

A EVOLUÇÃO DOS CORPORA: DE ALEXANDER CRUDEN AOS COMPUTADORES

VANDERLEI DOS SANTOS¹

1. Prof. Ms. Fatec Tatuí - Curso Tecnologia em Processos Gerenciais - vanderlei.santos@fatec.sp.gov.br

RESUMO

Os corpora podem ser definidos como uma coleção de textos considerados representativos de uma dada língua, utilizada para análise linguística. Os corpora, e as listas de palavras elaboradas a partir deles, são valiosos no ensino de línguas. São utilizados na escolha de vocabulário para livros-texto e textos elaborados para estudantes, na orientação para a aquisição e avaliação de vocabulário e como base para concordâncias. Desde Alexander Cruden e o primeiro corpus pré-eletrônico até os corpora modernos, compostos por centenas de milhões de palavras, muitos foram os avanços proporcionados pela evolução tecnológica à compilação dos corpora e listas de palavras. A apresentação dessa trajetória é o objetivo deste texto.

PALAVRAS-CHAVE: corpus. listas de palavras. ensino de línguas.

INTRODUÇÃO

Os corpora têm sido usados para o ensino de línguas há muito tempo. Listas de vocabulário para aprendizes, por exemplo, têm sido geradas a partir de corpora, e contagens de palavras derivadas de análises de corpora têm auxiliado na definição de objetivos na aquisição de vocabulário. Os criadores de dicionários e livros-texto têm usado corpora extensivamente. Mais recentemente, a tendência a se usar materiais autênticos no ensino de línguas tem valorizado o papel que compilações da linguagem escrita ou falada podem ter na aprendizagem de línguas. Os corpora são, afinal, enormes estoques da língua real em uso. O interesse no ensino de línguas para fins específicos amplia o uso dos corpora como meio de identificar os componentes específicos da língua a serem ensinados. Por essas razões torna-se interessante saber mais sobre os corpora e sua evolução.

1. CORPORA PRÉ-ELETRÔNICOS

O termo corpus, palavra do Latim para corpo ou conjunto, foi usado no século VI para descrever uma coleção de textos legais, Corpus Juris Civilis (FRANCIS, 1979). O termo corpus tem mantido esse significado, o de um corpo de texto; mas para a lingüística de corpus essa definição não é suficiente. Numa das cinco definições do *Oxford English Dictionary* um corpus é "O corpo de material escrito ou falado sobre o qual uma análise lingüística é baseada". O dicionário Collins COBUILD define um corpus como "uma grande coleção de textos escritos ou falados que é usada para a pesquisa lingüística". O dicionário Aurélio define corpus como "conjunto finito de materiais significantes constituído com vistas à análise semiológica". Portanto, o corpus não pode ser visto como apenas uma coleção de textos, mas sim uma coleção de textos considerados representativos de uma dada língua, dialeto ou outro sub-sistema de uma língua, para ser usada para análise lingüística.

Alguns termos freqüentemente utilizados no estudo dos corpora precisam ser conhecidos: **Ocorrências** são todas as palavras que aparecem em um texto, seja escrito ou falado. São contadas todas as vezes em que uma palavra aparece. Dessa forma a sentença: "Eu não sei se eu vou" conteria

seis palavras ou ocorrências. **Tipos** são ocorrências não repetidas. Essa mesma sentença teria 5 tipos, a palavra “eu” é contada como um tipo. **Lema** consiste de uma forma base e algumas de suas formas flexionadas e reduzidas (*n't* em inglês, por exemplo). Então *go, go, went, gone* contam como 4 ocorrências, 3 tipos e 1 lema (VERMEER, 2000). **Famílias** incluem a forma base, suas formas declinadas e suas formas derivadas próximas, o que inclui afixos como: *-ly, -ness, e un-*. Frequentemente, supõe-se que a pesquisa baseada em corpora tenha começado no início dos anos 60, com a disponibilidade de corpora eletrônicos. Entretanto, antes disso já havia uma considerável tradição de análise lingüística baseada em corpora.

Alexander Cruden, em meados da década de 1720, decidiu compilar a mais detalhada concordância (uma listagem, total ou parcial, das palavras de um texto ou corpus, mostrando o contexto de ocorrência de cada palavra) da Bíblia (*King James Version of the Bible - KJVB*). A primeira edição da Concordância da Bíblia de Cruden foi publicada em 1737. Esse trabalho, que sempre foi realizado por um grande grupo de pessoas e, mais tarde, pelos computadores, foi feito por Cruden em sua residência, sozinho e à mão. A KJVB tem 777.746 palavras. Cruden ainda escreveu textos explicativos para muitas das palavras; a palavra "Sinagoga", por exemplo, remetia a um texto de 4.000 palavras (LARSEN e MAIO, 2005).

O *Dictionary of the English Language*, publicado em 1775 por Samuel Johnson não foi, como muitos acreditam, o primeiro dicionário da língua inglesa, mas foi o primeiro a apresentar contextualização (150.000 citações ilustrativas), e foi o dicionário padrão de inglês por mais de 150 anos. O dicionário de Johnson permaneceu o mais confiável dicionário da língua inglesa até o aparecimento do dicionário *A New English Dictionary on Historical Principles*, publicado em 10 fascículos entre os anos de 1884 e 1928 e que foi reimpresso em 1933, quando recebeu o nome de *Oxford English Dictionary* (2005); 1.700 das definições de Johnson ainda se encontram nesse dicionário.

Em 1694, porém, a França já tinha publicado seu enorme dicionário, *Le dictionnaire de l'Academie francaise*, resultado de meio século de trabalho de 40 acadêmicos. Johnson acreditava que poderia fazer o trabalho em três anos mas, ao final, levou quase uma década, trabalhando com caneta e papel no sótão de sua casa em Londres. Johnson compilava longas listas de palavras, as quais eram cortadas em tiras e ordenadas alfabeticamente. Ele garimpou a literatura de muitos séculos em busca de citações ilustrativas para suas palavras (LYNCH, 2003).

A idéia para o *Oxford English Dictionary*, um dicionário extremamente complexo, surgiu em 1857. O plano era criar uma coleção vasta e abrangente de palavras inglesas, um léxico da língua mais completo do que jamais tinha sido tentado.

O plano foi formulado pela Sociedade Filológica Inglesa (*English Philologic Society*), um grupo que investigava a estrutura e a história das línguas. Reconhecendo as lacunas que existiam em outros dicionários, a sociedade acreditava que um novo projeto deveria ser executado para examinar o vocabulário da língua inglesa, inclusive aquelas palavras que tinham sido rejeitadas ou tinham passado despercebidas por outros lexicógrafos.

Em 1879, um acordo foi alcançado com os editores para começar o trabalho. O editor do dicionário, James Murray, pediu aos leitores dos países de língua inglesa para participarem.

Centenas de voluntários trabalharam como “detetives de palavras”, vasculhando textos históricos e contemporâneos para coletar tantas palavras quanto possível, analisando as diferentes maneiras em que cada uma tinha sido usada. Procuraram na literatura (popular e clássica), jornais, periódicos científicos e tecnológicos, letras de músicas, roteiros de teatro, livros de receitas, testamentos e documentos políticos, coletando uma infinidade de palavras e significados. Como resultado, Murray recebeu milhões de citações, as quais sua equipe passou a avaliar, ordenar e arquivar.

O dicionário viria a incluir palavras perdidas e fora de moda, assim como as mais recentes e termos técnicos; ele traçava a etimologia de cada palavra, mostrando seu uso mais antigo, mapeando como ela tinha mudado em uso ou significado através do tempo; mostrava também as famílias das palavras, pronúncia e múltiplos significados. Quase 50 anos se passariam antes que o último fascículo do dicionário fosse publicado, em 1928.

Nascido em Wet Hartford, Connecticut, em 1758, Noah Webster acreditava ardentemente no desenvolvimento da independência cultural dos Estados Unidos, e que uma linguagem americana distinta, com suas próprias características, pronúncia e estilo, era essencial.

Em 1806 Webster publicou *A Compendious Dictionary of the English Language*, o primeiro dicionário verdadeiramente americano. Imediatamente, começou a trabalhar na sua obra máxima, o *American Dictionary of the English Language*, para o que ele aprendeu 26 línguas, incluindo Anglo-Saxão e Sânscrito, a fim de pesquisar as origens da língua do seu país. Esse livro, publicado em 1828, incorporava um novo padrão da lexicografia; era um dicionário com 70.000 registros e considerado por muitos como tendo superado o dicionário de Johnson, não somente em escopo como também em confiabilidade.

Uma faceta da personalidade de Webster era sua disposição para inovar quando ele pensava que a inovação significava evolução. Ele foi o primeiro a documentar palavras distintamente americanas como: *skunk*, *hickory* e *chowder*. Acreditando que muitas convenções da escrita eram artificiais e desnecessariamente confusas, ele encorajou a alteração de várias palavras: *musik* para *music*, *centre* para *center* e *plough* para *plow*, por exemplo (NOAH, 2004).

Edward Lee Thorndike nasceu em Williamsburg, Massachusetts, E.U.A., em 1874. Esse psicólogo educacional americano ofereceu muitas contribuições para a pesquisa sobre o ensino e a aprendizagem. Ele foi um dos primeiros a desenvolver testes para medir a aprendizagem e aptidões. O interesse de Thorndike por contagem de frequência de palavras iniciou-se quando ele veio a saber que os professores de línguas na Alemanha e na Rússia estavam usando contagens de palavras. Eles acreditavam que quanto mais freqüentemente uma palavra fosse usada mais útil ela seria para os aprendizes de línguas.

Em 1911, Thorndike começou a contar a frequência das palavras em textos em inglês com a intenção de coletar palavras que seriam úteis para crianças americanas aprendendo a ler nessa língua. Em 1921 ele publicou *The teacher's word book*, o qual listava 10.000 palavras pela frequência de uso, e foi baseado em um corpus de 4.500.000 palavras. Aproximadamente 625.000 palavras de textos da literatura infantil, 3.000.000 de palavras da Bíblia e clássicos ingleses, 300.000 palavras de livros-texto do nível elementar, 50.000 palavras de livros sobre culinária, agricultura, costura, comércio, etc., 90.000 palavras de jornais e 500.000 palavras de correspondências foram usadas para compor esse corpus (MALLIKARJUN, 2002).

Em 1932 ele deu seguimento ao trabalho com a publicação de *A teacher's word book of 20,000 words* (DUBAY, 2004). Em 1944, com a ajuda de Irving Lorge, Thorndike publicou *The teacher's word book of 30,000 words*. Acrescentando a contagem de palavras de revistas que havia sido feita por Lorge, a contagem de Thorndike de 120 livros juvenis e a contagem semântica de Thorndike e Lorge à contagem geral de Thorndike, o corpus usado para a elaboração dessa lista chegou a 18.000.000 de ocorrências. O trabalho de Thorndike foi a base para as primeiras fórmulas de facilidade de leitura¹. Educadores, editores de livros-texto e professores usam listas de frequência de palavras, assim como fórmulas de facilidade de leitura para casar materiais de leitura com leitores de diferentes níveis.

Educado em uma escola pública em Christ Church, Oxford, Inglaterra, Michael Philip West foi diretamente da graduação, em 1912, para um posto no Serviço de Educação da Índia. Seus anos mais frutíferos e influentes foram passados em Bengala aonde ele veio a tornar-se Inspetor de Diretores de Escolas em Chittagong e Calcutá, Diretor da Faculdade de Treinamento de Professores em Dacca e Leitor Honorário em Educação na Universidade de Dacca (HOWATT, 1984).

A mais importante sucessão de eventos na carreira de West, entretanto, aconteceu 10 anos depois de sua nomeação. Em 1923, a Conferência de Educação Imperial tinha pedido uma investigação dos fatos do bilingüismo com referência ao desenvolvimento intelectual, emocional e moral da criança e a importância das questões práticas de educação advindas da investigação de tais fatos. A resposta de West a esse pedido foi um projeto experimental, escrito no relatório sobre bilingüismo. Nele, West desafiava a política educacional, que distribuía graduações e empregos no governo para um pequeno número de estudantes muito capazes, mas que resultava na desistência da maioria em diversos estágios, sem que qualquer benefício fosse auferido desse processo de educação incompleto (TICKOO, 1988). O desperdício era maior devido ao longo período dedicado ao desenvolvimento da habilidade de conversação. A habilidade de leitura, por outro lado, podia ser

¹ De acordo com Richards, Platt e Platt (1992), facilidade de leitura (readability) se refere à facilidade com que textos podem ser entendidos, e depende de vários fatores, incluindo a extensão das sentenças, o número de palavras novas e a complexidade da linguagem utilizada no texto.

adquirida mais rapidamente e mais agradavelmente, acreditava West, e não dependia da presença constante de um professor.

A abordagem de West permitia dar acesso a todos os bengaleses, e não só a alguns poucos privilegiados, a materiais de leitura sobre assuntos práticos que não estavam disponíveis na língua mãe. Isto significava o desenvolvimento de materiais melhorados, baseados em princípios de controle estrito do vocabulário. Havia duas maneiras, segundo West, pelas quais os textos existentes podiam ser melhorados. O primeiro era simplificar o vocabulário removendo palavras raras ou antiquadas e substituindo-as por equivalentes mais modernas e comuns. O segundo princípio aplicado por West consistia em distribuir as novas palavras de maneira tal que não ocorressem com muita freqüência, para que o leitor fosse capaz de absorvê-las e praticá-las a fundo. Nos novos textos que West adaptou ou escreveu, o número de palavras novas caiu, enquanto o número total de palavras aumentou acentuadamente. Em conseqüência, o ritmo pelo qual novas palavras eram apresentadas foi diminuído de 1 palavra desconhecida para 7,4 conhecidas para não mais do que 1 desconhecida para 44,7 conhecidas (HOWATT, 1984). As idéias de West estavam de acordo com as do movimento de controle de vocabulário. A idéia por trás desse movimento era a de que as palavras mais importantes de uma língua deveriam ser ensinadas aos aprendizes primeiro, e que o número dessas palavras deveria ser limitado.

Em meados dos anos 30, West envolveu-se, com outras figuras do movimento para o controle de vocabulário, na Conferência de Carnegie, a qual aconteceu pela iniciativa de West e sob os auspícios da Corporação Carnegie, primeiro em Nova Iorque, em 1934, e depois em Londres em 1935. O propósito principal da conferência era examinar o papel das listas de palavras no ensino de inglês como língua estrangeira. A escolha dos itens da *General Service List* foi confiada a West e Palmer. Essa lista viria a ser o maior componente do relatório da conferência, e seria publicada em 1953 separadamente, como a mais confiável e certamente a mais conhecida de todas as listas de palavras. Durante os anos 30, West se tornou um dos mais conhecidos e mais prolíficos autores no campo do inglês como língua estrangeira.

Sua carreira, porém, foi marcada pela controvérsia. Embora suas propostas para uma abordagem centrada no leitor fossem consideradas excelentes, elas não eram vistas com bons olhos nos círculos oficiais da Índia (TICKOO, 1988). Nos anos 50 e 60 ele permaneceu ativo, contribuindo regularmente para o periódico *English Language Teaching* entre outros. Escreveu também um pequeno manual, *Teaching English in Difficult Circumstances*, o qual incluía um apêndice com o que West considerava o vocabulário mínimo para permitir a comunicação. A GSL (West, 1953) é uma lista de aproximadamente 2.000 famílias de palavras escolhidas, em grande parte, com base em sua freqüência de uso. Os dados de freqüência usados na GSL vieram das contagens feitas por Thorndike e Lorge no início do século 20. Existem algumas significativas diferenças entre a GSL e listas anteriores. A primeira é que as palavras base e suas derivadas são apresentadas com exemplos ilustrativos. Exemplos com a palavra *extend* podem ser vistos na tabela 1. Além disso a lista é enriquecida pela inclusão de expressões idiomáticas e verbos frasais tais como: *break to pieces, break one's heart, break down, break up*.

Tabela 1 - Exemplo de um verbete da GSL Fonte: (WEST, 1953)

(3) EXTEND	
<i>extend</i> , v.	(1) (<i>stretch out, be stretched out</i>) The garden extends as far as the river (2) (<i>continue, enlarge, protract, lengthen</i>) Extend one's visit Extend a business
<i>extent</i> , n.	To the full extent of the garden
<i>extension</i> , n.	An extension of the hospital
<i>extensive</i> , adj.	Extensive repairs, enquiries
<i>extensively</i> , adv. (GSL)	

2. CORPORA DE PRIMEIRA GERAÇÃO

O *Brown Corpus of Standard American English* foi o primeiro corpus computadorizado. Ele foi compilado por W. Nelson Francis e Henry Kucera, da Universidade Brown, Providence, Rhode Island. O corpus consiste de 1 milhão de palavras extraídas de textos do inglês americano. Hoje, esse corpus é considerado pequeno e ligeiramente antiquado, porém o corpus ainda é usado. Muito da sua utilidade vem do fato de que sua estrutura tem sido copiada por outros compiladores de corpora. O Lancaster-Oslo/Bergen (LOB) Corpus (inglês britânico) e o Kolhapur Corpus (inglês indiano) são dois exemplos de corpora feitos para serem compatíveis com o *Brown Corpus*. A disponibilidade de corpora tão similar em estrutura é um recurso valioso para pesquisadores interessados em comparar diferentes variedades da língua (FRANCIS e KUCERA, 1979).

O *Corpus Lancastre-Oslo/Bergen* (LOB) foi compilado por pesquisadores de Lancaster, Oslo e Bergen, em 1961. Ele consiste de 1 milhão de palavras do inglês britânico extraídas de textos de 15 categorias diferentes como: jornais, textos de ficção adulta e infantil, biografias, revistas, etc. Cada texto tem pouco mais de 2.000 palavras (textos mais longos são cortados na primeira sentença após as 2.000 palavras) e o número de textos varia em cada categoria. O corpus é a contrapartida britânica ao *Corpus Brown*, o qual contém textos do mesmo ano, de forma que uma comparação entre as duas variedades pudesse ser feita (LOB, 2002).

Outros corpora gerais de primeira geração:

- *The Kolhapur Corpus of Indian English* - 1978, inglês indiano, segue a estrutura do Corpus Brown.
- *The Wellington Corpus of Written New Zealand English* - 1986, inglês da Nova Zelândia, segue a estrutura dos corpora *Brown* e LOB com algumas modificações.
- *The Australian Corpus of English* - 1986, mesma estrutura e tamanho do Brown/LOB, com algumas modificações.
- *The Corpus of English-Canadian Writing* - 1984, tem a mesma estrutura do Brown/LOB com a adição de textos das categorias feminismo e informática, sendo três vezes maior do que o Brown.
- *The Standard Corpus of Present-day English Language Usage* - 1970, é o Corpus Brown re-organizado de acordo com o número de letras das palavras.
- *The London-Lund Corpus* (LLC) - 1987, parte oral do *Survey of English Usage Corpus*, da *Lund University*. Ele contém 500.000 palavras do inglês britânico falado coletadas entre 1953 e 1987 (TONO e MAIO, 2003).

3. CORPORA DE SEGUNDA GERAÇÃO

Em 1980, na Universidade de Birmingham, o professor John Sinclair estabeleceu o projeto COBUILD para realizar uma análise por computador de um corpus totalizando 18 milhões de palavras. COBUILD é uma sigla de "CO" de Collins, a editora, e "BUILD" de *Birmingham University International Language Database*. Hoje esse corpus cresceu para mais de 400 milhões de palavras, havendo uma previsão de que mais 100 milhões sejam acrescentadas no futuro próximo.

Quase todo o material do corpus foi coletado a partir de 1990, e todo ano ele é revisado, visando substituir dados antigos por mais atuais. Os textos vêm de várias fontes, embora os jornais, sendo uma das fontes mais práticas atualmente, respondam por uma grande parte do corpus. Além dos jornais, o corpus inclui uma extensa coleção de textos de revistas gerais e especializadas e também publicações de negócios como o *Wall Street Journal* e *The Economist*. O inglês falado também é representado, contando com aproximadamente 20 milhões de palavras de transmissões de rádio americanas da *National Public Broadcasting* em Washington, 20 milhões de palavras da *BBC World Service* e rádios locais, além de entre 10 e 20 milhões de palavras de conversas informais, entrevistas e reuniões. Atualmente a equipe do COBUILD está trabalhando para dar ao corpus um caráter mais

internacional, com a inclusão de palavras de corpora australianos, canadenses e indianos (THE DICTIONARY, 2002).

O *Longman Corpus Network* compõe-se de um grupo de cinco bancos de dados. O *Longman Learners' Corpus*, composto pela produção escrita de aprendizes de inglês; o *Longman Written American Corpus*, composto por 100 milhões de palavras de livros e jornais americanos; o *Longman Spoken American Corpus*, composto por 5 milhões de palavras do inglês oral americano, extraídas de situações do cotidiano; o *Spoken British Corpus* (parte do *British National Corpus*), composto por palavras do inglês oral britânico e o *Longman/Lancaster Corpus* com mais de 30 milhões de palavras cobrindo um âmbito extenso de textos escritos que inclui desde a literatura até tabelas de horários de ônibus.

O *British National Corpus* contém mais de 100 milhões de palavras do inglês moderno, falado e escrito. Esse projeto foi executado e é mantido por um consórcio acadêmico/industrial liderado pela *Oxford University Press*. As palavras representativas da linguagem escrita (90%) foram extraídas de jornais regionais e nacionais, periódicos especializados, livros acadêmicos e ficção popular, cartas e memorandos publicados ou não, ensaios escolares e universitários, entre muitos outros tipos de textos. A proporção referente à linguagem oral (10%) provém de uma grande quantidade de conversação informal não ensaiada, gravada por voluntários selecionados de diferentes idades, regiões e classes sociais de uma forma demograficamente equilibrada, além de excertos orais coletados em diversos tipos de contextos, indo de reuniões formais de negócios ou governamentais a programas de rádio e ligações telefônicas. O trabalho de construção do corpus foi completado em 1994 (WHAT, 2004).

O *International Corpus of English* (ICE) teve início em 1990 com o objetivo de coletar material para estudos comparativos de inglês mundialmente. Quinze equipes de pesquisa ao redor do mundo estão preparando corpora eletrônicos de suas próprias variedades nacionais ou regionais de inglês. Cada corpus do ICE consiste de 1 milhão de palavras do inglês falado ou escrito produzido depois de 1989. Para muitos países participantes, o projeto ICE está estimulando a primeira investigação sistemática da variedade nacional. Para garantir a compatibilidade entre os corpora, todas as equipes seguem a mesma estrutura. Cada corpus contém 500 textos de aproximadamente 2.000 palavras, totalizando aproximadamente 1 milhão de palavras. Os textos do corpus são de 1990 ou mais recentes (CORPUS, 2003).

O *Cambridge International Corpus* (CIC) foi compilado pela *Cambridge University Press* nos últimos 10 anos. Os textos do corpus vem de jornais, livros de ficção e não-ficção sobre vários tópicos, sites da internet, revistas, correspondência, programas de rádio e televisão e conversas cotidianas.

Tabela 2. Composição do CIC Fonte: Cambridge (2005)

Inglês britânico	
Nº de palavras	Corpus
450 milhões	Inglês britânico falado
17 milhões	Inglês britânico falado
20 milhões	Inglês acadêmico britânico escrito
30 milhões	Inglês comercial britânico escrito
1 milhão	Inglês comercial britânico falado
Inglês americano	
Nº de palavras	Corpus
200 milhões	Inglês americano escrito
22 milhões	Inglês americano falado
7 milhões	Inglês acadêmico americano escrito
30 milhões	Inglês comercial americano escrito
Inglês de aprendizes	
Nº de palavras	Corpus
19 milhões	Inglês escrito por aprendizes (IEA)
8 milhões	IEA com erros codificados

4. CORPORA DE LÍNGUA PORTUGUESA

O Núcleo Interinstitucional de Lingüística Computacional (NILC) foi criado em 1993, na Universidade de São Paulo (USP) em São Carlos, para estimular a pesquisa e o desenvolvimento de projetos em Lingüística Computacional e Processamento da Linguagem Natural. Por essa ocasião, deu-se início à construção de diversas ferramentas computacionais, visando a dar suporte à tarefa de revisão textual; dentre esse material de suporte, foi elaborado um banco de textos o qual, mais tarde, se tornaria o Corpus de base para diversos aplicativos desenvolvidos pelo NILC, o Corpus NILC (CN) (PINHEIRO e ALUÍSIO, 2003). A construção do CN foi um empreendimento decorrente da ferramenta de revisão desenvolvida pelo NILC em parceria com a ITAUTEC/PHILCO a partir de 1993 – o revisor ReGra. O NILC, em setembro de 2002, contava com 34.092.630 palavras. O CN é composto de textos didáticos (1.147.325 ocorrências), jurídicos (761.852 ocorrências), literários (2.184.620 ocorrências), técnico-científicos (1.767.565 ocorrências), jornalísticos (27.203.360 ocorrências) e universitários (1.027.908 ocorrências).

O corpus de Araraquara teve sua montagem iniciada no começo dos anos 90, por Francisco S. Borba, lingüista especializado em dicionários e autor de diversas obras sobre a língua portuguesa. É um dos mais importantes bancos de dados de língua escrita no país, com aproximadamente 200 milhões de ocorrências de palavras em textos de literatura românica, dramática, técnica, oratória e, predominantemente, jornalística (FRANÇA, 2002).

5. LISTAS DE PALAVRAS ACADÊMICAS

A *University Word List* (UWL) é uma compilação, feita por Xue e Nation (1984) de quatro listas separadas, as de Campino e Ellery, de 1971, e a de Praninskas, 1972; baseadas em corpora e compostas por palavras que ocorriam em textos gerais; e as de Lynn, de 1973, e a de Ghadessy, de 1979, compiladas a partir das anotações de alunos em livros didáticos. A UWL consiste de 836 famílias, um total de 3.685 palavras, não encontradas na GSL, mas freqüentes em textos acadêmicos (NATION e HWANG, 1995). Ela exclui as primeiras 2.000 palavras da GSL e termos técnicos. Assim fazendo, a UWL pretende ser uma lista de palavras acadêmicas gerais² que ocorrem através de uma grande extensão de áreas como artes, ciências e direito. A lista contém 11 níveis, com o nível 1 contendo as palavras mais freqüentes e o nível 11 as menos freqüentes (COXHEAD, 2000).

Por considerar que a UWL era um amálgama de outras listas, sem critérios consistentes de seleção, baseada em corpora pequenos e não contendo uma extensão de tópicos ampla e equilibrada, (COXHEAD, 2000) compilou a *Academic Word List* (AWL), constituída de 570 famílias de palavras. O corpus para esse estudo foi composto por textos de 28 diferentes cursos, de 4 áreas: artes, comércio, direito e ciências, que juntos somaram 3,5 milhões de palavras. Para a seleção a autora seguiu três critérios: 1) ocorrência especializada, as famílias de palavras incluídas não podiam estar presentes na GSL; 2) alcance, uma palavra de cada família tinha de aparecer pelo menos 10 vezes em cada uma das quatro áreas e em pelo menos 15 dos 28 cursos que o corpus abrangia; e 3) freqüência, membros de uma família tinham que aparecer pelo menos 100 vezes no corpus. A lista foi dividida em famílias de acordo com o nível 6 da escala de Bauer e Nation.

Essa escala foi elaborada por Bauer e Nation (1993) para o agrupamento de palavras em famílias, na língua inglesa, com o objetivo de orientar os professores de inglês no ensino dos afixos. Essa classificação foi usada por Coxhead (2000) na elaboração da AWL, a qual utilizou o nível 6 da classificação de Bauer e Nation para o agrupamento em famílias. Bauer e Nation afirmam que “do ponto de vista da leitura, uma família de palavras consiste de uma palavra base e todas as suas formas derivadas e flexionadas que possam ser entendidas por um aprendiz sem que tenha que aprender cada forma separadamente”. Partindo dessa premissa, *watch*, *watches*, *watched* e *watching* podem ser membros da mesma família para um aprendiz que possua o conhecimento dos sufixos flexionais em inglês. Conforme o conhecimento do uso de afixos do aprendiz se desenvolve, o tamanho das famílias de palavras aumenta. Os autores deixam claro que o significado da palavra base na forma derivada

² O vocabulário acadêmico, ou sub-técnico, é o vocabulário comum a várias áreas acadêmicas, ao contrário do vocabulário técnico, definido como o vocabulário reconhecidamente específico de um tópico, campo ou disciplina.

deve estar claramente relacionado ao significado da palavra base quando encontrada em outras formas derivadas. Por exemplo, *hard* e *hardly* não seriam membros da mesma família. A classificação do agrupamento em famílias é apresentada pelos autores em sete níveis, cada nível inclui os níveis anteriores e acrescenta novos afixos.

- Nível 1 - Cada palavra é considerada uma família
- Nível 2 - As flexões da palavra base consideradas pertencentes à mesma família neste nível são: plural, terceira pessoa do singular no presente, passado, particípio passado, *-ing*, comparativo, superlativo e possessivo.
- Nível 3 - As famílias incluem as palavras formadas pela adição dos seguintes afixos: *-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-*.
- Nível 4 - Inclui os afixos: *-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-*.
- Nível 5 - Acrescenta os afixos: *-age, -al, -ally, -an, -ance, -ant, -ary, -atory, -dom, -eer, -en, -en, -ence, -ent, -ery, -ese, -esque, -ette, -hood, -i, -ian, -ite, -let, -ling, -ly, -most, -ory, -ship, -ward, -ways, -wise, anti-, ante-, arch-, bi-, circum-, counter-, en-, ex-, fore-, hyper-, inter-, mid-, mis-, neo-, post-, pro-, semi-, sub-, un-*.
- Nível 6 - Afixos: *-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-*.
- Nível 7: Afixos: *ab-, ad-, com-, de-, dis-, ex-, sub-*.

Então, usando como exemplo a palavra *develop* teríamos os seguintes agrupamentos em cada nível:

Tabela 3. Exemplo da classificação de Bauer e Nation

Nível	Agrupamento
	<i>develop</i>
2	<i>develops</i> <i>developed</i> <i>developing</i>
3	<i>developable</i> <i>undevelopable</i> <i>developer(s)</i> <i>undeveloped</i>
4	<i>development(s)</i> <i>developmental</i> <i>developmentally</i>
5	<i>developmentwise</i> <i>semideveloped</i> <i>antidevelopment</i>
6	<i>redevelop</i> <i>predevelopment</i>

No nível 3 uma família inclui todas as derivações dos níveis 2 e 3, no nível 4 todas as derivações dos níveis 2, 3 e 4; e assim por diante. Para maiores informações sobre os critérios para a elaboração dos níveis, limitações do estudo, usos sugeridos e sugestões para pesquisa adicional, vide Bauer e Nation (1993). Não encontrei estudos semelhantes sobre a língua portuguesa e acredito que tais estudos trariam grandes contribuições à área de ensino de português como língua estrangeira.

REFERÊNCIAS

- BAUER, L.; NATION, I. S. P. Word families. **International Journal of Lexicography**, v. 6, n. 2, p. 253-279, 1993.
- CAMBRIDGE International Corpus. Cambridge. 2005. Disponível em: http://www.cambridge.org/br/elt/catalogue/subject/custom/item3637700/Cambridge-International-Corpus-Cambridge-International-Corpus/?site_locale=pt_BR. Acesso em: 28 abr. 2005.
- CORPUS Design. International Corpus of English. 2003. Disponível em: <http://www.ucl.ac.uk/english-usage/ice/>. Acesso em: 2 mai. 2005.
- COXHEAD, A. A new academic word list. **TESOL Quarterly**, v. 34, n. 2, p. 213-238, 2000.
- THE DICTIONARY Revolution. Bangkok Post. 2002. Disponível em: <http://www.bangkokpost.net/education/site2002/cvap0202.htm>. Acesso em: 2 mai. 2005.
- DUBAY, W. H. The principles of readability. 2004. Disponível em: <http://www.nald.ca/fulltext/readab/02.htm>. Acesso em: 2 mai. 2005.
- FRANÇA, F. Autor fez pesquisa com textos jornalísticos. 2002. Disponível em: http://atica.com.br/imprensa/borba_londrina.asp. Acesso em: 2 mai. 2005.
- FRANCIS, W. N.; KUCERA, H. Brown Corpus manual. 1979. Disponível em: <http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM>. Acesso em: 2 mai. 2005.
- HOWATT, A. **A History of English Language Teaching**. Oxford: Oxford University Press, 1984.
- LARSEN, T.; MAI. The orderly product of a disordered mind. Christianity Today International, 2005.
- LOB Corpus. LOB. 2002. Disponível em: http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html. Acesso em: 2 mai. 2005.
- LYNCH, J. Samuel Johnson. 2003. Disponível em: <http://www.andromeda.rutgers.edu/~jlynch/Johnson>. Acesso em: 27 abr. 2005.
- MALLIKARJUN, B. Vocabulary education. 2002. Disponível em: <http://www.languageinindia.com/nov2002/vocabulary.html>. Acesso em: 2 mai. 2005.
- NATION, P.; HWANG, K. Where would general service vocabulary stop and special purposes vocabulary begin? **System**, v. 23, n. 1, p. 35-41, 1995.
- NOAH Webster and America's first dictionary. Merriam-Webster On Line. 2004. Disponível em: <http://www.m-w.com/info/noah.htm>. Acesso em: 28 abr. 2005.
- OXFORD English Dictionary. **History of the Dictionary**. Oxford University Press. 2005. Disponível em: <http://www.oed.com/about/history.html#supp>. Acesso em: 31 dez. 2005.
- PINHEIRO, G. M.; ALUÍSIO, S. M. **Corpus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web**. USP. São Carlos, SP, p.60. 2003.
- READ, J. Research in teaching vocabulary. **Annual Review of Applied Linguistics**, v. 24, p. 146-161, 2004.
- THE DICTIONARY revolution. Bangkok Post. 2002. Disponível em: <http://www.bangkokost.net/education/site2002/cvap0202.htm>. Acesso em: 2 mai. 2005.
- TICKOO, M. **ESP: State of the art**. Singapore: Singapore University Press / RELC, 1988.
- TONO, Y.; MAI. Learner corpora and SLA research. 2003. Disponível em: <http://leo.meikai.ac.jp/~tono/>. Acesso em: 2 mai. 2005.
- VERMEER, A. Coming to grips with lexical richness in spontaneous speech data. **Language Testing**, v. 17, n. 1, p. 65-83, 2000.

WEST, M. **A general service list of English words**. Essex: Longman Harlow, 1953.

WHAT is the BNC? British National Corpus. 2004. Disponível em:
<<http://www.natcorp.ox.ac.uk/what/index.html>>. Acesso em: 28 abr. 2005.

XUE, G.; NATION, I. A university word list. **Language Learning and Communication** , v. 3, n. 2, p. 215-229, 1984.