UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LETRAS MODERNAS

LUCIANA LATARINI GINEZI

CAFÉS DO BRASIL

Estudo de variantes em português e inglês na língua falada

Dissertação apresentada à Faculdade de Filosofia,
Letras e Ciências Humanas - Departamento de Letras
Modernas - Curso de Pós-Graduação em Estudos
Linguísticos e Literários em Inglês da Universidade
de São Paulo para obtenção do título de Mestre
em Estudos Linguísticos e Literários em Inglês

Área de concentração: Terminologia

Orientadora: Prof^a. Dr^a. Stella Esther Ortweiler Tagnin

SÃO PAULO
2007

G492 Ginezi, Luciana Latarini
 Cafés do Brasil : estudo de variantes em português e inglês na língua falada /
 Luciana Latarini Ginezi ; orientadora Stella Esther Ortweiler Tagnin. -- São Paulo,
 2007.
 270 f.

 Dissertação (Mestrado - Programa de Pós-Graduação em Estudos Lingüísticos
 e Literários em Inglês. Área de concentração: Terminologia) - Departamento de
 Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da
 Universidade de São Paulo.

 1. Lingüística de corpus. 2. Café (Terminologia). 3. Terminologia (Metodologia).
 4. Tradução. 5. Socioterminologia. 6. Língua inglesa. 7. Língua portuguesa. 8.
 Vocabulário técnico. I. Título.

21ª. CDD 413.028

FOLHA DE APROVAÇÃO

Luciana Latarini Ginezi

CAFÉS DO BRASIL - Estudo de variantes em português e inglês na língua falada

Dissertação apresentada à Faculdade de Filosofia,
Letras e Ciências Humanas - Departamento de Letras
Modernas - Curso de Pós-Graduação em Estudos
Linguísticos e Literários em Inglês da Universidade
de São Paulo para obtenção do título de Mestre
em Estudos Linguísticos e Literários em Inglês
Área de concentração: Terminologia

Aprovada em:

Banca Examinadora

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Dedicatória

Aos meus amores,

Isabella, Humbertinho

Marcelinho e Samuel

Agradecimentos

Aos meus pais, Antonio e Marlene, pelo amor e oportunidades que me proporcionam.

Aos meus irmãos, Jô, Humberto e Cris, pela amizade e presença constante em minha vida.

À minha grande família, Latarini e Ginezi, especialmente àqueles que passaram a fazer parte dela, Levi, Marcelo e Jozi, pelo incentivo aos meus estudos.

À minha orientadora, Prof^a. Stella Tagnin, pela dedicação, carinho e paciência.

Ao João Henrique Brando, professor e amigo, pelos ensinamentos valiosos.

A todos os professores e alunos do curso de Tradução e Interpretação da Uninove, especialmente às professoras Elaine Trindade e Márcia Fusaro, por acreditarem em meu trabalho.

Aos meus amigos, pelo apoio e compreensão: Carlos Gomes Luiz, Carmem Beatriz, Cristiane Wodevodsky, Di Baruque, Elisa Teixeira, Fábio Flores, Fernanda de Oliveira, Josimeire Martins, Lourdes Gonçalves, Luciana Bertassoli, Luciana Carvalho, Mary Falcão, Mary Helen, Sílvia Dias e Thaís Helena.

A todos os informantes das entrevistas, pela confiança no trabalho realizado.

Ao povo africano, pela receptividade e atenção.

Agradecimento Especial

Aos Srs. Carlos Brando e Paschoal Brando, da empresa P&A Marketing,
pela possibilidade de realizar as pesquisas em interpretação.

RESUMO

GINEZI, L. L. **Cafés do Brasil - Estudo de variantes em português e inglês na língua falada**. 2007. 270f. Dissertação (Mestrado) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

O objetivo deste trabalho é analisar a ocorrência de variantes terminológicas na linguagem de especialidade do Café e verificar a possibilidade de se construir um produto terminológico bilíngüe baseado na oralidade. Inspirado no trabalho profissional de interpretação consecutiva e intermitente, o estudo utiliza *corpora* falados, uma vez que podemos estabelecer uma relação clara entre a oralidade e a interpretação, modalidade oral da tradução. Todas as dificuldades encontradas na construção dos *corpora* falados são explicitadas e algumas sugestões são feitas para futuras pesquisas. A pesquisa segue os princípios da Lingüística de *Corpus* (LC), tanto na elaboração dos *corpora* como também na análise dos dados, essa com o uso da ferramenta computacional *WordSmith Tools*, agilizando o processo e dando a ele confiabilidade. O estudo justifica-se pela importância do conhecimento de variantes terminológicas nas línguas de especialidade e na sua modalidade falada, por intérpretes e por profissionais da área e, também, pelas possibilidades oferecidas pela LC para a pesquisa socioterminológica na oralidade. Assim, compilamos dois *corpora* falados monolíngües, um em português do Brasil e outro em inglês de países diversos, com o tema Café, subdividido em *colheita* e *processamento*, composto por entrevistas face-a-face da pesquisadora com profissionais da área cafeeira e por conversações entre profissionais, em ambas as línguas. Também construímos um *corpus* bilíngüe, composto por interpretações entre falantes de inglês e de português. Em seguida, analisamos os dados dos *corpora*, buscando encontrar variantes. Ao final do trabalho, elaboramos um vocabulário bilíngüe a partir dos dados coletados e das análises efetuadas.

Palavras-chave: Lingüística de corpus. Língua falada. Socioterminologia. Variantes. Café.

ABSTRACT

GINEZI, L. L. **Brazilian Coffee - Portuguese and English Variants in the spoken language**. 2007. 270f. Dissertação (Mestrado) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

The aim of this research is to analyze the presence of terminological variants in the specialty language of coffee and to verify the possibility of building a bilingual vocabulary based on spoken language. The study is guided by the consecutive or liaison interpreting and the use of spoken corpora, once we can establish a close relation between spoken language and interpreting, oral mode of translation. Several difficulties were faced in order to build the spoken corpora, and they are presented with some suggestions for future research. The principles of Corpus Linguistics are followed to the corpora design, as well corpora exploration, using Mike Scott's *WordSmith Tools*. The study is significant due to the knowledge of terminological variants in spoken language, by interpreters and professional workers at a specialty area, and also to the possibilities Corpus Linguistics offers to a socioterminological research at spoken variety. Two monolingual spoken corpora were compiled, one in Brazilian Portuguese and the other in English spoken world-wide. The main topic is *Coffee*, divided in *harvest* and *processing*, with face to face interviews as well as conversations among coffee professionals, in both languages. An interpreting corpus is also included in the work, between Portuguese and English speakers. After analysis, we present a bilingual vocabulary of spoken language, including the variants found for most of the terms.

Keywords: Corpus Linguistics. Spoken language. Socioterminology. Variants. Coffee.

LISTA DE ABREVIATURAS E SIGLAS

ABIC	- Associação Brasileira da Indústria de Café
BNC	- British National Corpus
CD	- Cereja descascado
CHAT	- Codes for Human Analysis of Transcripts
CHILDES	- Child Language Data Exchange System
CLAN	- Computerized Language Analysis
ICE	- International Corpus of English
KWIC	- Key Word in Context
LC	- Lingüística de <i>Corpus</i>
LLC	- London-Lund Corpus
MICASE	- Michigan Corpus of Academic Spoken English
NURC	- Projeto de Estudo da Norma Lingüística Urbana Culta
PLN	- Programa de Linguagem Natural
SBCSAE	- Santa Barbara Corpus of Spoken American English
TCT	- Teoria Comunicativa da Terminologia
TGT	- Teoria Geral da Terminologia
WST	- WordSmith Tools

LISTA DE FIGURAS

Figura 1 - Flores do café arábica na Região da Alta Mogiana	010
Figura 2 - Plantio de café entre árvores (Etiópia)	013
Figura 3 - Primeira florada de um pé de café.....	014
Figura 4 - Moega do lavador (Brasil).....	017
Figura 5 - Tanque do lavador (Brasil).....	017
Figura 6 - Carriolas de transporte de café do lavador para o terreiro (Brasil)	019
Figura 7 - Saída de café direto do lavador para o terreiro	019
Figura 8 - Sistema de terreiros de cimento (Brasil)	020
Figura 9 - Secadores rotativos na Etiópia	021
Figura 10 - Estrutura externa de secadores à lenha (Brasil)	021
Figura 11 - Terreiros suspensos (Etiópia).....	022
Figura 12 - Terreiros suspensos que podem ser cobertos à noite (Etiópia)	023
Figura 13 - Tela inicial do WordSmith Tools	043
Figura 14 - Tela da WordList – WordSmith Tools - versão 4.0.....	044
Figura 15 - Lista em ordem alfabética (WordList).....	045
Figura 16 - Lista em ordem de frequência (WordList)	045
Figura 17 - Lista estatística (WordList).....	046
Figura 18 - Tela de KeyWords – WordSmith Tools 4.0	047
Figura 19 - Lista de palavras-chave (KeyWords) - Por ordem de frequência.....	048
Figura 20 - Tela do Concord - WordSmith Tools - v.4.0	049
Figura 21 - Tela de lista de concordâncias (Concord).....	049
Figura 22 - Lista de colocados (Concord).....	050
Figura 23 - Corpora Falados.....	078
Figura 24 - Corpora Escritos.....	078
Figura 25 - Corpora de referência	079
Figura 26 - Mapa regional das regiões abrangidas no PB	086
Figura 27 - Lista estatística do corpus CAF_BO.....	089
Figura 28 - Lista estatística do CAF_IO	090

Figura 29 - Lista de palavras (estatística) do CAF_BI.....	091
Figura 30 - Configuração de extração de língua portuguesa para ignorar o PB	092
Figura 31 - Lista parcial de freqüência em inglês do corpus CAF_BI	092
Figura 32 - Lista estatística com problemas de leitura de arquivos	093
Figura 33 - CAF_BI_PB - Estatística.....	094
Figura 34 - CAF_BI_IN - Estatística.....	095
Figura 35 - Lista estatística do Corpus CAF_BE	101
Figura 36 - Lista estatística do Corpus CAF_IE	101
Figura 37 - Lista estatística do corpus de referência em português.....	102
Figura 38 - Lista estatística do corpus de referência em inglês	103
Figura 39 - Lista estatística do corpus NURC.....	104
Figura 40 - Árvore de domínio para a pesquisa terminológica.....	106
Figura 41 - <i>Stop List</i>	110
Figura 42 - CAF_BO - Lista de freqüência parcial	111
Figura 43 - Lista de Freqüência do Corpus CAF_IO	112
Figura 44 - Lista parcial de freqüência bilíngüe do corpus CAF_BI.....	115
Figura 45 - Lista parcial de freqüência CAF_BI_PB	116
Figura 46 - Lista parcial de freqüência CAF_BI_IN	116
Figura 47 - Lista parcial de freqüência CAF_ALL_PB	118
Figura 48 - Lista parcial de freqüência CAF_ALL_IN	119
Figura 49 - Tela do WordSmith Tools da ferramenta KeyWords.....	120
Figura 50 - Lista parcial de palavras-chave CAF_ALL_PB	121
Figura 51 - Lista parcial de candidatos a termos em português.....	123
Figura 52 - Lista parcial de candidatos a termos em inglês	124
Figura 53 - Lista final de candidatos a termo do corpus CAF_ALL_IN	125
Figura 54 - Palavra FOTO encontrada na lista de palavras-chave	126
Figura 55 - Linhas de concordância da palavra <u>foto</u>	127
Figura 56 - Linhas de concordância da palavra <u>mundo</u>	127
Figura 57 - Agrupamentos de palavras para <u>café</u> - 1ª página	128
Figura 58 - Agrupamentos de palavras para <u>café</u> - 2ª página	129

Figura 59 - Lista de colocados para a palavra <u>café</u>	130
Figura 60 - Linhas de concordância de <u>colheita</u>	131
Figura 61 - Linhas de concordância de colheita a dedo e colheita seletiva	132
Figura 62 - Colheita a dedo - Quatro ocorrências	132
Figura 63 - Colheita seletiva - 10 ocorrências	133
Figura 64 - Agrupamento de palavras para <u>cereja</u>	134
Figura 65 - Linhas de concordância de <u>cereja</u>	134
Figura 66 - Linhas de concordância de <u>água</u>	135
Figura 67 - Agrupamentos de palavras para <u>água</u>	136
Figura 68 - Linhas de concordância de verde - Parcial	136
Figura 69 - Agrupamentos de palavras para <u>verde</u>	137
Figura 70 - Linhas de concordância de <u>máquina</u>	138
Figura 71 - Agrupamentos de palavras para <u>máquina</u>	138
Figura 72 - Linhas de concordância da palavra <u>grão</u> (1ª página)	139
Figura 73 - Linhas de concordância da palavra <u>grão</u> (2ª página)	139
Figura 74 - Agrupamento de palavras para <u>grão</u>	140
Figura 75 - Linhas de concordância para <u>pé</u> - Parcial	141
Figura 76 - Agrupamento de palavras para <u>pé</u>	141
Figura 77 - Linhas de concordância da palavra <u>pé</u> - Parcial	142
<u>Figura 78</u> - Linhas de concordância para a palavra tipo - Parcial	143
Figura 79 - Linhas de concordância de <u>região</u> - Parcial	144
Figura 80 - Linhas de concordância de <u>região</u> - Parcial	144
Figura 81 - Lista de concordâncias da palavra <u>patio</u> (lado esquerdo)	146
Figura 82 - Lista de concordância da palavra <u>patio</u> (lado direito)	146
Figura 83 - Lista de concordâncias para a palavra <u>patios</u> (lado esquerdo)	147
Figura 84 - Linha de concordância onde aparece <u>terreirão</u>	148
Figura 85 - Lista de frequência CAF_ALL_PB	148
Figura 86 - Linhas de concordância de <u>terreirão</u> (lado esquerdo)	149
Figura 87 - Linhas de concordância de <u>terreirão</u> (lado direito)	150
Figura 88 - Lista de palavras CAF_BE	152

Figura 89 - Lista de palavras do CAF_IE	153
Figura 90 - Lista de frequência CAF_IE: lematização de <u>bean/beans</u>	154
Figura 91 - Candidatos a termo em português - Lista parcial.....	155
Figura 92 - Candidatos a termo final - Lista parcial	156
Figura 93 - Candidatos a termo do CAF_IE - Parcial	156
Figura 94 - Candidatos a termo final - Lista parcial	157
Figura 95 - Ficha terminológica.....	159

LISTA DE QUADROS

Quadro 1 -	Tipologia de Corpus	031
Quadro 2 -	Classificação de variantes terminológicas.....	070
Quadro 3 -	Normas para transcrição.....	083
Quadro 4 -	Subcorpora falados entrevistas/conversação, PB e EN	090
Quadro 5 -	Subcorpus falado Interpretações	096
Quadro 6 -	Nomeação de arquivos	097
Quadro 7 -	Modelo de cabeçalho em português.....	098
Quadro 8 -	Trecho de texto anotado	099
Quadro 9 -	Corpora escritos - Número de palavras.....	100
Quadro 10 -	Comparação de corpus de referência	105
Quadro 11 -	Comparação das listas de frequência em português	117
Quadro 12 -	Comparação das listas de frequência em inglês.....	118

SUMÁRIO

INTRODUÇÃO	001
CAPÍTULO 1 - CAFÉ	007
1.1 O Que Há Por Trás de Uma Xícara.....	008
1.2 Etapas Para a Produção do Café.....	012
CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA	026
2.1 Lingüística de Corpus	027
2.1.1 Breve definição.....	028
2.1.2 Corpora falados.....	033
2.1.2.1 Corpora desenvolvidos ou em desenvolvimento.....	033
2.1.2.2 Compilação de corpus falado e programas computacionais de análise.....	037
2.1.2.3 Extração terminológica e produção de vocabulários com a LC	051
2.2 Interpretação e Oralidade	051
2.2.1 A arte de interpretar	052
2.2.2 Panorama teórico da interpretação	056
2.2.2.1 A Teoria interpretativa da tradução ou modelo interpretativo.....	058
2.2.2.2 Modelo de esforço.....	059
2.2.3 Pesquisas em LC	060
2.2.4 Contrastes entre fala e escrita: a questão tradutória	061
2.3 Terminologia e Variantes - Uma Leitura Social.....	066
2.3.1 Terminologia e interpretação	072
2.3.2 Terminologia e lingüística de corpus	073
CAPÍTULO 3 - METODOLOGIA.....	075
3.1 A Construção dos Corpora	077
3.1.1 Coleta de dados	079
3.1.2 Transcrição de dados.....	081
3.1.3 Design dos corpora.....	084
3.1.3.1 Corpora falados “Café”.....	085
3.1.3.1.1 Marcação dos corpora falados.....	096
3.1.3.2 Corpora Escritos CAF_BE e CAF_IE	099
3.1.3.3 Corpora de referência escritos.....	102
3.1.3.4 Corpus de referência falado.....	103
3.2 Explorando os corpora.....	105
3.2.1 Corpora falados.....	109
3.2.1.1 Conhecendo os Corpora - Palavras mais freqüentes	110
3.2.1.2 Conhecendo os Corpora - Candidatos a termo.....	119
3.2.2 Conhecendo os corpora: o que dizem as linhas de concordância.....	126
3.2.3 Corpora escritos.....	151
3.2.3.1 Conhecendo o corpus - Palavras mais freqüentes	151
3.2.3.2 Conhecendo o corpus: candidatos a termo.....	154
3.3 Construção do Glossário	158
3.3.1 Seleção final de termos: verbetes para glossário	158
3.3.2 Protótipo do glossário ideal.....	160

CAPÍTULO 4 - ANÁLISE DOS DADOS	163
4.1 Língua Falada X Língua Escrita: O Que Dizem os Termos.....	164
4.2 Sociotermos: Análise de Casos.....	165
4.3 Vocabulário Bilíngüe - Café: Colheita e Processamento.....	169
 CAPÍTULO 5 - CONSIDERAÇÕES FINAIS	 176
REFERÊNCIAS BIBLIOGRÁFICAS.....	181
ANEXOS	190
BIBLIOGRAFIA CONSULTADA.....	261

INTRODUÇÃO

A presente dissertação de mestrado tem como objetivo analisar as ocorrências de variantes terminológicas na modalidade oral da língua de especialidade da cafeicultura, especificamente sobre a colheita e o processamento do café pós-colheita, além de elaborar um vocabulário sobre o tema, contendo as variantes acima citadas.

A idéia inicial para o projeto de mestrado partiu da observação da pesquisadora, durante sua atuação como intérprete de línguas estrangeiras (inglês <> português), da problemática de variantes e da falta de material terminológico de consulta na subárea da cafeicultura. A intérprete recebia delegações estrangeiras falantes de inglês de diversos países do mundo para conhecerem o funcionamento de usinas cafeeiras no Brasil. Para interpretar um falante estrangeiro há a exigência, entre outras coisas, do conhecimento do registro formal e informal das línguas envolvidas. Os participantes vêm ou podem vir de estratos sociais diferentes, ou seja, podem ser engenheiros agrônomos (com diploma universitário) ou trabalhadores rurais (sem instrução ou com instrução mínima), o que fica evidente em seu linguajar. Assim sendo, a preocupação com as variantes é pertinente para o trabalho desses profissionais.

Assim, pensando nessa questão e unindo o interesse pela área da cafeicultura, na qual trabalhamos com traduções e interpretações, decidimos investigar o assunto, propondo-nos a responder às seguintes questões:

- a) Quais são as ocorrências de variantes terminológicas na língua falada em um corpus de linguagem de especialidade do Café?
- b) Qual é a diferença entre a língua falada e a língua escrita em relação à riqueza lexical?

- c) Essas variantes devem constar de um vocabulário técnico da área? Se sim, como devem ser registradas?

Teremos, portanto, a divisão desta dissertação como segue abaixo.

Para tentar alcançar nosso objetivo, discutiremos, no primeiro capítulo, o tema, *Café*, escolhido devido à possibilidade de coletar dados em interpretações consecutivas e intermitentes realizadas nas empresas *P&A Marketing Internacional* e *Pinhalense Máquinas Agrícolas S.A.*¹. Também utilizaremos grande quantidade de dados obtidos em uma visita feita pela pesquisadora à Etiópia, país considerado berço do café, para participar do Congresso EAFCA (Eastern African Coffee Association) Conference & Exhibition, em Addis Ababa, em fevereiro de 2007, no qual estavam presentes pessoas diretamente ligadas ao café vindas do mundo todo, e cuja língua oficial era o inglês. Há, portanto, na pesquisa, profissionais da área cafeeira que falam português do Brasil, das micro-regiões Mogiana e Sul-Mineira, e inglês de países diversos, tais como: Tanzânia, Quênia, Estados Unidos da América, Sri Lanka, Ruanda, Uganda, entre outros. A seleção dos países está condicionada às relações deles com o tema *café* e ao fato de serem países que falam o inglês como língua internacional de negócios. No início, pensamos em elaborar o estudo considerando apenas falantes de inglês como língua nativa, pois segundo Cheshire (1991), há uma grande quantidade de variantes no inglês falado, mesmo como língua nativa. Porém, como língua internacional amplamente utilizada, não haveria razão para nos limitarmos à língua nativa, uma vez que já se fala em Inglês Internacional, ou seja, a variedade sem demarcação geográfica, uma variedade livre de regionalismos (Peters, 1998). Trata-se de um capítulo importante, pois pretende também fornecer um panorama da área ao leitor, de modo que possa melhor compreender os exemplos apresentados ao longo de todo o trabalho. Para isso, introduziremos o estudo com aspectos relacionados ao café, tais como sua função econômica e social no país e fora dele, além da definição de Colheita e Processamento, explicando várias técnicas, processos, assim como termos utilizados no decorrer do trabalho.

¹ Empresas estabelecidas na cidade de Espírito Santo do Pinhal, estado de São Paulo.

Quanto à base teórica, a pesquisa desenvolve-se com base em quatro disciplinas, devidamente explicadas no Capítulo 2 - Fundamentação Teórica. São elas:

- Lingüística de Corpus (LC), abordagem utilizada para a investigação de fenômenos prováveis da língua;
- Socioterminologia, ramo da Terminologia para o estudo de variantes terminológicas;
- Sociolingüística, voltada para os estudos de língua falada;
- Interpretação, área na qual identificamos a necessidade de pesquisas sobre variantes terminológicas.

As duas últimas, Sociolingüística/Oralidade e Interpretação, serão apresentadas em conjunto, devido à sua interação inseparável em nosso contexto.

Portanto, no segundo capítulo definiremos a teoria que fundamenta todo o trabalho. No decorrer da discussão teórica, procuraremos esclarecer os aspectos que são convergentes para esta pesquisa, ou seja, os aspectos que integram as quatro disciplinas. Apresentaremos em primeiro lugar a LC, pois sua linha metodológica servirá como alicerce desta pesquisa. A seguir, trataremos da Interpretação e de suas interações com a oralidade, discorrendo sobre a arte de interpretar e os requisitos necessários para isso. Teceremos, também, considerações a respeito de teorias que marcaram os estudos na Interpretação e citaremos alguns estudos feitos com a integração da LC. Comentaremos, ainda, o papel da Sociolingüística como metodologia responsável pelas pesquisas de língua falada. Por fim, apresentaremos, ainda nesse capítulo, a Terminologia, com foco na Socioterminologia, cujos princípios seguiremos para a definição de termos e elaboração do vocabulário.

A relevância dos estudos de variantes terminológicas na oralidade justifica-se por diversos fatores. O primeiro seria que o profissional da Interpretação, tanto de Interpretação Simultânea como Consecutiva, ou ainda Intermitente², trabalha com a língua falada e, portanto, deve ter

² As definições serão explicitadas com mais detalhes no Capítulo 2.

conhecimento das realizações terminológicas da fala. Sabemos que a pesquisa para a elaboração de glossários para intérpretes é normalmente feita em língua escrita (Gile, 1995a). Entretanto, a língua escrita possui características que a diferem da língua falada (Urbano, 2002) e, dessa forma, interferem no uso terminológico. Segundo, o intérprete deve ter domínio amplo das línguas de trabalho, tanto da língua geral como das línguas de especialidade, incluindo terminologias e suas variantes. A possibilidade de verificar essas ocorrências na língua em uso reflete a preocupação com o que ocorre de fato no âmbito socioprofissional (Faulstich, 1995a). Não poderíamos deixar de mencionar a quase inexistência de pesquisas em língua oral na Terminologia. É notável o trabalho de pesquisadores pioneiros como Urbano (1978) e Alves (2003). Alves (2003) avalia características apresentadas pelo léxico na língua falada, a partir da transcrição de uma aula universitária sobre Economia. Já Urbano (1978), em sua dissertação de Mestrado, analisa as variantes socioprofissionais do Automobilismo, sob um prisma que as diferencia entre terminologia técnica e popular. Essas pesquisas foram realizadas sem a utilização da metodologia da Lingüística de Corpus, mas atenderam prontamente aos objetivos que se propuseram. Outra pesquisa que podemos citar, quando falamos em uso de informática voltada à análise lingüística, é a de Camlong e Zapparoli (2002), demonstrando novas perspectivas de análises do discurso por meio do método matemático-estatístico-computacional (ferramenta STABLEX). Por último, como poucas pesquisas são realizadas sobre variantes utilizando a abordagem da LC, este estudo poderá abrir caminhos e direcionar outros pesquisadores a investigações na área. A utilização de corpora falados para estudos de variações lingüísticas têm sido amplamente disseminada nos últimos anos (Reppen, Fitzmaurice e Biber, 2002). Entretanto, estudos específicos de variações terminológicas são recentes e, mesmo utilizando corpora, não seguem os procedimentos da LC (Strehler, 1995; Miranda, 1995).

O terceiro capítulo apresenta a metodologia de nosso trabalho, explicitando a construção dos corpora da pesquisa e o método de análise

utilizado, com exemplos. Demonstraremos, passo a passo, o que foi feito para solucionar as questões que sempre surgem do trabalho com corpora orais: como coletar dados que possam ser utilizados e atendam às demandas da pesquisa; como dinamizar o procedimento de transcrição – utiliza-se software para esse trabalho ou não. Em seguida trataremos de questões gerais, abordadas tanto na escrita como na oralidade, como, por exemplo, quais são as formas mais eficientes para se obter candidatos a termos; como fazer a marcação de um corpus etc.

O quarto capítulo apresenta o resultado da análise em forma de variantes terminológicas, classificando-as segundo os critérios da Socioterminologia, e inserindo-as em um protótipo de vocabulário desenvolvido seguindo tais princípios terminológicos. A importância de elaborarmos um vocabulário sobre café em português do Brasil e inglês justifica-se pela falta de material de consulta terminológica nessa área no país, o maior produtor mundial de café. Não existe sequer um vocabulário ou dicionário bilíngüe sobre café, para o par de línguas português<>inglês. Tampouco encontramos grande diversidade de enciclopédias, dicionários ou vocabulários específicos para a cafeicultura em bibliotecas, como a Esalq - USP (Escola Superior de Agricultura Luiz de Queiroz), ou em grandes livrarias nacionais e internacionais. Embora haja diversos títulos sobre café, nenhum deles baseia-se em trabalho terminológico. Dentre as publicações que nos chamaram a atenção está o Guia do Barista (Bressani, 2007), com um pequeno glossário monolíngüe. Com o crescente interesse de exportadores brasileiros do produto para o mercado internacional e a constante visita de estrangeiros interessados em conhecer o manejo do café no Brasil, acreditamos que obras técnicas bilíngües auxiliariam todos os profissionais da área.

No quinto capítulo desenvolvemos a conclusão, partindo do trabalho analisado, e apresentamos sugestões para futuras pesquisas.

CAPÍTULO 1

CAFÉ

1.1 O Que Há Por Trás de Uma Xícara

Todas as vezes em que bebemos uma xícara de café estamos consumindo o trabalho de milhares e milhares de pessoas que o fizeram chegar até a nossa mesa. Estamos também degustando a história que há por trás dessa bebida tão popular. No mundo todo se produz uma média de 120 milhões de sacas/ano, das quais cerca de 32 milhões são fornecidas pelo Brasil (ABIC, 2007), maior produtor mundial da variedade chamada Arábica. São dois bilhões de xícaras consumidas por dia no mundo, quantidade essa que só tende a aumentar, principalmente no Brasil, considerado o próximo maior consumidor mundial de café até 2010. Junte-se a isso que o café é uma das maiores *commodities* brasileiras e mundiais, ao lado do açúcar, soja, trigo, entre outras, e teremos um produto agrícola de enorme poder social e econômico no decorrer dos tempos.

Há muitas histórias sobre a origem do café. A lenda mais conhecida é a do pastor Kaldi, que vivia na Abissínia, atual Etiópia, há cerca de mil anos. Kaldi observou que todas as vezes em que suas cabras comiam o fruto vermelho-amarelado de certa árvore, ficavam saltitantes e cheias de energia. Contou para um monge da região, que resolveu provar do fruto, fazendo uma infusão. O monge certificou-se, então, de que todas as vezes em que provava do fruto, era capaz de ficar horas e horas acordado, fazendo suas leituras noturnas, e assim a fama do produto espalhou-se por muitos monastérios. Dizem que o primeiro cultivo planejado foi em mosteiros do Yêmen, apesar de haver controvérsias quanto a isso. Na Etiópia, onde até hoje o café faz parte da vegetação nativa, encontraram as primeiras plantas, mas os registros de cultivo do café são datados de 575, no Yêmen. O café,

na época, era consumido *in natura*. Faziam uma infusão com os frutos e tomavam. Com o passar do tempo, na Pérsia, o café começou a ser torrado e moído, em uma preparação mantida em segredo, para que ninguém soubesse como era feito. Os árabes tampouco deixavam que estrangeiros se aproximassem das mudas, guardadas a sete chaves, ou que saíssem da Pérsia com elas. Para formar a muda, a semente do café deve estar com o pergaminho, que vem logo abaixo da casca. Os árabes, então, deixavam que as sementes saíssem sem o pergaminho, para que pudessem deter o poder sobre o cultivo da planta. Com o tempo, entretanto, o segredo foi sendo revelado, não se sabe exatamente como, pois o cultivo do café foi propagado entre os trópicos, região própria para seu cultivo. Atualmente há 70 países que cultivam o café.

Muitos acreditam que o nome “café” seja originário da palavra *Kaffa*, local de origem da planta. Porém, a verdadeira origem é do árabe *qahwa*, que significa vinho. Quando o café chegou à Europa, em meados do século XIV, era chamado de “vinho das Arábias”. Muitos “cafés” se abriram na Europa, para servir a famosa bebida. Desde então, tornou-se comum encontrarmos no velho continente locais onde o café é degustado, acompanhado de comidas ou não, onde ao mesmo tempo promovem-se reuniões para negócios ou lazer. Atualmente, no entanto, os “cafés” são também chamados “cafeterias” e podem referir-se a bares e restaurantes, que servem muito mais do que apenas a bebida café.

No Brasil, o café chegou pelas mãos de Francisco de Mello Palheta, em Belém, PA. Ele o trouxe via Guiana Francesa, durante uma missão em que fora enviado pelo Grão Pará. Há uma história de que Palheta conquistara a esposa do governador de Caiena e ela fugira para o Brasil com ele. Como era proibido sair com a planta de café do país, ela gentilmente lhe trouxera uma muda, colocando-a sob as saias de seu vestido e saindo do navio sem ser notada.

Após seu plantio no Pará, o café passou por diversos pontos no Brasil. Atualmente, o cultivo ocorre principalmente nas micro-regiões da Alta Mogiana, que inclui áreas do estado de São Paulo, e no sul de Minas

Gerais, além da do Cerrado Mineiro, Bahia, Rondônia, Paraná e Espírito Santo. A melhor qualidade, entretanto, pode ser encontrada nas produções da Alta Mogiana e Sul de Minas, que produzem a espécie Arábica (Figura 1).

Figura 1 - Flores do café arábica na Região da Alta Mogiana

A produção desses cafés de qualidade é uma conquista recente do mercado brasileiro. Atualmente, a maior preocupação da indústria nacional do café é que possamos oferecer qualidade, não apenas quantidade, item pelo qual o Brasil sempre foi muito conhecido devido à sua enorme produção anual. Diversas cooperativas e associações estão promovendo o desenvolvimento de cafés de qualidade, para que o Brasil não seja apenas o maior exportador, mas também um exportador de qualidade. Pesquisas recentes comprovam que até 2010 o Brasil será também o maior consumidor interno de café no mundo. Atualmente o maior consumidor em volume são os Estados Unidos da América. Vemos que mais e mais cafeterias são abertas em todos os lugares do Brasil, favorecendo a oferta de qualidade para os novos consumidores, apreciadores da bebida.

Mas para chegar até nossa mesa, o café percorre um longo caminho. Antes de o aroma tão característico invadir nosso desejo de um

cafezinho, o fruto já passou por vários processos de transformação. A semente transforma-se em muda. Plantando-a, temos uma árvore que dará frutos vermelhos ou vermelho-amarelados e, logo que colhidos, serão lavados, terão sua casca retirada ou não, serão secos, e enfim armazenados, aguardando apenas a reta final, do beneficiamento e torra, até virar pó.

Na terminologia encontramos também um longo caminho, menos saboroso, porém interessante como tal. A palavra café é polissêmica. Como reza a Enciclopédia Agrícola Brasileira (Inglez de Souza, 1998): “A palavra designa a planta, a plantação, o fruto, a semente, a bebida obtida da infusão feita com a semente torrada e moída, e também o estabelecimento comercial de preparo e venda do cafezinho” (p.46, vol. 2). Outro exemplo de polissemia é “verde”. Quando dizemos café “verde” estamos nos referindo ao café que está pronto para a venda, a um passo da torra. Já o grão “verde” é aquele que ainda não está bom para ser colhido, ou seja, é o contrário de grão “cereja” ou “maduro”. Se a palavra “verde” não estiver contextualizada, pois muitas vezes omite-se “café” ou “grão”, não saberíamos dizer ao que se refere. Caminhando agora na língua inglesa, há divergências. Por exemplo, no Quênia é comum que trabalhadores rurais refiram-se à tulha ou armazém como “*go down*”. Nos Estados Unidos só conhecem o local como “*warehouse*”.

Este capítulo, portanto, terá como finalidade explicar ao leitor um pouco sobre a planta, a plantação, o fruto, a semente e a bebida, cada qual compreendido em seu contexto. Pretendemos resumir como ocorre cada etapa, do início da plantação até a fase da bebida, procurando demonstrar uma visão geral da área a fim de compreender sua terminologia.

1.2 Etapas Para a Produção do Café

As etapas para a produção da bebida podem ser divididas, porém estão inter-relacionadas. Não podemos falar, por exemplo, em qualidade da bebida, sem que as condições propícias para a colheita estejam envolvidas. Elaboramos um roteiro simples, apenas elucidativo das várias etapas, sem termos a pretensão de fazer um “tratado” sobre o café. Há livros especializados na área que poderão oferecer informações mais precisas sobre cada uma dessas etapas, a saber:

- a) Plantio e produção
- b) Colheita
- c) Processamento
- d) Armazenamento
- e) Beneficiamento
- f) Torrefação
- g) Bebida

O Brasil é líder mundial em produção de café por ter uma geografia favorável: está situado entre os trópicos, possui regiões altas, entre 800m e 3000m, além de temperatura inferior a 37°C e superior a 4°C em muitas micro-regiões. As duas espécies mais cultivadas no mundo são Arábica e Robusta, sendo o Brasil o primeiro produtor de Arábica e segundo produtor de Robusta, superado nesta última espécie apenas pelo Vietnã.

A espécie Arábica possui 44 cromossomos, prefere locais de grandes altitudes e clima ameno, entre 15 e 22°C. Entre suas variedades no Brasil podemos citar: Bourbon, Rubi, Acaiá, Catuaí, Icatu, Tupi, Mundo Novo, Caturra, Catucaí, etc. Já a espécie Robusta possui 22 cromossomos. Pode ser plantada entre 800m ao nível do mar e aceita temperaturas entre 24 e 29°C. Suas variedades mais conhecidas são Conillon, Bukobenses e Oka. A espécie Arábica é considerada de melhor bebida, devido a suas características palatares; já a espécie Robusta é utilizada para fazer *blends* da bebida, ou seja, misturas que normalmente possuem 70% de Arábica e 30% de Robusta.

O plantio do café ocorre após um ano de formação da muda, que provém do desenvolvimento da semente com pergaminho. Para compreendermos bem as partes do fruto, vejamos o que diz Bressani (2007) sobre sua constituição: temos casca, polpa (muito carnuda, abaixo da casca), mucilagem (camada rica em açúcares, entre a polpa e o pergaminho), pergaminho (película que envolve a semente) e semente (cada fruto tem duas; o fruto com apenas uma é chamado moca). As mudas ficam armazenadas e são cuidadas em viveiros, recebendo diversos cuidados especiais para que possam dar origem a belas plantas de café. A área para o plantio deve ser bem escolhida em relação à fertilidade do solo. Em alguns lugares do mundo, como na Etiópia, planta-se o café entre árvores, para que a sombra auxilie a planta em seu crescimento e desenvolvimento. (Figura 2) Já em países como o Brasil, a sombra não é levada em consideração, uma vez que as variedades aqui plantadas sobrevivem bem à luz direta do sol.

Figura 2 - Plantio de café entre árvores (Etiópia)

Após o plantio, o café começará a produzir dentro de dois anos, com safras bianuais, ou seja, em um ano há produção forte e, no outro, a produção é mais fraca. Para monitorar a produção bianual, assim como o cultivo da lavoura, o agricultor trabalha por talhões, isto é, partes de uma

lavouira, manejando-a e renovando-a de acordo com os registros feitos em cada talhão. O nascimento do fruto é precedido pela florada (Figura3), que começa em meados de agosto, e normalmente ocorre em três etapas sucessivas. Devido a fatores climatológicos, como secas ou o aquecimento prévio da estação, poderá haver descontinuidade ou ausência das floradas. Em outros casos, agricultores lançam mão de fito-hormônios, para que a planta produza de uma vez todos os seus frutos.

Figura 3 - Primeira florada de um pé de café

Durante a produção do café, o agricultor deve tomar diversos cuidados com a lavouira em relação a pragas e insetos, para que não atinjam seu cafezal e prejudiquem sua produção. Entre as pragas mais comuns podemos citar o bicho mineiro, a ferrugem e a broca. Já entre os insetos causadores de males para a planta está a cigarra, que come a raiz do pé de café, pois vivem na terra ao redor da planta.

A colheita emprega muitos trabalhadores rurais e muitas vezes é feita pelos chamados agricultores familiares. Em muitas áreas do Brasil, o período de safra representa grande prosperidade para a região, assim como as pequenas safras determinam um período de oscilação no consumo e na produção industrial da região afetada. Segundo o Relatório “Café: Vida,

Produção e Trabalho”, (Observatório Social, 2004, p.19), o café é o segundo produto agroindustrial que mais emprega pessoas em nosso país. Alguns problemas no setor, como o preço baixo no mercado externo e a desregulamentação dos anos 90, fizeram com que a agricultura familiar sofresse forte diminuição e erradicação de cafezais, e conseqüentemente houve a redução do número de empregos, exclusão social e êxodo rural. Os grandes complexos agroindustriais tomam medidas de contenção de gastos e redução de custo, usando recursos tecnológicos para evitar gastos com mão-de-obra, que representa entre 40 e 60% dos custos totais. Há, pois, diversas maneiras de colher o café, e todas dependem em grande parte da estrutura geográfica do terreno onde está a plantação. A colheita manual, a forma mais comum de colher os grãos, também chamada de *derriça* manual, é aquela em que o trabalhador “puxa” os grãos de cada galho de dentro para fora, de maneira que retire todos os grãos, verdes ou maduros, acabando por retirar folhas e galhos secos também. O material colhido vai diretamente para o chão, ou às vezes sobre o pano que cobre o chão. No caso de ir diretamente para o chão, fazem a *varrição*, processo de varrer os grãos com vassouras próprias, feitas com bambuíra, ou com rastelo. Após a colheita, seja colheita no pano ou no chão, o café vai para a peneira, para ser abanado e ter impurezas leves, como folhas e galhos secos, retirados, e depois vai para o saco. Os sacos comportam 60 kg, e é essa a medida usada para os alqueires apanhados, muitas vezes considerada a forma de pagamento do apanhador. Há também a colheita a dedo, às vezes chamada *seletiva* ou *catação*. Esse processo é bem mais oneroso que o primeiro, pois é preciso selecionar apenas os grãos maduros durante a coleta. O apanhador retira da árvore apenas os grãos maduros, deixando os verdes para amadurecerem e serem colhidos em um repasse posterior. O produto colhido não possui impurezas e vai diretamente para o saco. O pagamento destes dois tipos de colheita normalmente é feito por produção, pois assim o proprietário garante maior agilidade ao processo, aumentando a produtividade. Em áreas de colheita seletiva, no entanto, o valor do alqueire costuma ser superior, uma vez que poucos apanhadores aceitam trabalhar

em áreas difíceis, em que a produção será pequena. Isso depende, é claro, do poder de barganha entre o trabalhador e o proprietário. E, por fim, temos a derriça mecanizada, feita por máquinas, normalmente utilizada em grandes áreas plantadas e planas. Podemos utilizar tanto as derriçadeiras quanto as colhedeiras ou colheitadeiras. Somente grandes propriedades custeiam máquinas como essas, pois o investimento é muito alto, justificando-se apenas para grandes safras. Dessa colheita, o fruto colhido vai diretamente para a carroceria dos tratores, e não para o chão, panos ou sacos.

Após a colheita e abanação, começa o processamento do café, que deverá ser feito no mesmo dia em que é colhido. Há dois tipos de processamento: por via seca e por via úmida. Os cafés brasileiros são mais conhecidos pela via seca. Já na Colômbia, o processo por via úmida é mais comum devido ao clima úmido, que poderá fermentar o grão mais do que o necessário caso seja utilizado outro processamento. A escolha do tipo de processamento varia de acordo com as necessidades do mercado, tais como demanda do produto, preço, etc., e também da capacidade de investimento do produtor na mecanização de sua produção. No momento, o Brasil tenta, como dito anteriormente, incentivar a produção de cafés de qualidade, que são elaborados mais facilmente por via úmida. Entretanto, algumas produções por via seca, se levados em consideração diversos itens para promover a qualidade da bebida, podem ter resultados tão satisfatórios como a produção por via úmida.

Na produção por via seca, após ser abanado e peneirado, retirando assim impurezas como folhas, galhos, pequenas pedras, etc., o café será encaminhado para o terreiro. Em propriedades com investimentos em qualidade, o café vai para um lavador, cuja estrutura pode ser mecânica ou de cimento. Há uma moega que recebe o café colhido (Figura 4). Dali o café é enviado para o tanque (Figura 5), onde ocorrerá a separação de impurezas.

Figura 4 - Moega do lavador (Brasil)

Figura 5 - Tanque do lavador (Brasil)

Os grãos chamados de bóia, que são mais leves, e também aqueles que já estão secos e parecem uma passa, flutuam e são transportados para o terreiro. Já os grãos mais pesados ficam no fundo, e constituem os grãos maduros e verdes (Bressani, 2007). Assim, separados no terreiro, poderão ter sua qualidade controlada. Esse é o café natural.

Já o processamento por via úmida prevê a utilização de máquinas para a obtenção do CD - cereja descascado - ou do despolpado. Após ser lavado, o que normalmente é feito em uma máquina chamada Lavador, o café passa pelo separador de verdes e despolpador. Nesse processo o café perde a casca vermelha, ficando somente a semente e o pergaminho. Se levado para o terreiro nesse momento, será o CD. Caso a intenção seja fazer o despolpado, o processo vai além: é preciso tirar a mucilagem, grudada ao pergaminho, até que se solte, utilizando então outra máquina (desmucilagem mecânica) ou tanques de fermentação (desmucilagem manual).

Os cafés cereja descascado e despolpado são considerados cafés com maior qualidade, devido ao cauteloso processo pelo qual passam e à bebida que produzem. Entretanto, o café natural poderá ter boa qualidade se forem observadas condições de cuidado, como colhê-lo na época certa, sem a mistura de verdes ou passados do ponto; processá-lo no mesmo dia em que é colhido; além da secagem bem feita, que é nosso próximo tópico.

Figura 6 - Carriolas de transporte de café do lavador para o terreiro (Brasil)

Do lavador, o café é transportado por carrinhos, que podem ser carriolas ou carrinhos mecânicos, próprios para o transporte de grãos (Figura 6). Outra possibilidade é seu transporte por manilhas subterrâneas, direto para o terreiro (Figura 7).

Figura 7 - Saída de café direto do lavador para o terreiro

Assim como no processamento, há três possibilidades de secar o café: naturalmente, no terreiro, ou mecanicamente, com secador, ou uma combinação das duas. A primeira forma leva mais tempo, pode ter a influência do clima, além de ocasionar a fermentação imprópria devido à umidade de acordo com o micro-clima da região.

Figura 8 - Sistema de terreiros de cimento (Brasil)

Também é importante lembrar que o terreiro deve ser de cimento ou asfalto, e nunca de terra, pois o sabor pode passar para o grão (Figura 8). Mesmo os terreiros feitos de alvenaria devem ter o seu preparo bem planejado, em estrutura ideal para escoamento correto de água da chuva, sem imperfeições que formam sulcos e acumulam os próprios grãos do café ou impurezas, e onde bata sol o dia todo. Ali, o café deverá ser movimentado a cada 15 min, acompanhando a posição do sol, com pás feitas de madeira. À noite, o café deverá ser coberto para que não apanhe umidade. De preferência, colocar pano entre a lona e o café, para que, ao transpirar, o café não umedeça novamente, uma vez que o fruto possui grande quantidade de água (cerca de 85%) e se ficar abafado obviamente transpirará. O ideal é que o café vá para o secador (Figuras 9 e 10) após três dias de terreiro, pois assim atingirá o nível de umidade máximo para que seja armazenado.

Figura 9 - Secadores rotativos na Etiópia

Figura 10 - Estrutura externa de secadores à lenha (Brasil)

Em muitos lugares da Etiópia, por exemplo, o café é seco em terreiros suspensos (Figura 11), para que não tenham contato com o chão.

Figura 11 - Terreiros suspensos (Etiópia)

No Brasil também encontramos tais telas suspensas, porém não são tão comuns, apesar do baixo investimento que exigem e da melhoria que proporcionam na bebida. É uma forma de melhorar a qualidade do café e conseguir uma colocação melhor daquele produto no mercado comercial, pois muitos atestam ser esta a melhor forma de secagem. Os terreiros suspensos, ou telas suspensas, além de evitar o contato com o chão, permitem maior ventilação ao grão, favorecendo sua secagem. Da mesma forma que os terreiros convencionais, o café deverá ser movimentado de acordo com a posição do sol e coberto à noite (Figura 12).

Figura 12 - Terreiros suspensos que podem ser cobertos à noite (Etiópia)

O passo seguinte à secagem é o armazenamento ou então o beneficiamento. Como o beneficiamento vai retirar o pergaminho do café e o deixar sem proteção, o ideal é que os lotes de café sejam armazenados e só beneficiados quando forem vendidos, ou seja, quando saírem da tulha, lugar onde se armazena o café. Há muitos cuidados para tomarmos com as tulhas, principalmente em relação à umidade, a fim de evitar o branqueamento ou mofo do café. Ele pode ser armazenado a granel ou em sacos de 60 kg.

Já no beneficiamento, os defeitos do grão e as impurezas finais são eliminados. O grão é limpo, retira-se o pergaminho e em seguida ele é selecionado. Muitos proprietários usam o sistema bica corrida, que não seleciona o café por tamanho de grão. Outros utilizam o sistema de classificação por cores, que separa o tamanho do grão. Normalmente apenas as cooperativas e os grandes complexos agroindustriais possuem a máquina de classificação, devido ao seu alto custo.

A partir daí o café será transportado para as torrefações ou será vendido verde para as exportadoras de café. O grande diferencial do preço está exatamente em questões como essa: o produtor poderá processar o café para que seja de qualidade, e vendê-lo com valor agregado, ou então

como *commodity*, que é o café verde, sem preocupação com qualidade de bebida e processamento de qualidade. Para produzir café de qualidade, o proprietário deverá investir em máquinas e equipamentos, ou então na mão-de-obra, e vender o seu produto por um preço que garanta seus investimentos. O Brasil, no entanto, por ser conhecido até hoje como país de grandes quantidades, não possui liderança em mercado comercial de produto com valor agregado, ou seja, vende o café verde para países como Alemanha, Itália, Estados Unidos e Japão, que depois farão a torra e a moagem e reexportarão o produto, muitas vezes para nós mesmos, países produtores. Precisariamos de maior espaço no mercado comercial mundial para conseguir vender o nosso produto pronto, alcançando então melhores preços e favorecendo a produção de cafés de qualidade. Mas ainda falta transpor barreiras comerciais existentes, especialmente com a Europa e com os Estados Unidos, detentores das grandes multinacionais da área de café, que em sua expansão mostram-se competitivas a ponto de não deixarem espaço para a exploração do segmento. Os resultados são claros: apesar de tanta produção, de tamanha exportação, o segmento não consegue se estabelecer como lucrativo em um todo, começando pelo trabalho braçal envolvido na agricultura, até a venda da bebida, mas um ramo que explora o lado mais frágil: trabalhadores e pequenos produtores rurais. Encontramos, em alguns países, auxílio de organizações como a *Fair Trade* – Comércio Justo, que se propõe a abrir caminho aos pequenos produtores que desejam vender diretamente ao consumidor, sem passar pelas etapas do comércio atual. Porém, o trabalho desses grupos ainda é tímido diante do percentual de produtores, devido até mesmo ao pouco conhecimento que agricultores têm da existência desses.

Segundo Bressani (2007), “diferenças de altitude, solo, condições climáticas, genética do grão e cultivo, por exemplo, determinam que tipo de bebida teremos na xícara, além da porcentagem de defeitos do café e dos diferentes pontos de torra a que ele é submetido”. Para classificar a bebida, há muitas divergências entre as entidades. A Associação Brasileira da Indústria de Café (ABIC), por exemplo, classifica os cafés em três

categorias: Tradicionais, Superiores e Gourmet ou Premium. Já a Alta Mogiana *Specialty Coffees*, a *Brazilian Specialty Coffees Association* e a Conselho das Associações dos Cafeicultores do Cerrado adicionam um outro tipo, que são os Cafés Especiais, a aqueles já mencionados anteriormente, e na ordem de qualidade.

Para a última fase, a torra é que determinará a qualidade final da bebida, lembrando que as condições anteriores são de suma importância em todo o processo. Há graus de torra, definidos pela *Specialty Coffee Association of America*, e pela Agtron, empresa norte-americana, que se constituem por discos com várias tonalidades de marrom indicando os graus de torra do café (Bressani, 2007). No Brasil, a preferência fica pela torra média clara ou moderadamente escura. Mas é importante pontuar que a preferência varia de região a região e de país a país. Assim, o café torrado irá acentuar as características do grão, como corpo da bebida, acidez e aroma.

O preparo da bebida começa com a moagem do grão. O café vendido em supermercados geralmente já é moído. Nas cafeterias, porém, o ideal é que seja moído na hora, pois assim conserva todas as suas características organolépticas e aquelas relacionadas ao sabor e aroma. O modo como o café será preparado dependerá do consumidor: há café coado, café espresso, café com leite, capuccino, etc. Os baristas são os responsáveis pelo preparo do café em cafeterias. Em suas residências, no entanto, cada um poderá escolher o tipo que mais lhe convém, desde o uso de um coador até o uso de uma máquina de espresso doméstica.

E assim concluímos nossas explicações sobre o objeto de nossa pesquisa, o café. No próximo capítulo falaremos sobre a fundamentação teórica do trabalho.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo abordaremos questões teóricas nas quatro disciplinas integradas às questões que pretendemos elucidar. A LC, aplicada a estudos de fenômenos lingüísticos diversos, estará sempre interagindo com outras disciplinas. Acreditamos, também, que a interdisciplinaridade de nosso projeto esteja relacionada à temática que nos propomos a pesquisar: a necessidade de intérpretes conhecerem variantes terminológicas produzidas na oralidade. Iniciaremos com a LC, em seguida discutiremos a Interpretação e a Oralidade e, finalmente, apresentaremos a Terminologia.

2.1 Lingüística de Corpus

Toda a problemática das questões envolvendo nosso projeto está em torno da LC que, com sua metodologia, nos proverá a análise de dados para atingirmos os objetivos propostos da pesquisa. Definiremos a LC e em seguida citaremos alguns corpora falados existentes, abordando suas características mais importantes. Daremos ênfase aos estudos da LC em corpora falados, uma vez que o escopo do projeto está relacionado à oralidade. Apesar de utilizarmos corpora escritos como parâmetros para uma das características da análise, que é a comparação das realizações de variações na língua falada com as realizações na língua escrita, não são eles o foco de atenção do presente trabalho. Por fim, apresentaremos as características do programa Wordsmith Tools (Scott, 2004-2007), que utilizaremos para a análise dos dados.

2.1.1 Breve definição

Para entendermos o que é e para quê serve a LC, começaremos pela definição de Corpus:

“Na lingüística, corpus é um conjunto de textos escritos ou de falas transcritas usado como base para a análise lingüística e descritiva”³⁴ (Kennedy, 1998).

Assim, a LC utiliza os corpora eletrônicos para fins de estudos lingüísticos. Dissemos corpora eletrônicos, pois a relação entre corpora e informática é inquestionável. No passado, os estudiosos utilizavam a análise manual, desfavorecendo a credibilidade dos resultados de seus estudos, uma vez que não podiam trabalhar com uma grande quantidade de textos ao mesmo tempo. O fato de trabalharem com pequenas amostras, no entanto, não significa que o trabalho não seja confiável, mas para que certos padrões encontrados sejam entendidos e considerados como parte da língua em uso, necessitamos mais do que um texto ou uma frase ou de apenas um autor. Tudo dependerá dos objetivos a que se propõe o corpus. A informática permite que grande volume de dados seja utilizado em cada pesquisa, proporcionando a vantagem da pesquisa automática e, ainda, favorecendo a armazenagem desses corpora para futuras pesquisas. As ferramentas computacionais que, aliadas aos modelos sugeridos pela LC, servem como facilitadoras para o processo de análise dos dados, tornam o trabalho do pesquisador comprovadamente mais dinâmico e confiável.

Segundo Tognini-Bonelli (2001), a LC poderia ser definida tanto como teoria quanto como metodologia. A autora afirma que, por um lado, a ligação da LC com a informática, estatística e cálculo algorítmico faz com que seja uma metodologia. Por outro lado, seus pressupostos teóricos, que levam a uma mudança qualitativa na maneira de como entendemos a linguagem, a classificam como teoria. A autora conclui que a LC pode ser definida tanto como uma nova maneira de realizar pesquisas, quanto como uma nova

³ “In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.”

⁴ Todas as traduções feitas neste trabalho são nossas.

abordagem filosófica da pesquisa lingüística. A LC pode ser aplicada a diversas áreas que estudam a língua, tais como lexicografia, ensino de línguas, tradução, análise do discurso, gramática, lingüística computacional, terminologia, etc.

Kennedy (1998) reflete que a LC não se preocupa apenas com estudos de palavras, estruturas e usos possíveis da língua, mas também com o que é provável, ou seja, com o que pode ocorrer.

“Qualquer pesquisa científica sobre a linguagem deve ser empírica no sentido de que possa ser comprovada ou não e, na análise final, as considerações feitas sobre a língua têm de mostrar a evidência da língua em uso”⁵ (Kennedy, 1998, p.8).

O autor continua, dizendo que essas evidências da língua em uso podem ser tanto intuídas pelos falantes da língua ou então observadas em um corpus. A diferença reside na riqueza da evidência, sua validade e confiabilidade.

O trabalho com a LC para a compilação de um corpus envolve uma grande gama de atividades, tais como seleção de textos, preparação deles para a inclusão no corpus, utilização de softwares para análises dos textos, exploração da lingüística descritiva baseada em corpus para desenvolvimento de áreas como ensino de línguas e PLN - Programa de Linguagem Natural. A LC vai além de ser uma forma mais rápida de descrever o funcionamento de uma língua; também revela aspectos lingüísticos antes não imaginados.

O desenvolvimento da LC está diretamente relacionado à tecnologia. Assim, antes da era da informática já havia corpora utilizados em pesquisas, porém somente a partir de 1980 é que os corpora tornaram-se populares e começaram a exercer grande influência na pesquisa lingüística (Berber Sardinha, 2004).

Os corpora são compilados de acordo com os propósitos de cada pesquisa, variando em relação ao tamanho, design e estrutura. Alguns

⁵ Any scientific enterprise must be empirical in the sense that it has to be supported or falsified on evidence and, in the final analysis, statements made about language have to stand up to the evidence of language use.

corpora são elaborados para atenderem pesquisas mais abrangentes, outros para pesquisas mais específicas. Entendemos as pesquisas abrangentes como em torno de aspectos gramaticais, prosódia, discurso, pragmática e léxico, ou seja, da Lingüística Geral. Específicas seriam, por exemplo, a observação de diferenças em uma língua em contextos geográficos diferentes (Berber Sardinha, 2004).

O que diferencia o corpus de um banco de textos (que podem ser livros, jornais de um determinado período, etc.) é o uso que fazemos desses textos (Kennedy, 1998). Berber Sardinha (2004) afirma que a grande diferença entre corpus e arquivos de textos é que o último não possui organização prévia, enquanto o corpus possui objetivos específicos e para isso é elaborado.

Os corpora desenvolvidos para este trabalho podem ser enquadrados nos pré-requisitos orientados por Berber Sardinha (2004):

- i os textos estão em linguagem natural, quer dizer, não foram produzidos com propósitos de pesquisa e nem criados em linguagem artificial, tal como linguagens de programação de computadores;
- ii o conteúdo foi escolhido para atender aos objetivos da pesquisa em variantes terminológicas;
- iii o conteúdo é considerado representativo da variedade lingüística que pretendemos pesquisar.

Quanto à tipologia, os corpora desenvolvidos para o projeto se enquadram na seguinte categorização, proposta por Berber Sardinha (2004) e Tagnin (2005a):

Quadro 1 - Tipologia de Corpus

TIPOLOGIA	FALADO	ESCRITO
Modo	falado	escrito
Tempo	contemporâneo	contemporâneo
Seleção	amostragem; balanceado	amostragem; balanceado
Conteúdo	especializado (categorias profissionais distintas)	especializado
Tipologia textual	Entrevistas e interpretações (conversação)	Textos informativos selecionados em sites da web relacionados ao tema Café (como produto agrícola)
Língua	Monolíngüe (entrevistas) – um em português e outro em inglês Bilíngüe (interpretações) – bidirecional: port < > en	Monolíngüe – em português e inglês
Finalidade	Estudo e comparação	Estudo e comparação
Tamanho	pequeno	pequeno-médio

Quando dizemos que um corpus é representativo da característica que pretendemos analisar, entendemos que é possível encontrar tais características no corpus. No campo do léxico, por exemplo, quanto maior o número de palavras do corpus, maiores as chances de encontrarmos as palavras que ocorrem na língua, mesmo as de baixa frequência. Leech (*apud* Kennedy, 1988) afirma que a representatividade está relacionada à possibilidade de generalizar os dados encontrados no corpus como representativos da língua geral ou de uma parte específica da língua. Berber Sardinha (2004) também menciona a diferença entre corpora pequenos (menos de 80 mil palavras), para fins específicos, e a utilização de partes de grandes corpora disponíveis (acima de um milhão de palavras). Segundo o autor, os corpora pequenos serão mais representativos da variedade específica que buscamos, pois foram elaborados para esse fim. Ele sugere que os corpora com até 80 mil palavras devem ser considerados “pequenos” e de 80 a 250 mil seriam “pequeno-médios”. Conclui que um corpus representativo está relacionado ao tamanho e qualidade do corpus, além do

objetivo da pesquisa. Bowker e Pearson (2002, p.48), no entanto, chamam a atenção para que o corpus, quando considerado pequeno, tenha conceitos, termos ou padrões lingüísticos suficientes para a investigação em curso. Para validar hipóteses levantadas pelo corpus, segundo as autoras, é preciso que tenhamos mais de três ocorrências, fazendo com que as estatísticas sejam confirmadas. As autoras afirmam que, em sua experiência, corpora em torno de dez mil palavras foram suficientes para comprovar hipóteses em estudos sobre linguagens de especialidade. O número de textos também deverá ser observado: é importante que cada corpus possua textos de autores variados, para que as mesmas características estudadas possam ser observadas em autorias diversas.

Outra consideração quanto à extensão de um corpus é a demanda de tempo para a pesquisa. Os corpora falados deverão levar em consideração para o planejamento de sua execução a disponibilidade de tempo e dificuldades de transcrição decorrentes da utilização de dados orais (Shlesinger, 1998). Pesquisar terminologia requer grande número de palavras gravadas em textos de linguagens de especialidade, ou seja, requer investimento de tempo. É imprescindível que a construção do corpus seja acompanhada por seu balanceamento em vista dos objetivos da pesquisa; e que esse corpus seja alimentado por dados enquanto necessário. Um exemplo de balanceamento seria avaliar, no caso de nosso trabalho, a quantidade de informantes em cada segmento profissional, para que haja proporção nos dados de cada categoria. Portanto, para cada corpus deverá haver um balanceamento específico, dinâmico por natureza, pois os corpora poderão ser modificados se as análises realizadas indicarem essa necessidade (Kennedy, 1998).

2.1.2 Corpora falados

Devido à ênfase da língua falada em nosso estudo, relataremos, agora, a existência de alguns corpora falados que foram desenvolvidos durante as últimas décadas e alguns que ainda estão em desenvolvimento. Também citaremos projetos que envolvem a língua falada que estão sendo executados no Brasil. Em seguida elucidaremos a fundamentação teórica de compilação de corpus, enfatizando corpora falados, sem deixar de lado os escritos, possibilitando ao leitor o conhecimento de alguns termos específicos da LC, bem como a utilização de certas ferramentas para a execução da análise.

2.1.2.1 Corpora desenvolvidos ou em desenvolvimento

Kennedy (1998) relata alguns estudos feitos em LC na língua falada, enfatizando que até a metade dos anos 90 poucos estudos foram realizados na língua inglesa devido à indisponibilidade de grandes corpora falados. O maior era o London-Lund Corpus⁶, com um milhão de palavras. Os estudos de variações em Inglês tornaram-se possíveis com o Brown Corpus⁷, escrito, construído em 1964. Os estudos começaram com análises simples de diferenças entre gêneros no léxico e na gramática; em seguida outros estudos foram conduzidos, como a comparação entre as variantes entre língua falada e língua escrita, entre variantes regionais de Inglês falado e escrito, entre registros e gêneros e do processo histórico de mudança da língua inglesa.

Há alguns corpora que compreendem tanto a língua falada quanto a escrita. Outros que tratam apenas da língua falada. Destacamos a seguir, corpora que envolvem língua falada, em inglês e português do Brasil⁸:

⁶ Disponível em <http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM>. Acesso 3 mar. 2002.

⁷ Disponível em http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html. Acesso 3 mar. 2002.

⁸ Fonte: Projeto COMET, 2005.

- Collins/Cobuild – composto de 56 milhões de palavras em inglês contemporâneo, escritas e faladas. Há concordanciador online que executa buscas de concordâncias nas falas transcritas em inglês britânico⁹.
- British National Corpus (BNC) possui 100 milhões de palavras em inglês Britânico das últimas décadas do século XX. A parte oral corresponde a 10% do Corpus e inclui conversação transcrita de informantes de idades e classes sociais diferentes, bem como habitantes de regiões diferentes. Há outros textos, como reuniões de negócios ou governamentais, programas de rádio e conversações por telefone¹⁰.
- American National Corpus - versão comparável aos gêneros do BNC, ainda em construção. Seu conteúdo será de inglês americano, com língua escrita e falada¹¹.
- International Corpus of English (ICE) - possui variedades regionais e nacionais da língua inglesa, com um milhão de palavras em língua falada e escrita. Estão disponíveis em CD-ROM comercializável (Grã-Bretanha e Nova Zelândia) ou online gratuitamente (Hong Kong, África do Leste, Filipinas, Cingapura), mediante cadastro com comprovação de uso para pesquisas sem fins lucrativos¹².
- The Wellington Corpus of New Zealand English - um milhão de palavras em língua falada e um milhão em língua escrita. Na língua falada, há discurso informal (75%), semi-formal (13%) e formal (12%), em diálogos ou monólogos. Desenvolvido pela Victoria University of Wellington, também é distribuído pelo ICE, em forma de CD-ROM¹³.
- Corpus of Spoken, Professional-American English - todos os subcorpora possuem um milhão de palavras, um sobre discussões acadêmicas e o outro sobre Conferências na Casa Branca.

⁹ Disponível em <<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>>. Acesso 3 mar. 2002.

¹⁰ Disponível em <http://sara.natcorp.ox.ac.uk/lookup.html>. Acesso em 3 mar. 2002.

¹¹ Disponível em <<http://americannationalcorpus.org/>>. Acesso em 3 mar. 2002.

¹² Disponível em <<http://www.ucl.ac.uk/english-usage/ice/avail.htm>>. Acesso em 3 mar. 2002.

¹³ Disponível em <<http://www.vuw.ac.nz/lals/research/corpora/wcs.aspx>>. Acesso em 3 mar. 2002.

Disponibiliza uma versão demonstrativa online; a versão completa é comercializada no próprio site ¹⁴.

- Corpus de linguagem falada de adolescente. Faz parte do BNC, possui 500 mil palavras. Disponível em CD-ROM ¹⁵.
- ICAME - International Computer Archive of Modern and Medieval English - um projeto que reúne alguns corpora, escritos ou falados ¹⁶.
- MICASE - Michigan Corpus of Academic Spoken English. Atualmente conta com 1.848.364 palavras, de 152 transcrições. Extremamente elaborado, permite buscas de acordo com as áreas acadêmicas ou com as características sociolinguísticas dos informantes ¹⁷.
- Santa Barbara Corpus of Spoken American English parte IV - desenvolvido e distribuído pelo Linguistic Data Consortium. Contém centenas de conversas espontâneas gravadas ao redor dos EUA, de pessoas de diferentes níveis sociais, etnias, idades e profissões. Alguns inquéritos foram gravados no trabalho, em cerimônias, casamentos, lojas, salas de aula, entre outros. O lançamento foi feito em 2003 ¹⁸.
- Spoltech Brazilian Portuguese. Parte do LDC. Distribuído em DVD (\$150). Contém transcrições fonéticas e ortográficas de diversas regiões do Brasil, divididas em leituras de textos (para análise fonética) e entrevistas (fala espontânea). São 477 informantes em 8080 sentenças. O trabalho foi executado por projetos da Universidade Federal do Rio Grande do Sul e da Universidade de Caxias do Sul ¹⁹.

¹⁴ Disponível em <<http://www.athel.com/cspa.html>>. Acesso em 3 mar. 2002.

¹⁵ Disponível em <<http://torvald.aksis.uib.no/colt/>>. Acesso em 3 mar. 2002.

¹⁶ Disponível em <<http://nora.hd.uib.no/icame/newcd.htm>>. Acesso em 3 mar. 2002.

¹⁷ Disponível em <<http://micase.umd.umich.edu/cgi/m/micase/micase-idx?type=revise>>. Acesso em 3 mar. 2002.

¹⁸ Disponível em <<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S25>>. Acesso em 5 mar. 2002.

¹⁹ Disponível em <<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S16>>. Acesso em 5 mar. 2002.

- Banco de Português - possui corpora escritos e falados, porém o acesso é restrito aos pesquisadores. Em 2004 possuía 223 milhões de palavras²⁰.

Há também outros projetos desenvolvidos ou em desenvolvimento com a utilização específica de corpora falados. No entanto, nem todos utilizam a metodologia da LC para suas análises. Podemos citar:

- VARPORT - Análise Contrastiva de Variedades do Português, da Universidade Federal do Rio de Janeiro e Universidade de Lisboa, sobre variações da língua portuguesa e investigações a partir dos contrastes gramaticais entre os diferentes falantes de português²¹;
- Português Falado - Variedades Geográficas e Sociais - CLUL - Centro de Lingüística da Universidade de Lisboa, que utiliza a LC como metodologia para sua pesquisa, investigando a língua falada de todos os países falantes de português²²;
- NURC-RJ - Projeto da Norma Urbana Oral Culta - Rio de Janeiro, com o estudo da norma culta do português falado no Rio de Janeiro²³;
- Núcleo de Ensino da Língua Portuguesa Rural (primeiro nome utilizado) ou Projeto Vertentes, da Universidade Estadual de Feira de Santana – BA, sobre a língua portuguesa falada em áreas rurais no sul da Bahia²⁴;
- Filologia Bandeirante, sobre o português falado no Brasil na época das bandeiras;
- Projeto do Atlas Etnolingüístico dos Pescadores do Estado do Rio de Janeiro; etc.

²⁰ Disponível em <<http://www2.lael.pucsp.br/corpora/bp/index.htm>>. Acesso em 20 mai. 2006.

²¹ Disponível em <http://www.letras.ufrj.br/lingbr/varport.htm>. Acesso em 7 mar. 2002.

²² Disponível em <http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php#disponibiliz>. Acesso em 7 mar. 2002

²³ Disponível em <http://www.letras.ufrj.br/nurc-rj>. Acesso em 7 mar. 2002.

²⁴ Disponíveis em www.vertentes.ufba.br/chave.htm. Acesso em 7 mar. 2002.

Como o material do Projeto de Estudo da Norma Lingüística Urbana Culta (NURC) foi desenvolvido em muitos estados brasileiros, tornou-se uma pesquisa de referência nacional. Suas transcrições estão disponibilizadas para projetos de língua falada de diversas maneiras: na USP, por exemplo, disponibiliza-se em CD-ROM, na UFRJ já encontra-se online.

Apesar de haver diversos projetos com utilização específica de corpora falados, nenhum deles utiliza línguas de especialidade para a verificação de variantes terminológicas, como apresentamos neste trabalho.

2.1.2.2 Compilação de corpus falado e programas computacionais de análise

Quando falamos em modalidade falada da língua, devemos seguir as orientações de sociolingüistas, como Rodrigues (1993), de que “a oralidade é uma característica essencial da língua falada, mas não suficiente...” e cita como exemplo da diferença entre oralidade e língua falada a transmissão de notícias pela televisão, que apesar de ser oral, não é falada, pois é baseada em texto escrito para ser falado. A língua falada é espontânea, natural, não segue roteiro. Por essa definição consideraremos os corpora construídos para este projeto como “corpora falados”, uma vez que contém conversações em entrevistas e interpretações espontâneas, naturais.

Uma das primeiras preocupações com a compilação de um corpus falado é a gravação dos textos, ou seja, a coleta de dados. A coleta de dados deve ser rigorosamente planejada, para que o pesquisador não perca seu tempo dedicado à tarefa. Na publicação do projeto Filologia Bandeirante, Cintra (2000, p.163-169) enfatiza a necessidade de o pesquisador seguir caminhos que o conduzam aos objetivos da pesquisa. Para o autor, há quatro considerações básicas para a obtenção e manutenção dos dados, que foram levadas em conta para o planejamento da compilação de nossos corpora:

- a) Delimitação da população a ser analisada
- b) Natureza dos dados que se pretende obter
- c) Transcrição dos dados obtidos
- d) Arquivamento dos dados obtidos

É preciso lembrar que na coleta de dados orais, gravar textos, em aparelhos de última geração ou não, torna possível reconhecermos apenas parte da realização da fala do informante. Todos os aspectos extralingüísticos são perdidos. Utilizar gravação em vídeo ajudaria nesse quesito, embora se observe a perda parcial ou total da espontaneidade do informante seja observada. Quando as gravações são feitas em aparelho digital, compatíveis com sistemas computacionais, o áudio fica disponível permanentemente, não só para averiguação da pesquisa, como um banco de dados para outras questões que possam surgir. Atualmente é possível fazer a edição do áudio, reduzindo ruídos locais ou apurando a qualidade do som. Isso, no entanto, significa outro investimento, desta vez em softwares apropriados ou trabalho conjunto com estúdios de gravação. O software Express Scribe²⁵ tem capacidade para executar esse trabalho, além de outras tarefas.

Outro aspecto importante é a transcrição do texto: como representar, de maneira clara, ao pesquisador e leigos, o que foi dito pelos informantes, incluindo características fonéticas e prosódicas? A transcrição, segundo Cintra (2000, p.165).

“... não é uma tarefa mecânica, sujeita apenas à representação de sons por símbolos gráficos, mas um posicionamento (ideológico) teórico, que permite estabelecer que fenômenos interessem à observação e por que, quais de seus aspectos são relevantes para os fins pretendidos, e que grau de variação deles se pode ou se poderia esperar.”

Antes de iniciarmos a transcrição, portanto, devemos decidir quais dados da oralidade serão relevantes para a análise posterior e, além disso, qual o software a ser utilizado para a análise, se algum for utilizado. A

²⁵ Distribuído gratuitamente pela NCH Swift Sound (2007).

representação ortográfica convencional é a mais utilizada para que não-especialistas em fonética possam também utilizar os dados transcritos. Mas como saber a entonação, as ênfases, as pausas formulaicas, entre outras características, nas realizações da fala do informante? Até certo nível, a transcrição consegue reproduzir essa realização do informante. Contudo, é certo que haverá algumas perdas. A solução para o quê transcrever dependerá dos objetivos a que se destina o corpus. Se pretendermos fazer uma averiguação de dados fonéticos, Sinclair (*apud* Thompson, 2005) afirma que poucos dados são suficientes; as transcrições, porém, deverão conter *links* para a gravação. Já a análise de frequência de palavras ou padrões léxicos da língua poderá ser feita em corpora transcritos em ortografia convencional, o que requer maior quantidade de textos. Quanto à análise de discurso, deveremos incluir no corpus todas as características contextuais possíveis. O documentador, portanto, deverá planejar cautelosamente todas as inclusões que deverá fazer, objetivando sua pesquisa. Alguns programas computacionais já disponibilizam recursos para as três necessidades acima apontada, como o Codes for Human Analysis of Transcripts (CHAT) ⁻²⁶ (MacWhinney, 2003a), por exemplo. Nesse software, integrante do projeto Child Language Data Exchange System (CHILDES)²⁷, podemos incluir as gravações em áudio ou vídeo, transcrição fonética, codificação detalhada da fala – incluindo recursos prosódicos, bem como informações a respeito do contexto de realização da fala. Apesar de o projeto ser destinado à pesquisa da fala infantil, qualquer tipo de transcrição pode ser feita com os programas desenvolvidos, infantil ou não. Após a transcrição, o software possui interface com o Computerized Language Analysis (CLAN)²⁸ (MacWhinney, 2003b), ferramenta que executa as análises (frequência, concordância, etc.), prevendo a codificação feita pelo CHAT. Entretanto, para dados em grande escala, torna-se uma tarefa demorada e árdua.

Segundo Leech, Myers & Thomas (*apud* Thompson, 2005), dos cinco estágios envolvidos na compilação de um corpus falado, que são: a)

²⁶ Símbolos para Análise de Transcrição da Fala.

²⁷ Sistema de Troca de Dados na Linguagem Infantil.

²⁸ Análise Lingüística Computadorizada.

gravação dos dados; b) transcrição; c) codificação; d) etiquetagem e e) aplicações, é preciso ficar atento para a diferença entre transcrição e codificação. Podemos verificar as palavras de MacWhinney (2003b, p.12):

“É importante reconhecer a diferença entre transcrição e codificação. A transcrição está focada na produção de um registro escrito que pode nos fazer compreender... o fluxo de interação original. (...) A codificação, por outro lado, é o processo de reconhecimento, análise e anotação de fenômenos percebidos em transcrições. A codificação pode ser feita por meio de transcrições apenas”²⁹.

A questão da autorização para utilização dos dados gravados é de extrema importância, para que os dados possam ser utilizados na íntegra e que outras pesquisas possam ser realizadas com os mesmos dados. A British Association for Applied Linguistics³⁰ (Thompson, 2005) recomenda que a autorização de dados orais seja gravada nos mesmos meios da coleta de dados (por vídeo ou áudio) ou na forma escrita.

A codificação dos dados deverá seguir os padrões dos softwares que serão utilizados. Entretanto, devemos sempre arquivar uma cópia da transcrição “crua”, quer dizer, a transcrição feita ainda sem objetivos de leitura computacional, mas sim leitura humana. Nossa sugestão é que os corpora transcritos sejam primeiramente salvos em arquivos Word, atualmente um processador de textos muito comum e de fácil utilização/interface. Outra sugestão é manter os originais gravados, preferencialmente, em arquivos digitais. Atualmente contamos com softwares capazes de transformar tapes em som digital, tais como Express Scribe³¹, já citado anteriormente, Sound Forge 9.0³², Cool Edit 2000³³ ou Audacity³⁴, além de gravadores digitais, MP3, MP4, etc. Tão logo esteja pronta a transcrição, os textos já poderão fazer parte dos respectivos corpora planejados.

²⁹ “It is important to recognize the difference between transcription and coding. Transcription focuses on the production of a written record that can lead us to understand... the flow of the original interaction. (...) Coding, on the other hand, is the process of recognizing, analyzing, and taking note of phenomena in transcribed speech. Coding can often be done by referring only to a written transcript.”

³⁰ Associação Britânica de Linguística Aplicada.

³¹ NCH Swift Sound, 2007.

³² Produzido pela Sony, 2007.

³³ Produzido pela Adobe, 2007.

³⁴ Creative Commons, 2007.

Há no mercado alguns softwares para a transcrição dos textos orais a partir de sua digitalização. É importante observar, porém, que os softwares não são capazes de identificar vozes diferentes, uma vez que trabalham com um sistema de treino para o reconhecimento de voz de apenas um falante.

O próximo passo da compilação é etiquetar o corpus, quer dizer, inserir automaticamente, semi - automaticamente ou manualmente, nos textos, códigos que indicam aspectos necessários à análise da pesquisa, como a classe de palavras, por exemplo. Segundo Teixeira [*no prelo*], “a informação acrescentada é em geral codificada de modo a permitir sua fácil identificação, seja pelo pesquisador, seja pelos programas computacionais empregados na análise.”

Algumas reflexões deverão ser consideradas antes de sua execução (Thompson, 2005):

- a) As informações são fáceis de extrair?
- b) As etiquetas são distintas entre si?
- c) Será fácil removê-las quando for necessário?
- d) Os detalhes sobre os informantes, sobre o contexto, enfim, sobre as condições de realização da fala, estão disponíveis?

A etiquetagem pode feita de diversos tipos, de acordo com os objetivos da pesquisa. Um dos tipos mais comuns (Berber Sardinha, 2004) é a etiquetagem morfossintática, que indica a classe gramatical das palavras. Mais uma vez, é importante manter o texto original em um arquivo separado, pois muitos etiquetadores não possuem a capacidade de retornar uma ação, ou seja, uma vez colocada a etiqueta automática, não há mais como voltar atrás.

Há no mercado programas etiquetadores, como o CLAWS, criado pela Universidade de Lancaster³⁵, utilizado para a etiquetagem do BNC e disponibilizado online. Para o português encontramos o Tree-Tagger na página <http://www2.lael.pucsp.br/corpora/etiquetagem/index.html>, também

³⁵ Disponível em <<http://www.comp.lancs.ac.uk/ucrel/claws/>>. Acesso em 5 mai. 2003.

disponível online, e etiquetadores automáticos do Projeto Lácio-Web, disponível em www.nilc.icmc.usp.br/lacioweb. Alguns programas, tais como WordSmith Tools (Scott, 2004-2007), são capazes de identificar as etiquetas para o procedimento da análise.

Muitas considerações deverão ser feitas antes da utilização de etiquetadores automáticos em corpora orais. Como os dados são transcritos conforme a fala poderá haver muitas variações e possíveis falhas na etiquetagem morfosintática, por exemplo, caso o programa não reconheça a palavra. Para tanto é preciso resolver anteriormente se o corpus deverá ser corrigido momentaneamente ou definitivamente, ou ainda quais as alternativas disponíveis para cada corpus. Em nossos corpora, diversas considerações foram feitas antes de pensarmos nas melhores etiquetas. Devido ao objetivo estar voltado à busca de variantes, desconsideramos a possibilidade de utilizarmos etiquetas morfosintáticas, uma vez que não poderíamos “corrigir” o corpus (Leech, Garside e Bryant, 1994). Etiquetamos, assim, manualmente, o início da fala de cada informante, bem como informações extra-textuais, como risos, comentários, etc.

As aplicações dos grandes projetos de corpora compilados estarão sujeitas aos seus objetivos e também à disponibilidade de direitos autorais. Se consultarmos os corpora anteriormente citados, tais como o MICASE, veremos as abrangentes possibilidades de aplicação, como a possibilidade de utilizar o corpus para projetos individuais, concordanciadores *online*, listas de frequências, etc. Os corpora específicos, compilados para pequenos projetos, também possuem aplicação sujeita aos objetivos e recursos tecnológicos da pesquisa. Há variados programas computacionais no mercado que podem ser utilizados para pesquisas, também de pequeno porte, em LC, como o CLAN (MacWhinney, 2003b) e o WordSmith Tools (Scott, 2004-2007), por exemplo.

Selecionamos o WordSmith Tools (WST) (Scott, 2004-2007), versão 4.0, devido aos seus recursos úteis e boa interface com o usuário. O programa trabalha com três ferramentas (Figura13) para o processo de

identificação de candidatos a termo, a saber: WordList³⁶, KeyWords³⁷ e Concord³⁸. É importante enfatizar que utilizamos os nomes dessas ferramentas em inglês, como citamos acima. Já os seus produtos, listas de palavras, palavras-chave e concordâncias, estão sempre em português.

Figura 13 - Tela inicial do WordSmith Tools

Para que um estudo de corpus seja feito utilizando o WST, é preciso ter um corpus de estudo e um corpus de referência. O corpus de estudo pode ser confeccionado pelo pesquisador, voltado ao tema da pesquisa, ou compor-se de corpus já existente, desde que atenda ao objetivo da pesquisa. O corpus de referência é usado como contraste ao corpus de estudo. Se o objetivo da pesquisa é uma determinada área terminológica, por exemplo, o corpus de referência deverá conter textos que não sejam da área objeto de estudo, pois assim o contraste entre o conteúdo dos dois corpora ficará evidente. Pode-se utilizar um corpus de

³⁶ Lista de palavras.

³⁷ Palavras-chave.

³⁸ Concordanciador.

língua geral como corpus de referência em casos de estudo de terminologia, por exemplo. O importante é que possua características diferentes do corpus de estudo.

Figura 14 - Tela da WordList – WordSmith Tools - versão 4.0

WordList (Figura14), é a ferramenta utilizada para a geração de listas de palavras que aparecem no corpus. Todas as palavras que constarem no corpus aparecerão com o uso dessa ferramenta. Há três listas possíveis, geradas pela ferramenta:

a) Lista de ordem alfabética (Figura15): as palavras são listadas de A-Z, com o número de sua freqüência no corpus na frente. Essa lista pode ser utilizada em reverso, ou seja, partir de seus sufixos, para encontrar o que se pretende analisar;

N	Word	Freq	%	Texto	% Lemmas	Sel
1	#	274	0,66	19	100,00	
2	A	1.141	2,74	19	100,00	
3	A	39	0,09	10	52,63	
4	ABABA	2		2	10,53	
5	ABANA	1		1	5,26	
6	ABANA	4		3	15,79	
7	ABANAÇÃO	4		1	5,26	
8	ABANAM	2		1	5,26	
9	ABANANDO	1		1	5,26	
10	ABANAR	3		2	10,53	
11	ABANDONO	1		1	5,26	
12	ABASTECER	1		1	5,26	
13	ABELHINHA	1		1	5,26	
14	ABENÇOOU	1		1	5,26	
15	ABERTO	2		2	10,53	
16	ABRAÇANDO	1		1	5,26	
17	ABRIL	1		1	5,26	
18	ABRINDO	1		1	5,26	
19	ABRIR	2		2	10,53	
20	ABRIU	1		1	5,26	
21	ABSORVE	2		1	5,26	
22	ABSORVIDO	3		2	10,53	
23	ABSORVIDOS	2		1	5,26	
24	ACABA	6	0,01	4	21,05	
25	ACABAM	3		2	10,53	
26	ACABAMOS	1		1	5,26	
27	ACABEI	1		1	5,26	
28	ACABO	2		2	10,53	
29	ACABÔ	1		1	5,26	
30	ACABOU	5	0,01	3	15,79	
31	ACAIA	2		2	10,53	

Figura 15 - Lista em ordem alfabética (WordList)

N	Word	Freq	%	Texto	% Lemmas	Sel
1	QUE	1.620	3,91	19	100,00	
2	O	1.379	3,31	19	100,00	
3	Ê	1.311	3,15	19	100,00	
4	A	1.141	2,74	19	100,00	
5	DE	1.138	2,73	19	100,00	
6	E	968	2,32	19	100,00	
7	CAFE	903	2,17	19	100,00	
8	NÃO	703	1,69	19	100,00	
9	EU	544	1,31	19	100,00	
10	DIO	519	1,25	18	94,74	
11	ELE	471	1,13	17	89,47	
12	UM	453	1,09	18	94,74	
13	NO	424	1,02	17	89,47	
14	TEM	415	1,00	19	100,00	
15	UMA	407	0,98	16	84,21	
16	PARA	364	0,87	13	68,42	
17	ENTÃO	353	0,85	17	89,47	
18	NA	353	0,85	19	100,00	
19	NE	352	0,85	15	78,95	
20	VOCÊ	341	0,82	19	100,00	
21	DA	327	0,79	19	100,00	
22	VAI	321	0,77	17	89,47	
23	MAIS	316	0,76	19	100,00	
24	COM	282	0,68	18	94,74	
25	POR	281	0,67	18	94,74	
26	#	274	0,66	19	100,00	
27	EM	271	0,65	19	100,00	
28	SE	258	0,62	16	84,21	
29	PORQUE	225	0,54	18	94,74	
30	ESSE	217	0,52	16	84,21	
31	ISSO	212	0,51	18	94,74	

Figura 16 - Lista em ordem de frequência (WordList)

b) lista de ordem de frequência (Figura16): as palavras são listadas da mais freqüente para a menos freqüente;

c) lista de informações estatísticas (Figura17): número de *types*, *tokens*, *type token ratio*, *standardised type-token*³⁹, número de parágrafos do texto, número de palavras com *n* letras, sendo:

Types: itens lexicais distintos que ocorrem no corpus;

Tokens: número total de palavras que ocorrem no corpus;

Type-Token ratio: riqueza lexical, ou seja, se a razão for alta, significa que o texto é variado, com poucas repetições.

Standardised Type-Token ratio: cálculo padronizado da riqueza lexical do texto. Essa estatística compara os trechos dos corpora a cada mil *tokens*, ou seja, se houver muitas repetições (gerando baixo valor de variedade lexical) em determinados trechos de um texto, há possibilidade de compará-los com trechos onde não há muitas repetições.

	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14
file size	250,400	24,000	19,775	23,194	40,340	19,209	6,949	12,957	20,091	15,669	10,057	14,274	2,439	1,767	1,954
tokens (running words) in text	41,639	3,908	3,424	3,622	6,795	3,364	1,112	2,000	4,735	2,760	1,796	2,480	575	244	209
tokens used for word list	41,364	3,698	3,367	3,699	6,767	3,357	1,103	1,991	4,725	2,747	1,773	2,471	366	237	203
types (distinct words)	3,848	943	751	770	1,074	757	332	516	1,007	705	539	641	190	119	107
type/token ratio (TTR)	9	24	22	21	16	23	30	26	21	26	30	26	52	50	53
standardised TTR	33,10	34,30	32,57	31,80	31,79	32,43	30,80	32,00	35,33	35,55	35,30	34,00			
standardised TTR std dev	63,96	50,00	51,30	52,32	96,97	61,00			61,71	45,57		46,57			
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00
mean word length (in characters)	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
word length std dev	2,52	2,57	2,44	2,53	2,40	2,69	2,32	2,62	2,61	2,77	2,45	3,39	2,33	2,38	2,39
sentences	3,021,00	309,00	290,00	404,00	658,00	134,00	125,00	122,00	272,00	118,00	152,00	41,00	22,00	18,00	11,00
mean (in words)	14	12	15	9	10	25	9	16	17	23	12	60	17	13	16
std dev	18,00	10,68	10,66	8,40	8,55	22,67	8,70	28,03	17,73	17,66	11,32	86,52	10,84	10,21	14,10
paragraphs	74,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	30,00	20,00	18,00	1,00
mean (in words)	659	3,898	3,397	3,699	6,767	3,357	1,103	1,991	4,725	2,747	1,773	124	18	13	203
std dev	1,304,36											538,41	68,67	45,85	
headings															
mean (in words)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
std dev															
sections	19,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	2,177	3,898	3,397	3,699	6,767	3,357	1,103	1,991	4,725	2,747	1,773	2,471	366	237	203
std dev	1,788,85														
numbers removed	274,00	10,00	27,00	23,00	18,00	7,00	9,00	9,00	10,00	13,00	25,00	9,00	9,00	7,00	6,00
stoplist tokens removed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
stoplist types removed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1-letter words	4,995	603	390	468	762	421	132	220	500	290	212	279	43	36	17
2-letter words	6,705	543	598	571	1,070	490	176	344	721	587	354	422	59	35	44
3-letter words	7,576	610	592	614	1,174	600	225	387	906	514	352	505	60	42	39
4-letter words	6,203	530	550	548	1,131	466	170	286	620	318	270	403	63	48	35

Figura 17 - Lista estatística (WordList)

³⁹ Devido ao seu uso consagrado, utilizaremos os termos em língua inglesa. Porém suas traduções são, respectivamente: formas, ocorrências, razão entre formas e ocorrências e razão entre formas e ocorrências padronizada.

É possível utilizar uma lista chamada *stop list* no programa WST, para que as palavras que não interessam à análise do estudo não entrem nas listas acima. Isso será mais bem explicado no capítulo dedicado à Metodologia (Capítulo 3).

As listas acima citadas não permitem que o programa identifique as diferenças entre as classes gramaticais ou significados de palavras homônimas, tal como *play*, que pode ser um verbo ou um substantivo e significar brincar, jogar ou mesmo brincadeira, jogo. Portanto, as listas de frequência nos informam apenas as palavras fora do contexto. Desta forma não podemos inferir os significados ou classes gramaticais se não observarmos as concordâncias para podermos, aí sim, proceder a uma análise eficaz.

Figura 18 - Tela de KeyWords – WordSmith Tools 4.0

KeyWords (Figura 18) é a ferramenta que compara uma lista de frequência do corpus de estudo com outra do corpus de referência, gerando lista de palavras que são consideradas “chave”, cujas frequências são estatisticamente diferentes nos corpora (Berber Sardinha, 2004). As

palavras-chave são então dispostas de acordo com sua “chavicidade”, ou seja, de acordo com as freqüências relevantes do corpus de estudo comparado ao corpus de referência. Como dito acima, o corpus de estudo é aquele feito para fins específicos, enquanto o corpus de referência é o parâmetro para a análise, ou seja, fornece uma norma para a comparação. Uma sugestão de Berber Sardinha (2004, p.102), é que ele deve ser maior que o corpus de estudo, pois:

... o tamanho do corpus de referência influencia a quantidade de palavras-chave obtidas. Os tamanhos críticos de corpora de referência são 2, 3 e 5 vezes o tamanho do corpus de estudo. Corpora de referência com essas dimensões retornam significativamente mais palavras-chave do que corpora de tamanhos menores...

A lista de palavras-chave obtidas nos oferece uma primeira visão dos candidatos a termo, pois sua ocorrência no corpus de estudo é alta. Essa seleção deverá ser observada em contexto e, finalmente, ser ou não considerada válida para o vocabulário (Figura 19).

N	Key word	Freq	%	Freq	RC %	Keywords	Planned	Ext
1	CAFE	1.003	1,98	0	100,00	000000		
2	VAI	416	0,82	13	612,09	000000		
3	GENTE	218	0,43	1	940,60	000000		
4	ESTA	172	0,34	0	541,02	000000		
5	TENHO	141	0,28	0	263,19	000000		
6	PODE	136	0,27	0	218,39	000000		
7	COLHEITA	128	0,25	0	146,70	000000		
8	CEREJA	104	0,20	0	931,64	000000		
9	VOÇES	102	0,20	1	902,40	000000		
10	VERDE	98	0,19	0	877,88	000000		
11	ANO	123	0,24	63	865,18	000000		
12	MÁQUINA	95	0,19	0	851,00	000000		
13	VOU	99	0,20	5	846,85	000000		
14	ÁGUA	103	0,20	14	837,30	000000		
15	GRÃO	93	0,18	0	833,08	000000		
16	FAZ	106	0,21	45	766,62	000000		
17	HOJE	105	0,21	42	742,39	000000		
18	FICA	94	0,19	25	720,28	000000		
19	FAZER	92	0,18	36	672,85	000000		
20	COISA	71	0,14	1	625,46	000000		
21	TER	107	0,21	166	595,70	000000		
22	PRODUTOR	70	0,14	5	580,39	000000		
23	VINTE	74	0,15	13	589,77	000000		
24	TERREIRO	65	0,13	0	582,23	000000		
25	DÁ	107	0,21	190	574,63	000000		
26	SECADOR	63	0,12	0	564,31	000000		
27	VAMOS	68	0,13	7	562,73	000000		
28	OLHA	62	0,12	0	555,35	000000		
29	CHÃO	62	0,12	0	555,35	000000		
30	ONDE	61	0,12	0	546,39	000000		
31	BEBIDA	65	0,13	5	546,32	000000		

Figura 19 - Lista de palavras-chave (KeyWords) - Por ordem de freqüência

Figura 20 - Tela do Concord - WordSmith Tools - v.4.0

Line	Text	Word	Count	%	File	%
1	informal cut which one to have (...) a tree which looks exactly like the source	tree	1,795	69.8%	0.2%	0.2%
2	should always try to keep the coffee tree not too high otherwise they would	tree	4,450	410.6%	0.7%	0.7%
3	This is about thirty years old, the coffee tree. Thirty. Sorry? This tree is about	tree	2,989	227.0%	0.2%	0.2%
4	coffee. It takes four years for the coffee tree to grow to its full size. In its 9th	tree	3,151	207.0%	0.0%	0.0%
5	over ripen now. Ah, ha (...) the coffee tree with the still have the coffee on the	tree	311	24.3%	0.7%	0.7%
6	No. No? ... a small tree... the coffee tree so, but I grow from the real tree	tree	1,079	84.3%	0.6%	0.6%
7	By the way selecting for the desirable tree, you want to be maintained in the	tree	2,155	73.0%	0.7%	0.7%
8	coffee tree so, but I grow from the real tree. Ah, right. (...) Ela ela cr	tree	1,087	84.0%	0.6%	0.6%
9	one... each one is gonna pick one. One tree? Yeah. Oh. Oh I see. (...) It	tree	1,109	108.0%	0.2%	0.2%
10	how many kilograms of fertilizer per tree every year? Quantos quilos de	tree	1,576	130.6%	0.8%	0.8%
11	it, and also establish into a productive tree, so this end up trade work including	tree	1,702	68.1%	0.8%	0.8%
12	so you don't know... (...) is a small tree... Ah, a árvore pequena... como é	tree	1,037	78.8%	0.5%	0.5%
13	ally coffee? No. No? ... a small tree... the coffee tree so, but I grow from	tree	1,077	84.1%	0.6%	0.6%
14	We cannot do anything to protect the tree. Quando acontece queda, eles não	tree	3,326	292.0%	0.9%	0.9%
15	After they pick the coffee on the tree, then they separate the others, ok?	tree	4,006	371.0%	0.6%	0.6%
16	advice farmers how to put up the tree, so that to not have too much	tree	949	64.0%	0.9%	0.9%
17	when the cherries being from the tree, which is picked when it's red, then	tree	271	19.0%	0.3%	0.3%
18	some (...) still have the coffee on the tree. Can we have the coffee on the tree	tree	345	26.0%	0.8%	0.8%
19	the production decreases, because the tree cannot (...). (...) Yeah. It's about	tree	744	69.2%	0.7%	0.7%
20	cut it. They cut it. How far from the tree? É qual é a distância... Cê viu o que	tree	4,398	407.0%	0.6%	0.6%
21	get that, and that will be part of the tree and, very long one, they can grow it,	tree	1,687	68.5%	0.8%	0.8%
22	they needed to cut the upper part of the tree and it's going to raise again. Oh,	tree	3,933	358.0%	0.8%	0.8%
23	the tree. Can we have the coffee on the tree but it is all ripe. Éh, tem algum lugar	tree	353	27.4%	0.8%	0.8%
24	tree with the still have the coffee on the tree? Excuse me. Now can you ask your	tree	320	24.0%	0.8%	0.8%
25	tree. Thirty centimeters from the tree. Oh ok. So when it gets up too long	tree	4,698	423.0%	0.9%	0.9%
26	not. You can see (...) The coffee on the tree now they all... cherry that... all very	tree	269	19.8%	0.6%	0.6%
27	it's a way of getting more from the tree. Oh, ok. And they I asked him, my	tree	5,267	547.0%	0.0%	0.0%
28	you keep doing it so you can see the tree is marked in many places. (...) but I	tree	2,642	233.3%	0.3%	0.3%
29	pouca em baixo. Ah... he said that this tree for example is producing too much	tree	713	66.9%	0.1%	0.1%
30	old, the coffee tree. Thirty. Sorry? This tree is about thirty years old. Ah...	tree	2,999	230.6%	0.2%	0.2%

Figura 21 - Tela de lista de concordâncias (Concord)

Concord (Figuras 20 e 21) é a ferramenta que fornece a palavra, ou palavras, em contexto. A concordância mais comum é a *Key Word in Context* (KWIC), com a palavra de busca no centro do texto, em diversas linhas de concordância. Podemos pedir a concordância das palavras-chave encontradas com a KeyWords, para analisar o contexto em que ocorrem. Essa ferramenta também produz listas de colocados (Figura 22).

The screenshot shows the IMFL software window with a menu bar (File, Edit, View, Compute, Settings, Windows, Help) and a toolbar. The main window displays a KWIC concordance list for the word 'THE'. The list consists of 12 rows, each representing a different context where 'THE' appears. The columns are: N (line number), Word (the word being searched), With (the word following 'THE'), relation (the relationship between 'THE' and 'With'), Total (total frequency), and various positional counts (L5, L4, L3, L2, L1, Centre, R1, R2, R3, R4, R5). The word 'THE' is highlighted in blue in the first row. The word 'COFFEE' is highlighted in red in the third row. The word 'ON' is highlighted in red in the twelfth row.

N	Word	With	relation	Total	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	TREE	0,000	36	30	6	3	5	0	7	15	0	1	1	0
2	TREE	TREE	0,000	33	1	2	0	0	1	0	0	0	1	1	0
3	COFFEE	TREE	0,000	12	10	2	0	0	5	0	0	1	0	0	1
4	TO	TREE	0,000	9	4	5	0	2	0	0	0	1	0	3	0
5	IS	TREE	0,000	7	1	6	0	0	1	0	0	2	1	2	0
6	A	TREE	0,000	7	5	2	1	0	0	3	1	0	0	1	0
7	HAVE	TREE	0,000	7	4	3	3	0	0	1	0	0	0	1	1
8	SO	TREE	0,000	6	1	5	1	0	0	0	0	0	3	0	1
9	FROM	TREE	0,000	6	5	1	0	0	1	4	0	0	0	0	1
10	THEY	TREE	0,000	5	0	5	0	0	0	0	0	0	2	0	2
11	IT	TREE	0,000	5	0	5	0	0	0	0	0	0	2	1	0
12	ON	TREE	0,000	5	5	0	0	0	0	5	0	0	0	0	0

Figura 22 - Lista de colocados (Concord)

2.1.2.3 Extração terminológica e produção de vocabulários com a LC

Após a compilação do corpus partiremos para a extração terminológica *per se*. Como dito anteriormente, utilizam-se ferramentas computacionais para a extração de termos, porém é necessário trabalhar com a análise humana após todo o processo. A geração de termos oriunda da extração feita pelo WST, por exemplo, apresenta-nos “candidatos” a termos, ou seja, é preciso haver análise humana do produto para que tais candidatos configurem-se como termos. Verificaremos os procedimentos da extração terminológica no Capítulo 2 – Metodologia.

2.2 Interpretação e Oralidade

Tradução e interpretação são atividades inseridas no mesmo escopo, porém divergem no modo de entrega, ou seja, a interpretação terá sempre a língua oral como produto. Assim, a maneira de atuar do tradutor e do intérprete é diferente, embora ambas as atividades “traduzam”.

A própria definição da palavra Interpretação une os dois conceitos. Assim reza o verbete “interpretar” no Dicionário Aurélio - Século XXI⁴⁰:

[Do lat. *interpretare, por interpretari.]

V. t. d.

1. Ajuizar a intenção, o sentido de
2. Explicar, explanar ou aclarar o sentido de (palavra, texto, lei, etc.)
3. Tirar de (sonho, visão, etc.) indução ou presságio
- 4. Traduzir ou verter de língua estrangeira ou antiga.**
5. Representar (no teatro, cinema, televisão, etc.)

V. t. d.

6. Inform. Executar (programa), convertendo suas instruções em código de máquina ao longo do processo. [Cf., nesta acepção compilar. Nesta acepção, der. de interpretador.]

V. transobj.

7. Julgar, considerar, reputar

⁴⁰ **Dicionário Aurélio - Século XXI.** Versão 3.0. Editora Nova Fronteira. 1 CD-ROM.

Apesar de a Interpretação ainda não ter atingido o patamar da Tradução no meio acadêmico encontramos, cada vez mais, pesquisas relacionadas ao tema. A perspectiva histórica das pesquisas é importante para podermos situar o que foi feito até agora, compreendendo melhor a profissão e suas exigências, mas também ter em mira o futuro.

É impossível afirmar se Tradução e Interpretação possuem mais semelhanças ou diferenças; mas é possível afirmar que, sob o paradigma da Tradução, as pesquisas em Interpretação podem contribuir de muitas maneiras para entendermos melhor ambas as artes.

A relevância da Interpretação para o desenvolvimento desta dissertação está na necessidade do intérprete em saber quais as variantes existentes nas áreas e línguas em que atua. Explanaremos, portanto, a atividade do intérprete, em suas modalidades e possíveis áreas de atuação, a fim de esclarecer pontos de intersecção com outras áreas exploradas por esta pesquisa de mestrado. Em seguida traçaremos um panorama teórico da Interpretação, delineando as pesquisas em LC aplicadas à Interpretação. Para finalizar, faremos um paralelo entre Interpretação/Oralidade e Tradução/Escrita.

2.2.1 A arte de interpretar

A interpretação envolve mais que a definição básica de tradução oral imediata. A interpretação envolve a comunicação (Jones, 1998). Apesar do desenvolvimento de teorizações sobre questões direcionadas somente para a Interpretação, Toury (1995) reconhece a Interpretação como extensão natural dos estudos descritivos da Tradução.

Faremos a seguir uma breve descrição dos modos de interpretar e das possíveis áreas de interpretação.

Há dois modos de interpretação: consecutiva e simultânea.

Interpretação Consecutiva é o modo de Interpretação realizado em pausas do falante/palestrante. Normalmente o intérprete utiliza anotações para conseguir produzir o discurso na língua de chegada. Envolve grande capacidade de análise e de memória.

Interpretação Simultânea é o modo de Interpretação realizado simultaneamente à fala do palestrante/falante. Normalmente o intérprete utiliza uma cabine com equipamento adequado para que os ouvintes possam receber a interpretação em fones de ouvido. Envolve agilidade e atenção dividida, pois ao mesmo tempo em que ouve, o intérprete monitora sua fala.

O trabalho do intérprete consecutivo e simultâneo está vinculado a sua capacidade de ouvir, compreender a mensagem, analisá-la e então reexpressá-la. Tratando-se de um ato comunicativo, normalmente a interpretação palavra por palavra impossibilita que o receptor compreenda a mensagem. Portanto, a fase de análise é valiosa para o intérprete. As técnicas utilizadas para um modo e outro diferem muito, pois na consecutiva o intérprete tem a chance de ouvir o período de fala completo, ou seja, se não entender uma palavra terá a chance de recuperar seu sentido no decorrer do discurso. Já na simultânea tal aspecto não existe: enquanto o discurso original é proferido, o intérprete deve transferi-lo para outra língua ao mesmo tempo.

Pagura (2003) sugere um terceiro tipo de interpretação: intermitente. Segundo o autor, seria um modo utilizado por falantes de língua estrangeira e não intérpretes, especialmente em reuniões de negócios, bastando o conhecimento e a fluência em língua estrangeira. Nesse modo, o intérprete fica entre dois falantes ou entre falante e público, traduzindo oralmente as frases ditas, normalmente curtas, para que a comunicação entre eles se estabeleça. Pode ser feita em reunião de negócios, acompanhamento de estrangeiros em turismo ou negócios, bem como em palestras. Acreditamos, no entanto, que a situação de amadorismo exista em todas as profissões relacionadas às línguas estrangeiras. Há professores não graduados em Língua Estrangeira, assim como há tradutores sem formação. Não cabe a nós julgar o comportamento do mercado e sua influência na teoria, tampouco é o objetivo do trabalho que se apresenta. As técnicas utilizadas pelo intérprete nesse modo são as mesmas da consecutiva: ouvir a mensagem, compreendê-la, analisá-la e

reexpressá-la. A capacidade de memória envolvida é, sim, variável. Entretanto, não há um critério estabelecido em relação ao período de tempo que o intérprete deverá armazenar dados em sua memória. Por outro lado, o intérprete da intermitente, quando funcionário bilíngüe de uma empresa, por exemplo, que não conhece as técnicas de interpretação, costuma interferir no conteúdo do texto oral. Muitas vezes percebemos que o intérprete vai além do texto a ser traduzido, ou então o adapta da maneira que mais convém, para que atinja seus objetivos de negócios. Dessa forma, podemos concluir que a intermitente é um submodo da consecutiva, ou seja, não possui características distintas suficientes que a diferenciem totalmente da consecutiva. O mesmo ocorre com a *Sight Translation*, ou Tradução à Prima Vista, e a *chuchotage*, ou cochicho. A *Sight Translation* consiste em um texto escrito que será interpretado oralmente para os ouvintes. Em nossa concepção, esse é um submodo da Interpretação Simultânea, uma vez que o intérprete fará a leitura oral em outra língua simultaneamente à leitura escrita na língua original. Já a *chuchotage* é uma interpretação simultânea, porém em voz baixa, para apenas um ouvinte, que ocorre normalmente em uma situação em que somente uma pessoa necessita da interpretação.

Dentre as muitas áreas em que o intérprete pode trabalhar estão as de conferências e seminários, judicial, de mídia, médica, comunitária e de negócios. Em todas elas, o intérprete poderá utilizar tanto a interpretação simultânea como a consecutiva. Atualmente, o modo mais utilizado é o simultâneo devido a diversos fatores, facilidade dos recursos disponíveis e possibilidade de trabalhar diversas línguas ao mesmo tempo.

Outro componente de importante elucidação é o tratamento dado às línguas de trabalho. De acordo com a Association of Conference Interpreters (AIIC, 1982), há três tipos de línguas de trabalho: língua A, B e C. Língua A é aquela de proficiência nativa; língua B é quase-nativa, quer dizer, a proficiência é quase a de um nativo e o intérprete pode fazer seu trabalho dela ou para ela; língua C é considerada passiva, ou seja, compreendida como nativa, porém o intérprete não traduz para essa.

É também importante considerarmos o tipo de conhecimento necessário para atuar como intérprete. Segundo Gile (1995, p.131-132a),

...intérpretes e tradutores de textos não-literários raramente possuem conhecimento de base com todas as informações para realizar seu trabalho, tornando a Aquisição de Conhecimento uma fase importante e regular para os procedimentos tradutórios⁴¹.

Gile segue citando quais são os passos que o tradutor e o intérprete devem seguir para adquirir o conhecimento necessário para textos escritos ou falados não-literários. Temos grande interesse no aspecto terminológico, uma vez que o assunto escolhido para a pesquisa é de uma área não-literária, a saber, a cafeicultura. O autor cita informações lingüísticas ou não, comuns aos dois gêneros da língua, falada e escrita: informação terminológica, conhecimento do estilo do texto e informação extralingüística. Sobre a informação terminológica, diz Gile (1995a) que o uso terminológico correto na tradução/interpretação pode ser considerado o elemento essencial de qualidade aos olhos dos usuários desses trabalhos. Discorrendo sobre os passos para preparar-se para uma interpretação, ele cita a importância de glossários para intérpretes e os coloca como menos confiáveis do que glossários feitos por terminógrafos ou tradutores. Como geralmente não há tempo para preparação de glossários completos, o intérprete faz apenas uma listagem terminológica apropriada para determinado evento. Gile refere-se a glossários de tradutores como mais confiáveis devido à questão de tempo para preparação. Acreditamos, porém, que essa questão deva ser alvo de maiores pesquisas e debates. Atualmente, com a informática nos auxiliando a superar as dificuldades de tempo, o intérprete poderia desenvolver um glossário tão rico e preciso quanto outros voltados à tradução escrita. A necessidade de precisão é tão importante em interpretação quanto em tradução. Além disso, intérpretes contam com mais um entrave em seu trabalho. Devido a variações que ocorrem com os termos em nível profissional e geográfico, nem sempre

⁴¹ "... the Knowledge Base of interpreters and translators of non-literary texts seldom contains all the information necessary for them to perform their work, so that Knowledge Acquisition is a regular and important part of their Translation operations."

registradas em livros, dicionários, glossários, vocabulários ou buscadores online (ex. Google), pesquisar essas variações pode se tornar uma tarefa árdua, porém necessária. Muitas variações ocorrem apenas na fala, daí a necessidade de pesquisas para elaboração de glossários que contenham variantes da língua falada.

2.2.2 Panorama teórico da interpretação

Apresentando agora um histórico das teorias mais importantes da Interpretação, sabemos que as pesquisas em Interpretação são muito recentes. Até mesmo a prática da Interpretação, comparada à da Tradução, só possui registros históricos a partir da Renascença (Delisle e Woodsworth, 1998). Há muitos fatores que contribuem para tal descaso com a área. Um deles é o fato de a Interpretação ser uma prática oral e, portanto, não necessitava de registros escritos. Assim, mesmo que trabalhos anteriores houvessem sido realizados, não haveria como nos reportarmos a eles.

Seguindo esta linha de comparação temporal, um breve histórico da pesquisa em Interpretação mostra-nos que os únicos trabalhos considerados válidos pela comunidade científica foram desenvolvidos por intérpretes ativos ou professores de tradução e interpretação. Mais uma vez, a Tradução tem um número maior de acadêmicos interessados na pesquisa, o que nos leva a questionar qual a verdadeira razão para isso. Há muitas considerações, entre as quais podemos citar a preferência dos intérpretes pelo mercado em vez da pesquisa acadêmica, fundamentada principalmente por motivos financeiros; a pesquisa em Tradução já possui um (pequeno) lugar nas Instituições de ensino, o que não é o caso da Interpretação; e até mesmo a falta de união entre pesquisadores de áreas científicas ligadas à Interpretação que, por não serem intérpretes, são muitas vezes criticados por seus trabalhos não terem aplicação prática.

Segundo Moser-Mercer (1991), os primeiros estudos válidos em Interpretação na comunidade científica foram realizados por cientistas cognitivos e behavioristas. Tais estudos eram desenvolvidos com intérpretes

amadores, fora do contexto real de trabalho, e foram amplamente criticados por toda a comunidade de intérpretes. Por essa razão, consideraremos apenas dois grupos de pesquisadores em Interpretação: a Comunidade de Artes Liberais, representada por Seleskovitch e Lederer, e a Comunidade de Cientistas Naturais, mais recente, com grandes nomes como Gile, Lambert, Pöchhacker, entre outros.

Seleskovitch e Lederer (1984) criaram a conhecida Teoria Interpretativa da Tradução, uma hipótese, como veremos abaixo, que nunca foi empiricamente verificada. Logo depois, essa teoria foi substituída por uma metodologia mais pedagógica, favorecendo a interdisciplinaridade. Segundo Shlesinger (1998), o maior problema da interdisciplinaridade é a ausência de um paradigma único para a Interpretação. Cada disciplina envolvida nas pesquisas possui seus próprios questionamentos, e não os faz de maneira integrada com as outras. O que é visto por psicólogos é bem diferente da preocupação dos lingüistas, por exemplo. Já na Tradução, não há tanta preocupação com a interdisciplinaridade, uma vez que o foco da atenção tem sido maior no produto do que no processo (Gile, 1995b). Essa segunda teoria acima mencionada busca explicações cognitivas para o processamento das informações, o que é natural devido às operações cerebrais envolvidas. O mais conhecido representante dessa fase contemporânea é Daniel Gile. Sua maior preocupação não é formular teorias que pareçam óbvias por sua prática, mas testar empiricamente as idéias que propõe.

Vejamos a seguir um breve sumário das principais idéias contidas na Teoria Interpretativa da Tradução, de Marianne Lederer, e em seguida no Modelo de Esforços, de Daniel Gile, dois dos principais representantes dos dois grupos de pesquisadores em Interpretação.

2.2.2.1 A Teoria interpretativa da tradução ou modelo interpretativo

A Teoria Interpretativa da Tradução foi elaborada por Danica Seleskovitch e sua equipe na Escola Superior de Intérpretes e Tradutores da Universidade Paris III – Sorbonne Nouvelle. A publicação, no entanto, deu-se em 1994, por Marianne Lederer, na obra denominada *La Traduction Aujourd'hui: le modèle interprétatif*.

Essa Teoria é baseada em três estágios: compreensão do discurso oral de partida; desverbalização de sua forma lingüística mantendo o que foi dito em idéias e fatos; e expressão dessas idéias e fatos em outra língua. Um de seus pontos positivos é que essa Teoria serviu e ainda serve como ferramenta para aprendizes a fim de que não fixem sua tradução em palavras, mas no sentido geral do texto.

Embora baseada na Interpretação de Conferências, a Teoria Interpretativa da Tradução faz menção à aplicabilidade dessa Teoria à Tradução (Lederer 1994). Segundo a autora, tanto a Interpretação como a Tradução podem ser realizadas por meio das três etapas da Teoria. Alguns requisitos, entretanto, são fundamentais para que as etapas possam ser aplicadas, como o conhecimento lingüístico e extralingüístico, as habilidades escritas (para o tradutor) e o conhecimento da língua de chegada, bem como do assunto em questão.

O desenvolvimento da Teoria Interpretativa da Tradução foi baseado em análises procedentes de observações feitas durante interpretações de conferência, quando intérpretes iam além do léxico ou da gramática do palestrante. Somente o significado interessava, como se o intérprete possuísse a mesma postura que o público-alvo da palestra, embora mais concentrado e menos envolvido com o assunto, por não ter de analisá-lo. Na Interpretação Consecutiva esse mesmo fenômeno pode ser observado: palavras ditas perdem sua importância em função do que significam, o jogo tradutório vai muito além da literalidade, buscando a equivalência contextual. (Schjoldager, 1995).

Assim, a crítica feita à Teoria Interpretativa da Tradução é exatamente a falta de pesquisa empírica. Seleskovitch e Lederer (1984) basearam sua hipótese em observações de seu próprio trabalho como intérpretes e

professoras de Interpretação e Tradução e em pesquisas de teóricos como Jean Piaget, Jacques Barbizet e Eugenio Coseriu. O problema vai além: essa Teoria afirma que a Interpretação é independente da língua, ou seja, não propõe dificuldades lingüísticas para a Interpretação. O fato de basearem suas pesquisas em seus próprios trabalhos não pode ser visto isoladamente como negativo, mas assim seria se houvesse falta de controle de observação e de sistematicidade.

2.2.2.2 Modelo de esforço

Representante da nova geração de pesquisadores na área da Interpretação, Gile (1995b) foi o primeiro a utilizar a capacidade de processamento na Interpretação como ferramenta conceitual em pesquisas na área. Essa capacidade refere-se à atenção dividida, ou seja, como o intérprete distribui sua atenção no decorrer de uma tarefa. Atribuiu-lhe esse nome por as tarefas interpretativas exigirem operações exaustivas e conscientes, como ouvir e analisar o texto de partida, produzir um texto de chegada e memorizá-lo para então entregá-lo, tudo isso em um curtíssimo espaço de tempo. Cada etapa é considerada um “esforço”. Se o intérprete não possui os requisitos para cada capacidade, quer dizer, se sua atenção for direcionada mais para um esforço do que outro, a atividade em conjunto poderá ser comprometida.

Esse modelo também pode explicar outros fatos relevantes ao processamento de tarefas na Interpretação, tais como os erros cometidos sem motivo aparente, que podem estar relacionados: a) às deficiências na administração de uma das etapas da capacidade de processamento; b) à razão de que certos pares de línguas apresentam mais dificuldades do que outros por possuírem estruturas lingüísticas diferentes; c) à interpretação da língua materna para a estrangeira.

Esse Modelo, entretanto, não discute as limitações da capacidade de processamento, sendo que tal estudo é sugerido por Shlesinger (1998), o qual apresentaremos a seguir.

2.2.3 Pesquisas em LC

Notamos que grande número de pesquisas sobre interpretação sob a ótica da LC são realizadas para o modo simultâneo. Entretanto, para o nosso trabalho, o material coletado foi originado pelo modo consecutivo/intermitente, que poderá ser utilizado por outras pesquisas para investigações específicas da área de Interpretação.

Apesar de não termos contemplado um número exaustivo de exemplos, temos abaixo dois exemplos selecionados por sua interdisciplinaridade com a LC.

Shlesinger (1998) observou as dificuldades encontradas para aplicar a Lingüística de Corpus aos estudos em Interpretação Simultânea. Segundo ela, a maioria dos estudos está baseada em dados coletados de interpretações ouvidas pelos pesquisadores, o que não deveria ser considerado sem as devidas gravações, no caso da Simultânea, tanto da fala-fonte quanto da fala do intérprete. Ainda há aqueles que utilizam recursos online para gravar sua própria interpretação, como o software Black Box Interpreting, da Melissi Multimedia, que permite a gravação de sua interpretação e posteriormente faz uma avaliação dela. Outros estudos são feitos com alunos ou estagiários de interpretação, dificultando a obtenção de resultados que possam ser generalizados como atuação do intérprete, uma vez que não possuem o mesmo desempenho de um intérprete profissional. Esse último, por sua vez, quando autoriza o estudo, geralmente não pode autorizar a publicação dos textos, pois não é o produtor deles e dessa maneira não detém os direitos autorais do conteúdo interpretado. A autora também cita o uso de corpora bilíngües e paralelos para as pesquisas em Interpretação, direcionadas para estudos que demonstrem as relações entre interpretação e outras formas de discurso oral e outras formas de tradução. Sugere então a descrição de qual seria a construção ideal de corpora para as pesquisas, a saber: corpora comparáveis, paralelos (ambos em três textos: interpretado, original e a tradução dos mesmos) e monolíngüe. A conclusão do estudo é de que a Interpretação já alcançou a posição de disciplina e, dessa maneira, necessita do apoio metodológico e técnico da LC.

Outro estudo baseado em LC foi feito por Wallmach (2002), que elaborou um corpus com as interpretações simultâneas feitas no Plenário Jurídico de Gaunteg, do inglês para o Africâner, o Zulu e o Sepedi. O corpus paralelo consistia da língua utilizada no local e das interpretações feitas por dois profissionais para cada combinação durante três Sessões Plenárias. Como o estudo ainda não havia terminado na ocasião da publicação do artigo, havia apenas quatro horas de gravações, com Africâner e Zulu. As gravações foram feitas em canal duplo. O objetivo do estudo é analisar a influência da técnica e do ritmo no desempenho do intérprete. Também pretende determinar qual a relação existente entre a expectativa do público e a produção do intérprete. A ferramenta utilizada para análise é o ParaConc, de Michael Barlow⁴². Os resultados encontrados no estudo piloto demonstram que, dentre outras estratégias utilizadas pelos intérpretes, estão a criação de palavras e uso de neologismos. Também observa que, de acordo com o aumento da complexidade e do ritmo da fala original, todos os intérpretes reduziram o tempo de espera entre as falas, bem como utilizaram estratégias de antecipação e de reestruturação de frases. Sob pressão, os intérpretes de Zulu utilizavam empréstimos do inglês.

2.2.4 Contrastes entre fala e escrita: a questão tradutória

Sabemos que a língua falada só pode ser assim considerada quando suas condições de realização respeitam a interação centrada entre os falantes, diferentemente dos textos em que a realização oral é em forma de leitura em voz alta de texto escrito. A interação centrada significa que os falantes possuem algo em comum sobre o que conversar (Marcuschi, 2003). O autor afirma que a conversação deve possuir cinco características básicas: interação entre pelo menos dois falantes; ocorrência de pelo menos uma troca de falantes; presença de uma seqüência de ações coordenadas; execução numa identidade temporal e envolvimento numa interação

⁴² Disponível em <<http://www.athel.com/para.html>> Acesso em 13 mai. 2006.

centrada. Nas entrevistas e interpretações coletadas para este trabalho, os falantes se envolvem, interagem, não se trata de uma estrutura monológica (Urbano, 1988). O processo de interpretação não é conversação *per se*, mas o texto falado interpretado sim, por tratar-se de uma consecutiva onde há sempre dois falantes de línguas distintas.

Apesar de entre as características da língua falada citarmos a informalidade, os corpora de entrevistas estão em um grau intermediário de formalidade. Alguns falantes, devido a características pessoais, tais como personalidade, sentem-se logo à vontade e produzem textos informais, como diálogos comuns, do dia-a-dia. Outros, talvez pela presença do gravador, ou por saberem que participam de uma entrevista gravada e têm “medo de errar”, já não são tão informais. Assim, a documentadora faz diversas tentativas de quebrar o formalismo, a fim de facilitar a naturalidade da fala e a interação com o informante. Pudemos perceber que os informantes falam mais que a documentadora, resultando em um intercâmbio esporádico. O que, porém, ocorre com frequência é a proximidade entre os informantes e a documentadora, uma vez que, já conhecida por eles, teve o acesso e as condições de gravação facilitadas.

Em um texto falado, podemos observar como o falante vai construindo seu texto, passo a passo, levando em conta o seu ouvinte, incluindo dados de acordo com a reação desse. Já no texto escrito não conseguimos perceber essas marcas abstratas de planejamento; o texto escrito é mais elaborado; seu planejamento ocorre concretamente, sem a intervenção do ouvinte (Rodrigues, 1993).

De acordo com Urbano (2002)⁴³, a língua oral possui características intrínsecas e extrínsecas por um lado; gerais e específicas por outro. As características **intrínsecas** referem-se à natureza sonora da língua oral, e não visual, além das condições de realização. Podemos também considerar aqui os recursos prosódicos utilizados para a realização dos textos orais. As palavras possuem sentidos além do lexical, contam com a entonação, com a

⁴³ O autor pretende rever as denominações utilizadas para a classificação. (fonte: comunicação pessoal).

elipse, utilizam recursos paralingüísticos. Há, nessa produção oral, uma limitação de desempenho do falante ou ouvinte quando o enunciado é muito elaborado. Já as características **extrínsecas** são as que ocorrem na fala e também na escrita, diferenciando-as apenas a intensidade e a frequência com que ocorrem. A língua oral, por exemplo, é mais concreta do que a escrita. Quando algo é dito, parece ser mais bem compreendido do que quando é escrito. As características **gerais** podem ocorrer em qualquer língua, tanto falada quanto escrita. Podemos dizer que a língua oral, mais informal e natural, possui conteúdos cognitivos simples, pouco elaborados, determinados pela naturalidade do cotidiano e destinados à praticidade e imediatismo da situação. Além do mais, ocorrem espontânea e simultaneamente à expressão do pensamento (Carvalho, 1970; Bally, 1951). As características **específicas** referem-se apenas ao país onde o assunto é tratado. Em nosso caso, as características específicas ficarão por conta do português falado no Brasil e do inglês falado nos países de onde vêm os falantes. Não podemos deixar de mencionar os fatores situacionais concretos (falante, ouvinte, assunto) envolvidos, que geram diversos matizes lingüísticos e graus de formalidade/informalidade. A língua oral adapta-se à situação, levando em conta a presença ou imagem do interlocutor. A intersubjetividade produz esse caráter social da língua.

Citaremos a seguir algumas diferenças também citadas por Urbano (2002) entre oralidade e escrita:

- a) Entonação: se não há codificação exata da entonação na escrita, ocorrem perdas ou confusões. Já a língua oral conta com a expressão concreta da entonação. Para que um texto em língua falada seja transcrito para fins terminológicos, a necessidade de representação da entonação não é relevante. Enquanto procede-se com a extração terminológica, saber qual foi a entonação não acrescentará dados aos resultados.
- b) Abstração: na oralidade, o discurso direciona-se para o assunto, o objetivo da fala. O dinamismo e a vulnerabilidade das situações é que direcionam a fala. Na escrita há abstração, pois não há necessidades imediatas a suprir, intelectualizando a expressão lingüística.
- c) Interação: ouvinte e falante interagem na produção oral, provocando mudanças de atitudes de acordo com a necessidade.

- d) Analitismo: a linguagem oral é mais analítica em algumas produções, tais como vou estudar em vez de estudarei.
- e) Sinteticismo: usado na língua oral em produções de conceitos abstratos, fundindo em um só pensamento os elementos objetivos e subjetivos. Exemplo: “Pobre homem!” (oral), “Eu lamento este homem porque ele é infeliz!” (escrita).
- f) Ênfase: a linguagem oral é predominantemente mais enfática que a escrita, mostrando o comportamento exagerado do falante.
- g) Elipse: na língua escrita, a omissão de enunciados leva à incompreensão do contexto. Já na oralidade há a possibilidade de entendimento, pois os meios paralingüísticos, o próprio contexto e a situação contribuem para a compreensão.
- h) Produção lingüística: na linguagem falada há economia de procedimentos lingüísticos como concordância, regência e a ordem, devido ao seu caráter funcional de expressão. Há extrema dependência entre escolha lingüística e situação de comunicação.
- i) Redundância e repetição: característica típica da linguagem oral, em razão da intenção de necessidade de reforço de uma idéia, de quaisquer reações psicológicas (nervosismo, por exemplo), da confirmação de questionamentos, da insistência, da ameaça, etc.
- j). Dinamismo: muitas palavras são usadas de forma múltipla e até incoerente na língua falada, pois pressupõe-se que exista pouca consciência de expressão devido ao dinamismo da oralidade, ou seja, “o imediatismo do coloquial dificulta a reflexão e o artifício”.

Já dissemos anteriormente que a grande diferença entre Tradução e Interpretação está na maneira como são feitas, ou seja, o produto de uma é escrito e da outra é oral. A intersecção entre Tradução e Interpretação pode facilmente ser observada na Interpretação de Conferências. Nesse discurso vemos como a manifestação concreta de uma fala ou texto passa pela mente do intérprete (e, por conseguinte, do tradutor), transformando-se em outra manifestação concreta, com a transferência de significado. Lederer (1994) vai além, afirmando que é o intérprete quem sabe como as pessoas pensam na língua de partida, pois as palavras estão vivas, não há que descobrir a intenção do autor, ela está ali. Na Interpretação, o público-alvo está presente e o intérprete pode captar as condições de produção do discurso oral, bem como sua recepção.

O Modelo dos Esforços em Interpretação, citado anteriormente, demonstra que os requisitos para a capacidade de processamento em

Tradução são poucos em relação à Interpretação, pois não há Esforços competindo entre si. Toda a atenção estará disponível alternadamente para a leitura analítica e para a escrita. A questão do tempo envolvido é quase nula na Tradução; na Interpretação ela é fundamental; o profissional utiliza os três esforços (ouvir/analisar, produzir e memorizar) em um curto espaço de tempo. Além do mais, o tradutor pode voltar ao texto original ou refazer o texto traduzido quantas vezes seja conveniente, o que é impossível com o discurso oral. Essa diferença de processamento de tarefas é crucial para entendermos que mesmo que um tradutor possua o domínio de ambas as línguas de chegada e de partida, isso não é suficiente para uma boa atuação como intérprete, pois importa saber trabalhar com a divisão de atenção envolvida.

Se na Tradução e produção da língua escrita em geral podemos consultar livros, internet, dicionários e afins, na Interpretação o máximo que fazemos é consultar o colega de cabine ou especialista presente; às vezes um dicionário ou glossário. A Tradução permite a revisão do texto, quando reavaliamos as decisões tomadas durante o ato tradutório. Já a Interpretação requer decisões imediatas e definitivas, exigindo grande conhecimento lingüístico e extralingüístico para prevenir “surpresas” durante o ato tradutório. Portanto, respondendo *grosso modo* aos questionamentos levantados acima, as grandes diferenças entre Tradução e Interpretação residem e se apóiam nas diferenças lingüísticas entre fala e escrita, podendo ambas serem vistas como exemplos dos dois modos de comunicação e usuárias de suas estratégias.

2.3 Terminologia e Variantes - Uma Leitura Social

Não há como negar a importância vital da Terminologia em um panorama de desenvolvimento mundial. Para que produtos sejam fabricados e vendidos internacionalmente, ou que conceitos sobre funcionamento tecnológico sejam divulgados, é preciso que exista uma mínima compreensão de significado entre as partes envolvidas. Assim, faz-se necessário estabelecer noções lingüísticas acerca de termos utilizados em áreas técnicas e científicas, para que a comunicação seja eficaz. Fazer compreender o outro, em seu domínio socioprofissional, é também tarefa da Socioterminologia, um ramo que desponta nos estudos terminológicos para debater questões de norma, uso e aceitação. O escopo deste subcapítulo é apresentar a Socioterminologia como ramo da Terminologia, explicitando a análise de variantes terminológicas e alguns conceitos básicos em Terminologia, bem como traçar um panorama de pesquisas terminológicas desenvolvidas no âmbito da LC. Os conceitos terminológicos básicos e a visão socioterminológica sobre variantes são utilizados posteriormente, na elaboração do Vocabulário Técnico em língua falada, apresentado no Capítulo 4.

Segundo Sager (1990) terminologia é “o estudo, e a área de atuação relacionada, da coleção, descrição, processamento e apresentação de termos, como, por exemplo, itens lexicais que pertencem a áreas de especialidade em uma ou mais línguas”⁴⁴. Portanto, a Terminologia estuda as linguagens de especialidade, que são um “sistema de comunicação oral ou escrita usado por uma comunidade de especialistas de uma área particular do conhecimento” (Pavel e Nolet, 2002, p. 124 *apud* Barros, 2004, p. 42).

Já o termo, unidade-padrão da Terminologia é definido como “designação, por meio de uma unidade lingüística, de um conceito, definida em uma língua de especialidade” (ISO 1087 *apud* Barros, 2004, p. 40).

⁴⁴ The study of, and the field of activity concerned with, the collection, description, processing, and presentation of terms, i.e., lexical items belonging to specialized areas of usage of one or more languages.”

Essas três definições, no entanto, não encerram as diferentes visões e amplas discussões existentes nesse campo teórico. A Terminologia existe desde o século XVIII (Cabré, 1993, *apud* Almeida, 2003), porém apenas no século XX surge a primeira teoria terminológica. As teorias da Terminologia foram elaboradas por pesquisadores interessados em favorecer a comunicação das ciências no plano internacional, ao começarem a desenvolver reflexões sobre o termo (Krieger e Finatto, 2004).

Fundador da Escola de Viena, Eugene Wüster estudou princípios normativos que deram origem à Teoria Geral da Terminologia (TGT), em 1931. A grande preocupação da TGT era com a normalização terminológica, ou seja, “padronizar os léxicos especializados para favorecer a eficácia das comunicações científicas e técnicas em plano internacional” (Krieger e Finatto, 2004).

Muito criticada, a TGT vê a Terminologia como prescritiva e como normativa. “A Terminologia normativa (...) ocupa-se da uniformização de conceitos e da atribuição de termos para designá-los. Tenta eliminar ou reduzir a sinonímia e a homonímia” (Barros, 2004, p. 56).

Segundo Wüster (1998, p.23),

... en terminologia, con la tremenda productividad de conceptos y términos, una evolución libre de la lengua llevaría a una confusión intolerable. Para evitar esta situación, desde principios del siglo, e incluso antes en el caso de algunas especialidades, los terminólogos empezaron a unificar, por consenso, algunos conceptos y términos. Esta unificación también se denomina norma y, para distinguirla de la norma descriptiva, se usará la expresión norma prescriptiva.”

Destarte, a terminologia pôde, com esses estudos, desenvolver outras reflexões acerca de termos e conteúdos, tratando principalmente das limitações impostas pela TGT. A autora Maria Teresa Cabré (1999) faz críticas à TGT por ser uma teoria que não dá conta de todos os aspectos envolvidos na Terminologia. Cabré (1999) afirma que em contextos prescritivos, fortemente estruturados e, a fim de garantir a padronização comunicativa, o uso da TGT é inquestionável. Já em situações de comunicação natural, de base social, torna-se insuficiente. Por isso a autora

foi buscar novos fundamentos que originassem uma nova teoria de termos baseada na linguagem e em seu caráter sociocultural. Surge, assim, a Teoria Comunicativa da Terminologia (TCT), que dá conta dos termos como unidades lexicais, ou seja, não há distinção entre termo e palavra, admitindo variações e levando em conta sua dimensão textual e discursiva.

Almeida (2003) expõe os pilares teóricos que servem de base para essa Teoria, com base nas definições de Maria Teresa Cabré. Um deles trata-se da variação terminológica, nosso enfoque neste trabalho. Segundo Cabré (1998 *apud* Almeida, 2003, p. 217), “a variação é entendida como fenômeno natural, inerente à linguagem. Nesse sentido, a variação deve ser considerada e deve estar adequadamente descrita na terminologia, e não eliminada nem reduzida a fim de favorecer uma comunicação especializada precisa e unívoca, sem ambigüidades”.

Essa perspectiva teórica, que em nosso trabalho nos permite descrever a diversidade da linguagem de comunicação especializada na subárea cafeeira, comporta a Socioterminologia. Inspirada na sociolinguística e na teoria da análise do discurso político, a Socioterminologia, que ainda não possui pilares para sustentar uma teoria, propõe novas bases comunicativas.

Considerada por Barros (2004, p. 68) como um dos ramos da Terminologia, seus fundamentos são considerar o uso dos termos e situar a comunicação especializada em seu lugar social. Para Gaudin (*apud* Barros, 2004, p. 69) a terminologia é analisada pela Socioterminologia como “práticas linguísticas e sociais concretas dos homens que a empregam”. Gaudin (1993) afirma que:

“... a socioterminologia, com o suposto de que deseja ultrapassar os limites de uma terminologia de “escrivão”, deve localizar a gênese dos termos, sua recepção, sua aceitação, mas também as causas do insucesso e do sucesso, no âmbito das práticas linguísticas e sociais concretas dos homens que empregam tais termos” (Gaudin, 1993 *apud* Faulstich, 2006, pág. 216).

Faulstich (1995a) defende a idéia de que a Terminologia é passível de variação porque faz parte da língua, porque é heterogênea por natureza e porque é de uso social. Portanto, um mesmo termo poderá ser usado em

contextos diferentes, bem como termos variantes poderão ser usados num mesmo contexto. É impossível determinar, por normalização, que um determinado termo X seja o correto em uma linguagem de especialidade, enquanto o termo Y deva ser deixado de lado.

Em Furlani e Silva (2003), as variantes são consideradas erros; variantes essas que “contaminam” a utilização da terminologia correta. Os autores gostariam que a normalização na área de mecanização agrícola na cultura do café fosse difundida, a fim de que não houvesse mais o uso errôneo dos termos “... às vezes por falta de conhecimento ou por regionalismo” (Furlani e Silva, 2003). Urbano (1978) explica que o desvio da norma é considerado, por alguns, como incorreção. Entretanto, é importante esclarecer que neste trabalho, devido a implicações socioprofissionais, tais como o contexto social, não haverá incorreção, mas sim variação.

A questão que se coloca, na Socioterminologia, é que o uso de variantes terminológicas pode provir de fontes discursivas variadas. Se considerarmos que os termos provêm do discurso científico, técnico e de vulgarização, não podemos normalizar um termo levando em conta apenas o discurso científico. Tampouco podemos pensar que os usuários das linguagens de especialidade sejam parte de um todo homogêneo. Segundo Faulstich (2006), a diversidade de uma língua pode ocorrer nos planos vertical, horizontal e temporal da língua.

Assim, a autora afirma que a Socioterminologia favorece a abordagem do termo na comunicação científica e técnica, levando em consideração, numa mesma área de conhecimento, os diferentes níveis de comunicação que dependem das circunstâncias de emissão e das características dos informantes. Classificou as variantes em três categorias: concorrentes, coocorrentes e competitivas, que serão utilizadas para a análise a ser feita no Capítulo 4:

Quadro 2 - Classificação de variantes terminológicas

Concorrentes	Coocorrentes	Competitivas
1. Lingüísticas => Fonológicas => Lexicais => Morfológicas => Sintáticas => Gráficas	Sinonímia	Empréstimos vs.: => formas vernaculares => termos híbridos => termos decalcados => estrangeirismos
2. Registro => Geográficas => Discursiva => Temporal		

A) Variantes Concorrentes

1a) Variantes terminológicas lingüísticas – a variação é determinada pelo fenômeno lingüístico.

- Fonológica - a escrita surge de formas decalcadas da fala.
- Lexical - há comutação no item lexical, sem, no entanto, interferir no significado terminológico.
- Morfológica - a ordem morfológica apresenta alteração, sem comprometer o significado.
- Sintática - duas construções sintagmáticas que funcionam como predicação das unidades terminológicas se alternam.
- Gráfica - variação entre oral e escrito.

2a) Variantes terminológicas de registro – a variação é determinada pelo ambiente de ocorrência, situação e no plano temporal em que se realizam os usos lingüísticos.

- Geográfica – toda língua é diversificada tanto pela origem geográfica quanto pela origem social dos locutores.
- De discurso – toda língua é estilisticamente diversificada; os locutores vão modificando sua maneira de falar de acordo com as situações sociais em que se encontram.

- Temporal – toda língua é historicamente diversificada e, dada a mudança lingüística, um estado de língua no tempo 1 é diferente de um estado de língua no tempo 2.

B) Variantes coocorrentes – há duas ou mais denominações para um mesmo referente, com equivalência de conteúdo.

a) Sinonímia terminológica

C) Variantes competitivas – relacionam significados entre itens lexicais de línguas diferentes. São os empréstimos vs. formas vernaculares, termos decalcados, termos híbridos ou outros estrangeirismos.

Finatto (2001) observa que a ocorrência de variantes em línguas de especialidade compartilhadas por países falantes do mesmo idioma é perfeitamente previsível. Segundo Kachru e Nelson (1996), as variações em inglês sob o ponto de vista geográfico existem, pois com a expansão do uso e do ensino de inglês ao redor do mundo, línguas como as asiáticas e africanas, por exemplo, influenciam o uso e escolhas da língua inglesa como língua internacional. Apesar de o inglês falado ser o mesmo basicamente, há muitas inovações lexicais, gramaticais, de pronúncia, discurso etc., devido às influências culturais e de línguas de povos falantes de inglês como língua internacional.

Utilizaremos, portanto, o conceito de termo como entidade variante, pois faz parte de diferentes situações comunicativas e se molda de acordo com a dinâmica das línguas (Faulstich, 2006).

2.3.1 Terminologia e interpretação

Sabemos que a interação entre Terminologia e Tradução/Interpretação está cada dia mais evidente no mercado de trabalho, devido às tarefas consideradas técnicas. As linguagens de especialidade, ou tecnoletos, constituem-se fundamentalmente de termos próprios da comunicação profissional. Assim sendo, quando traduz um texto de especialidade, o tradutor ou intérprete deve dominar a área de especialidade em questão. Segundo Krieger e Finatto (2004, p.66):

A primeira motivação do encontro que direciona a Tradução para a Terminologia relaciona-se ao fato de que os termos técnico-científicos são elementos-chave, nódulos cognitivos, dos textos especializados. É em relação a essa tipologia textual que se efetua a chamada tradução técnica ou especializada.

Assim, os termos dessas linguagens serão os elementos lingüísticos que permitirão a comunicação entre os profissionais. Tradutores e intérpretes devem conhecer os termos tanto na língua original quanto na língua de trabalho, tanto quanto as características de funcionamento textual, da área técnica e do gênero técnico na língua de chegada (Krieger e Finatto, 2004, p. 67). Podem utilizar, para adquirir esse conhecimento, glossários, vocabulários ou dicionários especializados. Esse material pode ou não estar disponível, embora normalmente o tradutor e o intérprete construam o seu próprio material.

Daqui por diante trataremos apenas da tarefa do Intérprete como construtor da Terminologia, devido aos objetivos de nosso trabalho. Para preparar-se para uma Interpretação em área técnica, o intérprete deverá, em princípio, além dos conhecimentos lingüísticos, pré-requisitos para qualquer interpretação, adquirir o conhecimento sobre a área em que atuará. Isso pressupõe pesquisas terminológicas em textos da área de conhecimento, glossários, vocabulários e dicionários especializados. Muitas vezes, o intérprete parte para a pesquisa em campo, conversando com especialistas da área. Embora grande parte das informações necessárias seja encontrada na forma escrita, não poderíamos deixar de questionar o que acontecerá na oralidade, considerada mais informal, durante o trabalho de Interpretação.

Conforme vimos nos princípios da TCT, a sinonímia é quantitativamente dependente do nível de especialização de um texto. Portanto, resta saber se os textos orais interpretados em conversações são tão especializados quanto os escritos, e qual seu grau de variação.

2.3.2 Terminologia e lingüística de corpus

Muitos projetos têm sido realizados em conjunto entre Terminologia e Lingüística de Corpus. Para citar alguns, há projetos abrangentes, tais como o COMET⁴⁵, o TextQuim (<http://www6.ufrgs.br/textquim/index.php>), e o e-Termos (www.nilc.icmc.usp.br/etermos), ainda em fase de elaboração.

Dentro do Projeto COMET, há pesquisas em Terminologia (Josimeire Martins; Guilherme Fromm; Luciana Latarini Ginezi; Luciana Carvalho Fonseca; Elisa Duarte Teixeira, Rosa Castanho) prática de tradução (Alvamar Lamparelli) e ensino de língua estrangeira (Helmara Moraes). O TextQuim permite investigar a leitura em língua estrangeira na Área de Química, relacionando-a à Tradução. Já o e-Termos propõe soluções para a extração automática de terminologia baseada em corpus em um ambiente computacional colaborativo.

Outros projetos, os quais não utilizam LC aplicada à Terminologia, são o Brasilterm, o AGROVOC e o AGRIS, que explicaremos em seguida.

O Projeto Brasilterm (Faulstich, 1995b), Banco de Dados Terminológico do Brasil, é divulgado em sites internacionais (ex: <http://www.termportal.de/organisationen2.html#suedam>) como o referencial terminológico e socioterminológico do Brasil. No entanto, acessando o site www.ibict.br, conforme indicação, não conseguimos encontrar nada relacionado ao nome Brasilterm. Entrando em outros links, como <http://www.canalciencia.ibict.br/>, encontramos *diretório de links*, que nos levou a *glossários*. Não há, porém, referência ao nome Brasilterm. É possível que o projeto, já finalizado, tenha gerado os referidos glossários.

⁴⁵ Disponível em <<http://www.ffflch.usp.br/dlm/comet>>. Acesso em 1 dez. 2006.

Porém, em sua proposta, o Brasilterm previa sete complexos temáticos para o desenvolvimento de Bancos Terminológicos, dentre eles “Beneficiamento de Café”, no Complexo 1. Nada encontramos para esse Complexo, mas em Strehler (1995) é desenvolvida a metodologia para a coleta de termos na área de autopeças, parte do Brasilterm. O trabalho é caracterizado por ser intuitivo, uma vez que se baseia no relato de vendedores de lojas, que indicavam as peças que eram chamadas de diferentes maneiras pelos fabricantes e pelos clientes.

O projeto AGROVOC, da Food and Agricultural Organization of The United Nations - FAO, de abordagem ontológica, é um Thesaurus que possui uma ferramenta de busca de termos de agricultura em diversas línguas, inclusive português. Esse projeto informa apenas o significado da palavra de busca em outras línguas, ou então palavras que possam pertencer ao mesmo campo semântico. O AGRIS, da mesma instituição, com sistema classificatório de abordagem. Nenhum dos dois faz menção ou à Lingüística de Corpus ou Socioterminologia, e estão restritos a termos mais comuns das áreas.

Ante todos os projetos apresentados, podemos afirmar que nossa pesquisa é pioneira no Brasil em conciliar a utilização dos princípios da LC aplicados à Socioterminologia na língua falada, na subárea do café.

CAPÍTULO 3

METODOLOGIA

No capítulo anterior apresentamos as disciplinas envolvidas em nosso trabalho e suas interações. Demonstramos que, para que a pesquisa possa ser realizada, necessitamos compreender cada parte do trabalho, relacionando-as aos estudos de suas áreas específicas.

Neste capítulo identificamos os procedimentos necessários para a realização da pesquisa, iniciando com a construção detalhada dos corpora.

O estudo é baseado nos princípios da Lingüística de Corpus aplicados ao campo socioterminológico, considerando que apenas a intuição sobre certas realizações lingüísticas nas pesquisas de muitos estudiosos não é suficiente para fundamentar seus resultados (Pearson, 1998). Dessa maneira, consideramos essencial seguir a metodologia de análise da Lingüística de Corpus para observarmos a ocorrência não só de termos possíveis, mas essencialmente dos prováveis na língua falada e na língua escrita. A combinação da Socioterminologia e da Lingüística de Corpus para a realização deste estudo só traz vantagens, pois o que antes era feito baseando-se na intuição do tradutor e do intérprete, hoje pode ser validado por generalizações a partir de evidências atestadas pelo corpus. A Lingüística de Corpus é, portanto, mais do que apenas uma maneira rápida e prática de realizar um estudo terminológico, pois possibilita a análise detalhada de realizações lingüísticas na oralidade e na escrita.

3.1 A Construção dos Corpora

Pensando na seleção terminológica em todos os corpora deste trabalho, a fim de realizar o levantamento de dados para análise das diversas variantes terminológicas, elaboramos um planejamento criterioso para a construção dos corpora a serem confeccionados:

- a) Construção de dois corpora falados monolíngües, um em português do Brasil (CAF_BO) e outro em inglês (CAF_IO), com o tema Café, composto por entrevistas face-a-face com profissionais da área cafeeira em ambas as línguas, bem como conversas entre diversos profissionais da área, também em ambas as línguas (Figura 23).
- b) Construção de corpus bilíngüe, (CAF_BI), híbrido, pois as falas se alternam, composto por interpretações feitas na área da cafeicultura envolvendo falantes de cada língua e intérpretes intermitentes, que traduzem ao mesmo tempo o idioma de origem e de destino, passando informações nos dois idiomas entre duas ou mais pessoas (Figura 23).
- c) Construção de corpus geral de línguas faladas, em português e inglês (CAF_ALL_PB e CAF_ALL_IN), unindo os corpora de interpretações e entrevistas/conversações em cada língua. (Figura 23).
- d) Construção de dois corpora escritos monolíngües (CAF_BE e CAF_IE), um em PB e o outro em EN, compostos por textos eletrônicos extraídos da Internet (Figura 24).

Explicaremos agora a construção dos corpora planejados, passo a passo, incluindo a coleta de dados, a transcrição, o design dos corpora falados e escritos, as dificuldades encontradas e sugestões que podem facilitar futuras pesquisas sobre oralidade.

Em primeiro lugar, trataremos das tarefas de coleta e transcrição de dados. Em seguida exporemos a construção dos diversos tipos de corpora de estudo e apresentaremos os corpora de referência para a pesquisa (Figura 25).

Figura 23 - Corpora Falados

Figura 24 - Corpora Escritos

Figura 25 - Corpora de referência

3.1.1 Coleta de dados

Coletar dados para um corpus falado parece uma tarefa bem simples, se comparada com a transcrição desses dados. No entanto, para que a coleta seja eficaz, é preciso estabelecer previamente quem serão os informantes. Em nossa pesquisa não houve necessidade de selecionar informantes de acordo com critérios como idade ou sexo, uma vez que o objetivo é o estudo socioterminológico. Optamos por estabelecer critérios segundo a profissão de cada informante, que deveria ser relacionada de alguma forma à subárea Café. Temos informantes dos seguintes grupos profissionais: trabalhadores rurais, incluindo apanhadores de café, abanadores e empregados de terreiro ou de secador; produtores de café, que normalmente são os proprietários das terras; engenheiros agrônomos, técnicos do maquinário, exportadores de café, provadores e consultores.

Os dados coletados foram gravados em sistema analógico ou digital. No corpus de entrevistas temos apenas áudio. Já em interpretações temos um vídeo. Porém, percebemos que os informantes são menos espontâneos quando se trata de gravação audiovisual do que em gravações de áudio.

Atualmente, devido à alta tecnologia, os gravadores são cada vez menores e mais discretos, possibilitando que fiquem até mesmo escondidos. Já as câmeras de filmagem, por menores que sejam, precisam ficar expostas.

Antes da gravação, a documentadora orientava os informantes apenas em relação ao objetivo da pesquisa. Os falantes estrangeiros receberam uma carta informativa apresentando o projeto antes de entrarem em contato com a documentadora (Anexo I). Essa observação é pertinente, pois segundo Kennedy (1998) "... a relação entre os interlocutores pode ter um efeito importante em algumas variáveis dos dados coletados"⁴⁶.

Um problema oriundo da coleta de dados foi o equipamento utilizado. Como o mini-gravador possuía sensor de voz, a documentadora percebeu que em algumas situações, especialmente quando o participante parava de falar momentaneamente para organizar sua fala, ou quando sua voz era muito baixa, a gravação era interrompida. Isso aconteceu inúmeras vezes e, quando percebido, as gravações eram retomadas a partir do ponto final gravado. A sugestão é que sejam utilizados gravadores sem sensor de voz. O uso de um microfone é indispensável, pois as vozes dos participantes são destacadas em detrimento dos ruídos do ambiente. Entretanto, muitos participantes pediram que não fosse utilizado, pois se sentiam inibidos com sua utilização.

Houve casos em que, no meio de uma gravação, o gravador de voz apresentou problemas técnicos. Assim, muitas gravações não foram utilizadas, quer por dificuldades na transcrição devido aos ruídos, quer por problemas técnicos de equipamento.

Apesar de o gravador digital utilizado não apresentar defeitos técnicos, ele possui tempo máximo de gravação de duas horas no modo mais nítido e de quatro horas em modo de qualidade inferior. Para gravar nova entrevista, é preciso baixar os arquivos gravados em um computador. Na coleta de dados feita na Etiópia, esse processo dificultou as gravações, pois quando se passavam duas horas, era preciso parar de gravar, baixar os

⁴⁶ "... the relation between the interlocutors may have a marked effect on the kind of data collected for some variables."

arquivos em *laptop*, e somente depois voltar a utilizá-lo. Muitas vezes sua utilização ficava restrita às primeiras duas horas, pois havia lugares de difícil acesso, em que carregar um *laptop* era absolutamente inviável.

As gravações de entrevistas são compostas de diálogos dirigidos pela documentadora sobre o tema café. Seleccionamos *colheita* e *processamento pós-colheita* por se tratarem de tópicos comuns a todos os profissionais da área. Porém, há momentos em que a conversa foge do tema selecionado, uma vez que na língua falada é difícil manter o controle do assunto que se planejou. Já as gravações de conversações não possuem interferência da documentadora. Diversas conversas foram gravadas durante visitas a fazendas no Brasil e na Etiópia, bem como a usinas de processamento de café. Quanto às gravações de interpretações, temos poucos exemplos, uma vez que não obtivemos permissão das empresas envolvidas para gravá-las. A justificativa das companhias foi de que o material era sigiloso, e não poderia ser divulgado.

3.1.2 Transcrição de dados

Muitos corpora orais foram analisados para embasamento da transcrição desta pesquisa, entre eles o MICASE⁴⁷, o SBCSAE⁴⁸, LLC⁴⁹, o Projeto NURC/SP⁵⁰, entre outros.

Devido aos nossos objetivos voltados ao plano lexicográfico, portanto contextual e privilegiando o sentido, os dados em português do Brasil e das variantes de inglês foram transcritos de acordo com a ortografia convencional, "... levando em conta a produção real, usando os sinais e, em parte, as normas convencionais da ortografia portuguesa" (Marcuschi *apud* Rodrigues e Ferreira Neto, 2000, p. 173), num parâmetro muito próximo ao desenvolvido pelo NURC, e por Marcuschi (2003), com algumas adaptações.

⁴⁷ Disponível em <www.hti.umich.edu>. Acesso em 16 mai. 2006.

⁴⁸ Disponível em www ldc.upenn.edu/Projects/SBCSAE. Acesso em 16 mai. 2006.

⁴⁹ Disponível em <http://khnt.hit.uib.no/icame> Acesso em 16 mai. 2006.

⁵⁰ Disponível em <http://www.fflch.usp.br/dlcv/nurc>. Acesso em 16 mai. 2006.

O programa utilizado para a análise do material, WordSmith Tools (Scott, 2004-2007), não exige padrão de transcrição específico, portanto não há regras determinadas a seguir. No Anexo II encontramos um exemplo de um inquérito transcrito do nosso trabalho.

As palavras pronunciadas de maneira diferente à norma culta não foram corrigidas ou normalizadas, para que assim não perdêssemos a autenticidade do texto. Por exemplo, se houvesse a ocorrência da palavra tuia (norma culta: tulha), seria mantida como foi falada. Também estão incluídas as pausas – representadas por reticências, hesitações – representadas por pela própria palavra, truncamentos – representados pela barra oblíqua /, bem como os turnos de cada participante, enumerados em ordem crescente segundo os participantes de cada entrevista. Quanto à pontuação, utilizamos a norma padrão. Os números são representados por extenso e as siglas (como USP, por exemplo) aparecem em letras maiúsculas. Também os nomes próprios iniciam-se por letra maiúscula. Quando a transcrição de determinado segmento não pôde ser feita por incompreensão, há um parêntese com reticências (...). Incluímos entre < > comentários do documentador sobre ocorrências durante as entrevistas que pudessem interferir na compreensão do assunto. Nossa maior preocupação foi, portanto, com a marcação de detalhes relevantes à pesquisa, como palavras ditas fora do padrão. Alguns detalhes foram deixados de lado, tais como a entonação fática de palavras ou frases, importantes para uma pesquisa sobre o significado do discurso, por exemplo. No Anexo III apresentamos um exemplo de transcrição do NURC com exemplos de entonação fática de palavras.

Vejamos abaixo o quadro com a codificação de nossa transcrição:

Quadro 3 - Normas para transcrição

ITENS REPRESENTADOS	CÓDIGOS DE TRANSCRIÇÃO	EXEMPLOS
pausas	...	A... a automotriz
hesitações	própria palavra ou reticências	Eééé...são máquinas especiais
siglas	Letras maiúsculas	USP
Comentários do transcritor	<>	Armazemanen/ armazema/<risos>
Representações de sons inaudíveis	(...)	A colher um (...)
números	Por extenso	Vinte e dois alqueires
Truncamentos de palavras	/	Você tem que pe/ colocar...

Houve um momento da pesquisa em que decidimos utilizar o programa computacional CLAN, interface do sistema de transcrição CHAT do projeto CHILDES. Devido ao seu moderno modelo de transcrição, que pode ser associado ao áudio, e também por permitir a análise das interações entre falantes, acreditamos que seria o ideal para o corpus de interpretações. Porém, os códigos de transcrição têm de ser treinados, devido ao seu alto nível de elaboração (Anexo IV). Isso dificultou sua utilização em nossa pesquisa, por termos um grande número de textos a digitar, que exigiriam tempo surpreendente.

De qualquer forma, as gravações foram todas digitalizadas, ou seja, há um arquivo das gravações em seu estado “cru”, pois sempre poderão ser consultadas quando houver dúvidas em relação à transcrição. O software utilizado para a digitalização de dados foi o Express Scribe, da NCH Swift Sound (2007).

Como transcrever demanda muito tempo, procuramos utilizar softwares de transcrição que pudessem nos auxiliar com o trabalho. Entretanto, como nossos áudios incluíam diversas vozes, o software não conseguia reconhecê-las. Para treiná-lo, teríamos novamente grande demanda de tempo. Dessa forma, decidimos utilizar a forma tradicional de transcrição. Felizmente, o Express Scribe permitiu que o trabalho fosse facilitado, pois podíamos reduzir a velocidade do áudio e assim trabalhar em suas partes mais difíceis.

No início do trabalho pensamos em fazer uma estimativa do tempo necessário para transcrever todas as gravações. Tivemos uma média pouco variável de 1000 palavras por hora em língua portuguesa, nas gravações de entrevistas. A conversação, em língua portuguesa, não conseguiu seguir a mesma média de tempo, tampouco as gravações em língua inglesa, que sofreram diversas influências, como ruídos do ambiente, falas sobrepostas, etc. Outra dificuldade enfrentada foi não fazer a transcrição logo após a gravação. Quando fizemos a transcrição tão logo gravamos, percebemos que a demanda de tempo era menor. Acreditamos que o motivo seja lembrar-se do conteúdo das falas.

Devido ao volume de trabalho, alguns transcritores foram contratados para que pudéssemos desenvolver mais rapidamente os corpora. Entre eles, uma agência especializada em transcrição. No entanto, em todos os casos foram encontrados problemas como palavras erradas, confusão com a nomeação do turno dos falantes, etc. Supõe-se que tais enganos se devam à falta de conhecimento da área cafeeira pelos transcritores, principalmente, e à qualidade dos áudios. Assim mesmo, a ajuda foi útil, pois fizeram o que chamamos de base do trabalho, ou seja, uma primeira transcrição do texto. Entretanto, o trabalho minucioso que exige tal transcrição, devido à presença de variantes, só pôde ser finalizado pela pesquisadora, que conhecia o contexto e os falantes.

3.1.3 Design dos corpora

Após a transcrição, começamos a trabalhar com a construção dos corpora falados e escritos. Apresentaremos primeiramente o design dos corpora falados e, em seguida, o dos corpora escritos.

3.1.3.1 Corpora falados “Café”

Os corpora falados “Café” visam fornecer dados para a pesquisa das variações terminológicas que ocorrem na subárea técnica da cafeicultura. A pesquisa foi feita em nível socioterminológico para o português do Brasil e inglês internacional. Por isso a necessidade de compilar um corpus especializado, que atendesse aos propósitos da pesquisa, e não um corpus de língua geral.

Os corpora falados “Café” são o que chamamos corpora de estudo, ou seja, foram construídos de acordo com nossas necessidades de pesquisa. Já os corpora de referência foram elaborados com base na língua geral, para servirem de termo de comparação de uma língua.

Atkins, Clear e Ostler (1992) sugerem alguns critérios para a construção de corpus:

- i. *delimitação dos temas abordados dentro da área*: a área cafeeira é abrangente, envolvendo aspectos relacionados à Agricultura em geral, bem como aspectos mercadológicos. Portanto, a definição dos temas foi feita pensando-se no caráter prático da questão. Se a problemática está voltada às variações terminológicas, o ideal seria abordar temas comuns às pessoas de níveis profissionais diferentes. Por isso foram escolhidos dois temas: colheita e processamento do café, embora outros assuntos com interferência direta nos temas sugeridos, como, por exemplo, a maneira como o processamento utilizado reflete na qualidade da bebida ou como o plantio interfere na colheita etc., tenham sido incluídos. Não havia, entretanto, perguntas pré-elaboradas, para que os participantes não fossem induzidos a determinados assuntos e termos.
- ii. *seleção dos representantes das diversas categorias de trabalho na área*: as categorias selecionadas foram os trabalhadores rurais, incluindo apanhadores de café, abanadores e empregados de terreiro ou de secador; produtores de café, que normalmente são os proprietários das terras; engenheiros agrônomos, exportadores, provadores e consultores. Todos eles podem falar com conhecimento sobre os temas “colheita” e “processamento”.

iii. *determinação de uma área geográfica única para todos os representantes:* para o português falado no Brasil, pessoas residentes ou nascidas nas regiões cafeeiras do sul de Minas Gerais e da Mogiana de São Paulo, aqui representadas pelas cidades de Andradas - MG, de Santo Antonio do Jardim - SP e Espírito Santo do Pinhal - SP. O fato de as cidades estarem localizadas em dois estados não interfere no resultado da pesquisa, uma vez que as cidades são vizinhas, com uma distância aproximada de 12 km entre Andradas e Santo Antonio do Jardim, 12 km entre Santo Antonio do Jardim e Espírito Santo do Pinhal e 24 km entre Andradas e Espírito Santo do Pinhal. Vejamos no mapa abaixo a localização dessas cidades:

Figura 26 - Mapa regional das regiões abrangidas no PB [Fonte: Escola de Parapente Ventomania http://www.ventomania.com.br/data/mapa_andradas.htm]

Para o inglês, contribuíram falantes dos seguintes países: Estados Unidos da América, Reino Unido, Tanzânia, Etiópia, Quênia, Sri Lanka, Vietnã, Tailândia, Burundi, Índia, Alemanha, Suíça, Colômbia, Argentina. Trata-se de um corpus oportunístico, pois reúne inquéritos coletados durante o trabalho de interpretação no Brasil, de países onde o inglês é considerado língua oficial de negócios e que tem relação comercial ou de produção com o café. Outros inquéritos foram coletados durante visita à Etiópia, em uma Convenção Internacional de Produtores de Cafés de Qualidade do continente Africano.

- iv. *definição da tipologia textual*: selecionamos a entrevista e a conversação para que dessa maneira a documentadora não interferisse substancialmente no discurso do informante. A condução da entrevista foi determinada totalmente pela documentadora de acordo com sua interação com o participante.
- v. *especificação do número de participantes de cada amostra de texto na entrevista*: uma vez escolhida a entrevista, estabeleceu-se que o número máximo de participantes de cada texto seriam dois, além da documentadora, para facilitar a transcrição. Nas conversações, porém, não foi possível manter esse critério, e várias pessoas participaram ao mesmo tempo. É comum, nos casos de interação, que os participantes sobreponham suas falas, o que dificulta o processo de transcrição.
- vi. *estabelecimento do local das gravações*: os participantes foram, em geral, entrevistados no local de trabalho, mas puderam escolher seus locais preferidos, pois pretendíamos evitar timidez, tensão ou nervosismo, comuns a essa situação. Acreditamos que a familiaridade com o local da entrevista garantiria uma interação mais espontânea.
- vii. *extensão de cada texto*: as entrevistas variaram em relação ao tempo, pois o que era importante nesta pesquisa era a qualidade das informações, ou seja, o conteúdo da fala individual.

viii. *todos os participantes concederam sua autorização para a utilização das gravações em projetos de estudos lingüísticos*, o que facilitou o trabalho de transcrição, pois não houve a necessidade de desgravar ou omitir os nomes dos informantes. Apenas um deles quis ouvir sua entrevista antes de assinar a autorização.

Os corpora falados estão compostos de subcorpora de entrevistas/conversação e de interpretações (vide Figura 23), em português e inglês.

A) Subcorpus de Entrevistas/Conversação

Os textos que o compõe são resultados das transcrições de entrevistas e conversações gravadas, que representam o falar de categorias profissionais, ou seja, são amostras das produções orais dessas pessoas dentro de sua área de atuação, a cafeicultura. Está composto por um corpus em Português falado no Brasil e outro em inglês falado em diversos países que utilizam o inglês como língua de negócios. O corpus em Português Falado é chamado CAF_BO – **BRASIL/ORAL**. O corpus em inglês é chamado CAF_IO – **INGLÊS/ORAL**.

Apesar de as entrevistas terem sido conduzidas pela documentadora, suas falas foram consideradas representativas, devido à sua familiaridade com a área cafeeira. Quando dizemos que as línguas de especialidade só são assim chamadas por se tratarem de:

“...variedade composta por um sistema de signos caracteristicamente construídos e/ou relacionados, empregados por um grupo de usuários determinado e utilizados para a descrição e veiculação de informação em uma área de conhecimento ou atuação profissional particular” (Finatto, 2001).

Vemos que o trabalho do intérprete corresponde ao que Finatto (2001) chama de profissional, uma vez que há simetria entre o seu conhecimento adquirido por pesquisas e o conhecimento do profissional da área. A pesquisadora atua na área cafeeira há alguns anos e possui conhecimento suficiente do assunto para manter um diálogo no mesmo nível do profissional.

Abaixo se encontra a lista estatística do corpus CAF_BO (Figura 27). A relação entre o número de palavras corridas (*tokens*) e as formas distintas (*types*) é alta, com média de índice 10 no corpus. Observando-se cada texto, percebemos que a razão *type/token* de cada texto gira ao redor do índice 40, porém quando analisados em conjunto, cai para 10. Isso significa que, se analisadas por textos individualmente, as palavras se repetem menos do que em conjunto, embora o percentual de 10 ainda seja demonstrativo de alta riqueza lexical. Supõe-se que, devido à entrevista ou conversação serem dinâmicas, os mesmos assuntos não se repetem em cada situação; porém, quando analisados em conjunto, tratam-se sempre do mesmo tema e acabam originando repetições. Já o corpus CAF_IO possui 13.390 palavras, portanto menor que o CAF_BO, de 39.859. Apesar de ser pequeno, é um corpus de grande riqueza lexical, com índice 14 (Figura 28).

	N	Overall	1	2	3	4	5	6	7	8	9	
text file		Overall	5pb.txt	4pb.txt	3pb.txt	2pb.txt	1pb.txt	0pb.txt	9pb.txt	8pb.txt	7pb.txt	6pb.txt
file size	250.480	12.880	6.287	4.076	2.854	3.662	1.554	1.757	2.439	14.274	10.	
tokens (running words) in text	39.942	2.081	874	530	419	131	153	188	310	2.411	1.	
tokens used for word list	39.859	2.081	874	530	404	131	153	188	310	2.411	1.	
types (distinct words)	3.801	612	265	214	179	87	83	92	158	616		
type/token ratio (TTR)	10	25	30	40	44	66	54	49	51	26		
standardised TTR	32,62	30,90								33,15	30	
standardised TTR std dev	63,67	48,86								47,27		
standardised TTR basis	1.000,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	0,00,00	
mean word length (in characters)	4	4	4	4	4	4	4	4	4	4	4	
word length std dev	2,49	2,63	2,02	2,32	2,07	2,26	1,97	1,92	2,14	2,36	2	
sentences	2.960,00	121,00	148,00	68,00	14,00	6,00	9,00	16,00	20,00	39,00	146	
mean (in words)	13	17	6	8	29	16	17	12	16	62		
std dev	18,01	14,06	4,56	9,10	28,36	14,40	16,12	10,75	11,36	88,62	11	
paragraphs	19,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1	
mean (in words)	2.098	2.081	874	530	404	131	153	188	310	2.411	1.	
std dev	1.810,12											
headings												
mean (in words)	0	0	0	0	0	0	0	0	0	0	0	
std dev												
sections	19,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1	
mean (in words)	2.098	2.081	874	530	404	131	153	188	310	2.411	1.	

frequency

alphabetical

statistics

filenames

notes

75

Type-in

CAFE

Iniciar

PT

17:11

Figura 27 - Lista estatística do corpus CAF_BO

	N	Overall	1	2	3	4	5	6	7	8	9
test file		Overall	01in.txt	02in.txt	03in.txt	1en.txt	02in.txt	03in.txt	04in.txt	05in.txt	06in.txt
file size	90,515	2,625	4,722	37,868	3,080	2,016	5,866	6,921	2,858	10,666	14,
tokens (running words) in text	13,427	296	471	5,993	449	233	818	1,121	363	1,771	2,
tokens used for word list	13,390	296	471	5,954	449	233	818	1,121	363	1,771	2,
types (distinct words)	1,925	125	179	1,182	155	125	256	357	157	447	
types/token ratio (TTR)	14	49	38	21	35	54	31	32	43	25	
standardised TTR	34,79			36,40				33,40		29,40	34
standardised TTR std dev	56,56			52,56							46
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00
mean word length (in characters)	4	4	4	4	4	4	4	4	5	4	
word length std dev	2,24	2,08	2,07	2,16	2,34	2,54	2,33	2,22	2,63	2,26	2
sentences	901,00	26,00	74,00	390,00	29,00	21,00	76,00	87,00	21,00	64,00	91
mean (in words)	15	9	6	14	15	11	11	13	17	21	
std dev	16,25	11,49	5,62	11,38	18,91	14,02	11,18	9,56	23,43	16,69	32
paragraphs	10,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1
mean (in words)	1,339	296	471	5,954	449	233	818	1,121	363	1,771	2,
std dev	1,641,83										
headings											
mean (in words)	0	0	0	0	0	0	0	0	0	0	
std dev											
sections	10,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1
mean (in words)	1,339	296	471	5,954	449	233	818	1,121	363	1,771	2,

Figura 28 - Lista estatística do CAF_10

Quadro 4 - Subcorpora falados entrevistas/conversação, PB e EN

Subcorpora Falados Entrevistas/Conversação	nº palavras	nº informantes
PB	26.269	29
EN	13.390	33
TOTAL	39.659	62

B) Subcorpus de Interpretações

A nomenclatura do corpus de Interpretações é CAF_BI (**B**ILÍNGÜE). Sua estrutura pode ser analisada pela lista estatística (Figura 29), onde observamos que possui em torno de 20.000, porém em ambas as línguas.

N	Overall	1	2	3	4	5	6	7	8	9
text file	Overall	m01.txt	m02.txt	m03.txt	m04.txt	m05.txt	m06.txt	m07.txt	m08.txt	m09.txt
file size	147,550	10,784	10,253	14,455	30,800	11,214	31,211	22,479	11,138	5,416
tokens (running words) in text	20,500	1,361	1,317	1,914	4,395	1,745	4,371	3,206	1,520	681
tokens used for word list	20,498	1,361	1,317	1,909	4,395	1,744	4,366	3,206	1,519	681
types (distinct words)	2,718	600	491	625	1,072	660	950	838	517	269
types/tokens ratio (TTR)	13	37	37	33	24	32	22	26	34	40
standardised TTR	38,96	39,60	40,60	40,30	39,88	38,60	36,65	39,27	40,20	
standardised TTR std. dev.	66,60				48,10		50,66	46,75		
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00
mean word length (in characters)	4	4	4	4	4	4	4	4	4	4
word length std. dev.	2,09	2,10	2,25	2,14	2,03	2,00	2,11	2,02	2,15	2,03
sentences	2,561,00	155,00	166,00	292,00	506,00	207,00	617,00	387,00	155,00	75,00
mean (in words)	8	9	8	7	9	8	7	8	10	9
std. dev.	7,62	10,18	6,66	5,43	8,66	8,27	6,88	6,93	9,18	7,97
paragraphs	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	2,276	1,361	1,317	1,909	4,395	1,744	4,366	3,206	1,519	681
std. dev.	1,368,05									
headings										
mean (in words)	0	0	0	0	0	0	0	0	0	0
std. dev.										
sections	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	2,276	1,361	1,317	1,909	4,395	1,744	4,366	3,206	1,519	681
std. dev.	1,368,05									

Figura 29 - Lista de palavras (estatística) do CAF_BI

Criamos, dentro do CAF_BI, as separações entre PB e EN, tendo assim o CAF_BI_PB e o CAF_BI_IN, cuja composição está explicada adiante.

Para a lista de palavras do corpus CAF_BI em português ou inglês, é preciso configurar o programa WordSmith Tools a fim de que seja capaz de compreender uma língua de cada vez.

Antes de colocarmos etiquetas separadas em cada informante de número de identificação e língua, havíamos planejado uma única etiqueta, com número e língua (ex: <2pb>.....</2pb>). Porém, a decisão de substituí-las (ex. <2> <pb>..... </2> </pb>) foi baseada no teste que fizemos com a configuração da WordList para buscar frases em línguas determinadas. Em nosso teste, tentamos fazer a configuração da seguinte maneira (Figura 30), para ignorar trechos em língua portuguesa:

Figura 30 - Configuração de extração de língua portuguesa para ignorar o PB

Os resultados obtidos estão apresentados abaixo (Figura 31):

N	Word	Freq	%	Texts	% Lemmas	Set
1	HERE	10	1,43	3	33,33	
2	RIGHT	7	1,00	1	11,11	
3	MIXED	6	0,86	1	11,11	
4	CHERRIES	5	0,72	2	22,22	
5	COFFEE	5	0,72	4	44,44	
6	HARVEST	5	0,72	2	22,22	
7	PER	5	0,72	1	11,11	
8	BAG	4	0,57	1	11,11	
9	COME	4	0,57	2	22,22	
10	DEFEITO	4	0,57	1	11,11	
11	GOING	4	0,57	2	22,22	
12	IDEA	4	0,57	1	11,11	
13	LIGHT	4	0,57	1	11,11	
14	LIKE	4	0,57	3	33,33	
15	OH	4	0,57	3	33,33	
16	SAY	4	0,57	3	33,33	
17	WANT	4	0,57	1	11,11	
18	YEAR	4	0,57	1	11,11	
19	ALLOWING	3	0,43	1	11,11	
20	ANY	3	0,43	1	11,11	
21	FIVE	3	0,43	2	22,22	
22	IMPROVE	3	0,43	1	11,11	
23	MATHS	3	0,43	1	11,11	

Figura 31 - Lista parcial de frequência em inglês do corpus CAF_BI

Como houve a ocorrência da palavra defeito, em um corpus que só deveria apresentar inglês, acreditamos que fosse decorrente de erro durante a marcação. Para saná-lo, procuramos o arquivo em que ocorreu a palavra e corrigimos a marcação errônea. Foi possível identificar o arquivo e descobrir que ele tinha as marcações de início, <*>, porém não as de fim </*>. Remarcado, o arquivo apresentava apenas palavras em inglês. Outra lista de palavras foi feita, para comprovar a eficiência do corpus. O problema que então se apresentou foi que alguns arquivos não estavam sendo lidos. Na tentativa de solucioná-lo, descobrimos que algumas etiquetas estavam em letras maiúsculas. Após a correção, o programa continuou a apresentar problemas, zerando as *tokens* do arquivo um e do arquivo três (Figura 32).

	N	Overall	1	2	3	4	5	6	7	8	9
test file		Overall	m01.txt	m02.txt	m03.txt	m04.txt	m05.txt	m06.txt	m07.txt	m08.txt	m09.txt
file size		148,318	10,685	10,246	14,448	30,592	12,263	31,204	22,472	11,131	5,387
tokens (running words) in text		675	0	97	0	5	17	41	17	83	415
tokens used for word list		675	0	97	0	5	17	41	17	83	415
types (distinct words)		248	0	88	0	5	14	33	13	53	163
typo/token ratio (TTR)		.37	0	.70	0	100	.82	.80	.76	.64	.39
standardised TTR											
standardised TTR std dev											
standardised TTR basis		1,000.00	0.000.00	0.000.00	0.000.00	0.000.00	0.000.00	0.000.00	0.000.00	0.000.00	0.000.00
mean word length (in characters)		4	0	5	0	3	4	4	4	4	4
word length std dev		2.04		2.40		1.84	1.77	1.82	2.08	2.13	1.84
sentences		75.00		14.00		1.00	3.00	10.00	2.00	7.00	38.00
mean (in words)		9	0	7	0	5	6	4	9	12	11
std dev		8.29		4.23			7.23	3.38	7.78	12.50	9.16
paragraphs		7.00		1.00		1.00	1.00	1.00	1.00	1.00	1.00
mean (in words)		96	0	97	0	5	17	41	17	83	415
std dev		144.75									
headings											
mean (in words)		0	0	0	0	0	0	0	0	0	0
std dev											
sections		7.00		1.00		1.00	1.00	1.00	1.00	1.00	1.00
mean (in words)		96	0	97	0	5	17	41	17	83	415
std dev		144.75									

Figura 32 - Lista estatística com problemas de leitura de arquivos

Foi preciso, dessa forma, voltar ao Manual do *WordSmith Tools*, para verificar se a opção com asterisco era válida nesta parte do programa. Infelizmente, há opção de excluir ou incluir apenas três marcadores. O uso do asterisco para tentar incluir todos os marcadores não funciona para esse atalho do programa, “Only part of file”. Já usávamos um campo para

<header> </header>, sobrando apenas dois. Uma vez que tínhamos um número maior que dois informantes nos arquivos de interpretações, seria impossível continuar com a marcação naquele formato.

Finalmente, a opção de deixar as marcas de língua <EN> ou <PB> sem números mostrou-se satisfatória. Todos os arquivos foram usados pelo programa com o novo critério de marcação. Vejamos as listas estatísticas do corpus CAF_BI_PB (Figura 33) e do CAF_BI_IN (Figura 34), para analisarmos sua constituição final:

	N	Overall	1	2	3	4	5	6	7	8	9
text file		Overall	m01.txt	m02.txt	m03.txt	m04.txt	m05.txt	m06.txt	m07.txt	m08.txt	m09.txt
file size	151.538	10.622	10.224	14.833	31.402	12.498	32.118	22.922	11.251	5.468	
tokens (running words) in text	10.324	1.160	523	1.322	2.319	776	2.088	1.365	721	60	
tokens used for word list	10.314	1.160	523	1.318	2.319	776	2.083	1.365	720	60	
types (distinct words)	1.729	386	222	414	675	311	571	480	293	40	
types/token ratio (TTR)	17	34	42	31	29	40	27	35	41	67	
standardised TTR	36,16	34,60		36,80	36,90		35,10	37,90			
standardised TTR std. dev.	54,87				44,62		45,89				
standardised TTR basis	1.000,00	000,00	000,00	000,00	000,00	000,00	000,00	000,00	000,00	000,00	
mean word length (in characters)	4	4	4	4	4	4	4	4	4	4	
word length std. dev.	2,22	2,12	2,49	2,17	2,14	2,21	2,31	2,14	2,33	2,01	
sentences	1.293,00	131,00	67,00	215,00	264,00	68,00	283,00	179,00	51,00	15,00	
mean (in words)	8	9	8	6	9	9	7	8	14	4	
std. dev.	7,46	10,33	5,96	4,89	7,90	7,61	6,36	6,20	11,60	2,51	
paragraphs	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	
mean (in words)	1.145	1.160	523	1.318	2.319	776	2.083	1.365	720	60	
std. dev.	726,06										
headings											
mean (in words)	0	0	0	0	0	0	0	0	0	0	
std. dev.											
sections	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	
mean (in words)	1.145	1.160	523	1.318	2.319	776	2.083	1.365	720	60	
std. dev.	726,06										

Figura 33 - CAF_BI_PB - Estatística

	N	Overall	1	2	3	4	5	6	7	8	9
text file		Overall	m01.txt	m02.txt	m03.txt	m04.txt	m05.txt	m06.txt	m07.txt	m08.txt	m09.txt
file size	151,538	10,622	10,224	14,833	31,402	12,458	32,118	22,922	11,251	5,468	
tokens (running words) in text	10,581	211	1,312	591	2,053	973	2,257	1,793	770	621	
tokens used for word list	10,579	211	1,312	590	2,053	972	2,257	1,793	770	621	
types (distinct words)	1,344	118	469	228	433	254	419	436	245	231	
types/token ratio (TTR)	13	66	37	39	21	26	19	24	32	37	
standardised TTR	29,65		40,70		27,90		26,40	28,80			
standardised TTR std. dev.	58,09				50,98		52,04				
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	
mean word length (in characters)	4	4	4	4	4	4	4	4	4	4	
word length std. dev.	1,98	1,96	2,26	2,07	1,90	1,83	1,91	1,93	1,98	2,03	
sentences	1,311,00	24,00	165,00	77,00	243,00	119,00	328,00	195,00	98,00	51,00	
mean (in words)	8	9	8	8	8	8	7	9	8	10	
std. dev.	7,91	9,26	5,67	7,44	9,33	9,15	7,35	7,76	7,08	8,40	
paragraphs	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	
mean (in words)	1,175	211	1,312	590	2,053	972	2,257	1,793	770	621	
std. dev.	718,15										
headings											
mean (in words)	0	0	0	0	0	0	0	0	0	0	
std. dev.											
sections	9,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	
mean (in words)	1,175	211	1,312	590	2,053	972	2,257	1,793	770	621	
std. dev.	718,15										

Figura 34 - CAF_BI_IN - Estatística

Por meio das listas estatísticas, fica claro que a interação de turnos entre os informantes acontece com relativa similaridade, pois o número de palavras corridas dos dois corpora está equilibrado, ao redor de 10.000 palavras em cada um.

As dificuldades para a composição desse Corpus estão relacionadas a diversos aspectos:

- Nem todas as empresas para as quais trabalhamos como intérprete permitem que gravemos seus visitantes, devido a questões de confidencialidade de dados, entre outras. As empresas que proporcionaram liberdade de gravação de todos os seus visitantes chamam-se P&A Marketing Internacional e Máquinas Agrícolas Pinhalense.
- Não são todos os visitantes que aceitam a gravação, mesmo com a explicação de tratar-se de um projeto investigativo na área lingüística, não comercial.
- Problemas técnicos da gravação, tais como os ruídos das máquinas de processamento, que impediram a transcrição.

- Falta de conhecimento profissional do visitante, ou seja, alguns visitantes ainda não fazem parte do negócio e, portanto, não dominam a área técnica. Limitam-se a concordar com as explicações dos procedimentos e, muitas vezes, apenas repetem o que foi dito, o que nos obrigou a descartar tais gravações.

Quadro 5 - Subcorpus falado Interpretações

Subcorpus Falado Interpretações	nº palavras	nº informantes
PB	10.314	10
EN	10.579	7
TOTAL	20.893	17

C) Corpora falados CAF_ALL_PB e CAF_ALL_IN

Ainda que as listas dos dois corpora de língua falada (ENTREVISTAS/CONVERSAÇÃO e INTERPRETAÇÃO) sejam muito ricas e nos forneçam material suficiente para estudo, resolvemos fazer uma lista única, em português e em inglês, que englobasse os dois subcorpora, para analisarmos a ocorrência de suas palavras em conjunto. Essas novas listas são assim formadas:

$$\text{CAF_ALL_PB} = \text{CAF_BO} + \text{CAF_BI_PB}$$

$$\text{CAF_ALL_IN} = \text{CAF_IO} + \text{CAF_BI_IN}$$

3.1.3.1.1 Marcação dos corpora falados

Cada arquivo dos corpora falados acima recebeu uma nomeação, a fim de identificar a tipologia textual. Assim temos:

- ENT para entrevista
- CON para conversaço
- INT para interpretaço

Quadro 6 - Nomeação de arquivosExemplos:

ENT02PB: 2º texto referente à entrevista realizada com informante de português

INT04: 4º texto referente à interpretação

CON01IN: 1º texto referente à conversação realizada entre informantes de língua inglesa

O número refere-se a uma organização interna de coleta, enquanto PB é língua portuguesa e IN inglês.

Os textos resultantes foram identificados individualmente com um cabeçalho para que, se necessário, o WordSmith Tools fizesse a busca de textos específicos. Consideramos dois tipos de dados: gerais e do participante. Nos dados gerais identificamos: nome do arquivo, data da coleta do texto, local de coleta, duração, nº de palavras e nome do transcritor. Para os participantes temos: nome, idade, sexo, grupo de trabalho, formação educacional e, para falantes de inglês, país de origem.

Vejamos a estrutura de um cabeçalho:

Quadro 7 - Modelo de cabeçalho em português

```

<header>
  <dadosgerais>
    Nome do arquivo: ENT05PB
    Data: Julho/04
    Local: residência do entrevistado
    Duração da entrevista: 20 min
    Nº de palavras: 2.694
    transcritor: Luciana
  </dadosgerais>
  <participantes>
    1pb.
    Nome: Luciana Latarini Ginezi
    Idade: 33 anos
    Sexo: f
    Grupo: documentadora
    Formação educacional: mestrado em andamento
    2pb.
    Nome: César (...)
    Idade: 24 anos
    Sexo: M
    Grupo: engenheiro agrônomo
    Formação educacional: Engenheiro Agrônomo – nível universitário
  </participantes>
</header>

```

O corpo do texto foi marcado manualmente, para a identificação dos falantes; o trabalho foi efetuado durante a transcrição dos dados. Assim, quando necessário estudar apenas as falas de determinado informante, é possível buscá-las com ajuda das anotações feitas. Para nosso trabalho, a anotação é extremamente útil no corpus de Interpretação, visto que podemos marcar as falas de língua inglesa ou língua portuguesa e trabalharmos apenas com aquela de que precisamos no momento. As identificações foram feitas utilizando o símbolo <> para início da anotação e </> para o final. Marcamos cada falante com um número, pois às vezes letras iniciais de nomes eram repetidas. Também adicionamos EN para inglês e PB para português. Partes do texto, trechos ou informantes que não pertenciam à área foram colocados entre <>, pois a configuração do programa permite definir trechos a ignorar, utilizando o símbolo <*>. Porém, caso haja necessidade de observar o contexto de determinadas ocorrências

próximas, a informação pode ser lida pelo pesquisador. Vejamos um exemplo de texto anotado:

Quadro 8 - Trecho de texto anotado

<1> <pb> Estamos aqui para fazer uma entrevista com o engenheiro agrônomo César Luiz Baruque Fuliaro de 24 anos. César, eu gostaria que você iniciasse a entrevista ahn... dizendo qual é a empresa para quem você trabalha. </1> </pb>

<2> <pb> Éh...Eu trabalho na Fuliagro Insumos Agrícolas, uma empresa que nasceu em Setembro de 2003, uma empresa que trabalha com todo tipo de insumo agrícola e também com todo tipo de assistência técnica para os mesmos que adquirem... </2> </pb>

Um índice com os nomes dos informantes, seus respectivos números e línguas utilizadas encontra-se no Anexo V. Os informantes com mais de uma participação também tiveram seus números repetidos, portanto o índice foi de importância fundamental do início ao fim da análise. Os intérpretes envolvidos no estudo receberam uma identificação em português e outra em inglês. Há o total de três intérpretes incluídos na pesquisa.

3.1.3.2 Corpora Escritos CAF_BE e CAF_IE

A fim de testar a hipótese de que a língua falada é menos rica do que a língua escrita em número de palavras que possam pertencer a um vocabulário específico, mesmo que a língua falada apresente maior número de variantes socioprofissionais, recorremos a um levantamento em corpora escritos neste trabalho. Normalmente, a terminologia registra apenas o discurso oficial dos profissionais mais graduados, deixando de fora os menos graduados, tais como técnicos agrícolas e trabalhadores rurais. (Strehler, 1995).

Os textos para esses corpora foram coletados em formato eletrônico, extraídos da web. Todos foram utilizados na íntegra, pois a localização dos

termos em um texto não é precisa; assim, se selecionado o início, poderíamos perder um termo que ocorre no final do texto, por exemplo.

Com o tema delimitado para “colheita” e “processamento”, buscamos textos eletrônicos direcionados para conhecedores do assunto ou envolvidos na área cafeeira, tais como engenheiros agrônomos e produtores de café. Só foram selecionados textos originais escritos em português do Brasil e inglês e que procedessem de sites confiáveis. Procuramos manter o balanceamento dos corpora em ambas as línguas com relação aos assuntos, para não haver discrepância nos resultados. Os sites consultados e seus respectivos assuntos estão no Anexo VI.

O tamanho dos corpora escritos, documento de Word, ou seja, antes de passá-los para texto sem formatação (.txt) é:

Quadro 9 - Corpora escritos - Número de palavras

Português	- 100.274 palavras
Inglês	- 102.379 palavras

Ao transformarmos os textos de Word para .txt, que é a maneira de o WordSmith Tools trabalhar com os arquivos, perdemos figuras, mapas, e algumas informações que não são passíveis de leitura pelo programa. No entanto, mantivemos um arquivo cru em Word para possibilitar ao pesquisador maiores detalhes sobre cada um, se necessário fosse. Devido ao objetivo da construção dos corpora escritos, ou seja, identificar a riqueza lexical de seus textos em comparação com os corpora falados, não houve necessidade de etiquetar e nomear cada texto coletado individualmente. Na transformação para .txt, utilizamos um só arquivo geral em cada língua, facilitando o trabalho e diminuindo a demanda de tempo para a pesquisa, já comprometida devido ao trabalho com a língua falada. A denominação ficou CAF_BE (**BRASIL – ESCRITO**) e CAF_IE (**INGLÊS - ESCRITO**).

Após a coleta dos textos e transformação em .txt, fizemos sua limpeza, eliminando símbolos ilegíveis ou espaços de linhas em branco, empecilhos para um bom funcionamento do WordSmith Tools. O resultado

da primeira visão geral do corpus pode ser visto abaixo (Figura 35), na Lista Estatística que segue. Observe que o número final de palavras utilizada em cada corpus pode ser vista na linha “*tokens used for wordlist*”.

	N	Overall	1
text file		Overall	cabra.txt
file size	647.914	647.914	647.914
tokens (running words) in text	105.234	105.234	105.234
tokens used for word list	101.830	101.830	101.830
types (distinct words)	9.966	9.966	9.966
type/token ratio (TTR)	10	10	10
standardised TTR	41,17	41,17	41,17
standardised TTR std. dev.	56,73	56,73	56,73
standardised TTR basis	1.000,00	1.000,00	1.000,00
mean word length (in characters)	5	5	5
word length std. dev.	3,29	3,29	3,29
sentences	3.769,00	3.769,00	3.769,00
mean (in words)	27	27	27
std. dev.	16,89	16,89	16,89
paragraphs	1,00	1,00	1,00
mean (in words)	101.830	101.830	101.830
std. dev.			
headings			
mean (in words)	0	0	0
std. dev.			
sections	1,00	1,00	1,00
mean (in words)	101.830	101.830	101.830
std. dev.			

Figura 35 - Lista estatística do Corpus CAF_BE

	N	Overall	1
text file		Overall	cofen.txt
file size	621.549	621.549	621.549
tokens (running words) in text	105.045	105.045	105.045
tokens used for word list	102.844	102.844	102.844
types (distinct words)	8.825	8.825	8.825
type/token ratio (TTR)	9	9	9
standardised TTR	41,68	41,68	41,68
standardised TTR std. dev.	56,88	56,88	56,88
standardised TTR basis	1.000,00	1.000,00	1.000,00
mean word length (in characters)	5	5	5
word length std. dev.	2,55	2,55	2,55
sentences	5.306,00	5.306,00	5.306,00
mean (in words)	19	19	19
std. dev.	11,20	11,20	11,20
paragraphs	1,00	1,00	1,00
mean (in words)	102.844	102.844	102.844
std. dev.			
headings			
mean (in words)	0	0	0
std. dev.			
sections	1,00	1,00	1,00
mean (in words)	102.844	102.844	102.844
std. dev.			

Figura 36 - Lista estatística do Corpus CAF_IE

3.1.3.3 Corpora de referência escritos

Os corpora de referência escritos foram compilados a partir de corpora já existentes. Em língua portuguesa utilizamos textos do Lácio Web⁵¹, tomando o cuidado de selecionar apenas textos de áreas que não tivessem relação com a cafeicultura. Assim selecionamos os textos de Sociais, Religião, Humanas, Generalidades e Saúde. O corpus de referência deve ser sempre maior que o corpus de estudo, para permitir uma melhor identificação da frequência dos termos, ou seja, maior precisão. Berber Sardinha (2004) recomenda que o corpus de referência seja cinco vezes o tamanho do corpus de estudo. Nosso corpus tem cerca de 4 milhões de palavras, como podemos observar na Figura 37. Em inglês, utilizamos corpora jornalísticos disponíveis do COMET, com cerca de um milhão de palavras (Figura 38), com os quais já havíamos trabalhado em disciplinas de Pós-graduação.

	Overall	1	2	3	4	5	6	7	8	9	
text file	Overall	des.zip	001.txt	002.txt	003.txt	004.txt	005.txt	006.txt	007.txt	008.txt	009.txt
file size	18,264,010	90,677	12,484	14,834	9,924	9,273	20,852	14,657	18,958	9,866	10,000
tokens (running words) in text	4,417,748	15,444	2,056	2,172	1,490	1,466	3,494	2,395	2,970	1,466	1,000
tokens used for word list	3,746,479	85,194	1,894	2,114	1,469	1,366	3,358	2,200	2,789	1,417	1,000
types (distinct words)	454,064	18,419	646	853	597	537	1,136	775	949	605	1,000
type/token ratio (TTR)	12	18	34	40	41	39	34	35	34	43	100
standardised TTR	46,04	45,98	35,90	49,80	44,90	41,70	44,10	43,80	45,15	47,00	37,00
standardised TTR std. dev.	53,91	53,33	36,35	36,35	36,35	36,35	42,47	38,74	38,78	38,78	38,78
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00
mean word length (in characters)	3	2	5	6	5	5	5	5	5	5	5
word length std. dev.	2,14	1,71	3,60	3,64	3,63	3,47	3,57	3,62	3,68	3,77	3,77
sentences	140,745,00	043,00	68,00	73,00	64,00	49,00	124,00	72,00	102,00	60,00	75,00
mean (in words)	27	27	28	29	23	28	27	31	27	24	24
std. dev.	25,44	28,16	12,63	18,29	14,97	15,75	10,70	18,82	16,18	11,30	11,30
paragraphs	130,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	28,819	85,194	1,894	2,114	1,469	1,366	3,358	2,200	2,789	1,417	1,000
std. dev.	168,819,25										
headings											
mean (in words)	0	0	0	0	0	0	0	0	0	0	0
std. dev.											
sections	128,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	28,289	85,194	1,894	2,114	1,469	1,366	3,358	2,200	2,789	1,417	1,000

Figura 37 - Lista estatística do corpus de referência em português

⁵¹ Disponível em <<http://www.nilc.icmc.usp.br/lacioweb/index.htm>>. Acesso 10 mai. 2006.

	N	Overall	1	2	3	4	5	6	7	8	9	10
text file		Overall	o-1.txt	o-2.txt	o-3.txt	o-4.txt	o-5.txt	o-6.txt	o-7.txt	o-8.txt	o-9.txt	o-10.txt
file size	8,051,486	3,283	8,818	7,567	15,426	8,806	16,452	11,079	8,748	12,001	5	
tokens (running words) in text	989,377	302	1,075	921	1,418	699	2,172	1,353	1,526	2,127		
tokens used for word list	940,580	290	978	890	1,392	673	2,011	1,302	1,476	1,962		
types (distinct words)	39,961	192	414	428	612	463	705	600	564	726		
type/token ratio (TTR)	4	66	42	48	44	53	35	46	38	37		
standardised TTR	46,80				48,80		42,20	48,10	42,20	44,40		
standardised TTR std. dev.	52,46						40,87					
standardised TTR basis	1,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00	0,000,00
mean word length (in characters)	5	5	5	4	5	5	4	4	4	5		
word length std. dev.	2,54	2,28	2,43	2,24	2,40	2,59	2,27	2,16	2,18	2,25		
sentences	47,296,00	17,00	47,00	58,00	89,00	37,00	102,00	103,00	66,00	75,00	36	
mean (in words)	20	17	21	15	16	24	20	13	22	26		
std. dev.	16,13	16,50	16,62	11,89	14,84	21,84	15,41	10,10	12,37	14,13	16	
paragraphs	1,166,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1	
mean (in words)	807	290	978	890	1,392	673	2,011	1,302	1,476	1,962		
std. dev.	862,32											
headings												
mean (in words)	0	0	0	0	0	0	0	0	0	0		
std. dev.												
sections	1,148,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1	
mean (in words)	819	290	978	890	1,392	673	2,011	1,302	1,476	1,962		

Figura 38 - Lista estatística do corpus de referência em inglês

3.1.3.4 Corpus de referência falado

No início da pesquisa, para o português falado no Brasil, utilizamos o corpus formado pelo material concedido pelo projeto NURC-SP, disponibilizado em CD-ROM, com textos falados de entrevistas, elocuições formais e diálogos entre dois informantes, com o total aproximado de 150.000 palavras. Os textos estavam com a extensão .pdf e foram transformados em .txt para que pudessem ser utilizados no WordSmith Tools (Scott, 2004-2007). Essa transformação, no entanto, deixou a desejar, pois o documento ficou sem acentos, cedilhas, perdeu páginas etc. Fazendo a busca por palavras-chave, pudemos observar que um corpus maior resultava em um número maior de *keywords*, além de apresentar mais variantes. Procuramos mais materiais disponíveis que pudessem servir como referência de língua falada em português, e que fosse de tamanho maior, porém não encontramos. Alguns sites disponibilizam listas de palavras, já em formato de wordlist, porém para serem usadas com a versão

3.0 do programa, que já não é disponível para compra. Apresentaremos os resultados da pesquisa com o corpus NURC, embora tenhamos utilizado o mesmo corpus de referência escrito do português, para ter acesso a um número maior de variantes. Vejamos a lista estatística do corpus NURC (Figura 39), com um total de 171.454 palavras usadas:

N	Overall	1	2	3
test file	Overall	me i.txt	se i.txt	se i.txt
file size	1 028 049	30 889	76 612	18 548
tokens (running words) in text	180 524	38 434	86 499	55 591
tokens used for word list	171 454	36 627	80 807	54 220
types (distinct words)	14 993	6 054	8 906	6 410
type/token ratio (TTR)	9	17	11	12
standardised TTR	37,46	39,51	37,30	36,33
standardised TTR std dev	60,65	57,93	69,85	62,61
standardised TTR basis	1 000,00	000,00	000,00	000,00
mean word length (in characters)	4	5	4	4
word length std dev	2,67	2,91	2,62	2,66
sentences	5 800,00	950,00	203,00	447,00
mean (in words)	31	39	25	37
std dev	47,55	71,93	33,63	51,97
paragraphs	1 888,00	73,00	757,00	58,00
mean (in words)	91	502	46	906
std dev	643,61	458,59	301,46	688,96
headings				
mean (in words)	0	0	0	0
std dev				
sections	3,00	1,00	1,00	1,00
mean (in words)	57,151	36 627	80 807	54 220
std dev	73 135,95			

Figura 39 - Lista estatística do corpus NURC

Como corpus de referência da língua inglesa, utilizamos primeiramente o MICASE, também falado, composto de mais de 2 milhões de palavras, com textos representativos de diversos discursos: acadêmico, conversas telefônicas, conversação, etc. Fazendo um teste entre o uso do MICASE como corpus de referência e o corpus jornalístico do COMET, observamos que não havia grandes variações entre o número de candidatos a termos selecionados entre os dois corpora de referência:

Quadro 10 - Comparação de corpus de referência

CORPUS	nº candidatos a termos	nº de candidatos a termos
	MICASE (ORAL)	corpus JORNALÍSTICO (ESCRITO)
CAF_ALL_IN	172	177
CAF_BI_IN	102	127
CAF_IO	111	102

Observando as listas de frequência não conseguimos identificar diferenças em relação a palavras de conteúdo ou de variantes, portanto utilizaremos também o corpus de referência escrito em inglês, para que nas duas línguas possamos fazer buscas com as mesmas diretrizes.

3.2 Explorando os corpora

Bowker e Pearson (2002) citam algumas “regras” práticas que todo aspirante a terminógrafo deve seguir para elaborar um glossário.

1. *Definir o domínio.* O tema escolhido está inserido no campo de domínio de Ciências Agrárias, subdividido em Engenharia Agrícola e Engenharia de Processamento de Produtos Agrícolas, de acordo com o CNPQ⁵², que foi subdividido por outros temas do nosso estudo. Na Figura 40 as três primeiras áreas são mencionadas pelo CNPQ. As restantes foram adicionadas de acordo com a abrangência do trabalho, contando com a ajuda do site www.coffee-ota.org. As subáreas abaixo de café, colheita e processamento, em azul, são foco de nosso trabalho. As subáreas em rosa fazem parte do domínio, porém não estão contempladas em nossa análise.

⁵² Disponível em <www.cnpq.br/areas/tabconhecimento/5.htm>. Acesso em 3 jun. 2005.

Figura 40 - Árvore de domínio para a pesquisa terminológica

2. *Consultar especialista da área para orientação sobre seleção de termos, definição dos mesmos, além do alcance de um termo comparado a outro.* Isso é feito no cotidiano do trabalho do tradutor e intérprete, visto que sempre há necessidade de pesquisas na área e conversas com especialistas, para esclarecimento de dúvidas e sugestões.
3. *Subdividir o domínio se necessário.* A subárea resultante deverá ser expandida ou restringida. Em nosso caso, subdividimos o domínio até chegarmos à colheita e ao processamento.

4. *Coletar textos para o projeto, que não sejam de usos específicos ou não-usuais.* Os textos coletados têm abrangência geral, ou seja, podem ser compreendidos e utilizados pelos diversos segmentos das áreas profissionais.
5. *Extrair termos.* Os candidatos a termo podem ser palavras simples ou complexas. Para isso utilizamos a ferramenta WordSmith Tools, com uso dos utilitários WordList e KeyWords. Nossa abordagem para a extração de termos é estatística (baseada em cálculos estatísticos de ocorrência), uma vez que a abordagem lingüística (identificação morfossintática) foi impossível, pois não há etiquetadores automáticos que consigam abranger as variantes morfológicas, como fazê (fazer), colheta (colheita), terrero (terreiro).
6. *Encontrar colocados para os candidatos a termo encontrados.* Nessa etapa utilizamos o utilitário Concord, fazendo buscas de colocados para os candidatos a termo e analisando as linhas de concordância de tais candidatos.
7. *Anotar informações sobre os termos em fichas terminológicas, que devem conter os seguintes dados: definição, área de domínio, informação gramatical, informação fraseológica, termos relacionados, notas de uso, sinônimos, antônimos, e o que mais for pertinente ao trabalho.* Para nossa pesquisa, no entanto, limitamos as anotações à definição, informações gramaticais, informações fraseológicas, notas de uso e variantes.

Barros (2004) propõe que antes mesmo desses passos, devemos definir as características tipológicas da obra e a metodologia a ser seguida, tais como objetivos da obra, público-alvo, conhecimento da área, limites da pesquisa, princípios metodológicos, organização interna da obra e cronograma de atividades.

No início do trabalho, havíamos determinado que faríamos um glossário bilíngüe bidirecional e que o apresentáramos completo ao final do estudo. Entretanto, no decorrer da pesquisa identificamos alguns percalços

que nos fizeram redefini-lo. Os corpora falados em português ficaram maiores do que os corpora falados em inglês, bem como trataram de assuntos diferentes, devido a diversos fatores, dos quais podemos citar a espontaneidade da língua falada, que leva o informante a falar sobre o que é conveniente no momento, não sobre o que o pesquisador deseja, e também às dificuldades com a transcrição, que se mostrou mais difícil em língua inglesa, pois seus representantes foram gravados em situações mais abertas, ao ar livre, ou em locais de trabalho, onde havia ruídos que impediram ou danificaram as gravações. Portanto, não conseguimos elaborar o glossário bidirecional, já que não encontramos informações suficientes nos corpora de língua falada para estabelecer os equivalentes. Quanto à apresentação final do glossário, o planejamento do tempo não foi adequado para a proposta, uma vez que, não tendo dados suficientes nos corpora em inglês falado, teríamos de buscar dados para definições, equivalentes e até mesmo variantes em outras fontes, levando, assim, mais tempo que o previsto.

Dessa forma, decidimos definir nosso produto final como um vocabulário bilíngüe português => inglês, restrito às subáreas de *colheita* e *processamento* do café, temas de nossa pesquisa. Faulstich (1995a) nos explica a definição de vocabulário: “Vocabulário multilíngüe: vocabulário que repertoria os termos acompanhados de seus equivalentes de uma ou várias línguas” (Faulstich, 1995a). Utilizaremos o direcionamento português => inglês. Quanto à determinação de termos, conforme já dito, seguimos os princípios da Lingüística de Corpus como metodologia para a coleta de dados, compilação do corpus e extração de termos com o auxílio de ferramentas computacionais. Para a determinação de variantes, consideramos a classificação dos sociotermos segundo Faulstich (2001), apresentando o vocabulário na seguinte forma:

<p>VERBETE em PB – EQUIVALENTE em EN</p> <p>VARIANTE (s) em PB (se houver)</p>

Esse vocabulário não possui definições, porém alcança o público-alvo de tradutores, intérpretes e profissionais da área, citando, conforme nossa preocupação principal, as variantes. Já para um público-alvo menos especializado, como estudantes de agronomia, por exemplo, tal glossário pode ser incompleto, pois não há definições. Para quem ainda não tem familiaridade com a área, o campo definitório é imprescindível. Por essa razão, elaboramos um protótipo do que seria o glossário ideal, trabalho que poderá ser realizado com a ampliação de nossos corpora falados.

VERBETE em português (INFORMAÇÃO GRAMATICAL): DEFINIÇÃO

NOTAS DE USO: (2 ou 3)

VARIANTES: (caso haja)

NOTAS DE USO: (2 ou 3)

COLOCADOS (todos os encontrados, não há limite, com ex. de uso):

EQUIVALENTE em inglês: (com exemplos de uso)

Essa organização interna da obra segue parte da estrutura da apresentação de verbetes do Glossário de Termos Neológicos da Economia (Alves, 2001), da qual excluímos alguns campos (sinônimos, antônimos e remissivas) e adicionamos o campo de colocados, extremamente pertinente para quem quer usar a língua com mais naturalidade e fluência (Tagnin, 2005b).

Dessa maneira, sabendo as informações a extrair, graças ao planejamento visto acima do que haverá na obra a ser confeccionada, começamos a buscar os termos com a ferramenta WordList em cada corpus.

3.2.1 Corpora falados

Ilustraremos a seguir o passo a passo da exploração dos corpora falados em ambas as línguas, utilizando as ferramentas do WordSmith Tools, versão 4.0 (Scott, 2004-2007).

3.2.1.1 Conhecendo os Corpora - Palavras mais freqüentes

As listas de palavras foram obtidas a partir do corpus coletado, transcrito e compilado. Para todas as listas de palavras utilizamos o recurso chamado *Stop List*, que consiste da eliminação de palavras com alta freqüência, geralmente gramaticais, que não terão utilidade em nosso corpus para a extração dos termos por não possuírem significado semântico. A *Stop List* pode ser alimentada pelo usuário, a seu critério. Conforme encontrávamos palavras sem significado semântico, acrescentávamos à *Stop List*, manualmente. Abaixo estão as palavras que fizeram parte de nossa *Stop List*. Nota-se que a *Stop List* contém palavras dos dois idiomas.

Figura 41 - *Stop List*

Abaixo apresentaremos uma visão das palavras que ocorreram em cada corpus confeccionado, utilizando a lista de freqüência.

A) Corpus oral CAF_BO

Observa-se na lista de frequência das palavras que ocorrem no corpus (Figura 42) que ainda há a ocorrência de palavras típicas da língua falada, tais como: vai, gente, está e coisa, mesmo usando a *Stop List*. Para eliminá-las, bastaria tê-las incluído na *Stop List*.

Além dessas palavras representantes da oralidade, há também verbos de alta frequência. As palavras de conteúdo mais frequentes nesse corpus, nesta primeira página, são: café, colheita, água, cereja, grão, verde, pé e região (Figura 43).

N	Word	Freq.	%	Texts	% Lemmas	Set
1	CAFE	857	2,17	19	100,00	
2	VAI	315	0,79	16	84,21	
3	GENTE	184	0,46	19	100,00	
4	ESTÁ	167	0,42	11	57,89	
5	TENHO	129	0,32	10	62,53	
6	COLHEITA	115	0,29	11	57,89	
7	PODE	109	0,27	15	78,95	
8	TER	97	0,24	12	63,16	
9	ÁGUA	93	0,23	10	52,63	
10	HOJE	91	0,23	11	57,89	
11	CEREJA	90	0,23	8	42,11	
12	VOU	90	0,23	8	42,11	
13	GRÃO	87	0,22	10	52,63	
14	MESMO	85	0,21	15	78,95	
15	ANO	84	0,21	16	84,21	
16	VERDE	83	0,21	13	68,42	
17	DIA	82	0,21	13	68,42	
18	FAZER	79	0,20	12	63,16	
19	FICA	76	0,19	15	78,95	
20	TIPO	76	0,19	11	57,89	
21	VOCÊS	75	0,19	8	42,11	
22	PÉ	73	0,18	11	57,89	
23	REGIÃO	72	0,18	8	42,11	

Figura 42 - CAF_BO - Lista de frequência parcial

B) Corpus oral CAF_IO

Segue-se a mesma orientação anterior para a exploração desse corpus. É interessante observar (Figura 43) que nessa primeira lista de palavras gerada já podemos contrastar o conteúdo dos textos de falantes de inglês e de português, ou seja, é nítida a maneira como uma *WordList* reflete o contexto da pesquisa. As primeiras palavras da lista, de conteúdo, são: coffee, arabica, price, farmers, country, world, crop, robusta.

N	Word	Freq	%	Texts	% Lemmas	Set
1	COFFEE	280	1,89	11	100,00	
2	LIKE	49	0,33	9	81,82	
3	KNOW	44	0,30	10	90,91	
4	ARABICA	43	0,29	6	54,55	
5	NOW	41	0,28	6	54,55	
6	GET	40	0,27	6	54,55	
7	HERE	39	0,26	6	54,55	
8	PRICE	37	0,25	2	18,18	
9	FARMERS	36	0,24	6	54,55	
10	PEOPLE	36	0,24	8	72,73	
11	COUNTRIES	30	0,20	6	54,55	
12	WORLD	30	0,20	4	36,36	
13	FIVE	29	0,20	6	54,55	
14	GOOD	29	0,20	9	81,82	
15	UP	29	0,20	7	63,64	
16	YEAR	28	0,19	7	63,64	
17	CROP	27	0,18	5	45,45	
18	ROBUSTA	27	0,18	5	45,45	
19	BETTER	26	0,18	5	45,45	
20	SEE	26	0,18	7	63,64	
21	USE	26	0,18	7	63,64	
22	T	25	0,17	8	72,73	
23	TALK	25	0,17	4	36,36	

Figura 43 - Lista de Frequência do Corpus CAF_IO

Ao montar um quadro com os substantivos mais freqüentes em inglês e português, em corpora compostos de entrevistas e conversação, obtêm-se os seguintes resultados:

Legenda:

palavras idênticas em ambos os corpora

* indica o mesmo lema

Português	Inglês
café	coffee
gente	arabica
colheita	price
dia	farmers*
região	people
tipo	countries
cereja	world
Brasil	year
qualidade	crop
produtor	robusta
secador	farmer
verde	market
bebida	production
ano	quality
coisa	area
água	birr
parte	years
máquina	money
pano	coffees*
anos	school
terreiro	time
tempo	Uganda
verdade	trade
mercado	problem

As diferenças de contexto em que foram gravadas as entrevistas e conversações ficaram claras no levantamento acima. Em português, enfatizaram-se assuntos sobre a colheita e o processamento do café. Já em inglês, o foco está em variedades de plantação (arábica, robusta), comercialização (market, trade, price, birr – moeda da Etiópia -, money). Também fica clara a diferença de menção a países: em português, Brasil, em inglês, Etiópia. Isso se explica facilmente devido à forma como se conduziu a coleta de textos. Os textos em inglês para esse corpus foram coletados durante a conferência internacional sobre café e visita a fazendas, todas na Etiópia. É comum que se tenha falado muito em valores, afinal todos os participantes de uma conferência, bem como os produtores rurais que receberam visitantes em suas terras etíopes, tinham maior interesse em o quanto um pé de café pode rentabilizar. Já no Brasil, o maior produtor mundial do café arábica, não se falou muito em variedades, mas em qualidade, processos de produção e produto.

C) Corpus de Interpretação CAF_BI

Para que o corpus de interpretação possa ser analisado com segurança, fizeram-se três listas de palavras: uma bilíngüe (Figura 44), uma em português (Figura 45) e outra em inglês (Figura 46).

N	Word	Freq	%	Texts	% Lemmas	Set
1	CAFE	132	0,84	8	88,89	
2	T	101	0,49	9	100,00	
3	COFFEE	100	0,49	9	100,00	
4	OH	95	0,46	9	100,00	
5	VAI	95	0,46	8	88,89	
6	HERE	89	0,43	9	100,00	
7	RIGHT	63	0,31	8	88,89	
8	DON	59	0,29	9	100,00	
9	LIKE	57	0,28	8	88,89	
10	DÁ	52	0,25	8	88,89	
11	SAID	51	0,25	6	55,56	
12	YEAR	42	0,20	6	66,67	
13	VINTE	40	0,20	8	88,89	
14	ANO	39	0,19	7	77,78	
15	FAZ	37	0,18	7	77,78	
16	SEE	37	0,18	7	77,78	
17	TEMPO	36	0,18	8	88,89	
18	DOIS	34	0,17	8	88,89	
19	HUNDRED	33	0,16	7	77,78	
20	MÁQUINA	33	0,16	6	66,67	
21	QUANTO	33	0,16	9	100,00	
22	TA	32	0,16	6	66,67	
23	TWENTY	27	0,13	6	66,67	

Figura 44 - Lista parcial de frequência bilingüe do corpus CAF_BI

A lista bilíngüe oferece algumas informações da composição do corpus, uma vez que alterna palavras em português e inglês pela frequência. Há muitos verbos (vai, like – do verbo ou comparação -, dá, said, see, don – da forma negativa do verbo auxiliar do presente-, faz, ta –do verbo “estar” ou marcador de fala) que marcam a condição de oralidade dos textos; quatro substantivos (café, coffee, year, ano) exatamente equivalentes, indicando a condição de tradução/interpretação dos textos; marcadores de fala (oh, right) e dêitico (here); pronome interrogativo (quanto), pois os visitantes tinham interesse em saber valores, tais como quantidades de café por área, de valores por saca, etc., cujo equivalente how much não ocorreu pois foi incluído na *Stop List*. Há também muitos números (vinte, dois, hundred, twenty), que devem se referir às quantidades, muito mencionadas no decorrer do texto, conforme dito anteriormente.

As palavras de conteúdo mais freqüentes no CAF_BI_PB são: café e máquina (Figura 45)

Figure 45 shows a screenshot of a software window titled "CAF_BI_PB.list". The window displays a table with columns: N, Word, Freq, %, Texts, % emmas, and Set. The table lists 22 words in Portuguese, sorted by frequency. The words are: CAFE, VAI, DÁ, VINTE, AND, FAZ, TEMPO, DOIS, MÁQUINA, QUANTO, MIL, GENTE, QUATRO, TA, VOCÊS, PODE, BOM, CENTO, OUTRO, PARTE, QUAL, CINCO, and O. The frequency values range from 127 for "CAFE" down to 19 for "O". The percentage values range from 1.23% for "CAFE" down to 0.19% for "O". The number of texts for each word is also listed, ranging from 8 to 9. The percentage of emmas ranges from 88.89% to 66.67%. The set column is empty for all words.

N	Word	Freq	%	Texts	% emmas	Set
1	CAFE	127	1,23	8	88,89	
2	VAI	95	0,92	8	88,89	
3	DÁ	52	0,50	6	88,89	
4	VINTE	40	0,39	8	88,89	
5	AND	39	0,38	7	77,78	
6	FAZ	36	0,35	7	77,78	
7	TEMPO	36	0,35	8	88,89	
8	DOIS	33	0,32	8	88,89	
9	MÁQUINA	33	0,32	6	66,67	
10	QUANTO	33	0,32	9	100,00	
11	MIL	30	0,29	8	88,89	
12	GENTE	28	0,27	6	66,67	
13	QUATRO	27	0,26	6	66,67	
14	TA	27	0,26	6	66,67	
15	VOCÊS	27	0,26	6	66,67	
16	PODE	26	0,25	7	77,78	
17	BOM	26	0,24	7	77,78	
18	CENTO	22	0,21	6	66,67	
19	OUTRO	22	0,21	5	55,56	
20	PARTE	22	0,21	6	55,56	
21	QUAL	21	0,20	6	66,67	
22	CINCO	20	0,19	6	66,67	
23	O	19	0,19	6	66,67	

Figura 45 - Lista parcial de frequência CAF_BI_PB

No CAF_BI_IN as palavras de conteúdo mais freqüentes são: coffee e cherries.

Figure 46 shows a screenshot of a software window titled "CAF_BI_IN.list". The window displays a table with columns: N, Word, Freq, %, Texts, % emmas, and Set. The table lists 22 words in English, sorted by frequency. The words are: I, COFFEE, HERE, OH, RIGHT, DON, LIKE, SAID, YEAR, SEE, HUNDRED, PEOPLE, TOO, TWENTY, CHERRIES, DRY, GET, GOING, GONNA, GOOD, WAY, MAKE, and I'D. The frequency values range from 101 for "I" down to 16 for "I'D". The percentage values range from 0.95% for "I" down to 0.15% for "I'D". The number of texts for each word is also listed, ranging from 9 to 8. The percentage of emmas ranges from 100.00% to 88.89%. The set column is empty for all words.

N	Word	Freq	%	Texts	% emmas	Set
1	I	101	0,95	9	100,00	
2	COFFEE	93	0,88	9	100,00	
3	HERE	86	0,81	9	100,00	
4	OH	86	0,81	9	100,00	
5	RIGHT	62	0,59	8	88,89	
6	DON	59	0,56	9	100,00	
7	LIKE	56	0,53	8	88,89	
8	SAID	51	0,48	6	55,56	
9	YEAR	41	0,39	6	66,67	
10	SEE	34	0,32	7	77,78	
11	HUNDRED	32	0,30	7	77,78	
12	PEOPLE	29	0,27	7	77,78	
13	TOO	29	0,27	6	66,67	
14	TWENTY	29	0,27	6	55,56	
15	CHERRIES	27	0,26	6	66,67	
16	DRY	27	0,26	7	77,78	
17	GET	27	0,26	7	77,78	
18	GOING	27	0,26	8	88,89	
19	GONNA	27	0,26	4	44,44	
20	GOOD	27	0,26	7	77,78	
21	WAY	27	0,26	8	88,89	
22	MAKE	26	0,25	7	77,78	
23	I'D	16	0,15	8	88,89	

Figura 46 - Lista parcial de frequência CAF_BI_IN

Uma observação a se fazer é que a primeira palavra do corpus em inglês é T, e a sexta é DON. Pesquisando nos arquivos em que apareciam, pôde-se perceber que correspondem parcialmente ao DON'T.

Por ser um corpus baseado em interpretações consecutivas ou intermitentes, ou seja, interpretações que não exigem literalidade das frases e que permitem que o intérprete adapte seu discurso à situação, não há tanta equivalência lexical ou terminológica na ordem de frequência das palavras. Uma vez que o intérprete é também profissional da área (especialista em café), muitas vezes ele responde às questões baseado em seu próprio conhecimento.

D) Corpora gerais CAF_ALL_PB e CAF_ALL_IN

Como esses corpora incluem os corpora anteriores, é óbvio que suas palavras serão as mesmas. Entretanto, devido às ocorrências, as palavras podem variar quanto à sua posição de frequência (Figuras 47 e 48).

Se fizermos uma comparação com a ocorrência de palavras de conteúdo, considerando apenas a primeira parte da lista de cada, teríamos:

Quadro 11 - Comparação das listas de frequência em português

PORTUGUÊS		
CAF_BO	CAF_BI_PB	CAF_ALL_PB
café	café	café
colheita	máquina	colheita
água		cereja
cereja		água
grão		verde
verde		máquina
pé		grão
região		

Quadro 12 - Comparação das listas de frequência em inglês

	INGLÊS	
CAF_IO	CAF_BI_IN	CAF_ALL_IN
coffee	coffee	coffee
arabica	cherries	arabica
price		
farmers		
country		
world		
crop		
robusta		

N	Word	Freq	%	Texts	% Lemmas	Set
1	CAFE	1,003	1,99	27	96,43	
2	VAI	416	0,82	26	89,29	
3	GENTE	218	0,43	26	89,29	
4	ESTÁ	172	0,34	16	53,57	
5	TENHO	141	0,28	16	57,14	
6	PODE	136	0,27	23	82,14	
7	COLHEITA	129	0,25	16	53,57	
8	ANO	123	0,24	23	82,14	
9	DÁ	107	0,21	23	82,14	
10	TER	107	0,21	16	57,14	
11	FAZ	106	0,21	26	89,29	
12	HOJE	105	0,21	19	67,86	
13	CEREJA	104	0,20	14	50,00	
14	MESMO	104	0,20	22	78,57	
15	ÁGUA	103	0,20	14	50,00	
16	VOCÊS	102	0,20	14	50,00	
17	DIA	99	0,20	19	67,86	
18	VOU	99	0,20	13	46,43	
19	VERDE	98	0,19	20	71,43	
20	MÁQUINA	95	0,19	15	53,57	
21	FICA	94	0,19	22	78,57	
22	GRÃO	93	0,18	12	42,86	
23	CAFE	91	0,18	16	64,79	

Figura 47 - Lista parcial de frequência CAF_ALL_PB

N	Word	Freq	%	Tests	% emmas	Set
1	COFFEE	370	1,54	19	100,00	
2	T	123	0,51	16	84,21	
3	HERE	122	0,51	14	73,68	
4	LIKE	104	0,43	16	84,21	
5	OH	95	0,40	16	78,95	
6	RIGHT	77	0,32	16	84,21	
7	DON	72	0,30	16	84,21	
8	YEAR	69	0,29	13	68,42	
9	GET	66	0,27	12	63,16	
10	KNOW	65	0,27	16	84,21	
11	PEOPLE	63	0,26	14	73,68	
12	SAID	61	0,25	8	42,11	
13	NOW	59	0,25	13	68,42	
14	SEE	58	0,24	13	68,42	
15	GOOD	56	0,23	16	84,21	
16	UP	55	0,23	15	78,95	
17	USE	52	0,22	13	68,42	
18	GOING	50	0,21	13	68,42	
19	THINK	48	0,20	11	57,89	
20	ARABICA	46	0,19	7	36,84	
21	FIVE	46	0,19	11	57,89	
22	THREE	46	0,19	14	73,68	
23	HINDEN	46	0,18	17	83,16	

Figura 48 - Lista parcial de freqüência CAF_ALL_IN

3.2.1.2 Conhecendo os Corpora - Candidatos a termo

Funcionando como candidatas a termo, as palavras-chave oferecem maior confiabilidade estatística, uma vez que comparam o corpus de estudo ao de referência, ou seja, a linguagem de especialidade será evidenciada, pois serão eliminadas todas as palavras de freqüência proporcionalmente similar ao corpus de língua geral (referência). Só permanecem na lista, e assim são chamadas palavras-chave, aquelas que apresentam uma diferença de ocorrência estatisticamente significativa.

Como critério de busca de palavras-chave, cruzamos a lista de palavras do corpus em estudo com a lista de palavras do corpus de referência, como ocorre na figura abaixo (Figura 49):

Figura 49 - Tela do WordSmith Tools da ferramenta KeyWords

A questão que se colocou nesse momento foi a real necessidade de analisar o CAF_BO e CAF_IN separadamente do CAF_BI, uma vez que ambos trazem resultados que podem ser analisados por conjunto de línguas. Se a WordList nos dá todas as palavras que ocorrem nos corpora, podemos fazer a análise dos corpora reunidos por conjunto de línguas. Utilizando os corpora de referência selecionados, em português o LÁCIO-WEB e em inglês o Jornalístico (COMET), conforme a Figura 25, demonstraremos os resultados de busca por palavras-chave para os corpora CAF_ALL_PB e CAF_ALL_IN.

A) Corpus CAF_ALL_PB

O primeiro resultado obtido foi de 500 candidatos a termo, limite da ferramenta utilizada (Figura 50):

N	Key word	Freq.	%	Freq.	RC	%	express	P	mmas	Set
1	CAFÉ	1.003	1,98	0			.002,67	000000		
2	VAI	416	0,82	13			.612,88	000000		
3	GENTE	218	0,43	1			.940,60	000000		
4	ESTÁ	172	0,34	0			.541,02	000000		
5	TENHO	141	0,28	0			.263,19	000000		
6	PODE	136	0,27	0			.218,38	000000		
7	COLHEITA	128	0,25	0			.146,70	000000		
8	CEREJA	104	0,20	0			.931,64	000000		
9	VOCÊS	102	0,20	1			.902,48	000000		
10	VERDE	98	0,19	0			.877,66	000000		
11	AND	123	0,24	63			.865,18	000000		
12	MÁQUINA	95	0,19	0			.851,00	000000		
13	VOU	99	0,20	5			.846,66	000000		
14	ÁGUA	103	0,20	14			.837,30	000000		
15	GRÃO	93	0,18	0			.833,08	000000		
16	FAZ	106	0,21	45			.766,62	000000		
17	HOJE	105	0,21	52			.742,39	000000		
18	FICA	94	0,19	25			.720,26	000000		
19	FAZER	92	0,18	36			.672,66	000000		
20	COISA	71	0,14	1			.625,46	000000		
21	TER	107	0,21	165			.596,70	000000		
22	PRODUTOR	70	0,14	5			.590,39	000000		
23	VANTE	74	0,15	13			.588,77	000000		

Figura 50 - Lista parcial de palavras-chave CAF_ALL_PB (lista completa Anexo VII)

Observa-se a presença de muitos verbos de apoio que não serão parte de nossa primeira seleção de candidatos a termos, pois podem não ter relação semântica com qualquer termo da área. Ainda há palavras como o pronome vocês que não foram incluídas na *Stop List*, por isso aparecem. Não podemos, entretanto, usar a intuição sem a comprovação da ocorrência contextual no corpus, o que faremos adiante, com o uso da ferramenta Concord.

Para que os candidatos a termo acima sejam validados, temos de analisá-los individualmente, uma vez que não há a etiquetagem morfossintática do texto, impossibilitando que retiremos apenas os substantivos e os adjetivos para constituírem nosso repertório terminológico. Para essa abordagem terminológica, poderemos excluir os seguintes itens, tanto no português como no inglês:

1. nomes próprios
2. numerais
3. pronomes
4. advérbios
5. marcadores de oralidade, como ah, sim, não, claro, com certeza, é verdade, etc.
6. marcadores de tempo (dia, mês, ano, hoje, etc.)
7. moeda (reais, dólares, birr, etc.)
8. conjunções e preposições
9. verbos - tomando cuidado para não eliminar adjetivos ou substantivos formados por verbos conjugados.

Alguns desses itens já haviam sido incluídos na *Stop List*, outros não. Portanto passaremos por todas as palavras do corpus, tentando encontrar os itens acima e eliminando-os com o **ZAP**, recurso do programa para excluir itens quando deletados. É preciso, porém, muito cuidado para não eliminar itens que pertençam ao nosso campo de interesse, pois o WordSmith Tools não tem a tecla “desfazer”.

Em seguida é preciso verificar se as palavras pertencem ao domínio ou não. Esse passo, no entanto, será explicado no item sobre a ferramenta Concord. Abaixo (Figura 51) encontra-se a lista de palavras-chave obtida após essa limpeza, a qual chamamos de CAF_ALL_PB_ZAP, para não confundirmos as nomenclaturas.

N	Key word	Freq	%	Freq	RC	%	express	P	emmas	Set
1	CAFÉ	1.003	1,98	0	002,67	000000				
2	COLHEITA	128	0,25	0	146,70	000000				
3	CEREJA	104	0,20	0	931,64	000000				
4	VERDE	98	0,19	0	877,88	000000				
5	MÁQUINA	95	0,19	0	851,00	000000	lâminas			
6	ÁGUA	103	0,20	14	837,30	000000				
7	GRÃO	93	0,18	0	833,08	000000	grãos			
8	PRODUTOR	70	0,14	5	590,38	000000	dutores			
9	TERREIRO	65	0,13	0	582,23	000000				
10	SECADOR	63	0,12	0	564,31	000000				
11	CHÃO	62	0,12	0	555,35	000000				
12	BEBIDA	65	0,13	5	546,32	000000				
13	REGIÃO	78	0,15	54	521,32	000000				
14	DESCASCADO	58	0,11	0	519,52	000000				
15	BOM	74	0,15	42	511,95	000000				
16	LAVOURA	53	0,10	0	474,73	000000				
17	CASCA	55	0,11	3	469,10	000000				
18	PÉ	87	0,17	177	448,68	000000	pés			
19	BÓIA	48	0,09	0	429,94	000000				
20	NATURAL	54	0,11	15	411,77	000000				
21	PÃO	46	0,09	1	402,36	000000				
22	MERCADO	48	0,09	5	396,93	000000				
23	TIPO	81	0,17	731	391,87	000000				

Figura 51 - Lista parcial de candidatos a termos em português (lista completa Anexo VIII)

Temos, no total, 209 candidatos a termo em português.

B) Corpus CAF_ALL_IN

Utilizam-se as mesmas diretrizes de construção do corpus CAF_ALL_IN já mencionadas para o português. Assim, temos a seguinte lista de candidatos a termo dos dois corpora falados, com o total de 177 palavras sem exclusões (Figura 52):

N	Key word	Freq.	%	Freq.	RC	%	Synness	P	emmas	Set
1	COFFEE	370	1,54	103			542,24	0,000000		
2	T	123	0,51	223	0,02		564,67	0,000000		
3	OH	85	0,40	67			560,60	0,000000		
4	DON	72	0,30	71			393,78	0,000000		
5	HERE	122	0,51	539	0,04		367,23	0,000000		
6	ARABICA	46	0,19	2			360,10	0,000000		
7	CHERRIES	33	0,14	1			261,20	0,000000		
8	TWENTY	40	0,17	21			249,68	0,000000		
9	HUNDRED	44	0,18	47			235,80	0,000000		
10	ROBUSTA	28	0,12	2			214,62	0,000000		
11	RIGHT	77	0,32	546	0,04		182,87	0,000000		
12	DRY	38	0,16	63			179,49	0,000000		
13	FARMER	28	0,12	13			178,47	0,000000		
14	SEVENTY	24	0,10	3			177,76	0,000000		
15	BIRR	21	0,09	0			171,93	0,000000		
16	GONNA	28	0,12	17			170,15	0,000000		
17	FARMERS	38	0,16	82			164,06	0,000000		
18	COFFEES	20	0,08	2			150,41	0,000000		
19	LL	19	0,08	2			142,41	0,000000		
20	THOUSAND	29	0,12	44			140,82	0,000000		
21	SIXTY	21	0,09	7			140,68	0,000000		
22	CROP	30	0,12	52			139,67	0,000000		
23	HECTARES	18	0,08	1			134,44	0,000000		

Figura 52 - Lista parcial de candidatos a termos em inglês (lista completa Anexo IX)

Excluindo os mesmos itens anteriormente mencionados no corpus em português, com cuidado para não excluir verbos que possam ser substantivados ou adjetivados, temos 78 candidatos em inglês, já lematizados (Figura 53):

N	Key word	Freq	%	Freq	RC	%	eyness	P	lemmas	Set
1	COFFEE	370	1,54	103			.542,24	0,000000		
2	ARABICA	46	0,19	2			360,10	0,000000		
3	CHERRIES	33	0,14	1			261,20	0,000000		
4	ROBUSTA	28	0,12	2			214,62	0,000000		
5	DRY	38	0,16	63			179,49	0,000000		
6	FARMER	26	0,12	13			176,47	0,000000	farmers	
7	COFFEES	20	0,08	2			150,41	0,000000		
8	CROP	30	0,12	52			139,67	0,000000		
9	PROCESSING	25	0,10	27			133,58	0,000000		
10	TREE	27	0,11	46			126,41	0,000000	trees	
11	HARVEST	23	0,10	24			123,98	0,000000		
12	SOIL	16	0,07	18			96,07	0,000000		
13	DRYER	13	0,05	2			94,72	0,000000		
14	KILO	11	0,05	0			90,05	0,000000	kilos	
15	QUALITY	32	0,13	183	0,01		87,26	0,000000		
16	PRODUCTION	37	0,15	281	0,02		83,69	0,000000		
17	PLANTING	14	0,06	11			80,89	0,000000		
18	PRUNING	11	0,05	2			76,96	0,000000		
19	PICKING	18	0,07	40			76,87	0,000000		
20	DRYING	11	0,05	3			76,61	0,000000		
21	BEAN	12	0,05	6			75,53	0,000000		
22	GOOD	66	0,23	784	0,06		73,41	0,000000		
78	VINT	30	0,12	217	0,03		70,73	0,000000		

Figura 53 - Lista final de candidatos a termo do corpus CAF_ALL_IN (lista completa Anexo X)

3.2.2 Conhecendo os corpora: o que dizem as linhas de concordância

O próximo passo para a extração de termos de nossos corpora é analisar as linhas de concordância, que nos dão o contexto de cada ocorrência, definindo se um candidato pode ser validado como termo ou não. Suponhamos que temos dúvidas sobre a palavra foto, que aparece como candidata a termo (Figura 54).

N	Key word	Freq.	%	Freq.	RC. %	express	P	lemmas	Set
57	DESCASCADO	58	INF	0	519,52	000000			
58	DESMUCILADO	15	INF	0	134,35	000000			
59	DESPOLPADO	24	INF	0	214,96	000000			
60	DESPOLPADOR	9	INF	0	80,61	000000			
61	DIFERENÇA	30	INF	160	106,61	000000			
62	DIFERENTE	18	INF	39	91,75	000000			
63	DINHEIRO	9	INF	2	70,22	000000			
64	DIRETO	26	INF	15	179,38	000000			
65	DURO	10	INF	1	82,66	000000			
66	ENGENHEIRO	15	INF	0	134,35	000000			
67	ENTREVISTA	10	INF	8	65,01	000000			
68	ÉPOCA	14	INF	14	86,89	000000			
69	ESPECIAL	20	INF	44	100,63	000000			
70	ESPRESSO	19	INF	0	170,17	000000			
71	EXEMPLO	46	INF	136	209,34	000000			
72	FAZENDA	42	INF	0	376,19	000000		zendas	
73	FERMENTAÇÃO	18	INF	0	161,22	000000			
74	FERRUGEM	10	INF	1	82,66	000000			
75	FLORADA	22	INF	0	197,04	000000			
76	FOLHA	20	INF	1	171,11	000000		folhas	
77	FORTE	23	INF	29	136,43	000000			
78	FOTO	12	INF	0	107,48	000000		fotes	
79	EDITO	74	INF	1	706,66	000000		fotos	

Figura 54 - Palavra FOTO encontrada na lista de palavras-chave

Pelas linhas de concordância (Figura 55) vemos que os contextos estão voltados para situações em que pessoas tiravam ou viam fotos. Por mais que pareça óbvio que a palavra não pertença ao domínio, é importante e recomendável não confiar apenas em nossa intuição. Outro exemplo será a palavra mundo. À primeira vista, não parece ter relação direta com o café. Analisando as concordâncias (Figura 56) teremos, já na primeira linha, Mundo Novo, um tipo de variedade do café arábica. Como é um colocado, poderia passar despercebido sem a análise das concordâncias.

The screenshot shows the FOTO.cnc application window. The main table displays concordance lines for the word 'foto'. The interface includes a menu bar (File, Edit, View, Compute, Settings, Windows, Help), a toolbar with icons for various functions, and a status bar at the bottom showing '12' and 'Set'.

N	Concordance	Set	Tag	Word #	t	f	os	#	os	#	os	t	f
1	vai ficar uma moedinha... Eu tenho uma foto aqui , que dá pra gente ver melhor.				4.227	392	2%	0	9%			0	
2	lavado, e aí eu vou secar. Essa foto aqui é de um temeiro (...) suspenso.				5.491	600	0%	0	6%			0	
3	(...) para resfriar. Essa aqui é uma foto do torrador de prova. O que é um				3.486	346	0%	0	5%			0	
4	grãos aqui e os verdinhos. Nessa foto estão todos vermelhinhos.				3.624	229	0%	0	0%			0	
5	Really? Yeah. (...) pra ele tirar uma foto . Eu vou pegá aqui a máquina, para aí				1.024	94	0%	0	7%			0	
6	dois dias, além da Júlia? Pode tirar foto ? Pode! Pode tirar foto, pode filmar,				2.710	207	0%	0	4%			0	
7	Júlia? Pode tirar foto? Pode! Pode tirar foto , pode filmar, pode fazer tudo que				2.716	209	3%	0	4%			0	
8	não tem luminosidade. Aqui, olha, essa foto que nós tiramos, estava em cima da				6.643	690	1%	0	2%			0	
9	de eu levar para o torrador. Essa foto que vocês estão vendo aqui são				2.432	262	9%	0	9%			0	
10	Afonso, cê vai lá pra Austrália aí é... na foto sô viu? Vamo embora ou cê que				1.974	177	8%	0	6%			0	
11	ser um dos mais bonitos... Aqui nessa foto , tenho outras fotos que devem				2.937	183	4%	0	1%			0	
12	Ele não fica meladinho, como na outra foto . Vocês se lembram que ele estava				5.465	496	0%	0	6%			0	

Figura 55 - Linhas de concordância da palavra foto

The screenshot shows the MUNDO.cnc application window. The main table displays concordance lines for the word 'mundo'. The interface includes a menu bar (File, Edit, View, Compute, Settings, Windows, Help), a toolbar with icons for various functions, and a status bar at the bottom showing '46' and 'Set'.

N	Concordance	Set	Tag	Word #	t	f	os	#	os	#	os	t	f
1	a produção de uma variedade do tipo Mundo Novo, por exemplo, foi uma coisa				4.519	233	0%	0	2%				
2	ser produzido por países do terceiro mundo né, ou países em				3.844	201	2%	0	6%				
3	região o que se mais utiliza ainda é o Mundo Novo desenvolvido na década de				4.571	236	6%	0	3%				
4	um direito cultivar dentro do nome Mundo Novo, aí então utiliza esse termo				4.605	238	2%	0	4%				
5	alguns cultivares (...), porque o Mundo Novo ele não é um cultivar bom				4.580	237	3%	0	3%				
6	ter concordância dos países do primeiro mundo né? Em compensação você pode				3.814	199	8%	0	7%				
7	Brasil. A (...) exporta a máquina para o mundo inteiro. É um processo				4.077	374	0%	0	6%				
8	que é o que o mercado comercializa no mundo inteiro. Então, não quer dizer que				2.423	229	5%	0	4%				
9	Legal! Muito bom. Ela tem clientes no mundo inteiro. - Hum, hum. É				4.139	363	6%	0	7%				
10	E tanto meu pai como eu, e como todo mundo aqui né, acho que a vida toda				3.649	191	0%	0	4%				
11	pra hoje? Vamos supor: hoje todo mundo se preocupa se vai machucar as				438	21	0%	0	9%				
12	esse termo de número, por exemplo, Mundo Novo 388/17, que é um cultivar				4.616	238	1%	0	4%				
13	é o... mundo novo... mundo... novo mundo , porque ele é melhor ou é				457	16	7%	23	6%				
14	e mais conhecido é o... mundo novo... mundo ... novo mundo, porque ele é				455	18	0%	23	6%				
15	vê lançou essa nova variedade né, o Mundo novo, acho que até por aí até o				505	20	2%	23	8%				
16	de olho no café brasileiro e hoje todo mundo tá partindo pra isso, os grandes				2.115	115	7%	0	4%				
17	participa disso ou não? Não, todo mundo participa... tanto mulher quanto				242	12	5%	0	1%				
18	disse que é o mais conhecido é o... mundo novo... mundo... novo mundo,				453	18	3%	23	6%				
19	técnicos na empresa pra assistir todo mundo né. Temos dois na empresa e um				1.066	39	5%	0	6%				
20	Éh... são diversos cultivares de Mundo Novo. Existe um cultivar que se				4.630	239	8%	0	4%				
21	é o volume do Brasil que equilibra todo mundo . Porque o Brasil dá uma... Por				677	66	0%	0	5%				
22	variedade de café... que é o café novo mundo ... é o que mais se ouve falar... e				370	14	4%	23	3%				

Figura 56 - Linhas de concordância da palavra mundo

Apresentamos, abaixo, algumas palavras que ocorrem no corpus de português. Com o uso da ferramenta Concord, analisamos suas linhas de concordância, colocados e os agrupamentos de palavras, apenas a título de ilustração das possibilidades da ferramenta para a extração terminológica. As palavras selecionadas foram as dez primeiras encontradas no corpus CAF_ALL_PB_ZAP:

1. Café

Com um total de 1.039 linhas de concordância, sabemos que é a palavra com maior frequência do corpus por ser o tema do trabalho, portanto foco principal de todas as conversações e entrevistas, bem como interpretações. Obtivemos 25 agrupamentos de palavras significativos (Figuras 57 e 58), usando a configuração de no mínimo três ocorrências com três palavras. As palavras que eram unidades de sentido incompleto foram excluídas com o ZAP. Vejamos as duas listas abaixo:

N	Cluster	Freq	Length
1	PE DE CAFÉ	37	3
2	FAZENDA DE CAFÉ	13	3
3	CAFÉ DE QUALIDADE	12	3
4	ESPECIALISTA EM CAFÉ	12	3
5	O CAFÉ CEREJA	12	3
6	CAFÉ DO CHÃO	12	3
7	O CAFÉ NATURAL	12	3
8	O CAFÉ VERDE	11	3
9	PÉS DE CAFÉ	9	3
10	BEBIDA DO CAFÉ	8	3
11	O CAFÉ ESPECIAL	8	3
12	PLANTAÇÃO DE CAFÉ	7	3
13	O CAFÉ BRASILEIRO	7	3
14	SACA DE CAFÉ	7	3
15	PRODUTOR DE CAFÉ	7	3
16	GRÃO DE CAFÉ	7	3
17	DE CAFÉ VERDE	7	3
18	LAVOURA DE CAFÉ	7	3
19	TIPO DE CAFÉ	6	3
20	CAFÉ DE VARRIÇÃO	6	3
21	CAFÉ DO PANO	6	3
22	QUALIDADE DO CAFÉ	6	3

Figura 57 - Agrupamentos de palavras para café - 1ª página

The screenshot shows the CAFÉ.cnc software window. The main table displays word clusters for the word 'café'. The table has four columns: N, Cluster, Freq, and Length. The rows are numbered 5 to 25. The bottom of the window shows a menu bar with options like 'concordâncias', 'colocações', 'plot', 'padrões', 'clusters', 'filanames', 'sources list', and 'notes'. The status bar at the bottom indicates '1.038' and 'CAFÉ CEREJA DESCASCADO'.

N	Cluster	Freq	Length
5	O CAFÉ CEREJA	12	3
6	CAFÉ DO CHÃO	12	3
7	O CAFÉ NATURAL	12	3
8	O CAFÉ VERDE	11	3
9	PÉS DE CAFÉ	9	3
10	BEBIDA DO CAFÉ	8	3
11	O CAFÉ ESPECIAL	8	3
12	PLANTAÇÃO DE CAFÉ	7	3
13	O CAFÉ BRASILEIRO	7	3
14	SACA DE CAFÉ	7	3
15	PRODUTOR DE CAFÉ	7	3
16	GRÃO DE CAFÉ	7	3
17	DE CAFÉ VERDE	7	3
18	LAVOURA DE CAFÉ	7	3
19	TIPO DE CAFÉ	6	3
20	CAFÉ DE VARRIÇÃO	6	3
21	CAFÉ DO PANO	6	3
22	QUALIDADE DO CAFÉ	6	3
23	VARIEDADE DE CAFÉ	6	3
24	CAFÉ NO BRASIL	5	3
25	CAFÉ CEREJA DESCASCADO	5	3

Figura 58 - Agrupamentos de palavras para café - 2ª página

Assim, a partir do termo café conseguimos encontrar outros termos, por sua vez complexos: pé de café, fazenda de café, café de qualidade, especialista em café, café cereja, café do chão, café de varrição, café natural, café verde, bebida do café, café especial, plantação de café, lavoura de café, café brasileiro, saca de café, produtor de café, grão de café, café do pano, qualidade do café, variedade do café, tipo de café.

Analizando também a relação de colocados, outra forma de extrair termos, somente com palavras de conteúdo, inclusive verbos, vejamos o que foi encontrado (Figura 59):

N	Word	With	relation	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	CAFÉ	CAFÉ	25,000	1,228	70	19	17	25	12	10	6	1,039	7	13	26	37	36
2	PÉ	CAFÉ	0,000	43	37	6	0	0	0	37	0	0	0	1	0	2	3
3	VERDE	CAFÉ	0,000	43	9	34	1	3	2	1	2	0	23	0	4	4	3
4	NATURAL	CAFÉ	0,000	36	5	31	2	0	0	2	1	0	20	2	2	3	4
5	CEREJA	CAFÉ	0,000	32	4	28	0	1	3	0	0	0	11	2	4	8	2
6	QUALIDADE	CAFÉ	0,000	32	11	21	0	1	3	7	0	0	0	13	5	3	0
7	CHÃO	CAFÉ	0,000	29	5	24	0	2	2	0	1	0	0	13	4	3	4
8	BEBIDA	CAFÉ	0,000	24	14	10	3	2	0	8	1	0	0	1	3	4	2
9	BÓIA	CAFÉ	0,000	20	3	17	0	2	1	0	0	0	9	2	0	3	3
10	PANO	CAFÉ	0,000	20	4	16	0	2	0	1	1	0	0	9	1	2	4
11	FAZENDA	CAFÉ	0,000	19	12	7	0	0	1	11	0	0	0	2	0	3	2
12	GRÃO	CAFÉ	0,000	18	12	6	1	1	1	9	0	0	0	3	1	1	1
13	BENEFICIAR	CAFÉ	0,000	15	8	7	0	1	1	4	2	0	0	4	1	2	0
14	DESCASCADO	CAFÉ	0,000	15	1	14	0	0	1	0	0	0	0	4	3	2	5
15	MELHOR	CAFÉ	0,000	15	5	10	0	2	0	1	2	0	2	3	2	2	1
16	TIPO	CAFÉ	0,000	15	9	6	0	2	0	7	0	0	0	1	3	1	1
17	COLHE	CAFÉ	0,000	14	5	9	1	1	0	3	0	0	0	0	1	3	5
18	COLHIDO	CAFÉ	0,000	14	0	14	0	0	0	0	0	0	5	1	6	1	1
19	FAZER	CAFÉ	0,000	14	7	7	1	1	1	2	2	0	1	2	0	2	2
20	LAVOURA	CAFÉ	0,000	14	11	3	0	0	1	10	0	0	0	1	0	2	0
21	PRODUTOR	CAFÉ	0,000	14	12	2	1	1	2	8	0	0	0	0	1	1	0
22	TERREIRO	CAFÉ	0,000	14	2	12	0	0	1	0	1	0	0	2	6	2	2
23	TOMAR	CAFÉ	0,000	14	10	4	0	1	0	3	6	0	0	1	1	2	0

Figura 59 - Lista de colocados para a palavra café

Observa-se, apenas nessa primeira página, que os verbos: beneficiar, tomar, colhe e fazer ocorrem bem próximos à palavra, ou seja, formam colocações verbais: beneficiar o café, tomar café, colher o café, fazer o café.

2. Colheita

Nas linhas de concordância para colheita (Figura 60) encontramos colheita seletiva e colheita a dedo, que significam a mesma coisa. Colheita seletiva foi o termo utilizado por um engenheiro agrônomo, enquanto colheita a dedo foi utilizada por um produtor rural, sem formação universitária.

The screenshot shows a window titled 'COLHEITA.cnc' with a menu bar (File, Edit, View, Compute, Settings, Windows, Help). Below the menu is a table of concordance results. The table has columns for line number, text, and frequency statistics. The text is a paragraph about coffee harvesting, with the word 'colheita' highlighted in blue in the original image. The frequency statistics are shown in a small table on the right of each line.

N	Concordance	Set	Tag	Word #	t	es	os	os
1	principalmente na colheita, põe-se aí na colheita, colheita de Minas, Paraná,	2.060	32	9%	23	1%		
2	cinco a dez por cento só pode iniciar a colheita, a gente, por enquanto, essa	624	22	0%	23	3%		
3	que você prepara a lavoura pro início da colheita? a lavoura antes da colheita eh...	810	24	0%	23	0%		
4	colheita do tipo Colombiana né, que é a colheita a dedo, vamos dizer assim.	2.896	153	4%	0	9%		
5	que eles usam em Minas Gerais que é colheita a dedo. A dedo, é. É essa	722	92	9%	0	8%		
6	a dedo. A dedo, é. É essa noção, é a colheita a dedo. Olha, colhi o café. O	735	94	6%	0	8%		
7	colheita seletiva uh huh que é a colheita a dedo coletando só os grãos	699	13	3%	0	1%		
8	fazer a colheita. Eles abanam, fazem a colheita, abanam o café na lavoura[?],	2.828	267	5%	0	9%		
9	sestrar alguma coisa, para a gente ver a colheita. Agora, por quê? Por causa do	1.252	126	0%	0	7%		
10	também o custo por causa da colheita agora nessa época, (...).	1.874	88	0%	0	6%		
11	quantidade de café quanto tempo leva a colheita? ah... de três a quatro meses...	675	22	0%	23	6%		
12	no caso é melhor do que a máquina na colheita? Ah, sim, é melhor que a	2.068	110	0%	0	4%		
13	tinha pouco café, para terminar a colheita. Além da produção ser menor	1.298	131	0%	0	8%		
14	estamos iniciando a mecanização da colheita. Algumas outras mecanizações	1.186	65	0%	0	4%		
15	ano o que você (...) pra trabalhar com a colheita? aqui na roça? É vámer...eu não	96	7	0%	21	2%		
16	da colheita. De que forma eu faço a colheita aqui no Brasil? Eu tenho... Nós	865	94	3%	0	2%		
17	quer chegar com isso. Logo depois da colheita, assim que eu tirei tudo do pé,	2.579	245	1%	0	6%		
18	e meia o que gosta de na... na... na... colheita assim... qual que é a parte que	235	15	2%	21	5%		
19	as pessoas que não faziam a colheita bem feita. Hoje você vê que com	627	33	6%	0	3%		
20	de empregos né... principalmente na colheita, põe-se aí na colheita a	2.055	32	5%	23	0%		
21	Totalmente mecanizado, inclusive colheita. Colheita, tudo. Tudo. E no	1.806	81	0%	0	2%		
22	e ele se viu ameaçado de perdê a colheita. com certeza né. São fatores	1.894	68	0%	0	1%		

Figura 60 - Linhas de concordância de colheita

Vejam os que o número de ocorrências de colheita a dedo é menor (Figura 61 e 62) do que o número de ocorrências de colheita seletiva (Figura 61 e 63) e foi usado somente por trabalhadores e produtores rurais, observando a fonte dos textos.

The screenshot shows the COLHEITA.cnc application window. The main table displays concordance lines between two texts. The columns are: N, Concordance, Set, Tag, Word #, t, os, #, os, #, os, t, e. The text in the 'Concordance' column is highlighted in blue.

N	Concordance	Set	Tag	Word #	t	os	#	os	#	os	t	e
1	colheita do tipo Colombiana né, que é a colheita a dedo, vamos dizer assim.			2.886	163	4%	0	9%			0	
2	que eles usam em Minas Gerais que é colheita a dedo. A dedo, é. É essa			722	92	9%	0	8%			0	
3	a dedo. A dedo, é. É essa noção, é a colheita a dedo. Olha, colhi o café. O			735	94	8%	0	6%			0	
4	colheita seletiva uh huh que é a colheita a dedo coletando só os grãos			689	13	3%	0	1%			0	
5	o cereja né. Cereja. Mas aí seria uma colheita do tipo Colombiana né, que é a			2.888	163	2%	0	9%			0	
6	seletiva né, porque na Colômbia a colheita é seletiva então muda um			1.455	55	9%	0	2%			0	
7	colheita, durante a colheita, que tipo de colheita é feita... é seletiva ou... como é			945	26	8%	23	6%			0	
8	um limite, então hoje quem pode faz colheita seletiva, faz despulpamento, faz			1.820	96	5%	0	1%			0	
9	fazer, aconselho é lógico a fazer uma colheita seletiva, que é um café melhor,			1.862	98	0%	0	3%			0	
10	Certo. E como é feita a colheita, é colheita seletiva ou não? Certo, é na			452	20	5%	0	3%			0	
11	da Pinhal aqui tem pessoas fazendo colheita seletiva uh huh que é a colheita			679	13	8%	0	1%			0	
12	saído do Brasil né, porque não tem colheita seletiva né, porque na na			1.447	55	7%	0	2%			0	
13	é na nessa região nós não fazemos colheita seletiva. Então na verdade nós			466	21	2%	0	4%			0	
14	da figura. Mas no Brasil não existe a colheita seletiva e eles vão se diferenciá			1.470	66	4%	0	3%			0	

Figura 61 - Linhas de concordância de colheita a dedo e colheita seletiva

The screenshot shows the COLHEITA.cnc application window. The main table displays a list of clusters. The columns are: N, Cluster, Freq, Length. The cluster 'COLHEITA A DEDO' is highlighted in blue.

N	Cluster	Freq	Length
1	DEPOIS DA COLHEITA	8	3
2	É A COLHEITA	6	3
3	A COLHEITA QUE	5	3
4	É FEITA A	4	3
5	FEITA A COLHEITA	4	3
6	A COLHEITA NO	4	3
7	A COLHEITA A	4	3
8	COLHEITA A DEDO	4	3
9	MECANIZAÇÃO DA COLHEITA	3	3
10	EPOCA DA COLHEITA	3	3
11	QUE É O	3	3
12	LOGO DEPOIS DA	3	3
13	COM A COLHEITA	3	3
14	A COLHEITA É	3	3
15	DA COLHEITA EM	3	3
16	FAZER A COLHEITA	3	3
17	ENTÃO A COLHEITA	3	3

Figura 62 - Colheita a dedo - Quatro ocorrências

N	Word	With	relation	Total	tal Left	tal Right	L5	L4	L3	L2	L1	Centre
1	COLHEITA	COLHEITA	0,000	147	7	12	2	3	1	1	0	123
2	A	COLHEITA	0,000	99	72	27	4	3	10	4	51	0
3	É	COLHEITA	0,000	50	24	26	3	8	4	7	2	0
4	QUE	COLHEITA	0,000	46	17	29	6	3	6	2	0	0
5	DA	COLHEITA	0,000	36	27	8	0	0	2	1	24	0
6	DE	COLHEITA	0,000	33	15	18	4	3	2	0	6	0
7	NA	COLHEITA	0,000	33	21	12	2	1	2	2	14	0
8	O	COLHEITA	0,000	27	14	13	5	0	8	1	0	0
9	E	COLHEITA	0,000	23	7	16	1	2	2	2	0	0
10	NO	COLHEITA	0,000	20	6	14	2	2	2	0	0	0
11	NÃO	COLHEITA	0,000	17	7	10	1	1	3	2	0	0
12	ENTÃO	COLHEITA	0,000	16	9	7	3	0	3	3	0	0
13	PORQUE	COLHEITA	0,000	13	6	7	4	0	2	0	0	0
14	NÊ	COLHEITA	0,000	13	5	8	1	2	2	0	0	0
15	UMA	COLHEITA	0,000	13	9	4	1	0	1	1	6	0
16	EU	COLHEITA	0,000	12	2	10	0	0	2	0	0	0
17	EM	COLHEITA	0,000	12	5	7	1	0	4	0	0	0
18	VOCÊ	COLHEITA	0,000	11	7	4	1	3	3	0	0	0
19	SELETIVA	COLHEITA	0,000	10	0	10	0	0	0	0	0	0
20	MAIS	COLHEITA	0,000	10	3	7	0	2	1	0	0	0
21	POR	COLHEITA	0,000	10	2	8	1	0	1	0	0	0
22	PRA	COLHEITA	0,000	9	7	2	1	2	2	0	2	0

Figura 63 - Colheita seletiva - 10 ocorrências

3. Cereja

Observamos, pelos Clusters de três palavras e pelas linhas de concordância, que é possível usarmos o cereja descascado, o café cereja, café cereja descascado e o cereja (Figura 64 e 65).

4. Água

Segundo especialistas, a palavra água não poderia ser termo, pois não é específica do tema café, mas um elemento que integra a natureza e vai interagir com a planta ou bebida. Analisando as ocorrências (103) nas linhas de concordância (Figura 66), percebemos que em diversos momentos o elemento água foi usado como parte do processamento do café, como no caso em que o café é lavado ou irrigado. Porém, usaremos apenas o colocado tanque com água (Figura 67) em nosso glossário, por fazer parte do processo de fermentação do café.

The screenshot shows a concordance search tool window titled 'ÁGUA.cnc'. The window has a menu bar (File, Edit, View, Compute, Settings, Windows, Help) and a toolbar. The main area displays a list of concordance lines, numbered 1 to 22. Each line shows a snippet of text with the word 'água' highlighted in blue. To the right of each line, there are statistics: 'Set', 'Tag', 'Word #', 't', 'os', 's', 'os', 't'. Below the list, there are buttons for 'concordance', 'collocates', 'plot', 'patterns', 'clusters', 'filenames', 'sources list', and 'notes'. At the bottom, there is a status bar showing '57' and a search bar with the text 'a disco pu pu lavador de café. Então o lavador nosso é muito simples, é com água, com clareza que o grão boba, a semente aboba e vai levar'.

Nº	Concordance	Set	Tag	Word #	t	os	s	os	t
1	maduro e o verde ele afunda na água...então esse é o cereja...ai cai	1.254	97.8%	0.5%					
2	Uh huh. O café dá mais... mais sem água. Então eu penso (...) Já aumentou.	381	30.0%	0.1%					
3	(...) They want bigger. (...) eu jogo a água naquele reservatório a hora que ele	521	39.8%	0.9%					
4	soil will act as a filter, it'll drain. E a água do despulpado é rica de potássio.	596	47.4%	0.3%					
5	ao lado das plantações pra pegar água e aqui no Brasil ele não viu.	1.690	136.6%	0.0%					
6	coffee will die. No Vietnam se não tiver água o café morre. But it is nice in Brazil	1.976	168.7%	0.7%					
7	water in the soil. Ah... o solo guarda a água, ele conserva a água? Ele tem	2.089	173.9%	0.9%					
8	solo guarda a água, ele conserva a água? Ele tem assim muita umidade né?	2.043	173.0%	0.9%					
9	may be spoiled. Ele achava que se a água caísse dentro do botão, ia estragar	3.067	274.6%	0.4%					
10	o (...) não tem uma gota d'água, com dez, onze metros de altura e	894	80.4%	0.4%					
11	Errou! Cerveja? Emul (...) Água? Whisky? Água (...) Água? Água.	124	12.0%	0.3%					
12	Água, claro! O (...) essencial (...). A água é a bebida mais vendável dos	154	20.0%	0.4%					
13	é café e noventa e oito por cento é água? É água noventa e oito por cento! E	1.127	111.0%	0.6%					
14	e noventa e oito por cento é água? É água noventa e oito por cento! E aí eu te	1.129	112.3%	0.6%					
15	Dai você pensa assim: — Como na água? Mas eu não tinha que fazer café.	1.163	114.0%	0.7%					
16	— É! Com certeza! Como é que a água em São Paulo (...) e... (...) tomar	1.308	129.8%	0.1%					
17	água em São Paulo (...) e... (...) tomar água, não é? É que é um pouquinho	1.316	129.8%	0.1%					
18	sobre uma determinada pressão, e a água que passa sobre o café que está	1.488	139.3%	0.5%					
19	café de qualidade inferior, a pressão da água sobre aqueles grãos, vai sair tudo	1.556	141.7%	0.7%					
20	é onde o café passa para... (...) café, a água passa espremida sobre o bolo de	3.069	246.9%	0.2%					
21	a outra coisa que eu (...) para vocês da água (...) do café. Por quê? Imaginem	100	4.0%	0.3%					
22	experiência. botar num alodão com água dentro de casa. (...) da natureza.	404	31.0%	0.1%					

Figura 66 - Linhas de concordância de água

The screenshot shows the AQUA.cnc software window. The main table displays word clusters for the word 'água'. The table has columns for N, Cluster, Freq, and Length. The clusters are listed in the 'Cluster' column, and their frequency and length are in the 'Freq' and 'Length' columns respectively. The 'N' column contains numbers from 1 to 10.

N	Cluster	Freq	Length
1	TANQUE COM ÁGUA	7	3
2	E A ÁGUA	4	3
3	UM TANQUE COM	4	3
4	QUE FICOU POR	3	3
5	VAMOS DIZER ASSIM	3	3
6	POR CIMA DA	3	3
7	CIMA DA ÁGUA	3	3
8	A ÁGUA ELE	3	3
9	FICOU POR CIMA	3	3
10	ESSA ÁGUA FORA	3	3

At the bottom of the window, there is a status bar with the text: '103 Set a disco pu pu lavador de café. Então o lavador nesse é muito simples, é com água, com densidade que o grão boba, a seteira abassa e vai levar'.

Figura 67 - Agrupamentos de palavras para água

5. Verde

The screenshot shows the VERDE.cnc software window. The main table displays concordance lines for the word 'verde'. The table has columns for N, Concordance, Set, Tag, Word #, t, s, os, #, ss, #, os, t, s, os. The concordance lines are listed in the 'Concordance' column, and their frequency and length are in the 't, s, os' and '#, ss, #, os' columns respectively. The 'N' column contains numbers from 1 to 22.

N	Concordance	Set	Tag	Word #	t	s	os	#	ss	#	os	t	s	os
1	que é da cor cereja mesmo. O café verde ele não consegue ser despulpado			605	35	5%	0	4%						
2	num viu lá, isso aqui tira folha do café verde , ele num viu isso ontem, que num			291	16	1%	0	5%						
3	o café no mundo inteiro: café verde , sem tomar, sem deixar (...). Só			437	58	4%	0	0%						
4	na verdade quando você seca o café verde junto com o café cereja ou o café			1.875	71	4%	0	5%						
5	é chamado no mundo inteiro de café verde . É a forma que se comercializa o			424	57	0%	0	0%						
6	casca, e é seco dessa forma. E o café verde e cereja, ele segue pro			783	34	1%	0	3%						
7	se chama via natural? café verde café verde ? ele fica quanto tempo? Ah ele fica			257	8	0%	0	2%						
8	ele é beneficiado para vender. E o café verde , é o fruto que não amadureceu no			2.449	232	8%	0	4%						
9	adstringente, não é (...). Porque o café verde para vender, comercializar, é esse			2.488	235	9%	0	4%						
10	café verde. Quando a gente fala café verde (...)? Olha... Virou uma confusão,			2.379	225	0%	0	3%						
11	Mas também é chamado de café verde , porque o café, depois que você			401	56	2%	0	9%						
12	Nossal (...). Tô ensacando o café verde porque eu não tenho tulinha. Ah, tá,			1.063	109	1%	0	9%						
13	distintas. Chama-se de café verde o café depois de beneficiado, (...),			2.408	229	4%	0	3%						
14	o seguinte: eles deixavam o café verde no terreiro, sozinho num cantinho			1.834	70	9%	0	4%						
15	fica virando. Sei. Você pode per o café verde molhado que ele não pára. Ele tem			663	60	4%	0	3%						
16	pra você produzir um café natural café verde pra produção do cereja			399	11	1%	0	6%						
17	para isso. Se compra também café verde . Quando a gente fala café verde			2.371	224	0%	0	3%						
18	outro, vai guardar o café verde . Vou deixar tudo separado. Você			6.489	671	0%	0	0%						
19	a casca, fora isso é melhor (...) de café verde , café beneficiado, (...), que é esse			554	39	6%	0	5%						
20	esses grãos que vocês viram aí do café verde , antes de beneficiar. O que vai			2.535	265	1%	0	2%						
21	aquela lavada, o café maduro, e o café verde , tudo junto. O café maduro e o			3.875	366	0%	0	4%						
22	como é que se chama via natural? café verde café verde? ele fica quanto tempo?			253	8	6%	0	2%						

At the bottom of the window, there is a status bar with the text: '97 Set despulpar, onde só é despulpado o café cereja que é da cor cereja mesmo. O café verde ele não consegue ser despulpado justamente por ele não'.

Figura 68 - Linhas de concordância de verde - Parcial

Os agrupamentos de palavras, obtidos pelo Clusters, que se apresentam mais freqüentes são: café verde, o verde, grão verde. Separa o verde e tirar o verde são colocações verbais, enquanto o maduro e o verde, o bóia, cereja e verde, o cereja e o verde são binômios e trinômios. É interessante que verde está ao final das colocações (Figuras 68 e 69).

N	Cluster	Freq	Length
1	O CAFÉ VERDE	11	3
2	E O VERDE	9	3
3	O GRÃO VERDE	4	3
4	MADURO E O	3	3
5	O MADURO E	3	3
6	O VERDE É	2	3
7	QUANTO MENOS VERDE	2	3
8	VERDE BOA SEPARAR	2	3
9	VIA NATURAL CAFÉ	2	3
10	VERDE PRA PRETO	2	3
11	SEPARA O VERDE	2	3
12	TIRAR O VERDE	2	3
13	SEPARAÇÃO DO VERDE	2	3
14	DE VERDE PRA	2	3
15	CAFÉ MADURO E	2	3
16	BÓIA CEREJA E	2	3
17	CEREJA E VERDE	2	3
18	CAFÉ VERDE É	2	3
19	O VERDE DO	2	3
20	ESSE GRÃO VERDE	2	3
21	GRÃO VERDE AQUELE	2	3

Figura 69 - Agrupamentos de palavras para verde

6. Máquina

Nos agrupamentos de palavras para máquina, o que chamou a atenção foi o verbo passar, como podemos observar nas linhas de concordância abaixo, usado diversas vezes (Figuras 70 e 71). Também temos ocorrências com tipo de máquina e máquina de café, que se refere ao local onde se leva o café para ser beneficiado e vendido.

The screenshot shows the MAQUINA.cnc application window. The main area displays a list of concordance lines for the word 'máquina'. Each line is numbered and includes the word and its frequency. The bottom of the window shows a toolbar with various icons and a status bar.

N	Concordance	Set	Tag	Word #	f	os	# os	# os
55	(Não. São as três.) Ah! - Mas essa máquina eletrônica nem todas as				1.925	221	4%	0.7%
56	nem pelos italianos, que montaram a máquina de café espresso. Por isso que				3.112	247	1%	0.3%
57	e a máquina não pára. Se o motor da máquina pifa? Pára. Pára a máquina				1.495	138	8%	0.3%
58	ele. Eu vou tirar essa parte de novo na máquina, tira só isso aqui ó... só isso				1.585	126	5%	0.8%
59	isso aí, então implica numa outra máquina junto a esse lavador, porque se				2.605	133	9%	0.1%
60	curso básico para você poder operar a máquina do espresso, poder preparar				1.978	164	8%	0.6%
61	o motor da máquina pifa? Pára. Pára a máquina. Então não compensa. (...) a				1.504	140	0%	0.4%
62	onze por cento a hora que ele passa na máquina pra limpá ele quebra... daí o				1.748	29	3%	23.6%
63	ele vai cá... ele cá e vai passa a máquina então pra roçá o mato (..)				1.342	101	8%	0.0%
64	na hora em que está passando na máquina do beneficiamento, ele quebra.				3.019	305	9%	0.3%
65	Ja. Faz seis anos que tá passando a máquina. Nesse mesmo café? Nesse				1.020	70	0%	0.3%
66	caso a máquina, mesmo passando a máquina, tem pessoas que não pagam				2.095	112	8%	0.6%
67	e olenta e quatro... eu passo na máquina de novo, tira esse café (...)				1.517	123	5%	0.4%
68	do café é feito lógico por uma máquina né, o esse beneficiamento				2.078	76	8%	0.1%
69	eu faço com eles? Eu passo por uma máquina, desço também ele pela água,				3.894	358	5%	0.4%
70	ele pela água, e passo por uma máquina, antes de eu levar para secar,				3.904	358	1%	0.4%
71	eu vou beneficiar esse café, porque a máquina tem que fazer (...). Vocês				2.978	302	1%	0.2%
72	o pengaminho e o grão. Quando a máquina aperta ele, tirou casca, vai				4.184	369	1%	0.6%
73	caso... Mas não tão bem quando uma máquina? Ou depende? De lavadores				2.014	102	0%	0.2%
74	A gente faz idéia (...). Será que a máquina tá funcionando agora? Cá				129	6	0%	0.7%
75	a seguinte: a partir do momento que a máquina vai passando, aquele café já				2.698	108	6%	0.6%
76	consegue fazer a colheita? Que essa máquina vai colé? É. Mais ou menos				940	77	5%	0.2%

Figura 70 - Linhas de concordância de máquina

The screenshot shows the MAQUINA.cnc application window. The main area displays a table of word clusters for the word 'máquina'. The table has columns for N, Cluster, Freq, and Length. The bottom of the window shows a toolbar with various icons and a status bar.

N	Cluster	Freq	Length
1	TIPO DE MÁQUINA	4	3
2	NENHUM TIPO DE	3	3
3	PASSANDO A MÁQUINA	2	3
4	PASSO TUDO NA	2	3
5	PASSO POR UMA	2	3
6	MÁQUINA PIFA PÁRA	2	3
7	MÁQUINA DE CAFÉ	2	3
8	MÁQUINA MESMO PASSANDO	2	3

Figura 71 - Agrupamentos de palavras para máquina

7. Grão

A palavra grão pode ser observada nas linhas de concordâncias abaixo (Figuras 72 e 73):

N	Concordance	Set	Tag	Word #	t	os	#	os	#	os	t
1	maior percentagem de grão seco... de grão bôia... por que porque você visando			531	12	1%	0	4%			
2	é com água, com densidade que o grão bôia, a sujeira abaixo e vai levando			1.240	64	8%	0	5%			
3	da noite a planta está dando seu grão de café e não precisa. Só com o			1.907	122	9%	0	3%			
4	grãos do café, e a breca que come o grão de café - a gente estava			2.555	166	6%	0	0%			
5	de falar para vocês é assim: cada grão de café contém lipídio - cada grão			3.066	308	8%	0	4%			
6	this part. Oh, ok. Por exemplo, esse grão de café aqui ele vai dá dezesseis			3.769	347	2%	0	5%			
7	ter uma nova forma de qualidade desse grão, de que forma tem que ser feito?			1.015	106	2%	0	4%			
8	no grão. Fica preto o grão. Eu vi o grão de café... Você já viu,			5.753	520	1%	0	0%			
9	assim a maior quantidade de de grão de maior qualidade uh huh e a			563	12	9%	0	6%			
10	são as camadas que compõem o grão de café. - (...) ? Quase sempre a			190	13	9%	0	5%			
11	grão de café contém lipídio - cada grão de café, hein, gente? (...) gordura			3.073	308	5%	0	5%			
12	(...) da natureza, não é? Quando sai o grão do chão, (...). Então na verdade a			416	32	3%	0	1%			
13	Então, se eu for olhar para o grão de café, para o fruto do café, eu vou			73	3	1%	0	2%			
14	um... (...) grande, que ela aperta o grão de café; (...) o café. Conforme ela			3.922	368	7%	0	4%			
15	o café, tem separação assim por grão de bôia... de verde? Não, aqui num			231	22	2%	0	1%			
16	quanto do Robusta é possível eu ter o grão Moca. O Moca é só a denominação			576	64	3%	0	6%			
17	num. Acaba acontecendo isso com o grão Moca. Olha aqui a variedade do			717	78	8%	0	0%			
18	fazer uma comparação, (...), eu tenho o grão natural, o café natural, que foi			4.498	418	0%	0	2%			
19	para ver bem essa condição, mesmo o grão preto, gente, é a segunda florada.			528	72	7%	0	3%			
20	vinte em vinte minutos só pra não dar grão preto, pra continuá na cor certa.			1.033	84	4%	0	8%			
21	que tem uma maior percentagem de grão seco... de grão bôia... por que			528	12	9%	0	4%			
22	café no dia seguinte, porque eu tenho grão seco, tenho grão vermelho e tenho			3.061	269	7%	0	3%			
23	de novo. Isso tem que ser bem... o grão verde, ele tipo, ele cozinha			1.673	29	7%	23	5%			
24	o café bem maduro, porque senão esse grão verde, na hora que ele seca, ele fica			581	22	7%	23	1%			
25	cereja descascado, eu já preparei... O grão verde já foi cuspidô lá, pela			4.895	444	0%	0	8%			
26	grão seco, tenho grão vermelho e tenho grão verde. Aquela grão que está			3.068	269	6%	0	3%			
27	Ricardo. Ricardo, nada se perde. O grão verde: aquele grãozinho que caiu no			2.301	217	0%	0	2%			
28	eu disse, nada se perde. Eu pego esse grão verde e também levo para secar no			4.921	447	4%	0	6%			
29	pela máquina. É o que eu faço com o grão verde? Como eu disse, nada se			4.910	445	0%	0	6%			
30	porque eu tenho grão seco, tenho grão vermelho e tenho grão verde.			3.064	269	9%	0	3%			
31	que acontece é o seguinte, o café é um grão vivo Uh huh. tá... Então até ele			430	23	6%	0	2%			

Figura 72 - Linhas de concordância da palavra grão (1ª página)

N	Concordance	Set	Tag	Word #	t	os	#	os	#	os	t
11	grão de café contém lipídio - cada grão de café, hein, gente? (...) gordura			3.073	308	5%	0	5%			
12	(...) da natureza, não é? Quando sai o grão do chão, (...). Então na verdade a			416	32	3%	0	1%			
13	Então, se eu for olhar para o grão de café, para o fruto do café, eu vou			73	3	1%	0	2%			
14	um... (...) grande, que ela aperta o grão de café; (...) o café. Conforme ela			3.922	368	7%	0	4%			
15	o café, tem separação assim por grão de bôia... de verde? Não, aqui num			231	22	2%	0	1%			
16	quanto do Robusta é possível eu ter o grão Moca. O Moca é só a denominação			576	64	3%	0	6%			
17	num. Acaba acontecendo isso com o grão Moca. Olha aqui a variedade do			717	78	8%	0	0%			
18	fazer uma comparação, (...), eu tenho o grão natural, o café natural, que foi			4.498	418	0%	0	2%			
19	para ver bem essa condição, mesmo o grão preto, gente, é a segunda florada.			528	72	7%	0	3%			
20	vinte em vinte minutos só pra não dar grão preto, pra continuá na cor certa.			1.033	84	4%	0	8%			
21	que tem uma maior percentagem de grão seco... de grão bôia... por que			528	12	9%	0	4%			
22	café no dia seguinte, porque eu tenho grão seco, tenho grão vermelho e tenho			3.061	269	7%	0	3%			
23	de novo. Isso tem que ser bem... o grão verde, ele tipo, ele cozinha			1.673	29	7%	23	5%			
24	o café bem maduro, porque senão esse grão verde, na hora que ele seca, ele fica			581	22	7%	23	1%			
25	cereja descascado, eu já preparei... O grão verde já foi cuspidô lá, pela			4.895	444	0%	0	8%			
26	grão seco, tenho grão vermelho e tenho grão verde. Aquela grão que está			3.068	269	6%	0	3%			
27	Ricardo. Ricardo, nada se perde. O grão verde: aquele grãozinho que caiu no			2.301	217	0%	0	2%			
28	eu disse, nada se perde. Eu pego esse grão verde e também levo para secar no			4.921	447	4%	0	6%			
29	pela máquina. É o que eu faço com o grão verde? Como eu disse, nada se			4.910	445	0%	0	6%			
30	porque eu tenho grão seco, tenho grão vermelho e tenho grão verde.			3.064	269	9%	0	3%			
31	que acontece é o seguinte, o café é um grão vivo Uh huh. tá... Então até ele			430	23	6%	0	2%			

Figura 73 - Linhas de concordância da palavra grão (2ª página)

As linhas de concordância demonstram grande ocorrência de colocados para grão, tais como grão bóia, grão de café, grão do chão, grão do café, grão Moca, grão natural, grão seco, grão verde e grão vermelho.

Ao observarmos os agrupamentos de palavras, percebemos que há a ocorrência de casca, polpa, pergaminho e grão, 4 vezes (Figura 74).

N	Cluster	Freq	Length
1	GRÃO DE CAFÉ	7	3
2	PERGAMINHO E O	4	3
3	O GRÃO DO	4	3
4	O GRÃO VERDE	4	3
5	E O GRÃO	4	3
6	O GRÃO ENTÃO	3	3
7	UM GRÃO ESTÁ	3	3
8	POLPA PERGAMINHO E	3	3
9	O GRÃO FICA	3	3
10	GRÃO SÓ ISSO	3	3
11	DO GRÃO SÓ	3	3
12	CASCA.POLPA.PERGAMINHO	3	3
13	O GRÃO DE	3	3
14	NO GRÃO FORAM	3	3
15	GRÃO VERDE AQUELE	3	3

Figura 74 - Agrupamento de palavras para grão

8. Pé

As linhas de concordância alinhadas à direita (Figura 75) demonstram que pé de café é um termo complexo, com 37 ocorrências (Figura 76). Ainda observando as linhas de concordância (Figura 77), encontramos na segunda casa à esquerda o colocado sec*: secando no pé, secar no pé, seco no pé, secou no pé.

N	Concordance	Set	Tag	Word #	f	os	f	os	f	os
1	foi secando como uma uva passa, no pé, eu treli tudo, coloquei na água,				4.511	418	0%	0	2%	
2	e cinco centímetro lenço de tóido do pé, aí ela brota tudo de novo... daí				3.685	339	6%	0	3%	
3	entendeu? (...) que ela anasta no pé ali, né? É. Ela dormia um pouco,				1.533	119	3%	0	0%	
4	este ano, ele secou muito rápido no pé, amadureceu muito rápido por causa				1.311	132	4%	0	8%	
5	em termos de colheita pra retirar do pé apenas o cereja né. Cereja. Mas aí				2.878	161	0%	0	8%	
6	on the... the (...) Todo mundo deixa no pé assim? É comum? É comum porque				864	63	8%	0	1%	
7	a mais, deixar um tempinho aí no pé, até os outros todos chegarem ao				1.972	190	8%	0	7%	
8	eu aproveito e (...) na terra (...) no pé de café. O pessoal até brinca na				2.862	176	5%	0	9%	
9	isso não é saudável para a planta. Um pé de café, que é induzido assim,				3.391	210	8%	0	3%	
10	e tempo para dar a flor. A florada no pé de café acontece mais ou menos em				3.034	169	6%	0	4%	
11	encerrada. Como acontece a florada no pé de café? Aparece com os ramos,				3.008	187	8%	0	3%	
12	da broca e também uma importante no pé de café. Então bicho-mineiro, broca e				2.825	173	4%	0	8%	
13	no pé de café. No ano seguinte, esse pé de café não vai ter produção de				2.674	170	5%	0	4%	
14	só e um paizinho, uma varzinha no pé de café. No ano seguinte, esse pé de				2.667	169	3%	0	3%	
15	ano. No ano seguinte, quando o pé de café estiver todo bonito, ela vai ter				2.601	173	1%	0	7%	
16	se a broca está aí e vem naquele pé de café e ele tirou todas as folhas ela				2.771	172	6%	0	6%	
17	natureza e não tem milagre. Então (...) pé de café, é um dos canônes mais				3.444	213	8%	0	5%	
18	o pé de café, e vai passar através do pé de café, vai chacoalhar os grãos, e				1.678	164	0%	0	3%	
19	se fosse um caminhão, que (...) o pé de café, e vai passar através do pé de				1.670	164	2%	0	3%	
20	(...), sendo que a... (...) sempre e meu pé de café, eu vou ter catagão para só				2.102	199	7%	0	9%	
21	um monte de pazinha, que ela entra no pé de café assim. - E daí onde? O?				1.739	170	3%	0	4%	
22	na similes maneira dele de cortar o pé de café ou sei lá né, fazê o... alguma				2.731	113	8%	0	8%	

Figura 75 - Linhas de concordância para pé - Parcial

N	Cluster	Freq	Length
1	PE DE CAFÉ	37	3
2	DO PÉ DE	8	3
3	O PÉ DE	7	3
4	NO PÉ DE	7	3
5	DE CAFÉ E	7	3
6	POR QUÊ QUE	5	3
7	PÉ POR QUÊ	4	3
8	UM PÉ DE	4	3
9	QUÊ QUE TEM	3	3
10	PEQUENO IGUAL ESSE	3	3
11	UM PÉ PEQUENO	3	3
12	QUE TEM PÉ	3	3
13	LÁ NO PÉ	3	3
14	PÉ PEQUENO IGUAL	3	3
15	O MÁXIMO POSSÍVEL	3	3

Figura 76 - Agrupamento de palavras para pé

Line	Text	Set	Tag	Word #	t	os	#	os	#	os	lt
63	um senhor tirando mato ao redor do pé de café. Hoje não temos aí toda a			2.011	129	6%	0	5%			
64	em termos de colheita pra retirar do pé apenas o cereja né. Cereja. Mas aí			2.828	161	0%	0	8%			
65	espera esse café estar secado no pé, sem fermentar nada, nem nada.			1.432	143	3%	0	0%			
66	não posso deixar esse fruto secar no pé, ele nunca vai secar. Ele vai			1.465	145	0%	0	1%			
67	eu só consigo fazer natural seco no pé e terminado de secar no tamoio, se			4.799	436	0%	0	6%			
68	Agora, quando ele já está seco no pé, o risco é muito menor. Então, eu só			4.785	435	4%	0	6%			
69	o prelinho, escolhidos, que já estão no pé. Tudo aqui, em cima desse pano,			2.655	251	0%	0	7%			
70	no pé de café. No ano seguinte, esse pé de café não vai ter produção de			2.674	170	5%	0	4%			
71	de mato porque o mato ia sufocar o pé de café, tanto que não é verdade, ele			2.064	133	0%	0	7%			
72	já viu o Maragogipe? Não. Eu tenho um pé um pé numa das lavouras daqui. E se			4.786	254	4%	0	7%			
73	eu aproveito e (...) na terra (...) no pé de café. O pessoal até brinca na			2.862	176	5%	0	8%			
74	o que que é denegar. Denegar é tirar do pé e jogar no pano. E... não tem muito			1.080	57	7%	0	6%			
75	depois da colheita, ele corta todo o pé de café, tira todas as folhas e fica só o			2.651	169	6%	0	3%			
76	exemplo, está granado, e está todo no pé. Ele está amarelo, porque é a			734	80	0%	0	0%			
77	e cinco centímetro longe do tokdo do pé, aí ele brota tudo de novo... daí			3.665	339	6%	0	3%			
78	da colheita, assim que eu tirei tudo do pé, eu tenho a abanção. E até hoje ela			2.586	245	5%	0	6%			
79	ai as bama produz, entende? Uh-huh. Pé de café desse, pequeno assim,			344	25	2%	0	1%			
80	só é um paucinho, uma varzinha no pé de café. No ano seguinte, esse pé de			2.667	169	3%	0	3%			
81	café... não produz... agora das vezes um pé pequeno igual esse é um café que			230	11	9%	0	4%			
82	se a breca está aí e vem naquele pé de café e ele tirou todas as folhas ele			2.771	172	6%	0	6%			
83	café... não produz... agora das vezes um pé pequeno igual esse até Está faltando			479	15	5%	0	6%			
84	café... não produz... agora das vezes um pé pequeno igual esse até Está faltando			614	24	5%	0	9%			

Figura 77 - Linhas de concordância da palavra pé - Parcial

9. Tipo

Apesar de ter número alto de ocorrências como candidata a termo, a palavra tipo não foi considerada válida pelo especialista por ser usada para nomear outros conceitos. Nas linhas de concordância (Figura 78) encontramos tipo de bebida, tipo de café, tipo de colheita, tipo de serviço, tipo de máquina. Segundo o especialista, seriam termos as palavras que denominassem os tipos de bebida, de café, de colheita, de serviço e de máquina.

The screenshot shows a concordance tool window titled 'TPO.cnc'. The main window displays a list of lines of concordance for the word 'tipo'. The table has columns for 'Set', 'Tag', 'Word #', 't', 'os', 's', 'os', 'lt'. The text in the concordance lines is as follows:

Set	Tag	Word #	t	os	s	os	lt
1		613	31	2%	0	7%	
2		2.483	89	5%	0	3%	
3		2.797	102	3%	0	3%	
4		4.043	397	7%	0	8%	
5		126	6	9%	0	5%	
6		2.150	32	8%	23	4%	
7		1.869	212	0%	0	5%	
8		1.286	163	3%	0	1%	
9		3.563	349	3%	0	7%	
10		2.095	236	0%	0	1%	
11		1.899	216	8%	0	6%	
12		2.196	119	6%	0	6%	
13		1.768	29	1%	23	9%	
14		2.208	119	5%	0	8%	
15		3.478	174	3%	0	1%	
16		466	12	3%	0	1%	
17		2.820	103	0%	0	3%	
18		4.703	246	3%	0	6%	
19		2.402	33	8%	23	4%	
20		339	12	9%	23	1%	
21		1.529	123	3%	0	5%	
22		3.724	346	6%	0	2%	

At the bottom of the window, there is a status bar showing the current line and the text: '84 Set que fazer duas ou três colheitas do mesmo café então a gente não pratica esse tipo, porque a gente não sabe se ainda é rendável ou não. Então:'.

Figura 78 - Linhas de concordância para a palavra tipo - Parcial

10. Região

A palavra região é seguida pelo colocado montanhosa (Figura 79) e refere-se a locais onde normalmente há plantações de café. Apesar de termos excluído nomes próprios que se referiam a lugares, tais como Alta Mogiana, Sul de Minas, Etiópia, Brasil etc., região montanhosa não especifica uma única localidade e nem é substantivo próprio, tornando-se um candidato válido a termo. Como termo simples, região significa a micro-área onde há o plantio do café, segundo o especialista e conforme podemos observar pelas linhas de concordância (Figura 80).

Nº	Concordance	Set	Tag	Word #	t	os	as	os	lt
1	cafeira aqui dessa parte, ela é uma região bem montanhosa... que é... por			155	8	3%	23	4%	
2	que é... por exemplo, uma região bem montanhosa às vezes a			163	8	5%	23	4%	
3	montanhosas... essa parte aqui... a região cafeeira aqui dessa parte, ela é			147	8	2%	23	4%	
4	produtores tipo da Zona da Mata, da Região da Zona da Mata, que ainda é...			841	35	8%	0	5%	
5	pessoas ah... principalmente na região da Pinhal aqui tem pessoas			672	13	5%	0	1%	
6	é cooperativa dos cafeicultores da região da Pinhal, né, então tem um			1.920	31	5%	23	5%	
7	né? Yes. Why? Por que? Porque a região de pedra é bom pa café. He Said			2.314	212	4%	0	0%	
8	né... mas a região não montanhosa, a região do cerrado, por aí já existe			1.124	27	7%	23	3%	
9	se eu tiver micro-clima pra isso, uma região favorável. Onde eu tenho isso? A			4.613	436	6%	0	7%	
10	região favorável. Onde eu tenho isso? A região Mogiana é muito favorável pra			4.620	438	0%	0	7%	
11	tipo de máquina tanto porque em região montanhosa não existe por			1.284	66	6%	0	7%	
12	plana, montanhosa... como é? É uma região montanhosa. Uh-huh... e como			145	8	6%	0	6%	
13	depois da colheita é... então é uma região montanhosa, o acesso não é			636	33	0%	0	6%	
14	como dizem, é o café onde a... da região montanhosa a gente produz um			186	8	8%	23	5%	
15	problema com a colheita por ser uma região montanhosa ou não? O problema			160	9	3%	0	7%	
16	lugares... em relação a [...] colheita pra região montanhosa a única colheita que			493	24	6%	0	2%	
17	não apanha menos por ser uma região montanhosa? Não é mais difícil			377	18	6%	0	7%	
18	que é acostumada a colher em região montanhosa então a gente se			189	10	6%	0	6%	
19	física do apanhador por ser uma região montanhosa: tem que ser mais			224	11	4%	0	0%	
20	o você dá assistência é na região, nessa região, na Alta Mogiana ou no Sul de			1.692	70	9%	0	7%	
21	colheiteira de de café né... mas a região não montanhosa, a região do			1.120	27	6%	23	3%	
22	no lugar onde você dá assistência é na região, nessa região, na Alta Mogiana ou			1.690	70	2%	0	7%	

Figura 79 - Linhas de concordância de região - Parcial

Nº	Concordance	Set	Tag	Word #	t	os	as	os	lt
30	que o senhor costuma usar? Bom, na região o que se mais utiliza ainda é o			4.562	236	2%	0	3%	
31	certeza, e tenho certeza também na região. Em oferecer não produtos mas			2.696	112	0%	0	7%	
32	você fala diversas vezes aqui na região, da região. Nós estamos na divisa			3.154	119	8%	0	3%	
33	da estôia do que já aconteceu na região e também da plantação como é			110	3	7%	0	2%	
34	É gosto ou é preço melhor? Ca vê na região nossa é quase tudo mata, né?			2.036	211	7%	0	5%	
35	no lugar onde você dá assistência é na região, nessa região, na Alta Mogiana ou			1.690	70	2%	0	7%	
36	o café que... que você produz? (...) na região nossa é o duro. É que existe			2.163	117	0%	0	7%	
37	com o cereja. Então a secagem na região é feita basicamente só em terreno			1.194	43	5%	0	5%	
38	você dá assistência é na região, nessa região, na Alta Mogiana ou no Sul de			1.692	70	9%	0	7%	
39	(...), um pouco mais para lá, nessa região (...). Se nós falamos (...) de café,			1.030	63	1%	0	8%	
40	que mais já se inicia a colheita, nessa região aqui sim a pra essa quantidade			660	22	8%	23	4%	
41	também atua, o zinco né porque nossa região é muito deficiente de zinco, e aí			656	22	6%	0	4%	
42	basicamente a cultura do café, nossa região nós estamos no município que é			212	8	5%	0	8%	
43	existia uma doença primária na nossa região, é a ferrugem do café. Uh-huh.			680	22	2%	0	5%	
44	aplicado tudo bem só que na nossa região por causa de ser pequenas			2.200	94	9%	0	9%	
45	muito importante. Mais é... na nossa região por ser uma região de clima bem			1.544	60	3%	0	6%	
46	uma carência muito grande na nossa região do quê? Informações e boas			967	36	2%	0	6%	
47	hoje, é... mesmo porque na nossa região, a quantidade de pés de café em			844	39	4%	0	7%	
48	né. Principalmente agora na nossa região como está aumentando muito a			1.665	68	4%	0	5%	
49	muito caprichada nós temos na nossa região condições de produzir café			3.068	169	9%	0	2%	
50	bebida por dois motivos: ele é na nossa região, todo café de melhor qualidade a			2.764	146	3%	0	6%	
51	de São Paulo na verdade a nossa região é tratada como a Mogiana, certo?			3.265	125	5%	0	6%	

Figura 80 - Linhas de concordância de região - Parcial

Dessa maneira, procuramos fazer uma análise cuidadosa de todos os candidatos a termo na língua portuguesa, inclusive daqueles que nossa intuição nos levava a crer não terem relação semântica com a área.

Já para a língua inglesa, não fizemos a análise com Concord para extrairmos candidatos a termos, mas para analisarmos os possíveis equivalentes e as possíveis variantes. As diretrizes são similares às aplicadas à extração terminológica, com a diferença de já sabermos o conceito que buscamos. Tagnin (2007) expõe diversas maneiras de se chegar aos equivalentes, enfatizando o uso de concordâncias. A partir do momento que se têm os colocados dos termos, é possível encontrar formas de se chegar até seu equivalente. Por exemplo, se quisermos saber qual é o equivalente de terreiro, procuramos por seus colocados. Encontramos: secar no terreiro. Vejamos a concordância:

N Concordance

3 cima da árvore e ponho para secar no terreiro. Ótimo. O segundo é o...? (...)

Pensamos na palavra dry, mas no corpus em língua falada não encontramos equivalente. Buscamos então no corpus de língua escrita e tivemos os seguintes resultados, com a solução: patio.

N Concordance

17 simply spread out on huge surfaces to dry in the sun. In order to prevent the

N Concordance

27 mucilage can be sent to the patios to dry as pulped natural coffees or can be

N Concordance

28 sun. The coffee is then shifted onto the dry portion of the patio, and the section

N Concordance

63 after harvest is placed in the sun to dry on tables or in thin layers on patios.

N	Concordance	Set	Tag	Word #	t	es	os	as	t	os	c
1	sun Raking coffee beans on a drying patio in the sun Once the beans are	26.801	316	5%	0	1%					0.1
2	floaters are usually sent directly to the patio to be dried and are often slated for	27.988	383	2%	0	5%					0.6
3	that regularly occurs during drying on a patio in humid conditions (Wigley). For	28.525	410	9%	0	7%					0.7
4	then shifted onto the dry portion of the patio, and the section where it was	28.765	424	6%	0	7%					0.7
5	to the old piles. Drying coffee solely by patio takes 6-7 days for washed coffees,	28.828	427	4%	0	8%					0.8
6	coffee beans are typically dried on a patio until they reach a moisture content	28.858	428	4%	0	8%					0.8
7	They are also more efficient than the patio drying technique because their	29.314	459	9%	0	9%					0.9
8	rainfall is too frequent for patio drying to occur, artificial drying	32.945	651	0%	0	0%					0.0

Figura 81 - Lista de concordâncias da palavra patio (lado esquerdo)

N	Set	Tag	Word #	t	es	os	as	t	os	File	%
1			26.801	316	5%	0	1%			escritorio\cofen.txt	27%
2			27.988	383	2%	0	5%			escritorio\cofen.txt	28%
3			28.525	410	9%	0	7%			escritorio\cofen.txt	28%
4			28.765	424	6%	0	7%			escritorio\cofen.txt	28%
5			28.828	427	4%	0	8%			escritorio\cofen.txt	28%
6			28.858	428	4%	0	8%			escritorio\cofen.txt	28%
7			29.314	459	9%	0	9%			escritorio\cofen.txt	29%
8			32.945	651	0%	0	0%			escritorio\cofen.txt	32%

Figura 82 - Lista de concordância da palavra patio (lado direito)

Ao buscarmos os textos de origem para o candidato, descobrimos que todas as ocorrências pertenciam ao mesmo texto e, segundo as diretrizes seguidas, os termos deveriam aparecer pelo menos em dois textos diferentes, de diferentes autores. Assim, tentamos encontrar o mesmo candidato, porém no plural, e os resultados foram:

PATIOS,enc

File Edit View Compute Settings Windows Help

N	Concordance	Set	Tag	Word #	t	os	#	as	#	os	t	e
1	sun, either on large concrete or brick patios or on mailing raised to waist			6,453	279	5%	0	5%				0
2	flat concrete or brick areas, known as patios , or on tables made of fine-mesh			7,288	316	5%	0	7%				0
3	and green cherries can be sent to the patios to be dried using the natural			27,969	364	1%	0	4%				0
4	slippery mucilage can be sent to the patios to dry as pulped natural coffees or			28,266	398	1%	0	5%				0
5	tanks, the beans are moved to drying patios and dried to 11-12% moisture			28,563	412	1%	0	6%				0
6	Coffee is typically dried on large patios made of asphalt or cement and			28,702	421	1%	0	6%				0
7	dryers. The coffee on the drying patios is shifted every 30-40 minutes and			28,719	422	1%	0	6%				0
8	Burlap bags should be aired on the patios before storing coffee to prevent a			30,502	618	2%	0	0%				0
9	most often spread out in rows on large patios where it needs to be raked every			43,398	232	0%	0	0%				0
10	sun to dry on tables or in thin layers on patios . It will take between ten days and			43,522	238	0%	0	0%				0

concordance collocations plot patterns clusters filename source text notes

10 Set

Iniciar

Figura 83 - Lista de concordâncias para a palavra patios (lado esquerdo)

Encontramos, inclusive, uma definição em inglês, ao buscarmos no texto:

“The sun-drying is done on extensive flat concrete or brick areas, known as patios, or on tables made of fine-mesh wire netting.”

“A secagem ao sol é feita em áreas de concreto planas e compridas, conhecidas como terreiros, ou em terreiros suspensos feitos de telas estreitas de arame”.

O mesmo procedimento pode ser seguido para buscarmos variantes.

Observamos nas linhas de concordância de secador, que aparecia na lista de colocados para terreiro, que havia uma variante, terreirão (linha 50, Figura 84), mas não estava entre as palavras-chave de CAF_ALL_PB.

N	Concordance	Set	Tag	Word #	t	os	#	os	#	os	it
31	muchã depois de uns dias e põe no secaador direto. E... eu não entendi direito	1.061	56	5%	0	7%					
32	Uh-huh. E nas suas propriedades tem secaador vertical ou horizontal? Nas	2.680	142	0%	0	4%					
33	É pra cá que vem. É depois do secaador ... So after the dryers that's the	3.444	357	0%	0	9%					
34	é o... (...). Uh hu. Aqui vem depois do secaador ? É pra cá que vem.	3.425	354	0%	0	8%					
35	after that? E depois do secaador ? Ai do secaador vem pro armazém. From the	2.159	228	0%	0	7%					
36	dryer, (...) Pal, você tem a... sorry... o secaador , um...sh...um...aquela que faz o	2.421	208	5%	0	6%					
37	Então, ela faz o desmucilado, e põe no secaador . Quem é que compra? Europeu	5.905	636	0%	0	2%					
38	para eu secar esse café. Imaginem um secaador do tamanho de uma sala, aqui	3.638	342	0%	0	0%					
39	ou no terreiro, com a luz solar, ou no secaador , que atinge quantos por cento	973	125	0%	0	4%					
40	then. Yeah. And after that? E depois do secaador ? Ai do secaador vem pro	2.153	227	0%	0	7%					
41	o caracol depois ela volta. É estilo um secaador . This is new and they're called	1.795	183	0%	0	6%					
42	Seis? Até aí tudo bem, mais pó... (...) secaador aqui, ó. Uh huh. Tem a rosca,	1.681	166	6%	0	1%					
43	no queimador... ele deixou cinco dias o secaador só viu no ar. No relativo? é	527	26	3%	0	7%					
44	(...) aonde? Lá dentro. Lá no secaador ? Vai ser alto esse elevador,	2.087	214	0%	0	6%					
45	À tarde... amanhã faz leite e faz o secaador . Leite é como... lá né? Lá, é,	2.094	216	0%	0	6%					
46	o mesmo tempo. Depois que vai pro secaador . Isso. Quem vê isso acha que	1.493	125	0%	0	6%					
47	don't use the drier? Vocês não usam o secaador ? Eu vou mostrar. Ah, they use	719	66	0%	0	2%					
48	esse café no terreiro para secar, ou no secaador , até atingir 11% de umidade,	6.093	547	0%	0	4%					
49	e o secaador . Você não fica só no secaador . Você olha o secaador , põe lenha	1.020	100	0%	0	2%					
50	alguém. Entre o terreiro e o secaador . Você não fica só no secaador .	1.014	99	0%	0	1%					
51	fica na horizontal? Ai depois ele sai do secaador e vai pra onde? Direto pra tuiha.	716	68	4%	0	7%					
52	não fica só no secaador . Você olha o secaador , põe lenha nele, você não	1.024	101	3%	0	2%					

Figura 84 - Linha de concordância onde aparece terreirão

Buscamos na WordList CAF_ALL_PB e encontramos a palavra, que porém apareceu apenas em um texto, com sete ocorrências.

N	Word	Freq	%	Texts	% emmas	Set
696	PINHAL	7	0,01	6	21,43	
697	PLANTIO	7	0,01	3	10,71	
698	PODEM	7	0,01	4	14,29	
699	PONTOS	7	0,01	2	7,14	
700	PRÁTICA	7	0,01	2	7,14	
701	PROCESSADO	7	0,01	5	17,86	
702	PROVAR	7	0,01	3	10,71	
703	PUXA	7	0,01	5	17,86	
704	ROÇA	7	0,01	6	21,43	
705	SAIR	7	0,01	3	10,71	
706	SUPOR	7	0,01	7	25,00	
707	TANÉM	7	0,01	5	17,86	
708	TERIA	7	0,01	3	10,71	
709	TERREIRÃO	7	0,01	1	3,57	
710	TIRÁ	7	0,01	5	17,86	
711	TIRANDO	7	0,01	4	14,29	
712	TURMA	7	0,01	3	10,71	
713	ÚNICA	7	0,01	4	14,29	
714	VALE	7	0,01	3	10,71	
715	VENDA	7	0,01	7	25,00	
716	VEREM	7	0,01	3	10,71	
717	VIDA	7	0,01	5	17,86	
718	VISÃO	7	0,01	3	10,71	

Figura 85 - Lista de frequência CAF_ALL_PB

Como cada texto pode ser representativo de mais de um falante, buscamos o texto no qual ocorreu, clicando sobre a palavra na lista, em seguida clicando no **COMPUTE** da barra de ferramentas e depois em **CONCORDANCE**. O programa responde com as linhas de concordância de terreirão. A partir daí, vemos o número do texto em que ocorre:

N	Concordance	Set	Tag	Word #	t	#	es	#	os	#	as	#	c
1	É secar o café. Eh, o café chega no terreirão, eh... eu espero, vou ajudar o			118	5	3%		0	0%				0 0
2	parte no caminhão, a gente põe no terreirão, tira o que vem no saca, tira, e			224	8	8%		0	9%				0 9
3	É. Consegue tocar uma média ali no terreirão de trezentos a quinhentos			1.003	98	7%		0	3%				0 3
4	a quinhentos alqueires. Entre o terreirão e o secador. Você não fica só			1.011	99	7%		0	4%				0 4
5	lenha de novo, e volta e faz o senão do terreirão. Certo. Uma pessoa consegue			1.051	102	0%		0	8%				0 6
6	consegue fazer muita senão no terreirão. Entendi. Tem alguma coisa			1.074	104	0%		0	9%				0 9
7	da manhã. Certo. O maior trabalho no terreirão é isso. Entendi. Tá ótimo			1.208	119	0%		0	0%				0 0

Figura 86 - Linhas de concordância de terreirão (lado esquerdo)

N	Set	Tag	Word #	t	A	es	.	A	os	#	as	t	A	os	File	%
1			118	5	3%	0	0%	0	0%	0	0%	0	0%	0%	txt\ter02pb.txt	11%
2			224	8	8%	0	9%	0	9%	0	9%	0	9%	0%	txt\ter02pb.txt	19%
3			1.003	98	7%	0	3%	0	3%	0	3%	0	3%	0%	txt\ter02pb.txt	60%
4			1.011	99	7%	0	4%	0	4%	0	4%	0	4%	0%	txt\ter02pb.txt	61%
5			1.061	102	0%	0	8%	0	8%	0	8%	0	8%	0%	txt\ter02pb.txt	64%
6			1.074	104	0%	0	9%	0	9%	0	9%	0	9%	0%	txt\ter02pb.txt	66%
7			1.208	119	0%	0	0%	0	0%	0	0%	0	0%	0%	txt\ter02pb.txt	96%

Figura 87 - Linhas de concordância de terreirão (lado direito)

Buscando pelo texto, descobrimos que todas as ocorrências são referentes a um único falante de nosso corpus, um trabalhador rural. Consultando o especialista da área, confirmamos a possibilidade de terreirão ser variante. Concluímos, naquele momento da pesquisa, que é necessário ter um número maior de textos (apesar de termos em nosso corpus da mesma tipologia textual cinco trabalhadores rurais) para que possamos visualizar mais variantes. Decidimos, então, fazer a busca da palavra na lista gerada pela KeyWords de língua escrita, cujo corpus é maior (100.000), e tampouco a encontramos. Para certificar-nos de que a palavra seria variante apenas da oralidade, procuramos por ela no Google⁵³, fazendo uma busca refinada com “café” e encontramos 1.200 ocorrências de terreirão. Em seguida procuramos terreiro, cruzando com “café”, e encontramos 181.000, demonstrando que a palavra terreirão tem uso limitado: não aparece em um corpus de 100.000 palavras e aparece pouco no Google (0,66% de ocorrência em relação à palavra de uso – terreiro). Ou seja, é possível considerar terreirão como variante, segundo todas as análises feitas, porém

⁵³ Disponível em <www.google.com.br>. Acesso em 3 mar. 2002.

será incluída na ficha terminológica com um asterisco, para indicar uma possibilidade, mesmo que não possamos comprovar sua ocorrência pelos meios definidos como metodologia.

As 209 palavras do corpus CAF_ALL_PB foram assim analisadas, para que pudéssemos selecionar os termos do vocabulário apresentado no Capítulo IV, bem como seus colocados e variantes.

3.2.3 Corpora escritos

Os corpora escritos, CAF_BO e CAF_IO, foram analisados seguindo diretrizes semelhantes aos corpora falados. Apresentaremos as listas de palavras mais freqüentes, seguidas dos candidatos a termos e, por fim, as análises de linhas de concordância dos candidatos.

3.2.3.1 Conhecendo o corpus - Palavras mais freqüentes

Para a obtenção de ocorrências de maior freqüência nos corpora de língua inglesa não fizemos nenhuma configuração especial no WordSmith Tools. Vejamos parte das listas de palavras obtidas para cada corpus:

A) Corpus escrito CAF_BE

N	Word	Freq	%	Texts	% Lemmas	Set
1	CAFÉ	1.783	1,69	1	100,00	
2	SECAGEM	475	0,45	1	100,00	
3	QUALIDADE	388	0,37	1	100,00	
4	COLHEITA	338	0,32	1	100,00	
5	CAFÉS	278	0,26	1	100,00	
6	GRÃOS	280	0,25	1	100,00	
7	PRODUTO	240	0,23	1	100,00	
8	PRODUÇÃO	233	0,22	1	100,00	
9	TERREIRO	197	0,19	1	100,00	
10	MAIOR	177	0,17	1	100,00	
11	FRUTOS	175	0,17	1	100,00	
12	BEBIDA	165	0,16	1	100,00	
13	BRASIL	165	0,16	1	100,00	
14	AR	164	0,16	1	100,00	
15	UNIDADE	158	0,15	1	100,00	
16	SEMENTES	149	0,14	1	100,00	
17	ÁGUA	145	0,14	1	100,00	
18	PROCESSO	144	0,14	1	100,00	
19	PRODUTORES	144	0,14	1	100,00	
20	SISTEMA	131	0,12	1	100,00	
21	TIPO	129	0,12	1	100,00	
22	SECADOR	128	0,12	1	100,00	
23	CEDEIRA	126	0,12	1	100,00	

Figura 88 - Lista de palavras CAF_BE

Podemos observar que a palavra de conteúdo de maior frequência no corpus acima é café, seguida por secagem. Como são corpora escritos, é comum que as primeiras palavras sejam aquelas oriundas da palavra usada na coleta dos textos, ou seja, as palavras que nos levaram aos textos que compõem os corpora. Entretanto, já podemos observar quais são as palavras mais comuns à área, e até inferir quais seriam os possíveis candidatos a termo, ou seja, as palavras que vêm logo abaixo, como qualidade, colheita, grãos, produto, terreiro, etc.

B) Corpus escrito CAF_IE

N	Word	Freq.	%	Texts	% lemmas	Set
1	COFFEE	3.433	3,27	1	100,00	
2	BEANS	547	0,62	1	100,00	
3	COFFEES	389	0,36	1	100,00	
4	WATER	276	0,26	1	100,00	
5	ESPRESSO	268	0,26	1	100,00	
6	QUALITY	261	0,25	1	100,00	
7	CHERRIES	231	0,22	1	100,00	
8	BEAN	215	0,20	1	100,00	
9	ROAST	188	0,18	1	100,00	
10	GREEN	187	0,18	1	100,00	
11	ARABICA	186	0,18	1	100,00	
12	FLAVOR	176	0,17	1	100,00	
13	CUP	175	0,17	1	100,00	
14	METHOD	171	0,16	1	100,00	
15	PROCESS	167	0,16	1	100,00	
16	HARVESTING	165	0,16	1	100,00	
17	WORLD	162	0,15	1	100,00	
18	HIGH	155	0,15	1	100,00	
19	TIME	155	0,15	1	100,00	
20	DRYING	154	0,15	1	100,00	
21	DRY	152	0,14	1	100,00	
22	OVER	147	0,14	1	100,00	
23	MANY	143	0,14	1	100,00	

Figura 89 - Lista de palavras do CAF_IE

Nessa lista, é interessante percebermos que a palavra de maior frequência é coffee, assim como no português, e na terceira posição está coffees. Apesar de terem o mesmo lema, não é possível uni-las, uma vez que não sabemos se fazem parte de contextos diferentes, o que poderá ser observado nas linhas de concordância. Para dar um exemplo de como ocorre a lematização, unimos beans e bean, fazendo com que, na busca terminológica, representem um único termo. Para que se possa ter a contagem geral das palavras lematizadas, é possível unir as duas formas clicando sobre a palavra e levando-a com o mouse para cima da outra (Figura 90).

N	Word	Freq	%	Texts	%lemmas	Set
1	COFFEE	3.433	3,27	1	100,00	
2	BEANS	762	0,62	1	100,00	an[215]
3	COFFEES	369	0,35	1	100,00	
4	WATER	276	0,26	1	100,00	
5	ESPRESSO	268	0,26	1	100,00	
6	QUALITY	261	0,25	1	100,00	
7	CHERRIES	231	0,22	1	100,00	
8	BEAN	215	0,20	1	100,00	
9	ROAST	188	0,18	1	100,00	
10	GREEN	187	0,18	1	100,00	
11	ARABICA	186	0,18	1	100,00	
12	FLAVOR	176	0,17	1	100,00	
13	CUP	175	0,17	1	100,00	
14	METHOD	171	0,16	1	100,00	
15	PROCESS	167	0,16	1	100,00	
16	HARVESTING	165	0,16	1	100,00	
17	WORLD	162	0,15	1	100,00	
18	HIGH	155	0,15	1	100,00	
19	TIME	155	0,15	1	100,00	
20	DRYING	154	0,15	1	100,00	
21	DRY	152	0,14	1	100,00	
22	OVER	147	0,14	1	100,00	
733	MANY	1,03	0,14	1	100,00	

Figura 90 - Lista de frequência CAF_IE: lematização de bean/beans

Examinando todas as listas de palavras relacionadas acima, dos corpora falados e dos escritos, já podemos observar que não será difícil encontrar termos e equivalentes para nosso vocabulário, pois eles estão evidentes até mesmo nas listas de palavras.

3.2.3.2 Conhecendo o corpus: candidatos a termo

Para encontrarmos candidatos a termo, utilizamos as listas de palavras dos corpora CAF_BE e CAF_IE e cruzamos com as listas de palavras dos corpora de referência do Lácio-Web (português) e o Jornalístico do COMET (inglês). Os resultados serão apresentados a seguir.

A) Corpus escrito CAF_BE

N	Key word	Freq.	%	Freq.	R.C. %	Gyness	P	eramas	Set
1	CAFÉ	1.783	1,69	0		440,48	0,000000		
2	SECAGEM	475	0,45	0		674,80	0,000000		
3	COLHEITA	338	0,32	0		543,32	0,000000		
4	QUALIDADE	388	0,37	115		384,29	0,000000		
5	CAFÉS	278	0,26	0		691,88	0,000000		
6	GRÃOS	260	0,25	3		923,55	0,000000		
7	PRODUTO	240	0,23	25		641,26	0,000000		
8	TERREIRO	197	0,19	0		482,08	0,000000		
9	PRODUÇÃO	233	0,22	83		392,99	0,000000		
10	FRUTOS	175	0,17	3		286,24	0,000000		
11	BEBIDA	165	0,16	5		196,42	0,000000		
12	UMIDADE	156	0,15	0		173,58	0,000000		
13	SEMENTES	149	0,14	2		699,73	0,000000		
14	PRODUTORES	144	0,14	3		654,14	0,000000		
15	BRASIL	165	0,16	46		622,17	0,000000		
16	ÁGUA	145	0,14	14		996,70	0,000000		
17	SECADOR	128	0,12	0		962,90	0,000000		
18	CEREJA	125	0,12	0		940,33	0,000000		
19	MERCADO	125	0,12	5		898,16	0,000000		
20	CHÃO	106	0,10	0		797,38	0,000000		
21	TEMPERATURA	113	0,11	10		781,16	0,000000		
22	MINAS	103	0,10	1		763,68	0,000000		
23	MANUAL	101	0,10	0		759,76	0,000000		

Figura 91 - Candidatos a termo em português - Lista parcial

Os candidatos a termo com conteúdo significativo encontrados acima são: café, secagem, colheita, qualidade, cafés, grãos, produto, terreiro, produção, frutos, bebida, umidade, sementes, produtores, água, secador, cereja, mercado, chão, temperatura e manual. Excluimos Brasil e Minas, por se tratarem de nomes próprios referentes a localidades. São 21 candidatos a termo na primeira página.

Usando os mesmos critérios para exclusões como dos corpora falados, temos ao total 396 candidatos a termo em português, língua escrita.

N	Key word	Freq.	%	Freq.	RC. %	eyness	P	emmas	Set
1	CAFÉ	1.783	1,89	0		440,46	0,000000		
2	SECAGEM	475	0,45	0		574,80	0,000000		
3	COLHEITA	338	0,32	0		643,32	0,000000		
4	QUALIDADE	388	0,37	115		384,28	0,000000		
5	CAFÉS	278	0,26	0		691,68	0,000000		
6	GRÃOS	260	0,25	3		923,66	0,000000		
7	PRODUTO	240	0,23	25		641,26	0,000000		
8	TERREIRO	197	0,19	0		482,09	0,000000		
9	PRODUÇÃO	233	0,22	83		382,99	0,000000		
10	FRUTOS	175	0,17	3		285,24	0,000000		
11	BEBIDA	165	0,16	5		195,42	0,000000		
12	UMIDADE	156	0,15	0		173,68	0,000000		
13	SEMENTES	149	0,14	2		699,73	0,000000		
14	PRODUTORES	144	0,14	3		654,14	0,000000		
15	ÁGUA	145	0,14	14		995,70	0,000000		
16	SECADOR	128	0,12	0		962,90	0,000000		
17	CEREJA	125	0,12	0		940,29	0,000000		
18	MERCADO	125	0,12	5		898,18	0,000000		
19	CHÃO	106	0,10	0		797,38	0,000000		
20	TEMPERATURA	113	0,11	10		781,16	0,000000		
21	MANUAL	101	0,10	0		759,76	0,000000		
22	PROCESSO	144	0,14	110		740,90	0,000000		
23	CAFÉ	101	0,10	5		740,16	0,000000		

Figura 92 - Candidatos a termo final - Lista parcial (lista completa Anexo XI)

B) Corpus escrito CAF_IE

N	Key word	Freq.	%	Freq.	RC. %	eyness	P	emmas	Set
1	COFFEE	3.433	3,27	109		634,32	0,000000		
2	BEANS	647	0,62	9		836,00	0,000000		
3	COFFEES	389	0,36	2		948,53	0,000000		
4	ESPRESSO	289	0,26	1		419,77	0,000000		
5	CHERRIES	231	0,22	1		222,18	0,000000		
6	BEAN	215	0,20	6		695,10	0,000000		
7	ROAST	188	0,18	2		983,07	0,000000		
8	ARABICA	186	0,18	2		972,41	0,000000		
9	FLAVOR	176	0,17	10		864,32	0,000000		
10	HARVESTING	165	0,16	3		852,32	0,000000		
11	QUALITY	261	0,25	183	0,01	819,78	0,000000		
12	DRYING	154	0,15	3		793,92	0,000000		
13	METHOD	171	0,16	19		759,26	0,000000		
14	ROASTING	141	0,13	0		763,67	0,000000		
15	CAFFEINE	139	0,13	0		737,63	0,000000		
16	WATER	276	0,26	291	0,02	731,55	0,000000		
17	ROASTED	143	0,14	5		721,37	0,000000		
18	ROBUSTA	132	0,13	2		685,05	0,000000		
19	ACIDITY	130	0,12	1		683,26	0,000000		
20	GRIND	125	0,12	4		633,04	0,000000		
21	GREEN	187	0,18	123		600,77	0,000000		
22	PROCESSING	136	0,13	27		584,46	0,000000		
23	NOV	153	0,14	63		661,43	0,000000		

Figura 93 - Candidatos a termo do CAF_IE - Parcial

Dessa lista, excluiríamos apenas roast e grind, que correspondem a verbos. O restante (21 palavras) poderiam ser consideradas candidatas a termo.

Se utilizarmos o ZAP, ficaremos com 417 candidatos a termo.

N	Key word	Freq	%	Freq	RC %	eyness	P	emmas	Set
1	COFFEE	3.433	3,27	109		534,32	0,00000		
2	BEANS	547	0,52	9		835,00	0,00000		
3	COFFEES	389	0,36	2		948,53	0,00000		
4	ESPRESSO	269	0,26	1		419,77	0,00000		
5	CHERRIES	231	0,22	1		222,18	0,00000		
6	BEAN	215	0,20	6		096,10	0,00000		
7	ARABICA	186	0,18	2		972,41	0,00000		
8	FLAVOR	176	0,17	10		864,32	0,00000		
9	HARVESTING	165	0,16	3		862,32	0,00000		
10	QUALITY	261	0,25	183	0,01	819,78	0,00000		
11	DRYING	154	0,15	3		793,92	0,00000		
12	METHOD	171	0,16	19		793,25	0,00000		
13	ROASTING	141	0,13	0		763,67	0,00000		
14	CAFFEINE	139	0,13	0		737,63	0,00000		
15	WATER	276	0,26	291	0,02	731,55	0,00000		
16	ROASTED	143	0,14	5		721,37	0,00000		
17	ROBUSTA	132	0,13	2		665,05	0,00000		
18	ACIDITY	130	0,12	1		689,26	0,00000		
19	GREEN	187	0,18	123		600,77	0,00000		
20	PROCESSING	136	0,13	27		584,46	0,00000		
21	DRY	152	0,14	63		561,42	0,00000		
22	GROWN	141	0,13	48		546,34	0,00000		
23	GIDE	07	0,06	6		470,35	0,00000		

417 Type-in FERMENTING

Figura 94 - Candidatos a termo final - Lista parcial (lista completa Anexo XII)

A expectativa para a extração de termos dos corpora escritos ficou dentro do esperado: mais de 500 candidatos poderiam ter sido encontrados, embora o programa apresente apenas 500. O tamanho dos corpora (ao redor de 100.000 palavras cada um) induz a um maior número de candidatos. Não podemos simplesmente comparar quantos candidatos foram extraídos na língua falada e na língua escrita, pois o tamanho dos corpora é que leva ao número de candidatos. Assim, no Capítulo 5 apresentaremos uma comparação final entre as duas modalidades, falada e escrita.

3.3 Construção do Glossário

3.3.1 Seleção final de termos: verbetes para glossário

Ao finalizarmos toda a extração terminológica em ambas as línguas, tanto em suas formas faladas como escritas, pudemos observar que, além das diferenças já citadas, devido ao tamanho diferente dos corpora falados em português e inglês, não seria possível encontrarmos todos os equivalentes necessários para concluir nosso trabalho, direcionado do português para o inglês, pois há mais candidatos na língua portuguesa. Como temos um corpus em português maior e um corpus de interpretação com assuntos muito distintos das entrevistas, além das entrevistas *per se* serem diferentes em cada língua, está claro que haveria diferença na constituição terminológica final.

Portanto, utilizaremos os corpora em português falado como fonte de construção de verbetes e os corpora em inglês para extração dos equivalentes e suas variantes. Também utilizaremos o corpus escrito como fonte de dados para equivalentes e variantes.

Abaixo, algumas diretrizes seguidas para a seleção dos termos dentre os candidatos:

- sua pertinência ao domínio;
- palavras que não têm sentido como termo simples podem ser termos complexos;
- verbos foram elencados junto a sua base, formando colocações verbais;
- as formas singular e plural foram consideradas como um só termo, desde que apresentem o mesmo sentido;

A ficha terminológica de cada termo foi preenchida da seguinte forma:

nº de entrada: X

Domínio: Ciências agrárias

Subdomínio: café/processamento pós-colheita

Língua: Português

Termo: terreiro

Exemplo de uso: "... eu levo (...) direto no **terreiro**, para secar, com casca, com tudo, do jeito..."

Categoria Gramatical: substantivo

Definição: lugar onde ocorre a secagem do café

Fonte: CAF_BO

N Concordance

10 Onde seca o café, que se chama **terreiro**, porque antigamente era de terra

Colocações:

a) verbais: *secar no terreiro*

b) nominais: *café de terreiro*

c) adjetivas: *terreiro tradicional*

Variante terminológica:

1) terrero;

Exemplo de uso: "... diz que é melhor como se fosse uma seca de terrero. Uh-huh. A seca dele."

Classificação: variante terminológica fonológica

2) terreirão*

Exemplo de uso: "Então você põe a lenha, deixa ele secando vinte minutos, trinta minutos, você vai lá põe lenha de novo, e volta e faz o serviço do terreirão."

Classificação: variante terminológica morfológica

Equivalente em inglês: *patio*

Exemplo de uso:

N Concordance

28 sun. The coffee is then shifted onto the **dry** portion of the patio, and the section

(fonte: CAF_IE)

Figura 95 - Ficha terminológica

3.3.2 Protótipo do glossário ideal

Para o glossário que consideramos ideal, a apresentação dos verbetes obedece a ordem alfabética e as formas são apresentadas da seguinte forma: substantivos e adjetivos no masculino singular. Alguns termos possuem outra definição quando no plural, dessa forma são inseridos como um novo verbete. A maioria dos termos elencados são sintagmas nominais. Há termos simples e complexos (colocações nominais, adjetivas e verbais). Incluímos colocações por serem de conhecimento indispensável ao tradutor e ao intérprete, evitando formas de “tradutês” (Tagnin, 2004) Os exemplos de uso foram extraídos dos corpora; o sinal [...] significa que uma parte foi retirada por interrupção do contexto.

Para as definições, Bowker e Pearson (2002) sugerem formulá-las por meio da análise das linhas de concordância, a partir de palavras como: “por exemplo”, “ex:”, “eg”, e do sinal de parênteses, que poderá conter explicações. No entanto, em corpus falado, a possibilidade de encontrarmos parênteses é nula, uma vez que as transcrições não usam parênteses para as falas, mas para observações.

Pretendíamos extrair as definições dos corpora falados. Porém, esses apresentam explicações superficiais, já que a situação permite completar o significado, inclusive com o uso de dêiticos, como aqui, ali, lá etc. Dessa forma, esses corpora não ofereceram material suficiente para a construção das definições. Procuramos, então, definições no corpus de língua escrita, que ofereceu definições simples e compreensíveis, mas suficientes para o público-alvo.

Apresentamos, abaixo, um protótipo de glossário ideal para dois termos escolhidos aleatoriamente e, em seguida, iniciamos outro capítulo onde temos o vocabulário finalizado e completo, porém mais simplificado, de termos das subáreas delimitadas.

Símbolos utilizados: s.m. (Substantivo Masculino); C (colocados); VAR (variantes):

BÓIA (s.m.): café que seca na planta, frutos verdes mal granados ou leitosos, que flutua quando lavado por ser menos denso.

“Após a colheita o café foi lavado, separado em cereja e bóia e seco em terreiro até 11% (b.u.) de teor de água”.

“... então na verdade nós ainda temos que descascar o bóia e esse verde...”

C: café bóia

grão bóia

Var : café bóia; grão bóia

“A lavagem é mais para tirar as pedra e impurezas do café né... quanto... mais pro meio do ano aí o café seca e cê separa ele também no lavador... sai o café bóia, que eles falam, e o café cereja ou verde... daí nisso... o café... por ex/ o café bóia sendo mais leve, cê consegue separar ele do verde...”

“... como você vai priorizar um produto de qualidade maior você tem que fazer uma amostragem nos talhões e começar pelo talhão que tem uma maior porcentagem de grão seco... de grão bóia... porque porque você visando a menor queda de grãos no chão...”

FLOATER

“The DC pulper receives unripe and ripe cherries which were separated from the floaters” in a mechanical washer-separator. or siphon tank.”

“This floater was processed yesterday.”

CAFÉ (s.m.): 1) fruto do cafeeiro; 2) bebida produzida a partir dos grãos torrados do fruto do cafeeiro, servido tradicionalmente quente, mas também pode ser consumido gelado. 3) *commodity* agrícola.

1) *“Depois de lavado, o café é secado no terreiro (ao sol) ou em secadores mecânicos.”*

2) *“Foi precisamente na França que, pela primeira vez, se adicionou açúcar ao café, o que aconteceu durante o reinado de Luís XIV, a quem haviam oferecido um cafeeiro em 1713”.*

3) *“Em alguns períodos da década de 1980, o café era a segunda commodity mais negociada no mundo por valor monetário, atrás apenas do petróleo.”.*

C: pé de café, café do chão, café de qualidade, café natural, café cereja, fazenda de café, bebida do café, café especial, especialista em café, café verde, lavoura de café, café brasileiro, grão de café, plantação de café, colher o café, qualidade do café, produzir café, beneficiamento do café, tomar café, variedade do café, café Bourbon, café espresso, secar o café, saca de café, sacos de café, café maduro, café comum, café queniano, café orgânico, café robusta, armazém de café, apanhadores de café, café arábica, café beneficiado, café do Brasil, plantar café, casca do café, café bóia.

COFFEE; COFFEE BEANS; COFFEE BERRIES

1) *“Workers sorting and pulping coffee beans in Guatemala”*

1) *“Coffee berries must be picked, defruited, dried, sorted, and, in some processes, also aged.”*

2) *“Studies are contradictory as to whether coffee has any specific health benefits, and results are similarly conflicting with respect to negative effects of coffee consumption.”*

3) *“There are coffee collectors, coffee suppliers, coffee exporters who are bidding for the coffee.”*

CAPÍTULO 4

ANÁLISE DOS DADOS

4.1 Língua Falada X Língua Escrita: O Que Dizem os Termos

Como há diferenças de tamanho entre os corpora de língua falada e língua escrita, não podemos identificar a riqueza lexical partindo da quantidade de candidatos a termo encontrados. Porém, se observarmos um dado estatístico fornecido pela WordList chamado STANDARDIZED TTR⁵⁴, podemos chegar a uma conclusão sobre a densidade das línguas falada e escrita. O *standardized TTR* significa, como já explicado, a padronização da proporção entre *types* (formas distintas) e *tokens* (número de palavras total do texto). Esse cálculo é feito em base de 1.000, ou seja, a cada 1000 palavras é calculado um índice da riqueza lexical do texto. Ao contrário do índice da riqueza lexical (*TTR*), que é o número de palavras total do texto dividido pelo número formas distintas, o *standardized TTR* calcula as formas distintas encontradas a cada 1.000 palavras. Assim é possível fazer a verificação de corpora de tamanhos diferentes, observando esse dado. No corpus escrito em português, o índice de palavras distintas que ocorrem no texto é de 51,17, enquanto no corpus falado (CAF_ALL_PB) é de 33, 27. Em inglês, o corpus escrito apresenta o índice da TTR padronizada em 1.000 de 41,68 e dos corpora falados, representados pelo CAF_ALL_IN, de 32,11.

Comprovamos, portanto, a hipótese de que a língua escrita é mais rica em nossos corpora especializados para o estudo da colheita e processamento do café.

Quanto às variantes, observamos que aquelas classificadas como variantes terminológicas de discurso (variantes de registro) apresentam

⁵⁴ Diferença entre palavras corridas e formas distintas encontradas em um corpus, a cada 1.000 palavras encontradas.

maior diferença entre os modos de língua falada ou escrita. Nos corpora escritos, por exemplo, não encontramos árvore do café ou árvore, para se referirem ao pé de café. São termos próprios do discurso de vulgarização científica, encontrados apenas em nossos corpora falados. Já derriça de chão foi encontrado apenas no corpus escrito, enquanto nos corpora falados encontramos café de chão. Outro detalhe interessante é o tratamento dos verbos: beneficiar café é encontrado em ambas as modalidades, falada e escrita. Sua variante, fazer beneficiamento do café, encontrada na língua falada, não aparece na língua escrita. Isso é um exemplo do quanto a língua escrita é mais “econômica” do que a língua falada.

4.2 Sociotermos: Análise de Casos

Todas as variantes analisadas (Anexo XIII) em nosso estudo foram incluídas no vocabulário que se apresentará adiante. Percebemos, por meio da análise, que as variantes encontradas em português no nosso trabalho podem ser enquadradas como:

A) Concorrentes

1. Lingüísticas:

- fonológicas
- morfológicas
- sintáticas
- lexicais

2. Registro:

- de discurso

B) Coocorrentes

- sinônimos

As classificações que não incluímos, tais como gráficas, geográficas, temporais e competitivas, não ocorrem no levantamento, devido às seguintes razões:

- **gráficas:** não há um levantamento de variantes encontradas na língua escrita presentes no vocabulário; as transcrições obedecem a norma culta de língua escrita;
- **geográficas:** todas as gravações foram feitas na mesma região;
- **temporais:** o período de gravação iniciou-se em meados do ano 2000, e a área não apresentaria muitas mudanças em tão curto espaço de tempo;
- **competitivas:** trata-se de um produto agrícola, que não possui influencia estrangeira em sua colheita ou processamento.

Faremos, abaixo, uma breve exposição para exemplificar cada tipo de variante, lembrando que a mesma variante pode ser classificada por mais de um critério.

A) Variantes Concorrentes	
1. Lingüísticas:	
- <u>fonológicas</u> :	colheta/colheita; panhador de café/apanhador de café; coiê café/colher café; lavora/lavoura; terreiro/terrero;
- <u>morfológicas</u> :	botão/botãozinho; saca de café/saco de café; terreiro/terreirão;
- <u>sintáticas</u> :	beneficiar café/fazer beneficiamento do café; colher com a mão/colheita manual; colheita mecanizada/colher com máquina;
- <u>lexicais</u>	arábica/café arábica; árvore/árvore do café; bóia/café bóia; Bourbon/café Bourbon; café de varrição/varrição; café de chão/café do chão; café cereja/cereja; café cereja descascado/cereja descascado; café de pano/café do pano; café desmucilado/desmucilado; café despulpado/despulpado; café lavado/lavado; café maduro/maduro; café robusta/robusta; café natural/natural; café cereja/cereja; fruto de café/fruto do café; grão moca/moca; grão verde/verde; secador/secador de café; terreiro/terreiro tradicional

Continua

Continuação

2. Registro:	
- <u>de discurso</u>	<p> árvore do café/pé de café; barra/saia; bóia/grão bóia; café catado a dedo/colheita a dedo; café catado no chão/café de varrição; café seca natural/café natural; panha de café/colheita de café; seca/secagem; </p>
B) Variantes Coocorrentes	
- <u>sinônimos</u>	<p> benefício/beneficiamento; café beneficiado/café verde/café limpo; colheita seletiva/colheita a dedo; café colhido no pano/café derriçado no pano; café de pano/ café derriçado no pano; café em coco/café sem beneficiar; colheita mecânica/colheita mecanizada; fazenda de café/propriedade; seca/secagem; tanque com água/tanque de fermentação; </p>

4.3 Vocabulário Bilíngüe - Café: Colheita e Processamento

Apresentaremos agora o produto final de nosso estudo, em forma de um vocabulário bilíngüe (PB – EN), da subárea do café, específico sobre colheita e processamento. A elaboração do referido vocabulário considerou como termo ambos os termos simples e os complexos, o que inclui os sintagmas nominais, colocações nominais, adjetivas e verbais. O Anexo XM apresenta a lista das colocações incluídas.

O vocabulário **CAFÉ: COLHEITA e PROCESSAMENTO** foi baseado principalmente na língua falada, utilizando a língua escrita apenas para a extração de equivalentes. Todas as variantes apresentadas foram também extraídas de corpora falados.

Pretendemos, com esse vocabulário, auxiliar tradutores, intérpretes e profissionais da área cafeeira com um produto técnico, visto que há poucos termos de vulgarização ou científicos. Há, no total, 178 verbetes incluídos, contando cada variante, em língua portuguesa.

Símbolos utilizados

s.m. – substantivo masculino

s.f. – substantivo feminino

v - verbo

PB – português

EN – inglês

var. – variação. As variações em *itálico* correspondem a usos que não pertencem à norma culta.

PB	EN
abandar café (v.)	to winnow
adubação (s.f.)	fertilization
alqueire (s.m.)	<i>sm 1 bras</i> land measure = 48,400 m ² in Minas Gerais, Rio de Janeiro and Goiás; 27,225 m ² in the Northern States; 24,220 m ² in São Paulo. 2 land into which one measure of 13.8 litres seed has been sown. 3 measure of contents = 13.8 litres (<i>port 8.4</i> litres). (fonte: Michaelis online – http://michaelis.uol.com.br)
apanhador de café (s.m.)	picker; coffee picker
var. panhador de café	
apanhar café (v.)	to pick coffee; to harvest coffee
var. colher café; coiê café; panhar café	
arábica (s.m.)	arabica
var. café arábica	
armazém de café (s.m.)	warehouse
árvore (s.f.)	tree
var. árvore do café; pé de café	
árvore do café (s.f.)	coffee tree
var. árvore; pé de café	
barra (s.f.)	bottom of the bush
var. saia	
bebida do café (s.f.)	coffee cup
beneficiamento (s.m.)	hulling
var. benefício	
beneficiar café (v.)	to hull coffee
var. fazer beneficiamento do café	
benefício (s.m.)	hulling
bóia (s.m.)	floater
var. café bóia; grão bóia	
botão (s.m.)	blossom
var. botãozinho	
botãozinho (s.m.)	little blossom
var. botão	
Bourbon (s.m.)	Bourbon
var. café Bourbon	
broto do café (s.m.)	sprout; coffee sprout
café (s.m.)	coffee; coffee berry; coffee bean
café arábica	Arabica coffee
var. arábica	
café beneficiado (s.m.)	hulled coffee; green coffee
var. café verde; café limpo	
café bóia (s.m.)	floater
var. bóia; grão bóia	
café Bourbon (s.m.)	Bourbon coffee
var. Bourbon	
café catado a dedo	hand-picked coffee
var. colheita a dedo; colheita seletiva	
café catado no chão	gleanings
var. café de varrição; café de chão; café do chão; varrição	

Continua

Continuação

café cereja (s.m.)	coffee cherry
var. cereja	
café cereja descascado (s.m.)	pulped natural
var. cereja descascado; descascado	
café colhido no pano (s.m.)	coffee harvested on a mat
var. café de pano; café derriçado no pano; café do pano	
café de chão (s.m.)	gleanings
var. café do chão; café de varrição; café catado no chão; varrição	
café de melhor qualidade (s.f.)	best quality coffee
café de pano (s.m.)	see "café colhido no pano"
var. café colhido no pano; café do pano; café derriçado no pano	
café de qualidade (s.m.)	quality coffee
café derriçado no pano (s.m.)	coffee stripped on a mat
var. café de pano; café do pano; café colhido no pano	
café desmucilado (s.m.)	semi-washed coffee
var. desmucilado	
café despulpado (s.m.)	fully washed coffee
var. despulpado	
café de terreiro (s.m.)	dry cherry
café de varrição (s.m.)	gleanings
var. café catado no chão; café de chão; café do chão; varrição	
café do chão (s.m.)	gleanings
var. café de chão; café catado no chão; café de varrição; varrição	
café do pano (s.m.)	see "café colhido no pano"
var. café colhido no pano; café de pano; café derriçado no pano	
café em coco (s.m.)	dry cherry
var. café sem beneficiar; café comum; café normal	
café especial (s.m.)	specialty coffee
café esqueletado (s.m.)	pruned coffee tree
cafeicultura (s.f.)	coffee farming
café lavado (s.m.)	washed coffee
var. lavado	
café limpo (s.m.)	clean coffee; green coffee; hulled coffee
var. café verde; café beneficiado	
café maduro (s.m.)	ripe cherry; ripe coffee cherry
var. maduro	
café natural (s.m.)	natural coffee
var. café seca natural; natural	
café orgânico (s.m.)	organic coffee
café robusta (s.m.)	Robusta coffee
var. robusta	
café seca natural (s.m.)	natural
var. café natural; natural	

Continua

Continuação

café sem beneficiar (s.m.)	dry cherry
var. café em coco	
café sujo	dirty coffee
café verde	green coffee
var. café beneficiado	
carreta do trator (s.f.)	tractor loader
casca do café (s.f.)	coffee husk; trash
catar café (v.)	to pick coffee
cereja (s.m.)	cherry
var. café cereja	
cereja descascado	pulped natural
var. café cereja descascado; descascado	
chão do terreiro (s.m.)	patio floor
clima (s.m.)	weather
coiê café (v.)	to harvest coffee; to pick coffee
var. colher café; apanhar café; panhar café	harvesting
colheita (s.f.)	
var. panha de café; colheita do café; colheta	
colheita a dedo (s.f.)	selective picking
var. café catado a dedo; colheita seletiva	
colheita do café (s.f.)	coffee harvesting
var. colheita; panha de café; colheta	
colheita manual (s.f.)	manual harvesting
var. colher com a mão	
colheita manual mecanizada (s.f.)	manual and mechanized harvesting
colheita mecânica (s.f.)	mechanized harvesting
var. colheita mecanizada; colher com a máquina	
colheita mecanizada (s.f.)	mechanized harvesting
var. colheita mecânica; colher com a máquina	
colheita seletiva (s.f.)	selective picking
var. café catado a dedo; colheita a dedo	
colher café (v.)	to harvest coffee; to pick coffee
var. coiê café; panhá café; apanhar café	
colher com a mão (v.)	to harvest manually
var. colheita manual	
colher com a máquina (v.)	to harvest mechanically
var. colheita mecânica; colheita mecanizada	
colher no pano (v.)	to harvest on a mat
colheta (s.f.)	harvesting
var. panha de café; colheita de café; colheita	

Continua

Continuação

cuidar da lavoura (v.)	to maintain the coffee crop
dar café (v.)	to produce coffee
derricha (s.f.)	stripping; strip-picking
derricha no pano (s.f.)	stripping on mat; strip-picking on a mat
derricar café (v.)	to strip coffee
derricar no pano (v.)	to strip on a mat
descascado (s.m.)	pulped
var. café cereja descascado; cereja descascado	
descascador de cereja (s.m.)	pulper
desmucilado (s.m.)	semi-washed
var. café desmucilado	
despolpado (s.m.)	fully washed
var. café despolpado	
despolpado natural (s.m.)	fully washed natural
despolpado por fermentação (s.m.)	fully washed by fermentation
despolpador de café (s.m.)	pulper
época da colheita (s.f.)	harvesting time
esqueletamento (s.m.)	pruning
fazenda de café	coffee farm
var. propriedade	
fazer a colheita (v.)	to harvest
fazer beneficiamento do café (v.)	to hull; to clean the coffee
var. beneficiar o café	
florada (s.f.)	flowering
folha (s.f.)	leaf
fruto (s.m.)	fruit
var. fruto de café; fruto do café	
fruto de café (s.m.)	coffee fruit; coffee bean
var. fruto; fruto do café	
fruto do café (s.m.)	coffee fruit; coffee bean
var. fruto; fruto de café	
geada (s.f.)	frost
grão bóia (s.m.)	floater
var. bóia; café bóia	
grão de café (s.m.)	coffee bean
grão maduro (s.m.)	ripe cherry
grão moca (s.m.)	peaberry
var. moca	
grão preto (s.m.)	black bean
grão seco (s.m.)	dry bean/cherry
grão verde (s.m.)	immature coffee
var. verde	
guardar café (v.)	to storage coffee
guardar na tulha (v.)	to storage in the warehouse
lavado (s.m.)	washed
var. café lavado	
lavador (s.m.)	washer
var. lavador de café	
lavador de café (s.m.)	washer
var. lavador	

Continua

Continuação

lavador mecânico (s.m.)	mechanized washer
lavagem (s.f.)	washing
lavar café (v.)	to wash coffee
lavora (s.f.)	coffee garden
var. lavoura de café	
lavoura de café (s.f.)	coffee garden
var. lavora	
maduro (s.m.)	ripe cherry
var. café maduro	
mão-de-obra (s.f.)	labor
máquina (s.f.)	machine
máquina do beneficiamento (s.f.)	huller
mecanização da colheita	harvest mechanization
meia seca (s.f.)	half drying
moca (s.m.)	peaberry
var. grão moca	
momento da colheita	harvesting time
mucilagem (s.f.)	mucilage
muda de café (s.f.)	sprout
natural (s.m.)	natural
var. café natural; café seca natural	
palha do café (s.f.)	husk; trash
<i>panha de café</i> (s.f.)	harvesting
var. colheita de café; colheta; panha de café	
<i>panhar café</i> (v.)	to harvest coffee; to pick coffee
var. apanhar café; colher café; coiê café	
<i>panhador de café</i> (s.m.)	coffee picker; picker
var. apanhador de café	
pé de café (s.m.)	coffee tree
var. árvore; árvore de café	
pedra (s.f.)	stone
peneira (s.)	sieve
pergaminho (s.m.)	parchment
plantar café (v.)	to plant coffee
plantação (s.f.)	coffee plantation
polpa (s.f.)	pulp
processamento (s.m.)	processing
processo mecânico (s.m.)	mechanized process
produção (s.f.)	production
produção do café (s.f.)	coffee production
produtor de café (s.m.)	coffee farmer; coffee grower
propriedade (s.f.)	coffee land
var. fazenda de café	
qualidade do café	coffee quality
rama (s.f.)	branch
robusta (s.m.)	Robusta
var. café robusta	
saca (s.f.)	bag
saca de café (s.f.)	coffee bag
var. saco de café	

Continua

Continuação

saca de peneira (tamanho) (s.f.)	bag sieve ...(size)
saco de café (s.m.)	coffee bag
var. saca de café	
safra (s.f.)	crop
saia (s.f.)	skirt
var. barra	
seca (s.f.)	drying
var. secagem	
secador (s.m.)	dryer
var. secador de café	
secador à gás (s.m.)	gas dryer
secador à lenha (s.m.)	wooden dryer
secador de café (s.m.)	coffee dryer
var. secador	
secador vertical (s.m.)	rotative dryer
secagem (s.f.)	drying
var. seca	
secagem no terreiro (s.f.)	sun drying
secar café (v.)	to dry coffee
secar no terreiro (v.)	to sun dry coffee
talhão (s.m.)	stripe
tanque com água (s.m.)	tank
var. tanque de fermentação	
tanque de fermentação	fermentation tank
var. tanque com água	
terreiro (s.m.)	patio; terrace
var. terrero, terreirão; terreiro tradicional	
terreiro coberto (s.m.)	covered tables
terreiro tradicional (s.m.)	patio; terrace
var. terrero, terreirão; terreiro	
terreno (s.m.)	land
terrero (s.m.)	patio; terrace
var. terreiro, terreirão, terreiro tradicional	
terreirão (s.m.)	patio; terrace
var. terreiro, terrero; terreiro tradicional	
tirar folha (v.)	to remove leaves
trator (s.m.)	tractor
tulha (s.f.)	warehouse
turmero (s.m.)	workers chief
varrição (s.f.)	gleanings
var. café de varrição; café catado no chão; café de chão; café do chão	
verde (s.m.)	unripe fruit
var. grão verde	

CAPÍTULO 5

CONSIDERAÇÕES FINAIS

O objetivo da presente dissertação de mestrado era analisar as ocorrências de variantes terminológicas na modalidade oral da língua de especialidade da cafeicultura, especificamente sobre a colheita e o processamento do café pós-colheita, além de elaborar um vocabulário bilíngüe sobre o tema, contendo variantes terminológicas nas subáreas acima citadas.

Propusemo-nos a investigar essas ocorrências, observando sua existência na língua falada em um corpus de linguagem de especialidade do Café, assim como a diferença entre a língua falada e a língua escrita em relação à riqueza lexical e a possibilidade de incluirmos tais variantes em um vocabulário técnico.

A fim de analisarmos a ocorrência de variantes na língua falada, construímos corpora falados específicos, utilizando entrevistas, conversações e interpretações efetuadas com profissionais da área cafeeira, em ambas as línguas. Muitas foram as dificuldades enfrentadas para se atingir esse propósito, devido à lentidão do trabalho envolvido na coleta e transcrição de dados da língua falada. A necessidade de deslocamento, contato com informantes, avaliação e seleção do material coletado, além da coleta de dados propriamente dita, faz com que o tempo para a pesquisa em língua falada seja multiplicado. Além disso, é preciso levar em conta que a transcrição exige treinamento e trabalho árduo, o que envolve muito tempo. Para a língua inglesa, os percalços foram ainda maiores: encontrar o mesmo número de informantes que tínhamos para o português, e que se dispusessem a participar do projeto, não foi possível apenas em nosso país. Os dados em inglês estavam escassos em relação

ao português, razão pela qual decidimos partir para a pesquisa em campo, em terras estrangeiras. Uma conferência voltada à área cafeeira, realizada em terras etíopes, foi uma excelente solução encontrada. Em primeiro lugar, porque a Etiópia é considerada o berço do café. Em segundo, porque, além da participação na conferência e contato com informantes de diversos países, pudemos visitar várias fazendas cafeeiras e entrar em contato com uma variedade de trabalhadores da área. Mais uma vez esbarramos nos entraves de coleta de dados e transcrição. Devido a ruídos externos, muitas gravações oferecem alto grau de dificuldade com o áudio; além disso, o equipamento disponível no mercado para gravações em áudio não favorece pesquisadores de “pequeno porte”. Os mais novos gravadores digitais gravam o máximo de quatro horas seguidas, além de exigirem um computador para arquivar as gravações feitas e recarregar a bateria necessária ao seu funcionamento. Imaginemos esse panorama em locais remotos, normalmente zonas rurais, a quilômetros da eletricidade... Apesar de tudo isso, conseguimos transcrever mais de 25 horas de gravação das 60 horas que vieram na bagagem.

Como a língua falada foge ao planejamento do tópico discursivo, nosso material acabou contendo referência também a outras áreas da cafeicultura. Após a marcação dos corpora com etiquetas de identificação de informantes, inclusão de cabeçalhos e extração terminológica, selecionamos os termos que poderiam ser elencados em nosso vocabulário.

Considerando aspectos metodológicos, o programa WordSmith Tools versão 4.0, utilizado para todas as análises dos corpora, mostrou-se satisfatório na maioria das vezes. Encontramos alguns problemas, porém todos puderam ser contornados com criatividade para que os objetivos fossem alcançados. Como a versão 5.0 encontra-se em desenvolvimento, acreditamos que tais problemas sejam sanados em curto prazo.

A Lingüística de Corpus, com seu ferramental tecnológico, mostrou-se um eficiente referencial teórico e prático para o trabalho, de modo que mais pesquisas deveriam ser feitas com seu apoio, a fim de garantir maior eficiência nos trabalhos com língua falada.

Como era de se esperar, a análise de nossos corpora demonstrou que a língua falada apresenta uma variedade de termos não presentes na língua escrita e vice-versa. Por isso, cremos que a construção de um produto terminológico voltado a profissionais que utilizem a língua falada em seu trabalho, como é o caso de intérpretes, deva ser baseada em corpora na modalidade falada, para permitir a observação de termos em condições e situações sociais distintas. Também acreditamos que seja importante para todos os profissionais de determinadas áreas de especialidade ter conhecimento das variantes utilizadas pelos diferentes níveis socioprofissionais; afinal, elas existem, mesmo que certas variantes não sejam empregadas por todos da mesma maneira. Assim, o produto terminológico ideal seria resultado da análise de corpora falados e escritos, unindo a riqueza lexical de um às possibilidades de variantes apresentadas pelo outro.

Nesse sentido, atingimos nosso propósito em apresentar um vocabulário que contivesse variantes terminológicas. Não foi possível, entretanto, elaborar um glossário bidirecional, que também apresentasse as variantes em língua inglesa. Apesar de o material coletado em língua inglesa ser extremamente rico, com diversas variantes geográficas devido à metodologia de coleta de dados, não conseguimos transcrever o suficiente dentro do tempo previsto para o trabalho, como já mencionamos, devido às dificuldades relacionadas aos processos envolvidos. Nada impede, no entanto, que tais corpora sejam ampliados, e que no futuro continuemos o trabalho em direção ao inglês.

Todos os dados coletados em português do Brasil e em inglês de diversos países poderão ser disponibilizados a pesquisadores da língua falada, uma vez que temos autorização dos informantes para sua utilização. Esse aspecto deveria ser amplamente considerado ao se fazer pesquisas em língua falada, já que o extenuante trabalho envolvido justifica a disponibilização para outros pesquisadores, seguindo o princípio da reusabilidade, economizando, dessa forma, tempo e recursos materiais.

Ainda há muito trabalho a ser desenvolvido na área cafeeira, principalmente em relação à produção terminológica bilíngüe, como ficou evidente no decorrer da pesquisa realizada: pouco se produz em material monolíngüe, e mais escassos ainda são os materiais bilíngües. Isso, para o país que é o maior exportador de café do mundo, não é necessariamente algo do que se orgulhar.

REFERÊNCIAS BIBLIOGRÁFICAS

ABIC - Associação Brasileira da Indústria de Café. Disponível em <<http://www.abic.com.br>>. Acesso em 8 set. 2007.

AIIC. Practical guide for professional interpreters. Geneva: AIIC, 1982.

Almeida, G.M.B. O percurso da terminologia: de atividade prática à consolidação de uma disciplina autônoma. **TRADTERM - 9 - Revista do CITRAT - FFLCH- USP.** São Paulo: Humanitas, 2003.

Alves, I.M. (Coord.). Glossário de termos neológicos da economia. In: Cadernos de Terminologia, nº 3. **CITRAT-FFLCH/USP.** São Paulo: Humanitas, 2001 (reimp. com correções).

_____. O léxico na língua falada. In: Preti, D. (Org.). **Análise de textos orais. Projeto NURC-SP.** São Paulo: Humanitas/FFLCH/USP, 2003.

Atkins, S.; Clear, J.; Ostler, N. Corpus design criteria. In: **Literary and linguistic computing**, vol. 7, nº 1. Oxford: Oxford University Press, 1992.

Bally, C. **Traité de stylistique française.** Georg, Geneve, Klincksieck, 1951.

Banco de português. Disponível em <<http://www2.lael.pucsp.br/corpora/bp/index.htm>>. Acesso em 20 mai. 2006.

Barros, L.A. **Curso básico de terminologia.** São Paulo: Edusp, 2004.

Berber Sardinha, T. **Linguística de corpus**. São Paulo: Manole, 2004.

Bowker, L; Pearson, J. Working with specialized language. **A practical guide to using corpora**. London and New York: Routledge, 2002.

Bressani, E. **Guia do barista. Da origem do café ao espresso perfeito**. São Paulo: Café Editora, 2007.

Cabré, M.T. **La terminología: representación y comunicación**. Barcelona: Universitat Pompeu Fabra, 1999.

Camlong, A.; Zapparoli, Z.M. **Do léxico ao discurso pela informática**. São Paulo: Edusp, 2002.

Carvalho, J.G.H. **Teoria da linguagem**. Coimbra: Atlântida, 1970.

Cheshire, J. **Introdução. English around the world**. Cambridge: Cambridge University Press, 1991.

Cintra, G. Filologia bandeirante: Registro de entrevistas. In: Megale, H. (Org.). **Filologia bandeirante. Estudos I** São Paulo: Humanitas/FFLCH, 2000.

Delisle, J.; Woodsworth, J. **Translators through history**. 1995. Tradução Sérgio Bath. Os Tradutores na História, 1998.

Faulstich, E. A socioterminologia na comunicação científica e técnica. **Ciência e Cultura**. São Paulo, v. 58, n. 2, 2006.

_____. Aspectos de terminologia geral e terminologia variacionista. **Tradterm 7**. São Paulo: Humanitas/FFLCH/USP: 2001, p.11-40.

_____. Socioterminologia: mais que um método de pesquisa, uma disciplina. **Ciência da Informação**. v. 24, n. 3, 1995a.

_____. Terminologia: o Projeto Brasilterm e a formação de recursos humanos. **Revista Ciência da Informação**. Brasília, v. 24, n. 3, p. 356-63, set. dez/1995b.

Finatto, M.J.B. Unidade e variação na língua portuguesa: a variação em terminologia. In: Krieger, M.G; Maciel, A.M.B. (Org.). **Temas de terminologia**. Porto Alegre e São Paulo: Universidade/UFRGS e Humanitas/USP, 2001. p.150-54.

Furlani, C.E.A.; Silva, F.M. Terminologias da mecanização agrícola na cultura do café. **Revista Cafeicultura**. 5ª Ed.. Jan./Fev., 2003.

_____. **Basic concepts and models for interpreter and translator training**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995a.

Gile, D. Mirror mirror on the Wall: An introduction. TARGET 7:1. Special Issue. **Interpreting Research**. Amsterdam: John Benjamins Publishing Company, 1995b. p. 1-6.

Inglez de Souza, J.S. (Coord.). **Enciclopédia agrícola brasileira**. C-D. v. 2. São Paulo: Edusp, 1998.

Instituto Observatório Social. **Café: vida, produção e trabalho**. Disponível em <http://www.observatoriosocial.org.br/download/cafe_maio2004BX.pdf>. Acesso em 8 set. 2007.

Jones, R. **Conference Interpreting Explained**. Manchester, UK & Northampton, MA: St Jerome Publishing, 1998.

Kachru, B.B.; Nelson, C.L. World Englishes. In: McKay, S.L (Ed.). **Sociolinguistics and language teaching**. Cambridge: Cambridge University Press, 1996.

Kennedy, G. **An introduction to corpus linguistics**. London and New York: Longman, 1998.

Krieger, M.G.; Finatto, M.J.B. **Introdução à terminologia: teoria & prática**. São Paulo: Contexto, 2004.

Lederer, M. **La traduction aujourd'hui: le modèle interprétatif**. Paris: Hachette Livre, 1994.

Leech, G.; Garside, R.; Bryant, M. CLAWS4: **The tagging of the British National Corpus**. In Proceedings of the 15th International Conference on Computational Linguistics (COLIN 94) Kyoto, Japan, 1994. p. 622-8.

Macwhinney, B. **Manual de CHAT**. Disponível em <<http://childes.psy.cmu.edu/manuals/CHAT.pdf>>. Acesso em 2 dez. 2003a

_____. **Manual de CLAN**. Disponível em <<http://childes.psy.cmu.edu/manuals/CLAN.pdf>>. 2 dez. 2003b.

Marcuschi, L.A. **Análise da conversação**. São Paulo: Editora Ática, 2003.

Miranda, L.M.C. Aplicação de base metodológica para pesquisa em socioterminologia na elaboração de um glossário demonstrativo em fitopatologia. **Ciência da Informação**. v. 24, n. 3, 1995.

Moser-Mercer, B. Research Committee Paradigms Gained, or the art of productive disagreement. **AIIIC Bulletin XIX**. n.:2, p. 11-5, 1991.

Pagura, R.A Interpretação de conferências: interfaces com a tradução escrita e implicações para a formação de intérpretes e tradutores. **D.E.L.T.A.** n. 19, Especial. São Paulo: EDUC, 2003. p. 209-36.

Pearson, J. **Terms in context**. Amsterdam, John Benjamins Publishing Company, 1998.

Peters, P. In quest of international English: mapping the levels of regional divergence. In: Renouf, A. (Ed.) Explorations in corpus Linguistics. Language and computers: **Studies in Practical Linguistics**. nº 23. Amsterdam - Atlanta: Rodopi, 1998.

Reppen, R.; Fitzmaurice, S.M.; Biber, D. **Using corpora to explore linguistic variation**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

Rodrigues, A.C.S. Língua falada e língua escrita. In: Preti, D. (Org.). **Análise de Textos Orais**. Projeto NURC-SP. São Paulo: FFLCH/USP, 1993.

Rodrigues, A.C.S.; Ferreira Neto, W. Transcrição de inquéritos: problemas e sugestões. In: Megale, H. (Org.) **Filologia bandeirante. Estudos I**. São Paulo: Humanitas/FFLCH, 2000.

Sager, J.C. **A practical course in terminology processing**. Amsterdam/Philadelphia: John Benjamins, 1990.

Schjoldager, A. Interpreting research and the "Manipulation School" of translation studies. **Target 7:1**. Special Issue. Interpreting Research. Amsterdam: John Benjamins Publishing Company, 1995. p. 29-45.

Scott, M. **WordSmith Tools 4.0**. Disponível em <<http://www.lexically.net/wordsmith/index.html>>. 2004-2007.

Seleskovitch, D.; Lederer, M. **Interpréter pour traduire**. Paris: Didier Érudition, 1984.

Shlesinger, M. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. **META XLIII**. v. 4, p.486-93, 1998.

Strehler, R.G. A socioterminologia como base para a elaboração de glossários. **Ciência da Informação**. v. 24, n. 3. 1995.

Tagnin, S.E.O. A **identificação de equivalentes tradutórios em corpora comparáveis**. Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em data 2007.

_____. **Construção de corpora**. Apresentação no curso de Pós-Graduação de Estudos Lingüísticos e Literários em Inglês. Disciplina: Lingüística de Corpus II. Em 14 set. 2005a.

_____. **O Jeito que a gente diz. Expressões convencionais e idiomáticas. Inglês e português**. São Paulo: Disal Editora, 2005b.

_____. **Uma metodologia contrastiva baseada em corpus paralelo para evitar o "tradutês"**. Apresentação em Congresso - GEL, 2004.

Teixeira, E.D. Etiquetagem em lingüística de corpus – possibilidades de aplicação. In: Gerber, R.M.; Vasilévski, V. **Um percurso para pesquisas com base em corpus**. Florianópolis: Editora da UFSC. [no prelo].

Thompson, P. Spoken language corpora. In: Wynne, M (Ed.). **Developing linguistic corpora: a guide to good practice**. Oxford: Oxbow Books, 2005. Disponível em <<http://ahds.ac.uk/linguistic-corpora/>>. Acesso em 21 dz. 2005.

Tognini-Bonelli, E. **Corpus linguistics at work**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001.

Toury, G. **Descriptive translation studies and beyond**. Amsterdam/Philadelphia: John Benjamins, 1995.

Urbano, H. Apresentação. In: Preti, D.; Urbano, H. (Org.) **A Linguagem falada culta na cidade de São Paulo. Vol. III - Entrevistas**. São Paulo: T. A. Queiroz Editor LTDA/ FAPESP, 1988. p. 1-7.

_____. **Oralidade na literatura (O caso Rubem Fonseca)**. São Paulo: Cortez, 2002. p.97-105

_____. **Vocabulários específicos: subsídios para seu estudo e compreensão**. Dissertação (Mestrado em Filologia e Língua Portuguesa). Universidade de São Paulo: São Paulo, 1978.

Wallmach, K. Using parallel corpora to determine interpreting strategies for languages of limited diffusion in South Africa. **Translation and Meaning**. Part 6, 2002. p. 503-9.

Wüster, E. **Introducción a la teoría general de la terminología y a la lexicografía terminológica**. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona, 1998.

ANEXOS

ANEXO I

UNIVERSIDADE DE SÃO PAULO

Mestrado em Estudos Lingüísticos e Literários em Inglês

Luciana Latarini Ginezi

Project Presentation

This study aims to analyze the linguistic varieties among different English speaking communities in the field of Terminology. Thus, the idea is to compare how the terms behave in context among different speakers. The topic for the present study is Coffee – harvesting and processing, as well as its related subjects that may come to be relevant.

The task will be carried out by the collection of oral texts that will compose the Corpus. These oral texts will be recordings done by the researcher using audio and visual aids. Later on, all the material will be transcribed and codified using specific computer tools and then analyzed.

A commented vocabulary with the definition of terms will be the final product, indicating the probable differences found in the Corpus

The research proposes to help Consecutive Interpreters when working in the same area, through identifying the different terms according to the country.

ANEXO II**INQUÉRITO 3**

Data: 12/07/2004

Duração: 14 min

Entrevistadora: 1

Entrevistado Fausto Aurélio Scanavachi: 2

1. Estou entrevistando Fausto Scanavachi.
2. *Fausto Aurélio.*
1. Fausto Aurélio Scanavachi quantos anos você tem Fausto?
2. *Vinte e quatro anos.*
1. Fausto... Éh... Faz quanto tempo que você atua como engenheiro agrônomo?
2. *Fazem dois anos no final de dois mil e quatro.*
1. Uh-huh. éh... Você trabalha ou você já tem as informações pra poder aplicá-las sobre U... Café a... Produção do cereja descascado você poderia ah... Falar um pouquinho a respeito?
2. *É a produção do cereja descascado seria a produção de um café com grão com grãos exclusivos no caso né que... Seria uma bebida bem mais suave só que com no caso do no caso de você produzir o café ensacado teria um tempo bem menor de armaze arma de armazema.*
1. De armazenamento.
2. *Armazenamento.*
1. Uh huh.
2. *No caso de um ano um ano e pouquinho.*
1. É o máximo que ele pode ficar.
2. *É o máximo que ele pode ficar armazenado.*
1. E o outro café que éh... Não sei como é que se chama via natural?
2. *Café verde.*

1. Café verde? Ele fica quanto tempo?
2. *Ah ele fica por cinco...*
1. Sem deteriorar.
2. *É a cada ano que passa ele vai perdendo a qualidade né.*
1. Ah é?
2. *Vai perdendo a propriedade organoléptica vai perdendo o tipo a cor.*
1. Uh huh.
2. *Então o... O aceitável assim essa diferença de tempo geralmente tanto como no cereja como no café normal é de um ano tanto é que no mercado de um ano pro outro o café do ano anterior ele vale mais que o café do ano corrente do ano que tá sendo colhido.*
1. Ah então esse armazenamento é bom pro café pro preço.
2. *Isso... Por causa do descanso tudo.*
1. Certo e... E essa... E quais seriam assim as etapas... Há diferenças pra você produzir um café natural café verde pra produção do cereja descascado em termos de colheita por exemplo?
2. *Há totalmente diferente... Porque geralmente quando você você vai co/começar a colher um talhão do café normal éh... Tecnicamente a gente diz que você tem que ter pelo menos de quatro a cinco por cento de grãos verdes.*
1. Certo.
2. *Para diminuir a porcentagem de grãos verdes na sua no seu tipo de café e também de grãos de grãos pretos né que é pra diminuir a quantidade de defeitos do... Do produto já no caso do cereja descascado... Você tem que pensar em outra forma... Como você vai priorizar um produto de qualidade maior você tem que fazer uma amostragem nos talhões e começar pelo talhão que tem uma maior porcentagem de grão seco... De grão bóia... Por que porque você visando a menor queda de grãos no chão.*
1. Certo.
2. *Diminuindo o café de chão e aumentando assim a maior quantidade de de grão de maior qualidade.*
1. Uh huh.

1. E a mão-de-obra para essa colheita ela tem que ser diferenciada? tem que ser pessoa especializada pra colher um...
2. *Tem tem pessoas que que colhem o café normalmente no pano tem que ser feito sempre no pano não pode ter contato nenhum com o chão.*
1. Certo.
2. *Tem pessoas que colhem normalmente no pano daí no lavador separa-Se o bóia do cereja né o cere/ o cereja mais o verde vão passar pelo descascador de cereja onde o descascador vai processar o cereja e vai tirar o verde...também agora tem tem pessoas ahn... Principalmente na re/ região de Pinhal aqui tem pessoas fazendo colheita seletiva.*
1. Uh huh.
2. *Que é a colheita a dedo coletando só os grãos que já estão cereja né dando aí uns dois a três repasses por por por talhão e aí você/ você vai ter bem não vai ter nada de café de chão e e pouquíssimo bóia.*
1. Certo.
2. *Também tem pessoas que que que colhem duas vezes no mesmo talhão coletando a planta da do meio pra cima e do meio pra baixo na segunda colheita.*
1. E... Para esse tipo de produção não pode ser usado colhe... A/ aquela máquina para colher como é que ela chama?
2. *A... A auto-motriz?*
1. Isso.
2. *Pode.*
1. Pode também.
2. *Pode ser só que daí você entra com máquina mais vezes no talhão e... E em épocas assim um pouco mais espaçadas uma da outra.*
1. Ah entendi.
2. *Que daí quando começar a amadurar o café você vai entrar com a máquina e aí você vai colher só o que está mal que está que está maduro e vai ficar o que está verde aí você passa de novo.*
1. Certo.

2. *Só que é o que vai ocasionar um ma/ maior desgaste para sua lavoura né.*
1. Por que?
2. *Porque o trânsito da máquina e a a própria colheita dela o mecanismo de colheita dela judia judia mais da... Deteriora mais a planta em relação à colheita manual ou até a a manual mecanizada né.*
1. Ah entendi e... Agora... Ahn você disse que nessa lavagem né que faz a separação do verde do do bóia éh... São máquinas especiais que são usadas?
2. *São são máquinas especiais é... Tem que se o... O... Lavador mecânico que vai que vai dar um fluxo contínuo de de café pra abastecer o descascador de cereja que que vai fazer o descascamento de cereja a sepa/ separação do verde e posteriormente éh...posteriormente opcional ou não, a o desmucilamento.*
1. Ah, entendi.
2. *O café, somente o descascado, ele conserva mais semelhantes com o café verde do que o desmucilado. Só que o desmucilado vai vai como produto final ter uma bebida mais suave ainda do que o cereja, que só o cereja descascado.*
1. Certo. Esse desmucilado é de que tipo, ele entra em que classificação?
2. *É, geralmente esses cafés são bebida mole né. Bebida mole.*
1. Ah, tá.
2. *Porque até os concursos que têm aí, se não for bebida mole não entra.*
1. Não entra.
2. *Aí dentro de mole pra cima vai ter as classificações até atingir o melhor né. Muitos desses concursos que as pessoas mais têm ido assim, dizem que é mais por marketing da marca. Vamos supor, o último concurso da Illy (...) foi mil e quatrocentos dólares a saca. Mas foi um lote de quarenta e cinco sacas. Então, mais é por marketing por dizer que a firma Illy Café que seja pagou mil quatrocentos e cinqüenta dólares de uma saca de café. Nem que, assim, isso não quer dizer que esse café é o melhor de todos, entendeu?*
1. Entendi.
2. *Mais é o marketing de dizer que eles pagaram isso do café.*

1. Certo. Você acha que então todos os cafés que acabam sendo produzidos eles têm mais ou menos a mesma qualidade, todos eles são de ótima qualidade?
2. *Eles são de ótima qualidade. É lógico que tem o... De todos os candidatos vai ter sempre o que vai se sobressair né?*
1. Sim. Entendi. E continuando sobre o cereja descascado, a secagem dele também é diferente da secagem do verde?
2. *A secagem dele tem que ser uma secagem mais branda do que a do verde, com menos fogo, vai ter um número maior de horas, um número maior de descanso até o beneficiamento, sempre com mais cuidado do que o verde porque ele é um produto que está mais exposto.*
1. Certo.
2. *Principalmente na secagem, que no caso do cereja descascado a temperatura da massa não pode ultrapassar os quarenta graus Celsius. Já no café normal não pode ultrapassar os quarenta e cinco graus.*
1. Ah, então é diferente também.
2. *É diferente também.*
1. E ele passa um tempo pelo terreiro também ou não?
2. *Passa. Em geral, em geral, assim dependendo do manejo do terreiro, da capacidade do terreiro, o recomendável é a partir é... Tem que ficar no terreiro por três dias.*
1. Certo.
2. *Uns três dias. Daí ele não pode ser amontoado, assim como o outro café, tem um (...) costume de fazer um enleiramento no fim do dia, ele não pode ser amontoado, porque cada vez, quanto mais você aglomera ele, mais aumenta a chance de fermentação e também o terreiro no qual é colocado o café não pode ter trinca...*
1. Tem que ser lisinho.
2. *Lisinho. Lisinho porque um grãozinho de café que fica dentro de uma trinca do terreiro que fermenta, ele pode estragar o lote inteiro de bebida por causa dessa fermentação dele, atrapalhar no na bebida do lote inteiro.*
1. No resto. Um grão apodrece apodrece...
2. *Um grão (...) maçã podre...*

1. É. Apodrece o saco né. E... Éh... O terreiro que você está falando tem que ser feito de qual material pra não ter nenhuma trinca, tem alguma...
2. *Tem. O mais indicado o mais indicado é o concreto zilado né. Tem muita gente que usa que usa o asfáltico, mas o piche pelo alcatrão libera um pouco de cheiro que pode ser absorvido pelo café, porque o café ele absorve tudo quanto é tipo de cheiro que você pode pensar. Até o pó de café dentro do armário ele absorve o cheiro que está em volta dele.*
1. Interessante. Muito bom. E... Depois, você trabalha... Numa... Numa... No lugar onde você dá assistência é na região, nessa região, na Alta Mogiana ou no Sul de Minas?
2. *É no Sul de Minas. A gente trabalha mais é... Eu trabalho na propriedade da família né e propriedade aqui em Santo Antonio do Jardim, tem mais perto do cerrado de Minas e aqui no Sul de Minas.*
1. Uh-huh. E você nessas na na nessas terras você éh... Tem muita mecanização, como é que é a mecanização, é meio a meio?
2. *Depende da região. Depende da região. Dependendo da região e da topografia né. Aqui em Santo Antonio do Jardim (...) mecanizado.*
1. Uh-huh.
2. *Mais perto do cerrado é totalmente mecanizado.*
1. Totalmente mecanizado, inclusive colheita.
2. *Colheita, tudo.*
1. Tudo.
2. *E no Sul de Minas ainda não.*
1. E você indica mecanização pra todos os produtores, mesmo para os pequenos produtores? Você aconselha?
2. *Aconselho porque é uma maneira de cortar custos né. Principalmente agora na nossa região como está aumentando muito a incidência da fiscalização de mão-de-obra e aumentando também o custo por causa da colheita agora nessa época, (...) cinquenta por cento o custo por causa de registro, de encargo trabalhista, tudo. Isso sem falar em IPI, tudo que você tem que fornecer pro empregado. Então na mecanização não fica um tipo de serviço, como é que eu posso dizer, igualado ao manual, mas você ganha nessa economia de mão-de-obra né?*

1. Então o serviço, a mão-de-obra, o manual é melhor do que o mecanizado?
2. *Melhor que o mecanizado, sem dúvida. Sem dúvida, só que...*
1. Uh-huh. Mas só na colheita? Ou nas outras partes, não tem como né? Você descascar sem máquina, nem lavar sem um mecanismo, não tem como né...
2. *Não, não. Lavar sim. Lavar tem lavador antigo que faz o serviço muito bem né. Mas no caso...*
1. Mas não tão bem quanto uma máquina? Ou depende?
2. *Os lavadores antigo?*
1. É.
2. *Não, faz do mesmo jeito. Só que eles não vão ter esse fluxo contínuo que é necessário pra pra pro descascador de cereja trabalhar né.*
1. Certo. Entendi. Agora, a mão-de-obra no caso é melhor do que a máquina na colheita?
2. *Ah, sim, é melhor que a máquina. Estraga menos a planta, deixa pouco café pra trás, que no caso a máquina, mesmo passando a máquina, tem pessoas que não pagam né, tem pessoas que pagam o repasse depois ainda.*
1. Certo.
2. *Aí nesse repasse ainda tira o resto de café que ficou na planta e junta com o café catado no chão.*
1. Muito bem. Fausto é isso então. Eu queria mesmo essas informações sobre o cereja descascado. Você autoriza o uso da sua entrevista para o meu projeto de pesquisa da USP e também para o projeto COMET, também da USP?
2. *Ah, autorizo.*
1. Muito obrigada.
2. *De nada.*

ANEXO III

PROJETO NURC/SP

INQUERITO N? 18 - BOBINA N? 07 - INF. N? 23

Tipo de Inquerito: dialogo entre informante e documentador (DID)

Duracao: 45 minutos

Data do registro: 01/12/71

Tema: A casa, o terreno, vegetais, agricultura, animais, gado

Informante: Homem, 31 anos, solteiro, advogado, natural de Sao Paulo, pais paulistanos, 11 faixa etaria.

1 *Doc.* Dr. J. nos entao queriamos perguntar ao senhor... alguma coisa sobre a sua experiencia de fazenda experiencia nesse caso... ih ah no sentido de descrever por exemplo a casa da fazenda... pra comecar se o senhor

5 se lembra

Inf. LEMbro perfeitamente... born entao nos poderiamos comecar dizendo que:... eu is pra fazenda na infancia... em tempo de ferias... a casa da fazenda... ela era... uma casa antiga... tipo colonial brasileiro... janelas

10 LARgas...

e:... mesmo as fechaduras das portas daquelas

fechaduras GRANdes et cetera... depois:: o chao era tudo de tabua larga... e os moveis urn pouco:: misturados quer dizer moveis antigos

17

moveis novos... e:: sempre...

15 precisando fazer restauracoes et cetera

Doc. e iluminacao?

Inf. a iluminacao ja era corn luz eletrica... isso na fazenda... onde eu is aqui em Campinas... agora 1a pra pro lado de Barretos... papai tinha uma fazenda onde nao havia

20 ainda luz elétrica... mas já a uma zona mais nova também... então a própria casa era de construção recente... mas não tinha luz elétrica

Doc. e a iluminação era feita como?

Inf. então a:: a iluminação era feita com:: lâmpada...

25 lâmpada daqueles tipo Aladin... com camisinha... de::... até não sei de que que era feita a camisinha... mas era assim

Doc. e o tipo de terreno onde situava a fazenda era um terreno plano ou um terreno mais acidentado?

30 *Inf.* o terreno era... aqui em Campinas... tem uma parte acidentada... a fazenda era vamos dizer tinha teria duas

ANEXO IV

MODELO DE TRANSCRIÇÃO DA FERRAMENTA CHAT

@Begin

@Languages: pt, en

@Participants: IN1 Joao Subject, IN2 Luciana Investigator, JAM James

Subject, ANT Manager Subject, PAS Paschoal Subject

@ID: pt en|cafe|IN1|45|male|||Subject|superior|

@ID: pt en|cafe|IN2|33|female|||Investigator|superior|

@ID: en|cafe|JAM|42|male|||Subject|superior|

@ID: pt|cafe|ANT|35|male|||Manager|highschool|

@ID: pt|cafe|PAS|40|male|||Subject|superior|

@Filename: cafeen1.chat

@Date: 02-JUL-2004

@Location: Barreiro Rehder Farm, Sao Joao da Boa Vista, SP, Brazil

@Time Duration: 09:00 - 12:00

@Tape Location: Tape 15 Side A

@Situation: The company guide is asking the manager some questions.

They

are in front of the processing machine, watching how it works.

@Comment: there is a lot of noise, it is difficult to record the conversation

@Warning: a entonação fática não foi codificada

@Bg: en

***IN1:** xx hundred hectars.

%add: JAM

***IN1:** or so.

***JAM:** +< uhuh.

@Eg: en

@Bg: pt

***IN1:** ahn: e tem mais café aonde agora?

%add: ANT

***ANT: tem são roque da fartura, lá onde fechá[*], isso vai dá mais
duzentos hectares lá.**

%err: fechá= fechar

@Eg: pt

@Bg: en

***IN1: and the: in the other farm ahn: when they finish will be another
two hundred hectars.**

%add: JAM

@Eg: en

@Bg: pt

***IN1: faz o processamento todo. xxx.**

%add: ANT

***ANT: ele 0deve provavelmente plantá[*] mais uns duzentos.**

%err: plantá= plantar

***IN1: certo.**

@Eg: pt

@Bg: en

***IN1: so they are planning to: ahn plant another two hundred a year.**

%add: JAM

***JAM: ah, ok.**

***IN1: so you're [///] more or less looking to half of what you're going to
be in the future.**

@Comment: James is surprised.

***IN1: in this place, they did have some slope, but they ought to use
partially the slope.**

***JAM: uhuh.**

***IN1: but as you can see, the whole equipment, the whole set up is
designed on a flat ground.**

***JAM: +< ok.**

***IN1: again, if there's no slope available, there's no need to create
slope.**

***JAM: yeah.**

***IN1: so this is a: # if they didn't have this part and the slope, it
would run the same way.**

***JAM: right.**

@Eg: en

@Comment: they are walking to the driers.

%com: laugh

@Bg: pt

***PAS: e resolveu o problema, né?**

***ANT: é: resolve o problema.**

@End

ANEXO VÍndice numérico dos informantes**PB:**

- <1PB> Luciana
- <2PB> Márcia
- <3PB> Marcelo
- <4PB> Fausto
- <5PB> Márcio
- <6PB> Humberto
- <7PB> César
- <8PB> João Henrique
- <9PB> Marco B.
- <10PB> José Rubens
- <11 PB> Ricardo M.
- <12PB> Roseli
- <13PB> Roseli C.
- <14PB> Maria das Graças de Souza
- <15PB> Roberto do Amaral
- <16PB> Aparecida Donizete Gregório
- <17PB> Antonio F.
- <18PB> Bil
- <19PB> Sebastião Junqueira Filho
- <20PB> professora de curso para baristas
- <21PB> especialista em café
- <22PB> especialista em café
- <23PB> Paschoal
- <24PB> gerente Rehder
- <25PB> João Roberto
- <26PB> Claudinei
- <27PB> João S.
- <28PB> Ricardo Andradas
- <29PB> mulher da visita dos tailandeses

EN:

- <1EN> Luciana - Brazil
- <2EN> James Wright – Sri Lanka
- <3EN> João Henrique - Brazil
- <4EN> João S. - Brazil
- <5EN> Vietnamita
- <6EN> Jim - EUA
- <7EN> Vivian - Colombia
- <8EN> Tailandês
- <9EN> Tailandesa
- <10EN> Pauline - Uganda
- <11EN> Nathan - Uganda
- <12EN> Jean Nicolas - Bélgica
- <13EN> Chrispine - Quênia
- <14EN> David - EUA
- <15EN> Peter - Tanzânia
- <16EN> Daniel Zerayacob - Etiópia
- <17EN>Ulf Kusserow - Etiópia
- <18EN>Joachim W. - Alemanha
- <19EN> Elias Tamu - Tanzânia
- <20EN>Pierre Claver – Ruanda
- <21EN> Agrochemical - Etiópia
- <22EN> Noah - Etiópia
- <23EN> agricultor etíope 1
- <24EN> Coffee expert - Etiópia
- <25EN> T. M. - Etiópia
- <26en> Farmers - Etiópia
- <27en> Farmer 1 - Etiópia
- <28en> Farmer 2 - Etiópia
- <29en> J. O’neill
- <30en> mulher - Etiópia
- <31en> Barista - Itália
- <32en> E. I. – Itália
- <33en> Fabiana - Itália
- <34en> mulher - Etiópia
- <35en> entrevistador - UK

ANEXO VI

Dados sobre o corpus escrito

PB

FONTE	Assunto
www.wikipedia.org	Geral
http://www.abic.com.br/gar_qcafe.html	Qualidade
http://www.abic.com.br/jcafe/jcafe_ed156_p64a69.pdf	Qualidade/Máquinas
http://www22.sede.embrapa.br/cafe//ctp/programacao.htm	Geral
http://www.embrapa.br/linhas_de_acao/alimentos/cafe/index_html/mostra_documento	Geral (20/10/2004)
http://www.abic.com.br/scafe_historia.html	Geral
http://www.pacafes.com.br/default_noticias_interna.asp?cod_area=5&idNoticia=3860	Geral
http://www.fae.edu/publicacoes/pdf/revista_fae_business/n9/12_agronegocio.pdf	Mercado e qualidade
http://www.cenargen.embrapa.br/cenargenda/divulgacao2007/ocafezal160307.pdf	Variabilidade genética
http://www.coffeebreak.com.br/ocafezal.asp?SE=8&ID=384	Colheita; arborização; máquinas
http://www.cnpab.embrapa.br/publicacoes/sistemasdeproducao/cafe/colheita.htm	Colheita e pós-colheita
http://www.ufv.br/poscolheita/cafe/CBP&D.htm	Pós-colheita
http://www.biologico.sp.gov.br/NOTICIAS/cafezal_ib.htm	Colheita
http://www.revistacafeicultura.com.br/index.php?mat=3701&tipo=ler	Processamento
http://www.pimentamaquinas.com.br/poscolheita.htm	Processamento
http://www.agrotrends.com.br/portugues/cafe.htm	Máquinas
http://www.aenoticias.pr.gov.br/modules/news/print.php?storyid=9063	Exportação
http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612003000300015&lng=&nrm=iso&tlng=	Secagem
http://www.es.gov.br/site/noticias/show.aspx?noticiald=99668472	Colheita
http://www.ufv.br/dea/poscolheita/boletins/manual_de_constru%C3%A7%C3%A3o_e_manejo_de.htm	Secagem
http://www.seag.es.gov.br/materia595.htm	Colheita
http://www.sebrae.com.br/revistasebrae/09/cafe_02.htm	Processamento
http://www.criareplantar.com.br/agricultura/cafe/cafe.php?tipoConteudo=texto&idConteudo=1267	Processamento
http://www.fazendaalianca.com.br/site/etapas.php	Processamento
http://www.pinhalse.com.br/index-port.htm#	Equipamentos
http://www.bsca.com.br/index.php	Venda de café
http://www.qualicafex.com	Qualidade de café
http://www.sindicafesp.com.br/dload/cafe_resolucaoSAA37e7.pdf	Qualidade de café

Total de palavras: 100.274

EN

FONTE	Assunto
http://www.coffeereview.com/article.cfm?ID=133	Market
http://en.wikipedia.org/wiki/Coffee	General
http://www.ico.org/field_processing.asp	Processing
http://www.cosic.org/background-on-coffee/history-of-coffee	General
http://www.indiacoffee.org/relatedlinks/default.htm	Coffee in india
http://www.scaa.org/whatis_sustainability.asp	General
http://www.ncausa.org/i4a/pages/index.cfm?pageid=71	Pos-processing
http://www.coffeeresearch.org/agriculture/harvesting.htm	Harvesting
http://www.wsu.edu:8080/~gmhyde/433_web_pages/coffee/student-pages/3harvest/harvest.htm	Harvesting
http://home.howstuffworks.com/coffee4.htm	Processing
http://www.ecf-coffee.org/	Processing
http://www.schaerer.ca/plantation.htm	Processing
http://www.starbucks.com/ourcoffees/coffee_edu5.asp?category%5fname=coffee+trees+and+beans	Harvesting and processing
http://www.2basnob.com/coffee-roasting.html	General
http://www.newcrops.uq.edu.au/newslett/ncnl7-71.htm	Mechanisation
http://www.nasa.gov/centers/ames/news/releases/2002/02_27AR.html	Harvest
www.coffee-ota.org/3_2_harvest.asp	General
www.teacoffeeasia.com/news.asp?id=527	Harvesting
http://www.itdg.org/docs/technical_information_service/coffee.pdf	General
http://www.mpechicago.com/coffee/	Machines
www.pinhalse.com.br	Machines
http://www.nationalgeographic.com/coffee/ax/frame.html	Consumption and market
http://www.peets.com/learn/coffee_tasting.asp	Taste
http://www.arabica.com/coffee/article00046/default_.asp	General
http://www.newcrops.uq.edu.au/newslett/ncnl7-71.htm	Harvesting
www.cafebelmondo.com	General
http://www.coffee-ota.org/4_1_prevention.asp	General
www.usaid.gov/press/frontlines/fl_apr06/dialogue.htm - 36k -	General
http://www.fao.org/docrep/003/X6939E/X6939e12.htm	General
http://agnews.tamu.edu/dailynews/stories/AGPR/Sep1107b.htm	General
http://www.batianpeakcoffee.com/coffee_101.html	General

TOTAL: 102.379 palavras

ANEXO VII

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	Freq.	RC.	%	eyness	P
1	CAFÉ	1.003	1,98	0			.002,67	000000
2	VAI	416	0,82	13			.612,89	000000
3	GENTE	218	0,43	1			.940,60	000000
4	ESTÁ	172	0,34	0			.541,02	000000
5	TENHO	141	0,28	0			.263,19	000000
6	PODE	136	0,27	0			.218,39	000000
7	COLHEITA	128	0,25	0			.146,70	000000
8	CEREJA	104	0,20	0			.931,64	000000
9	VOCÊS	102	0,20	1			.902,48	000000
10	VERDE	98	0,19	0			.877,88	000000
11	ANO	123	0,24	63			.865,18	000000
12	MÁQUINA	95	0,19	0			.851,00	000000
13	VOU	99	0,20	5			.846,85	000000
14	ÁGUA	103	0,20	14			.837,30	000000
15	GRÃO	93	0,18	0			.833,08	000000
16	FAZ	106	0,21	45			.766,62	000000
17	HOJE	105	0,21	52			.742,39	000000
18	FICA	94	0,19	25			.720,26	000000
19	FAZER	92	0,18	36			.672,85	000000
20	COISA	71	0,14	1			.625,46	000000
21	TER	107	0,21	166			.596,70	000000
22	PRODUTOR	70	0,14	5			.590,39	000000
23	VINTE	74	0,15	13			.589,77	000000
24	TERREIRO	65	0,13	0			.582,23	000000
25	DÁ	107	0,21	190			.574,63	000000
26	SECADOR	63	0,12	0			.564,31	000000
27	VAMOS	68	0,13	7			.562,73	000000
28	OLHA	62	0,12	0			.555,35	000000
29	CHÃO	62	0,12	0			.555,35	000000
30	ONDE	61	0,12	0			.546,39	000000
31	BEBIDA	65	0,13	5			.546,32	000000
32	DIZER	66	0,13	8			.540,67	000000
33	GRÃOS	61	0,12	3			.522,24	000000
34	REGIÃO	78	0,15	54			.521,32	000000
35	DESCASCADO	58	0,11	0			.519,52	000000
36	BOM	74	0,15	42			.511,95	000000
37	BRASIL	73	0,14	46			.496,16	000000
38	LAVOURA	53	0,10	0			.474,73	000000
39	CASCA	55	0,11	3			.469,10	000000
40	QUÊ	52	0,10	2			.448,71	000000

WordSmith Tools 4.0 -- 28/11/2007

41	PÉ	87	0,17	177	448,69 000000
42	HORA	62	0,12	27	446,73 000000
43	DIA	99	0,20	309	441,73 000000
44	ENTENDI	49	0,10	0	438,89 000000
45	MESMO	104	0,20	383	435,23 000000
46	BÓIA	48	0,09	0	429,94 000000
47	VÃO	54	0,11	11	424,83 000000
48	CENTO	51	0,10	6	418,59 000000
49	ALI	52	0,10	9	414,93 000000
50	MIL	56	0,11	20	414,45 000000
51	NATURAL	54	0,11	15	411,77 000000
52	NADA	50	0,10	7	405,55 000000
53	PANO	46	0,09	1	402,36 000000
54	MERCADO	48	0,09	5	396,93 000000
55	CERTO	49	0,10	7	396,86 000000
56	TIPO	84	0,17	230	392,97 000000
57	TIRA	43	0,08	0	385,15 000000
58	CADA	43	0,08	0	385,15 000000
59	VERDADE	53	0,10	22	384,46 000000
60	TRINTA	47	0,09	6	383,68 000000
61	SEU	42	0,08	0	376,19 000000
62	SEM	42	0,08	0	376,19 000000
63	FAZENDA	42	0,08	0	376,19 000000
64	FALA	43	0,08	1	375,62 000000
65	CHAMA	46	0,09	6	374,96 000000
66	PERGAMINHO	41	0,08	0	367,23 000000
67	TEMPO	78	0,15	215	364,03 000000
68	PARTE	71	0,14	157	356,74 000000
69	DOIS	81	0,16	272	351,50 000000
70	LAVADOR	39	0,08	0	349,32 000000
71	UMIDADE	39	0,08	0	349,32 000000
72	SECAR	39	0,08	0	349,32 000000
73	POUCO	58	0,11	69	345,99 000000
74	VER	47	0,09	18	344,69 000000
75	VEM	53	0,10	43	343,67 000000
76	MEU	40	0,08	2	342,24 000000
77	QUAL	68	0,13	155	338,35 000000
78	QUALIDADE	63	0,12	115	335,59 000000
79	SECO	37	0,07	0	331,40 000000
80	SECA	43	0,08	11	330,80 000000
81	NOVO	48	0,09	33	321,19 000000

WordSmith Tools 4.0 -- 28/11/2007

82	QUANTO	76	0,15	275	320,27 000000
83	SEI	38	0,07	3	318,96 000000
84	TINHA	39	0,08	6	314,11 000000
85	TRABALHA	36	0,07	1	313,27 000000
86	COLHER	36	0,07	1	313,27 000000
87	MELHOR	63	0,12	144	313,19 000000
88	CINCO	50	0,10	50	310,36 000000
89	MÃO	39	0,08	7	310,24 000000
90	OUTRA	55	0,11	84	307,96 000000
91	DENTRO	51	0,10	58	307,48 000000
92	MADURO	34	0,07	0	304,53 000000
93	OUTRO	71	0,14	252	301,50 000000
94	NOME	36	0,07	3	301,36 000000
95	COLHE	33	0,07	0	295,57 000000
96	MUNDO	42	0,08	22	294,32 000000
97	TIRAR	35	0,07	3	292,56 000000
98	MINHA	35	0,07	4	287,78 000000
99	ACHO	32	0,06	0	286,61 000000
100	PREÇO	34	0,07	3	283,77 000000
101	LENHA	31	0,06	0	277,66 000000
102	QUANTIDADE	44	0,09	42	275,89 000000
103	TRÊS	63	0,12	210	274,15 000000
104	PRODUTO	40	0,08	25	272,23 000000
105	UNS	33	0,07	4	270,31 000000
106	VU	31	0,06	1	268,78 000000
107	EXISTE	47	0,09	67	268,00 000000
108	QUANTOS	33	0,07	5	266,09 000000
109	FICAR	34	0,07	8	263,81 000000
110	CIMA	31	0,06	2	262,61 000000
111	PROPRIEDADE	34	0,07	9	260,61 000000
112	LAVADO	29	0,06	0	259,74 000000
113	POUQUINHO	29	0,06	0	259,74 000000
114	PLANTA	29	0,06	0	259,74 000000
115	SAI	29	0,06	0	259,74 000000
116	MUITA	33	0,07	9	252,13 000000
117	TORRA	28	0,06	0	250,79 000000
118	TULHA	28	0,06	0	250,79 000000
119	MENOS	62	0,12	248	250,76 000000
120	VEZES	61	0,12	251	243,79 000000
121	TAVA	28	0,06	1	242,11 000000
122	CHEGA	31	0,06	7	241,51 000000

WordSmith Tools 4.0 -- 28/11/2007

123	FAZENDO	33	0,07	13	241,09 000000
124	PAS SA	36	0,07	25	240,44 000000
125	QUATRO	47	0,09	99	239,77 000000
126	S BMP RE	45	0,09	85	237,29 000000
127	SABE	40	0,08	49	236,92 000000
128	SES SENTA	29	0,06	4	235,46 000000
129	MANUAL	26	0,05	0	232,87 000000
130	CINQÜENTA	29	0,06	5	231,46 000000
131	PRODUTORES	28	0,06	3	231,14 000000
132	QUEM	30	0,06	10	223,94 000000
133	ACONTECE	32	0,06	17	223,74 000000
134	PROCES SAMENTO	27	0,05	3	222,39 000000
135	ESTÃO	52	0,10	185	220,59 000000
136	FALAR	28	0,06	6	219,23 000000
137	PODER	37	0,07	48	216,09 000000
138	SACAS	24	0,05	0	214,96 000000
139	DAQUI	24	0,05	0	214,96 000000
140	DES POL PADO	24	0,05	0	214,96 000000
141	DAÍ	30	0,06	16	209,62 000000
142	VARIEDADE	26	0,05	4	209,40 000000
143	EXBMP LO	46	0,09	136	209,34 000000
144	NO RMALMENTE	30	0,06	17	207,57 000000
145	FRUTO	24	0,05	1	206,58 000000
146	SEJA	23	0,05	0	206,00 000000
147	ARÁBICA	23	0,05	0	206,00 000000
148	SEGUINTE	32	0,06	29	202,86 000000
149	PRECISA	26	0,05	7	198,92 000000
150	MÁQUINAS	23	0,05	1	197,71 000000
151	SACO	22	0,04	0	197,04 000000
152	MECANIZAÇÃO	22	0,04	0	197,04 000000
153	FLORADA	22	0,04	0	197,04 000000
154	ENTENDEU	22	0,04	0	197,04 000000
155	TERRERO	22	0,04	0	197,04 000000
156	COMEÇA	24	0,05	3	196,19 000000
157	ESSES	47	0,09	184	191,79 000000
158	ERA	43	0,08	138	189,83 000000
159	SIM	30	0,06	28	189,00 000000
160	COLHIDO	22	0,04	1	188,84 000000
161	CAFÉS	21	0,04	0	188,09 000000
162	SACA	21	0,04	0	188,09 000000
163	POLPA	21	0,04	0	188,09 000000

WordSmith Tools 4.0 -- 28/11/2007

164	PRODUÇÃO	37	0,07	83	185,03 000000
165	VÊ	45	0,09	177	183,27 000000
166	TODA	30	0,06	34	181,00 000000
167	TANTO	47	0,09	211	180,86 000000
168	INTEIRO	21	0,04	1	179,97 000000
169	MEIO	40	0,08	122	179,97 000000
170	DIRETO	26	0,05	15	179,36 000000
171	SOLO	20	0,04	0	179,13 000000
172	SECAGEM	20	0,04	0	179,13 000000
173	ROBUSTA	20	0,04	0	179,13 000000
174	ALGUMA	27	0,05	21	176,52 000000
175	NESSE	35	0,07	82	172,59 000000
176	CAI	22	0,04	5	171,28 000000
177	FOLHA	20	0,04	1	171,11 000000
178	CAMINHÃO	20	0,04	1	171,11 000000
179	CONSEGUE	23	0,05	8	170,78 000000
180	GRANDE	44	0,09	195	170,28 000000
181	FAÇO	19	0,04	0	170,17 000000
182	BENEFICIAMENTO	19	0,04	0	170,17 000000
183	FALOU	19	0,04	0	170,17 000000
184	CARO	19	0,04	0	170,17 000000
185	ESPRESSO	19	0,04	0	170,17 000000
186	SOL	21	0,04	3	170,07 000000
187	NESSE	31	0,06	55	166,48 000000
188	PESSOA	26	0,05	23	165,65 000000
189	POSSO	19	0,04	1	162,25 000000
190	OBRA	19	0,04	1	162,25 000000
191	TRABALHAR	19	0,04	1	162,25 000000
192	MOLE	20	0,04	3	161,39 000000
193	AQUELA	23	0,05	12	161,27 000000
194	PLANTAR	18	0,04	0	161,22 000000
195	FERMENTAÇÃO	18	0,04	0	161,22 000000
196	CHUVA	18	0,04	0	161,22 000000
197	AQUILO	18	0,04	0	161,22 000000
198	CATUAÍ	18	0,04	0	161,22 000000
199	OITO	27	0,05	32	161,19 000000
200	NEM	32	0,06	75	157,77 000000
201	FAZEM	25	0,05	24	156,55 000000
202	DESSE	33	0,07	87	156,40 000000
203	DESSA	31	0,06	68	156,14 000000
204	BENEFICIAR	22	0,04	11	155,28 000000

WordSmith Tools 4.0 -- 28/11/2007

205	TIVER	20	0,04	5	154,22	000000
206	DEPENDE	27	0,05	39	153,42	000000
207	FALANDO	18	0,04	1	153,40	000000
208	PÕE	19	0,04	3	152,72	000000
209	FAZ	17	0,03	0	152,26	000000
210	LAVA	17	0,03	0	152,26	000000
211	JEITO	17	0,03	0	152,26	000000
212	GOSTO	17	0,03	0	152,26	000000
213	CARA	17	0,03	0	152,26	000000
214	PEGA	17	0,03	0	152,26	000000
215	NUNCA	23	0,05	17	151,84	000000
216	REAIS	20	0,04	6	151,18	000000
217	DEKAR	19	0,04	4	149,01	000000
218	TERRA	19	0,04	4	149,01	000000
219	DEZ	23	0,05	19	148,59	000000
220	NOVENTA	18	0,04	2	148,26	000000
221	PROPRIEDADES	24	0,05	25	147,62	000000
222	TOMAR	21	0,04	11	147,15	000000
223	MOMENTO	28	0,06	54	146,73	000000
224	ESTAVA	26	0,05	39	146,27	000000
225	SEPARADO	19	0,04	5	145,72	000000
226	TRABALHO	40	0,08	205	144,88	000000
227	FAZÊ	17	0,03	1	144,56	000000
228	VERMELHO	18	0,04	3	144,06	000000
229	SACOS	16	0,03	0	143,30	000000
230	MONTANHOSA	16	0,03	0	143,30	000000
231	QUARENTA	19	0,04	6	142,76	000000
232	PESSOAS	36	0,07	151	142,69	000000
233	MAIOR	63	0,12	722	0,02 142,14	000000
234	JUSTAMENTE	18	0,04	4	140,44	000000
235	AGOSTO	20	0,04	11	139,06	000000
236	PASSAR	19	0,04	8	137,54	000000
237	FORTE	23	0,05	28	136,43	000000
238	MINAS	16	0,03	1	135,72	000000
239	QUER	22	0,04	23	135,21	000000
240	COLHETA	15	0,03	0	134,35	000000
241	CAFEICULTURA	15	0,03	0	134,35	000000
242	ENGENHEIRO	15	0,03	0	134,35	000000
243	DESMUCILADO	15	0,03	0	134,35	000000
244	VENDER	15	0,03	0	134,35	000000
245	BOURBON	15	0,03	0	134,35	000000

WordSmith Tools 4.0 -- 28/11/2007

246	GUARDAR	15	0,03	0	134,35	000000
247	ACREDITO	15	0,03	0	134,35	000000
248	TÃO	25	0,05	45	133,70	000000
249	ASISTÊNCIA	18	0,04	7	131,73	000000
250	OTENTA	16	0,03	2	130,79	000000
251	DISSE	16	0,03	2	130,79	000000
252	USA	19	0,04	13	127,24	000000
253	EXATAMENTE	18	0,04	9	127,05	000000
254	ACHA	18	0,04	9	127,05	000000
255	COISAS	15	0,03	1	126,89	000000
256	BARRA	15	0,03	1	126,89	000000
257	SELETIVA	16	0,03	3	126,80	000000
258	CASO	37	0,07	213	126,65	000000
259	ELAS	21	0,04	24	126,45	000000
260	CHEGAR	17	0,03	6	125,99	000000
261	ANOS	58	0,11	700	0,02 125,90	000000
262	PRODUZIR	22	0,04	31	125,81	000000
263	VOLTA	20	0,04	19	125,52	000000
264	BROCA	14	0,03	0	125,39	000000
265	GEADA	14	0,03	0	125,39	000000
266	ILLY	14	0,03	0	125,39	000000
267	TRATOR	14	0,03	0	125,39	000000
268	TREIS	14	0,03	0	125,39	000000
269	PEGAR	14	0,03	0	125,39	000000
270	COMPRA	14	0,03	0	125,39	000000
271	NEGÓCIO	14	0,03	0	125,39	000000
272	COOPERATIVA	14	0,03	0	125,39	000000
273	PESSOAL	21	0,04	25	125,24	000000
274	PROBLEMA	22	0,04	32	124,78	000000
275	SABER	19	0,04	15	123,85	000000
276	PRIMEIRO	24	0,05	49	123,61	000000
277	FORA	18	0,04	11	122,97	000000
278	QUANTAS	15	0,03	2	122,08	000000
279	PRODUZ	19	0,04	17	120,76	000000
280	NENHUM	24	0,05	55	119,19	000000
281	NUMA	22	0,04	38	119,05	000000
282	ESSAS	28	0,06	100	118,59	000000
283	VENDO	14	0,03	1	118,06	000000
284	ACONTECER	16	0,03	6	117,66	000000
285	PENEIRA	13	0,03	0	116,43	000000
286	MADUROS	13	0,03	0	116,43	000000

WordSmith Tools 4.0 -- 28/11/2007

287	TIREI	13	0,03	0	116,43 000000
288	TERRENO	13	0,03	0	116,43 000000
289	ALQUEIRES	13	0,03	0	116,43 000000
290	ÁRVORE	13	0,03	0	116,43 000000
291	BENEFICIADO	13	0,03	0	116,43 000000
292	FEITA	13	0,03	0	116,43 000000
293	OBRIGADA	13	0,03	0	116,43 000000
294	PROCESSO	28	0,06	110	114,08 000000
295	DUZENTOS	14	0,03	2	113,38 000000
296	DZ	19	0,04	23	112,85 000000
297	TODO	36	0,07	247	112,41 000000
298	PÉS	15	0,03	5	111,97 000000
299	SERVIÇO	22	0,04	48	110,99 000000
300	RÁPIDO	16	0,03	9	110,84 000000
301	CAUSA	28	0,06	121	109,56 000000
302	DEU	15	0,03	6	109,36 000000
303	DEKA	15	0,03	6	109,36 000000
304	PERDE	13	0,03	1	109,25 000000
305	SETENTA	13	0,03	1	109,25 000000
306	AMARELO	13	0,03	1	109,25 000000
307	SEIS	24	0,05	72	108,63 000000
308	ALGUM	20	0,04	35	107,83 000000
309	UMAS	12	0,02	0	107,48 000000
310	MUCILAGEM	12	0,02	0	107,48 000000
311	ANTIGAMENTE	12	0,02	0	107,48 000000
312	MM	12	0,02	0	107,48 000000
313	ESTOU	12	0,02	0	107,48 000000
314	CABRAS	12	0,02	0	107,48 000000
315	FOTO	12	0,02	0	107,48 000000
316	MECANIZADO	12	0,02	0	107,48 000000
317	PAI	16	0,03	11	107,05 000000
318	DOZE	15	0,03	7	106,98 000000
319	DAR	17	0,03	16	106,91 000000
320	CLARO	17	0,03	16	106,91 000000
321	DIFERENÇA	30	0,06	160	106,61 000000
322	TA	35	0,07	256	105,46 000000
323	PARTIR	26	0,05	104	105,14 000000
324	CAFEÍNA	13	0,03	2	104,70 000000
325	GÁS	14	0,03	5	103,60 000000
326	EXISTEM	24	0,05	84	102,46 000000
327	MAIO	16	0,03	14	102,17 000000

WordSmith Tools 4.0 -- 28/11/2007

328	FRUTOS	13	0,03	3	101,06 000000
329	COMEÇO	13	0,03	3	101,06 000000
330	ESPECIAL	20	0,04	44	100,63 000000
331	BRASILEIRO	17	0,03	21	100,48 000000
332	ENTRA	12	0,02	1	100,45 000000
333	PALHA	12	0,02	1	100,45 000000
334	CLIMA	12	0,02	1	100,45 000000
335	INTERESSANTE	18	0,04	29	99,32 000000
336	FRENTE	16	0,03	16	99,31 000000
337	CERTEZA	14	0,03	7	98,81 000000
338	BONITO	11	0,02	0	98,52 000000
339	TREZENTOS	11	0,02	0	98,52 000000
340	ALQUEIRE	11	0,02	0	98,52 000000
341	LENDA	11	0,02	0	98,52 000000
342	TORRAR	11	0,02	0	98,52 000000
343	MATO	11	0,02	0	98,52 000000
344	QUERIA	11	0,02	0	98,52 000000
345	NOIS	11	0,02	0	98,52 000000
346	LAVORA	11	0,02	0	98,52 000000
347	DEMORA	11	0,02	0	98,52 000000
348	MOCA	11	0,02	0	98,52 000000
349	ALÉM	11	0,02	0	98,52 000000
350	ADONDE	11	0,02	0	98,52 000000
351	MATA	11	0,02	0	98,52 000000
352	PORCENTO	11	0,02	0	98,52 000000
353	TORRADOR	11	0,02	0	98,52 000000
354	CLASSIFICAÇÃO	19	0,04	38	98,48 000000
355	CONTA	17	0,03	23	98,24 000000
356	RUM	13	0,03	4	97,97 000000
357	JUNTO	16	0,03	17	97,97 000000
358	GANHA	12	0,02	2	96,04 000000
359	VEZ	26	0,05	129	95,61 000000
360	TERMOS	19	0,04	43	94,74 000000
361	SUL	16	0,03	21	93,17 000000
362	FEZ	15	0,03	15	93,10 000000
363	LUGAR	15	0,03	15	93,10 000000
364	DUAS	30	0,06	209	92,89 000000
365	QUINZE	13	0,03	6	92,87 000000
366	NAQUELE	13	0,03	6	92,87 000000
367	CUSTO	18	0,04	38	91,75 000000
368	TEMOS	18	0,04	38	91,75 000000

WordSmith Tools 4.0 -- 28/11/2007

369	DIFERENTE	18	0,04	38	91,75 000000]
370	NOVECENTOS	11	0,02	1	91,66 000000]
371	COLHENDO	11	0,02	1	91,66 000000]
372	LAVAGEM	11	0,02	1	91,66 000000]
373	QUILOS	11	0,02	1	91,66 000000]
374	CHAMAMOS	11	0,02	1	91,66 000000]
375	PEDRA	11	0,02	1	91,66 000000]
376	FAZENDAS	11	0,02	1	91,66 000000]
377	BAIXO	22	0,04	83	91,14 000000]
378	QUASE	18	0,04	40	90,28 000000]
379	SEPARAR	12	0,02	4	89,57 000000]
380	PANHA	10	0,02	0	89,56 000000]
381	ARMAZÉM	10	0,02	0	89,56 000000]
382	AGRÔNOMO	10	0,02	0	89,56 000000]
383	BARISTA	10	0,02	0	89,56 000000]
384	FOTOS	10	0,02	0	89,56 000000]
385	ADUBAÇÃO	10	0,02	0	89,56 000000]
386	TANQUE	10	0,02	0	89,56 000000]
387	NÓIS	10	0,02	0	89,56 000000]
388	ONTEM	10	0,02	0	89,56 000000]
389	CORTA	10	0,02	0	89,56 000000]
390	ESTRAGA	10	0,02	0	89,56 000000]
391	BALCÃO	10	0,02	0	89,56 000000]
392	SECÁ	10	0,02	0	89,56 000000]
393	MEIA	16	0,03	25	89,03 000000]
394	LADO	25	0,05	133	88,95 000000]
395	DISSO	21	0,04	77	88,01 000000]
396	FEITO	19	0,04	55	87,12 000000]
397	FICOU	15	0,03	20	87,00 000000]
398	ESTAMOS	12	0,02	5	86,99 000000]
399	ÉPOCA	14	0,03	14	86,89 000000]
400	PRETO	13	0,03	10	85,17 000000]
401	FERRUGEM	10	0,02	1	82,88 000000]
402	SAIU	10	0,02	1	82,88 000000]
403	DURO	10	0,02	1	82,88 000000]
404	COLOCAR	12	0,02	7	82,63 000000]
405	SEGUNDA	15	0,03	25	81,99 000000]
406	ENTRAR	11	0,02	4	81,21 000000]
407	DIAS	25	0,05	161	80,77 000000]
408	VAMO	9	0,02	0	80,61 000000]
409	MECANIZADA	9	0,02	0	80,61 000000]

WordSmith Tools 4.0 -- 28/11/2007

410	ÓTIMO	9	0,02	0	80,61 000000
411	HEIN	9	0,02	0	80,61 000000
412	NASCER	9	0,02	0	80,61 000000
413	RAMA	9	0,02	0	80,61 000000
414	EMBAIXO	9	0,02	0	80,61 000000
415	DESPOLPADOR	9	0,02	0	80,61 000000
416	MUDAS	9	0,02	0	80,61 000000
417	BICHO	9	0,02	0	80,61 000000
418	SENHOR	9	0,02	0	80,61 000000
419	TIROU	9	0,02	0	80,61 000000
420	VARRIÇÃO	9	0,02	0	80,61 000000
421	PLANTAÇÃO	9	0,02	0	80,61 000000
422	VENDE	9	0,02	0	80,61 000000
423	PETRÓLEO	9	0,02	0	80,61 000000
424	TÊM	32	0,06	316	80,13 000000
425	VANTAGEM	12	0,02	9	79,00 000000
426	COMPENSA	10	0,02	2	78,79 000000
427	COMEÇOU	10	0,02	2	78,79 000000
428	INTERNO	11	0,02	5	78,76 000000
429	PASSADO	14	0,03	21	78,76 000000
430	GERALMENTE	19	0,04	74	77,69 000000
431	TEMPERATURA	12	0,02	10	77,39 000000
432	COMUM	17	0,03	54	75,33 000000
433	LEVAR	16	0,03	44	74,72 000000
434	MANEIRA	20	0,04	96	74,67 000000
435	USAM	11	0,02	7	74,62 000000
436	ALGUÉM	9	0,02	1	74,13 000000
437	DEDO	9	0,02	1	74,13 000000
438	SAFRA	9	0,02	1	74,13 000000
439	PEDRAS	9	0,02	1	74,13 000000
440	ONZE	9	0,02	1	74,13 000000
441	TODAS	19	0,04	83	73,99 000000
442	ALTO	20	0,04	99	73,62 000000
443	LEVA	16	0,03	46	73,55 000000
444	PRIMEIRA	18	0,04	71	73,21 000000
445	PARTES	10	0,02	4	72,90 000000
446	CAIU	10	0,02	4	72,90 000000
447	TARDE	11	0,02	8	72,84 000000
448	PRODUZIDO	11	0,02	8	72,84 000000
449	CALOR	11	0,02	8	72,84 000000
450	MESMA	21	0,04	119	72,44 000000

WordSmith Tools 4.0 -- 28/11/2007

451	UTILIZA	12	0,02	14	71,91 000000
452	LIMPO	8	0,02	0	71,65 000000
453	LAVÁ	8	0,02	0	71,65 000000
454	DEZESSEIS	8	0,02	0	71,65 000000
455	LAVAR	8	0,02	0	71,65 000000
456	SOU	8	0,02	0	71,65 000000
457	SECANDO	8	0,02	0	71,65 000000
458	TALHÃO	8	0,02	0	71,65 000000
459	FOLHAS	8	0,02	0	71,65 000000
460	ATRÁS	8	0,02	0	71,65 000000
461	LEVO	8	0,02	0	71,65 000000
462	SOBROU	8	0,02	0	71,65 000000
463	DEKÁ	8	0,02	0	71,65 000000
464	COMMODITY	8	0,02	0	71,65 000000
465	SÁBADO	8	0,02	0	71,65 000000
466	COMEÇAR	8	0,02	0	71,65 000000
467	COLÔMBIA	8	0,02	0	71,65 000000
468	PASCHOAL	8	0,02	0	71,65 000000
469	ACONTECEU	11	0,02	9	71,20 000000
470	CHEGOU	10	0,02	5	70,58 000000
471	SETE	14	0,03	31	70,30 000000
472	BAR	9	0,02	2	70,22 000000
473	VERTICAL	9	0,02	2	70,22 000000
474	SENÃO	9	0,02	2	70,22 000000
475	DINHEIRO	9	0,02	2	70,22 000000
476	SEPARA	9	0,02	2	70,22 000000
477	PONTO	20	0,04	110	70,02 000000
478	LÓGICO	11	0,02	10	69,68 000000
479	ESTAR	20	0,04	112	69,40 000000
480	SERIA	20	0,04	114	68,79 000000
481	CONSUMO	21	0,04	133	68,44 000000
482	PEQUENO	14	0,03	35	67,56 000000
483	MUNDIAL	12	0,02	18	67,51 000000
484	TRABALHAM	9	0,02	3	67,18 000000
485	GOSTARIA	9	0,02	3	67,18 000000
486	DELA	10	0,02	7	66,69 000000
487	DEFEITO	12	0,02	19	66,53 000000
488	SABOR	8	0,02	1	65,39 000000
489	INGLÊS	8	0,02	1	65,39 000000
490	JOGA	8	0,02	1	65,39 000000
491	GOSTA	8	0,02	1	65,39 000000

WordSmith Tools 4.0 -- 28/11/2007

492	TRABALHADOR	8	0,02	1	65,39 000000
493	PEGO	8	0,02	1	65,39 000000
494	OIA	8	0,02	1	65,39 000000
495	ENTREVISTA	10	0,02	8	65,01 000000
496	SEPARAÇÃO	10	0,02	8	65,01 000000
497	JOÃO	10	0,02	8	65,01 000000
498	BATATA	9	0,02	4	64,65 000000
499	ALGUMAS	17	0,03	79	64,41 000000
500	HORAS	19	0,04	114	63,68 000000

ANEXO VIII

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	. Freq.	RC.	%	eyness	P
1	CAFÉ	1.003	INF	0			.002,67	000000
2	COLHEITA	128	INF	0			.146,70	000000
3	CEREJA	104	INF	0			.931,64	000000
4	ÁGUA	103	INF	14	INF		.837,30	000000
5	VERDE	98	INF	0			.877,88	000000
6	MÁQUINA	95	INF	0			.851,00	000000
7	GRÃO	93	INF	0			.833,08	000000
8	PÉ	87	INF	177	INF		.448,69	000000
9	TIPO	84	INF	230	INF		.392,97	000000
10	REGIÃO	78	INF	54	INF		.521,32	000000
11	TEMPO	78	INF	215	INF		.364,03	000000
12	BOM	74	INF	42	INF		.511,95	000000
13	PARTE	71	INF	157	INF		.356,74	000000
14	PRODUTOR	70	INF	5	INF		.590,39	000000
15	TERREIRO	65	INF	0			.582,23	000000
16	BEBIDA	65	INF	5	INF		.546,32	000000
17	SECADOR	63	INF	0			.564,31	000000
18	QUALIDADE	63	INF	115	INF		.335,59	000000
19	MELHOR	63	INF	144	INF		.313,19	000000
20	MAIOR	63	INF	722	INF		.142,14	000000
21	CHÃO	62	INF	0			.555,35	000000
22	DESCASCADO	58	INF	0			.519,52	000000
23	CASCA	55	INF	3	INF		.469,10	000000
24	NATURAL	54	INF	15	INF		.411,77	000000
25	LAVOURA	53	INF	0			.474,73	000000
26	BÓIA	48	INF	0			.429,94	000000
27	MERCADO	48	INF	5	INF		.396,93	000000
28	NOVO	48	INF	33	INF		.321,19	000000
29	PÃO	46	INF	1	INF		.402,36	000000
30	EXEMPLO	46	INF	136	INF		.209,34	000000
31	QUANTIDADE	44	INF	42	INF		.275,89	000000
32	GRANDE	44	INF	195	INF		.170,28	000000
33	SECA	43	INF	11	INF		.330,80	000000
34	FAZENDA	42	INF	0			.376,19	000000
35	MUNDO	42	INF	22	INF		.294,32	000000
36	PERGAMINHO	41	INF	0			.367,23	000000
37	PRODUTO	40	INF	25	INF		.272,23	000000
38	TRABALHO	40	INF	205	INF		.144,88	000000
39	LAVADOR	39	INF	0			.349,32	000000
40	UMIDADE	39	INF	0			.349,32	000000

WordSmith Tools 4.0 -- 28/11/2007

41	MÃO	39	INF	7	INF	310,24	000000
42	SECO	37	INF	0		331,40	000000
43	PRODUÇÃO	37	INF	83	INF	185,03	000000
44	NOME	36	INF	3	INF	301,36	000000
45	MADURO	34	INF	0		304,53	000000
46	PREÇO	34	INF	3	INF	283,77	000000
47	PROPRIEDADE	34	INF	9	INF	260,61	000000
48	LENHA	31	INF	0		277,66	000000
49	DIFERENÇA	30	INF	160	INF	106,61	000000
50	PLANTA	29	INF	0		259,74	000000
51	LAVADO	29	INF	0		259,74	000000
52	TORRA	28	INF	0		250,79	000000
53	TULHA	28	INF	0		250,79	000000
54	MOMENTO	28	INF	54	INF	146,73	000000
55	PROCESSO	28	INF	110	INF	114,08	000000
56	PROCESSAMENTO	27	INF	3	INF	222,39	000000
57	MANUAL	26	INF	0		232,87	000000
58	VARIEDADE	26	INF	4	INF	209,40	000000
59	DIRETO	26	INF	15	INF	179,36	000000
60	PESOA	26	INF	23	INF	165,65	000000
61	VEZ	26	INF	129	INF	95,61	000000
62	LADO	25	INF	133	INF	88,95	000000
63	DESPOLPADO	24	INF	0		214,96	000000
64	FRUTO	24	INF	1	INF	206,58	000000
65	ARÁBICA	23	INF	0		206,00	000000
66	FORTE	23	INF	28	INF	136,43	000000
67	TERRERO	22	INF	0		197,04	000000
68	MECANIZAÇÃO	22	INF	0		197,04	000000
69	SACO	22	INF	0		197,04	000000
70	FLORADA	22	INF	0		197,04	000000
71	COLHIDO	22	INF	1	INF	188,84	000000
72	PROBLEMA	22	INF	32	INF	124,78	000000
73	SERVIÇO	22	INF	48	INF	110,99	000000
74	BAIXO	22	INF	83	INF	91,14	000000
75	POLPA	21	INF	0		188,09	000000
76	CAFÉS	21	INF	0		188,09	000000
77	SACA	21	INF	0		188,09	000000
78	INTEIRO	21	INF	1	INF	179,97	000000
79	SOL	21	INF	3	INF	170,07	000000
80	PESSOAL	21	INF	25	INF	125,24	000000
81	CONSUMO	21	INF	133	INF	68,44	000000

WordSmith Tools 4.0 -- 28/11/2007

82	SECAGEM	20	INF	0	179,13	000000
83	SOLO	20	INF	0	179,13	000000
84	ROBUSTA	20	INF	0	179,13	000000
85	FOLHA	20	INF	1	INF 171,11	000000
86	CAMINHÃO	20	INF	1	INF 171,11	000000
87	MOLE	20	INF	3	INF 161,39	000000
88	VOLTA	20	INF	19	INF 125,52	000000
89	ESPECIAL	20	INF	44	INF 100,63	000000
90	MANEIRA	20	INF	96	INF 74,67	000000
91	ALTO	20	INF	99	INF 73,62	000000
92	PONTO	20	INF	110	INF 70,02	000000
93	ESPRESSO	19	INF	0	170,17	000000
94	BENEFICIAMENTO	19	INF	0	170,17	000000
95	CARO	19	INF	0	170,17	000000
96	OBRA	19	INF	1	INF 162,25	000000
97	TERRA	19	INF	4	INF 149,01	000000
98	SEPARADO	19	INF	5	INF 145,72	000000
99	CLASSIFICAÇÃO	19	INF	38	INF 98,48	000000
100	TERMOS	19	INF	43	INF 94,74	000000
101	CATUAÍ	18	INF	0	161,22	000000
102	CHUVA	18	INF	0	161,22	000000
103	FERMENTAÇÃO	18	INF	0	161,22	000000
104	VERMELHO	18	INF	3	INF 144,06	000000
105	ASSISTÊNCIA	18	INF	7	INF 131,73	000000
106	INTERESSANTE	18	INF	29	INF 99,32	000000
107	CUSTO	18	INF	38	INF 91,75	000000
108	DIFERENTE	18	INF	38	INF 91,75	000000
109	JEITO	17	INF	0	152,26	000000
110	GOSTO	17	INF	0	152,26	000000
111	CARA	17	INF	0	152,26	000000
112	CLARO	17	INF	16	INF 106,91	000000
113	BRASILEIRO	17	INF	21	INF 100,48	000000
114	CONTA	17	INF	23	INF 98,24	000000
115	COMUM	17	INF	54	INF 75,33	000000
116	MONTANHOSA	16	INF	0	143,30	000000
117	SELETIVA	16	INF	3	INF 126,80	000000
118	RÁPIDO	16	INF	9	INF 110,84	000000
119	PAI	16	INF	11	INF 107,05	000000
120	MEIA	16	INF	25	INF 89,03	000000
121	BOURBON	15	INF	0	134,35	000000
122	COLHETA	15	INF	0	134,35	000000

WordSmith Tools 4.0 -- 28/11/2007

123	ENGENHEIRO	15	INF	0		134,35	000000
124	CAFEICULTURA	15	INF	0		134,35	000000
125	DESMUCILADO	15	INF	0		134,35	000000
126	BARRA	15	INF	1	INF	126,89	000000
127	LUGAR	15	INF	15	INF	93,10	000000
128	BROCA	14	INF	0		125,39	000000
129	NEGÓCIO	14	INF	0		125,39	000000
130	GEADA	14	INF	0		125,39	000000
131	TRATOR	14	INF	0		125,39	000000
132	CO OPERATMA	14	INF	0		125,39	000000
133	GÁS	14	INF	5	INF	103,60	000000
134	ÉPOCA	14	INF	14	INF	86,89	000000
135	PASSADO	14	INF	21	INF	78,76	000000
136	PEQUENO	14	INF	35	INF	67,56	000000
137	MADUROS	13	INF	0		116,43	000000
138	TERRENO	13	INF	0		116,43	000000
139	ALQUEIRES	13	INF	0		116,43	000000
140	PENEIRA	13	INF	0		116,43	000000
141	ÁRVORE	13	INF	0		116,43	000000
142	OBRIGADA	13	INF	0		116,43	000000
143	BENEFICIADO	13	INF	0		116,43	000000
144	AMARELO	13	INF	1	INF	109,25	000000
145	CAFEÍNA	13	INF	2	INF	104,70	000000
146	RUM	13	INF	4	INF	97,97	000000
147	PRETO	13	INF	10	INF	85,17	000000
148	MUCILAGEM	12	INF	0		107,48	000000
149	FOTO	12	INF	0		107,48	000000
150	CABRAS	12	INF	0		107,48	000000
151	MECANIZADO	12	INF	0		107,48	000000
152	CLIMA	12	INF	1	INF	100,45	000000
153	PALHA	12	INF	1	INF	100,45	000000
154	VANTAGEM	12	INF	9	INF	79,00	000000
155	TEMPERATURA	12	INF	10	INF	77,39	000000
156	MUNDIAL	12	INF	18	INF	67,51	000000
157	DEFEITO	12	INF	19	INF	66,53	000000
158	MATO	11	INF	0		98,52	000000
159	BONITO	11	INF	0		98,52	000000
160	LAVORA	11	INF	0		98,52	000000
161	MATA	11	INF	0		98,52	000000
162	PORCENTO	11	INF	0		98,52	000000
163	DEMORA	11	INF	0		98,52	000000

WordSmith Tools 4.0 -- 28/11/2007

164	MOCA	11	INF	0		98,52	000000
165	LENDA	11	INF	0		98,52	000000
166	ALQUEIRE	11	INF	0		98,52	000000
167	TORRADOR	11	INF	0		98,52	000000
168	QUILOS	11	INF	1	INF	91,66	000000
169	LAVAGEM	11	INF	1	INF	91,66	000000
170	PEDRA	11	INF	1	INF	91,66	000000
171	INTERNO	11	INF	5	INF	78,76	000000
172	CALOR	11	INF	8	INF	72,84	000000
173	PRODUZIDO	11	INF	8	INF	72,84	000000
174	LÓGICO	11	INF	10	INF	69,68	000000
175	AGRÔNOMO	10	INF	0		89,56	000000
176	BARISTA	10	INF	0		89,56	000000
177	PANHA	10	INF	0		89,56	000000
178	BALCÃO	10	INF	0		89,56	000000
179	ADUBAÇÃO	10	INF	0		89,56	000000
180	ARMAZÉM	10	INF	0		89,56	000000
181	TANQUE	10	INF	0		89,56	000000
182	FERRUGEM	10	INF	1	INF	82,88	000000
183	DURO	10	INF	1	INF	82,88	000000
184	ENTREVISTA	10	INF	8	INF	65,01	000000
185	SEPARAÇÃO	10	INF	8	INF	65,01	000000
186	BICHO	9	INF	0		80,61	000000
187	SENHOR	9	INF	0		80,61	000000
188	PLANTAÇÃO	9	INF	0		80,61	000000
189	RAMA	9	INF	0		80,61	000000
190	MECANIZADA	9	INF	0		80,61	000000
191	MUDAS	9	INF	0		80,61	000000
192	ÓTIMO	9	INF	0		80,61	000000
193	DESPOLPADOR	9	INF	0		80,61	000000
194	VARRIÇÃO	9	INF	0		80,61	000000
195	PETRÓLEO	9	INF	0		80,61	000000
196	DEDO	9	INF	1	INF	74,13	000000
197	SAFRA	9	INF	1	INF	74,13	000000
198	PEDRAS	9	INF	1	INF	74,13	000000
199	DINHEIRO	9	INF	2	INF	70,22	000000
200	BAR	9	INF	2	INF	70,22	000000
201	VERTICAL	9	INF	2	INF	70,22	000000
202	SEPARA	9	INF	2	INF	70,22	000000
203	BATATA	9	INF	4	INF	64,65	000000
204	LIMPO	8	INF	0		71,65	000000

WordSmith Tools 4.0 -- 28/11/2007

205	TALHÃO	8	INF	0		71,65	000000
206	COMMODITY	8	INF	0		71,65	000000
207	TRABALHADOR	8	INF	1	INF	65,39	000000
208	SABOR	8	INF	1	INF	65,39	000000
209	INGLÊS	8	INF	1	INF	65,39	000000

ANEXO IX

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	. Freq.	RC.	%	eyness	P
1	COFFEE	370	1,54	103			.542,24	000000
2	T	123	0,51	223	0,02		.564,67	000000
3	OH	95	0,40	67			.560,60	000000
4	DON	72	0,30	71			.393,78	000000
5	HERE	122	0,51	589	0,04		.367,23	000000
6	ARABICA	46	0,19	2			.360,10	000000
7	CHERRIES	33	0,14	1			.261,20	000000
8	TWENTY	40	0,17	21			.249,68	000000
9	HUNDRED	44	0,18	47			.235,80	000000
10	ROBUSTA	28	0,12	2			.214,62	000000
11	RIGHT	77	0,32	546	0,04		.182,87	000000
12	DRY	38	0,16	63			.179,49	000000
13	FARMER	28	0,12	13			.178,47	000000
14	SEVENTY	24	0,10	3			.177,76	000000
15	BIRR	21	0,09	0			.171,93	000000
16	GONNA	28	0,12	17			.170,15	000000
17	FARMERS	38	0,16	82			.164,06	000000
18	COFFEES	20	0,08	2			.150,41	000000
19	LL	19	0,08	2			.142,41	000000
20	THOUSAND	29	0,12	44			.140,82	000000
21	SIXTY	21	0,09	7			.140,68	000000
22	CROP	30	0,12	52			.139,67	000000
23	HECTARES	19	0,08	4			.134,44	000000
24	PROCESSING	25	0,10	27			.133,58	000000
25	LIKE	104	0,43	1.506	0,11		.131,35	000000
26	TREE	27	0,11	46			.126,41	000000
27	HARVEST	23	0,10	24			.123,98	000000
28	MÁQUINA	15	0,06	0			.122,80	000000
29	EIGHTY	17	0,07	3			.122,37	000000
30	KNOW	65	0,27	627	0,04		.122,17	000000
31	UGANDA	17	0,07	6			.112,98	000000
32	DOESN	14	0,06	1			.107,30	000000
33	FLOATERS	14	0,06	1			.107,30	000000
34	VE	15	0,06	3			.106,69	000000
35	NINETY	16	0,07	7			.102,96	000000
36	USE	52	0,22	490	0,03		.99,64	000000
37	HARVESTING	14	0,06	3			.98,87	000000
38	SOIL	18	0,07	18			.98,07	000000
39	FORTY	16	0,07	10			.96,68	000000
40	NINETEEN	14	0,06	4			.95,68	000000

WordSmith Tools 4.0 -- 28/11/2007

41	SEE	58	0,24	655	0,05	94,75	000000
42	DRYER	13	0,05	2		94,72	000000
43	RE	29	0,12	125		92,64	000000
44	KILO	11	0,05	0		90,05	000000
45	FIFTY	16	0,07	14		90,01	000000
46	QUALITY	32	0,13	183	0,01	87,26	000000
47	BAGS	20	0,08	45		85,02	000000
48	PRODUCTION	37	0,15	281	0,02	83,69	000000
49	VARIETIES	13	0,05	6		82,93	000000
50	PLANTING	14	0,06	11		80,69	000000
51	ETHIOPIAN	11	0,05	2		78,96	000000
52	PRUNING	11	0,05	2		78,96	000000
53	PICK	25	0,10	111		78,63	000000
54	PICKING	18	0,07	40		76,87	000000
55	DRYING	11	0,05	3		75,61	000000
56	BEAN	12	0,05	6		75,53	000000
57	KILOS	9	0,04	0		73,68	000000
58	RWANDA	12	0,05	7		73,47	000000
59	GOOD	56	0,23	784	0,06	73,41	000000
60	REALLY	36	0,15	326	0,02	71,24	000000
61	THINK	48	0,20	606	0,04	70,27	000000
62	KIND	30	0,12	217	0,02	70,23	000000
63	GREEN	24	0,10	123		69,79	000000
64	COUNTRIES	31	0,13	242	0,02	68,73	000000
65	GET	66	0,27	1.125	0,08	68,12	000000
66	MULCHING	9	0,04	1		67,21	000000
67	GOING	50	0,21	694	0,05	66,17	000000
68	FME	46	0,19	591	0,04	66,13	000000
69	REAS	8	0,03	0		65,49	000000
70	MUCILAGE	8	0,03	0		65,49	000000
71	VINTE	8	0,03	0		65,49	000000
72	WASHED	12	0,05	12		65,37	000000
73	BERRY	13	0,05	20		62,85	000000
74	BAG	14	0,06	28		62,09	000000
75	LITERS	8	0,03	1		59,25	000000
76	EXPORTERS	12	0,05	18		58,47	000000
77	CENTO	7	0,03	0		57,31	000000
78	PULPING	7	0,03	0		57,31	000000
79	MOKA	7	0,03	0		57,31	000000
80	MACHINES	14	0,06	35		57,16	000000
81	NORMALLY	15	0,06	46		56,30	000000

WordSmith Tools 4.0 -- 28/11/2007

82	DIFFERENT	32	0,13	335	0,02	56,00	000000
83	ETHIOPIA	9	0,04	5		55,60	000000
84	TANZANIA	9	0,04	5		55,60	000000
85	EXPORT	16	0,07	60		54,78	000000
86	EXPORTER	8	0,03	3		52,70	000000
87	ELEVATOR	8	0,03	3		52,70	000000
88	RIPEN	7	0,03	1		51,31	000000
89	FLOATER	6	0,02	0		49,12	000000
90	CHERRY	9	0,04	9		49,03	000000
91	COME	36	0,15	491	0,03	48,56	000000
92	ACTUALLY	21	0,09	157	0,01	48,03	000000
93	AREA	27	0,11	278	0,02	47,95	000000
94	DOING	25	0,10	238	0,02	47,48	000000
95	LOT	28	0,12	316	0,02	45,74	000000
96	NOWADAYS	9	0,04	12		45,40	000000
97	SELECTING	7	0,03	3		45,19	000000
98	THIRTY	9	0,04	13		44,35	000000
99	BETTER	34	0,14	485	0,03	43,63	000000
100	PICKERS	6	0,02	1		43,41	000000
101	HECTARE	6	0,02	1		43,41	000000
102	RAIN	16	0,07	94		42,91	000000
103	PRICE	43	0,18	753	0,05	42,78	000000
104	MACHINE	12	0,05	42		42,45	000000
105	THANK	11	0,05	32		42,23	000000
106	STORAGE	9	0,04	16		41,55	000000
107	FROST	8	0,03	10		41,10	000000
108	PINHALENSE	5	0,02	0		40,93	000000
109	ROASTERS	5	0,02	0		40,93	000000
110	PULPER	5	0,02	0		40,93	000000
111	QUATRO	5	0,02	0		40,93	000000
112	HARAR	5	0,02	0		40,93	000000
113	PRUNE	5	0,02	0		40,93	000000
114	DIDN	5	0,02	0		40,93	000000
115	SESSENTA	5	0,02	0		40,93	000000
116	PEABERRY	5	0,02	0		40,93	000000
117	MINAS	5	0,02	0		40,93	000000
118	DÁ	5	0,02	0		40,93	000000
119	IMPORTANT	24	0,10	261	0,02	40,58	000000
120	MEND	6	0,02	2		40,19	000000
121	SPECIALTY	8	0,03	11		40,00	000000
122	PLANT	19	0,08	161	0,01	39,61	000000

WordSmith Tools 4.0 -- 28/11/2007

123	DEPENDING	10	0,04	28		39,01	000000]
124	TAKE	44	0,18	836	0,06	38,98	000000]
125	GROW	14	0,06	78		38,77	000000]
126	FIFTEEN	7	0,03	7		38,13	000000]
127	OPERATME	8	0,03	13		38,02	000000]
128	FERTILIZER	6	0,02	3		37,76	000000]
129	FARM	13	0,05	68		37,36	000000]
130	BRAZIL	14	0,06	83		37,34	000000]
131	SEEDS	8	0,03	14		37,12	000000]
132	VARIETY	13	0,05	69		37,04	000000]
133	TONS	8	0,03	15		36,28	000001]
134	TO O	36	0,15	626	0,04	36,15	000001]
135	VARIABILITY	6	0,02	4		35,79	000001]
136	BRANCHES	9	0,04	25		35,22	000004]
137	TREES	11	0,05	48		34,91	000005]
138	PEOPLE	63	0,26	1.578	0,11	34,60	000011]
139	GOES	17	0,07	150	0,01	34,33	000017]
140	PLANTS	16	0,07	131		34,24	000020]
141	DEPENDS	9	0,04	27		34,10	000023]
142	COMES	21	0,09	241	0,02	33,77	000033]
143	CBD	4	0,02	0		32,75	000076]
144	COMPENSA	4	0,02	0		32,75	000076]
145	MUNDO	4	0,02	0		32,75	000076]
146	TRATOR	4	0,02	0		32,75	000076]
147	CUSTO	4	0,02	0		32,75	000076]
148	WASHER	4	0,02	0		32,75	000076]
149	HORA	4	0,02	0		32,75	000076]
150	DUZENTOS	4	0,02	0		32,75	000076]
151	MÉDIA	4	0,02	0		32,75	000076]
152	FERTILIZERS	4	0,02	0		32,75	000076]
153	DIA	4	0,02	0		32,75	000076]
154	PESSOAS	4	0,02	0		32,75	000076]
155	SEPARATOR	4	0,02	0		32,75	000076]
156	SIDAMO	4	0,02	0		32,75	000076]
157	TON	6	0,02	6		32,69	000079]
158	DISEASES	12	0,05	69		32,61	000084]
159	NOW	59	0,25	1.474	0,10	32,57	000086]
160	LEAF	6	0,02	7		31,41	000180]
161	CAFÉ	6	0,02	7		31,41	000180]
162	KENYA	7	0,03	14		31,04	000223]
163	PULPED	5	0,02	3		30,45	000314]

WordSmith Tools 4.0 -- 28/11/2007

164	PROBLEM	23	0,10	320	0,02	30,34	000333
165	CHAT	8	0,03	24		30,31	000339
166	PRODUCE	12	0,05	78		30,18	000364
167	NATURAL	14	0,06	120		28,92	000724
168	ORGANIC	7	0,03	17		28,90	000731
169	STARBUCKS	5	0,02	4		28,70	000815
170	WERE	17	0,07	2.943	0,21	-28,92	000724
171	CENT	5	0,02	1.679	0,12	-29,28	000597
172	BUSINESS	4	0,02	1.587	0,11	-30,30	000341
173	HER	3	0,01	1.543	0,11	-33,05	000061
174	HAD	15	0,06	3.362	0,24	-43,58	000000
175	SAYS	3	0,01	1.898	0,13	-43,76	000000
176	TIMES	12	0,05	3.436	0,24	-54,15	000000
177	HIS	20	0,08	5.615	0,40	-87,36	000000

ANEXO X

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	. Freq.	RC.	%	eyness	P	emmas
1	ARABICA	46	INF	2	INF	360,10	000000		
2	AREA	27	INF	278	INF	47,95	000000		
3	BAG	14	INF	28	INF	62,09	000000		bags
4	BEAN	12	INF	6	INF	75,53	000000		
5	BERRY	13	INF	20	INF	62,85	000000		
6	BETTER	34	INF	485	INF	43,63	000000		
7	BRANCHES	9	INF	25	INF	35,22	000004		
8	BUSINESS	4	INF	1.587	INF	-30,30	000341		
9	CBD	4	INF	0		32,75	000076		
10	CHAT	8	INF	24	INF	30,31	000339		
11	CHERRIES	33	INF	1	INF	261,20	000000		
12	CHERRY	9	INF	9	INF	49,03	000000		
13	COFFEE	370	INF	103	INF	.542,24	000000		
14	COFFEES	20	INF	2	INF	150,41	000000		
15	COUNTRIES	31	INF	242	INF	68,73	000000		
16	CROP	30	INF	52	INF	139,67	000000		
17	DIFFERENT	32	INF	335	INF	56,00	000000		
18	DISEASES	12	INF	69	INF	32,61	000084		
19	DRY	38	INF	63	INF	179,49	000000		
20	DRYER	13	INF	2	INF	94,72	000000		
21	DRYING	11	INF	3	INF	75,61	000000		
22	ELEVATOR	8	INF	3	INF	52,70	000000		
23	EXPORTER	8	INF	3	INF	52,70	000000		
24	EXPORTERS	12	INF	18	INF	58,47	000000		
25	FARM	13	INF	68	INF	37,36	000000		
26	FARMER	28	INF	13	INF	178,47	000000		farmers
27	FERTILIZER	6	INF	3	INF	37,76	000000		fertilizers
28	FLOATER	6	INF	0		49,12	000000		floaters
29	FROST	8	INF	10	INF	41,10	000000		
30	GOOD	56	INF	784	INF	73,41	000000		
31	GREEN	24	INF	123	INF	69,79	000000		
32	GROW	14	INF	78	INF	38,77	000000		
33	HARVEST	23	INF	24	INF	123,98	000000		
34	HECTARE	6	INF	1	INF	43,41	000000		hectares
35	IMPORTANT	24	INF	261	INF	40,58	000000		
36	KILO	11	INF	0		90,05	000000		kilos
37	KIND	30	INF	217	INF	70,23	000000		
38	LEAF	6	INF	7	INF	31,41	000180		
39	LITERS	8	INF	1	INF	59,25	000000		
40	MACHINE	12	INF	42	INF	42,45	000000		machines

1 / 2

WordSmith Tools 4.0 -- 28/11/2007

41	MOKA	7	INF	0	57,31	000000	
42	MUCILAGE	8	INF	0	65,49	000000	
43	MULCHING	9	INF	1	INF	67,21	000000
44	NATURAL	14	INF	120	INF	28,92	000724
45	OPERATIVE	8	INF	13	INF	38,02	000000
46	ORGANIC	7	INF	17	INF	28,90	000731
47	PEABERRY	5	INF	0	40,93	000000	
48	PEOPLE	63	INF	1.578	INF	34,60	000011
49	PICKERS	6	INF	1	INF	43,41	000000
50	PICKING	18	INF	40	INF	76,87	000000
51	PLANT	19	INF	161	INF	39,61	000000
52	PLANTING	14	INF	11	INF	80,69	000000
53	PLANTS	16	INF	131	INF	34,24	000020
54	PRICE	43	INF	753	INF	42,78	000000
55	PROBLEM	23	INF	320	INF	30,34	000333
56	PROCESSING	25	INF	27	INF	133,58	000000
57	PRODUCTION	37	INF	281	INF	83,69	000000
58	PRUNE	5	INF	0	40,93	000000	
59	PRUNING	11	INF	2	INF	78,96	000000
60	PULPED	5	INF	3	INF	30,45	000314
61	PULPER	5	INF	0	40,93	000000	
62	PULPING	7	INF	0	57,31	000000	
63	QUALITY	32	INF	183	INF	87,26	000000
64	RAIN	16	INF	94	INF	42,91	000000
65	RIPEN	7	INF	1	INF	51,31	000000
66	ROASTERS	5	INF	0	40,93	000000	
67	ROBUSTA	28	INF	2	INF	214,62	000000
68	SEEDS	8	INF	14	INF	37,12	000000
69	SEPARATOR	4	INF	0	32,75	000076	
70	SOIL	18	INF	18	INF	98,07	000000
71	SPECIALTY	8	INF	11	INF	40,00	000000
72	STORAGE	9	INF	16	INF	41,55	000000
73	TON	6	INF	6	INF	32,69	000079 tons
74	TREE	27	INF	46	INF	126,41	000000 trees
75	VARIABILITY	6	INF	4	INF	35,79	000001
76	VARIETY	13	INF	69	INF	37,04	000000 varieties
77	WASHED	12	INF	12	INF	65,37	000000
78	WASHER	4	INF	0	32,75	000076	

ANEXO XI

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	. Freq.	R.C.	%	eyness	P
1	CAFÉ	1.783	1,69	0			.440,48	000000
2	SECAGEM	475	0,46	0			.574,80	000000
3	COLHEITA	338	0,32	0			.543,32	000000
4	QUALIDADE	388	0,37	115			.384,29	000000
5	CAFÉS	278	0,26	0			.091,69	000000
6	GRÃOS	260	0,25	3			.923,55	000000
7	PRODUTO	240	0,23	25			.641,26	000000
8	TERREIRO	197	0,19	0			.482,09	000000
9	PRODUÇÃO	233	0,22	83			.392,99	000000
10	FRUTOS	175	0,17	3			.286,24	000000
11	BEBIDA	165	0,16	5			.196,42	000000
12	UMIDADE	156	0,15	0			.173,58	000000
13	SEMENTES	149	0,14	2			.099,73	000000
14	PRODUTORES	144	0,14	3			.054,14	000000
15	ÁGUA	145	0,14	14			.996,70	000000
16	SECADOR	128	0,12	0			.962,90	000000
17	CEREJA	125	0,12	0			.940,33	000000
18	MERCADO	125	0,12	5			.898,18	000000
19	CHÃO	106	0,10	0			.797,38	000000
20	TEMPERATURA	113	0,11	10			.781,16	000000
21	MANUAL	101	0,10	0			.759,76	000000
22	PROCESSO	144	0,14	110			.740,90	000000
23	CASCA	102	0,10	3			.740,18	000000
24	AR	164	0,16	236			.703,38	000000
25	TEOR	106	0,10	22			.680,95	000000
26	SECADORES	88	0,08	0			.661,96	000000
27	PREPARO	92	0,09	4			.658,99	000000
28	GRÃO	84	0,08	0			.631,87	000000
29	LIMPEZA	83	0,08	0			.624,35	000000
30	VERDES	80	0,08	0			.601,78	000000
31	OPERAÇÃO	83	0,08	3			.598,46	000000
32	SISTEMA	131	0,12	157			.595,95	000000
33	IMPUREZAS	79	0,08	0			.594,25	000000
34	BENEFICIAMENTO	78	0,07	0			.586,73	000000
35	REGIÃO	101	0,10	54			.561,91	000000
36	POLPA	74	0,07	0			.556,64	000000
37	SECA	81	0,08	11			.542,46	000000
38	SACAS	71	0,07	0			.534,07	000000
39	ARMAZENAMENTO	71	0,07	2			.515,83	000000
40	TIPO	129	0,12	230			.512,36	000000

WordSmith Tools 4.0 -- 28/11/2007

41	PRODUTOR	73	0,07	5		512,21 000000
42	SEPARAÇÃO	72	0,07	8		489,96 000000
43	MUNDO	78	0,07	22		482,39 000000
44	MÁQUINA	63	0,06	0		473,89 000000
45	DERRIÇA	63	0,06	0		473,89 000000
46	MAIOR	177	0,17	722	0,02	473,63 000000
47	AMOSTRAS	77	0,07	24		469,58 000000
48	ARÁBICA	62	0,06	0		466,37 000000
49	FORNALHA	62	0,06	0		466,37 000000
50	TERREIROS	62	0,06	0		466,37 000000
51	PROCESSAMENTO	65	0,06	3		464,49 000000
52	REGIÕES	75	0,07	21		464,29 000000
53	CUSTO	80	0,08	38		455,26 000000
54	DESCASCADO	60	0,06	0		451,32 000000
55	LAVOURA	58	0,06	0		436,28 000000
56	MÁQUINAS	59	0,06	1		433,67 000000
57	PENEIRA	57	0,05	0		428,75 000000
58	CAFEICULTURA	57	0,05	0		428,75 000000
59	PROPRIEDADES	71	0,07	25		425,14 000000
60	CAFEIEIRO	56	0,05	0		421,23 000000
61	LAVADOR	56	0,05	0		421,23 000000
62	MUCILAGEM	56	0,05	0		421,23 000000
63	PRODUTIVIDADE	57	0,05	1		418,70 000000
64	AÇÚCARES	55	0,05	0		413,71 000000
65	AGRICULTURA	55	0,05	0		413,71 000000
66	CAFEICULTORES	55	0,05	0		413,71 000000
67	MEIO	93	0,09	122		411,19 000000
68	PENEIRAS	54	0,05	0		406,19 000000
69	CONSTRUÇÃO	53	0,05	0		398,66 000000
70	SAFRA	54	0,05	1		396,24 000000
71	GRANDE	101	0,10	195		388,97 000000
72	VERDE	51	0,05	0		383,62 000000
73	FIGURA	77	0,07	68		381,95 000000
74	TEMPO	99	0,09	215		363,42 000000
75	PREÇOS	48	0,05	0		361,05 000000
76	PLANTAS	51	0,05	3		360,59 000000
77	ESTADO	80	0,08	99		360,31 000000
78	ÉPOCA	57	0,05	14		358,91 000000
79	DIFERENTES	90	0,09	159		358,67 000000
80	NOVO	64	0,06	33		358,58 000000
81	CONDIÇÕES	83	0,08	118		357,38 000000

WordSmith Tools 4.0 -- 28/11/2007

82	PRODUTOS	59	0,06	20	355,35 000000
83	PLANTA	47	0,04	0	353,53 000000
84	TRATAMENTOS	60	0,06	26	347,14 000000
85	PREÇO	49	0,05	3	345,78 000000
86	DEFEITOS	51	0,05	6	345,54 000000
87	TORRADO	45	0,04	0	338,49 000000
88	SECO	45	0,04	0	338,49 000000
89	INTERNACIONAL	53	0,05	12	337,05 000000
90	ÁREA	73	0,07	86	333,81 000000
91	EMPRESA	48	0,05	4	333,04 000000
92	MECÂNICA	47	0,04	3	330,97 000000
93	FRAÇÃO	65	0,06	53	329,07 000000
94	SABOR	45	0,04	1	328,90 000000
95	LAVAGEM	44	0,04	1	321,42 000000
96	COLHIDO	44	0,04	1	321,42 000000
97	MATERIAL	58	0,06	32	320,64 000000
98	NACIONAL	52	0,05	16	317,69 000000
99	FERMENTAÇÃO	42	0,04	0	315,92 000000
100	LAVOURAS	42	0,04	0	315,92 000000
101	SOLO	42	0,04	0	315,92 000000
102	CATUAÍ	42	0,04	0	315,92 000000
103	FINAL	83	0,08	167	314,41 000000
104	TEORES	43	0,04	1	313,94 000000
105	LOTES	41	0,04	0	308,40 000000
106	FOLHAS	41	0,04	0	308,40 000000
107	TRABALHO	86	0,08	205	303,27 000000
108	QUANTIDADE	58	0,06	42	302,20 000000
109	CONTATO	46	0,04	8	301,08 000000
110	ÚMIDA	40	0,04	0	300,87 000000
111	COLHEDORA	40	0,04	0	300,87 000000
112	ORIGEM	55	0,05	32	300,76 000000
113	CARACTERÍSTICAS	68	0,06	96	293,47 000000
114	VIÇOSA	38	0,04	0	285,83 000000
115	OPERAÇÕES	38	0,04	0	285,83 000000
116	BÓIA	38	0,04	0	285,83 000000
117	MELHOR	74	0,07	144	284,08 000000
118	MATERIAIS	39	0,04	1	284,05 000000
119	OBRA	39	0,04	1	284,05 000000
120	VALORES	92	0,09	297	280,45 000000
121	PAÍS	48	0,05	20	279,61 000000
122	ESPECIAIS	49	0,05	23	279,45 000000

WordSmith Tools 4.0 -- 28/11/2007

123	CULTURA	40	0,04	3	279,25 000000
124	MADUROS	37	0,04	0	278,31 000000
125	ROBUSTA	37	0,04	0	278,31 000000
126	TECNOLOGIA	39	0,04	2	277,46 000000
127	MATURACÃO	38	0,04	1	276,58 000000
128	ACORDO	66	0,06	111	267,87 000000
129	MAIORES	67	0,06	118	267,31 000000
130	PRÉ	72	0,07	153	266,70 000000
131	QUENTE	38	0,04	3	264,51 000000
132	CULTIVO	35	0,03	0	263,26 000000
133	AROMA	35	0,03	0	263,26 000000
134	PROPRIEDADE	41	0,04	9	261,68 000000
135	FAZENDA	34	0,03	0	255,74 000000
136	CUSTOS	44	0,04	19	254,72 000000
137	MAS SA	71	0,07	163	254,51 000000
138	CLASSIFICAÇÃO	49	0,05	38	251,15 000000
139	SOL	36	0,03	3	249,77 000000
140	LEIRAS	33	0,03	0	248,22 000000
141	PERGAMINHO	33	0,03	0	248,22 000000
142	CULTIVARES	33	0,03	0	248,22 000000
143	FORMA	80	0,08	250	247,99 000000
144	PANO	34	0,03	1	246,71 000000
145	SISTEMAS	44	0,04	24	243,79 000000
146	ESPECIAL	49	0,05	44	241,99 000000
147	COFFEA	32	0,03	0	240,70 000000
148	AMBIENTE	42	0,04	20	238,89 000000
149	TERRA	35	0,03	4	237,66 000000
150	TOTAIS	35	0,03	5	233,36 000000
151	MATÉRIA	31	0,03	0	233,17 000000
152	ABANAÇÃO	31	0,03	0	233,17 000000
153	PODA	31	0,03	0	233,17 000000
154	DESPOLPADO	31	0,03	0	233,17 000000
155	BENEFICIADO	31	0,03	0	233,17 000000
156	EFICIÊNCIA	39	0,04	15	230,25 000000
157	TIPOS	46	0,04	41	227,62 000000
158	MÃO	35	0,03	7	225,75 000000
159	LOTE	30	0,03	0	225,65 000000
160	MOÍDO	30	0,03	0	225,65 000000
161	EMBALAGENS	30	0,03	0	225,65 000000
162	PLANTIO	30	0,03	0	225,65 000000
163	CENTRO	40	0,04	21	223,32 000000

WordSmith Tools 4.0 -- 28/11/2007

164	UNIDADE	40	0,04	21	223,32	000000
165	VARIEDADE	33	0,03	4	223,06	000000
166	METROS	35	0,03	8	222,32	000000
167	PRIMEIRA	51	0,05	71	221,13	000000
168	DESENVOLVIMENTO	60	0,06	130	220,45	000000
169	GRANDES	48	0,05	56	220,13	000000
170	BOA	50	0,05	68	218,47	000000
171	COLHIDOS	31	0,03	2	218,18	000000
172	VARRIÇÃO	29	0,03	0	218,13	000000
173	CAFEICULTOR	29	0,03	0	218,13	000000
174	MOAGEM	29	0,03	0	218,13	000000
175	MECÂNICOS	30	0,03	1	216,86	000000
176	RESULTADOS	97	0,09	522	0,01 216,73	000000
177	SUPERIOR	53	0,05	88	216,12	000000
178	MÉTODO	58	0,06	122	215,75	000000
179	OBTENÇÃO	41	0,04	30	213,09	000000
180	VIA	54	0,05	98	213,00	000000
181	EQUIPAMENTOS	38	0,04	20	212,05	000000
182	COMERCIALIZAÇÃO	30	0,03	2	210,78	000000
183	PRIMEIRO	45	0,04	49	210,65	000000
184	CAFEZAS	28	0,03	0	210,61	000000
185	TORRA	28	0,03	0	210,61	000000
186	VARIEDADES	28	0,03	0	210,61	000000
187	LENHA	28	0,03	0	210,61	000000
188	ACIDEZ	28	0,03	0	210,61	000000
189	CONSUMIDOR	28	0,03	0	210,61	000000
190	CAFEIRA	28	0,03	0	210,61	000000
191	DESCASCADOR	28	0,03	0	210,61	000000
192	POSSÍVEL	59	0,06	138	209,78	000000
193	NATURAL	36	0,03	15	209,70	000000
194	CUIDADOS	38	0,04	22	208,01	000000
195	VALOR	61	0,06	160	205,97	000000
196	MECÂNICO	33	0,03	9	205,00	000000
197	ELÉTRICA	31	0,03	5	204,40	000000
198	VENDA	27	0,03	0	203,09	000000
199	XÍCARA	27	0,03	0	203,09	000000
200	RENDIMENTO	32	0,03	8	201,04	000000
201	COCO	29	0,03	3	198,36	000000
202	CASO	65	0,06	213	196,58	000000
203	UTILIZADO	45	0,04	62	195,78	000000
204	PASSADAS	26	0,02	0	195,56	000000

WordSmith Tools 4.0 -- 28/11/2007

205	DESIDRATADA	26	0,02	0	195,56	000000
206	GERMINAÇÃO	26	0,02	0	195,56	000000
207	RETIRADA	35	0,03	19	194,11	000000
208	PASSAR	31	0,03	8	193,97	000000
209	INFERIOR	46	0,04	70	193,49	000000
210	UTILIZAÇÃO	52	0,05	111	192,25	000000
211	TAMANHO	43	0,04	55	191,65	000000
212	PAÍSES	39	0,04	36	191,19	000000
213	OBTIDOS	49	0,05	92	190,76	000000
214	PROCESSOS	35	0,03	21	190,16	000000
215	HORAS	52	0,05	114	190,11	000000
216	ABANADORA	25	0,02	0	188,04	000000
217	LAVADO	25	0,02	0	188,04	000000
218	OUTROS	81	0,08	413	187,90	000000
219	MUNICÍPIO	32	0,03	13	187,20	000000
220	MICROONDAS	26	0,02	1	187,06	000000
221	ESPÉCIES	29	0,03	6	186,34	000000
222	ALTURA	37	0,04	31	186,03	000000
223	PESQUISA	43	0,04	63	183,25	000000
224	REDUTORES	28	0,03	5	182,77	000000
225	CHUVAS	24	0,02	0	180,52	000000
226	CAÍDO	24	0,02	0	180,52	000000
227	DESPOLPADOR	24	0,02	0	180,52	000000
228	VENTILADOR	24	0,02	0	180,52	000000
229	PODAS	24	0,02	0	180,52	000000
230	VEZ	52	0,05	129	180,12	000000
231	MENORES	40	0,04	50	179,57	000000
232	DEPARTAMENTO	29	0,03	9	176,95	000000
233	SEÇÃO	26	0,02	3	176,42	000000
234	FEDERAL	29	0,03	10	174,20	000000
235	FIM	36	0,03	35	174,02	000000
236	CÂMARA	25	0,02	2	173,88	000000
237	CAMADA	33	0,03	23	173,46	000000
238	MUDAS	23	0,02	0	173,00	000000
239	CEREJAS	23	0,02	0	173,00	000000
240	MERCADOS	23	0,02	0	173,00	000000
241	SACA	23	0,02	0	173,00	000000
242	FAZENDAS	24	0,02	1	172,17	000000
243	FRAÇÕES	31	0,03	17	171,58	000000
244	OBJETIVO	53	0,05	153	170,94	000000
245	PORCENTAGEM	34	0,03	29	170,17	000000

WordSmith Tools 4.0 -- 28/11/2007

246	PROJETO	26	0,02	5	168,41 000000
247	CONVENCIONAL	32	0,03	23	167,01 000000
248	EMPRESAS	22	0,02	0	165,48 000000
249	TULHAS	22	0,02	0	165,48 000000
250	DESPOLPAMENTO	22	0,02	0	165,48 000000
251	REVOLVIMENTO	22	0,02	0	165,48 000000
252	EXPORTAÇÃO	22	0,02	0	165,48 000000
253	GEADAS	22	0,02	0	165,48 000000
254	ÉPOCAS	22	0,02	0	165,48 000000
255	CLIMA	23	0,02	1	164,73 000000
256	PEQUENOS	29	0,03	14	164,52 000000
257	EXPRESSO	27	0,03	9	163,02 000000
258	MENOR	59	0,06	228	162,92 000000
259	TABELA	50	0,05	142	162,56 000000
260	MELHORIA	28	0,03	12	162,30 000000
261	TEMPERATURAS	24	0,02	3	161,83 000000
262	EQUIPAMENTO	26	0,02	7	161,79 000000
263	ESPÍRITO	25	0,02	5	161,24 000000
264	ELIMINAÇÃO	28	0,03	13	160,00 000000
265	PLENUM	21	0,02	0	157,96 000000
266	CONDUTIVIDADE	21	0,02	0	157,96 000000
267	GOVERNO	21	0,02	0	157,96 000000
268	PROTÓTIPO	21	0,02	0	157,96 000000
269	MOCAS	21	0,02	0	157,96 000000
270	SECOS	21	0,02	0	157,96 000000
271	FUNCIONAMENTO	25	0,02	6	157,86 000000
272	ÁREAS	30	0,03	21	157,54 000000
273	QUALIDADES	22	0,02	1	157,30 000000
274	PONTO	45	0,04	110	156,91 000000
275	MUNDIAL	29	0,03	18	156,42 000000
276	ALTA	48	0,05	139	154,58 000000
277	MOLE	23	0,02	3	154,54 000000
278	SOLÚVEIS	23	0,02	3	154,54 000000
279	FISIOLOGICA	28	0,03	16	153,68 000000
280	SACOS	20	0,02	0	150,43 000000
281	CAFEIRAS	20	0,02	0	150,43 000000
282	PLÂNTULAS	20	0,02	0	150,43 000000
283	TECNOLOGIAS	20	0,02	0	150,43 000000
284	PARCERIA	21	0,02	1	149,87 000000
285	CONSUMO	46	0,04	133	148,25 000000
286	SECRETARIA	22	0,02	3	147,27 000000

WordSmith Tools 4.0 -- 28/11/2007

287	NOME	22	0,02	3	147,27	000000
288	TRADICIONAL	22	0,02	3	147,27	000000
289	ANÁLIS ES	30	0,03	29	145,24	000000
290	COMÉRCIO	21	0,02	2	144,46	000000
291	OPERACIONAL	21	0,02	2	144,46	000000
292	NORMAS	25	0,02	11	144,24	000000
293	US O	75	0,07	493	0,01 144,04	000000
294	BICA	19	0,02	0	142,91	000000
295	SEPARADOR	19	0,02	0	142,91	000000
296	ORGÂNICO	19	0,02	0	142,91	000000
297	ROTOR	19	0,02	0	142,91	000000
298	COMBUSTÍVEL	19	0,02	0	142,91	000000
299	RECOLHIMENTO	19	0,02	0	142,91	000000
300	SÓLIDOS	19	0,02	0	142,91	000000
301	DIFERENÇAS	42	0,04	109	142,51	000000
302	MENOS	56	0,05	248	142,45	000000
303	FRUTO	20	0,02	1	142,44	000000
304	PEDRAS	20	0,02	1	142,44	000000
305	PONTOS	26	0,02	15	142,42	000000
306	ESTRUTURA	27	0,03	19	141,61	000000
307	CERCA	41	0,04	106	139,36	000000
308	CRESCIMENTO	30	0,03	34	138,78	000000
309	RECURSOS	26	0,02	17	138,65	000000
310	COR	34	0,03	58	137,27	000000
311	RELAÇÃO	71	0,07	470	0,01 135,58	000000
312	ARABICA	18	0,02	0	135,39	000000
313	ARCA	18	0,02	0	135,39	000000
314	CLIMÁTICAS	18	0,02	0	135,39	000000
315	CONSÓRCIO	18	0,02	0	135,39	000000
316	TULHA	18	0,02	0	135,39	000000
317	HÍBRIDO	18	0,02	0	135,39	000000
318	MADEIRA	18	0,02	0	135,39	000000
319	ALTITUDE	19	0,02	1	135,02	000000
320	OFICIAL	19	0,02	1	135,02	000000
321	MUNICIPAL	19	0,02	1	135,02	000000
322	PERÍODO	58	0,06	299	133,52	000000
323	LOCAL	32	0,03	50	133,36	000000
324	PAS SA	27	0,03	25	132,25	000000
325	PADRÃO	37	0,04	86	131,91	000000
326	MANEJO	27	0,03	26	130,86	000000
327	COMERCIAL	19	0,02	2	129,80	000000

WordSmith Tools 4.0 -- 28/11/2007

328	RURAL	19	0,02	2	129,80 000000
329	SENTIDO	31	0,03	48	129,60 000000
330	MÉDIA	62	0,06	367	129,21 000000
331	CAMP O	26	0,02	23	128,90 000000
332	UNIVERSIDADE	27	0,03	28	128,18 000000
333	TORREFAÇÃO	17	0,02	0	127,87 000000
334	CHUVA	17	0,02	0	127,87 000000
335	OPERADOR	17	0,02	0	127,87 000000
336	AGRÍCOLAS	17	0,02	0	127,87 000000
337	COFFEE	17	0,02	0	127,87 000000
338	FERMENTAÇÕES	17	0,02	0	127,87 000000
339	SECRETÁRIO	17	0,02	0	127,87 000000
340	EXPORTAÇÕES	17	0,02	0	127,87 000000
341	UNIFORMIDADE	17	0,02	0	127,87 000000
342	PRODUTORAS	17	0,02	0	127,87 000000
343	BLEND	17	0,02	0	127,87 000000
344	ARMAZENAGEM	17	0,02	0	127,87 000000
345	ROÇA	17	0,02	0	127,87 000000
346	GRAVITACIONAL	17	0,02	0	127,87 000000
347	CARVÃO	17	0,02	0	127,87 000000
348	RADIAL	18	0,02	1	127,60 000000
349	INSTITUTO	28	0,03	34	126,84 000000
350	VENTILAÇÃO	23	0,02	13	126,52 000000
351	INDÚSTRIA	20	0,02	5	125,65 000000
352	TRADICIONAIS	19	0,02	3	125,53 000000
353	INDUSTRIAL	19	0,02	3	125,53 000000
354	VANTAGENS	26	0,02	27	123,38 000000
355	SÉCULO	20	0,02	6	122,63 000000
356	EKO	19	0,02	4	121,85 000000
357	SETOR	19	0,02	4	121,85 000000
358	MESA	19	0,02	4	121,85 000000
359	MARCA	23	0,02	16	120,95 000000
360	INÍCIO	41	0,04	142	120,38 000000
361	MATAS	16	0,02	0	120,35 000000
362	FUNGOS	16	0,02	0	120,35 000000
363	LAVADORES	16	0,02	0	120,35 000000
364	LAVRAS	16	0,02	0	120,35 000000
365	REFUGO	16	0,02	0	120,35 000000
366	DESMUCILADO	16	0,02	0	120,35 000000
367	CHAPA	16	0,02	0	120,35 000000
368	SOLAR	16	0,02	0	120,35 000000

WordSmith Tools 4.0 -- 28/11/2007

369	FITOTECNIA	16	0,02	0	120,35 000000
370	AGRÍCOLA	17	0,02	1	120,19 000000
371	TRANSPORTE	25	0,02	25	119,90 000000
372	MEIA	25	0,02	25	119,90 000000
373	PROGRAMA	39	0,04	126	118,82 000000
374	RIO	28	0,03	42	118,36 000000
375	PERDAS	20	0,02	8	117,31 000000
376	ENERGIA	21	0,02	11	117,29 000000
377	MANEIRA	35	0,03	96	115,71 000000
378	SEPARADOS	18	0,02	4	114,72 000000
379	INTRODUÇÃO	27	0,03	40	114,63 000000
380	FOGO	15	0,01	0	112,82 000000
381	GOURMET	15	0,01	0	112,82 000000
382	CHATAS	15	0,01	0	112,82 000000
383	MECANIZADA	15	0,01	0	112,82 000000
384	RECEPA	15	0,01	0	112,82 000000
385	SAFRAS	15	0,01	0	112,82 000000
386	ABASTECIMENTO	15	0,01	0	112,82 000000
387	AMONTADO	15	0,01	0	112,82 000000
388	CERRADO	15	0,01	0	112,82 000000
389	ROTATIVO	15	0,01	0	112,82 000000
390	PISO	16	0,02	1	112,79 000000
391	CONTAMINAÇÃO	20	0,02	10	112,71 000000
392	DETALHES	18	0,02	5	111,54 000000
393	ANÁLISE	47	0,04	236	110,16 000000
394	PRINCIPAL	32	0,03	81	109,83 000000
395	MELHORES	22	0,02	19	109,75 000000
396	UNIFORME	20	0,02	12	108,66 000000

ANEXO XII

WordSmith Tools 4.0 -- 28/11/2007

N	Key word	Freq.	%	Freq.	RC.	%	eyness	P
1	COFFEE	3.433	3,27	103			.534,32	000000
2	BEANS	547	0,52	9			.835,00	000000
3	COFFEES	369	0,35	2			.948,53	000000
4	ESPRESSO	268	0,26	1			.419,77	000000
5	CHERRIES	231	0,22	1			.222,18	000000
6	BEAN	215	0,20	6			.095,10	000000
7	ARABICA	186	0,18	2			.972,41	000000
8	FLAVOR	176	0,17	10			.864,32	000000
9	HARVESTING	165	0,16	3			.852,32	000000
10	QUALITY	261	0,25	183	0,01		.819,78	000000
11	DRYING	154	0,15	3			.793,92	000000
12	METHOD	171	0,16	19			.793,26	000000
13	ROASTING	141	0,13	0			.753,67	000000
14	CAFFEINE	138	0,13	0			.737,63	000000
15	WATER	276	0,26	291	0,02		.731,55	000000
16	ROASTED	143	0,14	5			.721,37	000000
17	ROBUSTA	132	0,13	2			.685,05	000000
18	ACIDITY	130	0,12	1			.683,26	000000
19	GREEN	187	0,18	123			.600,77	000000
20	PROCESSING	136	0,13	27			.584,46	000000
21	DRY	152	0,14	63			.561,42	000000
22	GROWN	141	0,13	48			.546,34	000000
23	RIPE	97	0,09	5			.479,26	000000
24	MACHINE	122	0,12	42			.471,49	000000
25	TASTE	120	0,11	48			.447,25	000000
26	MOISTURE	84	0,08	2			.430,24	000000
27	BREWING	90	0,09	8			.426,75	000000
28	BLEND	93	0,09	13			.420,03	000000
29	SPECIALTY	91	0,09	11			.418,18	000000
30	CHERRY	89	0,08	9			.416,84	000000
31	AROMA	86	0,08	6			.416,14	000000
32	DRIED	98	0,09	21			.415,89	000000
33	PARCHMENT	75	0,07	0			.400,85	000000
34	PROCESS	167	0,16	221	0,02		.393,95	000000
35	FARMERS	123	0,12	82			.393,24	000000
36	BLENDS	78	0,07	3			.391,65	000000
37	HARVEST	89	0,08	24			.362,25	000000
38	CUP	175	0,17	296	0,02		.356,29	000000
39	DRIP	69	0,07	3			.344,26	000000
40	TREE	91	0,09	46			.318,09	000000

WordSmith Tools 4.0 -- 28/11/2007

41	FLAVOUR	66	0,06	5		317,28 000000
42	BERRIES	63	0,06	3		312,73 000000
43	FRUIT	80	0,08	30		302,96 000000
44	MACHINES	82	0,08	35		300,50 000000
45	PULP	64	0,06	7		297,33 000000
46	PROCESSED	63	0,06	6		296,79 000000
47	COFFEA	55	0,05	0		293,94 000000
48	WET	82	0,08	38		293,86 000000
49	TEMPERATURE	70	0,07	17		290,60 000000
50	TREES	85	0,08	48		287,22 000000
51	METHODS	67	0,06	15		282,20 000000
52	HARVESTED	56	0,05	2		282,18 000000
53	CHARACTERISTICS	69	0,07	19		279,68 000000
54	BEVERAGE	58	0,06	5		275,76 000000
55	TASTING	54	0,05	2		271,63 000000
56	MECHANICAL	60	0,06	9		268,52 000000
57	BREW	53	0,05	2		266,36 000000
58	FERMENTATION	53	0,05	2		266,36 000000
59	SOIL	65	0,06	18		263,17 000000
60	BREWED	50	0,05	1		257,52 000000
61	ROASTERS	48	0,05	0		256,53 000000
62	CUPPING	48	0,05	0		256,53 000000
63	MEDIUM	71	0,07	36		247,95 000000
64	GROUND	113	0,11	173	0,01	244,94 000000
65	VARIETIES	52	0,05	6		240,19 000000
66	MUCILAGE	44	0,04	0		235,15 000000
67	FLAVORS	43	0,04	0		229,81 000000
68	FAIR	81	0,08	79		222,44 000000
69	PLANT	103	0,10	161	0,01	220,43 000000
70	EXTRACTION	40	0,04	1		204,51 000000
71	AROMATIC	35	0,03	0		187,05 000000
72	HARVESTER	34	0,03	0		181,70 000000
73	FILTER	44	0,04	11		181,68 000000
74	CROP	62	0,06	52		181,64 000000
75	WASHED	44	0,04	12		178,68 000000
76	BODIED	39	0,04	5		177,99 000000
77	SEEDS	45	0,04	14		177,84 000000
78	PRODUCTION	115	0,11	281	0,02	177,69 000000
79	DECAFFEINATED	32	0,03	0		171,02 000000
80	SOLUBLE	32	0,03	0		171,02 000000
81	DARK	65	0,06	70		170,45 000000

WordSmith Tools 4.0 -- 28/11/2007

82	CONTENT	61	0,06	60		166,87	000000
83	GROWING	105	0,10	249	0,02	166,41	000000
84	PULPING	31	0,03	0		165,67	000000
85	ROASTER	31	0,03	0		165,67	000000
86	POT	44	0,04	17		165,40	000000
87	PLANTATIONS	37	0,04	6		163,84	000000
88	WASHING	38	0,04	9		158,46	000000
89	PULPED	33	0,03	3		156,14	000000
90	MILK	52	0,05	43		153,22	000000
91	COLOR	54	0,05	49		153,07	000000
92	COFFEE'S	28	0,03	0		149,64	000000
93	UNRIPE	28	0,03	0		149,64	000000
94	ORIGIN	37	0,04	11		147,64	000000
95	FRESHNESS	32	0,03	4		146,47	000000
96	HULLING	27	0,03	0		144,29	000000
97	CAFÉ	33	0,03	7		140,26	000000
98	REGIONS	51	0,05	50		139,72	000000
99	CREMA	26	0,02	0		138,95	000000
100	SPECIES	47	0,04	39		138,29	000000
101	MARTINI	28	0,03	2		135,23	000000
102	GRINDER	25	0,02	0		133,60	000000
103	BLENDING	25	0,02	0		133,60	000000
104	DECAFFEINATION	25	0,02	0		133,60	000000
105	BOURBON	27	0,03	2		130,02	000000
106	PRODUCING	48	0,05	52		125,51	000000
107	JAVA	25	0,02	1		125,27	000000
108	VOLCANIC	26	0,02	2		124,83	000000
109	GOURMET	26	0,02	2		124,83	000000
110	PICKING	44	0,04	40		124,62	000000
111	GROWERS	34	0,03	16		121,31	000000
112	RAINFALL	27	0,03	4		121,02	000000
113	DRINK	45	0,04	47		119,73	000000
114	ORGANIC	34	0,03	17		119,21	000000
115	SWEET	44	0,04	45		118,22	000000
116	PEABERRY	22	0,02	0		117,57	000000
117	PLANTATION	27	0,03	5		117,27	000000
118	DIFFERENT	101	0,10	335	0,02	115,79	000000
119	ACID	31	0,03	13		114,12	000000
120	FRESH	56	0,05	95		113,75	000000
121	BODY	80	0,08	216	0,02	113,04	000000
122	GRINDING	26	0,02	5		112,27	000000

WordSmith Tools 4.0 -- 28/11/2007

123	BARISTA	21	0,02	0	112,23	000000	
124	MONSOONED	21	0,02	0	112,23	000000	
125	SORTING	27	0,03	7	110,72	000000	
126	CERTIFIED	24	0,02	3	109,85	000000	
127	MOUNTAIN	39	0,04	36	109,73	000000	
128	KNOWN	95	0,09	313	0,02	109,72	000000
129	ALTITUDE	26	0,02	6	108,92	000000	
130	GROUNDS	44	0,04	54	108,05	000000	
131	STORAGE	31	0,03	16	107,88	000000	
132	DRYERS	20	0,02	0	106,88	000000	
133	SOLVENT	20	0,02	0	106,88	000000	
134	GRADING	20	0,02	0	106,88	000000	
135	CARBON	26	0,02	7	105,85	000000	
136	STORED	29	0,03	13	104,87	000000	
137	MILD	32	0,03	20	104,59	000000	
138	ALTITUDES	21	0,02	1	104,23	000000	
139	TYPICAL	21	0,02	1	104,23	000000	
140	OILS	22	0,02	2	104,09	000000	
141	FINE	62	0,06	139	102,88	000000	
142	EVALUATION	29	0,03	14	102,72	000000	
143	CONE	24	0,02	5	102,31	000000	
144	OVERRIPE	19	0,02	0	101,54	000000	
145	CULTIVATION	19	0,02	0	101,54	000000	
146	EXTRACTED	25	0,02	7	100,99	000000	
147	CONSUMPTION	33	0,03	26	99,12	000000	
148	DIOXIDE	21	0,02	2	98,92	000000	
149	LAYER	27	0,03	12	97,87	000000	
150	TEMPERATURES	30	0,03	19	97,61	000000	
151	SEED	34	0,03	30	97,53	000000	
152	NATURAL	56	0,05	120	96,30	000000	
153	LIBERICA	18	0,02	0	96,19	000000	
154	GRINDERS	18	0,02	0	96,19	000000	
155	PICKED	46	0,04	71	95,73	000000	
156	SUMATRA	21	0,02	3	94,57	000000	
157	FRESHLY	19	0,02	1	93,74	000000	
158	GRAMS	19	0,02	1	93,74	000000	
159	PRODUCT	57	0,05	131	92,70	000000	
160	SMALL	112	0,11	487	0,03	91,12	000000
161	COUNTRIES	76	0,07	242	0,02	91,09	000000
162	FLOWERING	24	0,02	9	90,88	000000	
163	MOCHA	17	0,02	0	90,85	000000	

WordSmith Tools 4.0 -- 28/11/2007

164	LAMPUNG	17	0,02	0		90,85 000000]
165	PLUNGER	17	0,02	0		90,85 000000]
166	DENSITY	24	0,02	10		88,50 000000]
167	CLIMATE	30	0,03	25		88,12 000000]
168	VARIETY	42	0,04	69		87,10 000000]
169	TYPES	35	0,03	42		86,96 000000]
170	ORIGINS	23	0,02	9		86,18 000000]
171	EXTRACTING	20	0,02	4		85,83 000000]
172	COLOUR	29	0,03	24		85,42 000000]
173	SUGAR	33	0,03	39		82,63 000000]
174	AMOUNT	44	0,04	85		81,75 000000]
175	SMELL	23	0,02	11		81,68 000000]
176	ACIDS	19	0,02	4		80,86 000000]
177	EXTRACT	23	0,02	12		79,63 000000]
178	SHADE	21	0,02	8		79,21 000000]
179	CONTAINER	21	0,02	8		79,21 000000]
180	LEAVES	41	0,04	76		78,43 000000]
181	HAND	66	0,06	212	0,01	78,36 000000]
182	CULTIVATED	17	0,02	2		78,35 000000]
183	PALATE	16	0,02	1		78,04 000000]
184	TANZANIA	19	0,02	5		77,69 000000]
185	HIGH	155	0,15	898	0,06	77,08 000000]
186	IMPORTANT	73	0,07	261	0,02	76,76 000000]
187	LIQUID	24	0,02	16		76,71 000000]
188	BEST	119	0,11	600	0,04	76,71 000000]
189	BITTER	29	0,03	31		76,31 000000]
190	COOPERATIVE	18	0,02	4		75,90 000000]
191	SKIN	41	0,04	80		75,63 000000]
192	CANEPHORA	14	0,01	0		74,82 000000]
193	VARIETALS	14	0,01	0		74,82 000000]
194	FERMENTED	14	0,01	0		74,82 000000]
195	PERCOLATOR	14	0,01	0		74,82 000000]
196	SUGARS	14	0,01	0		74,82 000000]
197	SEEDLINGS	14	0,01	0		74,82 000000]
198	STEAM	25	0,02	20		74,64 000000]
199	BERRY	25	0,02	20		74,64 000000]
200	BREWERS	17	0,02	3		74,37 000000]
201	SORTED	17	0,02	3		74,37 000000]
202	SHAPED	27	0,03	27		73,30 000000]
203	FLAVORED	16	0,02	2		73,23 000000]
204	UNIQUE	31	0,03	41		73,12 000000]

WordSmith Tools 4.0 -- 28/11/2007

205	BAGS	32	0,03	45		72,92 000000]
206	DESIRED	18	0,02	5		72,82 000000]
207	QUANTITIES	18	0,02	5		72,82 000000]
208	SUN	60	0,06	189	0,01	72,74 000000]
209	PLANTING	21	0,02	11		72,62 000000]
210	CLEAN	40	0,04	80		72,46 000000]
211	FARMS	24	0,02	19		71,95 000000]
212	SIZE	53	0,05	151	0,01	71,15 000000]
213	CHOCOLATE	21	0,02	12		70,68 000000]
214	LARGE	82	0,08	343	0,02	70,47 000000]
215	BASKET	22	0,02	15		69,76 000000]
216	ROASTS	13	0,01	0		69,47 000000]
217	ARABICAS	13	0,01	0		69,47 000000]
218	MAREEBA	13	0,01	0		69,47 000000]
219	ROBUSTAS	13	0,01	0		69,47 000000]
220	HULLED	13	0,01	0		69,47 000000]
221	ETHYL	13	0,01	0		69,47 000000]
222	SAMPLES	24	0,02	21		69,08 000000]
223	PH	21	0,02	13		68,85 000000]
224	SURFACE	33	0,03	54		68,60 000000]
225	SOILS	15	0,01	2		68,13 000000]
226	POWDER	17	0,02	5		67,98 000000]
227	MARTINIS	14	0,01	1		67,61 000000]
228	HARVESTERS	14	0,01	1		67,61 000000]
229	COMPOUNDS	18	0,02	7		67,55 000000]
230	GROW	38	0,04	78		67,52 000000]
231	OUTER	23	0,02	20		66,38 000000]
232	NITROGEN	16	0,02	4		66,06 000000]
233	BETWEEN	140	0,13	833	0,06	65,86 000000]
234	HEATED	18	0,02	8		65,24 000000]
235	SELECTIVE	22	0,02	18		65,10 000000]
236	GRINDS	12	0,01	0		64,13 000000]
237	DRIER	12	0,01	0		64,13 000000]
238	DAVIDS	12	0,01	0		64,13 000000]
239	RHINOS	12	0,01	0		64,13 000000]
240	FLAVOURS	12	0,01	0		64,13 000000]
241	SEPARATED	19	0,02	11		63,68 000000]
242	TANKS	18	0,02	9		63,11 000000]
243	REMOVES	14	0,01	2		63,05 000000]
244	CATURRA	13	0,01	1		62,41 000000]
245	IMMATURE	13	0,01	1		62,41 000000]

WordSmith Tools 4.0 -- 28/11/2007

246	AROMAS	13	0,01	1		62,41 000000
247	PLANTS	46	0,04	131		61,78 000000
248	FRUITS	18	0,02	10		61,13 000000
249	STAGE	56	0,05	194	0,01	61,10 000000
250	SOLAR	22	0,02	21		60,99 000000
251	EXPORT	32	0,03	60		60,73 000000
252	RICA	17	0,02	8		60,65 000000
253	BRANCHES	23	0,02	25		60,04 000000
254	DRUM	20	0,02	16		59,71 000000
255	FINER	14	0,01	3		59,40 000000
256	ELEVATIONS	11	0,01	0		58,78 000000
257	CULTIVARS	11	0,01	0		58,78 000000
258	PICKER	11	0,01	0		58,78 000000
259	CUPPER	11	0,01	0		58,78 000000
260	MOIST	11	0,01	0		58,78 000000
261	UNROASTED	11	0,01	0		58,78 000000
262	DRIERS	11	0,01	0		58,78 000000
263	DARKER	19	0,02	14		58,56 000000
264	FINEST	19	0,02	14		58,56 000000
265	CONTAINERS	15	0,01	5		58,38 000000
266	FLOWERS	25	0,02	34		58,06 000000
267	COARSE	13	0,01	2		57,98 000000
268	HUMIDITY	13	0,01	2		57,98 000000
269	ACETATE	13	0,01	2		57,98 000000
270	IRRIGATION	13	0,01	2		57,98 000000
271	SWEETNESS	13	0,01	2		57,98 000000
272	INSTANT	29	0,03	51		57,51 000000
273	PRACTICES	29	0,03	51		57,51 000000
274	TYPE	32	0,03	65		57,31 000000
275	CHLORIDE	12	0,01	1		57,22 000000
276	PICKERS	12	0,01	1		57,22 000000
277	CAPPUCCINO	12	0,01	1		57,22 000000
278	MESH	12	0,01	1		57,22 000000
279	PROCESSES	20	0,02	18		56,89 000000
280	LOW	85	0,08	425	0,03	55,57 000000
281	DEGREES	20	0,02	19		55,56 000000
282	RESISTANCE	17	0,02	11		54,90 000000
283	SAMPLE	21	0,02	23		54,61 000000
284	RUST	13	0,01	3		54,46 000000
285	PLANTED	18	0,02	14		54,34 000000
286	TRADE	92	0,09	490	0,03	53,81 000000

WordSmith Tools 4.0 -- 28/11/2007

287	SPRAY	15	0,01	7		53,64 000000]
288	CHARACTERISTIC	15	0,01	7		53,64 000000]
289	BOILING	14	0,01	5		53,63 000000]
290	CINNAMON	10		0		53,44 000000]
291	AIRTIGHT	10		0		53,44 000000]
292	PORTAFILTER	10		0		53,44 000000]
293	COARSER	10		0		53,44 000000]
294	FRAGRANCE	12	0,01	2		52,93 000000]
295	FROST	16	0,02	10		52,29 000000]
296	SUSTAINABILITY	11	0,01	1		52,04 000000]
297	MOULD	11	0,01	1		52,04 000000]
298	REFERRED	20	0,02	22		51,90 000000]
299	PREMIUM	23	0,02	33		51,80 000000]
300	CUPS	17	0,02	13		51,66 000000]
301	ARTICLE	33	0,03	80		51,32 000000]
302	ENDANGERED	14	0,01	6		51,24 000000]
303	EQUIPMENT	37	0,04	103		50,79 000000]
304	PERFECT	31	0,03	72		49,97 000000]
305	CHEMICAL	29	0,03	62		49,95 000000]
306	CLEANING	15	0,01	9		49,70 000000]
307	DISTINCTIVE	15	0,01	9		49,70 000000]
308	STAGES	21	0,02	28		49,31 000000]
309	EXPORTED	16	0,02	12		48,98 000000]
310	TONS	17	0,02	15		48,76 000000]
311	REGION	45	0,04	158	0,01	48,34 000000]
312	PULPERS	9		0		48,10 000000]
313	TABLELANDS	9		0		48,10 000000]
314	TAMP	9		0		48,10 000000]
315	SEPARATOR	9		0		48,10 000000]
316	TEAS	9		0		48,10 000000]
317	GRADED	9		0		48,10 000000]
318	MARO	9		0		48,10 000000]
319	MUNDO	9		0		48,10 000000]
320	BUKIT	9		0		48,10 000000]
321	DRAINED	11	0,01	2		47,91 000000]
322	UNIFORM	18	0,02	19		47,65 000000]
323	OPTIMAL	10		1		46,88 000000]
324	PATIOS	10		1		46,88 000000]
325	ONION	12	0,01	4		46,71 000000]
326	PREPARATION	15	0,01	11		46,31 000000]
327	LIGHT	54	0,05	227	0,02	46,08 000000]

WordSmith Tools 4.0 -- 28/11/2007

328	MAINTAIN	23	0,02	40		45,95 000000
329	DRINKS	22	0,02	36		45,73 000000
330	PLACED	27	0,03	59		45,72 000000
331	RESULTING	20	0,02	28		45,69 000000
332	POURED	18	0,02	21		45,37 000000
333	DEVELOPMENT	49	0,05	194	0,01	45,35 000000
334	EVALUATING	14	0,01	9		45,32 000000
335	VACUUM	14	0,01	9		45,32 000000
336	AIR	60	0,06	275	0,02	45,12 000000
337	HEAT	30	0,03	76		44,90 000000
338	FREEZE	16	0,02	15		44,71 000000
339	EXPORTING	11	0,01	3		44,67 000000
340	STRIPPING	11	0,01	3		44,67 000000
341	LEAF	13	0,01	7		44,58 000000
342	TEA	25	0,02	52		43,98 000000
343	BALANCED	17	0,02	19		43,78 000000
344	FLORAL	14	0,01	10		43,65 000000
345	VARY	14	0,01	10		43,65 000000
346	RICH	36	0,03	114		43,39 000000
347	EARTHY	10		2		42,91 000000
348	POLISHED	10		2		42,91 000000
349	FROTH	10		2		42,91 000000
350	BOIL	10		2		42,91 000000
351	GRANULES	8		0		42,75 000000
352	CAFFEOL	8		0		42,75 000000
353	CATUAI	8		0		42,75 000000
354	NYBOT	8		0		42,75 000000
355	SENSORY	8		0		42,75 000000
356	CUPPED	8		0		42,75 000000
357	SOLIDS	8		0		42,75 000000
358	JUG	8		0		42,75 000000
359	BARISAN	8		0		42,75 000000
360	BARISTAS	8		0		42,75 000000
361	PYROLYSIS	8		0		42,75 000000
362	BURR	8		0		42,75 000000
363	SUPERCritical	8		0		42,75 000000
364	FACTORS	25	0,02	54		42,72 000000
365	PROPER	22	0,02	40		42,65 000000
366	ENJOY	21	0,02	36		42,36 000000
367	DEFECTIVE	12	0,01	6		42,07 000000
368	CITRUS	12	0,01	6		42,07 000000

WordSmith Tools 4.0 -- 28/11/2007

369	RESISTANT	15	0,01	14		42,00 000000]
370	ACIDY	9		1		41,74 000000]
371	OPTIMUM	9		1		41,74 000000]
372	FLOATERS	9		1		41,74 000000]
373	FLAVORFUL	9		1		41,74 000000]
374	RIPEN	9		1		41,74 000000]
375	BATCHES	9		1		41,74 000000]
376	LIQUOR	9		1		41,74 000000]
377	TIP	18	0,02	25		41,31 000000]
378	POPULAR	39	0,04	139		41,16 000000]
379	SEPARATION	13	0,01	9		40,99 000000]
380	LIGHTER	14	0,01	12		40,65 000000]
381	LAYERS	12	0,01	7		40,12 000000]
382	SUNLIGHT	12	0,01	7		40,12 000000]
383	PROCTER	10		3		39,82 000000]
384	PACKAGED	10		3		39,82 000000]
385	FERTILIZER	10		3		39,82 000000]
386	COVERED	24	0,02	54		39,70 000000]
387	USE	83	0,08	490	0,03	39,68 000000]
388	CONSISTENCY	11	0,01	5		39,63 000000]
389	STEEP	14	0,01	13		39,29 000000]
390	FULL	71	0,07	390	0,03	39,20 000000]
391	TECHNIQUES	22	0,02	45		39,19 000000]
392	SPREAD	30	0,03	88		39,13 000000]
393	SUPPLY	37	0,04	132		39,00 000000]
394	VESSEL	16	0,02	20		38,91 000000]
395	FIELDS	19	0,02	32		38,77 000000]
396	BAG	18	0,02	28		38,63 000000]
397	AVAILABLE	43	0,04	176	0,01	38,07 000000]
398	IDEAL	19	0,02	33		37,99 000000]
399	ENZYMES	9		2		37,95 000000]
400	ACIDIC	9		2		37,95 000000]
401	CONSERVATION	13	0,01	11		37,94 000000]
402	BOTTOM	30	0,03	91		37,82 000000]
403	SOAKED	11	0,01	6		37,57 000000]
404	IMPURITIES	7		0		37,41 000000]
405	BURLAP	7		0		37,41 000000]
406	TROPICS	7		0		37,41 000000]
407	RIPENESS	7		0		37,41 000000]
408	METHODOLOGY	7		0		37,41 000000]
409	MESOCARP	7		0		37,41 000000]

WordSmith Tools 4.0 -- 28/11/2007

410	PULPER	7	0	37,41 000000
411	WINY	7	0	37,41 000000
412	CATMOR	7	0	37,41 000000
413	FERTILE	7	0	37,41 000000
414	PERGAMINO	7	0	37,41 000000
415	FLAVOURED	7	0	37,41 000000
416	FERMENTING	7	0	37,41 000000
417	SHIPMENT	7	0	37,41 000000

ANEXO XIII

PB	VARIANTES
apanhador de café (s.m.)	<i>panhador de café</i>
apanhar café (v.)	<i>panhar café</i>
arábica (s.m.)	<i>café arábica</i>
árvore (s.f.)	<i>árvore do café; pé de café</i>
árvore do café (s.f.)	<i>árvore; pé de café</i>
barra (s.f.)	<i>Saia</i>
beneficiamento (s.m.)	<i>Benefício</i>
beneficiar café (v.)	<i>fazer beneficiamento do café</i>
bóia (s.m.)	<i>café bóia; grão bóia</i>
botão (s.m.)	<i>botãozinho</i>
botãozinho (s.m.)	<i>Botão</i>
Bourbon (s.m.)	<i>café Bourbon</i>
café arábica	<i>Arábica</i>
café beneficiado (s.m.)	<i>café verde; café limpo</i>
café bóia (s.m.)	<i>bóia; grão bóia</i>
café Bourbon (s.m.)	<i>Bourbon</i>
café catado a dedo	<i>colheita a dedo; colheita seletiva</i>
café catado no chão	<i>café de varrição; café de chão; café do chão</i>
café cereja (s.m.)	<i>Cereja</i>
café cereja descascado (s.m.)	<i>cereja descascado; descascado</i>
café colhido no pano (s.m.)	<i>café de pano; café derriçado no pano; café do pano</i>
café de chão (s.m.)	<i>café do chão; café de varrição; café catado no chão</i>
café de pano (s.m.)	<i>café colhido no pano; café do pano; café derriçado no pano</i>
café derriçado no pano (s.m.)	<i>café de pano; café do pano; café colhido no pano</i>
café desmucilado (s.m.)	<i>desmucilado</i>
café despulpado (s.m.)	<i>despulpado</i>
café de varrição (s.m.)	<i>café catado no chão; café de chão; café do chão</i>
café do chão (s.m.)	<i>café de chão; café catado no chão; café de varrição</i>
café do pano (s.m.)	<i>café colhido no pano; café de pano; café derriçado no pano</i>
café em coco (s.m.)	<i>café sem beneficiar</i>
café lavado (s.m.)	<i>Lavado</i>
café limpo (s.m.)	<i>café verde; café beneficiado</i>
café maduro (s.m.)	<i>Maduro</i>
café natural (s.m.)	<i>café seca natural; natural</i>
café robusta (s.m.)	<i>Robusta</i>
café seca natural (s.m.)	<i>café natural; natural</i>
café sem beneficiar (s.m.)	<i>café em coco</i>
café verde	<i>café beneficiado</i>
cereja (s.m.)	<i>café cereja</i>
cereja descascado	<i>café cereja descascado; descascado</i>
coiê café (v.)	<i>colher café</i>
colheita (s.f.)	<i>panha de café; colheita do café; colheta</i>
colheita a dedo (s.f.)	<i>café catado a dedo; colheita seletiva</i>
colheita do café (s.f.)	<i>colheita; panha do café</i>
colheita manual (s.f.)	<i>colher com a mão</i>
colheita mecânica (s.f.)	<i>colheita mecanizada; colher com a máquina</i>
colheita mecanizada (s.f.)	<i>colheita mecânica; colher com a máquina</i>

Continua

Continuação

colheita seletiva (s.f.)	café catado a dedo; colheita a dedo
colher café (v.)	coiê café
colher com a mão (v.)	colheita manual
colher com a máquina (v.)	colheita mecânica; colheita mecanizada
colheta (s.f.)	panha de café, colheita de café, colheita
descascado (s.m.)	café cereja descascado; cereja descascado
desmucilado (s.m.)	café desmucilado
despolpado (s.m.)	café despolpado
fazenda de café	propriedade
fazer beneficiamento do café (v.)	beneficiar o café
fruto (s.m.)	fruto de café; fruto do café
fruto de café (s.m.)	fruto; fruto do café
fruto do café (s.m.)	fruto; fruto de café
grão bóia (s.m.)	bóia; café bóia
grão moca (s.m.)	Moca
grão verde (s.m.)	Verde
lavado (s.m.)	café lavado
lavora (s.f.)	lavoura de café
lavoura de café (s.f.)	Lavora
maduro (s.m.)	café maduro
moca (s.m.)	grão moca
natural (s.m.)	café natural; café seca natural
panha de café (s.f.)	colheita de café
panhar café (v.)	apanhar café
panhador de café (s.m.)	apanhador de café
pé de café (s.m.)	árvore; árvore de café
produção (s.f.)	Produção do café
propriedade (s.f.)	fazenda de café
robusta (s.m.)	café robusta
saca de café (s.f.)	saco de café
saco de café (s.m.)	saca de café
saia (s.f.)	Barra
seca (s.f.)	Secagem
secador (s.m.)	secador de café
secador de café (s.m.)	Secador
secagem (s.f.)	Seca
tanque com água (s.m.)	tanque de fermentação
tanque de fermentação	tanque com água
terreiro (s.m.)	terrero, terreirão
terrero (s.m.)	terreiro, terreirão
terreirão (s.m.)	terreiro, terrero
varrição (s.f.)	café de varrição
verde (s.m.)	grão verde

ANEXO XIV

TERMOS	COLOCADOS
adubação	
alqueire	
arábica	... arábica
árvore	... do café
barra	
beneficiamento	fazer ... do café máquina do ...
benefício	
bóia	café ... grão ...
botão	
botãozinho	
Bourbon	café ...
café	abandar ... apanhador de ... apanhar arábica armazém de ... árvore do ... pé de ... bebida do ... beneficiar... ... bóia ... Bourbon broto do beneficiado ... verde ... limpo ... catado a dedo ... de varrição ... de chão ... do chão ... cereja ... cereja descascado ... colhido no pano ... do pano ... de pano ... derriçado no pano ... comum ... de melhor qualidade ... de qualidade ... desmucilado ... despulpado ... de terreiro ... em coco ... sem beneficiar ... especial ... esqueletado ... lavado

	... limpo ... maduro ... natural ... normal ... orgânico ... robusta ... seca natural ... sujo ... verde casca do ... catar ... coiê ... colher ... dar ... derriçar ... despolpar ... despolpador de ... fazenda de ... fazer beneficiamento do... grão de... guardar... lavador de... lavar ... lavoura de... muda de... palha do... panha de ... panhar ... panhador de... plantar... produção do ... produtor de ... qualidade do... saca de... saco de... secar... tipo de ...
cafeicultura	
cereja	café ... café ... descascado ... descascado descascador de ...
clima	
coiê café	
colheita	... a dedo ... seletiva ... manual ... manual mecanizada ... mecânica ... mecanizada época da ... fazer a ...

	mecanização da... momento da ...
colher café	... com a mão ... com a máquina ... no pano
colheta	
derriça	... no pano
derriçar	... café ... no pano
descascado	café cereja ... cereja ...
desmucilado	café ...
despolpado	café natural ... por fermentação
despolpar	... café
esqueletamento	
florada	
folha	tirar ...
fruto	... de café ... do café
geada	
grão	... bóia ... de café ... maduro ... moca ... preto ... grão seco ... verde
lavado	café ...
lavador	... de café ... mecânico
lavagem	
lavora	
lavoura	... de café
mão-de-obra	
máquina	... de beneficiamento
mecanização	... da colheita
moca	grão ...
mucilagem	
natural	café ... café seca ...l
palha	... do café
pedra	
peneira	saca de ... (...)
pergaminho	
plantação	

polpa	
processamento	
processo mecânico	
produção	... do café
propriedade	
qualidade	... do café
rama	
robusta	
saca	... de café ... de peneira (...)
saco	... de café
safra	
saia	
seca	meia ... café de ... natural
secador	... à gás ... à lenha ... de café ... vertical
secagem	... no terreiro
terreiro	chão do ... secagem no ... secar no ...
talhão	
tanque	... com água ... de fermentação
terreiro	... coberto ... tradicional
terreirão	
terrero	
trator	carreta do ...
tulha	guardar na ...
turmero	
varrição	café de...
verde	café ... grão ...

BIBLIOGRAFIA CONSULTADA

Abdulaziz, M.H. East Africa. In: Cheshire, J. **English around the world**. Cambridge: Cambridge University Press, 1991.

Adobe. Cool Edit 2000 [software]. Disponível em <<http://www.adobe.com/special/products/audition/syntrillium.html>>. Acesso em 20 set. 2007.

Almeida, G.M.B.; Aluísio, S.M.; Oliveira Júnior, O.N.; Oliveira, L.H.M.; Di Felippo, A.; Genoves Júnior, L.C.; Soares, L.G.; Mattos, D.; Coleti, J.S. Recolha e sistematização de corpora para elaboração do primeiro dicionário-piloto em nanociência e nanotecnologia em língua portuguesa. Disponível em <<http://www.realiter.net/spip.php?article559>>. Acesso em 20 dez. 2006.

Almeida, G.M.B.; Souza, D.S.L.; Pino, D.H.P. A definição nos dicionários especializados: proposta metodológica. **Rlterm**, nº 3, 2007.

Alves, I.M.; Desmet, I. Neologia terminológica e variação geolectal (português europeu, português brasileiro e francês hexagonal): um projecto de cooperação. Disponível em <http://http://www.realiter.net/barcelona/alves_desmet.htm>. Acesso em 22 dez. 2005.

AMSC - Alta Mogiana specialty coffees. Disponível em <<http://www.amscom.br>>. Acesso em 8 set. 2007.

ANC - American National Corpus. Disponível em <<http://americannationalcorpus.org>>. Acesso em 20 mai. 2006.

Aubert, F.H. **Introdução à metodologia da pesquisa terminológica bilíngüe.** CITRAT FFLCH/USP. 2ª ed. São Paulo: Humanitas, 2001.

Banks, M.; McFadden, C.; Atkinson, C. **Manual enciclopédico do café.** Tradução Dolores Ferreira. Lisboa: Editorial Estampa, 2000.

Beeching, K. Synchronic and diachronic variation: the how and why of sociolinguistic corpora. In: Wilson, A.; Archer, D.; Rayson, P. **Corpus linguistics around the world.** Amsterdam- New York: Rodopi, 2006.

BNC - British National Corpus. Disponível em <<http://sara.natcorp.ox.ac.uk/lookup.html>>. Acesso em 20 mai. 2006.

Brown Corpus. Disponível em <http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html>. Acesso 16 mai. 2006.

BSCA - Brazilian Specialty Coffee Association. Disponível em <<http://www.bsca.com.br>>. Acesso em 8 set. 2007.

Cabré, M.T. A terminologia hoje: concepções, tendências e aplicações. In: Krieger, M.G.; Araújo, L. **Cadernos de tradução**, nº 17. Porto Alegre: Instituto de Letras, UFRGS, 2004.

_____. **La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.** Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona, 1998.

CACCEER - Conselho das Associações do Café do Cerrado. Disponível em <<http://www.cafedocerrado.org>>. Acesso em 08 set. 2007.

Caldas, S.T.; D'ALESSIO, V. **Café - Um grão de história.** São Paulo: Dialetto, 2006.

Chiaro, D.; Nocella, G. Interpreters' perception of linguistic and non-linguistic factors affecting quality: a survey through the World Wide Web. **Meta**, **XLIX**, 2, 2004.

CLAWS4. Etiketador online. UCREL - University Centre for Computer Corpus Research on Language. Lancaster University. Disponível em <<http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>>. Acesso em 20 mai. 2006.

Coffee Sustainable Farming Handbook. 2nd Ed. Kenya: Coffee Management Services Ltd. January 2007.

COLLINS/COBUILD. Disponível em <<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>>. Acesso em 20 mai. 2006.

COLT. The bergen corpus of London teenage Language. Disponível em <<http://torvald.aksis.uib.no/colt/>>. Acesso em 20 mai. 2006.

Corpus of Spoken, Professional - American English. Disponível em <<http://www.athel.com/cspa.html>>. Acesso em 20 mai. 2006.

Creative Commons. Audacity [software]. Disponível em <<http://audacity.sourceforge.net/copyright>>. Acesso em 20 set. 2007.

CSLU – Spoltech Brazilian Portuguese. Disponível em <<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S16>>. Acesso em 20 mai. 2006.

Dicionário Aurélio - Século XXI. Versão 3.0. Editora Nova Fronteira. 1 CD-ROM.

Du Bois, J.W.; Chafe, W.L.; Meyer, C.; Thompson, S. Santa Barbara corpus of spoken american English. Disponível em <<http://www ldc.upenn.edu>>. Acesso em 10 dez. 2002.

Escola de parapente ventomania. Disponível em <http://www.ventomania.com.br/data/mapa_andradas.htm>. Acesso em 21 mai. 2006.

Express scribe. NCH swift sound. Disponível em <<http://www.nch.com.au/scribe/>>. Acesso em 20 set. 2007.

Faulstich, E. **Entre a sincronia e a diacronia: variação terminológica no código e na língua.** Conferência apresentada em Havana, Cuba: 1998.

_____. Variações terminológicas: princípios lingüísticos de análise e método de _____ recolha. Disponível em <http://dt il.unilat.org/realiter_spip/spip.php?article265>. Acesso 2 jul. 2007.

FDA - Food and Agriculture Organization of United Nations. Glossary of coffee related terms. Disponível em <<http://www.coffee-ota.org/glossary.asp>>. Acesso em 15 nov. 2007.

Fonseca, L.C. **Compilação de corpus eletrônico de Direito Contratual – Instrumentos contratuais. Projeto CORTEC.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 1 dez. 2006.

Fromm, G. **Compilação de corpus eletrônico de informática.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 1 dez. 2006.

Ginezi, L.L. **Compilação de corpus eletrônico de hipertensão. Projeto CORTEC.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 2 dez. 2006.

Greenbaum, S.; Svartvik, J. **LLC - London-lund corpus.** Disponível em <<http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.htm>>. Acesso em 10 dez. 2002.

ICAME. Disponível em <<http://nora.hd.uib.no/icame/newcd.htm>>. Acesso em 20 mai. 2006.

ICE. International corpus of english. Disponível em <<http://www.ucl.ac.uk/english-usage/ice/avail.htm>>. Acesso em 20 mai. 2006.

ISO 1087. **Terminology work. Vocabulary. Part 1: Theory and application.** International Organization for Standardization, 2000.

ISO 5127. **Information and Documentation - Vocabulary. International.** Organization for Standardization, 2001.

Lácio-Web. Disponível em <<http://www.nilc.icmc.usp.br/lacioweb/index.htm>>. Acesso 10 mai. 2006.

Lamparelli, A. **A naturalidade na tradução: quem garante?** Dissertação (Mestrado em Estudos Lingüísticos e Literário em Inglês). Universidade de São Paulo: São Paulo, 2007.

Lederer, M. **The interpretive model. Translated by ninon larche.** Manchester: St. Jerome Publishing, 2003.

London-Lund corpus of spoken english. Disponível em <<http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM>>. Acesso 15 mai. 2006.

Macwhinney, B. **Childes (Child Language Data Exchange System).** Disponível em <<http://childes.psy.cmu.edu/>>. 2 dez. 2003.

Martins, J. **Compilação de corpus eletrônico de meio-ambiente. Projeto CORTEC.** Disponível em <<http://www.ffflch.usp.br/dlm/comet>>. Acesso em 1 dez. 2006.

Melissi software. Disponível em <<http://www.melissi.co.uk/index.htm>>. Acesso em 5 dez. 2006.

MICASE. Disponível em <<http://micase.umdl.umich.edu/cgi/m/micase/micase-idx?type=revise>>. Acesso 15 mai. 2006.

Moraes, H. F. **Os advérbios na tradução, na terminologia e no ensino - um estudo contrastivo baseado em corpus.** Projeto de doutorado. Disponível em <<http://www.ffflch.usp.br/dlm/comet>>. Acesso em 3 nov. 2007.

PARACONC. Disponível em <<http://athel.com/para.html>>. Acesso em 13 mai. 2006.

Petite, C. Evidence of repair mechanisms in simultaneous interpreting. **Interpreting 7:1**. Amsterdam: John Benjamins Publishing Company, 2005. p. 27-49.

Pretti, D. **Sociolingüística - Os níveis da fala**. São Paulo: Edusp, 1994.

Projeto COMET. Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 1 set. 2005.

Projeto NURC-SP. Documentação Eletrônica. 2002. 1 CD-ROM.

Revista Anuário Brasileiro do Café. Santa Cruz do Sul: Editora Gazeta. 2006.

Rey, A. Origins and development of terminology. In: Sager, J.C. (Ed.). **Essays on terminology**. v. 9. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

Sampaio, C.A.L. **Construção e exploração de um corpus eletrônico de linguagem bilíngüe infantil**. 2004. Dissertação (Mestrado em Lingüística Aplicada) - Universidade Estadual do Ceará, Fortaleza.

_____. **O uso do WordSmith tools na investigação da linguagem falada de duas crianças bilíngües**. Apresentação: Curso de Mestrado em Lingüística Aplicada. Universidade Estadual do Ceará, Fortaleza, 2002.

Santa Barbara corpus of spoken american language - Part IV. Disponível em <<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S25>>. Acesso em 20 mai. 2006.

SCAA - Specialty Coffee Association of America. Disponível em <<http://www.scaa.org>>. Acesso em 8 set. 2007.

Silva, L.A. Monitoramento na conversação: interferência do ouvinte. In: Urbano, U.; Dias, F.A.R.; Leite, M.Q.; Silva, L.A.; Galembeck, P.T. (Orgs.). **Dino Preti e seus temas: oralidade, literatura, mídia e ensino.** São Paulo: Cortez, 2001.

Sound Forge 9. Sony. Disponível em <<http://www.sonycreativesoftware.com/products/soundforgefamily.asp>>. Acesso em 20 set. 2007.

Souza, E.M.C.; Oppenheimer, M.J.P. **Vocabulário para ciências agrárias inglês/português.** Série Mil & Um Termos. São Paulo: SBS Editora, 2004.

Tagnin, S.E.O. **Corpora: o que são e para quê serve. Corpora para Terminologia.** Disponível em <<http://www.fflch.usp.br/dlm/comet/artigos/Lexicografia.pdf>>. 20 dez. 2004.

Tagnin, S.E.O.; Teixeira, E.D. **British vs. American English, Brazilian vs. European Portuguese: how close or how far apart? - A corpus-driven study.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 5 jan. 2006.

Teixeira, E. D. **Como usar o WordSmith tools.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 5 dez. 2003.

_____. **Compilação de corpus eletrônico de Culinária.** Disponível em <<http://www.fflch.usp.br/dlm/comet>>. Acesso em 1 dez. 2006.

Temmerman, R. **Towards a new way of terminology description. The socio-cognitive approach.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2000.

Teubert, W. Language as an economic factor: the importance of terminology. In: Barnbrook, G.; Pernilla, D.; Mahlberg, M. **Meaningful texts.** London, New York: Continuum, 2005.

The wellington corpus of New Zealand english. Disponível em <<http://www.vuw.ac.nz/lals/research/corpora/wcs.aspx>> Acesso em 20 mai. 2006.

VARPORT. Disponível em <<http://www.letras.ufrj.br/varport>>. Acesso em 20 mai. 2006.

Ynoguti, C.A.; Barbosa, P.A.; Violaro, F. A Large speech database for Brazilian Portuguese spoken language research. In: Mamede, N.J; Baptista, J.; Trancoso, I.; Volpe Nunes, M.G. (Eds.). **Computational processing of the Portuguese Language.** Berlim: Springer, 2003.

Zapparoli, Z. Um pouco da história da análise informatizada do léxico no Brasil. In: Nunes, J.H e Petter, M. (Orgs.). **História do saber lexical e constituição de um léxico brasileiro.** São Paulo: Pontes/Humanitas – FFLCH/USP, 2002.