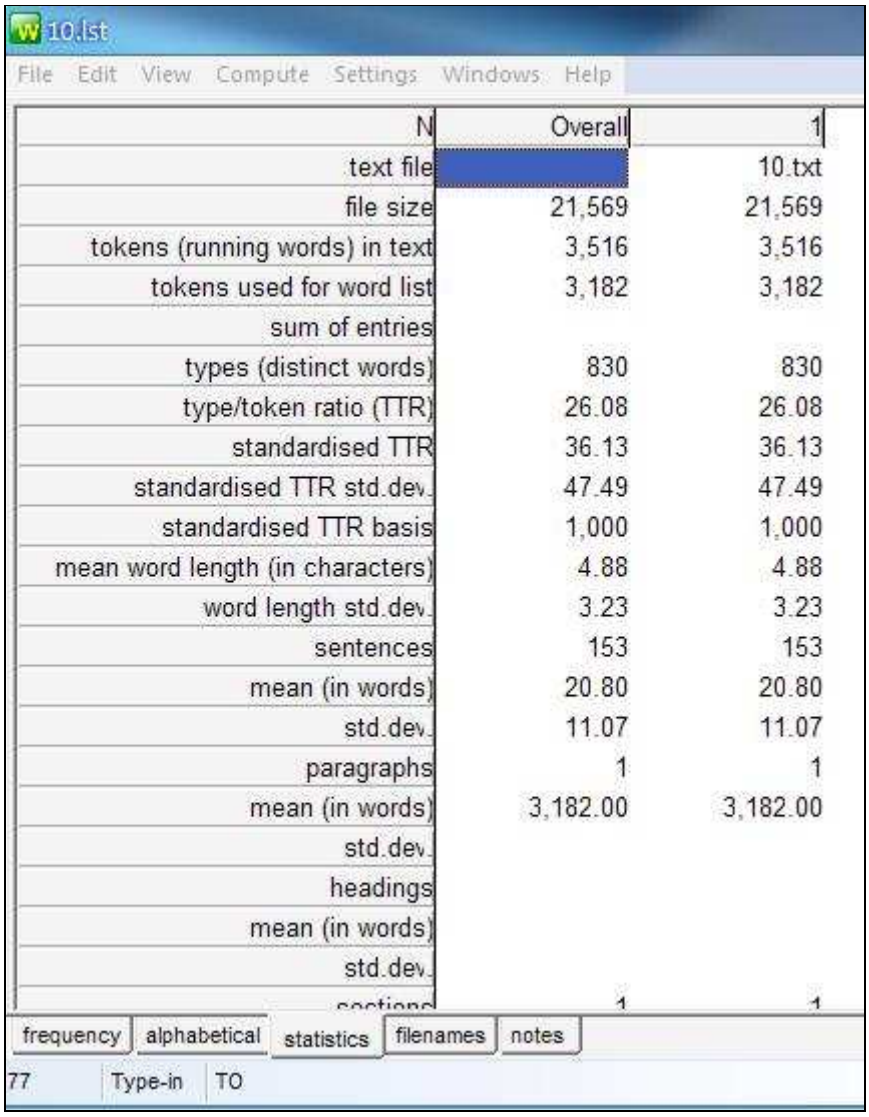
Lista de chavicidade de textos – organizada pela média keyword/tokens (em comparação à quantidade de keywords)							
Ordem Classif. por CHAVICIDADE Proporção Keywords /tokens	Ordem Classif. por KEYWORDS	Arquivo original do texto	Disciplina do PPGTA	Nr KW	Nr. Pág	Tokens/tokens usados para a lista	Proporção keywords/ tokens
1	9	04.txt	Recup. Areas. Degrad.	187	11	3,412 / 3,065	0.0548065
2	11	10.txt	Recup. Areas. Degrad.	178	6	3,516 / 3,182	0.0506257
3	6	sdarticle-06.txt	Recup. Areas. Degrad.	195	10	4,257 / 3,907	0.0458069
4	10	18.txt	Recup. Areas. Degrad.	185	10	4,297 / 3,910	0.0430532
5	8	ACVCOMPOSITOKENAF.txt	Trat. E rec.	190	10	6,010 / 5,679	0.0316139
6	5	5994 Characterization of compost-like (...).txt	Trat. E rec.	198	16	6,277 / 5,668	0.0315437
7	3	surfactante02.txt	Recup. Areas. Degrad.	265	18	9,270/8,687	0.0285868
8	12	8153 Potential ozone impacts of excess [NO.sub.txt	Contr. Poluição Atmosférica	177	28	9,030 / 8,166	0.0196013
9	4	surfactante03.txt	Recup. Areas. Degrad.	218	25	11,650/11,153	0.0187124
10	14	9199 Recycling hazardous wastes.txt	Trat. E rec.	176	23	9,526 / 8,813	0.0184757
11	1	wastefoodindustry.txt	Trat. E rec.	340	105	33859/32409	0.00100416
12	2	Emerging Opportunities for Environmental Auditing.txt	Gest. e Tecn. Amb.	279	183	45,118/42,761	0.0061837
13	15	09.txt	Recup. Areas. Degrad.	172	12	4,089 / 3,808	0.042064
14	13	20.txt	Recup. Areas. Degrad.	176	11	4,337 / 4,084	0.040581
15	7	4402 Electron beam technology for (...).txt	Contr. Poluição Atmosférica	194	12	4,809 / 4,501	0.040341

Ao selecionar um texto para a realização de atividades em sala de aula, a chavicidade não pode ser o único critério a ser considerado. O tamanho do texto também deve ser levado em conta, pois, do contrário, dificilmente se conseguirá concluir uma atividade dentro do tempo disponível. Assim, a escolha do texto-chave, para ser utilizado como base para o desenvolvimento das atividades aqui propostas, recaiu sobre o texto salvo no arquivo *10.txt,* classificado na segunda posição da tabela 9, de acordo com o critério da proporção palavras-chave/tokens. O texto selecionado apresenta chavicidade muito semelhante ao texto com ao maior grau de chavicidade, classificado na posição 1, conforme indica a tabela, além de conter apenas seis páginas, tamanho que o torna viável para a realização de atividades em sala de aula. Outro cuidado a ser tomado na seleção do texto, diz respeito ao perfil dos alunos, sobretudo, em se tratando de alunos que recém estão familiarizando-se com o estudo da língua inglesa, ou com a linguagem acadêmica, ou com ambos. Pode ser comum, em turmas de estudantes de curso de inglês instrumental, alunos com os mais variados níveis de conhecimento, tanto da língua estrangeira, quanto de conhecimento específico de sua área. Não é raro em turmas de ESP estarem matriculados estudantes de diferentes níveis acadêmicos, desde graduandos fazendo esta cadeira no primeiro semestre, alunos realizando o TCC (trabalho de conclusão de curso), até pós-graduandos de especialização e mestrado, principalmente.

4.2.1 Especificidades do texto-chave selecionado

Assim posto, o texto-chave selecionado a partir dos critérios acima referidos é o texto do arquivo *10.txt*, o qual contém o artigo “A Field Trial for an *ex-situ* bioremediation of a drilling mud-polluted site” de autoria de Rojas-Avelizapa et al. (2006), o qual foi selecionado como parte da bibliografia da disciplina Recuperação de Áreas Degradadas. A figura 10 apresenta os dados estatísticos desse artigo, ou seja, do texto-chave.

Figura 10 – Dados estatísticos do texto-chave



N	Overall	1
text file		10.txt
file size	21,569	21,569
tokens (running words) in text	3,516	3,516
tokens used for word list	3,182	3,182
sum of entries		
types (distinct words)	830	830
type/token ratio (TTR)	26.08	26.08
standardised TTR	36.13	36.13
standardised TTR std.dev.	47.49	47.49
standardised TTR basis	1,000	1,000
mean word length (in characters)	4.88	4.88
word length std.dev.	3.23	3.23
sentences	153	153
mean (in words)	20.80	20.80
std.dev.	11.07	11.07
paragraphs	1	1
mean (in words)	3,182.00	3,182.00
std.dev.		
headings		
mean (in words)		
std.dev.		
sections	1	1

Pela análise dos dados apresentados na figura 10, é possível obter diversas informações do texto-chave, entre elas: o texto contém 3516 palavras (tokens) em sua totalidade, porém apresenta apenas 830 palavras diferentes (types), o tamanho médio das palavras é em torno de 5 caracteres. Além disso, contém 153 frases, sendo que cada uma contém uma média de aproximadamente 20 palavras. A

contagem do número de parágrafos apresentada deve ser desconsiderada[146]. Outro dado importante diz respeito ao número total de palavras diferentes (830 types) desse texto, o que está muito abaixo do mínimo necessário, ou seja das 2000 palavras que Nation (2001) menciona, conforme citado no capítulo 2, ou em torno de 3000 palavras, caso seja adicionada a lista de vocabulário acadêmico (570 termos, segundo Coxhead (1998)). O artigo selecionado como texto-chave apresenta menos de 1/3 do mínimo de vocabulário necessário. Tais dados tornam-se muito importantes e também, interessantes, pois o texto analisado é um texto divulgado e distribuído através do portal ScienceDirect com o selo da editora Elsvevier. Em outras palavras, é um texto legitimado pela comunidade acadêmica e científica que, no entanto, foi produzido com uma quantidade não muito extensa de vocabulário, abaixo inclusive das estimativas mínimas segundo os estudiosos do vocabulário já mencionados.

A partir desta análise percebe-se a relevância da escolha deste artigo para ser utilizado como texto-chave e como o primeiro artigo para a leitura e estudo, diretamente com os alunos. Outro ponto importante desta análise diz respeito à própria eficácia da metodologia e instrumental da Linguística de *Corpus* para a seleção deste artigo, que demonstrou ser de um excelente tamanho para o trabalho em aula, mostrando que é possível escolher textos autênticos, adequado ao nível dos alunos, sem necessitar utilizar textos adaptados.

O texto-chave selecionado também foi comparado com a lista de vocábulos acadêmicos elaborada por Coxhead (1998) . A comparação é realizada com o comando "match words in list". Esse comando compara duas listas de palavras, e ao fazê-lo, permite que sejam deletados ou marcados tanto os termos semelhantes como os termos diferentes nas duas listas (matched/unmatched), de acordo com um critério definido de antemão. Para a elaboração dessa lista comparativa, foram utilizadas a lista de termos acadêmicos elaborada por Coxhead (1998) e a lista das palavras (types) do texto-chave. O resultado mostrou que o texto-chave contém 36 termos em comum com a lista de termos acadêmicos de Coxhead, e essa quantidade é aproximadamente 5% do total do texto-chave. Assim, conforme já estimado, o ensino do vocabulário acadêmico geral merece destaque

[146] Esse dado (apenas 1 parágrafo) é resultado das conversões de formato que desconfiguraram o layout padrão do arquivo. Operações, no entanto, necessárias para a organização do *corpus* e seu processamento no Wordsmith Tools 5.0.

em um curso de inglês instrumental. A tabela 10 apresenta os vocábulos acadêmicos encontrados na comparação das duas listas.

Tabela 10 - - Lista de palavras da ACADEMIC WORDLIST (Coxhead, 1998) presentes no TEXTO-CHAVE

Vocabulário acadêmico presentes no TEXTO-CHAVE			
N	Word	Freq.	%
1	CONCENTRATION	13	0.3
2	COMPOUNDS	7	0.2
3	AMENDED	6	0.1
4	ANALYSIS	5	0.1
5	APPROPRIATE	3	0.0
6	ATTRIBUTED	3	0.0
7	CONSTANT	3	0.0
8	ACHIEVE	2	0.0
9	ACHIEVING	2	0.0
10	ADJUST	2	0.0
11	AUTHORS	2	0.0
12	AVAILABILITY	2	0.0
13	AVAILABLE	2	0.0
14	CONSUMPTION	2	0.0
15	ABANDONED	1	0.0
16	ABSTRACT	1	0.0
17	ACHIEVEMENT	1	0.0
18	ADAPTED	1	0.0
19	ADJUSTED	1	0.0
20	ALTERNATIVE	1	0.0
21	AMENDMENTS	1	0.0
22	APPROACH	1	0.0
23	APPROACHES	1	0.0
24	BENEFIT	1	0.0
25	CHALLENGE	1	0.03
26	CHEMICAL	1	0.03
27	COMPOUND	1	0.03
28	CONDUCT	1	0.03
29	CONSEQUENTLY	1	0.03
30	CONSISTENT	1	0.03
31	CONSTRUCTED	1	0.03
32	CONSTRUCTION	1	0.03
33	CONSUMED	1	0.03
34	CONTRIBUTION	1	0.03
35	CORPORATION	1	0.03
36	INAPPROPRIATE	1	0.03

Além do mais, a análise do texto chave, elencando as frequências e quantidade do vocabulário utilizados pelos autores mostrou que, conhecer com precisão a trama do texto, pode possibilitar a professores de idiomas que organizem um programa de ensino totalmente focado nas necessidade específicas deste grupo de alunos. Com isso, evita-se o ensino de tópicos desnecessários para os alunos, prática muitas vezes comum em livros didáticos de língua inglesa geral. Tal análise, também surpreende ao mostrar a gama de vocabulário utilizada para a produção do artigo, relativamente pequena, quantidade essa que é possível de ser explorada com alunos ao longo de um semestre e que no entanto é a variação da língua utilizada pela comunidade intelectual à qual o aluno vincula-se.

4.2.2 Lista de palavras-chave do texto-chave

Selecionado o texto-chave fez-se uma lista das palavras-chave nele contidas. A lista consta da tabela a seguir - tabela 11. Esta lista pode ser utilizada pelo educador como um guia para estabelecer estratégias de ensino acerca do vocabulário específico do texto-chave, pois indica o campo semântico do artigo, e, além de fornecer subsídios para o professor, diversas atividades poderão ser elaboradas, a partir dela, para exploração do texto-chave.

Tabela 11 - Lista de palavras-chave do texto-chave

Listas de palavras-chave do texto-chave					
ADDITION	BIODEGRADATION	DUE	HIGHEST	PARAMETERS	SAMPLES
AERATION	BIOLOGICAL	DURING	IN	PARTIAL	SATURATION
AEROBIC	BIOPILE	E	INCREASED	PERFORMANCE	SCALE
AIR	BIOREMEDIATION	EFFECTS	KEYWORDS	PH	SHOWN
AIRFLOW	C	ENVIRONMENT	KG	PHASE	SIGNIFICANTLY
AL	CA	ENVIRONMENTAL	LIQUID	PILE	SITE
AMBIENT	CALCULATED	EQ	LOW	PLASTIC	SITU
AND	CM	EQUILIBRIUM	M	PROCESS	SOIL
APPLIED	COMPARED	ET	MASS	PROCESSES	SOLUTION
APPROXIMATELY	COMPONENTS	EXCESS	MAXIMUM	PRODUCTS	SPILL
AVERAGE	CONCENTRATION	EXTERNAL	MEASURED	PROFILES	STUDY
BASED	CONDITIONS	EXTRACTION	MEASUREMENTS	PROGRAM	SYSTEM
	CONDUCTED	FACILITIES	METHOD	PROPERTIES	SYSTEMS
	CONTAMINANT	FIG	METHODS	PW	TEMPERATURE
	CONTAMINATED	FIGS	MICROBIAL	RATE	TOTAL
	CONTAMINATION	FLOW	MICROORGANISMS	RATES	USED
	CONTENT	FRACTION	MIN	REACTION	USING
	CONTROL	FUEL	MODEL	REACTIONS	VALUE
	D	G	MODELED	REMEDIATION	VALUES
	DATA	GAS	MOISTURE	REQUIREMENTS	WATER
	DECREASE	GENERATED	MONITORING	RESIDUAL	WERE
	DECREASED	GROUNDWATER	N	RESPECTIVELY	
	DENSITY	H	NUTRIENT	RESULTS	
	DIAMETER	HIGH	OF	S	
	DIFFERENT	HIGHER	OXYGEN	SAMPLE	

4.2.3 Pacotes lexicais presentes no texto-chave

Durante o processo de análise e elaboração de listas dos vocábulos de um *corpus*, tanto de palavras individuais como de pacotes lexicais, um procedimento que pode ser utilizado pelo pesquisador é a lematização. Segundo Kennedy (1998) a "lematização é um processo que permite classificar, sob a base (raiz- grifo meu) de uma palavra, todas as formas identicas ou relacionadas[147]" (KENNEDY, 1998, p. 207). As palavras podem ser lematizadas de acordo com critérios estabelecidos a priori pelo linguista, tanto manualmente, quantro através de recursos automatizados oferecidos software. Assim, uma palavra no singular ou plural poderá ser processada pelo programa como a mesma palavra, desinências verbais diferentes de um mesmo verbo, também. Biber, Conrad e Reppen (1998) exemplificam o processo da lematização, lembrando que:

[147] Lemmatization is a process of clasifying together all the identical or related forms of a word under a common headword. (KENNEDY, 1998, p. 207)

> Ao estudar uma palavra, é com frequência útil considerar coletivamente as diferentes formas da palavra. Isto é, embora o computador identifique *deal, deals, dealing and dealt* como diferentes palavras, talvez se queira discuti-las de forma agregada ou investigar a frequência delas enquanto um grupo. O termo "lemma" é usado para referir à forma basilar da palavra, independentemente de oscilações gramaticais como das formas verbais e plurais[148]. (BIBER, CONRAD e REPPEN, 1998, p. 29).

O mesmo procedimento descrito em uma das seções anteriores, para extrair os pacotes-lexicais do *corpus* de TA, foi utilizado para para extraí-los do texto-chave. Optou-se por selecionar os pacotes lexicais que ocorreram no mínimo duas vezes no texto-chave, contendo de 3 a 8 palavras, o que totalizou 230 *clusters* pela análise do Wordlist. Essa lista de clusters foi manualmente depurada, isto é, os *clusters* semelhantes foram lematizados e assim se obteve o total de 153 pacotes lexicais. O resultado final do processo de lematização (manual) da lista de bundles é visualizado na tabela 12, a seguir.

Tabela 12 - Lista de pacotes lexicais presentes no texto-chave

Lista de pacotes lexicais do Texto-chave			
A C N P RATIO OF	AT THE END OF TREATMENT	DEMONSTRATED THAT THE	FOR HETEROTROPHIC AND HYDROCARBON
A FIELD TRIAL FOR	BACTERIA AND TOTAL FUNGI	DRILLING MUD POLLUTED SOIL	DEGRADING BACTERIA
A POLLUTED SOIL	BACTERIA REMAINED IN THE RANGE	DUE TO THE	FOR HYDROCARBON DEGRADATION
A WIDE RANGE	BEEN REPORTED BY	DURING THE EXPERIMENTATION	FOR HYDROCARBON DEGRADING BACTERIA
A WIDE RANGE OF	BEGINNING OF THE EXPERIMENTATION	DURING THE EXPERIMENTATION PERIOD	GAS CHROMATOGRAPHIC ANALYSIS
A WIDE RANGE OF C	BIOPILES EXPB AND BIOPILES WERE COVERED	DURING THE FIRST	HAS BEEN REPORTED
ADJUST AND MAINTAIN THE	BIOPILES WERE ESTABLISHED	DURING THE TREATMENT	HAS BEEN REPORTED BY
ALKYL DIBENZOTHIOPHENE TYPE	BY COMPOSTING IN BIO PILES	DURING THE WHOLE	HETEROTROPHIC AND HYDROCARBON
AND C P RATIOS OF	C N AND C P RATIOS OF	DURING THE WHOLE EXPERIMENTATION	DEGRADING BACTERIA
AND HYDROCARBON	C N P RATIO OF	END OF THE EXPERIMENTATION PERIOD	HIGH TPH CONCENTRATION
DEGRADING BACTERIA	C N RATIO	END OF TREATMENT	HYDROCARBON DEGRADING BACTERIA
AND MAINTAIN THE	C P RATIOS OF COMPOSTING IN BIOPILES	EXPB AND UAMB BIOPILES WAS	HYDROCARBON DEGRADING BACTERIA AND TOTAL
AND TOTAL FUNGI	D OF EXPERIMENTATION	EXPB AND UAMB RESPECTIVELY	HYDROCARBON DEGRADING BACTERIA AND TOTAL FUNGI
AND UAMB BIOPILES WAS	D OF TREATMENT	FIELD TRIAL FOR	IN A CHRONICALLY
AND UAMB RESPECTIVELY	DEGRADING BACTERIA AND TOTAL FUNGI	FOR EXPB AND UAMB RESPECTIVELY	IN A POLLUTED
AS CAN BE		FOR EXPB BIOPILES	IN A POLLUTED SOIL
AT A SOIL			IN BIO PILES
AT THE BEGINNING OF			
AT THE END OF			

[148] When studying a word, it is often useful to consider the different forms of the word collectively. That is, though the computer identifies *deal, deals, dealing* and *dealt* as different words, we may want to discuss them all together or investigate their frequency as a group. The term "lemma" is used to mean the base form of a word, disregarding grammatical changes such as tense and plurality. (BIBER, CONRAD e REPPEN, 1998, p. 29)

IN EXPB AND UAMB BIOPILES	OBSERVED DURING THE	THE BEGINNING OF THE	THE UNAMENDED BIOPILE
IN EXPB AND UAMB BIOPILES WAS	OBSERVED IN EXPB	THE BEGINNING OF THE	THE WHOLE EXPERIMENTATION
IN EXPB BIOPILES	OF SOIL IN	EXPERIMENTATION	TO ADJUST AND MAINTAIN THE
IN ORDER TO	OF THE EXPERIMENTATION	THE BIOPILES WERE COVERED	TO DETERMINE THE
IN OUR LABORATORY	PERIOD	THE BIOREMEDIATION OF	TO THE END OF
IN OUR STUDY	OF THE STRAW	THE C N RATIO	TOTAL PETROLEUM HYDROCARBONS
IN THE CASE	ON SITE DISPOSAL	THE CASE OF	TPH CONCENTRATION OF
IN THE CASE OF	P RATIOS OF	THE COMPOSITE SOIL	TPH REMOVAL WAS
IN THE LAST	PERIOD OF TIME	THE END OF	UAMB AND EXPB
IN THE LITERATURE	POLLUTED SITES IN THE	THE END OF THE EXPERIMENTATION	UAMB BIOPILES WAS
IN THE RANGE	POLLUTED SITES WITH	THE END OF THE EXPERIMENTATION PERIOD	UNAMENDED BIOPILE
IN UAMB BIOPILE	POTENTIAL	THE END OF TREATMENT	UAMB
IT HAS BEEN REPORTED	APPLICATION OF	THE EXPERIMENTATION PERIOD	UP TO THE END OF
IT IS WELL KNOWN THAT	PRESENT IN THE	THE OBJECTIVE OF	WAS ATTRIBUTED TO
MG TOC KG	RANGE OF C	THE PRESENCE OF	WAS PERFORMED BY
MG TPH KG	RATES AT D	THE RANGE OF	WELL KNOWN THAT
MICROBIAL GROWTH AND	RATIO IN EXPB	THE REMOVAL OF	WERE OBSERVED DURING THE
MIGHT HAVE BEEN	REMAINED IN THE	THE SOUTHEAST OF MEXICO	WHICH WAS ATTRIBUTED TO
MOISTURE CONTENT WAS	REMAINED IN THE RANGE	THE TPH REMOVAL	WIDE RANGE OF
MUD POLLUTED SOIL	RESIDUAL TPH	THE TREATMENT BY	WIDE RANGE OF C
N AND C P RATIOS OF	CONCENTRATION OF		WITHIN THE RANGE OF
N P RATIO OF	SITES IN THE SOUTHEAST OF MEXICO		
NATIVE C N	STRAW RATIO OF		
NITROGEN AND PHOSPHOROUS	TEMPERATURE AND MOISTURE		
NON EXCAVATED SOIL	THE ADDITION OF		

4.2.4 Comparação de lexical-bundles do *corpus* com aqueles presentes no texto-chave

Outra análise realizada foi a comparação da lista de pacotes lexicais do *corpus* de Tecnologia Ambiental (2636 itens) com a lista de pacotes lexicais do texto-chave (153 itens). Esse procedimento foi realizado pela ferramenta Wordlist através do comando “match words in list” descrito anteriormente. Para a elaboração dessa lista comparativa, foram utilizadas essas duas listas contendo os pacotes lexicais previamente extraídos.

O resultado foi uma listagem onde permaneceram apenas os pacotes lexicais idênticos nas duas listas de bundles. Todos os pacotes lexicais que não combinaram (*not matched*), foram eliminados, e o sistema gerou uma nova lista visualizada na figura 11. O resultado dessa operação totalizou 20 pacotes lexicais diferentes que o texto-chave e o *corpus* de TA tem em comum. A figura 11 mostra os resultados visualizados na interface do Wordlist após a comparação das duas listas de pacotes lexicais.

Figura 11 - Pacotes Lexicais recorrentes no Texto-chave e *corpus* de TA

MATCHED LIST DOS CLUSTERS DO TEXTO CHAVE COM OS DO CORPUS.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	IT IS WELL KNOWN THAT	10	0.06	1	100.00
2	THE C N RATIO	5	0.06	1	100.00
3	TO THE END OF	4	0.06	1	100.00
4	DURING THE FIRST	3	0.09	1	100.00
5	IN OUR STUDY	3	0.09	1	100.00
6	MOISTURE CONTENT WAS	3	0.09	1	100.00
7	OBSERVED DURING THE	3	0.09	1	100.00
8	PRESENT IN THE	3	0.09	1	100.00
9	WAS ATTRIBUTED TO	3	0.09	1	100.00
10	A WIDE RANGE OF	2	0.06	1	100.00
11	BEEN REPORTED BY	2	0.06	1	100.00
12	DEMONSTRATED THAT THE	2	0.06	1	100.00
13	DURING THE WHOLE	2	0.06	1	100.00
14	IN THE CASE OF	2	0.06	1	100.00
15	IN THE LAST	2	0.06	1	100.00
16	IN THE LITERATURE	2	0.06	1	100.00
17	REMAINED IN THE	2	0.06	1	100.00
18	SITES IN THE	2	0.06	1	100.00
19	THE BIOREMEDIATION OF	2	0.06	1	100.00
20	THE REMOVAL OF	2	0.06	1	100.00

O próximo capítulo detalhará os procedimentos envolvidos na produção da tarefa de ensino a ser utilizada com alunos do curso de Pós-graduação em Tecnologia Ambiental.

5 PLANEJAMENTO E APRESENTAÇÃO DAS TAREFAS

Este capítulo apresenta atividades de ensino produzidas segundo os pressupostos da Linguística de *Corpus*, anteriormente, discutidos, mantendo, ainda, consonância com o conceito de tarefa. O ponto inicial foi a análise dos dados do *corpus*, utilizando-se as listas de frequências dos vocábulos, geradas pelo software, dentre elas incluindo-se termos gerais (types), palavras-chave e pacotes lexicais. A análise das listas de palavras possibilitou determinar os termos de maior relevância para os alunos do curso de Tecnologia Ambiental.

O texto-chave foi decisivo para a organização das atividades aqui delineadas, sendo, de fato, o norteador de todas elas. A partir dele, da análise de seu vocabulário e de sua lexicogramática, foram elaboradas as atividades. Assim, a proposta é que se utilize o texto-chave, em aula, desde o primeiro encontro, ocasião em que os alunos devem receber, cada um, uma cópia. Com isso, intenciona-se que os aprendizes possam relacionar as pesquisas dos dados coletados do *corpus*, as amostras de linguagem extraídas e o texto em si. Como já explicado antes, os dados do *corpus* são todos autênticos, sendo fonte de farto input linguístico para o aprendiz. No entanto, esses dados são apresentados no concordanciador por meio de amostras da linguagem que, mesmo em grande abundância e buscando favorecer o aprendizado do aluno, podem parecer segmentados. Em vista disso, a utilização do artigo original, com sua diagramação e layout autênticos, possibilita colocar nas mãos do aluno o “discurso” em si, na sua totalidade, pois é dele que o curso pretende dar conta.

Essa modalidade de trabalho, que utiliza tanto os dados do *corpus* quanto o texto original, é produtiva por ter mão dupla, isto é, a partir da leitura do artigo, dúvidas (dos alunos) ou provocações do professor ou, ainda, perguntas focais concernentes aos enunciados do material de ensino podem levar a uma consulta aos dados do *corpus* de estudo, que fornecerá uma grande gama de amostras da língua para auxiliar o aprendiz a entender o discurso. Por outro lado, ao consultar o *corpus*, o aluno terá a oportunidade de retornar ao texto original e verificar o discurso, integralmente, comprovando como a língua funciona, como a língua se

comporta, de fato, seja por um viés pragmático, seja por alguma questão lexicogramatical. É no vai e vem entre texto e dados do *corpus*, na possibilidade de pensar a língua a partir de sua lexicogramática, na aplicação e utilização do conhecimento linguístico para construir o sentido do texto que talvez resida a riqueza de uma proposta de ensino, tendo por suporte o referencial da Linguística de *Corpus*.

O vai e vem entre *corpus* e discurso, entre concordâncias e texto, num primeiro momento, é estimulado a partir da análise das concordâncias e textos, ambos impressos. Posteriormente, à medida que o aluno apropria-se da estrutura das concordâncias, pensa-se, seja o momento de começar a utilizar o software. A utilização do software concordanciador, tanto o *Antconc* como o *Concord,* permite que, de forma rápida e instantânea, o aprendiz acesse o texto original das concordâncias consultadas. Ao simples clique do *mouse* sobre o nódulo (Key Word in Context) leva ao texto original. A única desvantagem do procedimento está em que o texto original, por estar gravado no formato TXT, não apresenta a configuração original de sua publicação.

Convém salientar, ainda, que na elaboração do *Corpus* de Tecnologia Ambiental procurou-se manter a 'quase' integridade do texto, eliminando-se apenas elementos gráficos e referências bibliográficas, justamente com o intuito de permitir que o aluno tivesse a possibilidade de consultar o texto original, para verificar, quando necessário, a utilização dos termos pesquisados num contexto maior, seja em nível frasal, em nível de parágrafo ou da totalidade do artigo. Esse foi um critério específico adotado na produção do *corpus* de TA. No entanto, é importante frisar, cada *corpus* tem seus critérios de elaboração, de acordo com seus propósitos e, nem todos, apresentam cópias integrais dos textos que lhes deram origem.

As primeiras atividades delineadas preveem a consulta aos dados do texto-chave, tanto em listas de palavras, quanto em linhas de concordâncias. A tarefa desenvolve-se, gradativamente. Assim, à medida que os alunos fazem a leitura das concordâncias do texto-chave, que apresentam uma quantidade pequena de dados, se comparadas à totalidade do *corpus*, incluem-se, aos poucos, atividades e tarefas que os estimulem, ou mesmo solicitem, a consultar a totalidade de dados do *corpus* de Tecnologia Ambiental.

As atividades foram planejadas, com base na expectativa de as turmas serem heterogêneas, sendo constituídas por alunos com diferentes níveis de conhecimento da língua e de sua área acadêmica. Em vista disso, a seleção de termos a serem explorados ocorreu, também, antevendo-se a possibilidade de, em uma determinada turma, haver alunos com excelente conhecimento da língua. Logo, mesmo nas primeiras aulas aparecem vocábulos e padrões da léxicogramática, possivelmente, classificados como “avançados”, em outras metodologias de ensino.

De início, as atividades propõem que o aluno realize um ‘sobrevoo’ sobre a estrutura textual. Nesse sentido, algumas tarefas voltam-se à exploração do conceito do gênero discursivo “artigo científico”, buscando o reconhecimento de sua estrutura e organização, bem como de alguns aspectos relacionados à sua produção, recepção e circulação. Gradativamente, as atividades passam a explorar elementos específicos da lexicogramática. O trabalho inicial, sistematizado, prevê que ao apropriar-se do instrumental de pesquisa, o aluno possa, por si mesmo, proceder a pesquisas de uso da língua, de forma autônoma, a partir de dúvidas surgidas em seus estudos.

Optou-se por não incluir neste capítulo a descrição de procedimentos relacionados ao ensino do manejo dos softwares, pois se entende não ser essa a principal questão desta pesquisa. A ênfase recai nas possibilidades de utilização das tarefas, em um sentido bastante amplo, tanto que todas as atividades, se o professor assim o quiser, podem ser realizadas utilizando concordâncias impressas. Na verdade, as primeiras atividades com as linhas de concordâncias, embora inicialmente planejadas para serem impressas, foram também desenhadas, dando ao professor opções de uso. Isto é, podem ser impressas, apresentadas em projetores ou através de um software como o power-point, em conjunto com o software concordanciador, bem como adaptadas para atividade em um ambiente de ensino a distância (EAD).

5.1 Atividades e tarefas

As atividades desenvolvidas são apresentadas a seguir, acompanhadas de explicações a respeito das escolhas feitas, tanto do ponto de vista da lexicogramática, quanto da utilização de conceitos da Linguística de *Corpus* e da

Linguística Cognitiva. Todas essas atividades estão inter-relacionadas e devem ser entendidas como fazendo parte de uma 'única atividade' a ser aplicada ao longo de 8-10 horas de aula, em torno de três ou quatro encontros, dependendo da organização de horários de cada instituição.

5.1.1 Atividade 1 – *Words, words, words*

Atividade de pré-leitura que procura despertar a atenção para a importância do vocabulário. Desde o primeiro momento, o aluno é colocado em contato com recursos da Linguística de *Corpus* direcionados para o ensino. Neste caso, utiliza-se parte da lista de palavras mais frequentes do texto-chave como uma atividade de preparação para a leitura do artigo e, ao mesmo tempo, de avaliação do conhecimento vocabular que os alunos possuem.

Atividade 1: Words, words, words...

A lista abaixo apresenta as palavras mais frequentes no texto que você lerá e ocorrem no mínimo 10 vezes. O sinal "#" indica a quantidade de algarismos presentes no texto.

Responda às seguintes peguntas:

1) A partir destas palavras, é possível prever o tema do texto?

2) A que tipo de texto/gênero esse vocabulário pode estar relacionado?

3) Há uma série de letras que não compõe palavras (A, C, N, P, X,G) e que são bastante recorrentes neste texto. Elas tem alguma significação especial?

4) Você conhece alguma dessas siglas/abreviações : TPH, EXPB, UAMB, FIG, MG, BA?

N	Word	Freq	16	BIOPILES	27	32	N	16	48	DEGRADING	11
1	#	334	17	EXPB	27	33	BIOPILE	15	49	HAVE	11
2	THE	178	18	ET	26	34	DURING	14	50	MOISTURE	11
3	AND	144	19	D	23	35	MICROBIAL	14	51	RATIO	11
4	OF	138	20	HYDROCARBON	22	36	OBSERVED	14	52	RESPECTIVELY	11
5	IN	94	21	UAMB	22	37	P	14	53	AFTER	10
6	TO	74	22	THAT	21	38	TEMPERATURE	14	54	BA	10
7	A	61	23	REMOVAL	20	39	X	14	55	BE	10
8	SOIL	55	24	WITH	19	40	CONCENTRATION	13	56	DRILLING	10
9	WAS	47	25	BY	18	41	FIG	13	57	FROM	10
10	TPH	40	26	AS	17	42	G	13	58	IT	10
11	FOR	36	27	IS	17	43	MG	13	59	ON	10
12	AT	30	28	POLLUTED	17	44	BACTERIA	12	60	OR	10
13	C	29	29	AN	16	45	TREATMENT	12	61	STRAW	10
14	WERE	28	30	EXPERIMENTATION	16	46	BIODEGRADATION	11	62	THIS	10
15	AL	27	31	KG	16	47	CONTENT	11			

5.1.2 Atividade 2 – Você é um leitor de artigos científicos?

Esta atividade pretende discutir com os alunos alguns conceitos atinentes ao gênero discursivo - artigo científico, contendo questões relacionadas ao seu uso, função social, produção, recepção, distribuição e, também, a alguns elementos de sua estrutura. A atividade prevê o reconhecimento dos principais elementos da estrutura do artigo científico. Tal atividade desenvolve-se a partir da entrega de quatro artigos científicos, a serem utilizados ao longo do semestre. Esses artigos apresentam o maior grau de chavicidade, isto é, a maior proporção de palavras-chave em relação ao total do número de palavras do artigo.

A critério do professor, dependendo do conhecimento prévio dos alunos, esta atividade pode ou não ser aplicada. Se a turma for heterogênea, composta de graduandos e pós-graduandos, com níveis de conhecimento muito diferentes entre si, pode ser necessária. No evento de a turma ser uniforme, composta unicamente de pós-graduandos, é provável que uma rápida conversa sobre esses pontos seja suficiente.

Atividade 2: Você é um leitor de artigos científicos?

Caso você já leia artigos acadêmicos, ótimo, já conhece um pouco deste gênero discursivo!

Do contrário, na aula de hoje, terá oportunidade de descobrir algumas características dos artigos científicos ou acadêmicos. Serão discutidos alguns

pontos sobre a definição e função de “artigo científico” e sobre a estrutura desse gênero do discurso. Essa disciplina desenvolverá várias atividades com o objetivo de auxiliá-lo na leitura de artigos acadêmicos em língua inglesa. Antes de começarmos a trabalhar com a língua propriamente, precisamos entender o que é esse tipo de texto. Reflita sobre as questões abaixo ou as discuta com um colega e, em seguida, as discutiremos no grupo.

Você já leu algum artigo científico/acadêmico?

Qual a função social de um artigo científico?

Quem o escreve?

Quem o lê?

Onde são publicados?

Onde localizar e encontrar artigos científicos? Você conhece algum meio/mídia ou instituição que os publica?

Como são acessados?

Como fazer para localizá-los no momento em que se necessita pesquisar sobre um determinado tema?

5.1.3 Atividade 3: Reconhecendo elementos da estrutura do artigo científico

A presente atividade é um desdobramento da atividade anterior e pretende que os alunos explorem a estrutura do artigo científico, reconhecendo elementos textuais e gráficos comuns, entre os diferentes artigos entregues, os quais, de fato, são características deste gênero discursivo. A atividade será realizada utilizando quatro artigos, selecionados do *corpus* de Tecnologia Ambiental. Os artigos escolhidos são aqueles que apresentaram maior chavicidade, de acordo com a análise realizada. O primeiro momento da atividade propõe que os alunos reconheçam e identifiquem os elementos e as características da organização textual semelhantes entre os quatro artigos. A segunda atividade centra-se em elementos textuais e gráficos de um único artigo, o artigo com a maior chavicidade.

As perguntas formuladas, na primeira parte desta atividade, dizem respeito aos elementos tipográficos do texto e objetivam estimular os alunos a perceberem que elementos não textuais e gráficos podem, também, caracterizar um artigo científico. Saber realizar uma leitura rápida, recolhendo informações da superfície do texto, é

uma habilidade que os alunos precisam desenvolver, pois ao longo da vida acadêmica, frente à grande quantidade de publicações, precisam aprender a garimpar textos para suas pesquisas e estudos. Assim, através da passagem rápida dos olhos pelo texto, é imprescindível obter dados que se prestem a identificar o texto e que ajudem o leitor a diferenciá-los de outros textos que podem assemelhar-se (por exemplo, artigos de divulgação científica, publicados na mídia) a um artigo científico.

O segundo momento desta atividade, denominado “Estrutura do Artigo”, propõe que os alunos observem a estrutura do artigo científico, prestando atenção à forma como a informação científica ou acadêmica é organizada e distribuída ao longo do texto. Isto é, espera-se que, ao comparar os quatro artigos distribuídos, percebam a estrutura comum entre eles (*abstract*, *keywords*, introdução, resultados, discussão, etc.) e produzam uma síntese, semelhante à representada na tabela 13. A propósito, esta tabela poderá, a critério do professor, ser utilizada como um complemento à atividade, para discutir e comparar a estrutura dos artigos selecionados.

Atividade 3: Reconhecendo elementos da estrutura do artigo científico

A partir da visualização dos quatro artigos entregues, realize uma leitura rápida (menos de 1 min para cada texto), isto é, prestando atenção a marcas e elementos gráficos, organização do texto, disposição de títulos e subtítulos, entre outras informações.

ELEMENTOS PRESENTES NO ARTIGO

É possível obter as informações a seguir:

- Quem são os autores de cada artigo? Você os conhece?
- Onde foi publicado o artigo?
- Quando foi publicado?
- Quem publicou o artigo?

É possível, em cada um dos artigos entregues, identificar:

a) Fórmulas científicas?

b) Algarismos?

c) Gráficos?

d) Ilustrações?

e) Tabelas?

f) Mapas?

g) Símbolos matemáticos?

h) Símbolos de notação científica?

i) Trechos do texto entre aspas?

j) Identificação de autores citados ao longo do texto?

k) Algum outro elemento que você percebeu e merece destaque? Qual?

A partir dessas respostas, é possível imaginar o leitor a quem o texto foi dirigido? Você seria um leitor deste artigo? Por quê?

ESTRUTURA DO ARTIGO

1) Há uma estrutura comum entre os quatro artigos? Qual?

2) Faça um esquema que sintetize a estrutura desses artigos, procurando indicar como a informação está organizada no texto.

3) A tabela que acompanha esta atividade apresenta um esquema que compara a estrutura dos quatro artigos. O esquema que você produziu, de alguma forma, assemelha-se a esse? Você identificou outros elementos que não estão presentes ali?

5.1.3.1 Tabela comparativa da estrutura dos 4 textos com maior chavicidade do *corpus* de Tecnologia Ambiental.

Esta tabela pode acompanhar a atividade 3 “Conhecendo a estrutura do artigo acadêmico/científico” e, ainda, a critério do professor, ser apresentada no *Power Point* ou entregue, na versão impressa, aos alunos. Além do mais, poderá ser utilizada como uma atividade complementar, ou mesmo, como uma atividade única e rápida para rever a estrutura organizacional do artigo científico. A opção depende de os alunos já possuírem experiência na leitura de artigos acadêmicos, ou da disponibilidade de tempo do professor para realizar uma discussão mais aprofundada sobre gênero discursivo.

Atividade 3.1: Análise da Tabela comparativa da estrutura dos 4 textos com maior chavicidade do corpus de TA

A tabela 13 apresenta um esquema sintético da estrutura de quatro artigos científicos diferentes. Como você definiria cada uma dessas partes? Qual a função de cada uma dessas partes e de seus elementos na estrutura e organização do texto?

Tabela 13 - Tabela comparativa da estrutura dos quatro textos com maior chavicidade do *corpus* de TA

Tabela comparativa da estrutura dos quatro textos com maior chavicidade do *corpus* de TA			
TEXTO-CHAVE 1 Arquivo 10.txt	TEXTO –CHAVE 2 Arquivo 04.txt	TEXTO –CHAVE 3 Arquivo sdarticles-06.txt	TEXTO –CHAVE 4 Arquivo 18.txt
CABEÇALHO DE IDENTIFICAÇÃO • Dados da editora e distribuidor • Título do artigo • Autores • Informações complementares sobre os autores • Datas de recebimento e de aceite do artigo **CORPO DO TEXTO DO ARTIGO** Abstract Keywords • Introduction • Materials and methods • Results and discussion • Aknowledgements • References	**CABEÇALHO DE IDENTIFICAÇÃO** • Dados da editora e distribuidor • Título do artigo • Autores • Informações complementares sobre os autores • Datas de aceite e de publicação do artigo **CORPO DO TEXTO DO ARTIGO** Abstract Keywords • Introduction • Materials and methods • Results • Discussion • Conclusions • Aknowledgements • References	**CABEÇALHO DE IDENTIFICAÇÃO** • Dados da editora e distribuidor • Título do artigo • Autores • Informações complementares sobre os autores • Data de aceite do artigo **CORPO DO TEXTO DO ARTIGO** Abstract Keywords • Introduction • Materials and methods • Results and discussion • Conclusions • Aknowledgements • References	**CABEÇALHO DE IDENTIFICAÇÃO** • Dados da editora e distribuidor • Título do artigo • Autores • Informações complementares dos autores • Data de aceite do artigo **CORPO DO TEXTO DO ARTIGO** Abstract Keywords • Introduction • Materials and methods • Results • Discussion • Conclusions • Aknowledgements • References

5.1.4 Atividade 4: Pistas textuais

A atividade quatro direciona a atenção do aluno para aspectos textuais a serem detectados em uma leitura rápida. Essa atividade afina estratégias de leitura, estimulando os alunos a perceberem elementos textuais que os auxiliarão a refinar o processo de, por exemplo, selecionar artigos para pesquisa. Assim, aplicarão conhecimentos relacionados à sua área, à estrutura e organização da informação do artigo, bem como à utilização de elementos tipográficos (aspas, sinal de parênteses, sobrescritos, notas de rodapé) para detectarem rapidamente informações específicas do artigo. Nomes de autores, autores citados, nomes instituições de pesquisa e universidades, referências bibliográficas, títulos e subtítulos utilizados pelo autor, podem dar pistas sobre o conteúdo de um artigo e auxiliar o leitor a decidir-se ou não pela leitura de determinado artigo. Mais ainda, essas informações textuais escaneadas, rapidamente, se agregadas ao conhecimento estrutural do artigo acadêmico, podem, inclusive, facilitar que se encontre uma informação muito específica. Por exemplo, um aluno do curso de Tecnologia Ambiental gostaria de saber se o autor do artigo, em seu experimento, utilizou determinado produto químico. Conhecendo a estrutura do artigo acadêmico, a probabilidade maior é que tal informação esteja contida na seção de materiais e métodos. Assim, o leitor acadêmico deve apropriar-se dessas estratégias de leitura para poder agir com mais rapidez e acerto.

À medida que o aluno passa a dominar maior quantidade de vocabulário acadêmico, obviamente, o processo de leitura será facilitado. Assim, a atividade 4 pode ser concluída com uma discussão a respeito da relação entre estratégias de leitura e conhecimento vocabular, possibilitando que os alunos descrevam e compartilhem estratégias por eles utilizadas no processo de seleção de informação.

Atividade complementar da atividade 4

SUGESTÃO
Antes iniciar a leitura de qualquer texto, a partir apenas da visualização de algumas palavras e elementos gráficos do texto, procure formular algumas hipóteses sobre o tópico textual. Com o prosseguimento da leitura, você pode comprovar ou não essas hipóteses. Pesquisas sugerem que estratégias como essas ativam os conhecimentos prévios e auxiliam no processamento da leitura. Você utiliza alguma outra estratégia, além das propostas acima, para extrair informações sobre o assunto do texto, antes de lê-lo na íntegra?

Available online at www.sciencedirect.com
ScienceDirect
CHEMOSPHERE
Chemosphere 66 (2007) 1595–1604
www.elsevier.com/locate/chemosphere

A field trial for an *ex-situ* bioremediation of a drilling mud-polluted site

N.G. Rojas-Avelizapa [a,b,*], T. Roldán-Carrillo [b], H. Zegarra-Martínez [a], A.M. Muñoz-Colunga [a], L.C. Fernández-Linares [b]

[a] Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas 152, Col. San Bartolo Atepehuacan, 07360 México D.F., Mexico
[b] Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada del IPN, José Siurob 10, Colonia Alameda, Santiago de Querétaro, Querétaro, Mexico

Received 31 May 2006; received in revised form 3 August 2006; accepted 7 August 2006
Available online 25 September 2006

Abstract

The remediation of drilling mud-polluted sites in the Southeast of Mexico is a top priority for Mexican oil industry. The objective of this work was to find a technology to remediate these sites. A field trial was performed by composting in biopiles, where four 1 ton soil-biopiles were established, one treatment in triplicate and one unamended biopile. Amended biopiles were added with nutrients to get a C/N/P ratio of 100/3/0.5 plus a bulking agent (straw) at a soil/straw ratio of 97/3. Moisture content was maintained around 30–35%. Results showed that, after 180 d, total petroleum hydrocarbon (TPH) concentrations decreased from 99 300 ± 23 000 mg TPH kg^{-1} soil to 5500 ± 770 mg TPH kg^{-1} for amended biopiles and to 22 900 ± 7800 mg TPH kg^{-1} for unamended biopile. An undisturbed soil control showed no change in TPH concentrations. Gas chromatographic analysis showed residual alkyl dibenzothiophene type compounds. Highest bacterial counts were observed during the first 30 d which correlated with highest TPH removal, whereas fungal count increased at the end of the experimentation period. Results suggested an important role of the straw, nutrient addition and water content in stimulating aerobic microbial activity and thus hydrocarbon removal. This finding opens an opportunity to remediate old polluted sites with recalcitrant and high TPH concentration.

Keywords: Hydrocarbon-degrading bacteria; Heterotrophic; Removal; Straw; Total petroleum hydrocarbons

1. Introduction

A large number of polluted sites with different levels of environmental impact have been found at the Southeast of Mexico as a result of more than 60 years of the oil industry activities (Adams et al., 1999). Most of these sites are located in marshes or flooded zones, situation that complicate the entrance to these sites and therefore their restoration. Drilling muds are among the most important hazardous wastes released to the environment at these sites. They help to lubricate and cool down the drill bit during oil well drilling and also to carry the drill cuttings to the surface for further screening and disposal. These hazardous wastes are produced by emulsifying oil and water with colloidal matter that may include organophilic clay, bentonite, barite and other additives (Darley and Gray, 1988).

The disposal of oil-based drill muds is a major environmental and operational issue in offshore drilling. Some alternative methods include on-site disposal, ship to shore and dispose or reinjection in abandoned wells or subsurface caverns. Only two reports are available in the literature concerning to the treatment by landfarming of drilling fluids or cuttings (Zimmerman and Rober, 1991; Lee et al., 2002). Nevertheless, the on-site disposal is the more attractive

* Corresponding author. Address: Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada del IPN, José Siurob 10, Colonia Alameda, Santiago de Querétaro, Querétaro, Mexico. Tel.: +52 442 2124011x216; fax: +52 442 2124011x103.
E-mail address: nrojas@imp.mx (N.G. Rojas-Avelizapa).

0045-6535/$ - see front matter © 2006 Elsevier Ltd. All rights reserved
doi:10.1016/j.chemosphere.2006.08.011

Atividade 4 - PISTAS TEXTUAIS
A imagem ao lado é uma cópia da primeira página de um dos artigos entregues. Este artigo será utilizado para a realização de uma série de atividades.

a) É possível identificar palavras que possam precisar a que área acadêmica vincula-se este artigo?
b) Que termos ou elementos desta página do artigo você relaciona com seu curso?
c) É possível associar este artigo, a partir da leitura desta página, a alguma disciplina do seu curso? Qual?

ABSTRACT E KEYWORDS
"Abstract" e "*keywords*" são partes componente de um artigo acadêmico e, geralmente, correlacionam-se.
a) Você sabe o que é o "*Abstract*"? Qual a sua função na composição do artigo?
b) Para que são utilizadas as *"keywords"?*
c) As *Keywords* do artigo são: "*Hydrocarbon-degrading bacteria; Heterothrophic; Removal; Straw; Total petroleum hydrocarbons".*
d) Vocês as conhece? Como são traduzidas para a língua portuguesa?

REFERÊNCIAS
a) Como identificar quando são citados outros autores?
b) Neste artigo, o que predomina, citações diretas ou indiretas? Como perceber isso ao passar os olhos pelo texto?
c) Você conhece alguma das referências feitas?
As referências bibliográficas estão organizadas por um formato que você conhece (ABNT, por exemplo)?

TEMA
a) A partir das perguntas acima, é possível deduzir o tema deste artigo?
b) A autoria do artigo, citações e outros autores referenciados no texto, dão alguma pista sobre o tema/ assunto do artigo?
c) Você concorda que a organização do texto adotada pelos autores é pertinente para a apresentação, comunicação, divulgação ou discussão de assuntos da sua área acadêmica? Por quê?

5.1.5 Atividade 5: Palavras-chave

A atividade 5 propõe-se a estimular os alunos a refletirem sobre alguns termos específicos da sua área de atuação. As palavras da atividade 5, utilizadas como base para essa tarefa, são palavras-chave extraídas do texto-chave, ou seja, do artigo “A field trial for an *ex-situ* bioremediation of a drilling mud-polluted site”. Foram escolhidas algumas das palavras mais frequentes da listagem gerada pelo software. O critério de chavicidade adotado pelo software para extração das palavras não permite que sejam diferenciados termos técnicos de linguagem geral. Assim, na lista de palavras-chave, surgem termos recorrentes na área de Tecnologia Ambiental que, por vezes, possuem um uso diferente daquele da palavra em outros contextos. Essa atividade pretende que os alunos reflitam sobre o uso do vocabulário, especificamente dos termos listados no enunciado da atividade 5. Ao longo do curso, quando já estiverem utilizando o software para consulta do *corpus*, serão incentivados a verificar com detalhe o uso desses termos.

Além disso, a atividade pretende alertá-los para a necessidade de organizarem seus estudos, pois é grande a quantidade de vocabulário que precisam dominar para se tornarem leitores competentes.

Atividade 5: Palavras-chave

As palavras abaixo são específicas no artigo “A field trial for an *ex-situ* bioremediation of a drilling mud-polluted site”. Você sabe o significado de todas elas?

HETEROTROPHIC
PILE
DEGRADING
DEGRADATION
ENVIRONMENT
MEASURED
SOILS
AERATION
AVERAGE
BEHAVIOR
HAZARDOUS
OPTIMAL
RATES
REMEDIATE
USED
WASTES
CHEMICAL
LEACHING
SCALE
SLUDGE
SOURCE
OPTIMAL
DRILL

2) Algumas dessas palavras possuem uso muito específico na sua área acadêmica. Outros termos são de uso geral da língua inglesa, isto é, podem ser encontrados em outros meios que não o acadêmico. Alguns deles são utilizados tanto na língua geral como em campos específicos, e às vezes, adquirem significados diferentes do usual. Classifique-os, na tabela abaixo, conforme os três critérios: termos de vocabulário específico da sua área, vocabulário geral e termos que são usados nas duas instâncias.

Vocabulário geral	**Vocabulário Geral e específico**	**Vocabulário específico**

5.1.6 Atividade 6: Caderno de vocabulário pessoal

Ademais, sugere-se aos alunos a organização de um caderno de vocabulário pessoal (digital ou em papel), o qual será organizado de acordo com necessidades individuais de cada um. Algumas orientações e sugestões serão fornecidas, inclusive a entrega de um guia produzido pela BBC, *Learning English,* sobre diferentes estilos de organização do léxico para o aprendizado.

O caderno será uma atividade continuada a ser realizada ao longo do período do curso. É importante destacar o papel do professor nesse processo, encorajando-os a tomarem notas dos termos e incentivando-os a pensarem estratégias para organizarem o caderno, bem como a utilizá-lo como um guia de consulta pessoal. Frisam-se algumas recomendações:

1) Anotar os termos sempre acompanhados de exemplos de uso, coletados de textos lidos ou do *corpus*, de preferência em uma frase completa.

2) Incentivar os alunos a anotarem e destacarem palavras que co-ocorrem com a palavra anotada. À medida que melhor entenderem o funcionamento da lexicogramática, procurar sempre fazer observações e notas sobre os padrões dos termos incluídos no caderno de vocabulário.

3) Anotar as incidências de utilização do termo, registrando dados a respeito de seus usos (pragmática).

4) Estimular o uso continuado do caderno de vocabulário como fonte de referência, revisando-o periodicamente e atualizando-o com novos dados sobre termos já vistos.

5.1.6.1 Atividade 6.1

Pesquisar, ao longo da primeira semana de aula, as 235 palavras mais frequentes do texto-chave desconhecidas dos alunos e organizá-las no caderno de vocabulário. Esta pesquisa e estudo serão individuais, realizado fora da aula, dando oportunidade de cada um investigar o vocabulário de acordo com suas necessidades. A lista será entregue aos alunos e a pretensão é de que, ao longo de um semestre, os alunos trabalhem com, pelo menos, as 2000 palavras mais frequentes da língua inglesa, além de parte do vocabulário específico de seu curso.

Alunos com maior conhecimento linguístico poderão receber os demais itens da lista de palavras do texto-chave para pesquisarem termos desconhecidos e iniciarem a organização do seu caderno de vocabulário.

5.1.7 Atividade 7 - Glossário de Termos Técnicos da área de Tecnologia Ambiental

A produção de um glossário com a terminologia especializada presente nos artigos lidos e estudados ao longo do curso, é outra atividade a ser realizada durante todo o curso. Este glossário será uma produção coletiva dos alunos e será realizada por algum meio virtual que permita o acesso e uso compartilhado dos arquivos produzidos, bem como de sua alteração por qualquer membro do grupo. As

possibilidades de produção digital são muitas e acessíveis a quase todos hoje em dia. Ferramentas como “Google docs”, ou “SkyDrive” da Microsoft, “Dropbox”, por exemplo, entre outras ferramentas de *cloud computing,* são gratuitas e poderiam ser utilizadas para o compartilhamento de informações, bem como para a produção compartilhada. O objetivo é que os alunos produzam traduções da terminologia específica, disponibilizando-a para todo o grupo, para que todos possam, ao mesmo tempo, acessar o glossário e produzir as alterações necessárias, contribuindo para o refinamento da qualidade do glossário. Ao final do semestre, o glossário poderá ser disponibilizado para livre acesso para outros interessados na temática, em formato a ser posteriormente definido (webpage, blog, arquivo autoexecutável para pendrive etc.). Também poderá ser um trabalho que irá se modificando ao longo de outros semestres, com novas contribuições e revisões.

O formato definitivo será estabelecido em cada turma, em razão das possibilidades de acesso do grupo aos recursos tecnológicos disponíveis.

5.1.8 Atividades 8 e 9: Estrutura das concordâncias

A proposta desta atividade é apresentar aos estudantes a estrutura da concordância, explorando estratégias de leitura, pois a leitura das concordâncias difere da leitura de um texto. Num primeiro momento, a atividade será realizada no papel. Posteriormente, poderá ser introduzido o uso do software concordanciador, para que os alunos pesquisem os dados no próprio *corpus*, de uma forma mais livre do que através de uma seleção pré-definida. Esta tarefa embasou-se em Gavioli (2005). A atividade 9 apresenta aos alunos a possibilidade de expandir a visualização das linhas de concordância.

Atividade 8: Estrutura das concordâncias

Abaixo você encontra uma lista contendo linhas retiradas de diversos textos acadêmicos. Essas linhas foram extraídas com um software que examinou uma série de artigos científicos de sua área de estudos e as extraiu tendo como critério a palavra: *sludge*. Cada uma dessas linhas é denominada concordância.

1) A maneira como a informação está organizada nesta lista, provavelmente, seja diferente de exemplos com os quais você costuma lidar. De que forma é diferente?

2) Como o “texto” está organizado?

3) Como se lê essas concordâncias?

4) Que tipo de informações é possível retirar dessas concordâncias?

5) As frases estão conectadas entre si?

N	Concordance
1	machinery was limited and some sludge could not be removed (ayora et al
2	started after the mine spill. the toxic sludge covering the ground and a major
3	treatment plants are treated in sludge thickeners & centrifuges and oil
4	from waste matrices such as soil, sludge and sediments. a carrier gas or
5	attributes the high levels of zinc in the sludge to a zinc plater that previously
6	flow to sbr aeration (do > 2 mg/l) sludge may be wasted time — typically
7	beginning of the experiments, activated sludge from local municipal wastewater
8	of groundwater in an up-low sludge blanket reactor (green et al.,
9	the performance of an upflow anaerobic sludge blanket (uasb) reactor treating
10	generators of hazardous electroplating sludge waste who generate less than 60
11	in k061 wastes (i.e., emission control dust/sludge from steel-producing
12	(e.g., spent pickle liquor used as a sludge conditioner in wastewater

A leitura de concordâncias, nem sempre exige que se leiam todas as linhas. A exigência varia de acordo como o que está sendo pesquisado e para que fim.

6) Faça uma leitura vertical observando os termos que antecedem e precedem a palavra *sludge*. Essas palavras o ajudam a compreender o significado de *sludge?*

7) A partir da leitura dessas linhas é possível deduzir o significado da palavra *sludge?*

8) As doze ocorrências da palavra *sludge* têm o mesmo significado?

Atividade 9: Estrutura das concordâncias 2

Às vezes, a quantidade de termos de uma única linha extraída do texto, de cada um dos extratos apresentados, não é suficiente para responder a uma pergunta de pesquisa, a algo que se queira saber sobre a língua estudada. O concordanciador (o software utilizado) permite que as linhas sejam expandidas,

acrescentando mais texto, como no exemplo abaixo. No caso da palavra *sludge*, esta expansão do texto, auxiliou no entendimento do significado de *sludge*?

N	Concordance
1	. in some sites, for example close to tree trunks, the accessibility to clean-up machinery was limited and some sludge could not be removed (ayora et al ., 2001). the first pca axis for the surface (0—25 cm) soil samples explained 55.
2	morena natural park, cma, 2001). an emergency soil clean-up procedure quickly started after the mine spill. the toxic sludge covering the ground and a major portion of the contaminated soil surface were mechanically removed and
3	total recovery of oil from oily sludges. (3) oily sludges generated in wastewater treatment plants are treated in sludge thickeners & centrifuges and oil recovered sludges having 3-5% oil is taken up for bioremediation. bioremediation
4	is a physical separation process applying heat to volatilise organic contaminants from waste matrices such as soil, sludge and sediments. a carrier gas or vacuum system sweeps the volatilised organic contaminants into the gas
5	the circuit board manufacturer is pretreated prior to entering the plant. the plant attributes the high levels of zinc in the sludge to a zinc plater that previously discharged to the plant. the sludge is stored in the third sbr prior to disposal.
6	time — half of the total cycle time minus the unaerated fill time react no influent flow to sbr aeration (do > 2 mg/l) sludge may be wasted time — typically — 1-2 h (varies widely) settle no influent flow to sbr no aeration sludge is wasted
7	by programmable timers (chrontrol, xt series). sbr operation and cyc/es at the beginning of the experiments, activated sludge from local municipal wastewater treatment plants was used to inoculate the sbr's. neither plant had a
8	g no^-n/(l day). the observed loading is some 90% of that found in denitriication of groundwater in an up-low sludge blanket reactor (green et al., 1992) and is, judging by the observed trend in the removal rate, expected to
9	recycling and from evaporator condensates. buzzini and pires (2007) evaluated the performance of an upflow anaerobic sludge blanket (uasb) reactor treating diluted black liquor from a kraft pulp mill, which simulates an unbleached kraft
10	, 65 fr 12378 the accumulation time period is extended from 90 to 120 days for generators of hazardous electroplating sludge waste who generate less than 60 tons/year. * conditional exclusions are finalized for 07/24/02 20:6; 2.1 1)
11	hazardous secondary materials. zinc is a regulated hazardous constituent only in k061 wastes (i.e., emission control dust/sludge from steel-producing electric furnaces). in addition, the definition for "underlying hazardous
12	in another process), or as an effective substitute for a commercial product (e.g., spent pickle liquor used as a sludge conditioner in wastewater treatment). a material is reclaimed if it is processed to recover a usable product,

5.1.9 Atividade 10: Comparando linhas de concordância e dicionário

A presente atividade deriva-se de uma proposta semelhante realizada por Gavioli (2005) e explora elementos peculiares da estrutura das concordâncias, sempre com o intuito de levar os alunos a refletirem sobre seu uso. Neste caso, a atividade aqui desenhada, estabelece uma comparação entre o uso de um dicionário e o uso das linhas de concordância para a consulta a um termo. A atividade prevê o reconhecimento das diferenças entre *exemplo e amostra (sample and example)*

apontadas por Gavioli (2005) e discutidas no capítulo três. Obviamente, ao realizar a atividade abaixo, não se intenciona gerar uma discussão teórica sobre o tema, mas que os alunos reflitam sobre algumas diferenças na utilização de um dicionário e de um concordanciador, para que possam, sempre, tirar o melhor proveito na utilização desses recursos, de acordo com o que cada um oferece de peculiar.

Atividade 10: Comparando linhas de Concordâncias e dicionário

A linguagem especializada nem sempre está disponível em dicionários e o uso de um concordanciador pode auxiliar neste processo. Abaixo está a definição da palavra *sludge* encontrada no *Longman Dictionary of Contemporary English*.

1) Algumas dessas definições correspondem aos extratos de texto acima?

2) Que tipo informação uma pesquisa no concordanciador traz e em que difere da consulta a um dicionário?

sludge /slʌdʒ/ *noun* [uncountable]
Word origin
1 soft thick mud, especially at the bottom of a liquid
2 the solid substance that is left when industrial waste or SEWAGE (=the liquid waste from toilets) has been cleaned
3 thick dirty oil in an engine
—**sludgy** *adjective*

5.2 Exploração da Lexicogramática

As atividades propostas, a partir deste ponto, centram-se na exploração de aspectos da lexicogramática de alguns itens selecionados do *corpus*, a partir da análise das frequências dos vocábulos, das palavras-chave e dos pacotes lexicais. Todas as atividades propostas procuram levar o aprendiz a pensar sobre a língua, a refletir sobre o seu uso, na condição de leitor, buscando apreender como o recurso linguístico foi utilizado pelo autor do texto para fazer sentido. Assim, ao propor atividades e tarefas de natureza inquiridora, acredita-se estar em consonância com as propostas do DDL (data driven learning) e com o conceito de tarefa discutido no capítulo 3, pois a meta é instigar o aprendiz a refletir sobre a língua para a obtenção do sentido, sendo prioridade formar um leitor bilíngue proficiente. Por isso, o aluno será estimulado a fazer o caminho de mão-dupla, realizando consultas no *corpus*,

verificando suas hipóteses, voltando ao texto para analisar a ocorrência no artigo completo, no discurso em sua completude.

As possibilidades de exploração do *corpus* de estudo para o ensino da lexicogramática são muitas e delas diversas atividades poderiam ter sido planejadas. No entanto, algumas escolhas tornam-se necessárias. Definir o que ensinar é crucial e, nesse sentido, o instrumental da Linguística de *Corpus* auxiliou a determinar o que tem relevância, estatisticamente, atestada para o aluno em formação. Os termos selecionados para a exploração dos aspectos lexicogramaticais consideraram a frequência comprovada nas listas obtidas na análise do *corpus*. A seleção desses itens lexicais teve como critério básico que fossem termos significativos dentro do texto-chave, incluindo-se nesse critério os termos “range”, “ratio” e “due to”. Um segundo critério considerado, somado ao anterior, foi que alguns padrões linguísticos tivessem alta representatividade no *corpus* de Tecnologia Ambiental, ou seja, padrões com alta recorrência em todos os textos do *corpus* e presentes majoritariamente em quase todos os textos. Nesse grupo incluiu-se a partícula “it” e padrões lexicogramaticais nos quais ela tem destaque, como “it is” e “it has been”. Um terceiro critério foi a seleção de um termo, embora com uma frequência não muito alta no texto-chave, mas com destaque no uso de relatos científicos, que é o verbo “to find” e o padrão contendo “attributed”. Por último, foram elencados alguns termos específicos da área, bastante frequentes no texto-chave e também no *corpus* de TA, entre eles: biopile, drilling/drill, bacteria, concentration. Embora alguns termos sejam cognatos e certamente do conhecimento dos alunos, a riqueza de seu estudo e exploração está na análise dos padrões por ele formados, que muitas vezes, agregam significações diversas do termo isolado.

Embora a presente pesquisa tivesse como foco o levantamento dos pacotes lexicais no *corpus* de Tecnologia Ambiental, no desenvolvimento das atividades aqui elencadas, optou-se por incluir outros padrões lexicogramaticais que não necessariamente se incluem na definição de pacote lexical proposta por Biber. Um desses exemplos é o binômio “due to” e a conjunção “however”, um termo isolado, se comparado às demais formações de outros itens lexicais. No entanto, essa conjunção precisa ter sua função muito bem definida dentro do texto e não pode ser negligenciada em um programa de ensino organizado a partir do léxico. De fato, as

atividades relacionadas aos aspectos lexicogramaticais se propõem a apresentar um olhar mais abrangente para os aspectos da lexicogramática na sua exploração didática, procurando despertar a atenção para a formação de padrões linguísticos, sejam eles pacotes lexicais ou não.

5.2.1 Atividade 11: lexicogramática de “range”

Atividade 11: Lexicogramática de “range”

As linhas abaixo, contendo as concordâncias para o termo “range”, foram extraídas do artigo “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”.*

1) Observe as palavras que antecedem e que sucedem “range” , que padrões elas formam? Quais você reconhece? Liste pelo menos 3 padrões encontrados nas concordâncias abaixo.

a)_____

b)_____

c)_____

2) Localize esses padrões no artigo, eles estão sublinhados no texto original. Faça a leitura do parágrafo onde ele se encontra.

vely. We have found in the literature that a wide range of C:N and C:P ratios have been reported, e.g.,
n was enough to favour TPH biodegradation. A wide range of C:P ratios has been reported for hydrocarbon
or hydrocarbon-degrading bacteria remained in the range (Fig. 2a). For UamB biopile, also heterothrophi
count for heterothrophic bacteria remained in the range of 1010 CFU g-1, whereas for hydrocarbon-degrad
opriate biodegradation conditions. The optimum pH range for hydrocarbon degradation in 50-,-1 Fig. 1.
incorporated to adjust and maintain the desirable range (30-35%) up to the end of the experimentation p
iod. 3.2. pH The pH of ExpB remained within the range of 7.5-8 throughout the 180 d of experimentatio
:P ratio in ExpB and UamB biopiles was within the range of 100-700 which was attributed to soil heterog
optimal moisture for TPH removal was within this range (Roldan et al., 2003). Non excavated soil in P3

3) Esses três padrões são muito recorrentes em textos acadêmicos. Veja abaixo outros exemplos selecionados do *corpus* de textos de Tecnologia Ambiental. Você detecta alguma outra variação nas concordâncias abaixo? Observe outras palavras que estão colocadas antes ou depois de cada um dos padrões (a wide range of, within the range of, in the range of).

N	Concordance
1	, with foliar Mn showing the widest range in content (Table 1). Foliar Mn
2	class of polymer that appeals to a wider range of applications (Consalves et al.,
3	. Chitosan has been used in a very wide range of applications, such as
4	was made to demonstrate the wide range of policies that could be optimal
5	environmental auditing provides wide range of information, which are not
6	with autotrophic nitrifiers under a wide range of conditions, play a significant
7	that grows very quickly under a wide range of weather condition. It grows
8	exhibit complex variability over a wide range of spatial and temporal scales
9	of the simulation results a wide range of different case studies are used
10	cross-linking or scission of a wide range of materials without dissolving the
11	allows separation of a wide range of materials from each other. The
12	auditing encompasses a wide range of management practices which
13	to favour TPH biodegradation. A wide range of C:P ratios has been reported
14	of the different heavy metals and a wide range of concentrations, between 0 and
15	most microbes can metabolized a wide range of c-compounds, Schwindinger
16	noncompliance ([f.sub.22]), with a wide range of policies in between. However,
17	have found in the literature that a wide range of C:N and C:P ratios have been
18	growth, and is naturally found on a wide range of soil types, being relatively
19	with self-policing can elicit a wide range of behavior. To evaluate the

N	Concordance
23	, the pH in both piles was within this range. Adequate moisture is essential
24	the value presented here falls within this range. Cellulose. Cellulose is the next
25	for TPH removal was within this range (Roldan et al., 2003). Non
26	and operated to remain within this range. Solar radiation and changes in
27	sample F1 falls comfortably within this range. Table 4. Dimambro measured
28	the study presented here lie within this range. The mixed MSW samples they
29	study of 15.2% lies just outside of this range but within 1 standard deviation.
30	presented here fall outside of this range, with F3 being lower and F4 higher
31	contents in foliage were within the range for healthy sugar maples (Table 1),
32	rates presented in Table 5 are within the range given by Beccari et al. (1983),
33	(2) enables calculating that, within the range of mean outdoor concentrations
34	pH The pH of ExpB remained within the range of 7.5-8 throughout the 180 d of
35	ExpB and UamB biopiles was within the range of 100-700 which was attributed
36	in PAH extraction within the range of 5-35 ℃. 2.4. Extraction with
37	in the summertime, igures within the range recommended by the 0zone
38	The CH value of 6.9 c/kg was within the range reported by Nilsson et al. (1998),

Concordance

2 and the nitrate removal rate in the range 0.3-0.5 on AN-D eluent indicate
at 15 rpm and at temperatures in the range 170-200 [degrees]C. A
the biomass concentration was in the range 4-6 g VSS/l (anoxic-reactor) and
of the tap water we used was in the range 518-570 mg CaCO3/l. In the
bacteria remained in the range (Fig. 2a). For UamB biopile, also
measured in both piles were in the range for a normal development of a
, a temperature increase in the range of 10-20 ℃ and an addition of
of simple unbranched n-alkanes in the range of C10-C22 (Heath et al., 1993), it
smell with an odour threshold in the range of 200-640 ug/m3 [4]. in 1987, the
heterothrophic bacteria remained in the range of 1010 CFU g-1, whereas for
levels spike dramatically upward in the range of 5000-10,000 (ig/m3. The mobile
microbial growth should be keep in the range of 6.5-8.5. But now, few
strength of the composites are in the range of 0.06-0.42 MPa at the same
that a total of 12 PAHs increased in the range of 13-56% in the soil samples due
above pH 9.0. The optimum pH is in the range of 6.5-8.0. A dissolved oxygen
obtain the optimal gas humidity in the range of 11-15 vol %, it is necessary to
result in styrene concentrations in the range of 0-800 (ig/m3 being generated in
exhibited copper concentrations in the range of parts per million to parts per
column effluent kept in the range of 5.1-8.6 indicating that this
of particle diameter takes a value in the range of 1.15 ~ 1.25 and results in the
the effect of acid concentration in the range of 0.81-2.0 M was studied at 25
carefully dried at temperatures in the range of 100-130 [degrees]C for at least
period was eliminated) was in the range of 4.1-5.2 mg/l with 92% removal
of composting in the chamber fell in the range of 20-60%, with further removal of
sections of the factory were in the range of 28-30℃. In the study, it was
case of BTXs, these doses were in the range of 4-8 kGy. Only benzene
time, which was optimized in the range of 4 s until 10 min, with eventual
(based 01 BOD) and at an SRT in the range of 25-30 days. J cycle time of 6 h
, DO levels in the effluent were in the range of 2.3-3.0 mg/L, and the
was not sensitive to temperature in the range of 25-55 ℃. However, a minor
The optimum pH for nitrification is in the range of 7.5-9.0. Below pH 7.0 and
by considerable decrease in the range of 30-40%. On the other hand, the
temperature effect was examined in the range of 25-55 ℃ under the conditions
occurs at temperatures in the range of 10-30℃. The rate of
be controlled, but generally was in the range of 60—80% as recommended by

5.2.2 Atividade 12: Lexicogramática de “due to”

Atividade 12: Lexicogramática de “due to”

Analise as linhas de concordância abaixo.

1) Observe o tipo de informação que vem após “due to”. O que você conclui?

2) Que outra palavra, na língua inglesa, é sinônimo deste padrão?

3) Como este padrão pode ser traduzido para a Língua Portuguesa, a partir dos usos que o autor fez no artigo “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”*?

is the more attractive approach (when feasible) due to the lower shipping and handling cost. However,
cating that their biodegradation could be limited due to their strong absorption into the soil (Nelson
e type compound with a retention time of 11.5 min due to its strong adsorption to soil particles or low
t TPH removal after 180 d of monitoring, probably due to strong soil compaction and clay texture that l
hers common mechanisms of removal appear unlikely due to the nature of pollution. Volatile losses did n

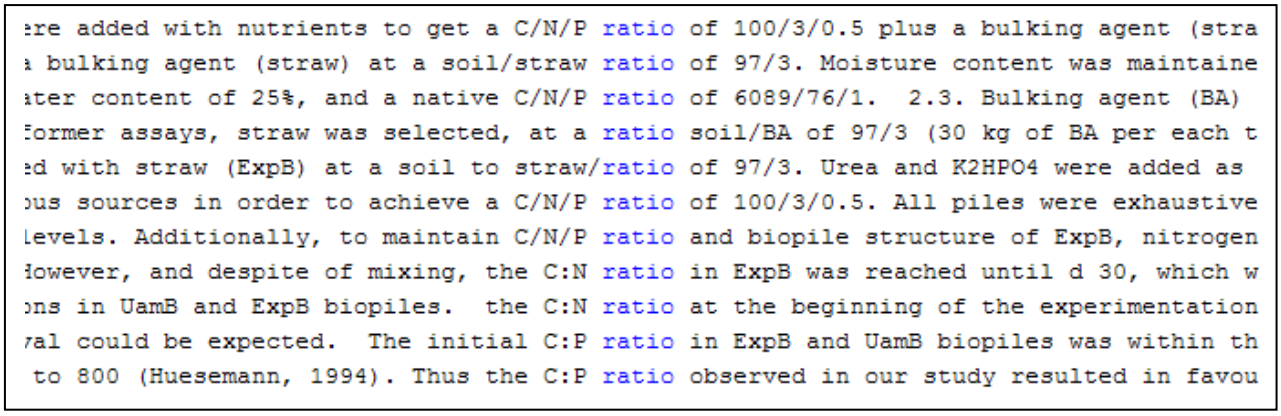
Observe outros padrões retirados do *corpus* de TA. Eles confirmam sua hipótese?

Concordance
removal from wastewaters has been considered difficult due to lack of available control parameters. It is known that
been absolutely no impact on the environment of Kudremukh due to large scale mining operations over the last two decades.
0.5 mg/l), a significant fraction of the floc mass may be anoxic due to limitation of oxygen diffusion within the loc, allowing
in dispersed cultures of bacteria in water and soil, probably due to limited oxygen difu-sion into the flocs. © 1999
piles had a higher temperature than the extremities, possibly due to local airflow rates and stimulated metabolism. However,
total fine particle exposure (10.7 /xgm-3) whereas exposure due to local traffic (including buses) and long-range transported
in relation to PM25, but in this study we considered mortality due to long-term exposure. Morbidity effects, such as lung
is sufficiently compelling to indicate that mortality risk due to long-term exposure would be expected, and
of pollutant and prevent proper air exchange. In fact due to low pollution level outdoor, the IO using any mechanical
, they concluded that the TPH removal in the test reactors was due to MF oxidation. This study indicated that CaO2 was a
to synthetic materials, while the better performances are due to mineral wools. However, the kenaf-fibre based products
on ore body and prevent erosion and wash-off. Dust generated due to mining operation is arrested by a tree belt so that the

5.2.3 Atividade 13: Lexicogramática de “ratio”

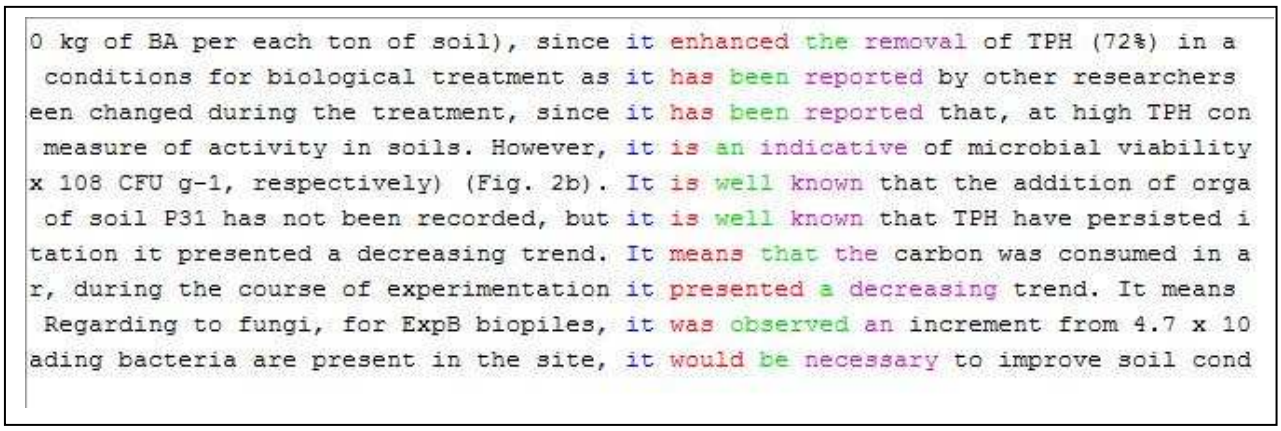
Atividade 13: Lexicogramática de “ratio”

No artigo “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”* a palavra “ratio” é muito repetida. Que termos se colocam junto com “ratio” nas linhas de concordâncias apresentadas abaixo? Que padrão você percebe neste extrato?



5.2.4 Atividade 14: Lexicogramática de “It ”

Atividade 14: Lexicogramática de “It ”



O pronome “it”, embora muito utilizado, tem seu sentido definido pelo seu cotexto (palavras que estão a sua volta) e juntos formam um pacote lexical. Abaixo estão listados alguns exemplos de “pacotes lexicais” formados com este pronome.

IT IS POSSIBLE TO
IT IS IMPORTANT TO
IT IS NECESSARY TO
IT IS OBSERVED THAT
IT IS OBVIOUS THAT

Uma consulta ao *corpus*, através do software concordanciador, apresenta muitas outras formações de padrões. Realize uma pesquisa para:

1) Encontrar dez outros padrões semelhantes, em que o pronome ‘it’ é utilizado.

2) Observar e descrever em que situações o autor os utiliza.

5.3 Introdução do software concordanciador para uso dos alunos

As atividades apresentadas até este ponto foram planejadas para serem realizadas com as concordâncias impressas, entregues aos alunos. Acredita-se que, pela quantidade de atividades apresentadas, caso o professor tenha seguido a sequência e as atividades tenham sido realizadas com os alunos, a partir deste ponto, provavelmente, os alunos já estejam habituados com a estrutura das concordâncias, sabendo lê-las. Em decorrência, já têm condições de interpretar os dados nelas contidos. Assim, o movimento em direção ao uso do software concordanciador diretamente com os alunos torna-se indispensável.

Ao utilizar o software concordanciador (Antconc ou Concord) os alunos terão uma liberdade muito maior para realizar suas pesquisas, para analisar e estudar a língua. Os exercícios impressos até aqui apresentados passaram pelo critério seletivo do professor. Conforme já pontuado, uma das riquezas do trabalho com os concordanciadores está justamente na possibilidade de os alunos manipularem diretamente os dados contidos no *corpus*, através da interface do software. Assim, como parte do programa do curso é necessário o professor realizar uma oficina de

introdução à utilização do software. A discussão sobre a introdução ao uso do software foge ao escopo deste estudo.

Embora, nas atividades seguintes, as linhas de concordâncias continuem sendo apresentadas, deve ficar claro que estão ali para ilustrar a atividade. A proposta é que os alunos realizem a pesquisa, interagindo diretamente com a interface do concordanciador adotado. Recomenda-se, para uso com alunos, o concordanciador Antconc, pois é gratuito e de utilização muito simples, não sendo sequer necessário instalá-lo no computador, pois é um software executável. Obviamente, em certos locais ou situações, o professor não terá acesso a computadores, mas, mesmo assim, poderá realizar a exploração da lexicogramática a partir de linhas de concordâncias impressas, como foi até aqui proposto. O professor linguista de *corpus* deve sempre priorizar a consulta diretamente aos dados da língua, estejam eles impressos ou acessíveis diretamente no computador. O mote é alterar a direção do ensino e romper com pressupostos baseados na intuição, acerca do funcionamento linguístico.

5.3.1 Atividades a serem realizadas com o software concordanciador

5.3.1.1 Atividade 15: Lexicogramática de “It has been”

Atividade 15: Lexicogramática de “It has been”

It has been...

As concordâncias abaixo foram retiradas do texto “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”*. Encontre essas estruturas no artigo e veja como são utilizadas.

nditions for biological treatment as it has been reported by other researchers (von Fahn
spectively (Fig. 3). A similar behavior has been reported by several authors (Kodres, 19
degradation. A wide range of C:P ratios has been reported for hydrocarbon degradation ra
changed during the treatment, since it has been reported that, at high TPH concentratio

1) Você conhece essa estrutura verbal?

2) Qual o sentido deste tempo verbal no texto?

3) Como você sintetizaria esse padrão? Veja logo abaixo outras concordâncias extraídas do *corpus* de TA contendo o mesmo padrão.

4) Realize você mesmo uma consulta ao *corpus* e detecte outros padrões possíveis

Concordance

of MSW (Pearson, 1996; Krochta et al., 1997). It has been estimated that 9.4 million tons of
stage, but there are numerous existing GCs. It has been estimated that the number of GCs
ions move to the positively charged anode. It has been experimentally proved that non-ionic
A. Review the Training matrix and ensure that it has been kept current with respect to any
nonparticle air pollution was not accounted for, as it has been minimal compared to the effects of fine
and cathode. After numerous experiments, it has been observed that the smaller the volume
range of 4.1-5.2 mg/l with 92% removal (average). It has been previously reported that SRT should be
and a base front will move out from the cathode. It has been proved by experiments that when

5.3.1.2 Atividade 16: Lexicogramática de “find/found”

Atividade 16: Lexicogramática de “find/found”

O verbo “to find” é muito utilizado em relatos de pesquisas e estudos. Segundo o dicionário, ele tem como principal tradução para a Língua Portuguesa “encontrar, achar, descobrir”.

1) Você concorda com essas possibilidades de tradução? De que forma você traduziria esse verbo nos exemplos abaixo, presentes no artigo “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”*?

stry. The objective of this work was to find a technology to remediate these sites.
evels of environmental impact have been found at the Southeast of Mexico as a result
d to 80 and 6000, respectively. We have found in the literature that a wide range of
94). In previous laboratory studies, we found than C:N and C:P ratios of 30 and 200,

1) Logo abaixo há uma seleção de linhas de concordância do *corpus* geral de TA com usos do verbo “to find”. Qual o padrão recorrente nas quatro seleções?

2) Em que situação os autores do artigo estão utilizando essa estrutura?

Concordance

l—1 at short sludge ages. It was also found out that many heterotrophic
on board properties. It was also found that composites with rHDPE
creating a layer of fresh air. It was also found that the levels are within 300
burden (Patel et al., 2000). It was also found that the energy consumed during
environmental regulations. It was also found that the lack of technical
pH rises from 4 to 7. It was also found that the GAC F-400 has a slightly
et al., 1994, 1995). It was also found that the weight loss percentages
open-air field [43]. Composting was also found to degrade 3- and 4-rings PAHs

Concordance

plants and soils have also been found in other studies (e.g., Markert,
, increasing soil moisture has been found to positively affect the removal of
treatment, as desorption rate has been found (Watts and Dilly, 1996) to be the
of environmental impact have been found at the Southeast of Mexico as a
million tonnes of primary ore have been found. Besides, there are other deposits
of problems that, on average, have been found in the audits are shown in Fig. 7.
Set 2. The pH, DO and ORP have been found to be very useful parameters for
in this study. Lichen species have been found to differ in their relative sensitivity
. This approach, however, has not been found to be a cost effective design for
facilities that disclose but have not been found to be in violation through an

Concordance

that facilities in [G.sub.2] that are found in compliance will transition to [G.
change. Facilities in [G.sub.2] that are found in compliance will move to [G.sub.
, especially submarines and ships, are found in excessive quantity near the
expansion rates ranging 1.52 ~ 1.95 are found to be too large to be realized only
are derived and wherever leaks are found to be more are being arrested on

Concordance

than external ones. External barriers found by Hillary are outlined in Table 26.
maintain high removal via nitrite. It is found that the monitoring and regulation
air intake and natural ventilation. It is found that the IO ratio is not specific
to different driving environment. It is found that the IO ratio is not specific
of abrupt heating in Figs. 9(d) ~ (f), it is found that the diameter at the point of
= 89.7 and 185???? respectively. It is found that the predicted unburnt rates
as two independent parameters. It is found that the temperature exhibits an
three pollutants at the tunnel site. It is found that the pollutant concentration
and flash gasification, it is found that the maximum increasing rate
travelling in a different environment. It is found that using fresh-air ventilation
different environment. For example, it is found that using fresh-air ventilation

5.3.1.3 Atividade 16: Lexicogramática de “however”

Atividade 16: Lexicogramática de “however”

HOWEVER,

e) due to the lower shipping and handling cost. However, inappropriate practices have generated soil po
4) and field scale (Iturbe et al., 2003, 2004). However, there is scarce information on the bioremediat
her researchers (von Fahnestock et al., 1998). However, the presence of high microbial counts (107CFUg
t is not a direct measure of activity in soils. However, it is an indicative of microbial viability or
ant since the beginning of the experimentation. However, and despite of mixing, the C:N ratio in ExpB w
700 which was attributed to soil heterogeneity. However, during the course of experimentation it presen
heterogeneity of the TPH pollution in soil P31. However, as can be noted by standard deviation and vari

Observe que a palavra HOWEVER em textos acadêmicos, predominantemente, é acompanha de uma vírgula e então o autor introduz outra informação.

1) Através da leitura dessas concordâncias, é possível perceber algum movimento (cognitivo) realizado pelo autor, em relação ao que foi dito anteriormente, quando foi introduzida a conjunção “however”? Isto é, que sentido ‘however’ introduz?

2) Como você descreve esse movimento?

3) Selecione no artigo, “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”* três exemplos que mostrem a movimentação cognitiva do autor. Esquematize seu argumento, indicando o tipo de transformação ocorrido frente a posição do autor ao que foi dito anteriormente (a however)?

4)Qual a importância deste vocábulo em artigos acadêmicos, considerando que sua frequência é alta. Consulte outros exemplos no *corpus*.

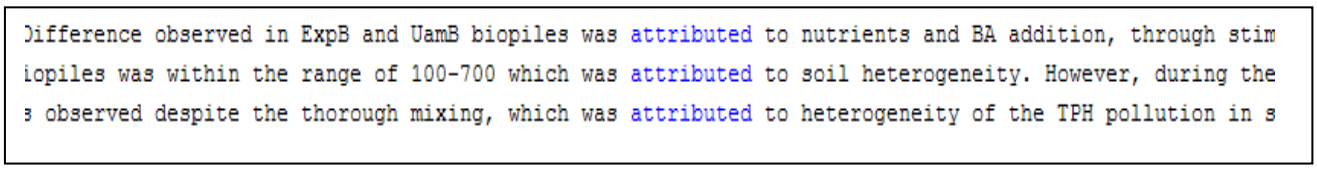
5.3.1.4 Atividade 17: Lexicogramática de “Attributed”

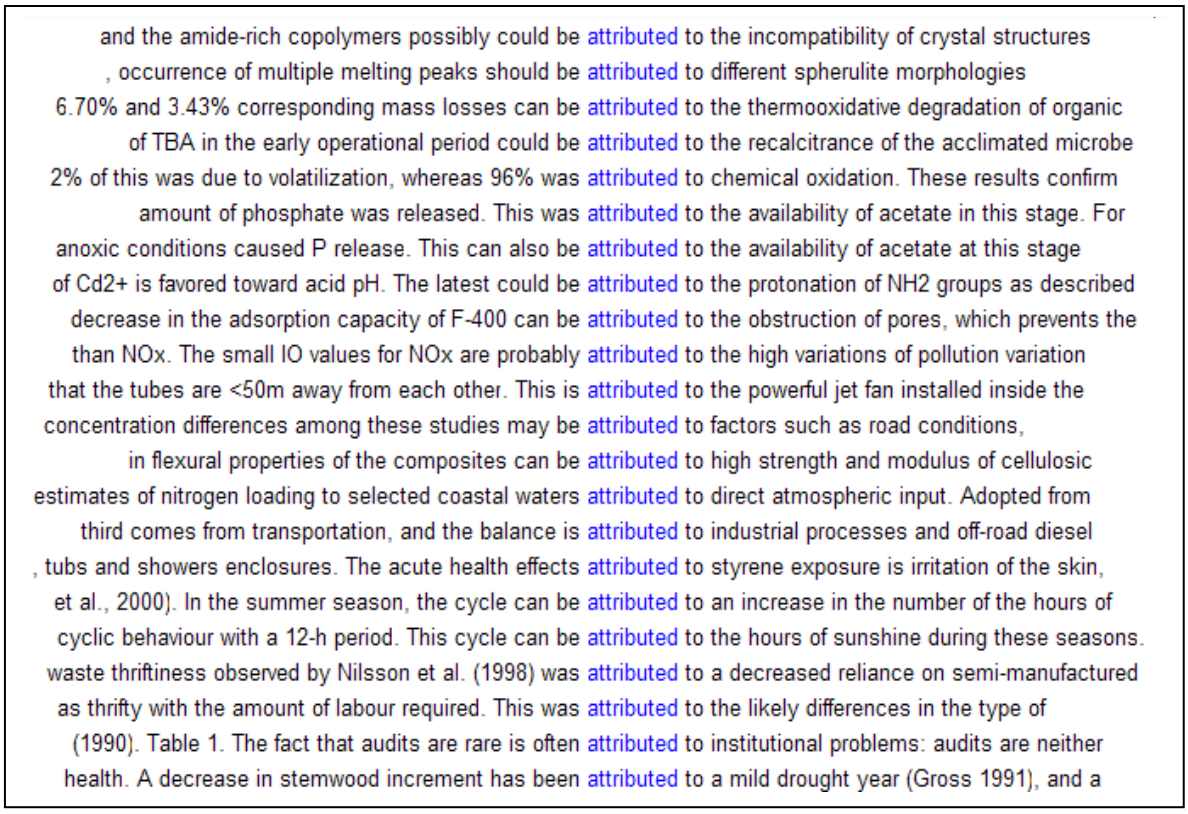
Atividade 17: Lexicogramática de “Attributed”

Attributed

A palavra “attributed” faz parte de uma estrutura verbal. No texto “*A field trial for an ex-situ bioremediation of a drilling mud-polluted site”* foi usada apenas no passado, conforme o excerto abaixo. No entanto, uma pesquisa no *corpus* de TA

trouxe vários resultados mostrando usos em diferentes tempos verbais. Há, no entanto, um padrão de uso. Identifique-o.





5.3.1.5 Atividade 18(a): Lexicogramática de palavras-chave da área de TA

Atividade 18(a): Lexicogramática de palavras-chave da área TA

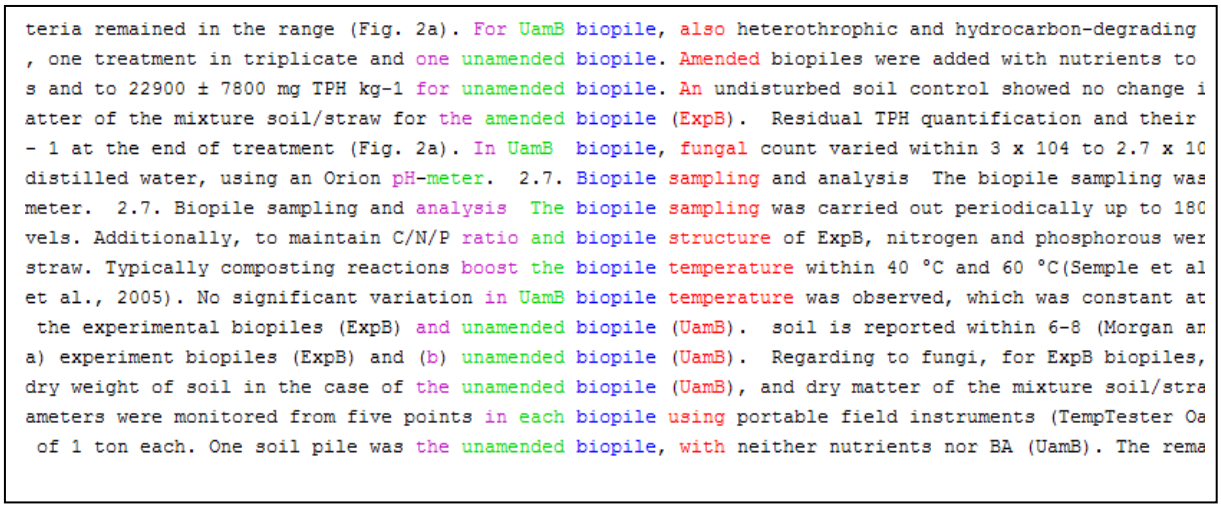
Lexical bundles with keywords

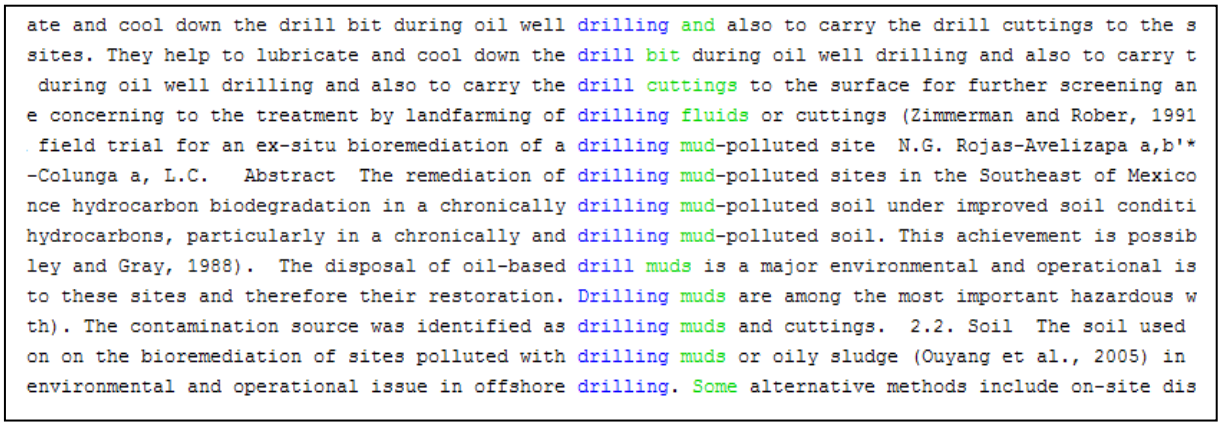
Os vocábulos centralizados (em azul), nas linhas das concordâncias abaixo, são palavras-chave presentes no artigo "*A field trial for an ex-situ bioremediation of a drilling mud-polluted site"* e foram detectadas pelo software Antconc.

1) Você concorda com a importância destes termos para a sua área de estudo?

2) Veja, se para cada uma dessas palavras-chave é possível reconhecer um padrão linguístico recorrente. Que padrões você destacaria?

3) Uma consulta ao *corpus* de Tecnologia Ambiental é recomendável e poderá fornecer mais dados.





5.3.1.6 Atividade 18(b): Lexicogramática de palavras-chave da área de TA

Atividade 18(b): Lexicogramática de palavras-chave da área de TA

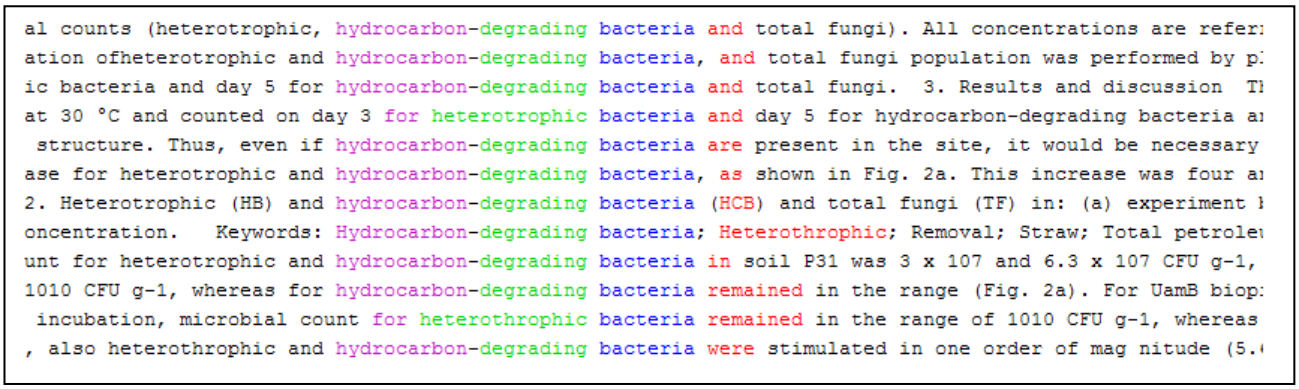
Lexical bundles with keywords

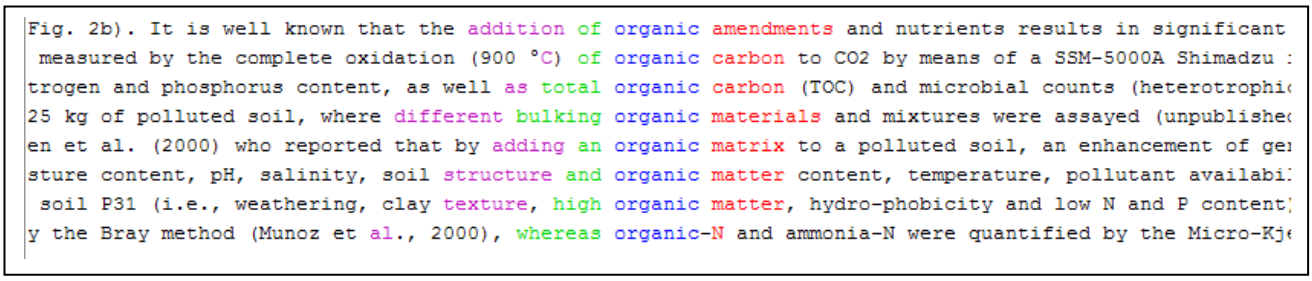
Os vocábulos centralizados (em azul), nas linhas das concordâncias abaixo, são palavras-chave presentes no artigo "*A field trial for an ex-situ bioremediation of a drilling mud-polluted site"* e foram detectadas pelo software Antconc.

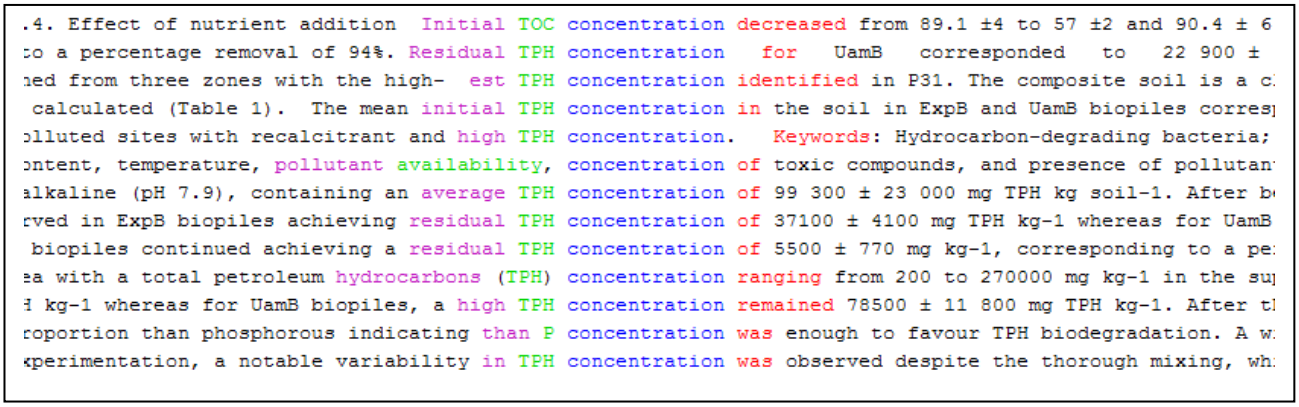
1) Você concorda com a importância destes termos para a sua área de estudo?

2) Veja, se para cada uma dessas palavras-chave, é possível reconhecer um padrão linguístico recorrente. Que padrões você destacaria?

3) Uma consulta ao *corpus* de Tecnologia Ambiental é recomendável e trará mais dados.







CONSIDERAÇÕES FINAIS

Neste ponto, faz-se o fechamento do trabalho, retomando os principais aspectos teóricos e metodológicos considerados. Em seguida, comentam-se os resultados e apresentam-se algumas críticas e limitações da presente investigação.

O objetivo deste estudo foi produzir material didático e atividades alternativas para o ensino de inglês instrumental para a área de Tecnologia Ambiental, utilizando-se para tanto o referencial teórico e técnico da Linguística de *Corpus*, em consonância com alguns princípios da Linguística Cognitiva. Percorreu-se uma jornada de mão-dupla, primeiramente, produzindo o *corpus* de estudo e realizando sua análise para, a partir daí, obter dados para a produção do material de ensino e desenvolvimento das tarefas. Inicialmente foram discutidas questões relativas ao *corpus* e, em seguida, à produção das tarefas e materiais de ensino.

A utilização de textos autênticos foi o primeiro critério estabelecido para a composição do *corpus* de estudos da área de Tecnologia Ambiental e foi, plenamente, alcançado. A estrutura e o tamanho do *corpus*, contendo 86 artigos acadêmicos e totalizando 450.565 *tokens*, permitiu que se respondesse a todas as questões de pesquisa. O fato de ser um *corpus* de tamanho pequeno (*small corpus*), além de especializado (Gavioli, 2005), por conter um único gênero discursivo e enfocar uma área teórica muito específica, parece ter facilitado a análise dos dados. Essa circunstância trouxe maior flexibilidade analítica, permitindo o uso de procedimentos automáticos e manuais para a o estudo da língua (lematização manual, por exemplo).

Quanto às perguntas de pesquisa, focalizaram o léxico da área de Tecnologia Ambiental. Mais precisamente, tais perguntas versaram sobre a presença de sequências formulaicas e de pacotes lexicais no *corpus* de Tecnologia Ambiental, além do conhecimento de vocábulos específicos desse campo.

A quantidade de termos específicos da área de Tecnologia Ambiental foi contrastada com o *corpus* BNC (British National *Corpus*). A comparação evidenciou que mais de 50% dos primeiros 200 itens mais frequentes eram palavras específicas

da área de Tecnologia Ambiental. A extração das palavras-chave também mostrou um alto índice de termos específicos do *Corpus* de Tecnologia Ambiental, o qual se constituiu de 3040 palavras-chave, quantidade significativa em relação ao tamanho total do *corpus*. O resultado obtido respondeu, positivamente, a uma das perguntas da pesquisa. A elaboração das listas de frequências, por sua vez, possibilitou que se determinasse com maior precisão o que incluir no programa de ensino e que se chegasse mais próximo daquilo que o aluno de ESP precisa, de fato, aprender.

As demais perguntas da pesquisa eram relativas à formação e quantidade de pacotes lexicais presentes no *corpus*. Tanto padrões linguísticos específicos da área de Tecnologia Ambiental, como pacotes lexicais que têm um uso mais amplo, isto é, que ocorrem também fora da área investigada, foram detectados. Em vista disso, o critério original que Biber et al. (1999) propõem para a extração dos pacotes lexicais de um *corpus* foi adaptado para a frequência mínima de 5 ocorrências no *corpus*, em pelo menos 3 textos diferentes, sendo cada *bundle* formado com o mínimo de 3 palavras e o máximo de 8. A partir da aplicação desse critério, a análise realizada com o Wordsmith Tools 5.0 obteve 2636 *clusters*, quantidade muito significativa em relação às dimensões do *corpus*. Além disso, procurou-se saber quantos pacotes lexicais continham palavras-chave, isto é, fez-se um levantamento, observando-se se cada pacote lexical identificado, tendo entre 3 e 8 palavras, continha pelo menos uma palavra-chave. O resultado indicou a presença de 2114 *clusters*, o que também é significativo.

Esses dados estatísticos atestam a alta frequência dos pacotes lexicais para a área de Tecnologia Ambiental, corroborando resultados apontados por outras pesquisas, citadas ao longo da dissertação, que afirmam que o índice de ocorrência das sequências formulaicas na língua é muito alto. Tais resultados indicam que as investigações sobre as co-ocorrências devem prosseguir, tanto no sentido de examinar com maior detalhe o comportamento dos padrões aqui elencados, como no de propor estudos semelhantes, a outras áreas do conhecimento. Seria interessante prosseguir a análise dos pacotes lexicais detectados no *corpus* de Tecnologia Ambiental, aprofundando a sua análise e verificando com maior precisão suas funções dentro do discurso acadêmico. Embora as quantidades identificadas na análise indiquem a abrangência dos pacotes lexicais no *corpus* de Tecnologia

Ambiental, talvez fosse útil encontrar um algoritmo preciso para calcular a percentagem de cada um dos pacotes lexicais, dentro do *corpus*.

Por fim, a última questão da pesquisa referia-se ao texto-chave. Conhecer o texto-chave do *corpus* era crucial, pois a partir dele, as atividades seriam desenhadas. Assim, além de ser representativo da linguagem contida no *corpus*, por apresentar um índice de cobertura de termos significativos estatisticamente calculados, ele também possuía uma função pedagógica importante: fora o texto escolhido para embasar a proposição de todas as atividades. Ou seja, esse texto permite que o aluno estabeleça a relação entre as amostras encontradas no *corpus* e o seu uso dentro do discurso. A ressalva é que o texto-chave selecionado não foi propriamente aquele com maior chavicidade, mas o que se classificou em segunda posição, cujo título é: “A Field Trial for an ex-site bioremedation of a drilling mud-polluted site”.

Uma das perguntas tidas em conta na elaboração das atividades dizia respeito a como vincular os fundamentos da Linguística Cognitiva com os fundamentos da Linguística de *Corpus*. O elemento comum foi a saliência do pacote lexical, que é crucial, sendo necessário marcá-lo para o aprendiz, tanto nas intervenções do professor, quanto na utilização de recursos tipográficos. Como já comentado, as sequências formulaicas e os pacotes lexicais não possuem um delimitador indicando seu início ou seu fim, tal como as palavras, sendo imprescindível introduzir atividades de iniciação para que o aprendiz de L2 aprenda a reconhecê-los no texto.

Por outro lado, imprimir as concordâncias favorece a leitura do termo pesquisado por colocá-lo em destaque, centralizado, e, ainda, com algum chamativo tipográfico. Ou seja, no desenrolar do trabalho com os alunos devem ser incluídas marcações tipográficas nos textos a serem lidos, bem como acrescentados enunciados de atividades que os provoquem a perceber os padrões linguísticos, a pensar a língua pelo viés da lexicogramática. Tais recursos foram utilizados na produção dos materiais e na execução das atividades e, o professor, em sua prática de aula, poderá auxiliar o aluno a perceber a função das sequências no discurso, até que consiga assimilá-las autonomamente. A instrução explícita deve ser utilizada quando necessário. Ela pode complementar as informações que os alunos por si próprios, em suas pesquisas, não conseguirem abstrair sobre o funcionamento, ou sobre o sentido da língua.

Já o uso da tecnologia em sala de aula, a partir da consulta ao *corpus*, em interface com o concordanciador, vai ao encontro do que preconiza a psicolinguística, para que haja a consolidação do aprendizado da língua: a exposição. Conforme visto, consultar o *corpus* possibilita exposição à língua em uso.

O uso de língua autêntica, segundo o "*usage based approach*", em situação real, ou o mais próxima possível da forma que o aluno usaria fora da sala de aula, parece contribuir também com o aprendizado. Esse ponto de vista é comum a linguistas de *corpus* e a psicolinguistas, conforme referido no capítulo dois. A escolha da "Abordagem baseada em tarefas" como uma metodologia para o desenvolvimento do ensino embasou-se nesse entendimento.

De outra parte, as tarefas apresentadas são ainda elementares e merecem aprimoramento, desde um melhor tratamento gráfico, até o desenvolvimento de muitas outras propostas, mais inovadoras, incentivando os alunos a fazerem uso real da língua e integrarem-na a diversas práticas da vida acadêmica. Mesmo com essas limitações, entende-se que o objetivo proposto foi plenamente alcançado.

Assim, este estudo buscou contribuir para a multiplicação das pesquisas de Linguística de *Corpus*, em especial, para a área de Linguística de *Corpus* e Ensino, descrevendo a utilização da Linguística de *Corpus* e de seus procedimentos e ferramentas no ensino de Língua Estrangeira. Almeja-se que os procedimentos sejam adaptados e utilizados em outros estudos, que levem a reformulações de práticas de ensino tradicionais e possibilitem que o conhecimento científico produzido na academia possa chegar aos alunos e contribuir, de fato, para a formação de leitores bilíngues proficientes.

REFERÊNCIAS

ADELE E. GOLDBERG; DEVIN CASENHISER. Construction Learning and Second Language Acquisition. In: ROBINSON, P.; ELLIS, N. C. **Handbook of Cognitive Linguistics and Second Language Acquisitions**. New York: Routledge, 2008.

ALISON WRAY; MICHAEL R. PERKINS. The Functions of formulaic language: an integrated model. **Language & Communication**, n. 20, p. 1-28, 2000.

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

BERBER SARDINHA, T. Preparação de material didátio para Aprendizagem Baseada em Tarefas com Wordsmith Tools e *corpora*. **Calidoscópio**, v. 4, n. 3, p. 148-155, set/dez 2006.

BERBER SARDINHA, T. **Pesquisa em Linguística de *Corpus* com Wordsmith Tools**. Campinas: Mercado de Letras, 2009.

BERBER SARDINHA, T. Como usar a Linguística de *Corpus* no ensino de Línguas estrangeiras. Ou por uma Linguística de *Corpus* educacional brasileira. In: VIANA, V.; TAGNIN, S. E. O. ***Corpora* no ensino de línguas estrangeiras.** São Paulo: Hub Editorial Ltda., 2011.

BERBER SARDINHA, T. Lexicogrammar. In: CHAPELLE, S. (.). **The Encyclopedia of Applied Linguistics**. Malden, CT: Wiley, no prelo.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at.: Lexical bundles in university teaching and textbooks. **Applied Linguistics**, v. 25, p. 371-405, 2004.

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

BISHOP, H. The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In: SCHMITT, N. **Formulaic Sequences:** acquisition, processing and use. Amsterdam: John Benjamin Publishing, 2004. p. 227-248.

BRITT ERMAN; BEATRICE WARREN. The Idiom Principle and the open choice principle. **Text**, v. 20, n. 1, p. 29-62, 2000.

BYBEE, J. Usage-based grammar and Second Language Acquistion. In: ELLIS, N. C.; ROBINSON, P. **Handobook of Cognitive Linguistics and Second Language Acquisition**. New York: Routledge, 2008. p. 216-236.

CARTER, R.; MCCARTHY, M. **The Cambridge Grammar of English:** a comprehensive guide : spoken and written English, grammar and usage. Cambridge: Cambridge University Press, 2006.

CHARTIER, R. **A aventura do livro:** do leitor ao navegador. São Paulo: Fundação Editora da UNESP, 1998.

CHRIS TRIBBLE; GLYN JONES. **Concordances in the classroom:** a resource book for teacher. Harlow: Longman Group UK, 1990.

COBB, T. **Why & how to use frequency lists to learn words**., 1997. Disponivel em: <www.lextutor.ca/research/>. Acesso em: nov. 2009.

COBB, T. Computing the vocabulary demands of L2 reading. **Language Learning & Technology**, 11, n. 3, 2007. 38-63.

COBUILD. **Collins cobuild dictionary.** London: HarperCollins, 2000.

COXHEAD, A. **An academic wordlist.** Occasional Publication Number 18, LALS, Victoria University of Wellington, New Zealand, 1998.

COXHEAD, A. **A new Academic Word List.** TESOL Quarterly, 34, 2, 212-238.
Nation, I. S. P.. *Learning Vocabulary in Another Language (Cambridge Applied Linguistics)*. 1 ed. New York: Cambridge University Press, 2001a. Print.

CRYSTAL, D. **Dicionário de Linguística e Fonética**. Rio de Janeiro: Jorge Zahar, 2000.

DE BEAUGRANDE, R. Large *corpora*, small *corpora*, and the learning of the "language". In: GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. ***Small corpus* Studies and ELT**. Amsterdam: John Benjamins Publishing, 2001. p. 3-30.

DOUGLAS BIBER ET AL. **Longman Grammar of Spoken and Written English**. Harlow: Pearson Education Limited, 1999.

ELLIS, Nick. C. Sequencing in SLA: phonological memory, chunking and points of order. **Studies in Second Language Acquisition**, 1996, 18. 91-126.

ELLIS, Nick. C. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In: Ellis, Nick C., and Peter Robinson. **Handbook of Cognitive Linguistics and Second Language Acquisition**. 1 ed. New York: Routledge, 2008b. Print. p. 372-405.

ELLIS, Nick C. Usage-based and form-focused SLA: The implicit and explicit learning of constructions. In: Tyler, Andrea; Kim, Yijoung; Takada, Mari (ed.) **Language in the context of use**: discourse and cognitive approaches do language. Mouton de Gruyter, Berlin, 2008a. p. 93-120.

ELLIS, Nick C., and Peter Robinson. **Handbook of Cognitive Linguistics and**

Second Language Acquisition. 1 ed. New York: Routledge, 2008. Print.

ERMAN, B. Cognitive processes as evidence of the idiom principle. **International Journal of *Corpus* Linguistics**, v. 12, n. 1, p. 25-53, 2007.

FLÔRES, O. C. Compreender e interpretar. In: (ORG.), O. C. F. **Linhas e entrelinhas:** leitura na sala de aula. Santa Cruz do Sul: EDUNISC, 2008. p. 26-48.

FLOWERDEW, J. Concordancing as a tool in course design. In: GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. ***Small corpus* Studies and ELT**. Amsterdam: John Benjamins Publishing, 2001. p. 71-92.

FRANKENBERG-GARCIA, A. Compilação e uso de *corpora* paralelos. In: TAGNIN, S. E. O.; VALE, O. A. **Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Humanitas, 2008. p. 117-136.

GASS, SUSAN M.; SELINKER, LARRY. **Second Language Acquisition:** an introductory course. New York: Routlede, 2008.

GAVIOLI, L. **Exploring *corpora* for ESP learning**. Amsterdam: John Benjamins Publishing, 2005.

GEOFFREY UNDERWOOD; NORBERT SCHMITT; ADAM GALPIN. The eyes have it: An eye-movement study into the processing of formulaic sequences. In: SCHMITT, N. **Formulaic Sequences:** acquisition, process and use. Amsterdam: John Benjamin Publishing, 2004. p. 153-172.

GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. ***Small corpus* Studies and ELT**. Amsterdam: John Benjamins Publishing, 2001.

GRANGER, S. **Learner English on Computer**. London: Longman, 1998.

HUDSON, Richard. Word Grammar, Cognitive Linguistics, ans second language learning and teaching. In: Ellis, Nick C., and Peter Robinson. **Handbook of Cognitive Linguistics and Second Language Acquisition**. 1 ed. New York: Routledge, 2008. Print. p. 89-113.

HUNSTON, S. ***Corpora* in Applied Linguistics**. Cambridge: Cambridge University Press, 2002.

HUNSTON, SUSAN; FRANCIS, GIL. **Pattern Grammar:** a *corpus* driven-approach to the lexical grammar od English. Amsterdam: John Benjamins Publishing , 1999.

HYLAND, K. **English for Academic Purposes:** an advanced resource book. New York: Routledge, 2006.

JOHNS, T. Should you be persuaded - two samples of data-driven materials. **ELR Journal**, Birmingham, v. 4, n. Scanned 2010 by M. Scott, p. 1-16, 1991.

JOHNS, T. From Printout to handout: Grammar and Vocabulary Teaching in the Context of Data-driven learning. In: ODLIN, T. (.). **Perspectives on Pedagogical Grammar**. Cambridge: Cambridge University Press, 1994. p. 27-45.

JONES, Martha; HAYWOOD, Sandra. **Facilitating the acquisition of formulaic sequences: an exploratory study in an EAP context**. In: SCHMITT, N. **Formulaic Sequences:** acquisition, processing and use. Amsterdam: John Benjamin Publishing, 2004. p. 269-292.

KENNEDY, G. **An Introduction to *Corpus* Linguistics**. London: Longman, 1998.
LANGACKER, R. W. Cognitive Grammar and Language Instruction. In: ROBINSON, P.; ELLIS, N. **Handbook of Cognitive Linguistics and Second Language Acquisition**. New York: Routledge, 2008. p. 66-88.

KUIPER, K. Formulaic performance in conventionalised varieties of speech. In: SCHMITT, N. **Formulaic Sequences:** acquisition, processing and use. Amsterdam: John Benjamin Publishings Company, 2004. p. 37-54.

LAUFER, B. The lexical plight in second language reading. In: COADY, J.; HUCKIN, T. **Second Language Vocabulary Acquisition:** a Rationale for Pedagogy. Cambridge: Cambridge University Press, 1997.

LAUFER, B.; NATION, P. Passive vocabulary size and speed of meaning recognition. **EUROSLA Yearbook** , n. 1, p. 7-28, 2001.

LEWIS, M. **The Lexical Approach**. Boston: Thomson-Heinle, 1993.

LEWIS, M. Pedagogical implications of the lexical approach. In: COADY, J.; HUCKLIN, T. **Second Language Vocabulary Acquisition**. Cambridge: Cambridge University Press, 1997. p. 255-270.

LONGMAN. **Longman dictionary of contemporary English.** 5. ed. Harlow: Pearson Longman, 2009.

MIKE SCOTT, CHRISTOPHER TRIBBLE. **Textual Patterns:** key words and *corpus* analysis in language education. Amsterdam: John Bejnamins Publishing, 2006.

NATION, Paul. **Learning Vocabulary in Another Language** *(Cambridge Applied Linguistics)*. 1 ed. New York: Cambridge University Press, 2001.

NATION, P. Using small *corpora* to investigate learner needs: two vocabulary research tool. In: GHADESSY, M.; HENRY, A. &. R. R. L. (. B. ***Small corpus* Studies and ELT:** theory and practice. Amsterdam: John Benjamin Publishing, 2001b.

NATION, Paul. Vocabulary. In: (ED), N. D. **Practical English Language Teaching**. New York: McGraw Hill, 2003. p. 129-152.

NATTINGER, JAMES R.; DECARRICO, JEANETTE S. **Lexical Phrases and Language Teaching**. Oxford: Oxford University Press, 1992.

NUNAN, D. **Task-based language teaching**. Cambridge: Cambridge University Press, 2004.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. **From *Corpus* to Classroom:** language use and language teaching. Cambridge: Cambridge University Press, 2007.

PAWLEY, A; SYDER, F. H. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J. C., Schmidt, R. W. (Eds.), **Language and Communication**. New York: Longman, pp. 191-226, 1983.

READ, J.; NATION, P. **Measurement of formulaic sequences**. [S.l.]: [s.n.].

RÖMER, U. The inseparability of lexis and grammar. **Annual Review of Cognitive Linguistics**, v. 7, p. 141-163, 2009.

SCHMITT, NORBERT; UNDERWOOD, GEOFFREY. Exploring the processing of formulaic sequences through a self-paced reading task. In: SCHMITT, N. **Formulaic Sequences:** acquisition, processing and use. Amsterdam: John Benjamin Publishing, 2004. p. 173-190.

SCHMITT, NORBERT; GRANDAGE, SARAH; ADOLPHS, SVENJA. Are *corpus*-derived recurrent clusters psycholinguistically valid? In: SCHMITT, N. **Formulaic Sequences:** acquisition, processing and use. Amsterdam: John Benjamin Publishing Company, 2004. p. 127-151.

SCHMITT, N. **Formulaic sequences:** acquisition, processing and use. Amsterdam: John Benjamin, 2004.

SCHMITT, N. **Vocabulary in Language Teaching**. Cambridge: Cambridge University Press, 2008.

SCHMITT, N.; CARTER, R. Formulaic Sequences in action: an introduction. In: SCHMITT, N. **Formulaic sequences:** acquisition, processing and use. Amsterdam: John Benjamin, 2004. p. 1-22.

SINCLAIR, J. ***Corpus*, Concordance, Collocation**. Oxford: Oxford University Press, 1997.

SINCLAIR, J. Preface. In: GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. ***Small corpus* Studies and ELT**. Amsterdam: John Benjamins Publishing, 2001. p. vii-xv.

SINCLAIR, J. *Corpus* and Text - Basic Principles. In: (ED), M. W. **Developing Linguistic *Corpora*:** a Guide to Good Practice. Oxford: Oxbow Books, 2004. p. 1-16. Disponível em http://ahds.ac.uk/creating/guides/linguistic-*corpora*/chapter1.htm, acesso em 06.01.10.

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

TRIBBLE, C. Genres, Keywords, teaching: towards a pedagogic account of the language pf project proposals.. In: BURNARD, L.; MCENERY, T. (. **Language Pedagogy from a *Corpus* Perspective**. Frankfurt: Peter Lang, 2000. p. 51-63.

VANPATTEN, Bill. Processing Matters in Input Enhancement. In.: Piske, Thorsten , and Martha Young-Scholten (eds.). **Input Matters in SLA (Second Language Acquisition)**. Clevedon: Multilingual Matters, 2008. Print.

WARREN, B.; ERMAN, B. The Idiom Principle and the Open Choice Principle. **Text**, v. 20, n. 1, p. 29-62, 2000.

WILLIS, D. **The Lexical Syllabus**. London: Colins ELT, 1990.

WILLIS, D. **Rules, Patterns and words:** Grammar and Lexis in English Language Teaching. Cambridge: Cambridge University Press, 2009.

WILLIS, D.; WILLIS, J. **Doing Task-based teaching**. Oxford: Oxford University Press, 2007.

WOOD, D. **Formulaic Language in Acquisition and Production**: implications for teaching., v. 20, n. 1, p. 1-15, 2002.

WRAY, A. **Formulaic Sequences in Second Language Teaching**: Principle and Practice, 2000.

WRAY, A. **Formulaic Language and the Lexicon**. Cambridge: Cambridge University Press, 2002.

WRAY, A.; PERKINS, M. R.. The Functions of Formulaic Language: an integrated model. **Language & Communication** , n. 20, p. 1-28, 2000.

ANEXOS

ANEXO A – TEXTO CHAVE SELECIONADO

Available online at www.sciencedirect.com
ScienceDirect
Chemosphere 66 (2007) 1595–1600
CHEMOSPHERE
www.elsevier.com/locate/chemosphere

A field trial for an *ex-situ* bioremediation of a drilling mud-polluted site

N.G. Rojas-Avelizapa [a,b,*], T. Roldán-Carrillo [a], H. Zegarra-Martínez [a], A.M. Muñoz-Colunga [a], L.C. Fernández-Linares [a]

[a] *Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas 152, Col. San Bartolo Atepehuacan, 07360 México D.F., Mexico*
[b] *Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada del IPN, José Siurob 10, Colonia Alameda, Santiago de Querétaro, Querétaro, Mexico*

Received 31 May 2006; received in revised form 3 August 2006; accepted 7 August 2006
Available online 25 September 2006

Abstract

The remediation of drilling mud-polluted sites in the Southeast of Mexico is a top priority for Mexican oil industry. The objective of this work was to find a technology to remediate these sites. A field trial was performed by composting in biopiles, where four 1 ton soil-biopiles were established, one treatment in triplicate and one unamended biopile. Amended biopiles were added with nutrients to get a C/N/P ratio of 100/3/0.5 plus a bulking agent (straw) at a soil/straw ratio of 97/3. Moisture content was maintained around 30–35%. Results showed that, after 180 d, total petroleum hydrocarbon (TPH) concentrations decreased from 99300 ± 23000 mg TPH kg^{-1} soil to 5500 ± 770 mg TPH kg^{-1} for amended biopiles and to 22900 ± 7800 mg TPH kg^{-1} for unamended biopile. An undisturbed soil control showed no change in TPH concentrations. Gas chromatographic analysis showed residual alkyl dibenzothiophene type compounds. Highest bacterial counts were observed during the first 30 d which correlated with highest TPH removal, whereas fungal count increased at the end of the experimentation period. Results suggested an important role of the straw, nutrient addition and water content in stimulating aerobic microbial activity and thus hydrocarbon removal. This finding opens an opportunity to remediate old polluted sites with recalcitrant and high TPH concentration.

Keywords: Hydrocarbon-degrading bacteria; Heterothrophic; Removal; Straw; Total petroleum hydrocarbons

1. Introduction

A large number of polluted sites with different levels of environmental impact have been found at the Southeast of Mexico as a result of more than 60 years of the oil industry activities (Adams et al., 1999). Most of these sites are located in marshes or flooded zones, situation that complicate the entrance to these sites and therefore their restoration. Drilling muds are among the most important hazardous wastes released to the environment at these sites. They help to lubricate and cool down the drill bit during oil well drilling and also to carry the drill cuttings to the surface for further screening and disposal. These hazardous wastes are produced by emulsifying oil and water with colloidal matter that may include organophilic clay, bentonite, barite and other additives (Darley and Gray, 1988).

The disposal of oil-based drill muds is a major environmental and operational issue in offshore drilling. Some alternative methods include on-site disposal, ship to shore and dispose or reinjection in abandoned wells or subsurface caverns. Only two reports are available in the literature concerning to the treatment by landfarming of drilling fluids or cuttings (Zimmerman and Rober, 1991; Lee et al., 2002). Nevertheless, the on-site disposal is the more attractive

* Corresponding author. Address: Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada del IPN, José Siurob 10, Colonia Alameda, Santiago de Querétaro, Querétaro, Mexico. Tel.: +52 442 2124111x216; fax: +52 442 2124111x103.
E-mail address: nrojasa@ipn.mx (N.G. Rojas-Avelizapa).

0045-6535/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2006.08.011

approach (when feasible) due to the lower shipping and handling cost. However, inappropriate practices have generated soil pollution and potential risks to human health.

In the last years, several research centers in Mexico, including the Mexican petroleum corporation and the Mexican Institute of Petroleum have studied effective technologies in terms of cost-benefit, to restore the most problematic and hydrocarbon polluted sites in the country, particularly at the states of Veracruz, Tabasco, Chiapas and Campeche. Treatability studies have demonstrated the feasibility and potential application of bioremediation processes to restore hydrocarbon-polluted areas at laboratory (Rojas-Avelizapa et al., 2003; Molina-Barahona et al., 2004; Rivera-Cruz et al., 2004) and field scale (Iturbe et al., 2003, 2004). However, there is scarce information on the bioremediation of sites polluted with drilling muds or oily sludge (Ouyang et al., 2005) in spite of the well-proven toxicity and negative effect on the environment (Holdway, 2002).

Previous studies performed in our laboratory showed that hydrocarbons present in the target site (Paredon 31, P31) corresponded to C_{10}–C_{21} aliphatic compounds, alkyl polycyclic aromatic hydrocarbons (PAHs) and alkyl-substituted sulfur polycyclic aromatic compounds (PAS) (Arce-Ortega et al., 2004) indicating that their biodegradation could be limited due to their strong absorption into the soil (Nelson et al., 1996; Hwang and Cutright, 2002) and their chemical structure. Thus, even if hydrocarbon-degrading bacteria are present in the site, it would be necessary to improve soil conditions and to increase their biological removal. Additional soil characteristics such as nutrimental status, texture, hydrophobicity and low air diffusion can also limit the degradation of such compounds.

Consequently, the objective of the present study was to conduct a field trial of bioremediation by composting in biopiles to stimulate and enhance hydrocarbon biodegradation in a chronically drilling mud-polluted soil under improved soil conditions (C/N/P, bulking agent, moisture, etc.), which were previously assayed and established in our laboratory.

2. Materials and methods

2.1. Polluted site

The polluted site is a mudpit located near an oil well drilled in the early 1970s, located at the Mexican State of Tabasco. The average environmental temperature observed during the experimentation period in Tabasco was within 26 °C and 35 °C. The mudpit is known as Paredon 31 (P31) and is a 7 ha area with a total petroleum hydrocarbons (TPH) concentration ranging from 200 to 270000 mg kg^{-1} in the superficial level (0–60 cm depth). The contamination source was identified as drilling muds and cuttings.

2.2. Soil

The soil used for this work (4 ton) is a composite of three soil samples obtained from three zones with the highest TPH concentration identified in P31. The composite soil is a clay loam and slightly alkaline (pH 7.9), containing an average TPH concentration of $99\,300 \pm 23\,000$ mg TPH kg $soil^{-1}$. After being excavated and exposed to air, the soil has a water content of 25%, and a native C/N/P ratio of 6089/76/1.

2.3. Bulking agent (BA) selection

The selection of BA was performed in previous treatability tests using 1.9 and 25 kg of polluted soil, where different bulking organic materials and mixtures were assayed (unpublished data). Based on these former assays, straw was selected, at a ratio soil/BA of 97/3 (30 kg of BA per each ton of soil), since it enhanced the removal of TPH (72%) in a polluted soil containing 134000 mg TPH kg^{-1}. Additionally, straw prevents soil compaction over a long period of time.

2.4. Treatment facilities

Unamended (UamB) and amended (ExpB) biopiles were established in a roof protected and asphalted enclosure, where four cement containers (2 m wide × 3 m large × 20 cm high) were constructed with a separation of 1 m among each one. Sand was placed into the containers with a 5° inclination; in the front of container a pipeline system (PVC) was adapted to collect possible leachates. A PVC impermeable geomembrane was placed above the sand to support and avoid the migration of the material.

2.5. Field implementation of biopiles

The composite soil was divided into four piles of 1 ton each. One soil pile was the unamended biopile, with neither nutrients nor BA (UamB). The remaining three biopiles were amended with straw (ExpB) at a soil to straw/ratio of 97/3. Urea and K_2HPO_4 were added as nitrogen and phosphorous sources in order to achieve a C/N/P ratio of 100/3/0.5. All piles were exhaustively mixed using shovels to achieve homogeneous material and then placed over the geomembrane, supported by the cement containers. They were covered afterwards with the geomembrane to maintain temperature and moisture constant.

During the experimentation period, moisture content was adjusted at 30–35%. Previous studies demonstrated that the optimal moisture for TPH removal was within this range (Roldan et al., 2003). Non excavated soil in P31 was used as the undisturbed control.

2.6. Monitoring and maintenance

Daily temperature and moisture parameters were monitored from five points in each biopile using portable field instruments (TempTester Oakton and Delmhorst KS-D1, respectively). All biopiles (UamB and ExpB) were turned and watering as necessary based on *in situ* monitoring of

the moisture levels. Additionally, to maintain C/N/P ratio and biopile structure of ExpB, nitrogen and phosphorous were added at d 30 and 90, and BA at d 90 and 150 (1.5% w/w, respectively).

The pH in soil samples was measured in a suspension of 1 g of soil in 9 ml of distilled water, using an Orion pH-meter.

2.7. Biopile sampling and analysis

The biopile sampling was carried out periodically up to 180 d. Each time, 30 soil/compost samples per pile were withdrawn from different locations from the upper layer (5–10 cm depth), in order to determine the overall behaviour and properties of the piles. The samples were transported in sterile dark bottles and preserved at 4 °C in a laboratory refrigerator for further analysis. Several analysis of each sample were performed to determine the TPH content, pH, moisture, nitrogen and phosphorus content, as well as total organic carbon (TOC) and microbial counts (heterotrophic, hydrocarbon-degrading bacteria and total fungi). All concentrations are referred to the dry weight of soil in the case of the unamended biopile (UamB), and dry matter of the mixture soil/straw for the amended biopile (ExpB).

Residual TPH quantification and their identification were performed by GC–MS according to Arce-Ortega et al. (2004). TOC was measured by the complete oxidation (900 °C) of organic carbon to CO_2 by means of a SSM-5000A Shimadzu infrared spectrophotometer. The phosphorous available in each sample was determined by the Bray method (Muñoz et al., 2000), whereas organic-N and ammonia-N were quantified by the Micro-Kjeldahl method (AOAC, 1970).

Enumeration of heterotrophic and hydrocarbon-degrading bacteria, and total fungi population was performed by plate-count method; in selective media according to Alef and Nannipieri (1995) and expressed as colony-forming units per g of dry soil or matter (CFU g^{-1}). The plates were incubated at 30 °C and counted on day 3 for heterotrophic bacteria and day 5 for hydrocarbon-degrading bacteria and total fungi.

3. Results and discussion

The TPH biodegradation in soil depends on several factors, such as oxygen and nutrient availability, moisture content, pH, salinity, soil structure and organic matter content, temperature, pollutant availability, concentration of toxic compounds, and presence of pollutant-degrading microorganisms (Walworth and Reynolds, 1995; Penberthy and Weston, 2000; Margesin and Schinner, 2001). According to previous information, some properties of soil P31 (i.e., weathering, clay texture, high organic matter, hydrophobicity and low N and P content) could be considered as non appropriate conditions for biological treatment as it has been reported by other researchers (von Fahnestock et al., 1998). However, the presence of high microbial counts (10^7 CFU g^{-1}) suggests the potential application of bioremediation if other physicochemical parameters could be overcome. Therefore, the bioremediation of P31 becomes a challenge which would increase our knowledge and raise new approaches for *in-situ* and *ex-situ* treatments.

3.1. Moisture and temperature monitoring

During the first 15 d of experimentation, the moisture content was relatively low (23–25%). After this period of time, enough water was incorporated to adjust and maintain the desirable range (30–35%) up to the end of the experimentation period. The hydrophobic nature of the soil might have been changed during the treatment, since it has been reported that, at high TPH concentrations, the polluted soil to water repellency increases (Callaham et al., 2002).

The temperature profile for ExpB biopiles after 24 h of experimentation showed a distinct increase of pile temperature, which reached a maximum of 45 °C at d 15 (Fig. 1). The temperature rise was related with the addition of the straw. Typically composting reactions boost the biopile temperature within 40 °C and 60 °C (Semple et al., 2001; Ouyang et al., 2005). The temperature in ExpB decreased to 30 °C between d 30 and 45. Afterwards, the addition of more BA caused only a slightly increase in temperature (<2 °C). Others authors have reported a significant increase of temperature after the second addition of manure (Ouyang et al., 2005). No significant variation in UamB biopile temperature was observed, which was constant at 30 °C during the whole experimentation period.

3.2. pH

The pH of ExpB remained within the range of 7.5–8 throughout the 180 d of experimentation, thus suggesting the presence of appropriate biodegradation conditions. The optimum pH range for hydrocarbon degradation in

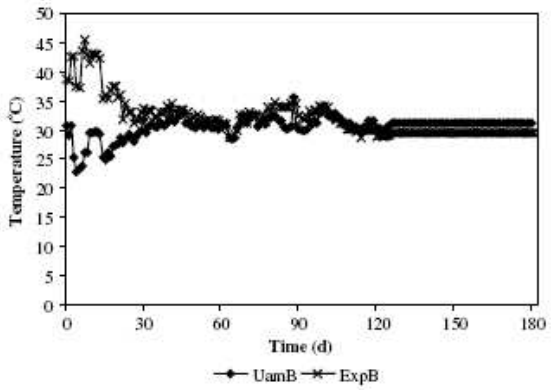


Fig. 1. Temporal variation of temperature in the experimental biopiles (ExpB) and unamended biopile (UamB).

soil is reported within 6–8 (Morgan and Watkinson, 1989; Cunningham and Philip, 2000). In the case of UamB biopile, no significant changes were observed during the whole experimentation time. The pH was maintained at 7.9.

3.3. Microbial counts

Microbial count is not a direct measure of activity in soils. However, it is an indicative of microbial viability or biodegradation potential in a polluted soil (Bossert and Kosson, 1997). The native population count for heterotrophic and hydrocarbon-degrading bacteria in soil P31 was 3×10^7 and 6.3×10^7 CFU g^{-1}, respectively, suggesting that bioremediation could be feasible (Mishra et al., 2001).

During the d 1, the microbial count in ExpB biopiles presented an increase for heterotrophic and hydrocarbon-degrading bacteria, as shown in Fig. 2a. This increase was four and three orders of magnitude higher (2×10^{11} and 5.5×10^{10} CFU g^{-1}, respectively) than that observed for native soil microbial population (3×10^7 and 6.3×10^7 CFU g^{-1}). In general during the entire incubation, microbial count for heterothrophic bacteria remained in the range of 10^{10} CFU g^{-1}, whereas for hydrocarbon-degrading bacteria remained in the range of 10^9 CFU g^{-1} (Fig. 2a). For UamB biopile, also heterothrophic and hydrocarbon-degrading bacteria were stimulated in one order of magnitude (5.6×10^8 and 9×10^8 CFU g^{-1}, respectively) (Fig. 2b). It is well known that the addition of organic amendments and nutrients results in significant changes in the number of microorganisms (Song and Bartha, 1990). This behavior was clearly observed after 1 d of treatment in our study (Fig. 2a) where microbial counts in Exp B presented a 3- or 4-fold increment. These agree with Jorgensen et al. (2000) who reported that by adding an organic matrix to a polluted soil, an enhancement of general microbial counts and also the activity of specific degraders was observed.

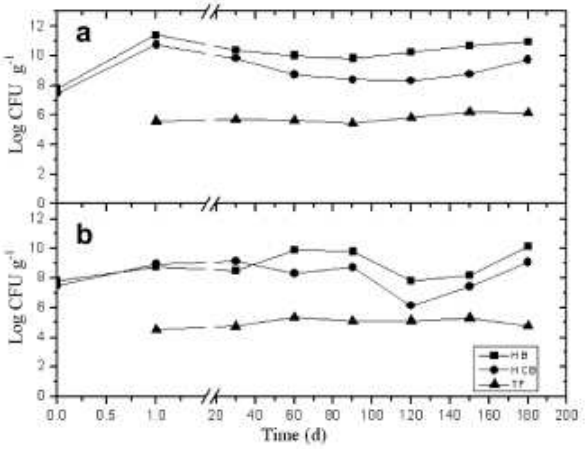


Fig. 2. Heterotrophic (HB) and hydrocarbon-degrading bacteria (HCB) and total fungi (TF) in: (a) experiment biopiles (ExpB) and (b) unamended biopile (UamB).

Regarding to fungi, for ExpB biopiles, it was observed an increment from 4.7×10^5 CFU g^{-1} (d 1) to 2×10^6 CFU g^{-1} at the end of treatment (Fig. 2a). In UamB biopile, fungal count varied within 3×10^4 to 2.7×10^5 CFU g^{-1} (Fig. 2b). Difference observed in ExpB and UamB biopiles was attributed to nutrients and BA addition, through stimulation of microbial growth and the contribution of native microflora present in the BA.

3.4. Effect of nutrient addition

Initial TOC concentration decreased from 89.1 ± 4 to 57 ± 2 and 90.4 ± 6 to 63.5 ± 6 g TOC kg^{-1} for ExpB and UamB, respectively. Two stages were identified during its consumption: (i) higher consumption rates at d 0–120 with 275 and 190 mg TOC kg^{-1} d^{-1} for ExpB and UamB, respectively, and (ii) low rates at d 120–180 with 75 and 42 mg TOC kg^{-1} d^{-1} for ExpB and UamB, respectively (Fig. 3). A similar behavior has been reported by several authors (Kodres, 1999; Cunningham and Philip, 2000; Gray et al., 2000; Jensen et al., 2000; Al-Daher et al., 2001; Zytner et al., 2006) pointing out that biodegradation of lighter molecular weight compounds occurs during the first d of treatment.

Nitrogen and phosphorous are essential macronutrients for microbial growth and metabolic activity, but hydrocarbon-polluted soils are often nutrient limited (Brook et al., 2001). In our study, the native C:N and C:P ratios of soil P31 corresponded to 80 and 6000, respectively. We have found in the literature that a wide range of C:N and C:P ratios have been reported, e.g., 9–200 and 60–800, respectively, for hydrocarbon biodegradation (Huesemann, 1994). In previous laboratory studies, we found than C:N and C:P ratios of 30 and 200, respectively, for soil P31 were optimal to enhance TPH removal (Roldan et al., 2003). Therefore, in this study, we tried to adjust and maintain the C:N and C:P ratios constant since the beginning of the experimentation. However, and despite of mixing, the C:N ratio in ExpB was reached until d 30, which was maintained within 30 and 40 up to the end of experimentation. This fact can be related to the recycling of N by microorganisms, together with a decrease of TOC in ExpB. For UamB,

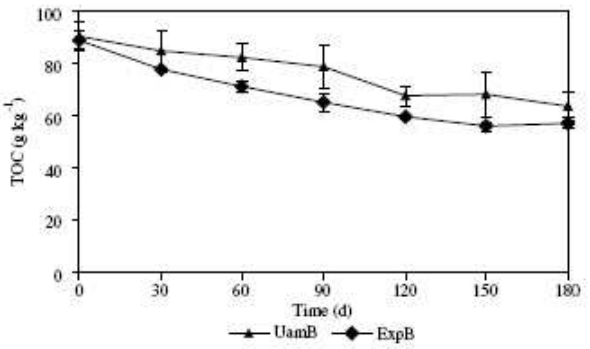


Fig. 3. Reduction in TOC concentrations in UamB and ExpB biopiles.

the C:N ratio at the beginning of the experimentation was 80, decreasing to 45 at the end of treatment, indicating that microbial activity takes place at native nitrogen content. Therefore, TPH removal could be expected.

The initial C:P ratio in ExpB and UamB biopiles was within the range of 100–700 which was attributed to soil heterogeneity. However, during the course of experimentation it presented a decreasing trend. It means that the carbon was consumed in a higher proportion than phosphorous indicating than P concentration was enough to favour TPH biodegradation. A wide range of C:P ratios has been reported for hydrocarbon degradation ranging from 60 to 800 (Huesemann, 1994). Thus the C:P ratio observed in our study resulted in favourable conditions to support TPH biodegradation process.

3.5. Hydrocarbon removal

The pollution history of soil P31 has not been recorded, but it is well known that TPH have persisted in the soils for many years. During the treatment by composting in biopiles, the residual TPH in soil were measured each 30 d and the reduction of TPH in ExpB and UamB is shown in Table 1. As can be observed therein, at the beginning of experimentation, a notable variability in TPH concentration was observed despite the thorough mixing, which was attributed to heterogeneity of the TPH pollution in soil P31. However, as can be noted by standard deviation and variation coefficient (Table 1) the heterogeneity decreases through the experimentation period. To have a consistent comparison of the removal between the biopiles, the percentage of TPH removal was also calculated (Table 1).

The mean initial TPH concentration in the soil in ExpB and UamB biopiles corresponded to 99300 ± 23000 mg TPH kg^{-1}. After 30 d of treatment an extensive hydrocarbon removal was observed in ExpB biopiles achieving residual TPH concentration of 37100 ± 4100 mg TPH kg^{-1} whereas for UamB biopiles, a high TPH concentration remained 78500 ± 11800 mg TPH kg^{-1}. After that time, the TPH removal for ExpB biopiles continued achieving a residual TPH concentration of 5500 ± 770 mg kg^{-1}, corresponding to a percentage removal of 94%. Residual TPH concentration for UamB corresponded to 22900 ± 7800 mg kg^{-1} or 77% of removal at the end of 180 d of experimentation. In the last case, the drainage of excavated soil and the resulting aeration might have been sufficient to permit substantial hydrocarbon removal.

Gas chromatographic analysis of TPH at the beginning of treatment evidenced that aliphatic hydrocarbons prevail in soil at an extent of 80% followed by PAHs at 15% and PAS at 5%. Amended biopiles (ExpB) were able to remove most of those recalcitrant compounds (Fig. 4), remaining an alkyl dibenzothiophene type compound with a retention time of 11.5 min due to its strong adsorption to soil particles or low solubility in aqueous media then limited biodegradation (Lindstrom and Braddock, 2002).

Additional data showed that undisturbed and non-excavated soil control did not show a significant TPH removal after 180 d of monitoring, probably due to strong soil compaction and clay texture that limit aerobic conditions (Fernández et al., 2004). Others common mechanisms of removal appear unlikely due to the nature of pollution. Volatile losses did not appear to be the major mechanism, as the biopiles were covered. Leaching of hydrocarbons was not substantial, because the biopiles were covered and no leachate was observed after construction of the biopiles. Therefore, the TPH removal in ExpB and UamB biopiles was mostly atributted to biodegradation processes.

Results demonstrated that the application of composting in biopiles is feasible in a field trial for enhancing the removal of recalcitrant hydrocarbons, particularly in a chronically and drilling mud-polluted soil. This achievement is possible based on the through balance of nutrimental status, the improvement of soil aeration, the watering maintainance, and the addition of an appropriate BA.

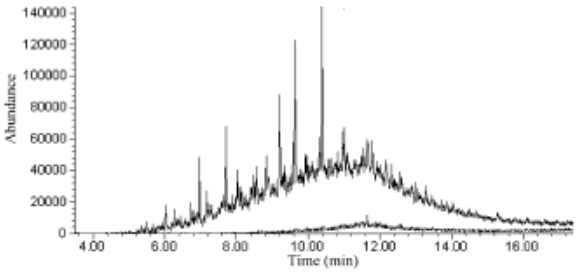


Fig. 4. Chromatogram for the samples collected from ExpB on d 1 and 180.

Table 1
Time-dependent variations of the TPH content in biopiles

Days	UamB			ExpB		
	mg TPH kg^{-1}	Variation (%)	Removal (%)	mg TPH kg^{-1}	Variation (%)	Removal (%)
0	99300	23	0	99300	23	0
30	78500	15	21	37100	11	63
60	65400	8	34	25500	16	74
90	54500	24	41	14000	36	85
120	33800	24	66	7100	34	92
150	26300	43	73	5300	11	94
180	22900	33	77	5500	14	94

Acknowledgements

The authors thank the many people who assisted in the field experiments helping to construct and maintain the biopiles at Paredon 31. This research was supported by the Grant D.00023 Atenuación Natural de Suelos Contaminados financed by the Mexican Institute of Petroleum.

References

Adams, R., Domínguez, V.I., García, L., 1999. Bioremediation potential of oil impacted soil and water in the Mexican tropics. Terra 117, 159–174.

Al-Daher, R., Al-Awadhi, N., Yateem, A., Balba, M.T., 2001. Compost soil piles for treatment of oil polluted soil. Soil Sediment Contam. 10, 197–209.

Alef, K., Nannipieri, P., 1995. Methods in Applied Soil Microbiology and Biochemistry. Academic Press, San Diego California.

AOAC, 1970. Official Methods of Analysis. Association of Official Analytical Chemists, Washington DC.

Arce-Ortega, J.M., Rojas-Avelizapa, N.G., Rodríguez-Vázquez, R., 2004. Identification of recalcitrant hydrocarbons present in a drilling mud-polluted soil. J. Environ. Sci. Health A 39, 1535–1545.

Bossert, I.D., Kosson, D.S., 1997. Methods for measuring hydrocarbon biodegradation in soils. In: Hurst, C.J., Knudsen, G.R., McInerney, M.J., Stetzenbach, M.V. (Eds.), Manual of Environmental Microbiology. ASM Press, Washington, DC, pp. 738–745.

Brook, T.R., Stiver, W.H., Zyther, R.G., 2001. Biodegradation of diesel fuel in soil under various nitrogen addition regimes. Soil Sediment Contam. 10, 539–553.

Callaham, M.A., Stewart, A.J., Alarcon, C., McMillen, S.J., 2002. Effect of earthworm (*Eisenia fetida*) and wheat (*Triticum aestivum*) straw additions on selected properties of petroleum-polluted soils. Environ. Toxicol. Chem. 21, 1658–1663.

Cunningham, C.J., Philip, J.C., 2000. Comparison of bioaugmentation and biostimulation in *ex situ* treatment of diesel polluted soil. Land Contam. Reclam. 8, 261–269.

Darley, H.C.H., Gray, G.R., 1988. Composition and Properties of Drilling and Completion Fluids. Gulf Professional Publishing, Houston TX.

Fernández, L.L.C., Ramírez, I.M.E., Rojas, A.N.G., Roldán, C.T.G., Zegarra, M.H.G., 2004. Final Report: Atenuación Natural de Suelos Contaminados con Hidrocarburos. Instituto Mexicano del Petróleo. ©2004 IMP.

Gray, M.R., Banerjee, D.K., Dudas, M.J., Pickard, M.A., 2000. Protocols to enhance biodegradation of hydrocarbon contaminants in soil. Bioremed. J. 4, 249–257.

Holdway, D.A., 2002. The acute and chronic effects of muds associated with offshore oil and gas production on temperate and tropical marine ecological processes. Mar. Pollut. Bull. 44, 185–203.

Huesemann, M.H., 1994. Guidelines for land-treating petroleum hydrocarbon-polluted soils. J. Soil Contam. 3, 1–204.

Hwang, S., Cutright, T.J., 2002. Impact of clay minerals and DOM on the competitive sorption/desorption of PAHs. Soil Sediment Contam. 11, 269–291.

Iturbe, R., Flores, R.M., Torres, L.G., 2003. Subsoil polluted by hydrocarbons in an out-of-service oil distribution and storage station in Zacatecas, Mexico. Environ. Geol. 44, 608–620.

Iturbe, R., Flores, R.M., Flores, C.R., Torres, L.G., 2004. TPH-polluted Mexican refinery soil: health risk assessment and the first year of changes. Environ. Monit. Assess. 91, 237–255.

Jensen, T.S., Arvin, E., Svensmark, B., Wrang, P., 2000. Quantification of compositional changes of petroleum hydrocarbons by GC/FID and GC/MS during a long-term bioremediation experiment. Soil Sediment Contam. 9, 549–577.

Jorgensen, K.S., Puutinen, J., Suortti, A.M., 2000. Bioremediation of petroleum hydrocarbon-contaminated soil by composting in biopiles. Environ. Pollut. 107, 245–254.

Kodres, C.A., 1999. Coupled water and air flows through a bioremediation soil pile. Environ. Modell. Softw. 14, 37–47.

Lee, B., Visser, S., Fleece, T., Krieger, D., 2002. Bioremediation and ecotoxicology of drilling fluids used for land-based drilling. In: Proceedings of AADE 2002 Technology Conference Drilling and Completion Fluids and Mud Management, Houston, TX April 2–3.

Lindstrom, J.E., Braddock, J.F., 2002. Biodegradation of petroleum hydrocarbons at low temperature in the presence of the dispersant Corexit 9500. Mar. Pollut. Bull. 44, 739–747.

Margesin, R., Schinner, F., 2001. Biodegradation and bioremediation of hydrocarbons in extreme environments. Appl. Microbiol. Biotechnol. 56, 650–663.

Mishra, S., Jeevan, J., Ramesh Chander, K., Banwari, L., 2001. *In situ* bioremediation potential of an oily sludge-degrading bacterial consortium. Curr. Microbiol. 43, 328–335.

Molina-Barahona, L., Rodriguez-Vazquez, R., Hernandez-Velasco, M., Vega-Jarquin, C., Zapata-Perez, O., Mendoza-Cantu, A., Albores, A., 2004. Diesel removal from polluted soils by biostimulation and supplementation with crop residues. Appl. Soil Ecol. 27, 165–175.

Morgan, P., Watkinson, R.J., 1989. Hydrocarbon biodegradation in soils and methods for soil biotreatment. Crit. Rev. Biotechnol. 8, 305–333.

Muñoz, I.D.J., Mendoza, C.A., López, G.F., Soler, A.A., Hernández, M.M.M., 2000. Manual de Análisis de suelo. UNAM, México.

Nelson, E.C., Walter, M.V., Bossert, I.D., Martin, D.G., 1996. Enhancing biodegradation of petroleum hydrocarbons with guanidinium fatty acids. Environ. Sci. Technol. 30, 2406–2411.

Ouyang, W., Lui, H., Murygina, V., Yu, Y., Xiu, Zengde, Kalyuzhnyi, S., 2005. Comparison of bio-augmentation and composting for remediation of oily sludge: a field-scale study in China. Proc. Biochem. 40, 3763–3768.

Penberthy, J., Weston, R., 2000. Remediation of diesel and fuel oil hydrocarbons in high clay content soils: a field comparison of amendment performance conducted at the Mare Island naval shipyard. In: Proceedings of the National Defense Industrial Association, Long Beach, CA, March 27–30.

Rivera-Cruz, M.D., Ferrera-Cerrato, R., Sánchez-García, P., Volke-Haller, V., Fernández-Linares, L., Rodríguez-Vázquez, R., 2004. Decontamination of soils polluted with crude petroleum using indigenous microorganisms and aleman grass (*Echinochloa polystachya*), *BK Hichc*. Agrociencia 38, 1–12.

Rojas-Avelizapa, N.G., Martínez-Cruz, J., Zermeño-Eguia, J.A., Rodríguez-Vázquez, R., 2003. Levels of polychlorinated biphenyls in Mexican soils and their biodegradation using bioaugmentation. Environ. Contam. Toxicol. 70B, 63–70.

Roldan, T., Rojas, N., Muñoz, A., Zaragoza, D., Fernandez, L.C., 2003. Bioremediation of a drilling muds polluted soil using agricultural muds. In: Proceedings of Second International Conference of Petroleum Biotechnology: The Development and Perspectives of Biotechnology Applied to the Oil Industry. México. November 5–7.

Semple, K.T., Reid, B.R., Fervor, T.R., 2001. Impact of composting strategies on the treatment of soils polluted with organic pollutants. Environ. Pollut. 112, 269–283.

Song, H.G., Bartha, R., 1990. Effects of jet fuel spills on the microbial community of soil. Appl. Environ. Microbiol. 56, 646–651.

von Fahnestock, F.M., Wickramanayake, G.B., Kratzke, R.J., Major, W.R., 1998. Biopile Design, Operation and Maintenance Handbook for Treating Hydrocarbon-Polluted Soils. Batelle Press, Columbus, OH.

Walworth, J.L., Reynolds, C.M., 1995. Bioremediation of a petroleum-polluted cryic soil: effects of phosphorous, nitrogen and temperature. J. Soil Contam. 4, 299–310.

Zimmerman, P.K., Rober, J.D., 1991. Oil-based drill cuttings treated by landfarming. Oil Gas J. 12, 81–84.

Zytner, R.G., Salb, A.C., Stiver, W.H., 2006. Bioremediation of diesel fuel contaminated soil: comparison of individual compounds to complex mixtures. Soil Sediment Contam. 15, 277–297.