Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Uma análise introdutória de ferramentas para produção de dicionários em ambiente MS Windows

José Luiz De Lucca

Maria das Graças Volpe Nunes

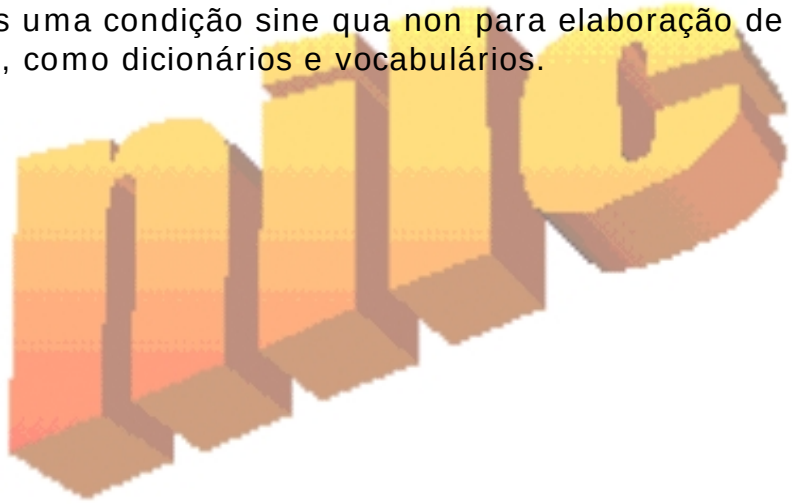
NILC-TR-02-20

Outubro 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta uma análise técnica e sucinta de alguns dos principais traços de softwares não web que manipulam corpora. Nossa análise não visa apontar qual a melhor ferramenta, mas como estas ferramentas trabalham e quais suas características que as diferenciam umas das outras. Os resultados servirão para orientar-nos na construção de uma ferramenta que atenda as necessidades de muitos autores de obras de referência, cuja base de dados origina-se em corpora, o que consideramos uma condição sine qua non para elaboração de certas obras de referência, como dicionários e vocabulários.



Índice

Uma análise introdutória de ferramentas para produção de dicionários em ambiente MS Windows

1. Introdução

No momento não existe, pelo que sabemos, uma análise das ferramentas de software que se destinam, direta ou indiretamente, à produção de obras de referência, o que prejudica sensivelmente a escolha da ferramenta certa. Embora exista listagem de produtos específicos da área, pouco se conhece deles no nível de desempenho e interface. Para preencher esta lacuna, partimos do pressuposto de que o manuseio de ferramentas computacionais não é relevante apenas para lingüistas e especialistas em lingüística computacional, mas também para professores e estudantes de língua materna e estrangeira, tradutores etc. Faz-se necessária, portanto, uma análise introdutória de alguns dos produtos do mercado, objeto e tema deste relatório - uma análise das ferramentas para produção de dicionários, embora nem todas tenham isso como fim, acabam por colaborar na confecção de obras de referência.

A maioria dos autores de obras de referência - dicionários, vocabulários, glossários, almanaques, etc. - utilizam, em geral, duas ferramentas: um banco de dados e um processador de textos; alguns apenas um banco de dados; outros somente um processador de textos. Não discutiremos, contudo, este tipo de ferramenta, digamos básica, de todo (ou quase todo) autor de obra de referência.

Atualmente lingüistas e programadores têm procurado desenvolver inúmeras ferramentas para os mais diversos fins na área conhecida como PLN (Processamento de Linguagem Natural), e é sobre algumas destas obras que iremos centrar nossa análise.

As ferramentas que analisamos[1], são as seguintes:

WORDSMITH TOOLS 3.0 (Mike Scott), KWIC Concordance 4.6 (Satoru Tsukamoto), CORPUS WIZARD (Takashi Hamaguchi), RANGE andWORD (Alex Heatley. Paul Nation and Averil Coxhead), MONOCONC Pro 2.0 (Michael Barlow Athelstan), CONCORDANCE 2.0 (R.J.C. Watt), TATOE –

[1] Para a análise dos produtos acima, foi utilizado um microcomputador ATHLON 1.1GHZ, 128RAM e 20GB harddisk, durante os meses de agosto/setembro de 2002.

Text Analysis Tool with Object Encoding (Melina Alexa, Lothat Rostek), TACT (TACT Group, University of Toronto) e DICTGEN 1.0b (Martin S. Mamo)

2. Critérios de Análise

Os critérios de análise foram baseados nas características em comum que os produtos possuem, como por exemplo, layout, limitações, concordanceador, etc. Também foi usado como critério de escolha aquilo que a maioria dos programas não possuem, como por exemplo, a lematização, que embora não seja um recurso presente na maioria dos programas, linguisticamente falando é muito importante em nosso ponto de vista, daí a razão de tê-la incluído.

2.1.0 Layout

TABELA I - LAYOUT

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
RUIM	REGULAR	REGULAR	BOM	REGULAR	BOM	RUIM	RUIM	RUIM

Análise: TACT - Não pelo fato de ser em MSDOS, mas por não haver um menu tipo pull-down incorporando todas as funções, é o de mais difícil interatividade com o usuário e sem apoio visual.

KWIC - Não apresenta os resultados na tela, é preciso abrir um processador de textos para ver os resultados.

RANGE AND WORD trabalham com duas telas para executar suas tarefas. Uma abre e salva os arquivos de trabalho (RANGE) e a outra (WORD) conta e ordena as palavras do(s) arquivo(s) aberto(s). Como são programas diferentes, isto obriga o usuário a entrar duas vezes com o nome do arquivo ou nomes dos arquivos, caso queira compará-los (RANGE) e Contar palavras e sort (WORD).

WORDSMITH TOOLS - A organização das telas inclui um forte apoio visual baseado em ícones, o que, de certa forma, economiza o acesso do usuário a determinadas funções.

DICTGEN - Além das poucas funções, a falta de um mouse obriga o usuário a muita digitação, além de obrigar o usuário a ter seus arquivos de trabalho no diretório do software.

CORPUS WIZARD - Embora tenha um número razoável de botões, estes não possuem HINTS. Deste modo quando o mouse passa pelo botão, não aparece nenhum HINT sob eles, o que os faz perder sua razão de ser.

MONOCONC - Tem a grande vantagem de abrir no próprio menu os arquivos de trabalho, normalmente .txt, ou de tipo html, contudo seu menu de navegação tem apenas duas barras, uma de funções, outra de informações sobre o produto e ajuda.

TATOE - Talvez para fugir do tradicional, possui um menu lateral e as telas em azul.

CONCORDANCE - Na tela principal apresenta de um lado a lista de palavras com sua freqüência e, de outro, as concordâncias. Diferentemente de WORDSMITH TOOLS e outras ferramentas similares, esta não faz a busca por palavras, o usuário escolhe no ListBox à esquerda a palavra a procurar.

2.1.1. Limitações

TABELA II - LIMITAÇÕES

	KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEI	TACT
Sentences	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM

Análise:

WORDSMITH TOOLS tem como limite qualquer arquivo com mais que 16.368 sentenças (o arquivo, em si, pode conter muito mais sentenças), arquivos ou diretórios. Individualmente os limites de cada ferramenta seriam os seguintes:

Concord permite que a palavra ou frase buscada tenha no máximo 80 caracteres, entretanto o usuário pode especificar até 300 palavras de busca ou frases em um arquivo de palavra de busca ou frase. Cada concordância pode acumular até 16.368 colocações e dispõe no máximo 25 palavras à esquerda e à direita da palavra-guia ou headword. Quanto aos Tags é possível processar até 2.000 em cada 30.000 caracteres de texto processados.

Wordlists individuais podem conter mais de 8 milhões de entradas. Listas de Consistência detalhadas podem controlar até 50 arquivos.

KeyWords lists podem conter até 16.368 palavras-chave separadas. Um banco de dados de palavras-chave pode conter dados de 16.368 arquivos de texto separados.

Splitter - Cada linha de um arquivo de texto grande pode ter até 10.000 caracteres em comprimento.

Text Converter - Converte até 16.368 arquivos separados em um lote. Pode haver até 500 cadeias para buscar ou substituir. Cada cadeia pode ter até 80 caracteres de comprimento. Um asterisco não deve ser o primeiro ou último caráter da cadeia de busca. Quando o asterisco é usado para reter informação o limite é de 1.000 caracteres.

Text Viewer – Ao teclar o View Button, ao escolher textos, este chamará os primeiros 9 arquivos de texto selecionados. Cada um pode ter até 16.368 orações/sentenças. Cada oração pode estar até 10.000 caracteres (aproximadamente 1.600 palavras).

KWIC Não informa seus limites, mas através dos testes realizados não consegue trabalhar com um arquivo com mais de 49.000 sentenças.

CORPUS WIZARD – Também não informa seus limites. Os testos foram prejudicados tendo em vista que a versão que possuimos processa apenas 3.900 concordâncias.

RANGE AND WORD – Trabalha com arquivos-texto muito grandes, mais de 2.400.000 palavras ou 102.000 sentenças. Seu limite é o tamanho das sentenças, no máximo 254 caracteres.

MONOCONC – Em sua versão demo, seu limite é de 20 sentenças.

CONCORDANCE 2.0 – Trabalha com mais de 17.000 concordâncias.

TATOE – Não foi possível avaliar seus limites.

DICTGEN – Não trabalha com arquivos grandes.

TACT – Ele abre arquivos grandes, mas trabalha apenas as primeiros 147.000 palavras.

2.1.2. Recursos

TABELA III - RECURSOS

	KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
Keywords	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM
Concord	SIM	SIM	NÃO	SIM	SIM	SIM	SIM	NÃO	SIM
Colloc	SIM	NÃO	NÃO	NÃO	SIM	SIM	NÃO	NÃO	NÃO
Estatist	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	NÃO	SIM

Análise: DICTGEN e **TACT** trabalham em ambiente MSDOS, as demais em ambiente Windows. **KWIC**, **WORDSMITH TOOLS e CONCORDANCE** são os softwares que apresentam a maior gama de recursos. **RANGE** e **DICTGEN**, por terem fins distintos, não apresentam todas as ferramentas acima citadas.

KWIC ainda apresenta Setup para várias línguas incluindo um Tipo de Corpus, além de um índice de rimas.

CORPUS WIZARD apresenta uma chave de busca por grupo de palavras, em que o usuário indica a palavra mais a classe gramatical da seguinte palavra a ser associada, como na Fig. 1 abaixo:

"get" + "Preposition."

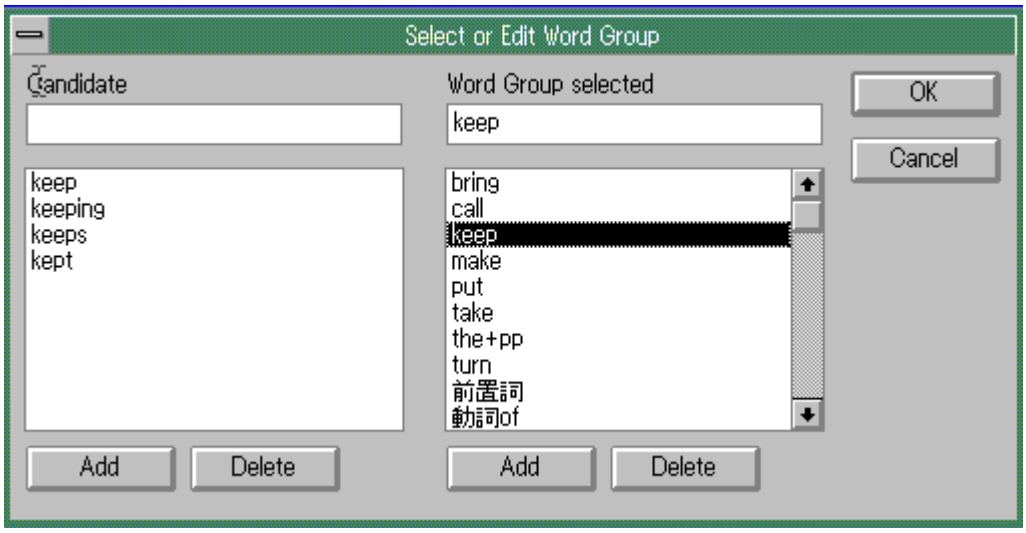


Fig. 1. Tela ilustrativa do Corpus Wizard:

Tanto **RANGE,** quanto **DICTGEN** apresentam como característica principal o vocabulário comum entre listas de palavras, especialmente **RANGE**, o qual possibilita a comparação de até 32 listas de palavras diferentes.

MONOCONC permite configurar a linguagem e realizar a busca através de Expressões Regulares, por texto ou por tags.

WORDSMITH TOOLS

Além das três ferramentas básicas: concordanceador, manuseio de palavras-chave e listas de palavras, possui, também, quatro utilitários: um gerenciador de arquivos, um conversor de textos, um divisor, e um visualizador e alinhador. O gerenciador de arquivos é o programa que "administra" todas as ferramentas do software; Splitter ou divisor divide ou fraciona grandes arquivos-texto em pequenos; Conversor de textos realiza tarefas tais como, editar textos, renomeá-los, busca e troca etc.; visualizador é a ferramenta que possibilita examinar os arquivos de trabalho em diversos formatos. Também em Ambientes é possível configurar a linguagem e formato do arquivo a ser aberto. WORDSMITH TOOLS também faz um índice de rimas através de finais de palavras.

CONCORDANCE – Possibilita a escolha da linguagem ou alfabeto. Possui um múltiplo editor de textos e um breve lematizador em inglês.
A um simples clique é possível ver a palavra-guia no contexto. Organiza os contextos pela desinência da palavra, criando assim como KWIC, um índice de rimas.

TATOE – Como é um produto em elaboração, falta muitas informações sobre ele. Dentre seus traços particulares salientam-se um esquema de categorização e um segmento de textos marcados. Em Tipos de palavras apresenta a palavra no contexto. Seu concordanceador apresenta as palavras-guia, tanto alinhadas verticalmente, como no contexto, coforme Fig. 2 abaixo.

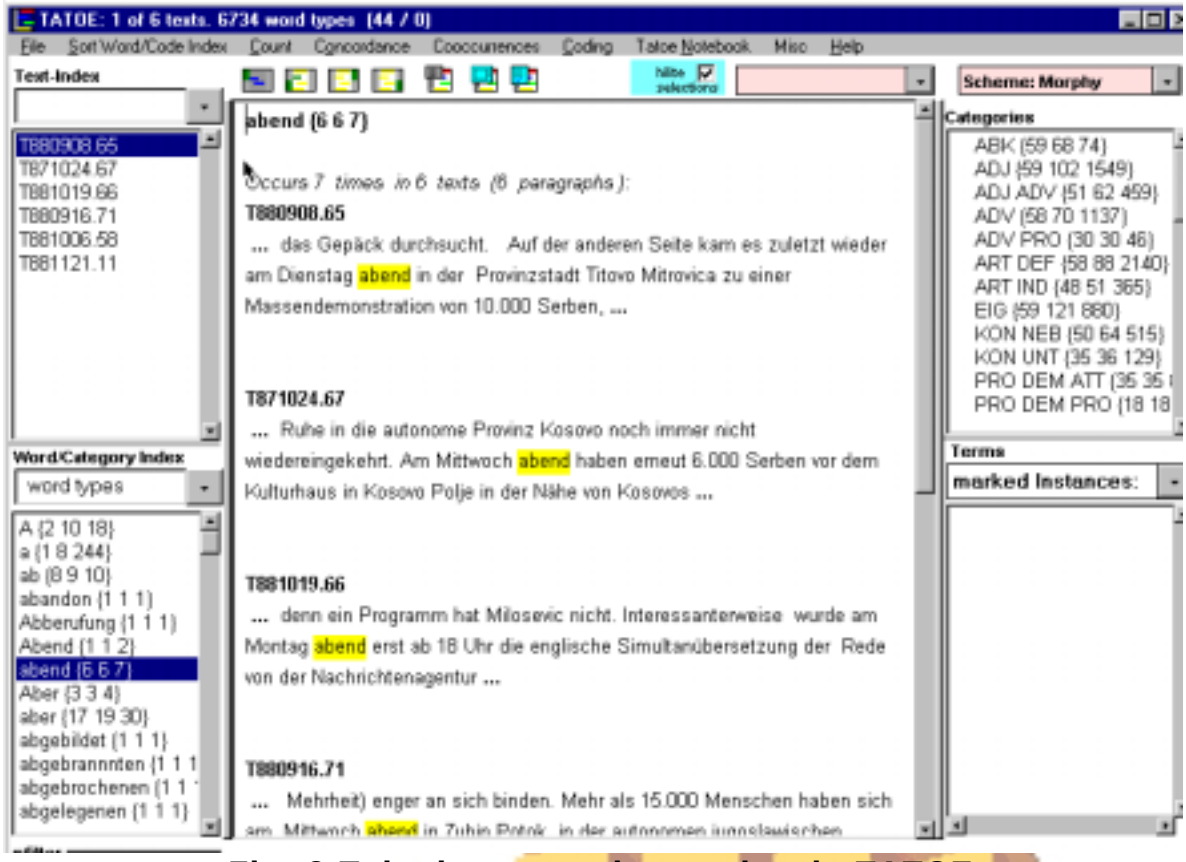


Fig. 2 Tela do concordanceador do TATOE

TACT – **Fica bastante transparente em** TACT**, diferentemente dos demais, a forma como trabalha. Importa arquivos em formato ASCII e os converte em um banco de dados de onde começa sua manipulação dos dados.**

2.1.3 Função de Concordanciador

TABELA IV - CONCORDANCE

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
SIM	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO	SIM

Análise: Todos, com exceção de **RANGE** e **DICTGEN** fazem concordância.

KWIC

Não chega a dois segundos até KWIC mudar do diretório padrão da abertura de seus documentos de trabalho para a raiz de “D:\”.

O caminho até efetuar a concordância é o seguinte.

Corpus setup

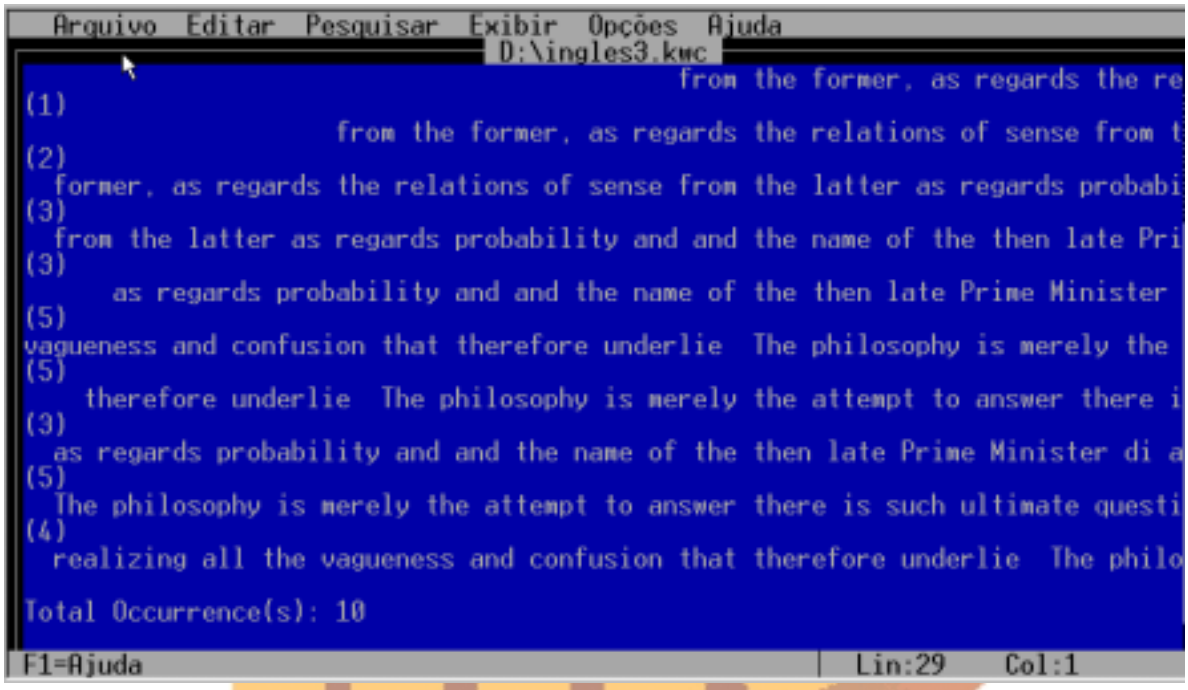


Fig. 3 Arquivo de Concordâncias de KWIC

Input files/Option (na mesma tela) – Option significa Corpus options e language.
Kwic Concordance (Menu Principal)
Escolha da Keyword.

Ele não distingue strings de substrings, assim, procurando por 'the' em um texto de língua inglesa, considerará em sua busca, then, there, therefore etc.

MONOCONC

O sistema de abertura do documento de trabalho é bastante rápido. Para abri-lo, o usuário escolhe a linguagem e depois o corpus, se um arquivo ou se um documento HTML, tudo na mesma tela. Feito isto, abre a janela de busca onde introduz a palavra a ser buscada, tendo para isto várias opções.

Escolhida a palavra, passa a processar a concordância. Só considerará a palavra buscada quando esta for uma string, assim como os demais programas, considerará os casos de duplicidade da palavra buscada na mesma frase, procedendo com as combinações existentes na mesma frase. Os dados são apresentados em duas telas, na tela superior o texto completo (em janela), na de baixo, as Concordâncias. Ver Fig. 4.

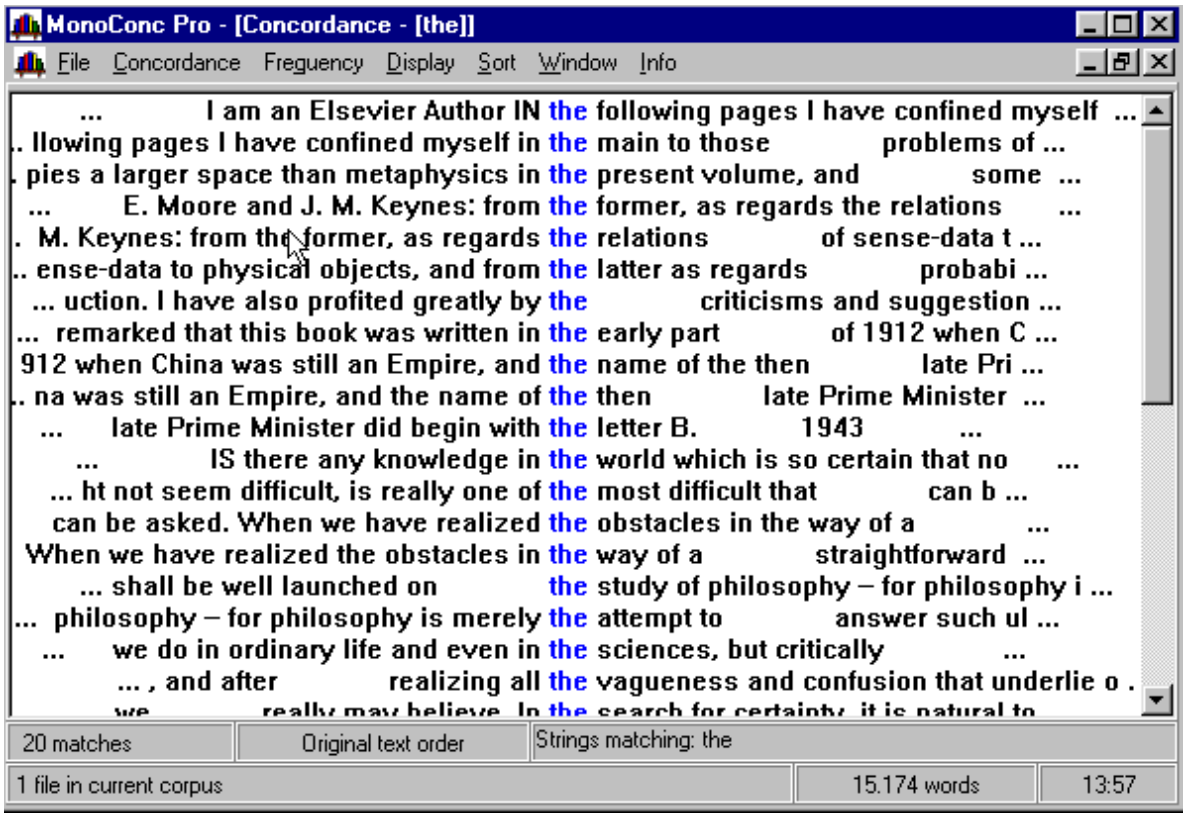


Fig. 4 Concordâncias de MONOCONC

CORPUS WIZARD – É preciso abrir o Kwic Search, fornecer a palavra-chave a ser buscada e o nome do arquivo. Os resultados são apresentados na tela, conforme Fig. 5, sem necessidade de ir para o prompt ver o resultado em um arquivo, como em KWIC, por exemplo.

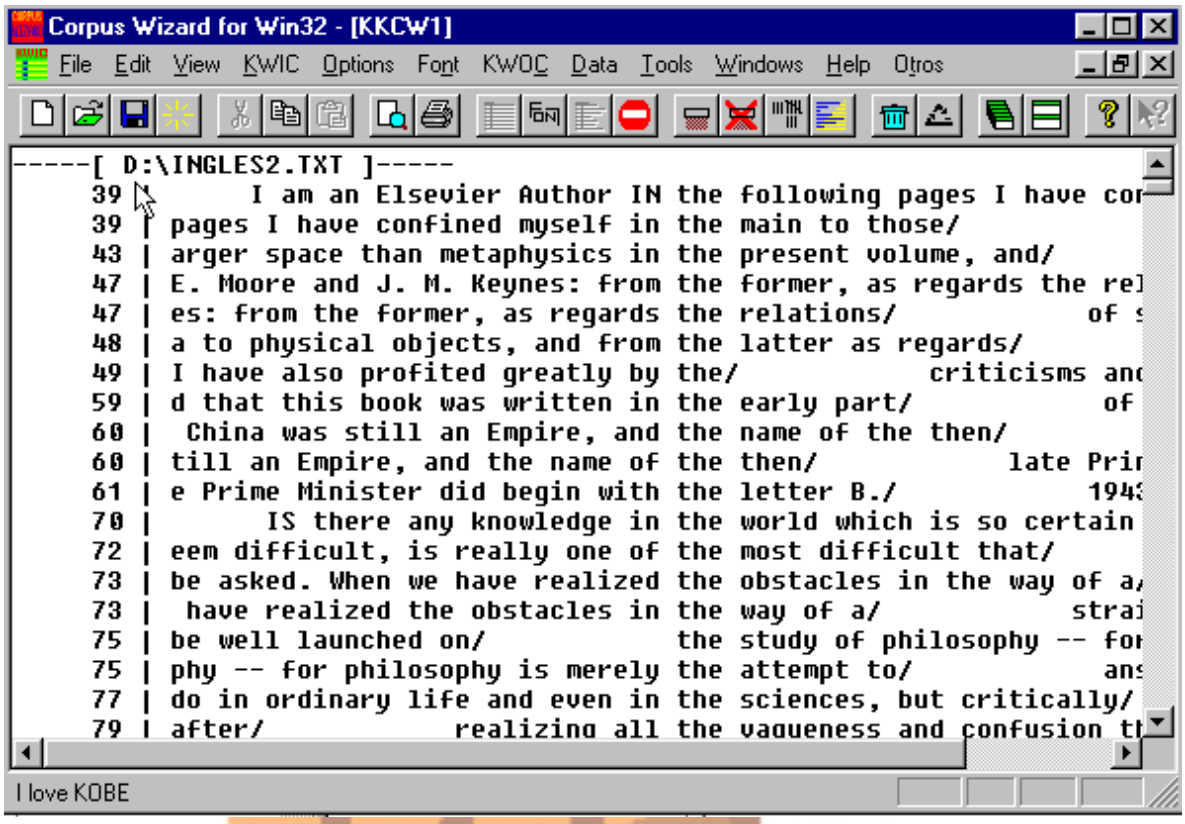


Fig. 5 Tela de Concordâncias do Corpus Wizard

WORDSMITH TOOLS

Ele precisa de nove segundos para mudar do diretório padrão C:\ para D:\ carregando os arquivos deste.
Seguem-se os passos:

Abre WordSmithTools Controller
Escolhe o texto ou os textos
Abre outra tela, a tela do Concord.
O usuário vai para Open e Start.
Espoecifica a palavra a buscar
Vai para a tela de Concordance Settings
Volta à tela anterior
Autoriza o programa a fazer as concordâncias
Abre outra tela onde são apresentados os resultados da concordância.

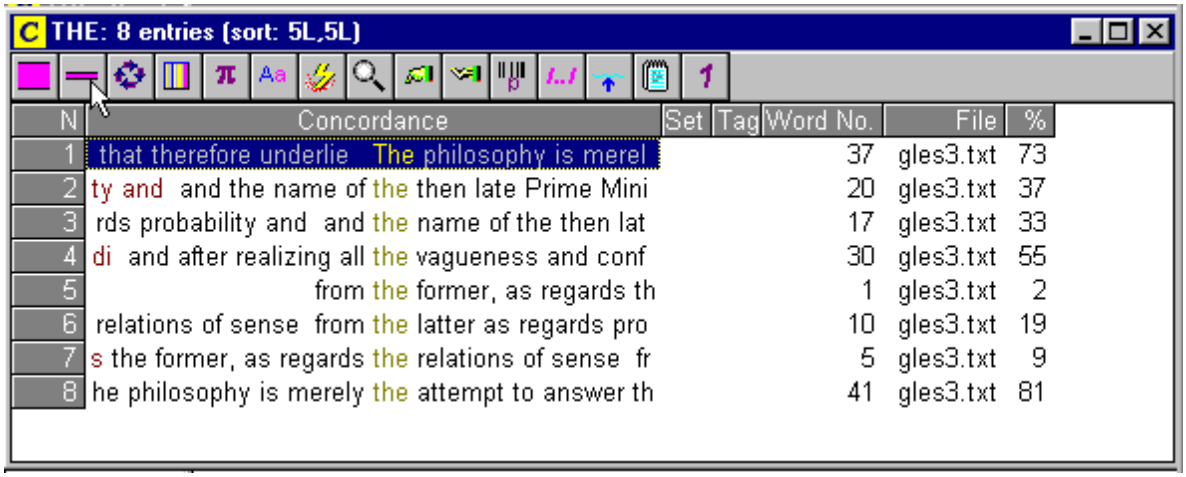


Fig. 6 Tela de concordâncias do WordSmith Tools

O **WORDSMITH TOOLS** apresenta os dados formatados da seguinte forma: Concordância, número da palavra no texto, nome do arquivo e (%) representando a proporção de Type/Token na linha.

O Concordanceador informa no máximo, as cinco palavras anteriores e as cinco posteriores à palavra buscada. As palavras das extremidades geralmente são excluídas.

Se na linha existir uma repetição da palavra buscada, ele registrará todas as ocorrências, não importando quantas sejam. No exemplo acima, “the name of the then”, “the” é mostrado em duas linhas diferentes.

CONCORDANCE

O primeiro comando é escolher entre Concordância total ou rápida. Na segunda tela, o usuário terá quatro opções: adiciona arquivos, remove arquivo, limpa lista de arquivos, salva lista de arquivos e carrega lista de arquivos. Ao escolher, por exemplo, adiciona arquivos, o usuário precisará de cerca de um a dois segundos para mudar para o disco (D:\), por exemplo, onde estão os arquivos de trabalho. Logo a seguir pressionará o botão de make full concordance ou make fast concordance. **CONCORDANCE**, além de fazer as concordâncias informará o tempo que precisou para organizar e criar as concordâncias, também informará se está carregando ou se já foi concluído o processo.

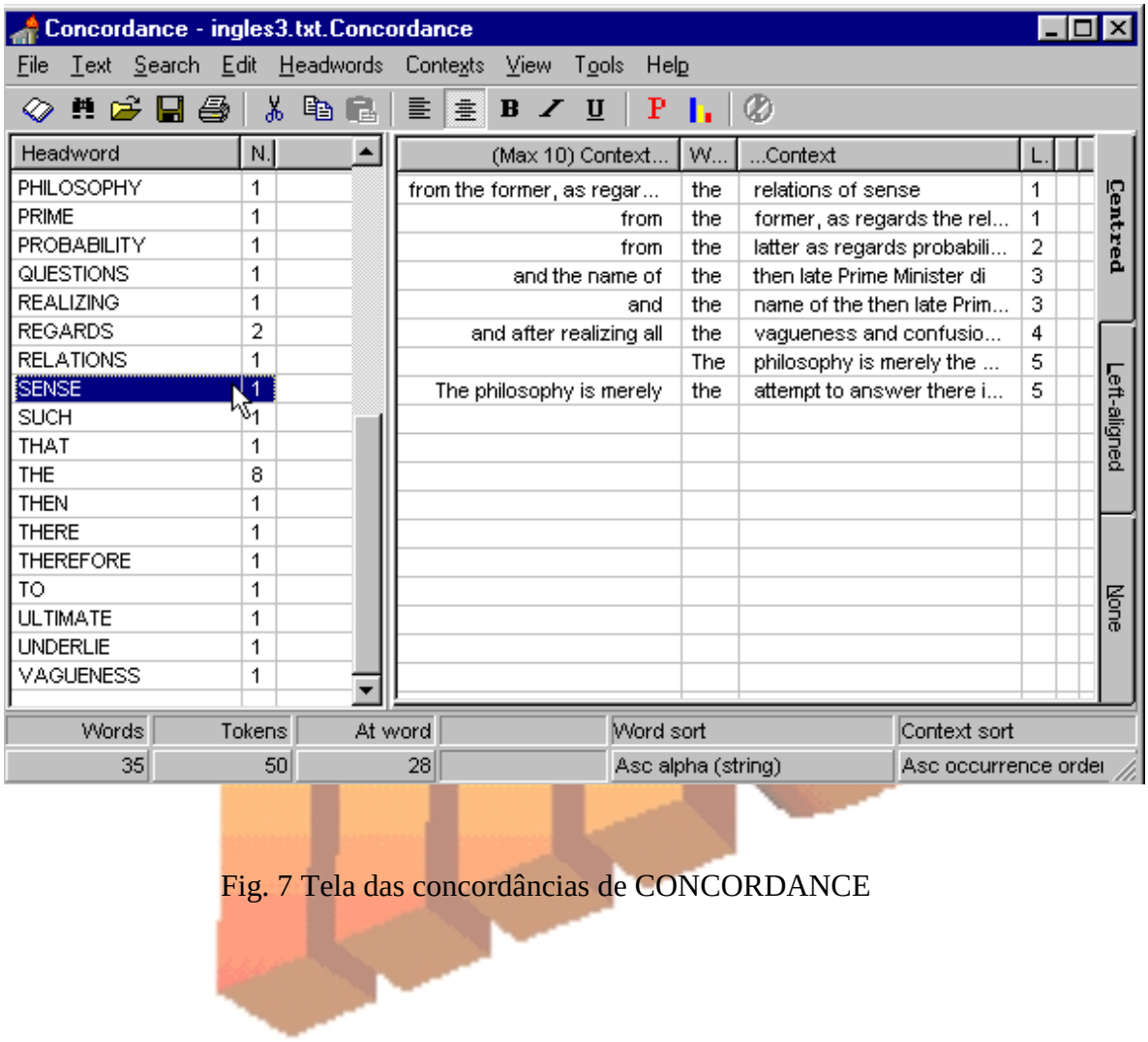


Fig. 7 Tela das concordâncias de CONCORDANCE

Conforme Fig. 7 acima, os contextos podem ser apresentados em dois formatos, com a palavra procurada no centro ou alinhada à esquerda. Ele também computa as ocorrências repetidas da palavra procurada, mesmo que na mesma frase. A linha mostrada, entretanto, varia de acordo com a posição da palavra. Informa, também, o número da linha onde aparece a palavra procurada. Diferentemente do WORDSMITH TOOLS**, a palavra a ser procurada fica na mesma tela de onde aparece a concordância, daí a economia de tempo. As palavras-guia estão alinhadas em ordem alfabética, juntamente com sua freqüência.**

TACT

Através do executável COLLGEN, produz uma lista de todas *fixed phrases*, onde *phrase* significa Seqüências repetidas de duas ou mais palavras

encontradas em .TDB file. Os arquivos .TDB são criados a partir de arquivos tipo texto abertos pelo usuário.

Conforme Fig. 8, na primeira coluna, o número de repetições; na Segunda coluna as *fixed phrases*.

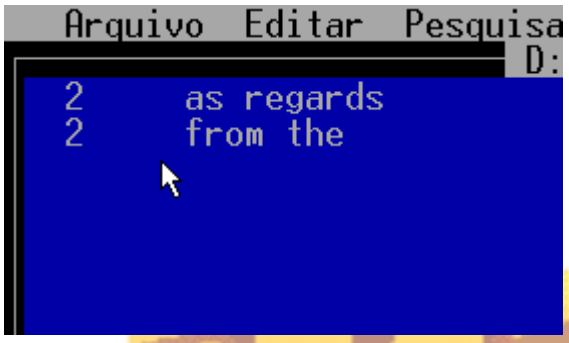


Fig. 8 Fixed Phrases em TACT

TATOE

Para acessar ao concordanceador é preciso primeiro:

Abrir uma base de textos já criada, à qual podem ser adionados outros textos. Na mesma tela, o usuário opta pelo modo de Concordância. Concordância total será apresentada conforme Fig. 9.

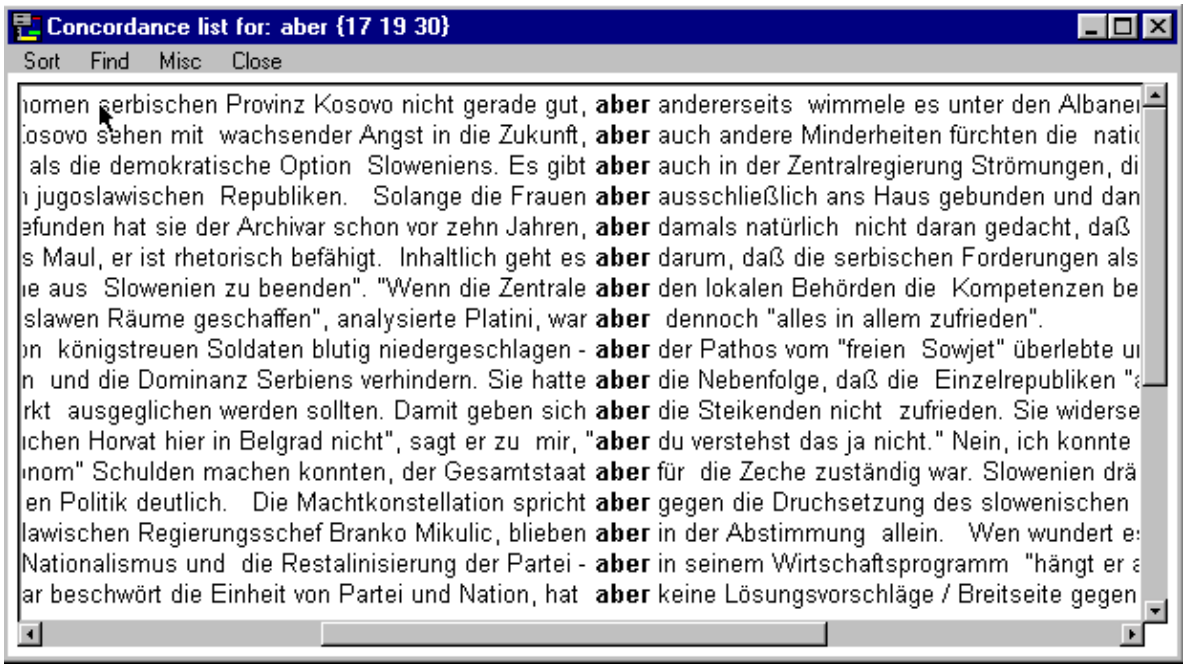


Fig. 9 Lista de Concordâncias de TATOE

2.1.4 Tempo do Concordanciador

TABELA V - CONCORDANCE PROCESSING TIME

Sentences	KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
			tempo em segundos						
10000	260		3 NÃO	LIMITED	NO	170	NÃO	NÃO	NO
17000	480		4 NÃO	LIMITED	10	190	NÃO	NÃO	NÃO
25000	705		4 NÃO	LIMITED	10	540	NÃO	NÃO	NÃO
49000	NO		4 NÃO	LIMITED	NO	1200	NÃO	NÃO	NÃO

Análise:

CORPUS WIZARD apresenta o melhor tempo de processamento, contudo, na verdade, não são processadas todas as concordâncias, como deveria ser. No caso de 17000 sentenças, por exemplo, **CORPUS WIZARD** indica que há cerca de 3900 concordâncias para 'de', apenas 'de' delimited with blanks. Na verdade, há mais de 8000 concordâncias para 'de', o que desautoriza o resultado.

CONCORDANCE Precisou de 119 segundos para fazer as concordâncias totais de 17000 sentenças e 190 segundos para carregá-las para a tela. **CONCORDANCE** faz apenas concordância total (full concordance).

KWIC **possui o pior desempenho. Nos testes efetuados, 49.000 sentenças pareceram estar além do limite de processamento do software.**

WORDSMITH TOOLS apresenta um excelente desempenho, contudo apresenta apenas 24 resultados, fruto de ser um produto demo, contudo o arquivo com 49.000 sentenças não foi aberto devido aos limites operacionais do sistema, pois para WORDSMITH TOOLS, a palavra procurada 'de' tinha mais de 17000 sentenças, entre as 47.000 do arquivo.

2.1.5 Determinação de Colocações

TABELA VI - COLOCAÇÕES

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
SIM	NÃO	NÃO	SIM	SIM	SIM	NÃO	NÃO	SIM

Análise:

Colocações são aquelas palavras que aparecem próximo ou perto de. ??? (Definir melhor o que são colocações) Colocações que acontecem com alta freqüência são indicadores poderosos de um padrão de significação em um texto. Eles também são úteis para estudar fraseologia e idioma.

KWIC – Como pode ser visto pela Fig.10, registra as colocações até cinco posições (palavras) à esquerda e outro tanto à direita. Na primeira coluna, as palavras; na segunda coluna, a freqüência absoluta (nas cinco posições); nas demais colunas a freqüência pela posição da palavra no texto.

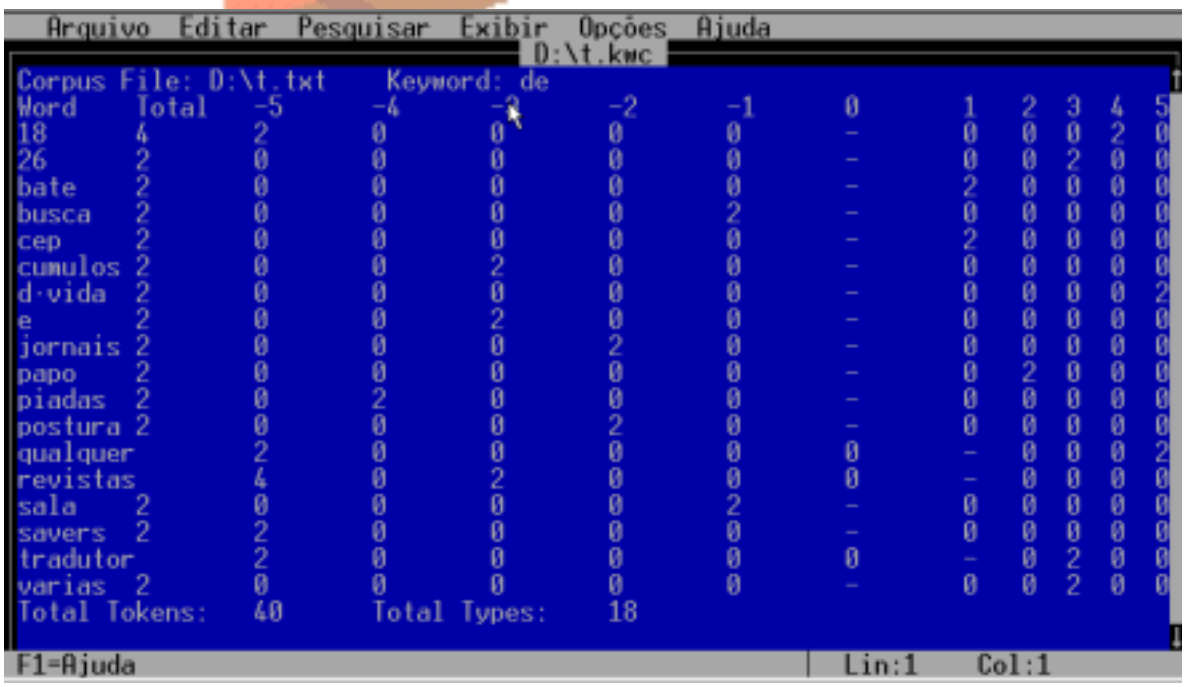


Fig. 10 Arquivo de Colocações do KWIC

MONOCONC

Colocação avançada contém dois modos de extrair colocações baseado na palavra de procura. No primeiro método o programa calcula a freqüência de três ou quatro palavras por colocação baseado nas posições das palavras (3L, 2L, 1L, Termo de Procura, 1R, 2R, 3R) especificadas pelo usuário. O segundo método é a colocação personalizada que permite ao usuário entrar com um maior número de posições. Por exemplo, entrando em 3L-3R produz uma lista de freqüência de colocações de 7 Colocações-palavra. Ver Fig. 11.

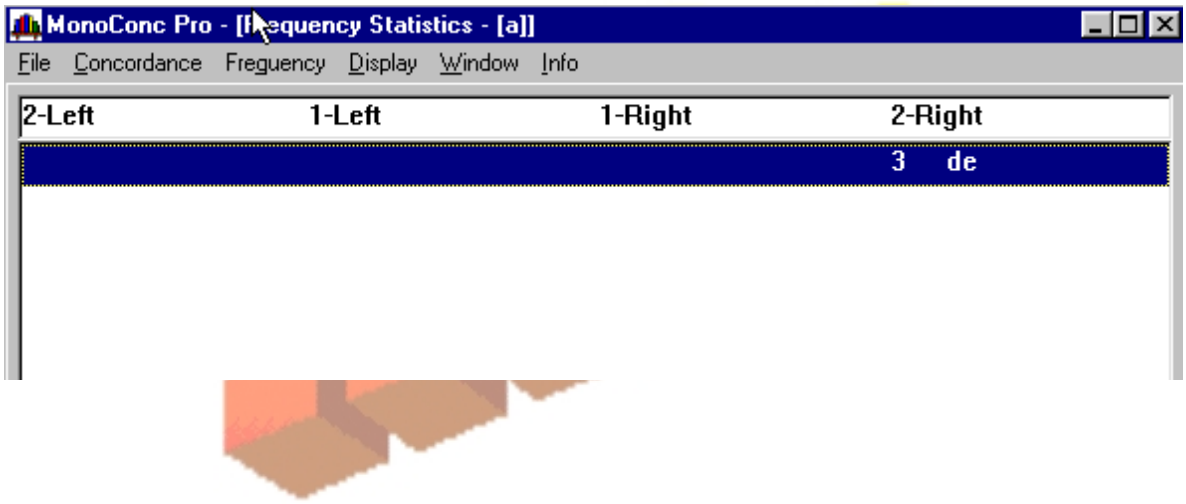
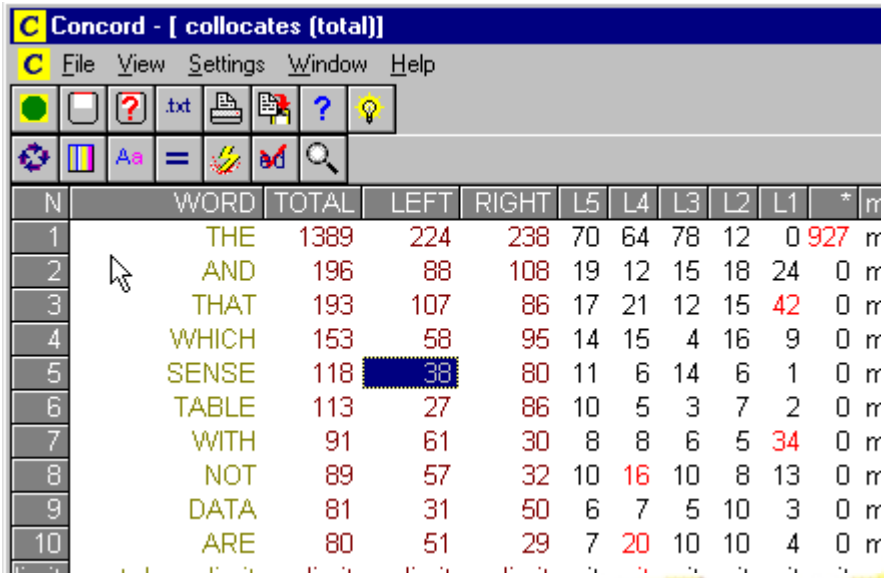


Fig. 11 Apresentação das colocações em MONOCONC

WORDSMITH TOOLS

A Fig. 12 mostra na primeira coluna, as palavras, inclusive a palavra-guia “the”. Na segunda coluna, a freqüência total. Na terceira e quarta coluna, a freqüência absoluta, ora do lado esquerdo da palavra-guia “**THE”**, ora do lado direito. Da quinta até a nona coluna, a freqüência absoluta nas primeiras cinco posições à esquerda da palavra-guia. Na última coluna, mostra a freqüência absoluta da palavra “**THE**”, na mesma linha da palavra-guia **“THE”**, ou seja, em casos em que havia mais de uma ocorrência de “**THE”** na mesma linha.



N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*
1	THE	1389	224	238	70	64	78	12	0	927
2	AND	196	88	108	19	12	15	18	24	0
3	THAT	193	107	86	17	21	12	15	42	0
4	WHICH	153	58	95	14	15	4	16	9	0
5	SENSE	118	38	80	11	6	14	6	1	0
6	TABLE	113	27	86	10	5	3	7	2	0
7	WITH	91	61	30	8	8	6	5	34	0
8	NOT	89	57	32	10	16	10	8	13	0
9	DATA	81	31	50	6	7	5	10	3	0
10	ARE	80	51	29	7	20	10	10	4	0

Fig. 12 Colocações em WordSmith Tools

CONCORDANCE

CONCORDANCE deixa o usuário ver listas de colocações para qualquer palavra-guia nas listas de palavras. Mostra e dispõe colocações da palavra-guia com todas as palavras que ocorrem até quatro palavras antes e quatro depois do palavra-guia.

Fig. 12 Colocações em CONCORDANCE

2.1.6 Critério: Arquivos de Entrada[2]

TABELA VII - Arquivos de Entrada

	KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGE	TACT
Text	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM
HTML	SIM	NÃO	NÃO	SIM	SIM	NÃO	SIM	NÃO	NÃO
DOC	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO
XML	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	SIM	NÃO	NÃO
PDF	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO
SGML	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO

Análise:

Embora sete, entre os nove softwares analisados, rodem em plataforma Windows, apenas quatro trabalham com outro tipo de texto que não texto (.txt). **KWIC**, **MC**, **TATOE** e **WORDSMITH TOOLS** importam arquivos no formato XML E HTML. **WORDSMITH TOOLS** suporta os três tipos de mark up texts (Html, Sgml e XML). Embora **WORDSMITH TOOLS** trabalhe com mark up texts, sua interface com o usuário não é tão fácil quanto **TATOE**, cuja interface é na principal tela, onde aparecem os formatos disponíveis para importação de textos/documentos, como pode ser visto na Fig. 13 abaixo.

[2] Na Tabela VII, Text aqui significa ASCII text, ANSI text, Text Only e DOS Text. Html, Sgml e XML são mark up texts, os quais tem sua informação entre tags..

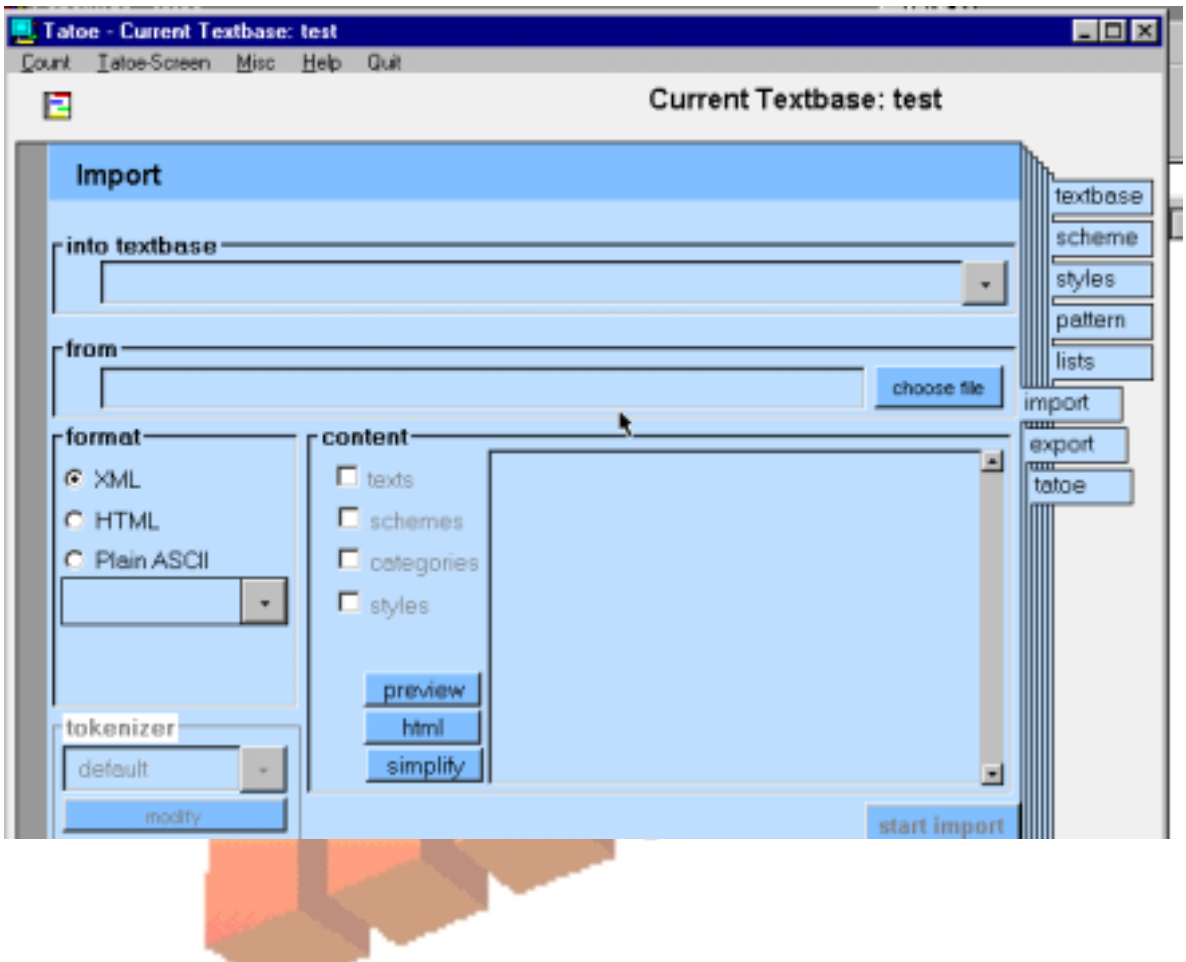


Fig. 13 Tela de interfaces do TATOE

Vale ressaltar, porém, que **KWIC** trabalha com vários formatos de corpus, como por exemplo:

Texto ASCII. Formatos COCOA, SGML (incluindo BNC), fixed block length (incluindo Brown,LOB,FROWN,FLOB,WC), Corpus Helsinki + Cocoa + formatos originais de CEECS e Escocês antigo. Parsed Corpus Penn-Helsinki do Médio Inglês. Formatos de Corpus do ICE, BNC, ACE e ICAMET.

2.1.7 Listas de Palavras

TABELA VIII - WORDLISTS

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM

Análise: KWIC e **RANGE** não apresentam na tela os resultados, apenas em arquivo. **KWIC** apresenta assim seu CORPUS FREQUENCY LIST em um arquivo, cuja extensão é (.KWC).

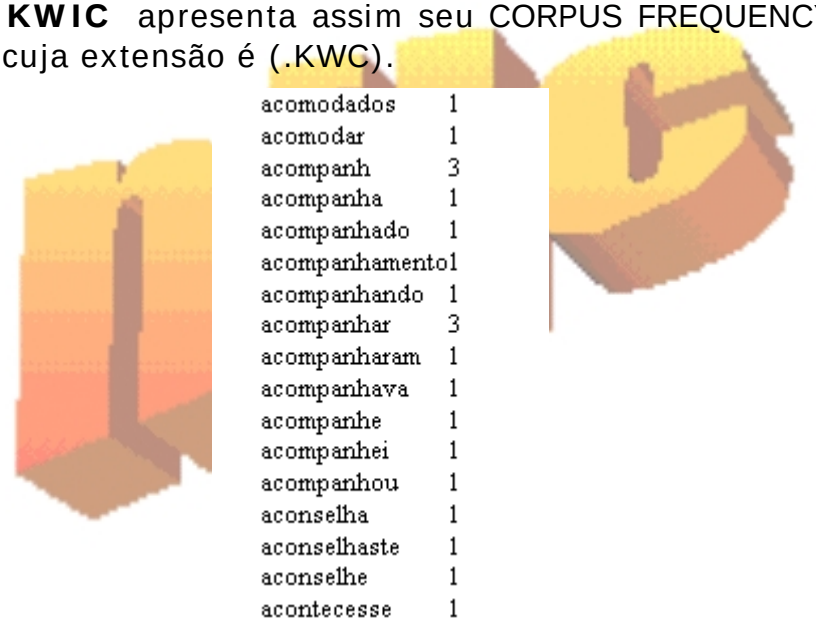


Fig. 14 Lista de palavras (KWIC)

As **WORDLISTS** de **RANGE** são formadas a partir da abertura de, no mínimo, um arquivo-texto, até no máximo 32. Os objetivos destas WORDLISTS são diferentes dos restantes softwares aqui analisados. **RANGE** busca encontrar o vocabulário comum a estes textos e também compará-lo com as três listas de vocabulário básico que possui para saber qual o nível do texto analisado. Os arquivos .**dat** quardam as listas do inglês básico, o que os autores denominam **RANGE**.

RANGE possui três arquivos .dat. O primeiro (BASEWRD1.DAT) inclui as 1000 palavras mais freqüentes do Inglês. O segundo (BASEWRD2.DAT) inclui as seguintes 1000 palavras mais freqüentes do Inglês, e o terceiro (BASEWRD3.DAT) inclui mais 1000 palavras não presentes entre as 2000 primeiras, mas que são frequentes nos textos do ensimo médio e

superior. Todas estas listas básicas incluem a forma básica e suas derivadas.

As primeiras 1000 palavras consistem de cerca de 4000 formas diferentes. As fontes de onde foram extraídas são: A General Service List of English Words by Michael West (Longman, London 1953) para as primeiras 2000 palavras e The Academic Word List by Coxhead (1998, 2000) contendo 570 famílias de palavras. As primeiras 1000 palavras de A General Service List of English Words são aquelas com uma freqüência maior do que 332 ocorrências por 5 milhões de palavras, mais meses, dias e semanas, números, títulos e frequentes saudações. As listas incluem ortografia britânica e americana.

Os resultados não são vistos na tela, mas em um arquivo pré-indicado pelo usuário, onde as informações são armazenadas na ordem
De (f1..fn) e distribuídas da seguinte forma: Lista das palavras (TYPE), seguido do (RANGE) número de arquivos abertos onde aparece a palavra (TYPE), a freqüência total e em ordem (F1,F2..Fn), a freqüência por arquivo, conforme Fig. 15.

TYPE	RANGE	FREQ	F1	F2	F3	F4
A	3	10642	0	222	1257	9163
Y	1	10429	0	0	0	10429
NO	3	4101	0	52	174	3875
S	3	2625	0	23	337	2265
N	3	2606	0	1	65	2540
O	3	2319	0	5	964	1350
M	3	2126	0	3	187	1936
DON	1	1817	0	0	0	1817
ME	3	1668	0	12	176	1480
AS	3	1590	0	190	220	1180
L	3	1052	0	3	78	971
E	3	1047	0	15	968	64
D	2	980	0	0	298	682
THE	1	927	0	927	0	0
OR	3	818	0	109	1	708
IS	3	770	0	455	5	310
AN	3	700	0	53	3	644
OF	2	543	0	542	1	0
DO	3	517	0	26	239	252
WE	1	421	0	421	0	0
TO	2	416	0	413	3	0

Fig. 15 Lista de palavras com freqüência comparada (RANGE)

Para cada texto informa, o número total de palavras, lemas e "famílias", que são as palavras comuns aos textos.

Tanto MONOCONC **quanto** WORDSMITH TOOLS **apresentam um CORPUS FREQUENCY LIST com três dados, a palavra, a freqüência absoluta e a freqüência relativa, conforme figuras 16 E 17 abaixo, respectivamente.**

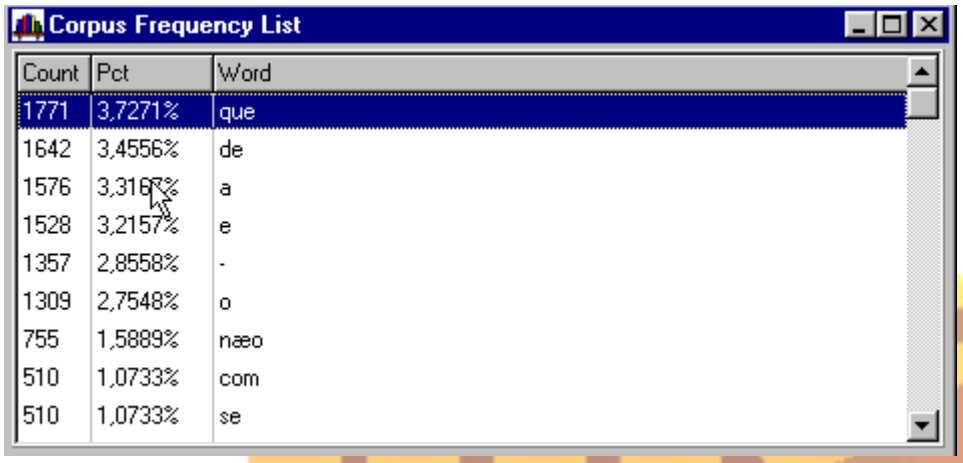


Fig 16 Lista de palavras (MONOCONC)

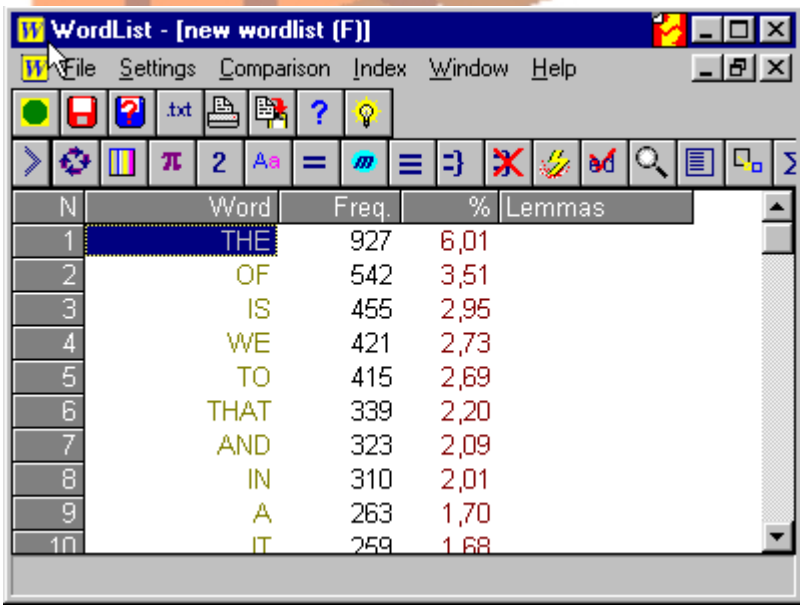


Fig. 17 Lista de palavras (WORDSMITH TOOLS)

CONCORDANCER apresenta sua WORDLIST de diversos modos, um deles é organizar por ordem alfabética ascendente.

Os dados são apresentados na tela, ao mesmo tempo em que, no lado direito da tela, visualiza-se o contexto em que a palavra aparece, conforme figura 18 abaixo.

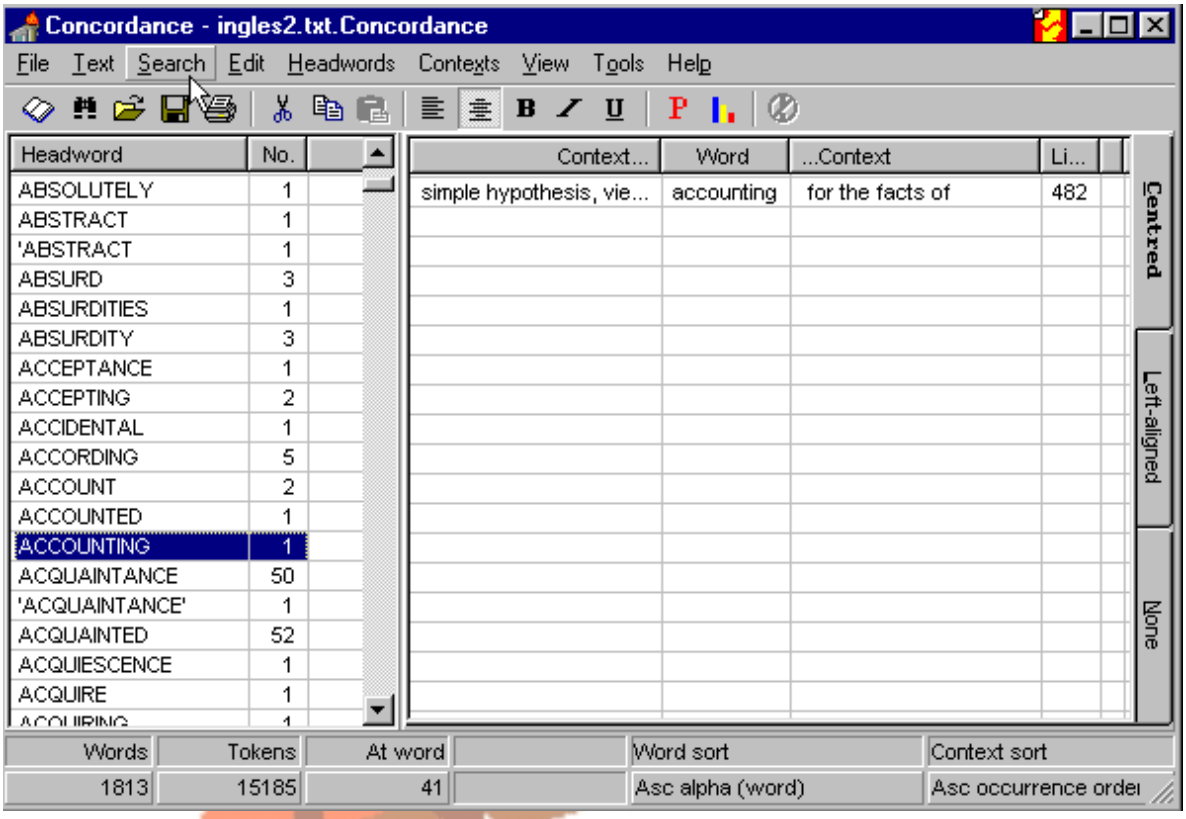


Fig. 18 Lista de palavras (CONCORDANCER)

Entre as três ferramentas para ambiente MSDOS, TACT **apresenta um executável para cada função. A lista de palavras somente é gerada por TACTFREQ após ter sido criado o banco de dados, através do executável MAKEBASE, conforme Fig. 19.**

Fig. 19 Lista de Palavras (TACT)

DICTGEN, assim como **RANGE** tem como objetivo a criação de dicionários ou listas de palavras que possam servir de base para a criação de dicionários, workbooks ou para outras finalidades. **DICTGEN** trabalha com dois arquivos ao mesmo tempo, da análise de ambos, sempre listas de palavras, não frases, ele extrai um arquivo que poderá ser uma das três opções:

1) Output a dictionary made up from the words in file one and file two.
2) Output the words in file one, but not in file two.
3) Output the words in file two, but not in file one.

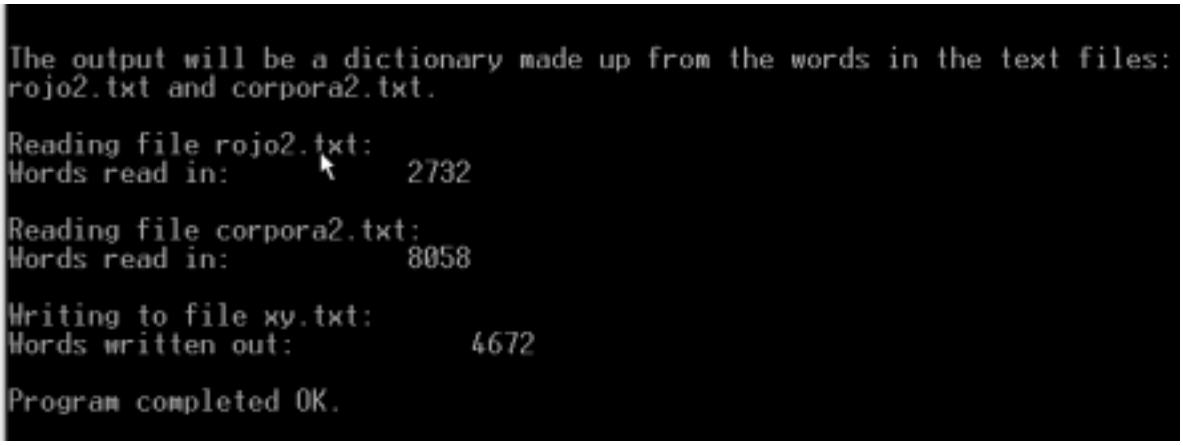


Fig. 20 Lendo e copiando arquivos (DICTGEN)

TATOE

Para **TATOE** não há uma simples lista de freqüência, mas um WORD/CATEGORY INDEX, no qual cada palavra aparece, ao mesmo tempo, no index e também no texto.

Na colunas da esquerda, em cima, aparece o nome do arquivo, embaixo, o WORD/CATEGORY INDEX; na coluna (mais larga) da direita, a palavra em seu contexto. As concordâncias aparecem no formato tradicional, em coluna, veja Fig. 21 abaixo ou como na figura 9.

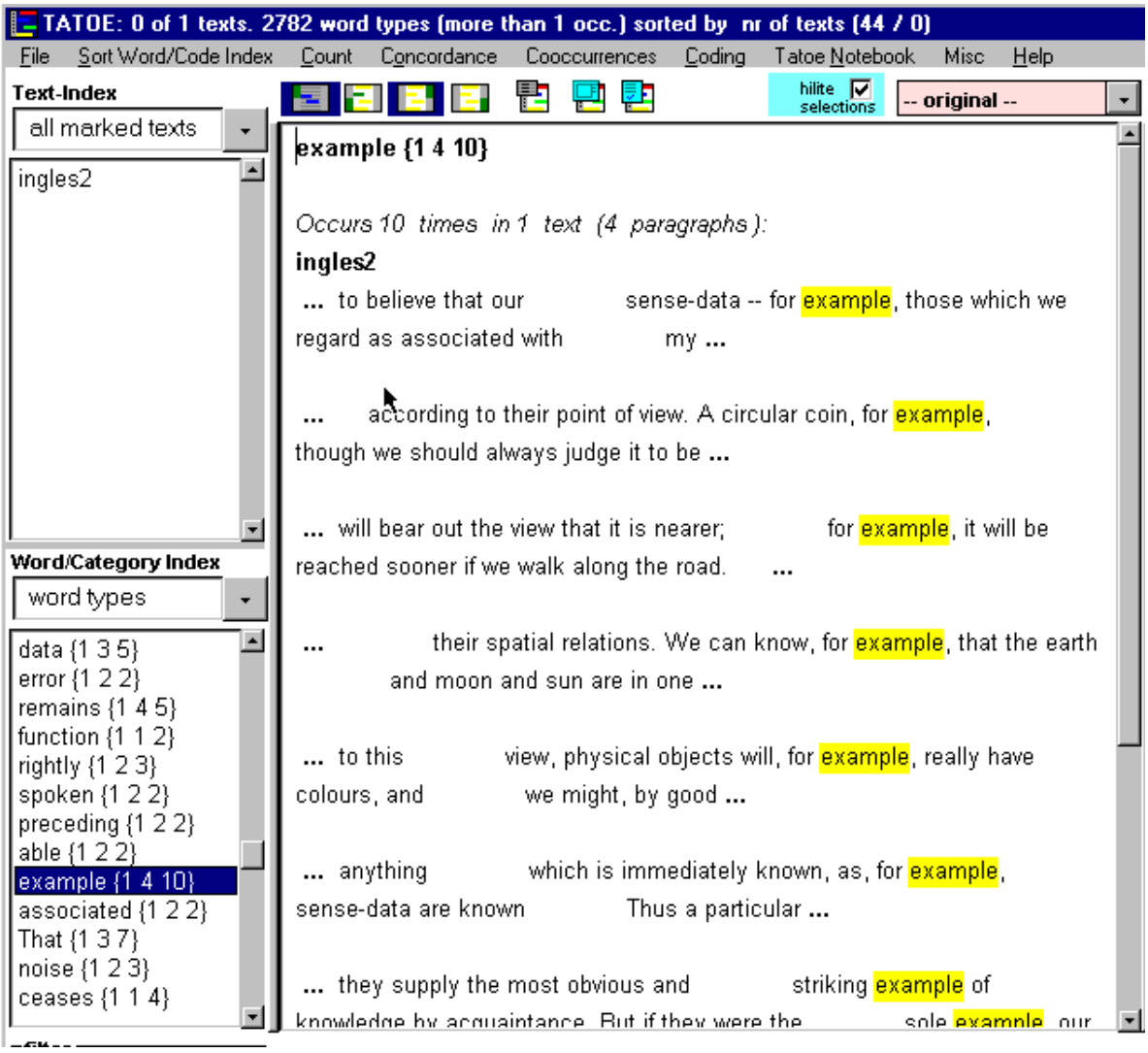


Fig. 21 Lista de palavras e concordâncias (TATOE)

2.1.8 Comparação de Listas de Palavras

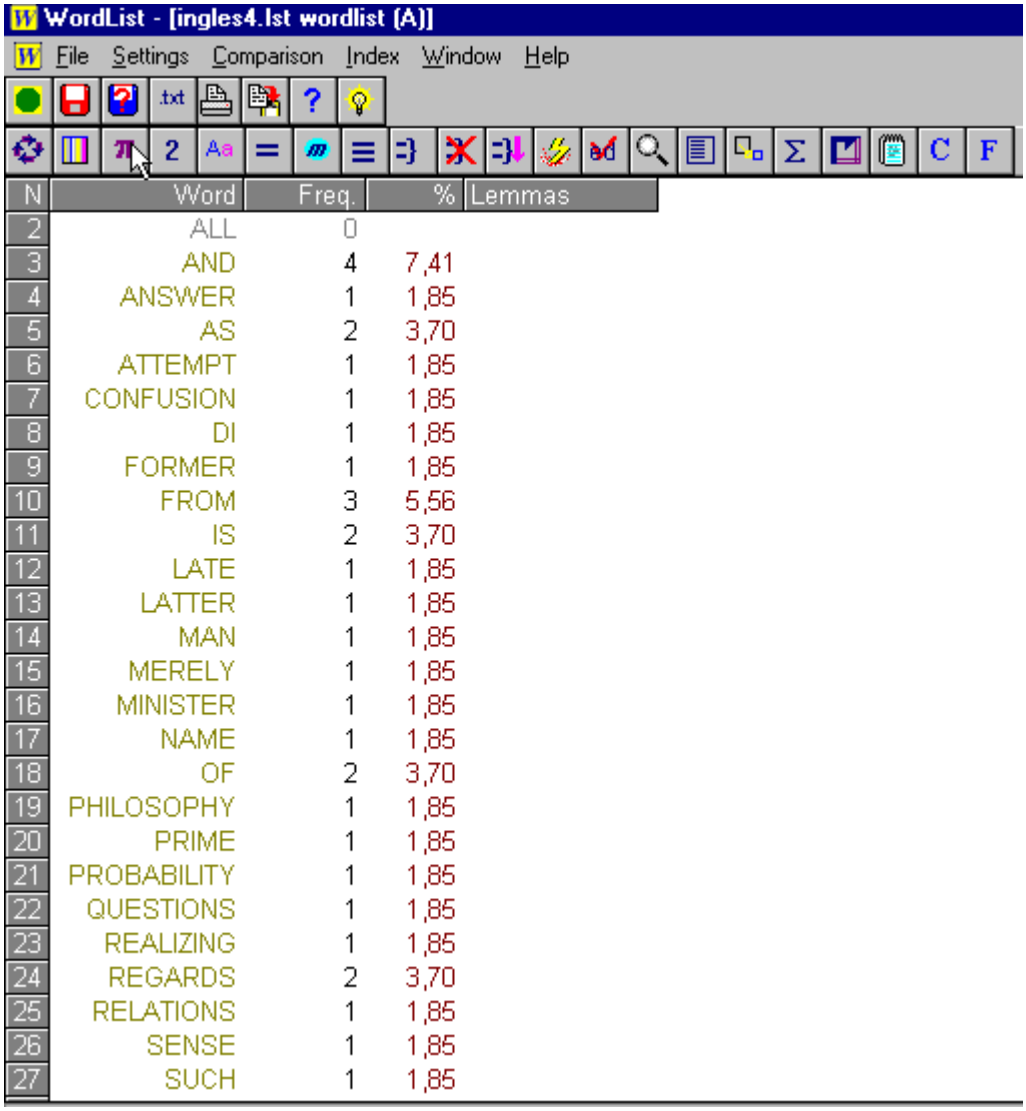
TABELA IX - COMPARAÇÃO DE LISTAS DE PALAVRAS

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
NÃO	NÃO	SIM	NÃO	SIM	NÃO	NÃO	SIM	SIM

Análise:

Um dos programas de **TACT** (FCOMPARE) compara as linhas de dois arquivos tipo ASCII criando um arquivo que contém as linhas em comum entre estes dois arquivos.

WORDSMITH TOOLS abre até 50 listas de palavras ao mesmo tempo. **RANGE**, 32. **WORDSMITH TOOLS** compara apenas duas listas de palavras de cada vez, ao passo que **RANGE** compara até 32.



N	Word	Freq.	%	Lemmas
2	ALL	0		
3	AND	4	7,41	
4	ANSWER	1	1,85	
5	AS	2	3,70	
6	ATTEMPT	1	1,85	
7	CONFUSION	1	1,85	
8	DI	1	1,85	
9	FORMER	1	1,85	
10	FROM	3	5,56	
11	IS	2	3,70	
12	LATE	1	1,85	
13	LATTER	1	1,85	
14	MAN	1	1,85	
15	MERELY	1	1,85	
16	MINISTER	1	1,85	
17	NAME	1	1,85	
18	OF	2	3,70	
19	PHILOSOPHY	1	1,85	
20	PRIME	1	1,85	
21	PROBABILITY	1	1,85	
22	QUESTIONS	1	1,85	
23	REALIZING	1	1,85	
24	REGARDS	2	3,70	
25	RELATIONS	1	1,85	
26	SENSE	1	1,85	
27	SUCH	1	1,85	

Fig. 22 Lista de palavras (WORDSMITH TOOLS)

RANGE não se limita apenas a comparar estatisticamente os dados dos textos. Ele compara o vocabulário dos textos abertos com suas listas de vocabulário.

Os arquivos .dat são onde se encontram the basic word list of English language, o que os autores denominam RANGE. Tokens são as palavras dos textos com suas desinências e flexões. Types seriam as palavras em

sua forma canônica ou lema. Familias ou Families, como os autores explicam, são as derivadas de uma palavra-guia, por exemplo, a palavra-guia AID tem como membros de sua família: AIDED, AIDING, AIDS e UNAIDED.

A lista abaixo é um exemplo de como as listas são comparadas em RANGE**.**

```
 Processing file: C:\WINNT\Profiles\nationp\Personal\range\widi.txt

  Number of lines: 6
  Number of words: 78

 Processing file: C:\WINNT\Profiles\nationp\Personal\range\joni.txt


  Number of lines: 6
  Number of words: 83

  Reading: basewrd1.dat
  Reading: basewrd2.dat
  Reading: basewrd3.dat
```

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	143/88.8	81/87.1	74
two	6/ 3.7	5/ 5.4	5
three	0/ 0.0	0/ 0.0	0
not in the lists	12/ 7.5	7/ 7.5	?????
Total	161	93	79

```
 Number of BASEWRD1.DAT types: 3126  Number of BASEWRD1.DAT families: 999
 Number of BASEWRD2.DAT types: 2721  Number of BASEWRD2.DAT families: 986
 Number of BASEWRD3.DAT types: 2540  Number of BASEWRD3.DAT families: 570

Table of Ranges: Types

   160  Words appear in  1 input files
    26  Words appear in  2 input files

Table of Ranges: Families

    71  Words appear in  1 input files
    15  Words appear in  2 input files


Types Found In Base List One
```

TYPE	RANGE	FREQ	F1	F2
AND	2	7	4	3
FOR	2	2	1	1
IT	2	3	2	1
OF	2	3	2	1
PLEASE	2	2	1	1
SAID	2	2	1	1
SCHOOL	2	2	1	1
SOME	2	2	1	1
STUDENTS	2	3	1	2
THE	2	14	10	4
TO	2	7	2	5
WAS	2	4	3	1
A	1	1	0	1
ALL	1	1	0	1
ALREADY	1	1	1	0
ANSWER	1	1	1	0

ARE	1	2	0	2
BECAME	1	1	1	0
BEGAN	1	1	1	0
BOOKS	1	1	0	1
BUT	1	1	1	0
CLASS	1	1	1	0
COST	1	1	0	1
DAY	1	1	1	0
DAYS	1	1	0	1
END	1	2	2	0
ENGLISH	1	1	1	0
EVERYBODY	1	1	1	0
FACE	1	2	2	0
FORGOTTEN	1	1	1	0
GAVE	1	1	1	0
GOING	1	2	0	2
HAD	1	1	1	0
HAPPY	1	1	0	1
HE	1	3	0	3
HER	1	3	0	3
HIS	1	1	1	0
HOT	1	2	2	0
IN	1	2	0	2
IS	1	1	1	0
LAUGH	1	1	1	0
LAUGHED	1	1	1	0
LEFT	1	1	1	0
LISTEN	1	1	1	0
ME	1	2	0	2
MISS	1	2	2	0
MONEY	1	2	0	2
NAMES	1	1	0	1
NEAR	1	2	0	2
NEXT	1	1	0	1
NOT	1	1	1	0
OTHERS	1	1	1	0
PAY	1	2	0	2
PLACES	1	1	0	1
RANG	1	1	1	0
RED	1	1	1	0
ROOM	1	1	1	0
S	1	2	2	0
SIT	1	1	0	1
SOON	1	1	0	1
SPOKE	1	1	0	1
STAY	1	1	0	1
STILL	1	1	1	0
STOOD	1	1	0	1
TALK	1	1	1	0
TELL	1	1	0	1
THAT	1	2	2	0
THEM	1	1	0	1
THEN	1	1	1	0
THREE	1	1	0	1
THURSDAY	1	1	0	1
TOO	1	2	0	2
UP	1	2	0	2
VISIT	1	2	0	2
WALKED	1	1	0	1
WANTED	1	1	0	1
WE	1	3	0	3
WENT	1	1	0	1
WILL	1	2	0	2
WITH	1	1	0	1
YOUR	1	1	0	1

Types Found In Base List Two

TYPE	RANGE	FREQ	F1	F2
BELL	1	1	1	0
CORRECT	1	1	1	0
LESSON	1	2	2	0
MISTAKE	1	1	1	0
PICKED	1	1	0	1

Types Found In Base List Three

TYPE RANGE FREQ F1 F2

LIST OF FAMILY GROUPS

BASE ONE FAMILIES	RANGE	TYFREQ	FAFREQ	F1	F2
AND	2	7	7	4	3
BE	2	0	9	6	3
DAY	2	1	2	1	1
FOR	2	2	2	1	1
HE	2	3	4	1	3
IT	2	3	3	2	1
OF	2	3	3	2	1
PLEASE	2	2	2	1	1
SAY	2	0	2	1	1
SCHOOL	2	2	2	1	1
SOME	2	2	2	1	1
STUDENT	2	0	3	1	2
THE	2	14	14	10	4
TO	2	7	7	2	5
A	1	1	1	0	1
ALL	1	1	1	0	1
ALREADY	1	1	1	1	0
ANSWER	1	1	1	1	0
BECOME	1	0	1	1	0
BEGIN	1	0	1	1	0
BOOK	1	0	1	0	1
BUT	1	1	1	1	0
CLASS	1	1	1	1	0
COST	1	1	1	0	1
END	1	2	2	2	0
ENGLISH	1	1	1	1	0
EVERY	1	0	1	1	0
FACE	1	2	2	2	0
FORGET	1	0	1	1	0
GIVE	1	0	1	1	0
GO	1	0	3	0	3
HAPPY	1	1	1	0	1
HAVE	1	0	1	1	0
HOT	1	2	2	2	0
I	1	0	2	0	2
IN	1	2	2	0	2
LAUGH	1	1	2	2	0
LEFT	1	1	1	1	0
LISTEN	1	1	1	1	0
MISS	1	2	2	2	0
MONEY	1	2	2	0	2
NAME	1	0	1	0	1
NEAR	1	2	2	0	2
NEXT	1	1	1	0	1
NOT	1	1	1	1	0
OTHER	1	0	1	1	0
PAY	1	2	2	0	2
PLACE	1	0	1	0	1
RED	1	1	1	1	0
RING	1	0	1	1	0
ROOM	1	1	1	1	0
SHE	1	0	3	0	3
SIT	1	1	1	0	1
SOON	1	1	1	0	1
SPEAK	1	0	1	0	1
STAND	1	0	1	0	1
STAY	1	1	1	0	1
STILL	1	1	1	1	0
TALK	1	1	1	1	0
TELL	1	1	1	0	1
THEN	1	1	1	1	0
THEY	1	0	1	0	1
THIS	1	0	2	2	0
THREE	1	1	1	0	1
THURSDAY	1	1	1	0	1
TOO	1	2	2	0	2
UP	1	2	2	0	2
VISIT	1	2	2	0	2
WALK	1	0	1	0	1
WANT	1	0	1	0	1
WE	1	3	3	0	3

```
WILL                                    1       2       2       0       2
WITH                                    1       1       1       0       1
YOU                                     1       0       1       0       1

BASE TWO FAMILIES                     RANGE  TYFREQ  FAFREQ      F1      F2
BELL                                    1       1       1       1       0
CORRECT                                 1       1       1       1       0
LESSON                                  1       2       2       2       0
MISTAKE                                 1       1       1       1       0
PICK                                    1       0       1       0       1

BASE THREE FAMILIES                   RANGE  TYFREQ  FAFREQ      F1      F2


Types Not Found In Any List

TYPE                                  RANGE    FREQ      F1      F2
JONI                                      2       4       3       1
HOMEWORK                                  1       1       1       0
JAKARTA                                   1       2       0       2
PRIYADI                                   1       1       0       1
RUPIAHS                                   1       1       0       1
SUHARDI                                   1       2       2       0
WIDI                                      1       1       0       1


 time taken was : 1 Seconds
Number of Nodes Read:              0
Number of Cache Nodes Read:        31175
Number of Nodes Written:           0
Number of Cache Nodes Written:       2045
Number of nodes per second,          0
Number of words per second,        161
Number of unique words in tree,          93
Number of unique words per second,       93


...Finished
```

Estatística básica dos três arquivos abertos: número de linhas e número de palavras.

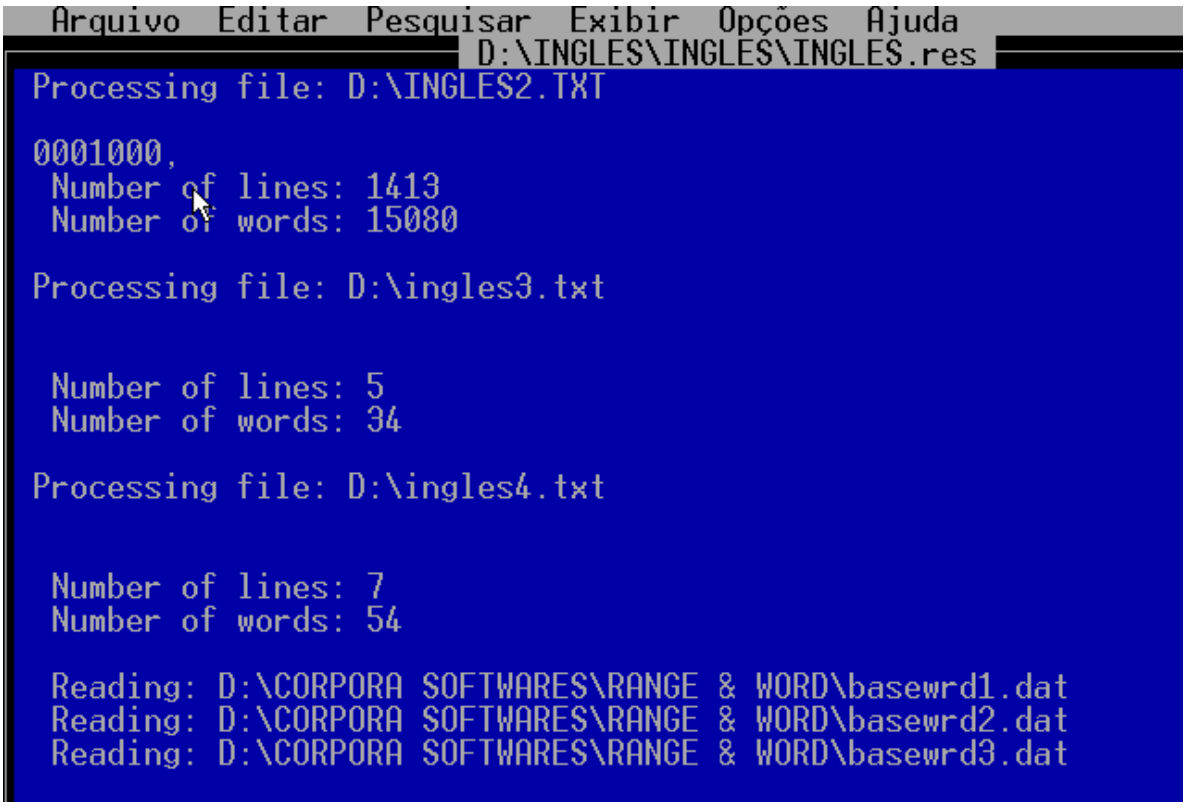


Fig. 23 Resumo das listas de palavras comparadas (RANGE)

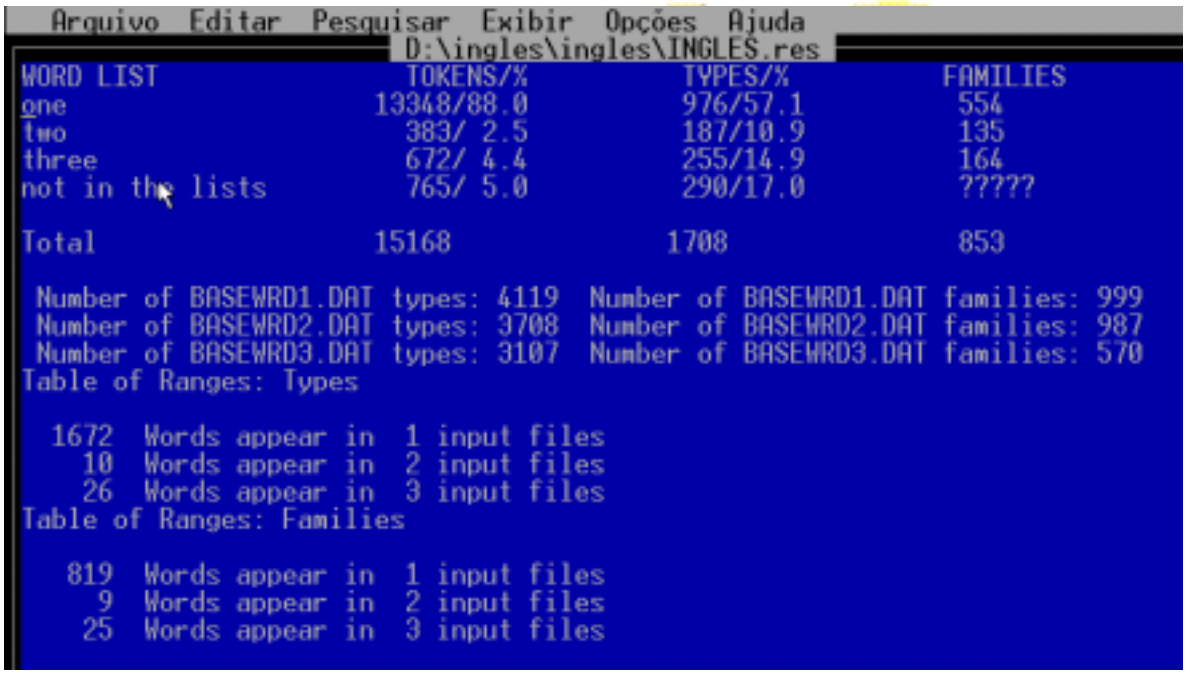


Fig. 24 Dados obtidos a partir das listas de palavras comparadas (RANGE)

Acima, na figura 24, o número de TOKENS, TYPES e FAMILIES. Para TOKENS e TYPES, freqüência absoluta e relativa.

No segundo parágrado, o número de TYPES e FAMILIES nos três arquivos de vocabulário-base.

No terceiro e quarto parágrafos, a tabela de **RANGES** de types e families relativo a cada arquivo aberto.

2.1.9 Estatísticas

TABELA X - ESTATÍSTICA

	KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGE	TACT
WC	SIM	SIM	SIM					SIM	
WF		SIM	SIM	SIM				SIM	
AWL		SIM							
CS		SIM		SIM					

WC = WORD COUNT
WF = WORD FREQUENCY
AWL = AVERAGE WORD LENGTH
CS = COLLOCATION STATISTICS

Análise:

Em termos de estatística, KWIC concordance tem pouco a oferecer, apenas a freqüência de palavras.

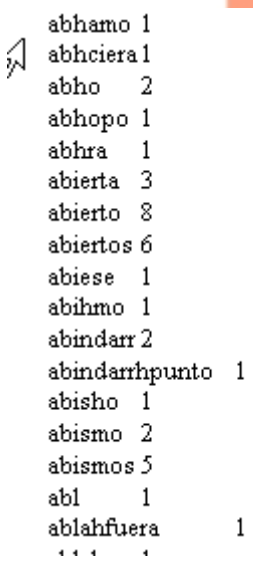


Fig. 25 Lista de palavras mais freqüência (KWIC)

CORPUS WIZARD – Oferece apenas tamanho médio da palavra, contagem de palavras e freqüência da palavra.

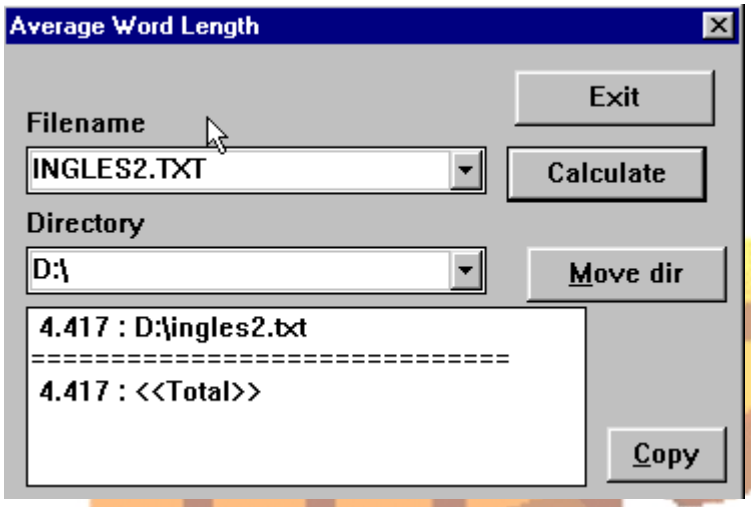


Fig. 26 Tamanho médio da palavra (CORPUS WIZARD)

Estatística Colocacional (I) – Ordem de freqüência – Cabe ao usuário escolher em qual posição da linha ele quer a ordem de freqüência das palavras. Ver Fig. 27.

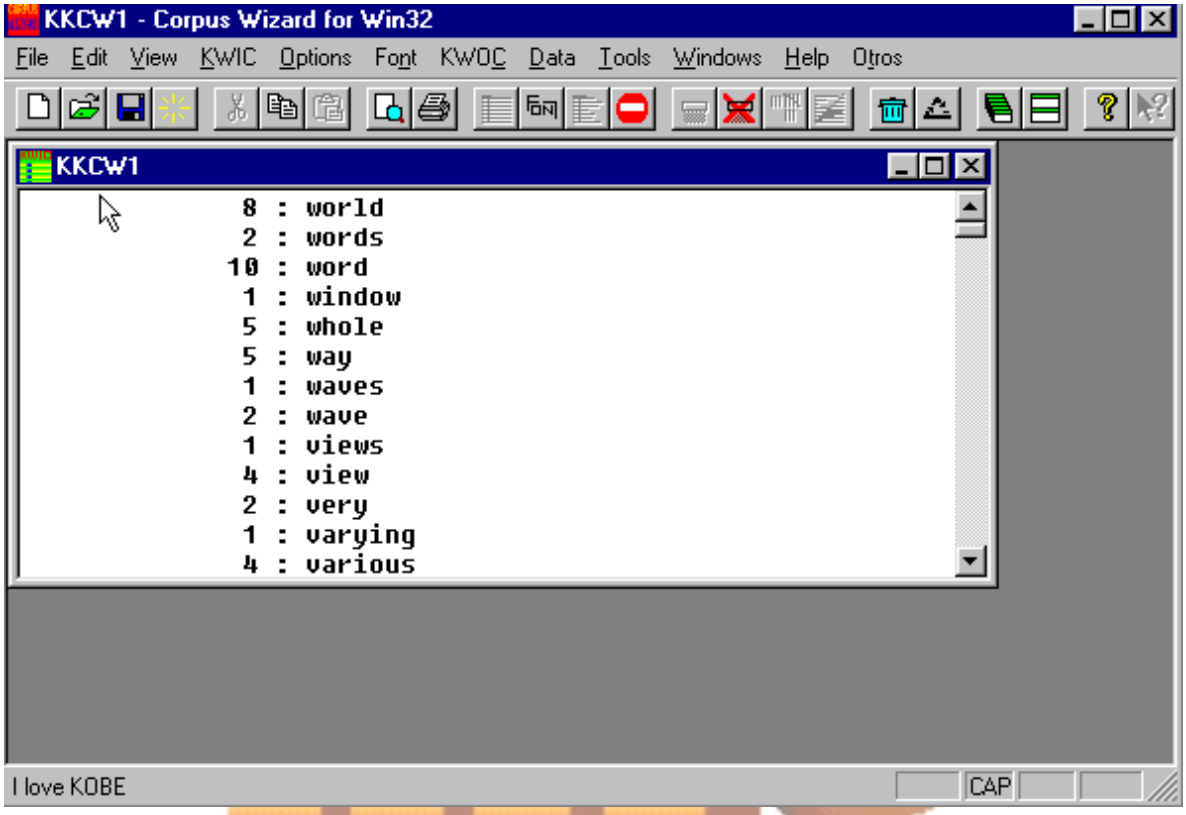
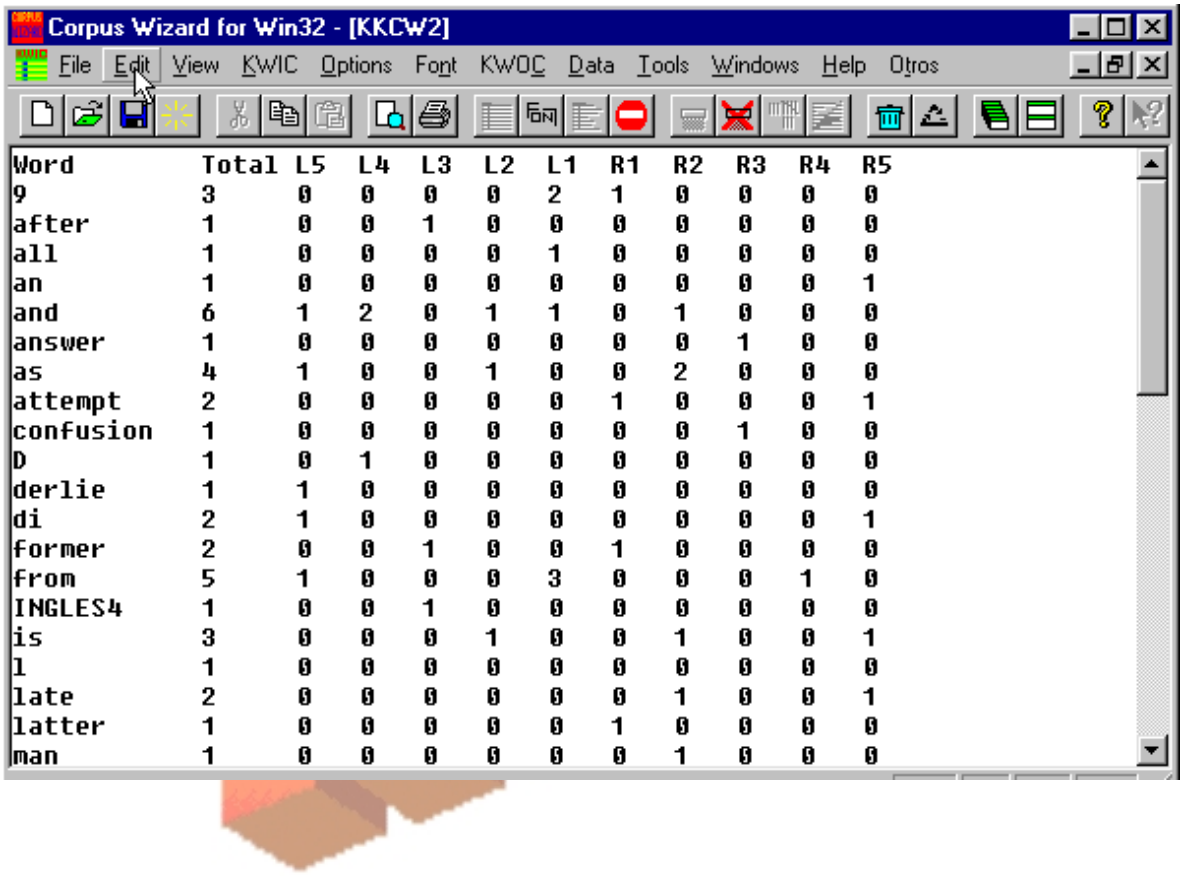


Fig. 27 Estatística colocacional (CORPUS WIZARD)

COLLOCATIONAL STATISTICS(2)

Distribuição das palavras de acordo com sua posição na linha, até cinco posições à esquerda e à direita para todas as COLLOCATIONS de **THE.**



Word	Total	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
9	3	0	0	0	0	2	1	0	0	0	0
after	1	0	0	1	0	0	0	0	0	0	0
all	1	0	0	0	0	1	0	0	0	0	0
an	1	0	0	0	0	0	0	0	0	0	1
and	6	1	2	0	1	1	0	1	0	0	0
answer	1	0	0	0	0	0	0	0	1	0	0
as	4	1	0	0	1	0	0	2	0	0	0
attempt	2	0	0	0	0	0	1	0	0	0	1
confusion	1	0	0	0	0	0	0	0	1	0	0
D	1	0	1	0	0	0	0	0	0	0	0
derlie	1	1	0	0	0	0	0	0	0	0	0
di	2	1	0	0	0	0	0	0	0	0	1
former	2	0	0	1	0	0	1	0	0	0	0
from	5	1	0	0	0	3	0	0	0	1	0
INGLES4	1	0	0	1	0	0	0	0	0	0	0
is	3	0	0	0	1	0	0	1	0	0	1
l	1	0	0	0	0	0	0	0	0	0	0
late	2	0	0	0	0	0	0	1	0	0	1
latter	1	0	0	0	0	0	1	0	0	0	0
man	1	0	0	0	0	0	0	1	0	0	0

Fig. 28 Estatística Colocacional (II) (CORPUS WIZARD)

EXTENDED COLLOCATION STATISTICS – informa as frequências absoluta e relativa do documento aberto, conforme Fig. 29

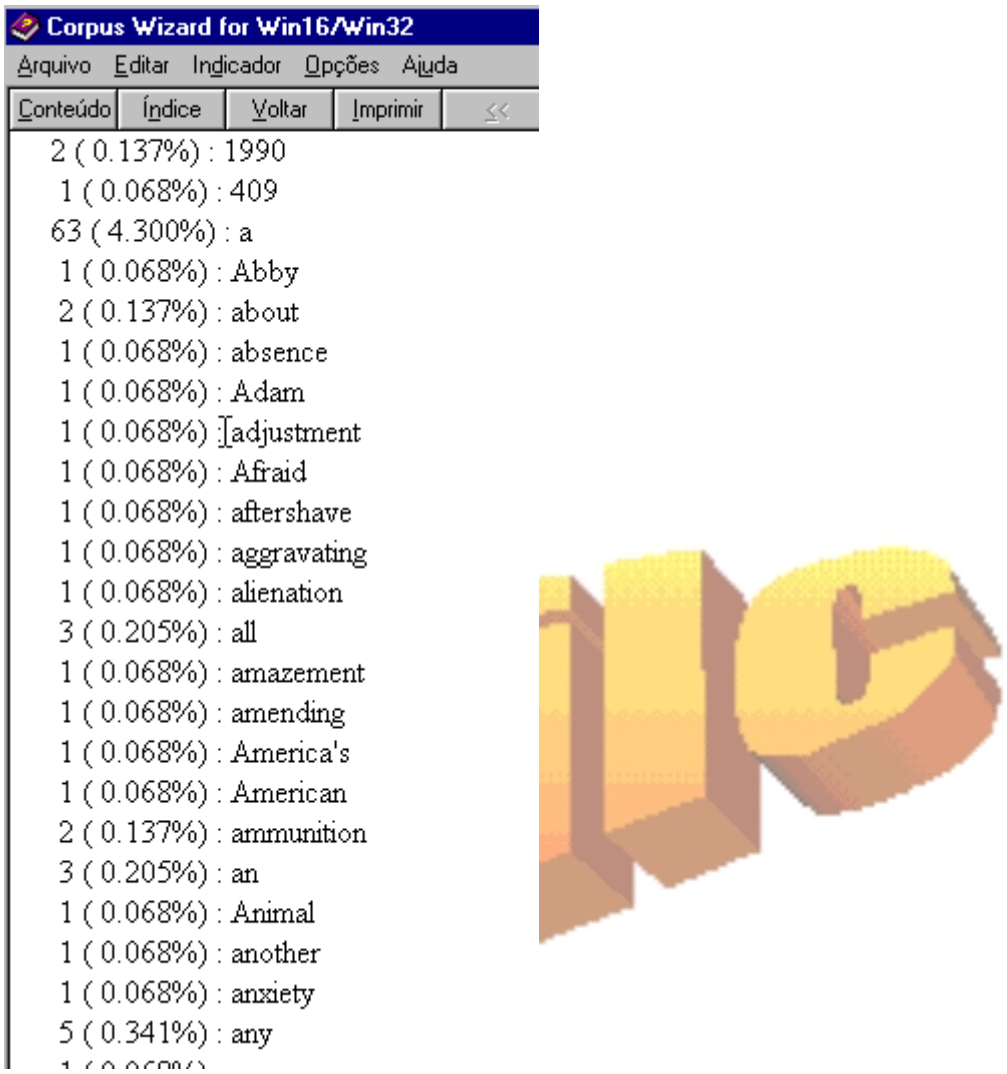


Fig. 29 Freqüências absoluta e relativa (CORPUS WIZARD)

MONOCONC

Possui Lista de Freqüências do Corpus e por colocação.

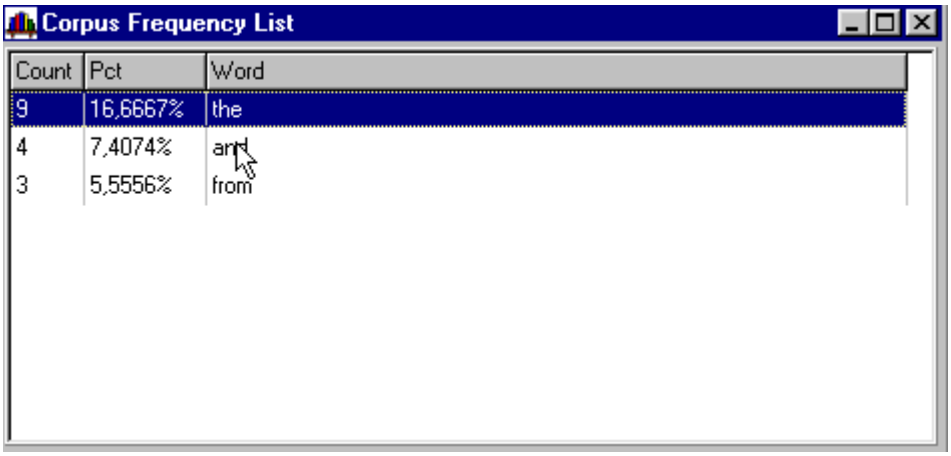


Fig. 30 Lista de freqüencias absoluta e relativa (MONOCONC)

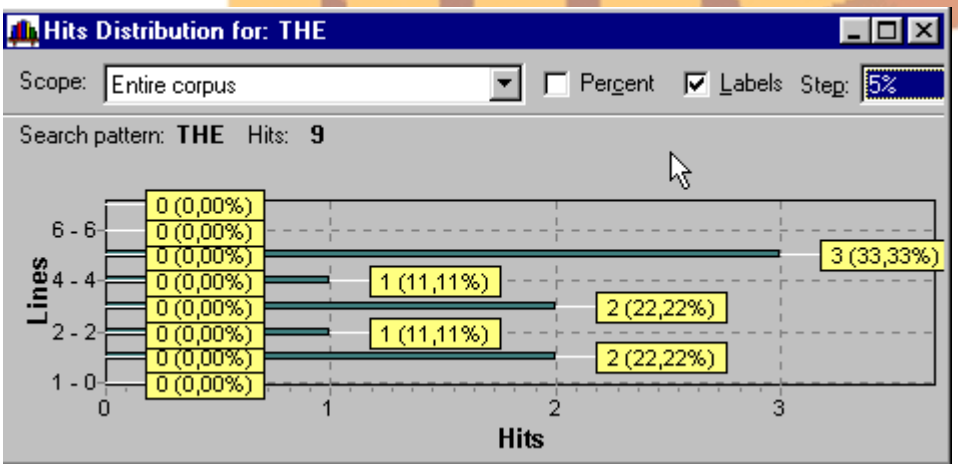


Fig. 31 Distribuição por HITS (MONOCONC) – por linhas

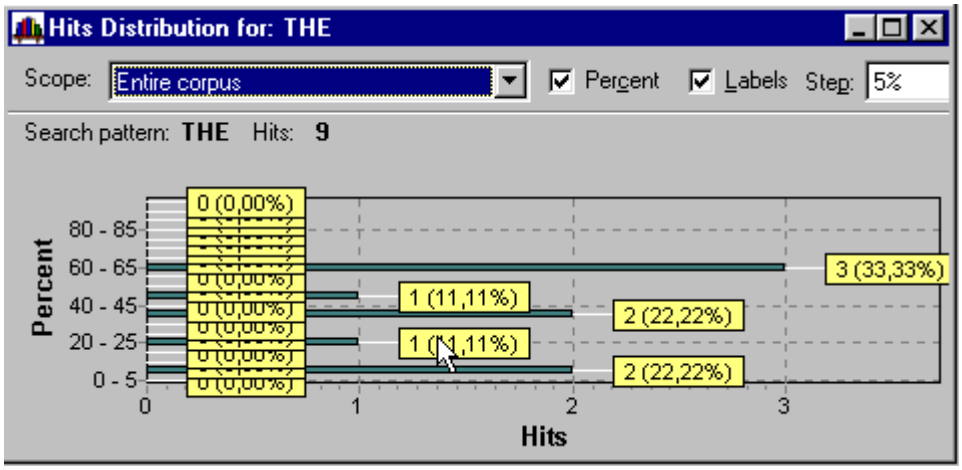


Fig. 32 Distribuição por HITS (MONOCONC) – por percentagem

Distribuição por Hits : Palavra "THE" (neste exemplo) – Distribuição por Hits é a representação gráfica da distribuição das concordâncias da palavra-guia, neste caso, "THE", no texto original. As Figuras 31 e 31, demonstra, portanto, que "**THE"** aparece em cinco linhas do texto. Uma vez aparece em duas linhas (11,11%); Duas vezes em outras duas linhas (22,22%) e, finalmente, em uma linha (na quinta linha), três vezes (33,3%)

WORDSMITH TOOLS

A freqüência das WORDLISTS são obtidas assim: Na primeira coluna, as palavras, tokens e types, a seguir, a frequência absoluta, a freqüência relativa e lemas. Os lemas precisam ser criados pelo usuário, o sistema não dispõe desta informação pronta.

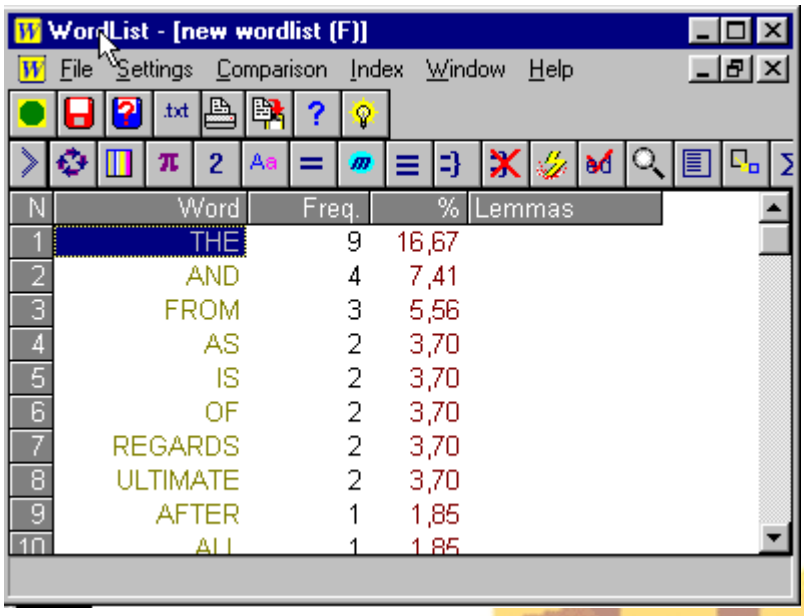
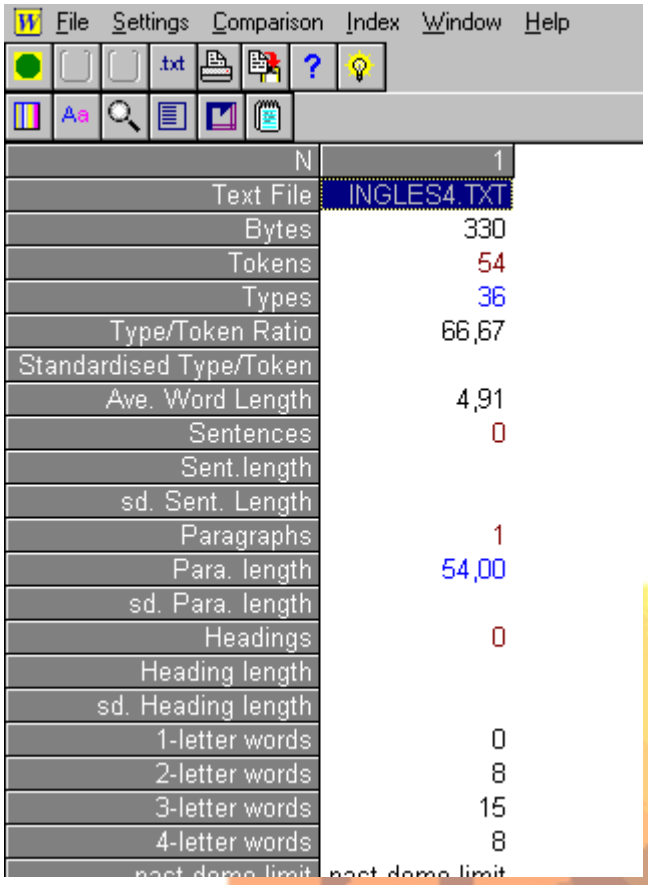


Fig. 33 Lista de palavras juntamente com as freqüências absoluta e relativa (WORDSMITH TOOLS)



N	1
Text File	INGLES4.TXT
Bytes	330
Tokens	54
Types	36
Type/Token Ratio	66,67
Standardised Type/Token	
Ave. Word Length	4,91
Sentences	0
Sent.length	
sd. Sent. Length	
Paragraphs	1
Para. length	54,00
sd. Para. length	
Headings	0
Heading length	
sd. Heading length	
1-letter words	0
2-letter words	8
3-letter words	15
4-letter words	8

Fig 34 Estatísticas para palavras, sentenças e parágrafos (WORDSMITH TOOLS)

WORDSMITH TOOLS oferece uma lista de estatísticas, tanto para as palavras (TOKENS e TYPES), quanto para os parágrafos e sentenças, especialmente em relação ao tamanho destas. Também dispõe da quantidade de Palavras com 1,2,3 e 4 letras presentes no arquivo de trabalho.

CONCORDANCE 2.0 – As estatísticas de CONCORDANCE 2.0, resumem-se a informar o número de linhas, palavras, caracteres e sentenças. O percentual de TYPE/TOKEN no texto e a média de palavras por sentença.

TATOE - As estatísticas de **TATOE** são um pouco diferentes, como trabalha com marcação de texto, possui estatísticas de palavras-conceito e lemas-conceito. Além disso informa o número de esquemas (squemes) utilizados, categorias, estilos e segmentos de texto marcado. As estatísticas, em resumo, podem ser vistas nas duas figuras 35 e 36.

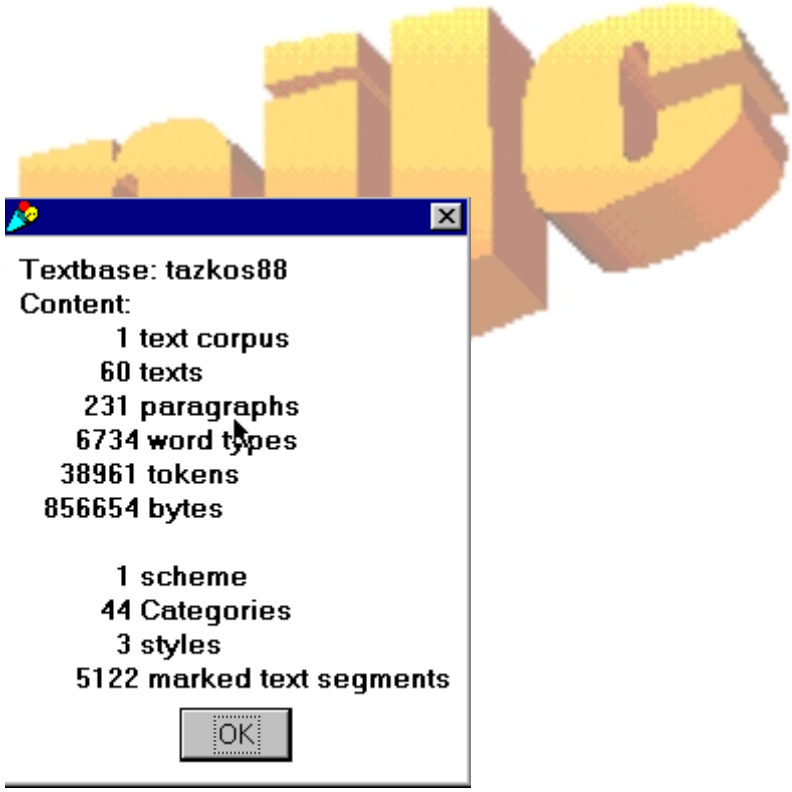


Fig.35 Resumo das estatísticas (TATOE)

Fig. 265

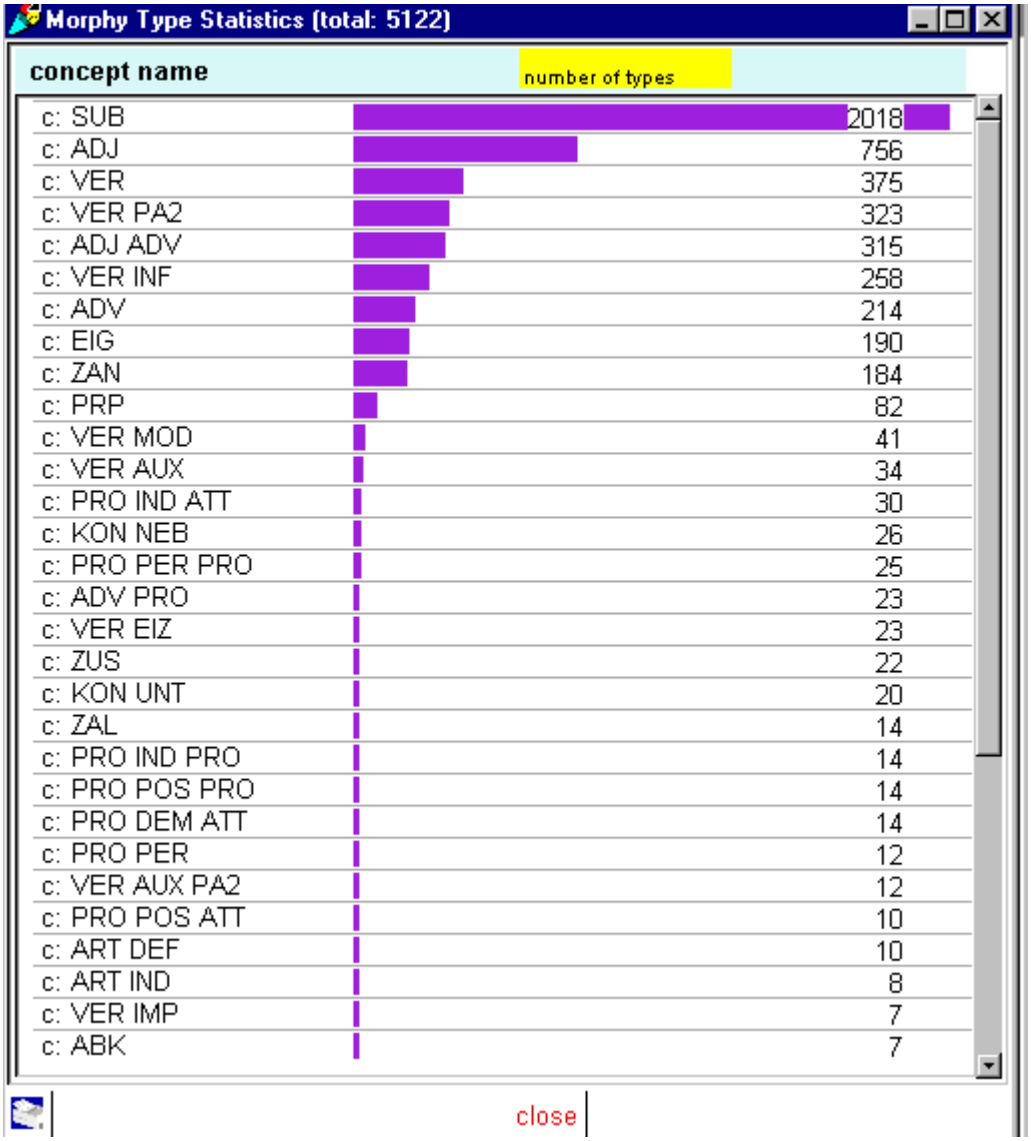


Fig. 36 Número de word types por concept name (TATOE)

2.2.0 Lematização

TABELA XI - LEMATIZAÇÃO

KWIC	CW	RANGE	MC	WS	CONC	TATOE	DICTGEN	TACT
NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO

Análise: CONCORDANCE 2.0 é o único dos softwares analisado que inclui lematização, embora seja parcial e da língua inglesa somente, como pode ser visto na figura abaixo. **WORDSMITH TOOLS** disponibiliza a lematização, mas por parte do usuário, não que esteja incorporada no produto

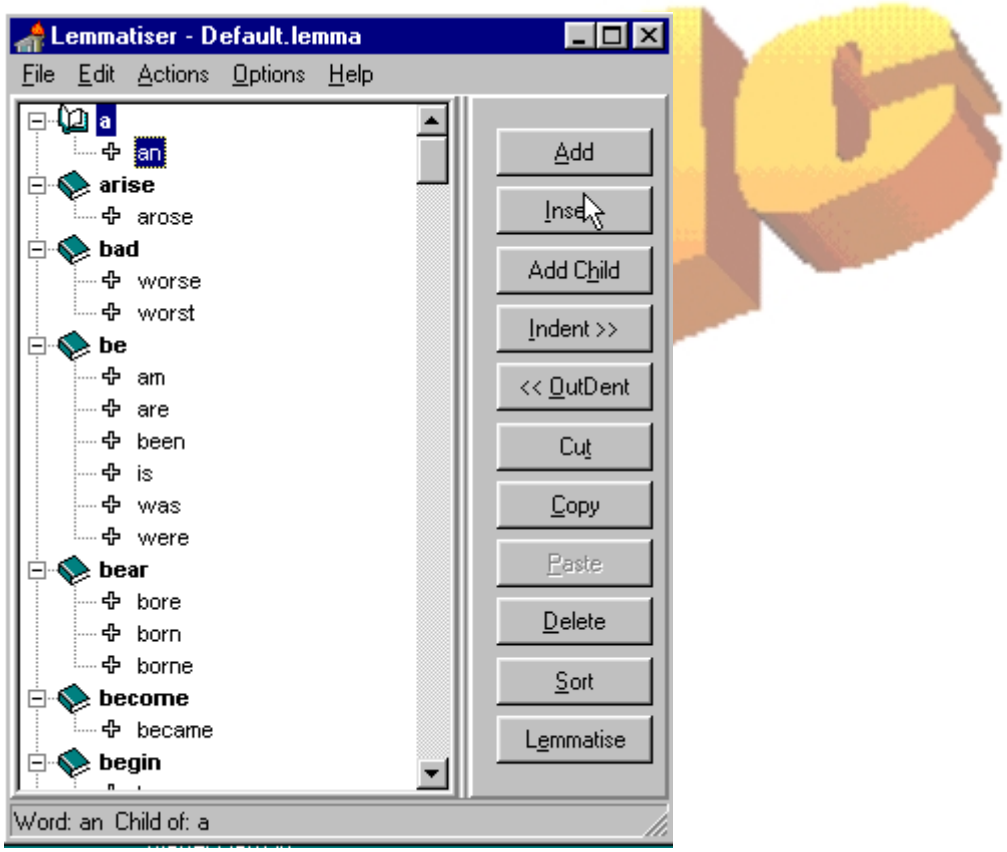


Fig. 37 Lematizador (CONCORDANCE)

Conclusão

Todas as ferramentas analisadas, em maior ou menor grau satisfazem algumas das necessidades de um autor de obras de referência, mas não muitas. Um software para criação de obras de referência necessita ter, ao menos, ferramentas como spelling checker e lematizador, e estas ferramentas não existem nestes produtos de forma satisfatória, em que pese, algumas como WORDSMITH TOOLS, TATOE e CONCORDANCE possuirem um conjunto de recursos muito bons para a criação de dicionários.

Na parte de estatística, há uma carência de exemplos na área de inferência estatística; não há cálculo de desvio-padrão, o que eu considero essencial para validar médias; não há análise de correlação, nem do coeficiente de Pearson.

Embora os recursos pareçam bons, em alguns casos eles são insuficientes. Nenhum dos softwares analisados importam arquivos do tipo .DOC , .RTF ou .PDF.

Bibliografia

ALEXA, M. E ROSTEK, L. 1999. TATOE Text Analysis Tool with Object Encoding v.0987. Darmstadt. GMD-IPSI

ATHELSTAN, MICHAEL B. 1996-2000. MonoConc Pro v. 2.0. Houston. ELF Ltd.

HAMAGUCHI**, TAKASHI. 1995-1999. Corpus Wizard for WIN32E .Kobe. Kobe Phoenix Laboratory.**

MAMO, MARTIN S. 1996-1997. DICTGEN V. 1.0b. London. Martin Mamo.

NATION, PAUL ET ALL. Range and Words – Programs for Windows based PCs. Wellington.
School of Linguistics and Applied Language Studies. Victory University

SCOTT, MIKE . 1999. WordSmith Tools v.3.00. Oxford. Oxford University Press.

TACT GROUP. 1995. TACT v.2.1. Toronto. University of Toronto.

TSUKAMOTO, SATORU. 2001. KWIC Concordance v. 4.6. Nihon. Satoru Tsukamoto (Nihon University).

WATT, R.J.C. 2000. Concordance v. 2.0. Dundee. RJCW.