GERAÇÃO DO SISTEMA CORPOR − SISTEMA GERENCIADOR DE BANCO DE DADOS RELACIONAL

Armazenamento das informações

A constituição das *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo* vem passando por diversas etapas de acordo com a disponibilidade dos recursos de *hardware* e de *software*. A primeira etapa dedicou-se ao armazenamento dos dados em *mainframes*; a segunda, ao armazenamento dos dados anteriormente processados em *mainframes* em planilhas do *Excel* – conversão da plataforma *mainframe* para plataforma PC; a terceira, à conversão dos dados para *Sistema Gerenciador de Banco de Dados Relacional*; a quarta, à expansão do Sistema para acesso remoto através de *download* via FTP, ou seja, para transferência de arquivos remotamente em redes de computadores. A quinta etapa pretende tornar disponível o *Sistema CorPor* em ambiente Web, para acesso universal.

A fim de melhor esclarecer, seguem os procedimentos metodológicos das diversas etapas, com exposição dos ambientes de programação, bancos de dados e *softwares* utilizados.

– Plataforma *Mainframe*

A pesquisa para o doutorado, concluída em 1980, foi estruturada em bases de dados gerados em *mainframes*, o que, como já observado na seção 1 – Contextualização −, embora tenha representado avanço em relação ao trabalho de compilação e análise manuais de *corpora*, exigiu uma equipe de profissionais da computação – analistas de sistemas, programadores, operadores, perfuradores de cartões −, além de estatistas.

– Plataforma PC

As *Bases de Informações Ortográfico-Fonéticas* processadas em *mainframes* foram convertidas para PCs, inicialmente, com sistema operacional da *Microsoft Windows* através do *software Excel* (versão 2007) − ferramenta do conjunto de programas do *Office* da *Microsoft*, destinada ao planilhamento de dados, apropriada, portanto, para

operações com dados dispostos em forma de tabelas, trabalhos sobre dados com as facilidades básicas de um banco de dados relacional e apresentações de dados numéricos sob a forma de gráficos.

Esclarecemos que, em virtude do volume grande de dados, a quantidade de registros da Base – cerca de 180 mil – excedeu a capacidade de processamento da versão do Excel 2003 − versão então disponível, limitada a 65.536 linhas −, tendo sido necessária a utilização da versão beta do Excel 2007. Esse procedimento possibilitou a recuperação de dados autênticos do português falado paulista da década de 1970, coletados para a tese de doutorado.

Posteriormente, os dados em planilhas do *Excel* foram implementados em um ambiente específico de Banco de Dados, ou seja, foram convertidos para um *Sistema de Banco de Dados Relacional*[1]. As *Bases de Informações* estão, assim, armazenadas no *Sistema CorPor*, Sistema de Banco de Dados Relacional, e são manipuladas por meio do Sistema Gerenciador de Banco de Dados (SGBD) − um conjunto de programas computadorizados (*software* ou ferramenta), desenvolvidos numa determinada linguagem, que possibilitam o gerenciamento das funções de edição, consulta, controle e remoção de registros, campos ou tabelas de um Banco de Dados. O SGBD utilizado é o *Firebird*[2]. Esse procedimento acrescenta, em relação à etapa anterior, a possibilidade de se estabelecerem relacionamentos entre os dados do Banco para a extração e análise de novas informações.

A estrutura dos dados segue o modelo relacional, conforme Diagrama de Registro do Informante (Figura 2), havendo uma correspondência entre os campos da tabela

[1] Um Banco de Dados Relacional é um banco de dados que segue o Modelo Relacional – armazena, manipula e recupera dados estruturados unicamente na forma de tabelas que constroem um banco de dados. O termo pode ser aplicável aos próprios dados, quando organizados dessa forma, ou a um Sistema Gerenciador de Banco de Dados Relacional – um programa de computador.

2 Firebird, algumas vezes chamado de FirebirdSQL, é um sistema gerenciador de banco de dados. Roda em *Linux*, *Windows*, *Mac OS* e em uma variedade de plataformas *Unix*. Baseado no código do *InterBase* da *Borland*, quando da abertura de seu código na versão 6.0 (em 25 de Julho de 2000), alguns programadores em associação assumiram o projeto de identificar e corrigir inúmeros problemas da versão original, surgindo aí o *Firebird 1.0*, que se tornou um banco com características próprias e com aceitação imediata entre os programadores. A versão mais recente estável é a 2.1.2. A versão 2.5 está em fase beta e trará uma nova arquitetura chamada *SuperClassic*, que fará a ponte para a versão 3.0. Gratuito em todos os sentidos − não há limitações de uso, e seu suporte é amplamente discutido em listas na Internet, o que facilita a obtenção de ajuda técnica −, o produto *Firebird* é seguro e confiável, suportando sistemas com centenas de usuários simultâneos e bases de dados com dezenas/centenas de *gigabytes*. É bastante utilizado em todo o mundo, com a maior base de usuários no Brasil, Rússia e Europa.

principal do banco de dados e os do diagrama. As Bases constituem, assim, uma coleção de dados ortográficos e fonéticos do português falado de São Paulo, organizados, relacionados e armazenados em função de anotações linguísticas e extralinguísticas, com as diferentes relações existentes entre os dados armazenados.

Como a criação do SGBD pressupõe o uso de um ambiente de programação, ou linguagem de programação, buscamos uma ferramenta de desenvolvimento capaz de acessar um banco de dados.

O ambiente de programação utilizado é o *Delphi*, produzido pela *Borland Software* Corporation. O *Delphi* é um compilador[3] e um ambiente integrado de desenvolvimento de aplicações - IDE - *Integrated Development Environment*[4], que utiliza a Linguagem Pascal com extensões orientadas a objetos[5] – *Object Pascal*[6] –, associada a recursos da Linguagem Estruturada de Pesquisa[7] (*Structured Query Language* – SQL [8]).

A linguagem estruturada de pesquisa SQL não é uma linguagem especificamente criada para desenvolver sistemas, como o são as linguagens de programação, a exemplo do *Delphi*. A SQL é tão-somente uma linguagem utilizada para facilitar o acesso às informações – por meio de consultas, atualizações, extrações e manipulações de dados – armazenadas em Bancos de Dados do tipo relacional. Possibilita o uso de linhas de comandos de pesquisa sem ser preciso o uso de programação, o que facilita a sua utilização por usuários não-especializados em programação.

Por incluir um conjunto de funções pré-implementadas que gerenciam as operações de inserção, remoção, atualização e consulta dos dados armazenados e demais tarefas próprias do gerenciamento de Banco de Dados, o *Delphi* possui um ambiente completo para as atividades de manipulação e de acesso às estruturas de Bancos de Dados Relacionais.

---

[3] Um compilador é um programa que transforma um código escrito em uma linguagem – código-fonte (do inglês *source code*) – em um programa equivalente em outra linguagem – código-objeto (do inglês *object code*).
[4] Ambiente Integrado de Desenvolvimento.
[5] Linguagem de alto nível − as sintaxes são mais próximas da língua natural.
[6] O *Object Pascal*, a partir da versão 7, passou a se chamar *Delphi Language*.
[7] Ou Linguagem Estruturada de Consulta.
[8] Comercialmente implementada pela IBM, tornou-se um padrão de linguagem de acesso a dados em vários bancos de dados relacionais, como Oracle, DB2, SQL Server, Sybase, Interbase etc.

Além de recursos de pesquisa – para o acesso às informações das Bases –, o Sistema abrange recursos de um editor de textos – para os trabalhos de edição dos resultados das pesquisas às Bases de Informações.

Para acesso dos usuários e para pesquisas por meio dos comandos da linguagem SQL, não somente as Bases de Informações Ortográfico-Fonéticas, como também *Corpora* e Léxicos (Dicionários) gerados a partir delas, integram o Sistema CorPor, cada um deles compondo um módulo (ou componente) com seus registros e campos.

O armazenamento das Bases em formato específico de Banco de Dados Relacional tem o propósito de oferecer a estudiosos do português facilidade, rapidez e confiabilidade na pesquisa (consulta), na recuperação (acesso) e no tratamento (exploração) automáticos de extensos e variados dados autênticos do português paulista para o desenvolvimento de estudos de aspectos diversos da língua – fonéticos, fonológicos, lexicais, morfológicos, sintáticos, textuais e discursivos − e para o desenvolvimento de sistemas de processamento da fala.

Trata-se da versão para *desktop* do *Sistema CorPor* em plataforma compatível com *Windows XP* da *Microsoft*.

Para acesso aos pesquisadores, a versão *desktop* do *Sistema CorPor* está hospedada no *site* http://www.corpor.fflch.usp.br, para que o *Sistema* possa ser acessado via FTP e, pois, para que o seu *download* possa ser feito para a máquina do pesquisador através de transferência de dados em redes de computadores. Trata-se de um protocolo genérico independente de *hardware* e de sistema operacional, que transfere arquivos por livre arbítrio, tendo em conta as suas propriedades e restrições de acesso. A transferência de dados em redes de computadores envolve normalmente transferência de arquivos e acesso a sistemas de arquivos remotos com a mesma interface usada nos arquivos locais.

O Sistema Corpor atual é, pois, um sistema monousuário, de computador de mesa, sem disponibilidade na Web − o usuário faz o *download* do Sistema, instala-o em sua máquina, onde é feito o processamento.

O Sistema está disponível para *download*, em versão beta, para a comunidade acadêmica em geral, para, de um lado, com ela compartilhar parte dos muitos anos de utilização de tecnologias informatizadas nos estudos linguísticos; de outro, para que os usuários possam reportar dificuldades e problemas encontrados, e apresentar sugestões para a sua melhoria.

– Plataforma Web

Apesar de a disponibilidade do *Sistema CorPor* através de FTP representar avanços significativos em relação aos produtos das etapas anteriores, destacam-se algumas desvantagens, como a ocorrência de problemas na instalação no micro do usuário devido à incompatibilidade de sistemas operacionais, de interfaces e de requerimentos de memória.

Para tornar o *Sistema CorPor* acessível aos interessados de maneira mais fácil, rápida, segura e amigável, tem-se a intenção de disponibilizá-lo em plataforma Web, seguindo as tendências atuais de produção, armazenamento e distribuição de conteúdos, o que significa converter o sistema atual em outro sistema com ferramentas web, de forma a viabilizar a sua utilização e pesquisa *on-line*, em tempo real.

Assim sendo, uma das grandes vantagens da versão Web consistirá na possibilidade de acesso universal às informações do *Sistema CorPor* com maior rapidez, segurança e confiabilidade, por não depender de recursos do computador do usuário – o usuário opera o sistema *on-line*. O Sistema, nesse ambiente, poderá ser acessado através de qualquer dispositivo (microcomputador, *notebook*, *netbook*, celular) que tenha comunicação com a *Internet*, portanto, de forma democrática, com controle centralizado e independência geográfica.

Recuperação das informações

O Sistema de Banco de Dados no ambiente *Delphi* – com o Banco Informatizado do Português Falado de São Paulo, numa estrutura de anotação de variáveis linguísticas e extralinguísticas por níveis e subníveis, e com recursos de pesquisas da linguagem SQL e de um editor de textos – oferece a vantagem de facilidade e flexibilidade para a

recuperação automática de um bom número de léxicos e de *corpora* de estudo. É possível extrair desde o *corpus* integral e conjunto, constituído pelo total das informações das 432 interações dialógicas realizadas com os 216 informantes, até diferentes *subcorpora* quantas são as variáveis que foram controladas e indexadas – *corpus* por regiões, por sexos, por faixas etárias, por níveis de escolaridade, por níveis socioeconômicos, por condições extraverbais de produção do diálogo –, *corpora* menores – constituídos pelo conjunto das informações de seis inquéritos – e muitos outros, de dimensões várias, pelas possibilidades de cruzamentos das variáveis anotadas.

Ou seja, como já observado e se pôde visualizar nos dois diagramas apresentados – *Diagrama de Distribuição dos Informantes* e *Diagrama de Registro do Informante* –, a estrutura do Sistema permite a recuperação dos dados por quaisquer campos ou pelo cruzamento deles – todos os campos podem ser cruzados –, do que resultam tantos *corpora* quantas são as variáveis linguísticas e extralinguísticas anotadas e suas diferentes possibilidades combinatórias, para posterior tratamento por programas de análise linguística.

Seguem algumas possibilidades de composição de *corpora*:

- *Corpus* integral: inquéritos formais e informais de 216 informantes.
- *Corpus* por condições extraverbais de produção dos diálogos, com 216 inquéritos cada um: (a) formal; (b) informal.
- *Corpus* por região, com 72 inquéritos formais e 72 inquéritos informais cada um: (a) São Paulo; (b) Campinas; (c) Itu.
- *Corpus* por sexo, com 108 inquéritos formais e 108 inquéritos informais cada um : (a) masculino; (b) feminino.
- *Corpus* por nível de escolaridade, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) superior completo; (b) superior incompleto – último ano; (c) ensino médio – último ano; (d) ensino fundamental II – último ano; (e) ensino fundamental I – último ano; (6) analfabeto.

- *Corpus* por faixa etária, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) a partir de 25 anos – superior completo; (b) 20 a 24 anos; (c) 17 a 19 anos; (d) 14 a 16 anos; (e) 10 a 13 anos; (f) a partir de 25 anos –.analfabeto.
- *Corpus* por faixa etária para os informantes de curso superior completo e para os analfabetos – subdivisão em seis faixas etárias, com uma dupla (homem / mulher) em cada cidade, num total de 12 informantes de curso superior completo e de 12 analfabetos em cada região, com os dois tipos de inquéritos: (a) 25 a 29 anos; (b) 30 a 34 anos; (c) 35 a 39 anos; (d) 40 a 44 anos; (e) 45 a 49 anos; (f) 50 a 54 anos.
- *Corpus* por nível socioeconômico, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) classe alta alta; (b) classe alta; (c) classe média alta; (d) classe média baixa; (e) classe baixa; (f) classe baixa baixa.

Alguns cruzamentos possíveis:

- *Corpus* por região / tipo de diálogo, com 72 inquéritos cada um: (a) São Paulo / formal; (b) São Paulo / informal; (c) Campinas / formal; (d) Campinas / informal; (e) Itu / formal / (f) Itu / informal.
- *Corpus* por região / sexo, com 36 inquéritos cada um: (a) São Paulo / masculino; (b) São Paulo / feminino; (c) Campinas / masculino; (d) Campinas / feminino; (e) Itu / masculino; (f) Itu / feminino.
- *Corpus* por região / nível de escolaridade, com 12 inquéritos cada um: (a) São Paulo / superior completo; (b) São Paulo / superior incompleto – último ano; (c) São Paulo / médio – último ano; (d) São Paulo / fundamental II – último ano; (e) São Paulo / fundamental I – último ano; (f) São Paulo / analfabeto; (g) Campinas / superior completo; (h) Campinas / superior incompleto – último ano; (i) Campinas / médio – último ano; (1j) Campinas / fundamental II – último ano; (k) Campinas / fundamental I – último ano; (l) Campinas / analfabeto; (m) Itu / superior completo; (n) Itu / superior incompleto – último ano; (o) Itu / médio – último ano; (p) Itu / fundamental II – último ano; (q) Itu / fundamental I – último ano; (r) Itu / analfabeto;

- *Corpus* por região / sexo / nível de escolaridade, com 6 inquéritos cada um: (a) São Paulo / masculino / superior completo; (b) São Paulo / feminino / superior completo; (c) São Paulo / masculino / superior incompleto – último ano; (d) São Paulo / feminino / superior incompleto – último ano; assim por diante.
- *Corpus* por região / tipo de diálogo / sexo / nível de escolaridade, com 6 inquéritos cada um: (a) São Paulo / formal / masculino / superior completo; (b) São Paulo / informal / masculino / superior completo; (c) São Paulo / formal / feminino / superior completo; (d) São Paulo / informal / feminino / superior completo; (e) São Paulo / formal / masculino / superior incompleto – último ano; (f) São Paulo / informal / masculino / superior incompleto – último ano; (g) São Paulo / formal / feminino / superior incompleto – último ano; (h) São Paulo / informal / feminino / superior incompleto – último ano; assim por diante.
- *Corpus* por região / nível socioeconômico, com 12 inquéritos cada um: (a) São Paulo / classe alta alta; (b) São Paulo / classe alta; assim por diante.
- *Corpus* por região / sexo / nível socioeconômico, com 6 inquéritos cada um: (a) São Paulo / masculino / classe alta alta; (b) São Paulo / feminino / classe alta; assim por diante.

A exploração de cada *corpus* por programas gerenciadores do léxico, além de apresentar o léxico por variável ou conjunto de variáveis selecionadas na sua extração, permite estudos contrastivos com os tratamentos efetuados para outras composições.

Assim, a língua portuguesa, observada numa perspectiva sincrônica, é passível de ser avaliada na sua diversidade: diferenças entre comunidades regionais, diferenças entre sexos, diferenças entre níveis de escolaridade, diferenças entre gerações, diferenças entre meios sociais, diferenças ligadas às condições de produção do diálogo.

O Sistema também viabiliza a cópia automática para transferência de resultados de pesquisa em função de interesses de estudo para outros *softwares*.

Tratamento das informações

Para o tratamento (exploração) automático de *corpora*, o pesquisador pode contar, hoje em dia, com uma grande variedade de programas computacionais, que facilitam o estudo, a análise e a aplicação dos dados informatizados.

Há programas disponíveis que são indexadores e servem para a busca textual – permitem a indexação das palavras de um texto, ou seja, a identificação de sua localização no texto, a recuperação por listagens em forma de concordâncias (o conjunto de ocorrências de cada palavra, em ordem alfabética, com seu contexto imediato e sua localização). Possibilitam, também, a busca de colocados (de combinações de palavras - listas de palavras que ocorrem à esquerda e à direita da palavra de busca selecionada, em ordem de frequência) e de padrões de colocados (frases comuns - palavras que co-ocorrem com outras com certa frequência), bem como a pesquisa de grupos de palavras (com o uso de coringas e expressões lógicas, é possível a busca de palavras que guardam alguma relação). Os programas ainda permitem um tratamento quantitativo dos dados e alguns, quantiqualitativo.

A presente pesquisa dá destaque ao emprego do programa computacional *Stablex* – geração de léxicos, indexação, extração de sequências e concordâncias, lematização, tratamento estatístico (André Camlong[9] e Thierry Beltran, Universidade de Toulouse II). Desenvolvido inicialmente para a plataforma *Macintosh* (1991), o programa *Stablex* conta, a partir de 2004, com a sua versão PC, que inclui novas funções estatísticas para a análise de textos.

Por contemplar uma confluência de áreas – Linguística, Matemática, Estatística, Computação –, o *Stablex* facilita e otimiza não somente a busca, organização e quantificação, mas também a análise de dados linguísticos – realiza uma análise preliminar dos dados a partir de um tratamento lexical quantiqualitativo. A análise

[9] André Camlong, de formação filosófica, linguística, matemática e estatística, é Professor Titular na Universidade de Toulouse II e Diretor no CRIC, Maison de la Recherche, da mesma universidade. Desde 1994, a partir de visita de colaboração à FFLCH/USP, a nosso convite e através de auxílio da Fundação de Amparo à Pesquisa do Estado de São Paulo, vem prestando assessoria a pesquisadores brasileiros na utilização do programa e do método de sua autoria.

quantitativa de textos é ponto de partida para a análise qualitativa. Assim sendo, esse programa atende às necessidades do pesquisador cujo objeto de trabalho é o texto e o discurso.

Destaca-se o fato de o programa ter sido desenvolvido em função de um modelo de análise lexical, textual e discursiva – *método matemático-estatístico-computacional de análise de textos* de André Camlong. Trata-se, por conseguinte, não apenas da aplicação de um programa computacional, mas, de forma mais ampla, de um programa que serve de ferramenta para um método de análise de textos e que, em função disso, reúne recursos computacionais, matemáticos e estatísticos.

O método é fundado na matemática e na estatística paramétrica (estatística descritiva); possibilita o estudo descritivo, objetivo e indutivo do texto; permite a análise quantiqualitativa do léxico, que indica apontamentos para a análise textual e discursiva.

O método matemático-estatístico-computacional de tratamento e análise de textos de André Camlong está descrito nas obras *Méthode d'analyse lexicale textuelle et discursive,* de André Camlong (1996) e *Do Léxico ao Discurso pela Informática,* de Zilda Maria Zapparoli e André Camlong (2002).

A fim de que o *corpus* possa receber tratamento estatístico, a aplicação do programa pressupõe uma preparação[10] do *corpus* em função dos objetivos do estudo – tarefa manual, que exige tempo e cuidado:

- harmonização gráfica do texto: como se trata de *corpus* de língua oral, uniformização das diferentes representações gráficas de um mesmo item, decorrentes de transcrições de especificidades da língua falada, como diferentes grafias por variações de pronúncia – quando as duas formas são dicionarizadas (*líquido* e *líquido* para *líquido*; *quatorze* e *catorze* para *catorze*; *cousa* e *coisa* para *coisa*, *contacto* e *contato para contato*, *questão* e *questão* para *questão*;[11]

[10] Com base em Camlong, 1996, p. 9-12.

[11] São mantidas ocorrências de formas alomórficas não-dicionarizadas, empregadas em desacordo com a norma culta, como *abuzinava, alembrado, contrareia, esteje, ponhar, vareia,* por refletirem, de um lado,

- reconstituição de sintagmas referenciais: para preservar-se a unidade referencial, evitando-se, no processamento, mais de uma entrada para o mesmo item lexical, recuperam-se as lexias compostas, complexas ou textuais, ou seja, expressões referenciais fixas, como nomes próprios de pessoas, cidades, obras, instituições, expressões ou frases feitas, através da substituição dos espaços em branco por traços de união[12]. Para esse trabalho, utiliza-se, quando necessário, a técnica da comutação e a consulta a dicionários. Por exemplo**:** *Serra-Negra, São-Paulo, Rio-de-Janeiro, Curso-de-Complementação-Pedagógica, Maria-Aparecida, Editora-Abril, ponto-de-vista, regra-geral, hoje-em-dia, mil-novecentos-e-quarenta-e-oito, por-que, por-quê*;
- reconstituição de *a-gente* no emprego de 1ª pessoa do plural, dada a sua função no estudo dos sentidos e dos papéis das categorias de pessoa;
- exclusão de comentários descritivos do transcritor, por não pertencerem ao discurso do informante.

A aplicação da abordagem de análise quantiqualitativa do léxico na exploração de *corpora* inclui:

- levantamento lexical com constituição de *léxicos de frequência*, em que os itens lexicais são classificados por ordem alfabética, por ordem crescente de frequência e por ordem decrescente de frequência;
- criação da *Tabela de Distribuição de Frequências – TDF* – cálculo aritmético – tratamento quantitativo;
- criação da *Tabela de Desvios Reduzidos – TDR* – cálculo algébrico – tratamento quantiqualitativo;
- determinação do grau de normalidade da distribuição lexical das variáveis pela aplicação do *teste estatístico do $\chi 2$ de Fisher*;

---

falta de conhecimento do item lexical adequado, de outro, uso restrito de certas coletividades, o que justifica entrada independente no léxico.

[12] O apóstrofo e sinais de pontuação (ponto, ponto-e-vírgula, dois-pontos etc.) são considerados separadores, enquanto que o traço de união é considerado unificador.

- criação de Léxicos Preferenciais, isto é, de *Tabelas de Valores Lexicais* – distribuição preferencial dos itens lexicais, ou seja, ordenação dos itens lexicais por ordem decrescente de preferência de emprego no texto;
- constituição de vocabulários preferenciais, básicos, diferenciais, exclusivos, a partir da estratificação do léxico preferencial em diferentes vocabulários, que destacam as características de emprego dos itens lexicais e os elementos fundamentais da estrutura temática e articuladora do discurso;
- constituição de vocabulários específicos pela técnica da lematização – em função de finalidades do estudo, destacam-se itens lexicais por associação léxica, semântica ou temática, e calcula-se o valor do novo vetor obtido pela lematização;
- extração de sequências textuais, ou seja, de recortes discursivos, por recurso aos textos;
- aplicação do teste da correlação para análise do grau de ligação existente entre as variáveis.

Em função de interesses de estudo, recorre-se, também, ao programa *WordSmith Tools*, em especial ao uso de suas ferramentas que possibilitam a busca de colocados e de padrões de colocados.

De autoria de Mike Scott, Universidade de Liverpool, o programa *WordSmith Tools* é publicado pela Oxford University Press e distribuído via *World Wide Web* (www.liv.ac.uk/~ms2928). Disponível para PC/Windows 98, NT, 2000 e XP em sua quarta versão. Possui interface gráfica.

WordSmith Tools começou a ser disponibilizado aproximadamente em 1995, quando era composto por ferramentas separadas – W*ordlist, Concord, Keywords*. Hoje, consiste num conjunto integrado de recursos para análise lexical: três ferramentas (as três citadas, que ainda são o centro do programa) e quatro utilitários (*Splitter, Text Converter, Dual Text Aligner, Viewer*), que, juntos, reúnem dezessete instrumentos de análise.

Cada um dos recursos do programa é usado para tarefas específicas de análises de textos:

- *Wordlist*: gera listas de palavras em ordem alfabética e em ordem de frequência, e listas de estatísticas dos textos (dimensões e densidade lexical);
- *Concord*: ferramenta, por excelência, para análise lexical, cria concordâncias das palavras de busca, gera listas de colocados, listas de padrões de colocados, listas de agrupamentos lexicais, e exibe um mapa gráfico que mostra onde a palavra ocorre no *corpus*;
- *Keywords*: lista palavras-chave de um dado texto através de comparações entre listas de palavras de arquivos diferentes quanto à sua frequência relativa. Exibe um mapa gráfico que mostra onde cada palavra-chave ocorre no *corpus*;
- *Splitter*: divide grandes arquivos em diversos menores;
- *Text Converter*: recurso de procura e substituição, reformata um número grande de textos;
- *Dual Text* Aligner: alinha dois textos, possibilitando a sua comparação por períodos ou parágrafos;
- *Viewer*: exibe o texto de origem.