PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
PUC-SP

Luciene Novais Mazza

Os *lexical bundles* na busca por semelhanças em um documento do setor farmacêutico.

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para a obtenção do título de MESTRE em Lingüística Aplicada e Estudos da Linguagem, sob orientação da Profa. Doutora Rosinda de Castro Guerra Ramos.

SÃO PAULO
2009

Ficha Catalográfica

MAZZA, Luciene Novais. Os *lexical bundles* na busca por semelhanças em um documento do setor farmacêutico: s.n, 2009.

Dissertação (Mestrado) – Pontifícia Universidade Católica de São Paulo

Área de concentração: Lingüística Aplicada e Estudos da Linguagem

Orientadora: Professora Doutora Rosinda de Castro Guerra Ramos

Palavras-chave: *Lexical bundles*; Lingüística de *Corpus*; *Site Master File*.

Banca Examinadora

_____

_____

_____

Autorizo, exclusivamente para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação por processos fotocopiadores ou eletrônicos.

Assinatura :_____Local e data: São Paulo, ____/____ / 2009.

Para José Luiz Mazza.

Agradecimentos

Ao CNPq pelo apoio financeiro.

À Profa. Dra. Rosinda de Castro Guerra Ramos pela amizade e dedicação.

Ao Prof. Dr. Tony Berber Sardinha pela atenção dispensada a cada etapa deste trabalho.

À Profa. Dra. Tânia Shepherd pelo interesse e suporte teórico.

Às Professoras Dra. Stella Tagnin e Dra. Elisa Teixeira por terem aceitado o nosso convite a participar da banca examinadora desta dissertação.

Às Professoras Dra. Beth Brait, Dra. Sumiko Ikeda e a todos os professores do LAEL por terem contribuído na ampliação dos meus conhecimentos nos estudos da Linguagem.

Às secretárias do LAEL, Maria Lúcia e Márcia Martins, pela paciência e colaboração.

A minha mãe Marta e ao meu pai Antonio (*in memorian*) pelo amor incondicional.

A minha irmã Simone e minha sobrinha Laura pelo carinho e incentivo.

Aos meus colegas pesquisadores do LAEL pela cooperação.

Luciene Novais Mazza. Os *lexical bundles* na busca por semelhanças em um documento do setor farmacêutico.

RESUMO

O objetivo deste trabalho foi examinar o documento *Site Master File* do setor farmacêutico a partir da investigação de uma combinação de palavras denominada *lexical bundles* (Biber et al. 1999) com o propósito de verificar o grau de conformidade com elementos lingüísticos que um documento com a mesma organização estrutural, escrita por diferentes autores em diferentes partes do mundo pode atingir. A presente pesquisa teve como principal suporte teórico e metodológico a Lingüística de *Corpus* (Stubbs, 1996; Scott e Tribble, 2006; Berber Sardinha e Barbara, 2008; entre outros), uma abordagem que permite investigar como a língua ocorre naturalmente no discurso por meio de ferramentas computacionais. Para esta investigação foram compilados quinze exemplares do documento *Site Master File* pertencente a um mesmo grupo farmacêutico multinacional com sede na Europa e com unidades de negócios espalhadas em mais de 100 países. O documento *Site Master File* é um conjunto de textos produzidos pelas indústrias farmacêuticas para atender as exigências de garantia e controle da qualidade dos medicamentos, a fim de se obter certificação internacional junto aos órgãos de vigilância sanitária. Ademais, todos os documentos devem ser oficialmente produzidos em língua inglesa. Para a análise dos dados foram utilizadas as linguagens de programação *Perl* e *Cygwin*, como também foi desenvolvido um aplicativo para gerar a extração dos *lexical bundles* de três palavras. Os resultados da análise dos dados indicaram, que embora o documento *Site Master File* apresente semelhanças em sua organização, não há uma regularidade de *lexical bundles* recorrentes entre as amostras dos quinze exemplares. Assim, dessa ausência de *bundles* semelhantes, foi possível observar traços característicos do tipo de negócio que cada unidade da empresa está envolvida, dos processos e produtos fabricados e, ainda, a relação da proximidade geográfica com as escolhas lingüísticas feitas pelos autores. Portanto, este estudo além de contribuir para o conhecimento das variações de uso da língua inglesa por autores de diferentes localidades na elaboração do documento *Site Master File*, também implica em futuras pesquisas no ensino de línguas para fins específicos baseado em *corpora* e nos estudos sobre terminologia.

Palavras-chave: *Lexical bundles*; Lingüística de *Corpus*; *Site Master File*.

Luciene Novais Mazza. Lexical bundles searching similarities in a document of pharmaceutical sector.

ABSTRACT

The present study explored a specific document of the pharmaceutical segment called Site Master File through the investigation of words combinations defined as lexical bundles (Biber et al.,1999). The aim of the study was to draw out the bundles so that to verify the degree of conformity of the linguistic features the use of lexical bundles may achieve, as being part of a document organized in a similar way, produced by different authors at different locations around the world. The theoretical-methodological approach was developed on the principles of Corpus Linguistics (Stubbs, 1996; Scott and Tribble, 2006; Berber Sardinha and Barbara, 2008; amongst others), an approach that makes use of a vast variety of authentic texts of language in use supported by computational tools. We compiled for this study fifteen samples of the Site Master File document stored in machine-readable form that belong to the same multinational pharmaceutical company based in Europe, which has more than a hundred of plants situated across the world. The Site Master File is a document prepared by pharmaceutical manufacturers that contains specific information about the quality assurance, the production and quality control of pharmaceutical manufacturing operations carried out at a named site/plant in order to be submitted to a regulatory authority. In addition, all documents must be officially certified in English. The analysis of the corpus data was performed to extract three-word bundles by using scripting languages such as Perl and Cygwin. Besides, a computer application was also designed to provide the cross-reference of data. The results of data analysis showed that although the samples of Site Master File bring a large range of similarity in its organization, we have not found regularity on the use of recurrent lexical bundles across the Site Master File documents. Thus, considering the absences of common lexical bundles across documents, we observed that, in each operating area of the pharmaceutical business unit there are some typical characteristics in relation to the type of product manufactured in the site, the processes engaged in the unit pharmaceutical operations as well as the geographic nearness relationships to the linguistic choices made by the different authors. Therefore, this study offers a contribution to the knowledge of variation in English use in preparing the Site Master File by authors allocated in a specific site. Moreover, the present study involves further research into the field of English for Specific Purposes based on corpora and into the studies of terminology.

Keywords: Lexical bundles; Corpus Linguistics; Site Master File.

SUMÁRIO

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

INTRODUÇÃO

Esta pesquisa tem por objetivo examinar um tipo de documento do setor farmacêutico, o *Site Master File* (definido adiante), a partir da investigação de *lexical bundles* recorrentes para um levantamento dos aspectos semelhantes entre quinze localidades do mundo nas quais o documento circula.

No contexto dos estudos de gêneros e tipo textuais, a dinâmica decorrente das mudanças que ocorrem no cotidiano social e da evolução tecnológica, cada vez mais demanda dos setores empresariais diversas formas de comunicação para efetivar as negociações internacionais. Para Bazerman (2006), há um grande número de gêneros altamente regulados e padronizados, os quais estabelecem uma hierarquização de poder com a escrita, e exigem uma relação com os atores na produção de sentidos. Por sua vez, essas relações, segundo o autor, estão inseridas de algum modo em um sistema mediado por documentos.

Com o início da globalização nos anos de 1970, tivemos a ampliação dos mercados e a fusão de capitais entre empresas locais e internacionais para a formação de uma cadeia produtiva mundial, promovendo, como resultado, uma transformação nos aspectos sócio-econômicos e culturais, exigindo, assim, um sistema de comunicação rápido e padronizado, ou seja, uma linguagem “universal” que ultrapassasse fronteiras.

Em um breve retrospecto histórico, podemos afirmar que devido ao grande poderio econômico da Inglaterra a partir do século XIX, impulsionado pela Revolução Industrial e pela expansão do colonialismo britânico, que alcançou uma grande abrangência geográfica, agregado ao evento do domínio norte-americano depois da Segunda Guerra Mundial, a língua inglesa atingiu uma vasta disseminação, que, por conseguinte, tem sido adotada como lingua franca (*ELF English as a Lingua Franca*) por quase todas as empresas multinacionais espalhadas pelo mundo (Crystal, 1997; Graddol, 2006). Desse modo, a língua inglesa passou a ser uma ferramenta de comunicação entre as instituições, com o fim de viabilizar transações comerciais, especialmente referentes à circulação e troca de documentos, e às informações relacionadas aos interesses de negócios corporativos e governamentais. Corroborando essas asserções, Graddol (2006:92) comenta que:

> [...] Na prática, em muitas das grandes empresas, e mesmo em parte das instituições governamentais européias, a língua inglesa tornou-se uma característica comum no local de trabalho [...] [1].

Entretanto, esse processo acelerado de difusão dos mercados globais transcendeu às condições e situações locais, no sentido de que as empresas multinacionais, quando instaladas em determinados países, impõem práticas de organização que, muitas vezes, podem impactar nas condições locais, principalmente as situações que envolvem pessoas, ou seja, a mão de obra especializada contratada para escrever, pensar e agir de acordo com as políticas estabelecidas pela cultura organizacional da empresa matriz. Logo, essas condições podem desencadear diferenças no modo de redigir os diversos tipos de textos de uma dada especialidade em uma língua globalizada (a língua inglesa), isto é, pode haver interferências, quer decorrente de uma comunicação intercultural, quer da diversidade nas relações interempresas.

Nas ciências da saúde, em particular no campo das ciências farmacêuticas do segmento industrial, existem muitos documentos técnicos, por exemplo: laudos, métodos analíticos, guias direcionados às práticas inerentes à produção, análises e controles de drogas farmacêuticas, entre outros. Esses documentos servem de base aos especialistas que atuam em laboratórios e controle de qualidade, como meio de realizar as suas atividades profissionais. Muitos outros tipos de documentos de natureza jurídica, regulamentado por normas e procedimentos instituídos por convenções governamentais e internacionais do setor, bem como políticas globais da empresa, também devem ser produzidos pelos profissionais que desempenham funções administrativas na empresa.

Percebe-se, no entanto, que esses profissionais não possuem habilidades lingüísticas adequadas para atender às atividades que demandam uma produção escrita, em virtude, talvez, de possuírem formações acadêmica e profissional específicas, ou seja, grande parte dos funcionários envolvidos e atuantes em empresas do setor farmacêutico possui, basicamente, títulos acadêmicos e experiências profissionais em áreas específicas, tais quais: farmacêutica, química, biológica, médica, engenharia, de auditoria, entre outras. Assim, a pouca habilidade lingüística desses profissionais pode muitas vezes exercer influência sobre a prática da escrita.

---

[1] [...] *In practice, within many large companies, and even in parts of the European governmental institutions, English has become a common working place* […]

Dessa variedade de documentos técnicos, produzidos em língua inglesa e veiculados nas indústrias farmacêuticas, temos o documento *Site Master File*[2] (doravante SMF). O SMF é um dos principais documentos que uma empresa fabricante de medicamentos deve portar. Trata-se de um conjunto de textos produzido especificamente para cada empresa fabricante, independente de sua localidade, com o objetivo principal de certificar a garantia e qualidade dos produtos fabricados. Essa certificação é guiada por princípios estabelecidos mundialmente pelos órgãos governamentais de vigilância sanitária.

Desse modo, cada localidade ou unidade de negócios (*Business Unit)* de cada empresa farmacêutica deve elaborar e preparar o seu próprio SMF, e, em seguida, apresentá-lo à agência reguladora vigente de cada país em que essas estiverem instaladas. Os funcionários responsáveis pela elaboração desse documento estão distribuídos entre os diversos departamentos da empresa, e a equipe, responsável pela aprovação e expedição do documento SMF, está alocada no departamento de Garantia da Qualidade (em inglês *Quality Assurance*), o departamento que concentra todas as informações referentes ao atestado de qualidade do produto e ao controle das operações de produção.

Para a elaboração do SMF é necessário, além de um conhecimento técnico da estrutura organizacional e das operações da empresa, um conhecimento lingüístico na área da especialidade farmacêutica, o qual possibilite aos profissionais dos diversos departamentos redigirem, de forma coesa e coerente, os documentos que circundam a sua comunidade discursiva, ou seja, o contexto no qual estão inseridos.

Entre essas e outras razões apontadas acima, emerge a motivação para investigar o documento *Site Master File*, e melhor conhecer como os recursos lingüísticos podem ser articulados nas diferentes localidades em que o documento foi produzido. Este estudo justifica-se, pela necessidade profissional da pesquisadora como tradutora técnica em traduzir documentos do setor farmacêutico, e pelo interesse nas pesquisas a respeito de ensino de línguas para fins específicos baseadas em *corpus* de textos autênticos, como também em estudos sobre terminologia.

[2] A palavra *site* em língua inglesa corresponde em português à local, sítio. Porém, no contexto do documento SMF, a palavra *site* quer dizer planta ou fábrica. Portanto, *Site Master File* corresponde a Arquivo Principal da Planta.

Outrossim, a pesquisadora participa do grupo de pesquisa Abordagem Instrumental e o Ensino-Aprendizagem de Línguas em Contextos Diversos (denominado GEALIN)[3], que tem como foco central o ensino-aprendizagem de línguas para fins específicos, conforme excerto abaixo extraído da WWW (WordWideWeb) do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq):

> [...] Em razão das novas necessidades do mercado, dos novos recursos tecnológicos e das posturas didático-pedagógicas, faz-se necessária uma pesquisa detalhada dessas demandas. Contextos e ambientações diversos sugerem questões, quer para a identificação das necessidades dos diversos públicos alvo; para a descrição dos gêneros textuais por eles utilizados nos vários contextos; para o reconhecimento de estratégias associadas ao desenvolvimento de habilidades lingüísticas apropriadas às situações-alvo; para o desenho de cursos que venham a articular necessidades e expectativas dos alunos e suas respectivas instituições; para a análise, implementação e avaliação de materiais didáticos [...].

Um dos objetivos específicos do grupo de pesquisa GEALIN é desenvolver pesquisas para identificar as necessidades dos alunos no mercado de trabalho, e identificar e descrever gêneros textuais que esses alunos deverão utilizar ou utilizam. Dessa forma, buscamos formas de contribuir para a identificação de padrões léxico-gramaticais encontrados em um documento específico, utilizado por um público-alvo do contexto farmacêutico.

Assim, escolhemos investigar no contexto de trabalho da pesquisadora, um documento autêntico (o SMF) que estivesse agregado às atividades dos profissionais que atuam nas indústrias farmacêuticas, possibilitando uma análise criteriosa dos aspectos lingüísticos que tal material pudesse oferecer.

No entanto, na literatura acadêmica pesquisada, não encontramos registros de pesquisas voltadas ao estudo da linguagem no contexto empresarial farmacêutico. Na Pontifícia Universidade Católica de São Paulo, localizamos trabalhos desenvolvidos por pesquisadores do Projeto DIRECT (*Development of International Research in English for Commerce and*

[3] O GEALIN é certificado no CNPq pela PUC-SP e coordenado pelas professoras Dra.Rosinda de Castro Guerra Ramos e Dra. Maximina Freire.

Technology)[4]. Dentre os Direct Papers que abordam o discurso empresarial podemos destacar: Berber Sardinha, 1994a, 1994b, 2003, 2004c; Freire, 1995; Thompson e Ramos, 1995; Sobhie, 2003; entre outros.

Além das pesquisas do projeto DIRECT, encontramos outras pesquisas que abordam o discurso empresarial (Santos, 1999; Ramos, 1997; Bressane, 2000; Lima-Lopes, 2001; entre outros), e o estudo de gêneros em contextos empresariais diversos (Vian Jr., 1997; Santos, 2002; Carvalho, 2003; Rossini, 2005; entre outros). Tais pesquisas concentram-se em examinar: (i) a organização retórica que define como os documentos na empresa se organizam, e (ii) a léxico-gramática de textos específicos.

Portanto, diante dos estudos mencionados acima, esta pesquisa difere-se quanto à especialidade em questão, a área farmacêutica a qual o documento pertence, e aos pressupostos levantados com relação a uma possível padronização encontrada no discurso compartilhado entre os documentos, ou seja, a busca, por meio de *lexical bundles* recorrentes, pelas semelhanças encontradas nos SMFs das quinze diferentes localidades da empresa.

Conforme Biber et al. (1999:990), *lexical bundles são seqüências de palavras que co-ocorrem naturalmente no discurso*[5]. Em outras palavras, os estudos de *lexical bundles* se baseiam na combinação de palavras recorrentes no texto com base em *corpora* específicos, sejam elas extraídas de registros escritos ou orais. O termo *lexical bundle* foi cunhado por Biber et al. (1999), embora existam, segundo Cortes (2002), outras denominações para o termo na língua inglesa (detalhados no próximo capítulo). Ademais, temos a tradução do termo em língua portuguesa, atribuída a Berber Sardinha (2003), como 'pacotes lexicais'. Porém, devido à variedade de rótulos na literatura, e ao fato de o termo não ter sido consagrado no Brasil, optamos por utilizar neste trabalho, o termo *lexical bundle*.

A presente pesquisa está inserida nos princípios da Lingüística de *Corpus* (LC), que tem como um dos seus principais objetivos, a identificação da freqüência de padrões sintáticos e léxico-gramaticais em *corpus* autêntico da língua em uso, por meio de ferramentas computacionais.

---

[4] O DIRECT é um projeto iniciado em 1989 e desenvolvido pelo Programa de Pós-Graduação em Lingüística Aplicada e Estudos da Linguagem (LAEL) da PUC-SP em parceria com o Departamento de Língua Inglesa da Universidade de Liverpool, na Inglaterra. Os DIRECT Papers são publicações que envolvem trabalhos de pesquisa direcionados as habilidades de comunicações específicas no uso das línguas inglesa e portuguesa para propósitos de negócios. Tais trabalhos estão disponíveis eletronicamente em http://www2.lael.pucsp.br/direct/direct_papers.htm.

[5] No original: […] *sequences of words that commonly go together in natural discourse* […]

Assim sendo, dentro da abordagem da LC com base nas pesquisas de *lexical bundles*, destacamos os seguintes trabalhos: (i) Hyland (2008), que examinou um *corpus* a partir de registros acadêmicos escritos de quatro diferentes disciplinas (Engenharia Elétrica, Biologia, Administração de Empresas, e Lingüística Aplicada), a fim de extrair *lexical bundles* baseados numa análise contrastiva das formas e funções dos *bundles*; (ii) Scott e Tribble (2006), que realizaram, a partir de textos escritos extraídos do BNC, um levantamento do conjunto ou agrupamento de palavras (*cluster)* para verificar a variação estatística e a função de determinadas palavras na formação dos *bundles*; (iii) Cortes (2006), que investigou o uso das combinações recorrentes de palavras em registros escritos do contexto da disciplina de História, visando à apreensão, por parte de alunos universitários, de *lexical bundles* recorrentes em futuros trabalhos acadêmicos; (iv) Berber Sardinha (2003), que estudou *lexical bundles* (segundo tradução do autor, pacotes lexicais) recorrentes nas unidades internas de um gênero da área financeira para estabelecer os elos (*links*) de coesão com as sentenças do texto; (v) Levy (2003), que se baseou na gramática de Biber et al. (1999) para pesquisar a ocorrência de *lexical bundles* em uma variedade de registros das áreas profissionais e acadêmicas, com o objetivo de estabelecer um parâmetro de proficiência de alunos nativos e não nativos de língua inglesa, provenientes de uma universidade norte-americana; dentre outros trabalhos.

Este estudo busca trazer contribuições para o grupo de pesquisa GEALIN e para a área de ensino-aprendizagem instrumental de língua inglesa, no sentido de realizar um levantamento das características apontadas em textos produzidos por profissionais da área de negócios, devido à maioria dos profissionais, tanto de empresas nacionais como multinacionais, necessitar de um aperfeiçoamento para a produção escrita de gêneros e tipos de textos específicos em língua inglesa, de tal forma a suprir as necessidades de comunicação internacional nas transações entre seus clientes e parceiros comerciais.

Com efeito, é importante ressaltar que nesta pesquisa não nos aprofundamos em descrever *lexical bundles*, como o fez Biber et al. (1999) e outros autores (como veremos no Capítulo 1 adiante), mas fizemos uso dessa categoria com o propósito de examinar como os quinze exemplares do documento SMF assemelham-se entre as localidades, conforme dito no início desta apresentação, ou seja, percorremos as etapas de pesquisa com a finalidade de descobrir qual grau de conformidade com elementos lingüísticos, um documento normatizado e regulamentado, e escrito por diferentes autores em diferentes partes do mundo pode atingir.

Além do mais, devido aos documentos apresentarem uma estrutura organizacional semelhante, isto é, seções e tópicos numa mesma ordem organizacional, hipotetizamos, utilizando a categoria dos *lexical bundles*, que poderia haver uma regularidade nos padrões léxico-gramaticais a partir de uma comparação entre os textos. Dessa forma, decidimos focalizar o estudo somente nos *lexical bundles* semelhantes. Assim sendo, para esta investigação, fizemos uso de quinze documentos *Site Master File*, em circulação em uma indústria farmacêutica com unidades instaladas em vários países do mundo. Esses documentos estão divididos em unidades de negócios, cada um deles pertence a uma localidade onde a empresa possui uma fábrica para a produção de insumos e produtos farmacêuticos.

Inserida nesse contexto, esta pesquisa foi desenvolvida para atender os objetivos elencados a seguir:

- verificar a estrutura organizacional do documento;
- identificar os *lexical bundles* semelhantes entre os SMFs das quinze localidades;
- analisar a relação coesiva dos *lexical bundles* semelhantes com o tópico das seções dos documentos.

A metodologia de análise dos dados desta investigação se baseou na abordagem da LC, a qual permite investigar a língua em uso por meio de ferramentas computacionais, facilitando o processamento de um grande volume de dados em um tempo relativamente pequeno, reduzindo-os, segundo Scott e Tribble (2006:5), a padrões potenciais (*potential patterns*).

Como procedimento de análise, utilizamos para esta investigação a linguagem de programação Perl (Wall, 1987) e a ferramenta Cygwin, que é um emulador do sistema operacional Unix para Windows, ambiente para digitação de linhas de comando para processar a limpeza dos textos. Ainda, com o suporte de um especialista em *software*, foi desenvolvido um aplicativo denominado *Análise Lingüística*, e criado um Banco de Dados em SQL para realizar a comparação entre os *lexical bundles*.

Portanto, para atingir os objetivos propostos, as seguintes perguntas de pesquisa foram elaboradas:

(1) Quais são as semelhanças da estrutura organizacional do *Site Master File* entre as quinze localidades em que o documento circula?

(2) Quais são os *lexical bundles* semelhantes entre as seções dos quinze documentos SMF?

(3) Como os *lexical bundles* semelhantes refletem os tópicos das seções do documento SMF?

Em resumo, o presente estudo organiza-se da seguinte forma: no Capítulo 1 - Fundamentação Teórica - temos uma discussão a respeito dos aportes teóricos que fundamentam o trabalho. No início desse capítulo, discutimos as perspectivas de Análise de Texto e Discurso na visão de Stubbs (1996) e baseados no trabalho de Barbara e Berber Sardinha (2008).

Em seguida, apresentamos o conceito de *lexical bundles* defendido por Biber et al. (1999), abordando alguns autores como: Hyland (2008); Cortes, (2006), (2008); e Scott e Tribble, (2006), os quais desenvolveram pesquisas com *lexical bundles*. Logo após, discutimos os trabalhos de Ramos (1997) e Berber Sardinha (1997) quanto ao tópico textual (o *aboutness* do texto).

No Capítulo 2 - Metodologia de Pesquisa - apresentamos o processo metodológico incorporado à descrição do *corpus* de pesquisa, à coleta dos dados, aos procedimentos de análise, e à respectiva contextualização em que este estudo se insere.

No Capítulo 3 - Apresentação e Discussão dos Resultados - discutimos a análise dos resultados em três partes, a saber: (i) a estrutura organizacional do documento; (ii) os resultados apurados na extração dos *lexical bundles* semelhantes, e a comparação desses *bundles* com um *corpus* de referência (o BNC); e (iii) a associação dos *lexical bundles* semelhantes com o tópico das seções do SMF.

Por fim, nas Considerações Finais são apresentadas as conclusões deste estudo, juntamente com as contribuições que implicam em futuras aplicações para a área dos estudos da linguagem. As referências bibliográficas e os anexos seguem no encerramento da dissertação.

CAPÍTULO 1

FUNDAMENTAÇÃO TEÓRICA

O propósito deste capítulo é oferecer um embasamento teórico a esta pesquisa. Assim, o organizamos em três seções. Na primeira seção, discutimos alguns conceitos básicos a respeito da Lingüística de *Corpus* (doravante LC), especialmente nos aspectos relacionados aos diferentes *corpora* existentes, como também questões relacionadas à freqüência no uso das palavras, e a representatividade da extensão do *corpus* de estudo.

Ainda nessa seção, discutimos o trabalho de Michael Stubbs (1996) referente à análise de textos baseada em *corpora* com o auxílio de ferramentas computacionais, da qual abordamos a relação existente entre texto, *corpus* e significado. Embora o trabalho de Stubbs tenha como foco os estudos em *corpora* de inglês britânico (*the Standard English)*, e a maneira como a língua repercutiu no discurso da sociedade britânica das últimas décadas, foi possível nesta pesquisa replicar alguns princípios e pressupostos adotados pelo autor. Dentro dessa abordagem, apresentamos a proposta de Barbara e Berber Sardinha (2008) referente à associação entre Análise de Discurso e Lingüística de *Corpus*.

Na segunda seção, introduzimos o conceito de *lexical bundles* defendido por Biber et at. (1999), e as categorias funcionais e estruturais dos *bundles*, desenvolvida pelos autores no *The Longman Grammar of Spoken and Written English* (LGSWE), uma gramática descritiva da língua inglesa baseada em *corpora* de uma variedade de registros escritos e falados. Além do mais, apresentamos alguns trabalhos realizados com base nesses pressupostos, dando ênfase às pesquisas com *corpora* empreendidas por Cortes (2002, 2006, 2008); Biber et al. (2004); Scott e Tribble (2006); Hyland (2008); entre outros.

Para finalizar, na terceira seção, apresentamos e discutimos a relação coesiva dos *lexical bundles* com o tópico do texto (*aboutness*), apoiados nos trabalhos de Ramos (1997) e Berber Sardinha (1997).

1.1 A Lingüística de *Corpus*

A Lingüística de Corpus é um estudo com base em textos extraídos de *corpora* autênticos da língua em uso. Berber Sardinha (2004a:3) define a LC da seguinte maneira:

> [...] A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador [...].

Do ponto de vista do autor, o estudo da LC além de priorizar evidências empíricas, não limita o pesquisador na análise dos seus dados, ao contrário, a LC possibilita obter por meio de ferramentas computacionais, resultados diversos para o analista na formulação das suas idéias e prováveis respostas às suas questões de pesquisa.

Provavelmente, segundo Berber Sardinha (2004b), os primeiros estudos baseados em *corpora* lingüísticos foram os referentes às citações dos livros sagrados (a Bíblia), que foram compilados por monges da Idade Média com o propósito de se extrair as transcrições *tal qual aparecia no texto original* (2004b:2). Scott e Tribble (2006:4) afirmam que no final do século XIX, o dicionário de língua inglesa Oxford (*Oxford English Dictionary*) foi compilado a partir da coleta de fragmentos de um grande número de amostras autênticas da língua em uso, a qual, na opinião dos autores, foi considerada uma das iniciativas de pesquisas mais expressivas levada a cabo por colaboradores e estudiosos da linguagem.

Embora o uso de textos autênticos com base em *corpora* vem sendo explorado desde a antiguidade, houve, segundo Aijmer e Altenberg (1991:1), uma grande expansão da LC nas últimas décadas. Segundo os autores, essa expansão é devida a dois importantes eventos ocorridos na década de 1960. O primeiro evento foi o lançamento de um *corpora* do inglês britânico, o *The Survey of English Usage* (SEU), cujo *corpora*, na época, não era eletrônico. O SEU continha um milhão de palavras de textos de língua escrita e falada, e foi coordenado por Randolph Quirk e Greenbaum. O segundo evento foi o advento da computação, que viabilizou o armazenamento e processamento de um grande volume de dados lingüísticos.

Conforme esses autores, nessa mesma década, paralelo a esses acontecimentos surgiram duas referências à prospecção dos *corpora* de língua inglesa: o *The Brown Corpus*, um

corpora de um milhão de palavras do inglês americano, compilado por Nelson Francis e Henry Kučera, e o *The Lancaster-Oslo/Bergen Corpus* (LOB), um *corpora* no mesmo formato do *Brown Corpus,* contendo um milhão de palavras do inglês britânico, compilado e computadorizado pelos então pesquisadores da Lancaster University, Leech, Oslo e Bergen.

Já na década de 1970, o *The London-Lund Corpus of Spoken English* (LLC), coordenado por Jan Svartvik, empreendeu a tarefa de disponibilizar o SEU *Corpus* em formato eletrônico. Para Aijmer e Altenberg (1991:2), o trabalho de Svartvik e seus colaboradores estimularam os estudos referentes à língua falada, contribuindo para os inúmeros projetos de pesquisas que advieram da utilização de *corpora.*

Segundo Berber Sardinha (2004a:7-8), um dos principais *corpora* eletrônicos da língua inglesa, além do pioneiro *Brown Corpus* citado anteriormente, são o *British National Corpus* (BNC) e o *Bank of English*. O BNC foi lançado em 1995 e contém 100 milhões de palavras, além de ser o único *corpora* comercializado e de permitir a coleta de *corpora* de outros idiomas. Entretanto, conforme Berber Sardinha (2004a:9), ambos os *corpora Brown* e BNC são *corpora de amostragem, planejados e fechados*. Enquanto que o *Bank of English,* lançado em 1987, é um *corpus* dinâmico, ou seja, em expansão, o qual, segundo definição do autor, até o período de fevereiro de 2002 possuía 450 milhões de palavras da língua inglesa.

Para Tagnin (2007:160-162), o uso dos diferentes tipos de *corpora* existentes depende do objetivo da pesquisa. A autora classifica os *corpora* da seguinte maneira:

- fechados e aberto, como, por exemplo: o BNC e o *Bank of English* (nesse período apurados em 500 milhões de palavras de textos em inglês britânico, americano, australiano e canadense);
- monolíngües ou multilíngües, ou seja, em uma só língua, como o BNC, e em duas ou mais línguas, como o COMPARA[6];
- comparáveis ou paralelos: os comparáveis são entendidos como aqueles formado por originais em línguas distintas de um mesmo gênero e tipo de texto, e os paralelos são formados por originais e suas respectivas traduções;
- língua geral ou de língua de especialidade: os de língua geral são compostos por diversos gêneros para assegurar a representatividade da língua usada no

[6] Segundo Tagnin (2007:161), o COMPARA é um corpus paralelo contendo trechos de obras literárias em variantes das línguas portuguesa e inglesa com suas respectivas traduções.

> cotidiano, e os de especialidade são compilados para fins específicos, sendo bastante restrito.

No caso desta pesquisa, de acordo com a classificação de Tagnin (2007), podemos classificar o nosso *corpus* de estudo na categoria dos *corpora* de especialidades, compilados para fins específicos, pois este *corpus* de estudo foi baseado na especialidade farmacêutica e coletado a partir de um tipo de documento (o *Site Master File*), com acesso restrito aos usuários do setor farmacêutico. O documento somente pode ser manuseado e veiculado entre e pelos departamentos da empresa, ou por instituições governamentais vinculadas à saúde da população. Portanto, estamos tratando de uma língua de especialidade representativa do seu domínio de atuação.

Acerca das características dos *corpora* existentes para as pesquisas lingüísticas, é importante enfatizar que o uso de *corpora* computadorizados promoveu novas metodologias para a análise dos diferentes tipos de textos, possibilitando identificar padrões sintáticos e léxico-gramaticais no uso da língua, dentro de uma variedade de discursos. Por conseguinte, as novas tecnologias para o Processamento de Linguagem Natural (PLN)[7] trouxeram modernos *softwares* com ferramentas capazes de processar uma grande massa de textos em um curto espaço de tempo. Para citar um desses programas computacionais, temos o WordSmith Tools (WSTools), um programa desenvolvido em 1996 por Mike Scott da Universidade de Liverpool - Inglaterra. De acordo com Berber Sardinha (no prelo), esse programa possui um grande número de usuários no mundo, inclusive no Brasil, pois é um programa escrito em ambiente Windows® com abrangência mundial.

Conforme definição do autor, o WSTools é um conjunto de aplicativos integrados (suítes) para análise lingüística, o qual permite realizar análises de freqüência e recorrência de palavras em *corpora*, além de realizar o pré-processamento dos arquivos do *corpus*, isto é, extrair partes desnecessárias de cada texto, organizar arquivos, inserir etiquetas, etc.

Outrossim, de conformidade com o avanço de novas ferramentas para o processamento de língua natural, a metodologia deste trabalho fez uso de um aplicativo computacional projetado, exclusivamente, para analisar os *lexical bundles* no documento SMF, denominado *Análise Lingüística*. O *Análise Lingüística* foi baseado em uma plataforma de banco de dados relacional (SQL), a qual permite consultar, cruzar informações, fazer contagens, estabelecer

[7] Segundo Berber Sardinha (2004:5), o PLN é uma disciplina ligada a Ciências da Computação que compartilha assuntos com a LC, mas ambas as áreas mantêm-se independentes.

freqüências, e outras funções necessárias à análise dos dados (ver capítulo Metodologia – sub-seção 2.4.2.1).

Assim, muitos esforços estão sendo realizados a favor do desenvolvimento de tecnologias voltadas ao processamento de linguagem natural, tanto concernente a aspectos comerciais como acadêmicos. Scott e Tribble (2006:5) afirmam que desde a década de 1960, engenheiros de *hardware* e seus parceiros comerciais, por um lado, investiram em máquinas com maior potencial, e engenheiros de *software* do outro lado, implantaram uma série de ferramentas para a integração de ambientes computacionais, como, por exemplo: compiladores, interpretadores, *debuggers*, *profilers*, etc.[8], com a finalidade de oferecer recursos para o desenvolvimento de ferramentas apropriadas às necessidades de análise lingüística.

Dessa forma, corroborando Leech (1991:13):

> [...] podemos observar, num retrospecto histórico, como a disponibilidade de instrumentos via computador para análise de *corpus*, possibilitou aos fenômenos sintáticos e lexicais de uma língua dar vazão às investigações empíricas numa escala antes inimaginável [...] [9].

Uma outra premissa que também serve de base nos estudos de Lingüística de *Corpus* é a freqüência de uso das palavras, assunto que discutimos na sub-seção seguinte.

1.1.1 Freqüência de uso das palavras

Para Halliday (1991:31), o *sistema* lingüístico foi sempre inerentemente probabilístico e, por conseqüência, a freqüência no texto representa a probabilidade na gramática. O autor acredita que o estudo de *corpus* ocupa uma posição central na investigação da linguagem, e que um dos muitos caminhos a serem explorados diz respeito às evidências das freqüências relativas presentes na gramática, da qual se pode traçar um perfil probabilístico de sistemas gramaticais. Conforme Halliday (1991:41):

---

[8] Compiladores são ferramentas de tradução entre linguagens, isto é, entre linguagens de programação; entre ambientes para o processamento de linguagens naturais, dentre outras.

[9] Todas as citações mencionadas neste trabalho, na versão em língua inglesa, foram traduzidas pela pesquisadora. No original: [...] *we can see, in historical retrospect, how the availability of vastly increasing computer corpus resources has enabled syntactic and lexical phenomena of a language to be open to empirical investigation on a scale previously unimagined* […]

> [...] Sistematicamente, a gramática de uma língua pode ser representada, paradigmaticamente, como uma rede de escolhas, cada escolha consiste de um pequeno número de alternativas relacionadas à construção da 'informação' [...] [10].

A despeito dos padrões quantitativos na gramática estabelecerem uma relação de significado entre o léxico e a gramática, segundo concepções de Halliday (1991:37), é válido tratar de probabilidades globais na linguagem, ainda que, cada texto encontre-se situado em algum contexto específico, uma vez que podemos definir registros e tipos de textos por meio de referências obtidas das probabilidades gramaticais.

Halliday (1991:30), citando Svartvik, afirma que usando dados de *corpus* para estabelecer uma taxionomia de classes gramaticais, é possível calcular a proporção das categorias encontradas e relacioná-las a uma extensão de registros variados. Essa proposta de taxionomia de classes foi explorada por Biber et al. (1999) em seu trabalho referente às formas estruturais e funcionais do elemento lingüístico *lexical bundles* (ver sub-seção 1.2.1 neste capítulo).

Tal qual ocorre nos textos do SMF, a probabilidade de ocorrência de certas estruturas gramaticais encontradas nos *lexical bundles*, apontam para a função que esses elementos lingüísticos exercem no contexto específico do documento, ou seja, podemos relacionar os *bundles* com os seguintes elementos: os tópicos do texto, os participantes, os destinatários, e, consequentemente, com quaisquer outros traços envoltos pelo contexto farmacêutico.

Entretanto, em virtude do sistema probabilístico da linguagem, no qual certos traços lingüísticos são mais freqüentes que outros (Berber Sardinha, 2004a), vale ressaltar a necessidade de se considerar o significado da extensão do *corpus*, ou seja, a sua representatividade que também está relacionada à probabilidade, assunto que abordamos a seguir.

1.1.2 Representatividade

Berber Sardinha (2004a:23) acredita que devido à algumas palavras serem difíceis de ocorrer no *corpus* é necessário trabalhar com uma grande quantidade de palavras para que haja probabilidade com relação a sua freqüência. Na opinião do autor:

---

[10] No original: [...] *Systematically, the grammar of a language can be represented paradigmatically as a network of choices, each choice consisting of a small number of alternatives related to the construing of ´information´*[…]

> [...] O *corpus* é uma amostra de uma população cuja dimensão não se conhece...Desse modo, não se pode estabelecer qual seria o tamanho ideal da amostra para que represente essa população. Uma salvaguarda é tornar a amostra a maior possível, a fim de que ela se aproxime ao máximo da população da qual deriva, sendo, portanto, mais representativa [...]

Ainda, Berber Sardinha (2000) afirma que:

> [...] A linguagem é de caráter probabilístico, e, sendo assim, há a possibilidade de estabelecer uma relação entre traços que são mais comuns e menos comuns em determinado contexto. O conhecimento da probabilidade de ocorrência de traços lexicais, estruturais, pragmáticos, discursivos, etc., está no cerne da Lingüística de Corpus [...]

No entanto, há polêmicas entre os pesquisadores de *corpus* com relação à delimitação de um *corpus* adequado para o estudo da língua a qual se quer representar. Vasilévski (2007), citando Rocha (2007), diz que a questão da representatividade não é simples, pois a *representatividade de um corpus é seguramente uma ilusão e que é fácil contestá-la, ao passo que é difícil assegurá-la* (2007:53), conquanto essa dificuldade não impossibilite uma aproximação confiável para a representação de um *corpus*. Segundo a autora:

> [...] Se o pesquisador delimita propriamente o alcance dos resultados que obtém do *corpus*, já está contribuindo para assegurar a representatividade dele. Ainda, sendo a língua infinita, a única forma de representá-la para fins de estudo é por meio de sua incontestável realização. Então, considera-se que essa polêmica não deve ser tão preocupante quanto se quer fazer parecer [...] (Vasilévski, 2007:53).

Quanto à probabilidade de ocorrências, Rocha (2007:200) explica que na Lingüística de *Corpus* as medições são baseadas em freqüências, e os dados de freqüência:

> [...] constituem a forma de caracterização de um *corpus*, registro ou texto. Porém, as freqüências com que uma determinada característica aparece em dois *corpora* diferentes só podem ser comparadas se houver algum tipo de normalização que permita a comparação [...]

Essa normalização, a qual o autor se refere no excerto acima, diz respeito a um cálculo com base pré-determinada, ou seja, um cálculo que estabeleça, por exemplo, a probabilidade de um *bundle*, ou de diferentes classes de palavras, ocorrerem a cada mil palavras. Desse modo, adaptando de Rocha (2007:201) uma demonstração desse procedimento, utilizando a amostra *according to the* extraída do *corpus* do SMF (total de 110.496 *bundles*), e comparando-a ao *corpus* do BNC (total de 4.423.944 *bundles*), teríamos o seguinte cálculo:

SMF (*according to the*): 58/ 110.496 x 1.000 = 0,525 a cada mil palavras
BNC (*according to the*): 4.640/ 4.423.944 x 1.000 = 1,049 a cada mil palavras

Segundo Rocha (2007), é possível utilizar uma base 100 para as normalizações, base essa mais comumente conhecida entre o público em geral, a qual demonstra percentagens ou proporções. Porém, Rocha acredita que para a Lingüística de *Corpus*, a base em 1.000 oferece uma comparação mais evidente. Isso significa dizer, que mesmo que a demonstração em percentagens seja de uso mais comum, a comparação expressa no exemplo acima, de que a cada mil palavras o *bundle according to the* ocorre em média 1,049, é mais fácil de interpretar do que se a mesma comparação fosse expressa em percentuais, isto é, 0,105% do total de 4.423.944 *bundles.*

Dessa maneira, estando a representatividade ligada a questões de probabilidades, e, conforme afirma Berber Sardinha (2004a:23-24), uma amostra deve conter um número considerável de textos e palavras que determinem a sua representatividade numa dada extensão de *corpus*, o autor aponta para duas questões: o corpus deve ser *representativo do quê e para quem*?

Na primeira questão, Berber Sardinha (2004) suscita que estimar a dimensão da população total da linguagem não é possível, portanto, não podemos afirmar que um determinado *corpus* seja representativo, mas podemos descobrir, por meio da probabilidade de ocorrências, quais elementos lingüísticos (lexicais, estruturais, pragmáticos ou discursivos) são mais comuns ou menos comuns dentro de um contexto específico. Na segunda questão, Berber Sardinha (2004a) refere-se à responsabilidade do usuário do *corpus* pela representatividade dos seus achados de pesquisa. Nas palavras do autor *tem-se falado em representatividade como um ato de fé* (2004a:25), ou seja, o usuário deve ser cauteloso em

considerar ou acreditar que uma determinada amostra possa ser representativa de uma população inteira.

Diante dessa perspectiva, Berber Sardinha sugere três abordagens para se definir a representatividade de um *corpus*, a saber: (1) impressionística; (2) histórica; e (3) estatística. Para o autor, a abordagem impressionística baseia-se na criação e exploração de *corpora* feita por autoridades da área, a histórica diz respeito aos *corpora* monitorados e usados pela comunidade, e a estatística se baseia na aplicação de dados estatísticos, mais propriamente a aplicação de fórmulas matemáticas.

Em se tratando do *corpus* deste trabalho, não podemos afirmar que as duas primeiras definições fornecidas acima pelo autor se refiram ao nosso *corpus* de estudo. Entretanto, conforme foi possível verificar no capítulo Metodologia - seção 2.4.4. deste trabalho, este *corpus* de estudo se tornou representativo na medida em que estabelecemos um *baseline* com um *corpus* de larga extensão (o BNC), possibilitando analisar estatisticamente, por meio de aplicações computacionais, a probabilidade de ocorrências dos *lexical bundle*s encontrados no SMF, comparados aos *lexical bundles* encontrados no *corpus* do BNC. Dessa forma, a abordagem utilizada para medir a representatividade do *corpus* utilizado nesta pesquisa se baseou em métodos estatísticos, terceiro item sugerido por Berber Sardinha (2004a:25) quanto a considerar uma amostra representativa de determinado domínio específico.

Para Berber Sardinha (2004a), é de se esperar que um *corpus* compilado de um tipo de texto específico apresente menor variação e maior padronização nos níveis lingüísticos, visto que possibilita um *maior grau de fechamento* (:28) do que os textos de domínio específico encontrados nos *corpora* gerais, como o BNC, *Bank of English*, *Brown*, etc. Assim sendo, conforme afirma Berber Sardinha (2004a:28), *normalmente, corpora compilados em pequena escala por pesquisadores individuais acabam sendo mais representativos do que os respectivos subcorpora dos corpora gerais*.

Desse modo, com a finalidade de discorrermos sobre algumas das asserções a respeito do significado da relação entre texto-significado-instituição inerentes a *corpora* autênticos, apresentamos na sub-seção seguinte, a proposta defendida pelo lingüista de *corpus* Michael Stubbs.

1.1.3 Análise textual de *corpora* na visão de Michael Stubbs

O conceito de texto adotado por Stubbs (1996:4) define texto como um exemplo da língua em uso (tanto falada como escrita), uma parte do funcionamento da língua que ocorre naturalmente sem a intervenção do lingüista[11]. Stubbs (1996) apresenta uma abordagem das pesquisas realizadas com amostras de *corpus* autêntico e oferece métodos para analisar como esses dados são aplicados pelos usuários na vida real, em particular, como os padrões das palavras encontradas nos textos e a gramática podem conduzir ao significado do texto em determinados contextos ou instituições. Para ilustrar essa relação de texto-significado-instituição, Stubbs (1996:6) levanta algumas questões que surgem das diversas formas presentes em *corpor*a de textos autênticos, como por exemplo:

> [...] Onde está localizado o significado de um texto? O significado está no interior do próprio texto? O significado está na memória das pessoas as quais dão sentido ao texto? O significado está em alguma parte da comunidade discursiva? – Talvez esteja numa interpretação consensual, com a qual todos pudessem concordar. [...] [12]

Podemos acrescentar a essas perguntas formuladas por Stubbs, o seguinte: estaria o significado de um texto vinculado ao conhecimento específico de língua de uma dada área do conhecimento, isto é, no entendimento de uma terminologia específica? É provável que essa questão possa ser respondida por meio de uma apreciação do texto *a partir da observação das escolhas léxico-gramaticais* (Barbara e Berber Sardinha, 2008:291).

Embora a questão de autoria não seja, por ora, um dos propósitos deste trabalho, levantamos uma breve discussão a respeito do assunto a fim de viabilizar a possibilidade de uma análise textual por meio da qualidade de redação dos escritores e, por conseguinte, o entendimento por parte dos seus interlocutores. De acordo com Barbara e Berber Sardinha (2008), uma das tarefas na apreciação de um texto seria a abordagem crítica utilizando a Análise do Discurso, *buscando notar, por exemplo, a ideologia que se apresenta nele, o posicionamento do autor, entre outras questões* (:291)

---

[11] No original: […] *By text, I mean an instance of language in use, either spoken or written: a piece of language behaviour which has occurred naturally, without the intervention of the linguist* […].

[12] No original: […] *where is the meaning of a text located? Is the meaning inside the text itself? Or inside the mind of the person who makes sense of it? Or is it in the speech community somewhere – perhaps in the form of a consensus interpretation on which we could all agree*? […]

Ademais, no mundo atual, é possível levarmos em consideração os textos produzidos por instituições empresariais, caso desta pesquisa. Nesses, os autores na maioria dos casos não são identificados, ou porque são múltiplos, ou porque são preparados ou treinados para produzirem os seus textos conforme os interesses da instituição (cf. Stubbs, 1996:7), instituições essas que, por certo, assumirão a autoria. No entanto, muitos desses autores não são profissionais da escrita, tais como: jornalistas, roteiristas, revisores, profissionais acadêmicos, entre outros, que provavelmente possuem, por natureza da profissão, um melhor desempenho na escrita, ao contrário daqueles profissionais, como, por exemplo, os participantes do contexto desta pesquisa, que necessitam redigir por exigência da instituição de negócios para qual representam e na qual atuam, implicando muitas vezes na qualidade dos textos.

Nesse sentido, para exemplificar os escritores do documento SMF, eles são múltiplos porque atuam em diferentes departamentos da empresa, e cada parte constituinte do documento é redigida por um especialista da área de atuação, entretanto, não são anônimos nem tampouco são fantasmas (cf. Stubbs, 1996:7), são colaboradores responsáveis em reportar todas as informações inerentes às atividades da empresa para a fabricação dos produtos farmacêuticos.

A partir desses pressupostos, corroboramos a proposta de Barbara e Berber Sardinha (2008:314) de que a Análise do Discurso combinada aos estudos de *corpus* auxiliam:

> [...] no entendimento das escolhas que perfazem a construção de um texto, tanto do ponto de vista da constituição das escolhas individuais, referentes à redação do texto, quanto das escolhas mais amplas, que revelam o posicionamento do autor e do meio em que o texto é veiculado [...]

Com efeito, para investir nessa empreitada, são necessários instrumentos eficazes que possibilitem explorar os textos na suas formas lingüísticas (e.g. léxico-gramaticais) e nos seus aspectos descritivos. Por exemplo, explorar textos do tipo: documentos específicos, artigos, textos acadêmicos, entre outros, já que o volume de dados extraídos de um *corpus* oferece resultados precisos ao analista. Para tanto, um dos recursos que auxiliam na apuração dos dados é o uso de ferramentas computacionais, aliadas à coleta de *corpora* eletrônicos de textos autênticos, dos quais os lingüistas de *corpus* fazem uso e defendem há mais de duas décadas. Segundo Barbara e Berber Sardinha (2008:292):

> [...] Conquanto a ênfase dos estudos baseados em *corpora* seja a descrição de línguas, tem havido, cada vez mais, incursões de lingüistas de *corpus* em áreas como a do ensino de língua estrangeira, em que *corpora* eletrônicos são criados e empregados, por exemplo, para a produção de materiais de ensino e para o estudo exploratório centrado no aprendiz, sem falar da própria descrição da língua do aprendiz [...]

Contribuindo ao excerto de Barbara e Berber Sardinha, podemos dizer que além dos usos de *corpora* autênticos, listados acima, serem empregados tanto na produção de materiais de ensino como nos estudos e descrição da língua de aprendizes, também podem oferecer meios de descrever a língua de especialistas (médicos, farmacêuticos, biólogos, engenheiros, advogados, e outros), com o fim de oferecer-lhes um melhor conhecimento lingüístico, ou mesmo de identificar, a partir das escolhas lingüísticas, possíveis variações que ocorrem não somente no nível léxico-gramatical, mas também no nível *sócio-situacional*, caso desta pesquisa.

Stubbs (1996:20) classifica esses dois níveis como sendo a relação entre a micro estrutura textual e a macro estrutura das instituições sociais. Na visão do autor, uma vez combinados os elementos: texto ↔ tipo de texto ↔ *corpora* ↔ instituições, esses podem garantir um estudo das bases empíricas da sociedade. Segundo o autor:

> [...] Os textos, escritos ou falados, abrangem uma grande parte das bases empíricas da sociedade, eles ajudam a construir a realidade social e a análise textual é uma perspectiva de observação da sociedade, ela concretiza as estruturas ideológicas [...] [13]. (Stubbs, 1996:21).

Conforme os fundamentos levantados por Stubbs, uma análise textual baseada no uso da língua do mundo real, ao contrário de dados lingüísticos intuitivos, e assistida por instrumentos tecnológicos, como, por exemplo, *softwares* computacionais ou métodos estatísticos, revelam características da língua que, de maneira geral, a intuição dos lingüistas não é capaz de alcançar sem o auxílio de *corpora* eletrônicos.

Um exemplo disso foi possível observar na metodologia deste trabalho, pois sem o auxílio das ferramentas computacionais utilizadas, encontraríamos um grau de dificuldade

[13] No original: [...] *Texts, spoken and written, comprise much of the empirical foundation of society: they help to construct social reality. And textual analysis is a perspective from which to observe society: it makes ideological structure tangible* […]

acentuado para cruzar o volume dos dados de pesquisa, a fim de que pudéssemos obter resultados de análise confiáveis e precisos.

Retomando os níveis micro e macro estruturais do texto, defendidos por Stubbs (1996), verificamos que no nível micro-estrutural, isto é, léxico-gramatical, é possível examinar as colocações (*collocations*) das palavras individuais que ocorrem no texto, as palavras-chave (*keywords*), e o conjunto ou pacotes de palavras (*lexical bundles* ou *clusters*) recorrentes. No nível macro-estrutural, isto é, discursivo, é possível estabelecer uma associação semântica entre os elementos léxico-gramaticais extraídos do *corpus*, e os contextos cultural, social, situacional ou ideológico no qual determinado tipo de texto está inserido.

Nessa relação micro↔macro, Stubbs (1996:21) afirma que texto e análise de *corpus* fornecem métodos para estudar os dados empíricos que permeiam a sociedade. Do mesmo modo, fundamentados no contexto pedagógico, Barbara e Beber Sardinha (2008:314) acreditam que a utilização dos instrumentos e da metodologia da Lingüística de *Corpus* para uma Análise do Discurso pode centrar-se em dois aspectos que beneficiam tanto aprendizes de língua quanto professores e analistas. O primeiro aspecto é o aproveitamento desses recursos lingüísticos para um exame criterioso e detalhado das escolhas lingüísticas feitas por redatores de textos diversos, e o segundo é o uso desses dados para o ensino de redação em língua materna em cursos universitários, ou até mesmo aos profissionais da escrita.

Com relação ao aproveitamento dos recursos lingüísticos para o ensino de redação, discutidos acima pelos autores, podemos considerar que a exploração lingüística com base em dados autênticos e com o auxílio da computação, decerto, pode ser, apesar da resistência ou mesmo da falta de interesse por parte de alguns lingüistas, um caminho mais prático e confiável a percorrer.

Segundo Sinclair (1994, *apud* Barbara e Berber Sardinha, 2008), a Lingüística de *Corpus* e a Análise do Discurso lidam com dimensões de padrões além do que a lingüística está acostumada, *ambas gerenciam a evidência de que precisam e mesmo para achar um pouco dessa evidência é preciso se fazer valer do poder da informática moderna* (2008:291).

Assim, encerramos essa seção, e em seguida apresentamos os conceitos que fundamentam a abordagem com *lexical bundles.*

1.2 O conceito de *lexical bundles*

Um dos principais fundamentos propostos para este trabalho se baseia nos estudos das combinações de palavras recorrentes no texto com base em *corpus* específico. Esse tipo de combinação de palavras foi nomeado por Biber et al. (1999) como *lexical bundles*, termo adotado para a presente pesquisa. Desse modo, destacamos algumas definições e explanações apresentadas por lingüistas que tiveram suas pesquisas direcionadas por *corpora* para a extração de *lexical bundles*, tanto da língua inglesa como das línguas portuguesa e espanhola.

Segundo Biber et al. (1999:990), *lexical bundles são seqüências de palavras que co-ocorrem naturalmente no discurso*[14], ou seja, são formados por expressões recorrentes, independente de sua idiomaticidade ou de sua condição estrutural. Biber et al. (1999) consideram expressões tais como: *the end of the, in addition to the, the point of view of*; e muitas outras, como exemplos de *lexical bundles.*

Para Berber Sardinha (2003), que utiliza o termo em português 'pacotes lexicais', esses *se referem a uma seqüência de palavras de tamanho determinado, retirado do texto ou corpus por um programa de computador* (2003:6).

Por sua vez, Cortes (2006:392) define *lexical bundles* como seqüências de três ou mais palavras, identificadas empiricamente em um *corpus* de língua natural. Embora do ponto de vista da autora, a aquisição e o uso apropriado dessas seqüências não seja um processo tão natural, dado a importância em considerar o significado que essas expressões apresentam em determinadas disciplinas.

Stubbs (2007) apóia-se no conceito de seqüência múltipla de palavras (*multi-word sequence*) para referir-se aos estudos baseados na extração do conjunto de palavras ininterruptas recorrentes no texto por meio de programas computacionais. Stubbs (2007:90) afirma que uma das possíveis explanações para definir essa seqüência ininterrupta de palavras está fundamentada nos modelos de linguagem *n-grams*, o qual fornece, com o auxílio de instrumentos computacionais, um determinado número de palavras em conjunto ao mesmo tempo, ordenando-as alfabeticamente ou por freqüência. Para o autor, não existe um termo padrão para essa seqüência de palavras, as quais podem ser atribuídas às expressões: *statistical phrases*; *recurrent word combinations*; *lexical bundles, cluster; chains; multi-word sequences;* ou mesmo *n-grams*.

[14] No original: […] *sequences of words that commonly go together in natural discourse* […]

Os modelos de linguagem *n-grams* foram desenvolvidos no âmbito da Estatística pelo matemático russo Andrey Markov (1856-1922), a fim de reconhecer padrões estatísticos do uso da língua baseados em cadeias, conhecidas como cadeias de Markov (*Markov chains*). Manning e Schütze (1999:192-193) explicam que os modelos estatísticos *n-grams* modelam a probabilidade de encadeamento das palavras, isto é, nessa cadeia a palavra que antecede poderá prever a palavra que sucede, possibilitando conhecer quais palavras tendem a acompanhar outras palavras[15]. Para isso, segundo os autores, é necessário aplicar alguns dos métodos de agrupamento das cadeias de Markov. Por exemplo: se construirmos um modelo onde uma determinada palavra recorra frequentemente como última palavra (n-1)[16], então essas palavras serão posicionadas em uma categoria de equivalência, e teremos, dessa forma, um modelo de palavra *n-gram* ou modelo de seqüência de Markov (*order Markov model*). Para exemplificar a lógica desse método de agrupamento, podemos encontrar maiores detalhes na seção 2.4.2.1 do capítulo Metodologia desta pesquisa.

Segundo Manning e Schütze (1999:193), o termo *gram* tem suas raízes no idioma grego e deveria estar atrelado aos prefixos dos números gregos, como: *di, tri, penta, tetra, hexa,* e assim por diante. No entanto, segundo os autores, devido à miscelânea ocorrida nas últimas décadas com relação à nomeação dos termos científicos, influenciada principalmente pelos idiomas grego, latim e inglês, esse uso não sobreviveu. Para os autores, essa influência de idiomas resultou, nos casos de *n-grams* (2, 3, 4 ou demais palavras), denominações do tipo: modelos *bigram, trigram, four-gram* ou mesmo *quadrigram*, esse último com prefixo em latim.

É provável que dessa imprecisão em nomear as seqüências ininterruptas de palavras, surge uma das razões que levaram outros pesquisadores de áreas afins, como, por exemplo, os lingüistas aplicados, a utilizarem diferentes denominações para esse conjunto de palavras, conforme referido anteriormente por Stubbs (2007). Para ilustrar, temos alguns lingüistas como Biber et al. (1999), que ao invés de utilizarem a denominação *gram* precedido do prefixo por extenso (*bi, tri, four* - conforme os modelos estatísticos), nomearam de maneira diferente essas combinações, de forma numérica e por extenso: *three-four-five-six word bundles* ou 3-4-5-6 *word bundles*, termos que frequentemente encontramos nos estudos referentes às seqüências ininterruptas de palavras.

---

[15] Corroborando a famosa citação de Firth (1951 *apud* Hyland, 2008:5) que diz: *you shall judge a word by the company it keeps*.

[16] Esses dados estatísticos poderão ser estudados com maiores detalhes na obra dos autores entitulada *Foundations of Statistical Natural Language Processing* (Manning & Schütze, 1999). Nesta seção apenas apresentamos alguns conceitos básicos para o entendimento dos *lexical bundles.*

Visando contribuir com essa discussão, Cortes (2002:12) comenta, que uma das principais *questões que devem ser levadas em consideração nos estudos sobre combinações de palavras referem-se ao problema dos* rótulos[17], e que muitos nomes foram inventados para denominar essas combinações, e outros tantos estão sendo criados. A autora apresenta em seu trabalho, uma lista por ordem alfabética de alguns dos termos em língua inglesa utilizados na literatura para descrever as combinações recorrentes de palavras, os quais estão listados no quadro a seguir:

Quadro 1. Termos utilizados na literatura para *lexical bundles* (extraído de Cortes, 2002:12).

building blocks	idiomatic expressions	ready-made expressions
chunks	idioms	ready-made formulas
clichés	lexical bundles	ready-made utterances
collocations	lexical phrases	routine formulaes
conventionalized forms	lexical units	sentence builders
fixed expressions	multiword units	
formulaic language	prefabricated routines	
formulas/formulae	prefabricated patterns	
frozen phrases		

Diversos lingüistas aplicados como DeCock (1998), Cortes (2002, 2006, 2008), Stubbs (2002, 2007), Berber Sardinha (2003), Partington e Morley (2004), Scott e Tribble (2006), Hyland (2007, 2008), entre outros, desenvolveram trabalhos a respeito de seqüências ininterruptas de palavras. Conforme já citado, destacamos Biber, o qual, dentre os seus diversos trabalhos baseados em *corpora* do inglês americano (AmE) e britânico (BrE), dedicou-se à recorrência de seqüência de palavras nos discursos oral e escrito, em particular, nos registros acadêmicos. Ademais, Biber é responsável por diversas pesquisas e obras apontadas como referência em metodologias que abordam *lexical bundles* (Biber et al.,1999; Biber et al., 2003; Biber et.al., 2004; Biber, 2006; entre outros).

Para Biber et al. (1999:990-991), as combinações de palavras por seqüências ininterruptas não são unidades estruturais completas ou bem formadas, do mesmo modo que não são expressões lexicais fixas ou idiomáticas. Por outro lado, segundo os autores, para que

[17] No original: [...]..*issues that must be considered when reviewing studies on word combinations…deals with the problem of labels* […].

as combinações sejam consideradas *lexical bundles*, essas devem recorrer frequentemente entre uma larga extensão de textos, entre cinco ou mais textos, distribuídos entre registros variados para evitar tendências idiossincráticas por parte do usuário da língua.

Porém, esta pesquisa se diferencia dos pressupostos de Biber et al. (1999), nos seguintes aspectos: (i) quanto ao tamanho do *corpus* compilado e (ii) quanto à recorrência entre registros variados. Para justificar, no primeiro aspecto, temos um *corpus* de estudo com média de 110.000 palavras, portanto, considerado um *corpus* de pequeno-médio porte, segundo classificação de Beber Sardinha (2004:26). No segundo aspecto, nosso *corpus* de estudo foi coletado a partir de um único tipo de texto (composto de quinze exemplares) para investigar os *lexical bundles* recorrentes. Dessa forma, esses dois aspectos diferenciais, mesmo que não estejam de acordo com alguns dos critérios estabelecidos por Biber et al. em seu trabalho, não invalidam as combinações extraídas do SMF como sendo *lexical bundles*, porque, em menor escala, adotamos critérios na metodologia desta pesquisa que possibilitaram observar as ocorrências dentro de um mesmo documento, bem como identificar expressões de um domínio específico do contexto farmacêutico.

Com relação à metodologia adotada na pesquisa com *lexical bundles*, apresentada na gramática descritiva de língua inglesa desenvolvida por Biber, et al. (1999) (o *Longman Grammar of Spoken and Written English* - LGSWE), os autores estabeleceram uma freqüência na extração dos *bundles* com critério de corte de no mínimo dez vezes por milhão de palavras, ou seja, um *lexical bundle* deveria recorrer, em um determinado *corpus*[18] selecionado pelo pesquisador, pelo menos dez vezes em cada milhão de palavras, para então ser considerado representativo num dado contexto. Contudo, Biber at al. (2004:376) acreditam que esse critério de corte *é relativamente arbitrário*[19], pois essa escolha dependerá dos objetivos apontados na pesquisa e das questões levantadas pelo pesquisador e, principalmente, do tamanho dos *corpora* explorados.

Nos estudos de Biber et al. (1999:994), os pesquisadores optaram por *bundles* de quatro palavras (*four-word bundles*), utilizando um *corpora* compilado a partir de registros acadêmicos e de conversações. Segundo os autores, essa opção por *bundles* de quatro palavras se deve ao fato de serem mais comum de ocorrer dos que os de cinco ou mais palavras, além de apresentarem estruturas e funções mais elucidativas.

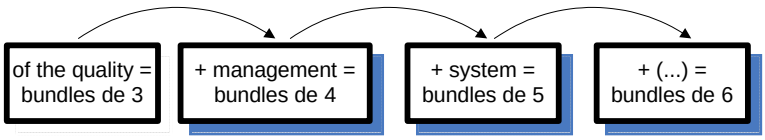
---

[18] Esse é um procedimento estatístico para ajustar as freqüências dos diferentes tamanhos de *corpora* e estabelecer uma comparação confiável na execução da contagem do número de palavras no *corpus* e o corte estipulado pelo pesquisador na extração dos *bundles*. Esse procedimento é denominado "normalização". (detalhes consultar Biber et al., 1999 e Rocha, 2007).

[19] No original: [...] *The actual frequency cut-off used to identify lexical bundles is somewhat arbitrary* […].

Inseridos nessa lógica, durante o processamento eletrônico dos *lexical bundles*, os autores identificaram que os *bundles* de três palavras (*three-word bundles*) ocorriam em média 80.000 vezes por milhão de palavras em conversações e 60.000 vezes por milhão de palavras em registros acadêmicos. Todavia, os *bundles* de quatro palavras (*four-word bundles*) ocorriam em média 8.500 vezes por milhão de palavras em conversações e 5.000 vezes por milhão de palavras em registros acadêmicos.

Dessa maneira, notou-se que os *lexical bundles* de três palavras ocorriam com maior freqüência por serem considerados um tipo estendido. Um tipo estendido significa que se extraíssemos um *lexical bundle* de três palavras, como, por exemplo: *of the quality* (do *corpus* SMF), em uma dada seqüência de encadeamento teríamos um *lexical bundle* de quatro palavras, como, por exemplo: *of the quality management*, e sequencialmente um *lexical bundle* de cinco palavras, por exemplo: *of the quality management system*. Por esse motivo é que Biber et al. (1999:992) afirmam que *lexical bundles* mais longos, como os de cinco ou seis palavras, são mais fraseológicos, em virtude de encapsularem *bundles* de três palavras em sua formação. Portanto, de acordo com o exemplo deste *corpus*, demonstrado acima, a partir do conjunto de três palavras, dá-se início à formação de seqüências maiores, explicando o porquê de os *bundles* de três palavras serem considerados um tipo estendido.



Dessa forma, à medida que a seqüência de palavras aumenta, a probabilidade de recorrência de *bundles* diminui, e assim, o ponto de corte pode variar, conforme a extensão do *bundle* que o pesquisador determinar e do tamanho do *corpus* disponível para a análise. A extração do tamanho do *bundle* fica a critério do pesquisador, sendo que tanto tamanhos menores como tamanhos maiores de *bundles* contribuem para a análise. No caso desta pesquisa, o fato de o *corpus* de estudo ser de pequeno porte, optamos pela extração de *bundles* de três palavras, porque esse critério favoreceu extrair uma quantidade maior de conjunto de palavras. Por outro lado, se optássemos por *bundles* de quatro ou cinco palavras teríamos uma quantidade menor para a análise.

Conforme Berber Sardinha (2003), um *lexical bundle* de duas palavras (um *bigram),* por exemplo, pode conter apenas fragmentos que são partes de *bundles* maiores, e

corroborando Biber et al.(1999), o autor acredita que trabalhar com *bundles* com seqüências maiores (três ou mais palavras) os tornam mais informativos.

Berber Sardinha (2003:7), em seu estudo sobre a análise de *lexical bundles* de três palavras, recorrentes nas unidades internas de um gênero da área financeira, afirma que o tamanho de três palavras é adequado para uma análise textual, pois evita incorrer em muitos fragmentos, como é o caso dos de duas palavras, não limitando a quantidade de *bundles* para extração, razões apontadas anteriormente para a escolha de *bundles* de três palavras nesta pesquisa.

Quanto às questões estruturais e funcionais dos *lexical bundles*, Biber et al. (1999:1001) verificaram que apesar de essas combinações recorrentes de palavras não representarem unidades estruturais completas, elas poderiam ser agrupadas gramaticalmente dentro de categorias associadas às suas estruturas correspondentes. Assim sendo, os autores dividiram os *lexical bundles* em dois grupos: os de conversações e os de discursos acadêmicos. Em conversação, os autores elegeram 14 categorias estruturais para análise, e em discurso acadêmico, 12 categorias. Diante dessas categorias de análise, Biber et al. (1999) estabeleceram algumas funções pertinentes aos *bundles*, com o propósito de sustentar os significados dessas seqüências de palavras nos textos.

É importante esclarecer, que as categorias de análise de Biber et al. (1999) se referem aos estudos com análise de *lexical bundles*, as quais estão contempladas na gramática de autoria dos pesquisadores (o LGSWE), já mencionada anteriormente.

1.2.1 Formas estruturais e funcionais dos *bundles*

De acordo com as categorias estruturais e funcionais de análise de *lexical bundles*, proposta por Biber et al. (1999), os autores afirmam que a maioria dos *bundles* encontrado nos registros acadêmicos e de conversações, não representa unidades estruturais completas. Porém, ambos os registros apresentam *lexical bundles* com estruturas gramaticais distintas.

Nos registros acadêmicos, na maioria das vezes os *bundles* revelaram duas unidades estruturais acopladas, sendo: um grupo nominal (*noun phrase*) + o início de um grupo preposicional (*prepositional phrase*), por exemplo: *as a result of, in the case of, the presence*

of a, entre outros[20], caso dos *bundles* encontrados nesta pesquisa, conforme será visto ainda. Ainda mais, frequentemente essas unidades terminam em artigo ou preposição.

Biber et al. (1999:999) mencionam, que quando os *bundles* recorrentes nos registros acadêmicos representam estruturas completas, esses por sua vez se apresentam, tipicamente, como grupos preposicionais que funcionam como um aparelho sinalizador do discurso[21], tais quais: *in the present study, in the next chapter*, *in the same way*, entre outros.

Para Biber et al. (1999:1000), os *bundles* de conversações, em contraste com os de registros acadêmicos, na sua grande maioria se apresentam como segmentos frasais (*clause segments*), quer dizer, como estruturas de tipo declarativo (*declarative structures*), com um pronome sujeito seguido de um grupo verbal estendido, ou como estruturas de tipo interrogativo (*interrogative structures*). Alguns exemplos de *bundles* dos registros de conversações desses segmentos frasais encontrados em Biber et al. (1999:1000) são os seguintes:

Declarative clause:	*Interrogative clause*:
I don´t know what	*Can I have a*
I said to him	*What are you doing*
Well you´ll have to	*I thought it was*

Biber et al. (1999) afirmam, que ao contrário da dependência de elementos frasais dos *bundles* extraídos dos registros de conversações, os *bundles* dos registros acadêmicos são formados com maior freqüência por elementos nominais e preposicionais.

Como neste estudo, o foco não foi em registros de conversações, destacamos os *lexical bundles* de quatro palavras, encontrados nos registros acadêmicos de Biber et al. (1999), para servir de base para a nossa análise, os quais foram classificados em 12 categorias, na seguinte taxinomia ilustrada no quadro a seguir:

[20] Alguns dos exemplos oferecidos por Biber et al. (1999), quanto a sua categoria estrutural, são questionáveis, pois muitos deles não representam grupos nominais.

[21] No original: [...] *When a lexical bundle is structurally complete in academic prose, it is typically a prepositional phrase that functions as a discourse signalling device* […]

Quadro 2. As 12 categorias de *lexical bundles* de registros acadêmicos (extraídas e traduzidas de Biber et al., 1999, pp.1014-1024).

Estrutura	Função	Exemplos
Sintagma nominal com fragmento frasal-*of*	Essa categoria contempla uma série de significados, os quais são utilizados para: (1) descrições físicas, incluindo identificações de lugares; tamanhos e quantidades; (2) existência; (3) qualidades abstratas; (4) processos ou eventos em um dado período de tempo.	*the surface of the; the shape of the; the position of the; the total number of; the presence of the; the nature of the; the course of the.*
Sintagma nominal com outro fragmento pós-modificador	Nessa categoria, a maioria dos *bundles* descreve como um processo ocorre e as relações entre entidades, utilizando especialmente os pós-modificadores *in which*, *between*, e o complemento frasal *that.*	*the way in which; the difference between the; the fact that the.*
Sintagma preposicional com fragmento frasal-*of* incorporado	A maioria dos *bundles* encontrados nessa categoria indica: (1) relações lógicas e abstratas com o uso das preposições *as* e *in*; (2) relações temporais com o uso da preposição *at*; período de tempo ou processos iniciando com a preposição *in*.	*as a result o f; in the case of; at the end of; in the process of.*
Sintagma preposicional sem *of* incorporado	Nessa categoria é freqüente o uso da preposição *in* no início da seqüência dos *bundles* para identificar uma localização específica ou período de tempo. Dois *bundles* são muito comuns nessa categoria: *at the same time* e *on the other hand*, tendo um significado relativamente idiomático, com a função de comparar dois assuntos ou eventos, ou ainda dois argumentos.	*in the present study; in the next chapter; at the same time; on the other hand.*
Antecipatório-*it* + sintagma verbal/ frase adjetiva	Na maioria das vezes ocorrem na forma de voz passiva. Os *bundles* dessa categoria reportam a posição do autor/escritor: possibilidade/probabilidade; importância; necessidade. Em muitos casos com fragmentos *that* identificam informações indubitáveis.	*it is possible to; it is important to; it should be noted; it can be seen that.*
Verbo na passiva + fragmento de sintagma preposicional	A maioria dos *bundles* desse grupo preposicional incorpora voz passiva. Os *bundles* indicam relações de locação ou lógicas, sem o uso de *by*. Esses *bundles* indicam gráficos e tabelas de dados, como também base de algumas afirmações ou achados.	*are shown in table; is shown in figure; is based on the.*
Verbo de ligação-*be* + sintagma nominal/ frase adjetiva	A maioria dos *bundles* nessa categoria utiliza o verbo de ligação *be* ou *may be*. Também, os *bundles* dessa categoria sobrepõem outras categorias com *of* inserido em sua formação. Nessa categoria os *bundles* indicam relações causativas ou comparativas.	*is one of the; is part of the; be the result of; may be due to; is equal to the.*
(Sintagma verbal+) fragmento oracional-*that.*	A maioria dos *bundles* nessa categoria é formado por sentenças verbais seguidas de *that*, como também parte de estruturas mais extensas. Em contraste, os *bundles* que compreendem as sentenças iniciadas por *that* ocorrem em estruturas declarativas não marcadas.	*should be noted that; that there is a; that it is not.*
(Verbo/adjetivo+) fragmento oracional-*to*.	Os *lexical bundles* formados por adjetivos predicativos guiando o fragmento *to* são utilizados para: (1) indicar possibilidade/habilidade; e (2) identificar informações ou conhecimentos prévios.	*is likely to be; is not possible to; has been shown to; was found to be.*
Fragmento adverbial oracional	Os *lexical bundles* nessa categoria, iniciados com o subordinador *as* são utilizados como referência demonstrativa a outros segmentos do discurso.	*as shown in Figure; as we have seen; as we shall see.*
Pronome/ sintagma nominal + *be* (+…)	Nessa categoria os *bundles* tem o verbo de ligação *be* como principal verbo, iniciando com o pronome demonstrativo *this* e o pronome existencial *there*. O pronome *this* é utilizado para conectar informações do discurso prévio, já, o pronome *there* é utilizado para fornecer informações.	*this is not the, there was no significant; there has been a; there are a number.*
Outras expressões	Alguns *bundles* nos registros acadêmicos não se ajustam claramente em nenhuma das categorias.	*as well as the; than that of the, may or may not; the presence or absence.*

No entanto, as categorias de *lexical bundles* proposta por Biber et al. (1999) podem ser passíveis de adaptações e/ou modificações, devido aos traços característicos de cada *corpus* de estudo, e ao tipo de análise empreendida pelo pesquisador. Por exemplo, em alguns trabalhos (Stubbs, 2002; Biber et al., 2004; Scott e Tribble, 2006; Cortes 2006, 2008; Hyland, 2008; entre outros), os pesquisadores desenvolveram estudos a respeito da função de *lexical bundles* em *corpora* de diferentes disciplinas (Hyland, 2008; Cortes 2006, 2008), apoiados nas categorias descritas por Biber et al. (1999), entretanto, com propósitos e metodologias de análises diferentes. Os resultados obtidos dessas pesquisas apresentaram aspectos variantes quanto à função exercida pelos *bundles* nos textos com relação ao trabalho de Biber et al. (1999), e ofereceram contribuições para a aplicação das seqüências ininterruptas de palavras no ensino de línguas em contextos específicos. Para um detalhamento dessas pesquisas, apresentamos na próxima sub-seção as diferentes abordagens com *lexical bundles*.

1.2.2 Diferentes abordagens *com bundles*

Nesta sub-seção, elegemos alguns autores que utilizaram a extração de *lexical bundles* para análises lingüísticas, e contribuíram para alavancar e conduzir alguns dos pressupostos que guiaram a presente pesquisa. Esses autores, diferentemente do propósito desta pesquisa, tiveram como objetivo principal realizar um levantamento de traços característicos do uso da língua por aprendizes e profissionais dentro do contexto acadêmico, implicando em futuras aplicações no ensino-aprendizagem de língua estrangeira. Dessa forma, na seqüência, apresentamos as pesquisas de Hyland (2008), Cortes (2006, 2008), e Scott e Tribble (2006).

Hyland (2008:4) afirma que os *lexical bundles são componentes importantes na produção lingüística e um ponto chave para o sucesso na aprendizagem de línguas*[22]. O autor acredita que essa seqüência de palavras recorrentes auxilia a recortar o significado do texto, contribuindo para identificar as características distintas de registros específicos. Vale ressaltar, que o conceito de *lexical bundles* para o autor não representa somente uma seqüência ininterrupta de palavras, mas também *uma associação psicológica entre as palavras, refletindo a realidade das experiências comunicativas dos usuários* (2008:5) [23]. Isso quer dizer, que nessa relação, os usuários ou aprendizes de uma língua, são capazes de reunir

[22] [...] *multi-word expressions are an important component of fluent linguistic production and a key factor in successful language learning* [...]

[23] [...] appear to represent a psychological association between words and reflect a very real part of users´ communicative experiences […]

um repertório das palavras referentes às práticas das comunidades específicas das quais eles pertencem.

Sob o mesmo ponto de vista, analisando o SMF, podemos dizer que as práticas dos profissionais que atuam na indústria farmacêutica, refletem na escrita do documento, a descrição de suas atividades e experiências adquiridas dentro de uma comunidade específica que lhes é familiar, e, por sua vez, reúnem um repertório comum que favorece uma comunicação adequada e coerente entre os participantes dessa comunidade. Conforme Hyland (2008:5), a presença de um conjunto de palavras (como os *lexical bundles*) no texto pode contribuir na identificação de um registro em específico e, por conseguinte, *ajudam a delinear os significados do texto, contribuindo para o nosso senso de distinção em um registro* [24].

Em sua pesquisa, Hyland (2008) examinou um *corpus* de 3.500 milhões de palavras, compilado a partir de registros acadêmicos escritos de quatro diferentes disciplinas (Engenharia Elétrica, Biologia, Administração de Empresas, e Lingüística Aplicada). O objetivo do autor foi extrair *lexical bundles* de quatro palavras (*four-word bundles*), baseado numa análise contrastiva das formas e funções dos *bundles*. A escolha do autor por *bundles* de quatro palavras fortalece as asserções de Biber et al. (1999) e Berber Sardinha (2003), conforme discutido na seção 1.2, de que *bundles* de três e quatro palavras é um tamanho adequado para uma análise textual, oferecendo *bundles* mais informativos, e uma maior clareza para a análise de suas estruturas e funções.

Entretanto, apesar de Hyland (2008) ter categorizado os seus *bundles* estruturalmente, ou seja, de acordo com os tipos gramatical e funcional das categorias de Biber et al. (ver Quadro 2), o autor propôs algumas modificações na classificação das funções dos seus *bundles*, a fim de melhor representá-los no seu *corpus* de estudo. Essa classificação se deu de forma indutiva, isto é, de forma probabilística, organizada em três categorias funcionais, as quais Hyland (2008) definiu como funções orientadas, a saber: orientada pela pesquisa (*research-oriented*), orientada pelo texto (*text-oriented*), e orientada pelo participante (*participant oriented*).

O estudo de Hyland confirmou, que a maioria dos *lexical bundles*, identificados nos registros acadêmicos, fazia parte de grupos nominais e preposicionais, no entanto, apresentavam variações interessantes entre as disciplinas. Uma dessas variações foi quanto à quantidade de *bundles* diferentes encontrados nos textos de Engenharia Elétrica, o que não foi identificado nas outras três disciplinas. Hyland (2008:9) afirma que essa diferença significa

[24] [...] *helping to shape text meanings and contributing to our sense of distinctiveness in a register* [...]

que pode existir, consideravelmente, uma grande dependência de estruturas pré-fabricadas na disciplina de Engenharia do que em outras áreas. Possivelmente, essa diferença se justifique por ser a Engenharia uma área técnica de natureza gráfica e com vocabulário próprio.

Por exemplo, se tomarmos como exemplo esta pesquisa, podemos verificar que os quinze SMFs analisados, provenientes de diferentes localidades, mas pertencentes a um mesmo registro, apresentaram diferenças significativas, mesmo sendo de uma mesma área técnica (farmacêutica). É evidente que devemos levar em consideração (conforme será discutido no Capítulo 3 deste trabalho) que essas diferenças podem implicar em diversos fatores, dependendo do contexto em que o *corpus* de estudo esteja inserido.

Embora Hyland (2008) tenha direcionado os seus esforços em identificar as funções dos *bundles* que permeiam a escrita acadêmica, por meio das ciências exatas, biológicas e sociais, e por mais que esses *bundles* tenham indicado funções argumentativas diferentes, pelas quais os autores se utilizam para fazer jus à sua credibilidade na persuasão de seus leitores, fica ainda uma pendência para futuras pesquisas: se analisássemos um *corpus* de domínio específico, Engenharia Elétrica, por exemplo, e buscássemos ao invés de *lexical bundles* diferentes, *bundles* semelhantes entre eles, talvez não encontrássemos uma grande dependência de seqüências recorrentes nos textos.

Essa hipótese foi testada nesta pesquisa por meio dos *lexical bundles* semelhantes, e os resultados mostraram, que mesmo os quinze exemplares do SMF pertencerem a uma mesma área técnica (farmacêutica), não houve uma grande dependência de estruturas pré-fabricadas, conforme ocorreu nos achados de Hyland (2008) com relação a área de Engenharia, independente do tamanho do *corpus* de ambas as pesquisas.

Por isso, Hyland (2008:20) recomenda que outros trabalhos de diferentes disciplinas ou de diferentes gêneros e registros sejam realizados, com o objetivo de oferecer uma visão mais ampla das práticas lingüísticas aplicadas a comunidades específicas.

Em Cortes (2008), a autora investiga dois *corpora* comparáveis[25], um de língua inglesa e outro de língua espanhola, extraídos de artigos publicados em periódicos argentinos e norte-americanos sobre a disciplina de História. Cortes analisou a recorrência de *lexical bundles* de quatro palavras (*four-word bundles*) encontrados em um milhão de palavras em cada idioma, num total de dois milhões de palavras. A autora optou pelo ponto de corte estatístico vinte, ou

[25] Segundo Tagnin (2007:161), não há ainda consenso para o uso desses termos. Conforme o projeto da USP, o COMET, *corpora* comparáveis são aqueles formados por originais em línguas distintas, obedecendo ao mesmo recorte, isto é, ao mesmo gênero, tipo de texto e período, e *corpora* paralelos são aqueles formados por originais e suas traduções.

seja, a seqüência ininterrupta deveria recorrer vinte vezes a cada milhão de palavras para ser considerada um *lexical bundle*.

Nesse estudo, Cortes focalizou a sua análise na recorrência da categoria gramatical formada por grupos nominais e preposicionais baseada em Biber et al., 1999, e descobriu que há muitos traços em comuns entre as estruturas de ambos os idiomas. Um desses traços é a recorrência de *bundles* que indicam lugares, como: *Rio de la Plata* e *The United States of* (sintagma nominal com fragmento frasal-*of*), e muitos outros exemplos. Para Cortes (2008), essa pesquisa contribuiu para um melhor entendimento na relação entre as línguas, tanto cultural como funcional, bem como auxiliou nas questões de tradução e versão. Ademais, a autora acredita que os *lexical bundles* podem dizer sobre o que se fala no texto (o tópico textual[26]), e cita alguns dos exemplos encontrados nos *corpora*, por exemplo: *La Revolución del Paraguay*, *The Second World War*, entre outros.

Em um outro estudo de Cortes (2006) sobre *lexical bundles*, a autora apresenta uma pesquisa desenvolvida com alunos matriculados em um curso intensivo de redação, na grade da disciplina de História de uma universidade de médio porte localizada no sudeste dos Estados Unidos. O objetivo dessa pesquisa foi ensinar o uso das combinações recorrentes de palavras em registros escritos, visando à apreensão dessas recorrências por parte dos alunos, em futuros trabalhos acadêmicos. A pesquisa foi conduzida em três etapas: na primeira etapa, a pesquisadora coletou um *corpus* de 800.000 palavras extraídas de periódicos de duas revistas norte-americanas referente à história americana. Segundo Cortes (2006:394), a maioria dos artigos coletado dessas duas revistas faziam parte das leituras do curso, daí a escolha como base de análise. Assim, por meio de um programa computacional, a autora empregou uma busca por *lexical bundles* de quatro palavras (*four-word bundles*), e obteve como resultado 35 *bundles*, os quais ela considerou *bundles*-alvo (*target bundles*). Esses *bundles* também foram classificados funcionalmente no contexto específico, de acordo com a proposta das categorias de Biber et al. (1999).

Em seguida, na segunda etapa da análise de Cortes (2006), foi necessário alinhar com o professor da disciplina de História, quais seriam os períodos para a aplicação do conteúdo das tarefas com *bundles*, ressaltando que a duração desse curso intensivo seria de 10 semanas com 2 encontros semanais. Para atender esse cronograma, foi destinado à pesquisa com *bundles*, 20 minutos de mini-lições (*micro-lessons*). Essas mini-lições seriam distribuídas 5 vezes

[26] Berber Sardinha (2003) sob um outro viés, também analisou as relações semânticas entre os tópicos textuais de um determinado gênero contábil empresarial a partir de *lexical bundles* de três palavras.

durante o semestre. No primeiro encontro, o projeto foi apresentado aos alunos, seguido de uma breve descrição do conceito de *lexical bundles*, e como esses elementos eram utilizados nos artigos de História publicados nos periódicos. Logo, no segundo, terceiro e quarto encontros, os alunos trabalharam em exercícios elaborados e aplicados pelo pesquisador. Já no último encontro, foi feita uma discussão em grupo sobre o uso dos *lexical bundles*. Nesse último encontro, os alunos realizaram atividades para inferir sobre a função que os *bundles* encontrados nos artigos exerciam nos textos produzidos pelos autores dos periódicos.

Enfim, na terceira etapa, foram coletadas amostras de artigos produzidos pelos alunos participantes da pesquisa, e na seqüência, cada produção textual dos alunos foi processada via um programa concordanciador, com o propósito de identificar quais palavras ou expressões poderiam estar relacionadas aos 35 *bundles*-alvo definidos *a priori.*

Como resultado, o estudo de Cortes (2006) reportou que não houve progresso dos alunos quanto ao uso dos *bundles* após o processo de aprendizagem, ou melhor, da aplicação das mini-lições. Durante o processamento dos dados, um dos resultados mostrou que apenas um único *bundle* foi utilizado somente uma vez por um dos alunos participantes. A partir desses resultados, a autora concluiu que possíveis razões referentes à baixa freqüência de uso dos *bundles* pelos alunos, poderiam estar relacionadas à curta duração do curso, ou porque as atividades não eram adequadas para despertar a autonomia dos alunos quanto ao uso dos *bundles*.

Destarte, pressupomos que esses resultados vêm ao encontro da qualidade das produções escritas dos alunos, ou mesmo do pouco conhecimento da área. De acordo com Hyland (2008:5), com relação à identificação de determinados textos por meio de *lexical bundles*, o autor afirma:

> [...] Esses *bundles* são familiares a escritores e leitores que regularmente partilham um discurso comum, sua própria 'naturalidade' sinaliza uma participação apropriada numa dada comunidade. Ao contrário, a ausência de tais *bundles* pode mostrar a falta de fluência de um aprendiz ou de um recém-chegado a essa comunidade [...][27]

[27] No original: [...] *These bundles are familiar to writers and readers who regularly participate in a particular discourse, their very 'naturalness' signalling competent participation in a given community. Conversely, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community* […]

Refletindo a respeito da pesquisa de Cortes (2006), inferimos o seguinte, que se a freqüência de seqüências de palavras pode ser comprovada em estudos com base em *corpora* de registros acadêmicos profissionais, a lacuna de *bundles* equivalentes em *corpora* de aprendizes pode ser um primeiro indício no diagnóstico da carência de repertório dos alunos, dentro das práticas discursivas de sua comunidade específica.

Um outro estudo que investiga as seqüências de palavras recorrentes em produções escritas acadêmicas entre profissionais e aprendizes é o trabalho de Scott e Tribble (2006). Lembrando que esses autores preferem a denominação *cluster* à *lexical bundles*. Scott e Tribble (2006) coletaram um *corpus* de dissertações de mestrado desenvolvidas por alunos do programa de Filologia do departamento de língua inglesa da Universidade de Poznań na Polônia. Para essa pesquisa, os autores trabalharam com um *corpu*s de estudo de 352.258 palavras, extraído de dissertações sobre literatura, e um *corpus* de referência, o *British National Corpus World Edition* (doravante BNC) [28], denominado pelos autores BNC_ALL, com aproximadamente 100 milhões de palavras [29].

O objetivo desse estudo foi o de descobrir meios de descrever o uso da língua inglesa por aprendizes poloneses com alto nível de proficiência em registros acadêmicos. Outrossim, Scott e Tribble (2006) atribuíram a esses aprendizes a definição de especialistas em língua estrangeira (*foreign language experts*). Segundo os autores, esse tipo de análise justificou-se pelo fato de que o uso sistemático de amostras de produções escritas, desenvolvidas por usuários proficientes de uma segunda língua, pode servir de apoio à escrita de alunos iniciantes, bem como oferecer informações para a prática no ensino da escrita acadêmica.

Ademais, Scott e Tribble (2006:131-132) afirmam que os estudos com *lexical bundles* contribuem para diversos fins, como, por exemplo, oferecem:

- meios de diferenciar textos dentre diferentes *corpora*;
- conhecimentos de aspectos fraseológicos importantes em contextos específicos, especialmente aos pesquisadores, professores e aprendizes de línguas estrangeiras;
- informações para o entendimento de como os aprendizes constroem os seus textos, e qual o nível de identificação das diferenças e semelhanças esses

[28] Ver seção 1.1 deste capítulo para detalhes sobre a composição desse banco de dados da língua inglesa.
[29] Esse número de palavras, segundo Scott & Tribble (2006:133) inclui os *sub-corpus* distribuídos no BNC, isto é, 500.000 palavras de artigos sobre estudos de literatura (BNC_LIT), 3 milhões de palavras entre conversação (BNC_CONV), escrita acadêmica e ficção (BNC_ACAD).

aprendizes atingem com relação aos textos de especialistas ou de profissionais acadêmicos;

- possibilidades de mostrar o que funciona em tipos particulares de textos e o que é melhor aceito por leitores experientes em contextos específicos.

Scott e Tribble (2006) empreenderam sua pesquisa fazendo uso de mais de um instrumento de análise, ou seja, em complemento à extração dos *clusters*, os autores interagiram lista de palavras (*word-list*) e palavras chaves (*key-words*), recursos de análise lingüística oferecidos por programas computacionais. Para isso, os autores utilizaram o aplicativo *WordSmith Tools*, desenvolvido por Mike Scott (1996)[30], aplicando duas etapas no percurso de análise dos *clusters*. Na primeira etapa, com o auxilio da ferramenta WSTools, foi processada uma lista de palavras (*word-list*) dos 20 mais freqüentes *clusters* extraídos do *corpus* de referência BNC_ALL e do *sub-corpus* BNC_ACAD (somente os registros acadêmicos). Os resultados nessa fase mostraram que os 20 mais freqüentes *clusters* no banco de dados do BNC_ALL e do BNC_ACAD foram muito semelhantes, embora alguns casos de diferenças tenham ocorridos devido aos pronomes de primeira, segunda ou terceira pessoa. Em seguida, foram extraídos *clusters* de duas, três e quatro palavras.

Na extração dos *clusters* de duas palavras, Scott & Tribble (2006) consideraram os 20 mais recorrentes, e nos *clusters* de três e quatro palavras os 40 mais recorrentes, contrastando esses achados entre os bancos do BNC_ALL e o BNC_ACAD. A distinção entre *clusters* de três e quatro palavras foi uma das variações que Scott & Tribbe (2006) consideraram importantes, pois, conforme as asserções de Biber et al. (1999), a escolha por seqüência de quatro palavras são mais adequadas, ou seja, apresentam uma maior clareza do que as seqüências de duas ou três palavras, motivo pelo qual Biber et al. (1999) optaram pela extração de *bundles* de quatro palavras nos registros acadêmicos.

No entanto, neste estudo, não optamos pela seqüência de quatro palavras porque o tamanho do *corpus* de estudo comprometeria a quantidade de extração dos conjuntos de palavras, ou seja, se estendêssemos a seqüência de três palavras para quatro, a quantidade diminuiria automaticamente e, assim, teríamos um volume de dados menor para a análise. Um outro motivo pela escolha de seqüências de três palavras nesta pesquisa, foi a possibilidade dos resultados implicarem em estudos sobre terminologia específicas e em estudos de línguas para fins específicos. Assim, hipotetizamos que um conjunto de três palavras possa ser

[30] Ver detalhes sobre ferramentas computacionais na seção 1.1 deste capítulo.

apropriado para tais propósitos, em particular, para a área pedagógica, por facilitar uma melhor apreensão por parte dos alunos, questões levantadas no parágrafo seguinte por Scott e Tribble (2006).

Scott e Tribbe (2006), comparando seus achados com a categoria estrutural de Biber et al. (1999), descobriram que os *bundles* formados de sintagmas nominais com fragmentos frasais-*of* (*noun phrase with of-phrase fragment*), extraídos do BNC_ALL e comparados ao BNC_ACAD, apresentaram algumas freqüências que deveriam ser levadas em consideração, como, por exemplo, o *bundle* de quatro palavras *one of the most*. Embora na maioria dos registros acadêmicos, o potencial combinatório dessa seqüência seja a palavra *most*, nos resultados de Scott e Tribble (2006) ocorreram seqüências do tipo: *one of the main.* Dessa forma, pedagogicamente falando, os autores consideraram importante o fato de se investigar uma extensa coleta de textos acadêmicos para a extração de *bundles* de três palavras do tipo: *one of the*, facilitando ainda mais a apreensão do conjunto de informações pertencentes ao discurso acadêmico.

Com isso, diante desses resultados, Scott e Tribble (2006:141) questionam:

> [...] quais dados serão mais úteis aos alunos: uma análise limitada a elementos frasais robustos, ou uma análise que inclua associações de colocações estendidas, das quais possamos oferecer meios pelos quais os alunos descubram extensões de combinações menos freqüentes, mas não menos importantes [...].[31]

Na segunda etapa, Scott e Tribble (2006) analisaram os textos dos aprendizes, o *corpus* das dissertações de mestrado. O principal objetivo nessa etapa foi o de comparar a escrita dos alunos com a escrita dos artigos acadêmicos publicados por profissionais, com o intuito de buscar as equivalências ou diferenças dentro de um mesmo campo de atuação. Conforme os autores, essa etapa também possibilitou descobrir qual o valor pedagógico que esse tipo de descrição pode abarcar. Isso posto, foi feita uma comparação entre o *corpus* da escrita dos profissionais (BNC_ALL com o BNC_ACAD) e a escrita dos alunos (dissertações). Logo após, foi feita uma nova comparação entre a escrita dos alunos e os *sub-corpus* de registros acadêmicos (BNC_ACAD) e de estudos literários (BNC_LIT). Nessas fases, os autores optaram em trabalhar na extração dos 40 mais freqüentes *clusters* de três palavras, por

[31] No original: [...] *which data is going to be more useful to learners: an analysis limited to strongly phrasal elements, or one which includes "extended collocational associations" and which we feel offer learners a means of discovering a range of less frequent, but no less valuable phrasal combinations* […]

acreditarem que no contraste com contextos específicos de produção escrita esse tamanho apresentaria variações, caso os *bundles* se estendessem para a seqüência de quatro palavras, como discutimos anteriormente.

Em conclusão, Scott e Tribble (2006:157) avaliaram que essa análise poderia ser parte de algumas questões pedagógicas, mas não oferecia meios de avaliar claramente o desempenho dos aprendizes, nem mesmo dos profissionais acadêmicos nos textos. No entanto, o estudo permitiu verificar algumas diferenças que indicaram que na escrita dos aprendizes, muito pouco da sofisticação encontrada na escrita dos profissionais acadêmicos foi aplicada. Com base nas funções dos *clusters*, os autores ofereceram dois exemplos dessa lacuna que foram: (1) uma baixa freqüência no uso da estrutura *it*-antecipatório + sintagma verbal/ frase adjetival (*anticipatory it + verb phrase/adjective phrase*) em *clusters* de quatro palavras, indicando um nível avaliativo menor nas dissertações, ou indicando que a avaliação foi feita apropriadamente, mas de forma diferente; (2) uma orientação descritiva aparente nos *cluster* de três palavras, inseridos na estrutura sintagma nominal com fragmento frasal-*of* (*noun phrase with of phrase fragment*), indicando ou não uma falta de sofisticação na análise.

Por fim, Scott e Tribble (2006) acreditam que pesquisas com seqüências recorrentes de palavras comparáveis entre textos, podem contribuir para que os escritores avaliem o próprio desempenho com o desempenho daqueles que eles desejam alcançar.

Embora o *corpus* utilizado neste trabalho seja dirigido pelo contexto profissional do segmento farmacêutico, as pesquisas com *corpora* compilados de textos acadêmicos, conduzidas pelos autores acima citados, serviram de embasamento prático-teórico para a fundamentação desta pesquisa, a fim de:

- evidenciar aspectos, de certa forma fraseológicos, do contexto de produção do documento SMF (o segmento farmacêutico);
- oferecer meios para o entendimento dos textos escritos por diferentes autores de uma mesma comunidade específica (as quinze diferentes localidades da empresa);
- contribuir na escolha dos *bundles* de três palavras como uma alternativa pedagógica no ensino de línguas para fins específicos.

Portanto, apresentadas as pesquisas com *lexical bundles*, e baseados na afirmação de Cortes (2006), de que os *bundles* podem revelar sobre o que se fala o texto, consideramos a

possibilidade de uma relação coesiva dos *lexical bundles* com o tópico das nove seções do documento SMF, tendo como suporte as teses de doutorado de Ramos (1997) e Berber Sardinha (1997), que apontam para as teorias desenvolvidas nesse sentido, e que são apresentadas na próxima seção.

1.3 O tópico textual

Ramos (1997:77) afirma que apesar das controvérsias a respeito da noção de tópico e das diferentes denominações empregadas ao termo, "tópico" pode ser definido *como aquilo de que o texto trata.* A autora apóia suas asserções em algumas pesquisas como as de Brown e Yule (1983) e Phillips (1989) para discorrer sobre o assunto. Ramos (1997:77) referindo-se às concepções de Brown e Yule (1983), diz que os autores caracterizam o "tópico" em níveis mais altos com relação à hierarquia do conteúdo do discurso, ou seja, *a noção de tópico não é dependente do texto mais sim do escritor ou do falante*. Por outro lado, segundo Ramos (1997:77), outros trabalhos dedicam-se ao estudo do tópico a partir de levantamento lexical. Nas palavras da autora, *esses estudos têm uma preocupação maior em ver quais palavras são usadas juntas, ao invés de como as palavras são usadas*.

Já se referindo a Phillips (1989), Ramos (1997:77) explica que o autor não utiliza o termo tópico, mas sim "*subject matter*" ou "*aboutness*", porque, segundo Ramos, Phillips acredita que o termo "*aboutness*" parece ser um fenômeno de larga escala e não apenas expressões lingüísticas. Na interpretação de Ramos (1997:78):

> [...] para esse autor, a percepção de que o texto é sobre alguma coisa vem da apreciação do leitor dessa organização textual em larga escala, ou seja, as redes lexicais de colocação e intercolocação que levam a um padrão geral de conexão entre os capítulos possibilitam a identificação da macroestrutura lexical do texto, e é essa macroestrutura que possibilita o leitor entender sobre o que é o texto [...].

Quanto às questões referentes às relações lexicais na organização textual, Ramos (1997:80) comenta que muitos trabalhos têm sido desenvolvidos e debatidos na área de coesão textual com o propósito de se avaliar a importância do léxico. Um dos principais propósitos para tais estudos é a verificação das cadeias lexicais que se formam na construção do texto por meio de uma seqüência de eventos (sentenciais ou agrupamento de palavras),

relacionando-se ao texto e ao contexto específico de forma consistente e coerente. Segundo a autora, o trabalho de Hoey (1991) é um dos que melhor reflete esses princípios, norteados pelos aspectos de repetição no texto, seja por repetição de significado, ou pela repetição da mesma palavra ou conjunto de palavras.

Atentamos ao fato de que em seu trabalho, Ramos (1997) analisa as relações coesivas não seqüenciais para descobrir o assunto (tópico) do texto, e nesta pesquisa analisa-se as relações coesivas dos *lexical bundles* (seqüência de três palavras) diretamente com o tópico prescrito nos títulos e subtítulos de cada uma das nove seções do documento, visto que as seções e sub-seções que compõem o SMF remetem às informações principais exigidas na escrita do documento. Com isso, no levantamento das recorrências dos *lexical bundles* semelhantes encontrados nos textos do SMF, foi possível identificar os elos coesivos entre as nove seções dos quinze diferentes documentos, isto é, determinados *lexical bundles*, em sua formação, possuíam elementos lingüísticos que, ora repetiam-se no título de cada seção, ora reiteravam o assunto tratado nas seções e sub-seções do SMF, revelando, por meio de uma cadeia coesiva, sobre o que tratavam os textos do documento *Site Master File*.

Para um detalhamento desses conceitos de repetição nas relações coesivas do texto, na sub-seção seguinte contemplamos uma discussão a respeito do assunto, ainda sob o enfoque das contribuições de Berber Sardinha (1997) e Ramos (1997).

1.3.1 Repetições no texto como relações coesivas

Berber Sardinha (1997:129) se baseou no trabalho dos lingüistas Halliday e Hasan (1976) para ilustrar a definição de coesão lexical, dos quais o autor faz uso da seguinte citação: *coesão lexical é a recorrência do mesmo item lexical ou a seleção de dois itens que estão em relação de proximidade*[32]. Segundo Berber Sardinha, essa definição vem ao encontro dos estudos baseados em *corpora* com o uso de instrumentos computacionais, que exploram a freqüência e recorrência de palavras por meio de análises estatísticas dos dados, ainda que Halliday e Hasan não tenham se apropriado de tais métodos analíticos quantitativos na formulação de suas teorias. De acordo com Berber Sardinha (1997:130), Halliday e Hasan determinaram dois tipos de coesão lexical: a *reiteração*, quando há ocorrência de um item

[32] No original: [...] *They define lexical cohesion as ´selecting the same lexical item twice, or selecting two that are closely related´* […].

idêntico ou de uma palavra que se relaciona ao item, e a *colocação*, quando há uma associação de itens lexicais que co-ocorrem regularmente.

Ramos (1997), aponta para o trabalho de Hoey (1991) a respeito das relações lexicais por repetição. Segundo Ramos (1997:84), Hoey (1991) acredita que repetição deve ser entendida *no contexto do significado de um item por outro item, seja a palavra a mesma ou não.*

Conforme Berber Sardinha (1997:131), quanto ao texto as ligações coesivas *são percebidas através das sentenças, pois essa é a única fonte de textura através das sentenças, visto que as sentenças são estruturalmente independente uma das outras*[33]. Além do mais, Berber Sardinha (1997) acredita que estudar coesão lexical por meio das repetições no texto, favorece e também permiti expandir os horizontes das investigações lingüísticas com o uso de ferramentas computacionais, já que, segundo o autor, *os computadores podem ser programados precisamente para identificar repetições, mas não podem ser facilmente programados para identificar outros tipos de coesão lexical*[34] (:132).

Todavia, apesar das discussões em torno das sentenças, acreditamos ser possível realizar estudos das cadeias coesivas, como o fez Ramos (1997) em seu trabalho sobre as relações entre textos baseadas nas escolhas lexicais. Para esse tipo de estudo, Ramos (1997) contou com o apoio teórico de Eggins (1994) e Martin (1992). Conforme resenhado por Ramos (1997:90-91), Eggins (1994) afirma que:

> [...] a análise de relações lexicais é um meio de descrever sistematicamente como as palavras num texto se relacionam uma às outras, como elas se juntam para construir conjuntos lexicais ou cadeias lexicais. A descrição dessas cadeias lexicais num texto pode, pois, nos dizer: (a) sobre o que se fala no texto: cada cadeia lexical principal indica um tópico ou parte de um tópico; (b) quantas coisas diferentes são faladas no texto, i.e., saber quantas cadeias lexicais ocorrem no texto, se elas são ligadas ou não, se são simultâneas ou seqüenciais; (c) que tipos de relações lexicais operam entre itens em cadeias, i.e., que tipos de taxonomias são desenvolvidos dentro de um texto [...].

[33] No original: [...] *Cohesive links are perceived across sentences because this is the only source of texture across sentences, given that sentences are structurally independent of each other* [...].

[34] No original: [...] *since computers can be programmed to reliably identify repetition but they cannot be easily made to identify other types of lexical cohesion* [...].

Nessa perspectiva, Berber Sardinha (1997) traz em seu trabalho a abordagem de Hoey (1991), embasado na noção de que a coesão lexical forma conjunto de palavras entre as sentenças. Essa noção foi base para o estudo de Berber Sardinha (1997), o qual investigou a coesão lexical de um grande número de textos, utilizando uma estratégia de segmentação das sentenças dos textos (por meio de recursos computacionais), com o intuito de mostrar como os elos coesivos são capazes de organizar o texto. Berber Sardinha (1997:150) explica que, segundo a proposta de Hoey, um dos principais critérios aplicados aos estudos sobre coesão lexical está na hipótese de que a *função principal da coesão lexical é repetir*. Desse modo, segundo Berber Sardinha, Hoey propõe duas noções chaves para uma análise de coesão lexical: os *links* (elos) e os *bonds* (encadeamentos)[35]. Em um breve resumo, Berber Sardinha (1997:155-160) expõe que os *links* ocorrem quando há repetição de um item em duas sentenças separadas, e os *bonds* quando há conexões entre sentenças. Normalmente, segundo Berber Sardinha, Hoey considera que são necessários três *links* para constituirmos um *bond.*

Posto que esta pesquisa não tenha tido como propósito realizar uma análise coesiva por meio das sentenças dos textos, como o fez Berber Sardinha (1997), e sim investigar a relação dos *lexical bundles* semelhantes com os títulos, ou seja, os tópicos do documento SMF, interessa-nos as asserções referentes à repetição dos elementos lingüísticos no texto, segundo as noções dos autores apontados nesta seção. Assim sendo, na sub-seção seguinte, damos seqüência aos comentários extraídos dos trabalhos de Berber Sardinha (1997) e Ramos (1997) com relação aos tipos de coesão.

1.3.2 Tipos de repetição e coesão lexical

Para Beber Sardinha (1997:155-156), os principais tipos coesivos defendidos por Hoey (1991) são: (i) a repetição simples; (ii) a repetição complexa; (iii) a paráfrase simples; (iv) a paráfrase complexa; e (v) a repetição como superordenados e hipônimos.

Ramos (1997:85-87) explica esses tipos de repetições da seguinte maneira: (i) a repetição simples ocorre quando um item é repetido sem alterações significantes, como um *paradigma gramatical fechado* (por exemplo: gato - gatos); (ii) a repetição complexa ocorre quando dois itens compartilham um mesmo morfema lexical, mas não possuem formas gramaticais idênticas (por exemplo: processo-substantivo, processar-verbo), ou ainda, no caso dos nossos dados de pesquisa, *description-described*; (iii) a paráfrase simples ocorre com

[35] Tradução dos termos feita pela pesquisadora.

itens que podem ser paráfrases em determinados contextos numa relação de sinonímia (por exemplo: estadista-político); (iv) a paráfrase complexa ocorre entre dois itens que compartilham ou não o mesmo morfema lexical ou gramatical, mas devem estar restritos a três características: quando antônimos (por exemplo: frio-quente); quando antônimo de um terceiro (por exemplo: escritor - autor); quando *há ausência de um mediador*, isto é, um item pode ser substituído por outro item sem comprometer a interpretação do texto (por exemplo: ensino-instrução); e (v) a repetição como superordenados e hipônimos ocorre quando dois itens possuem o mesmo referente mas não relacionam-se como itens lexicais (por exemplo: Augustus - o imperador). Na presente pesquisa, o que podemos abstrair *a priori* é a repetição simples, pois em nossos achados obtivemos *lexical bundles* que se repetem igualmente, sem alterações, nos tópicos das seções (ver capítulo Metodologia).

Quanto à reiteração, Ramos (1997:82) explica que essa se faz por repetição *de um mesmo item lexical, um sinônimo ou quase-sinônimo, um superordenado ou um nome genérico*. Segundo a autora, um dos trabalhos que levantam essas questões é o de Hasan e Halliday (1989). Para Ramos (1997:83), esses autores abordam três tipos de relações de significados estabelecidas com o texto, os quais estão resumidos abaixo:

(i) a co-referencialidade: uma relação situacional de identidade entre os membros (por meio de recursos de referência);
(ii) a co-classificação: uma relação na qual coisas, processos ou circunstâncias referidas entre os membros, pertencem a uma mesma classe idêntica (por meio de substituição e elipse);
(iii) co-extensão: uma relação onde os membros apresentam semelhanças gerais, isto é, dentro *de um mesmo campo geral de significado* (por meio de palavras de conteúdo).

Referindo-se a um outro trabalho de Halliday e Hasan (1976), Ramos (1997) apresenta uma discussão a respeito da noção de cadeia, defendida pelos autores. Ramos (1997) diz que a noção de cadeia, segundo Halliday e Hasan, é *um encadeamento de termos coesivos, ou seja, refere-se ao elemento que vem antes e depois e assim sucessivamente* (:88). Além do mais, segundo Halliday & Hasan (1989 *apud* Ramos, 1997:88): *uma cadeia é formada por um conjunto de relações semânticas de co-referência, co-classificação, e/ou co-extensão*, podendo ser subdivididas em dois tipos: cadeias de identidade e cadeias de similaridade.

Assim, utilizando os pressupostos teóricos dos autores citados (Halliday e Hasan, 1976, 1989; Hoey, 1981; Berber Sardinha, 1997; e Ramos, 1997) pudemos estabelecer as relações de coesividade entre os *bundles* e o tópico das seções do documento SMF. Conquanto, os critérios de análise deste trabalho não possuam um foco em classes gramaticais, tal qual Ramos (1997), que concentrou sua análise na classe dos substantivos, com o objetivo de verificar como as escolhas lexicais sinalizam os tópicos do texto e levam *a indicar um possível gênero*, podemos dizer, que os *lexical bundles*, identificados no *corpus* desta pesquisa, podem fornecer subsídios ao(s) escritor(es) construir(em) o seu texto de forma coesiva. Esses subsídios oferecem ao escritor meios de identificar em qual seção ou sub-seção ocorre determinado conjunto de palavras. Por conseqüência, as relações coesivas que se estabelecem entre esses conjuntos de palavras (os *lexical bundles*) e os títulos das seções, garantem a esses escritores o reconhecimento das partes que compõem o documento.

Dessa forma, encerramos este capítulo, o qual forneceu suporte à fundamentação deste trabalho, e em seguida apresentamos o capítulo da metodologia de pesquisa.

CAPÍTULO 2

METODOLOGIA

Neste capítulo, apresentamos a metodologia de pesquisa utilizada neste estudo. No início, retomamos os objetivos e perguntas de pesquisa, seguidos da descrição do *corpus* de estudo e o detalhamento dos dados de análise. Finalizando, apresentamos a escolha da metodologia, os procedimentos de análise, e as ferramentas computacionais utilizadas no processamento dos dados.

2.1 Objetivos e perguntas de pesquisa

Conforme mencionado no capítulo introdutório deste trabalho, o objetivo desta pesquisa é examinar um tipo de documento do setor farmacêutico, denominado *Site Master File* (SMF), a partir dos *lexical bundles* recorrentes entre quinze exemplares, bem como verificar a relação de significado com os assuntos tratados nos textos. Para isso, por meio dos *lexical bundles* encontrados, observamos o que há de comum nas ocorrências lingüísticas entre as localidades em que o SMF circula.

Uma vez que o documento *Site Master File*, de um lado, exige da indústria farmacêutica, profissionais especializados e capacitados a produzir os textos de acordo com as atividades e práticas estabelecidas pela empresa, e de outro, possibilita a certificação internacional junto aos órgãos governamentais de vigilância sanitária, podemos pressupor que o documento em questão implica uma análise descritiva que nos conduza à identificação de alguns aspectos fundamentais que contemplem, segundo Swales (1990:45-47): (i) a comunidade discursiva que compartilha determinadas posições e atividades sociais; (ii) os propósitos comunicativos de eventos específicos; e (iii) a organização retórica dos textos.

Tendo em vista essa perspectiva, as seguintes perguntas foram formuladas para atingir os objetivos desta pesquisa:

(1) Quais são as semelhanças da estrutura organizacional do *Site Master File* entre as quinze localidades em que o documento circula?

(2) Quais são os *lexical bundles* semelhantes entre as seções dos quinze documentos SMF?

(3) Como os *lexical bundles* semelhantes refletem o tópico das seções do documento SMF?

Apresentadas as três perguntas de pesquisa, na seção seguinte, elencamos a descrição geral do *corpus* e as dimensões contextuais que envolvem o presente trabalho.

2.2 Descrição e coleta do *corpus* de estudo

O documento *Site Master File* (SMF) é um conjunto de textos produzidos pelas indústrias do setor farmacêutico que circula em todas as unidades de negócios da empresa espalhadas pelo mundo, e cujos objetivos principais são: (i) atender às exigências de garantia e qualidade dos medicamentos e produtos fabricados e/ou importados; e (ii) obter a certificação internacional junto aos órgãos governamentais que regulamentam a vigilância sanitária de cada país.

O SMF tem sua origem no GMP (*Guide to Good Manufacturing Practices*), em português Guia de Boas Práticas de Fabricação (BPF). Um guia instituído em 1992 pela Comunidade Econômica Européia (CEE), o qual estabelece os princípios das boas práticas de fabricação dos medicamentos humanos e veterinários. Esse Guia é um conjunto de normas reconhecidas e regulamentadas pelos órgãos de saúde pública mundial, com a finalidade de certificar a qualidade dos produtos farmacêuticos. As regras estabelecem os procedimentos e as práticas que visam a uma padronização nos métodos de fabricação quanto: às condições de instalações de uma empresa; aos equipamentos e respectivas manutenções; aos critérios de segurança; às matérias-primas, embalagens, equipamentos; às condições de estocagem; e aspectos relacionados ao meio ambiente.

Esse regulamento garante a qualidade e a segurança dos produtos, visando à proteção da saúde da população. Entretanto, cada país pode, de acordo com os seus recursos, adequar os seus procedimentos para dar cumprimento às exigências das BPF ou GMP. O objetivo principal das BPF ou GMP é diminuir os riscos de toda produção farmacêutica, tais como: contaminação

cruzada; contaminação por partículas; troca ou mistura de produto; rotulagem incorreta; e demais riscos.

O documento é produzido com a finalidade de atender à legislação sanitária local estabelecida pelas agências reguladoras de cada país. Nos Estados Unidos temos o FDA (*Food and Drug Administration*), na Europa, o EMEA (*European Medicines Agency*), e no Brasil, a ANVISA (Agência Nacional de Vigilância Sanitária). Todos esses procedimentos devem, obrigatoriamente, obedecer à padronização instituída pelas BPF (Boas Práticas de Fabricação) ou GMP (*Good Manufacturing Practices*).

Cada empresa do segmento farmacêutico é responsável pela produção do documento. Os funcionários do Departamento da Garantia da Qualidade (em inglês *Quality Assurance Department*) são os que, normalmente, elaboram e revisam todo o conteúdo. Alguns dos textos que fazem parte das seções do documento são fornecidos pelas áreas específicas da empresa, como, por exemplo, as seções: 3 *Premises and Equipment* e 5 *Production* (comentadas adiante), pelo fato de exigir um conhecimento específico do pessoal qualificado das áreas de Engenharia Industrial e de Processos. Ademais, o departamento da Garantia da Qualidade ou *Quality Assurance Department* (QA) deve incorporar as BPF (Boas Práticas de Fabricação) e assegurar que os produtos sejam consistentemente produzidos e controlados, segundo procedimentos adequados e padronizados.

Originariamente, o documento SMF é produzido no idioma local em que a empresa está situada, e, posteriormente, o documento é traduzido oficialmente para a língua inglesa. O trabalho de tradução e versão dos documentos é executado por tradutores especializados na área farmacêutica, contratados e homologados por cada unidade de negócio. A versão do documento para o idioma inglês deve-se ao esquema de certificação da qualidade de produtos farmacêuticos como objeto de comércio internacional, acordado por órgãos de proteção à saúde do consumidor, os quais instituem as resoluções legais, como o fim de monitorar e regulamentar as operações farmacêuticas no mundo.

O departamento de Documentação da empresa é responsável pelo controle de todos os registros e documentos referentes aos processos de produção farmacêutica da localidade. Também, organiza, distribui e providencia o arquivo de todos os registros e formulários específicos que competem a cada área. Dentre esses documentos está o *Site Master File*, o qual deve ter as suas respectivas versões e atualizações em conformidade com as exigências locais.

As indústrias farmacêuticas devem seguir o mesmo padrão na elaboração do documento SMF, ou seja, todas as seções que constituem o documento devem conter as informações exigidas pelas autoridades competentes e cumprir, rigorosamente, com as práticas estabelecidas para a fabricação de medicamentos. O documento deve ser produzido uma única vez, quando da instalação da fábrica e início das operações de produção.

Caso haja alguma modificação nas instalações da fábrica, nos equipamentos ou na produção de novos fármacos, a empresa deve gerar uma nova versão do documento em substituição a anterior, especificando no cabeçalho a data de substituição e o número seqüencial da edição. Além disso, devem constar no documento todas as assinaturas das pessoas-chave (*key-person*) designadas como responsáveis pela empresa.

O SMF é veiculado somente entre as áreas envolvidas nas operações da empresa, isto é, entre os departamentos responsáveis pelo trâmite das informações inerentes à certificação da qualidade de fabricação. O arquivamento do SMF é feito por meio eletrônico e físico, em tempo determinado pelas diretrizes e políticas operacionais da empresa.

Conforme os procedimentos padrões dos manuais da qualidade e das auditorias de inspeções nacionais e internacionais, estabelecidos para as indústrias farmacêuticas, o documento deve possuir no máximo entre 25 e 35 páginas. Embora tenhamos identificado, nos quinze documentos analisados, que em algumas localidades[36] esse número de páginas excedeu e em outras obedeceu à média estabelecida. Essa diferença encontrada no número de páginas ocorreu devido à escolha na formatação do texto feita por cada localidade, decorrente da quantidade de gráficos, figuras e organogramas encontrados em cada um dos exemplares (ver Quadro 4).

Quanto às informações técnicas, o SMF descreve aspectos referentes ao Sistema de Qualidade (*Quality System*), às Operações de Produção Farmacêutica (*Pharmaceutical Operations*), e às condições das instalações onde ocorre a fabricação. Ademais, está contemplada no documento, a descrição dos prédios e das atividades realizadas ao redor da fábrica, ou seja, outras empresas que estejam alocadas próximas à fabricação dos produtos farmacêuticos.

Desse modo, conforme os procedimentos padrões estabelecidos para a elaboração do SMF, o documento deve, obrigatoriamente, conter nove seções que devem estar sequencialmente numeradas, abordando todas as informações exigidas para a certificação de qualidade de

[36] Neste trabalho, utilizamos a palavra "localidade" para nos referir: (i) a fábrica onde a empresa possui uma unidade de negócio operacional farmacêutico (BU *Business Unit*); e (ii) a cada um dos quinze documentos analisados.

fabricação. Assim, os assuntos do documento SMF são divididos em seções, conforme descritas no quadro abaixo.

Quadro 3. A divisão das nove seções do documento *Site Master File*.

Número da seção	Nome da seção
1	*General Information*
2	*Personnel*
3	*Premises and Equipment*
4	*Documentation*
5	*Production*
6	*Quality Control*
7	*Contract Manufacturing and Analysis*
8	*Distribution, Complaints and Product Recall*
9	*Self Inspection*

O *corpus* deste trabalho foi formado pela compilação de quinze exemplares do documento SMF que pertencem a uma única empresa. No entanto, por tratar-se de documentação exclusiva da empresa, isto é, com acesso restrito, obtivemos somente a quantidade disponível para consulta, limitando-nos a esses quinze exemplares. Todos esses documentos são redigidos, oficialmente, em língua inglesa, e circulam em uma multinacional do segmento farmacêutico com sede na Europa, e com unidades instaladas em mais de 100 países, tendo os seus principais negócios voltados: à saúde do consumidor (*Consumer Health*); aos medicamentos genéricos (*Generic Drug*); aos cuidados dos olhos (*Eye-Care*); e à saúde animal (*Animal Health*).

A seleção dos quinze documentos para compor o *corpus* de estudo desta pesquisa baseou-se nos documentos SMF, os quais foram produzidos pelas unidades operacionais farmacêuticas, distribuídas entre os países da Europa, Ásia, América do Norte e América do Sul. Os exemplares utilizados neste estudo são datados do período de 2004 a 2006, e a coleta dos dados foi realizada no ano de 2007.

A fonte de coleta desta pesquisa foi extraída de uma das unidades de negócios operacionais da empresa instalada no Brasil. Tal empresa estabeleceu algumas restrições quanto à divulgação do conteúdo na íntegra dos documentos, por tratar-se de informações que dizem respeito a sigilos

comerciais. Essas restrições não comprometeram, tampouco foram relevantes para esta análise lingüística.

Dessa forma, para assegurar essas restrições, por meio de ferramentas computacionais, os textos foram tratados de maneira a ser feita uma limpeza nas informações que deveriam ser substituídas ou omitidas. Essas substituições ocorreram principalmente no nome da empresa e nos nomes das pessoas responsáveis. O nome da empresa recebeu o pseudônimo de *PhamaCo*, e as pessoas responsáveis receberam uma numeração, conforme a quantidade de funcionários envolvidos nos processos de cada localidade da empresa.

Dos quinze documentos analisados, dois são provenientes de localidades que possuem o idioma inglês como língua nativa, os quais são: o SMF *England* e o SMF *USA*. Os outros treze documentos são provenientes de localidades da Europa, Ásia e Américas, os quais utilizam a língua inglesa como meio de comunicação no contexto de negócios. Essas localidades são as seguintes:

SMF *Austria*

SMF *Germany*

SMF *Brazil*

SMF *Annonay FR* [37]

SMF *NyonFR*

SMF *HuningueFR*

SMF *Netherlands*

SMF *Italy*

SMF *Japan*

SMF *Puerto Rico*

SMF *BaselSWIT* [38]

SMF *HettlingenSWIT*

SMF *Turkey*

No quadro a seguir, apresentamos uma síntese das quinze localidades às quais os documentos analisados pertencem, contendo os números de páginas e palavras de cada SMF.

[37] A abreviação FR que acompanha as localidades *Annonay, Nyon e Huningue* quer dizer *France* (França).
[38] A abreviação SWIT que acompanha as localidades *Basel* e *Hettlingen* quer dizer *Switzerland* (Suíça).

Quadro 4. Número de páginas e palavras por localidade do documento SMF.

Documento	Empresa	Localidade	Nº Palavras	Nº Páginas
SITE MASTER FILE (SMF)	PHARMACo[39].	*England*	9.295	43
		USA	4.098	20
		Austria	8.638	25
		Germany	10.695	36
		Brazil	7.700	40
		AnnonayFR	7.093	42
		NyonFR	5.901	52
		HuningueFR	9.599	27
		Netherlands	6.867	45
		Italy	9.158	59
		Japan	2.250	18
		Puerto Rico	5.851	35
		BaselSWIT	5.299	25
		HettlingenSWIT	7.135	45
		Turkey	11.185	69
		TOTAL	110.766	581

Ressaltamos, porém, que mesmo que o número de páginas não seja indicativo de número de palavras, a intenção de inserir a coluna *Nº de Páginas* acima foi para indicar as diferenças entre a extensão do conteúdo de cada documento, pois, conforme comentamos adiante, alguns documentos divergem quanto à apresentação de gráficos, desenhos, *lay-out*, diagramas e organogramas. Esse tipo de representação gráfica em documentos é típico de áreas técnicas especializadas, como engenharia, química, etc., visto que, dessa maneira, os autores, inseridos em contextos específicos de produção escrita, acreditam veicular as informações dos processos com maior precisão e acuidade. Portanto, para os estudos da linguagem é possível que essas informações possam ser utilizadas em futuras pesquisas sobre verbo-visual.

[39] Utilizamos o nome fictício PharmaCo. para processar a identidade da indústria farmacêutica mencionada nesta pesquisa em anonímia.

Feita a coleta dos dados e a devida limpeza nos textos, foram extraídas dos documentos, as figuras que não seriam necessárias para o tipo de análise lingüística deste estudo, como, por exemplo: os desenhos; *layout* das instalações; e dos equipamentos da fábrica; as tabelas com fórmulas químicas; os fluxogramas e os organogramas organizacionais da empresa; entre outras. Logo, após a constituição do *corpus* de estudo, os documentos foram gravados em arquivos texto para então viabilizarmos o processamento dos dados. Para isso, utilizamos linguagem de programação e rotinas do Cygwin (*shell* de comandos do *Unix*), ferramentas computacionais detalhadas na seção 2.4.

Na próxima seção, prosseguimos com a apresentação dos procedimentos e ferramentas utilizadas nesta análise.

2.3 Procedimentos de análise

2.3.1 Escolha da metodologia

Conforme apresentado na Fundamentação Teórica, a metodologia de análise dos dados desta proposta de investigação está baseada no estudo de *corpus*, uma *abordagem* que possibilita investigar como os escritores e falantes de uma língua exploram os recursos lingüísticos em sua comunidade. Segundo Biber et al. (1998:1), ao invés de observar o que é possível teoricamente, o pesquisador estuda a linguagem em uso de acordo como ela ocorre naturalmente nos textos. Ademais, baseamo-nos na proposta de Stubbs (1996), que defende os métodos para analisar como os padrões das palavras encontrados em *corpora* autênticos podem conduzir ao significado do texto em contextos ou instituições específicas.

A escolha da LC como metodologia deste trabalho deve-se a dois motivos: (i) permitir ao pesquisador realizar um levantamento quantitativo por meio de programas computacionais, e ao mesmo tempo; (ii) possibilitar um levantamento qualitativo, pelo qual o analista é capaz de interpretar os achados lingüísticos que ocorrem em situações autênticas de uso da língua em diferentes tipos de textos que circulam em diferentes comunidades discursivas. Segundo Biber et at. (1998:5), uma das características da metodologia com base em *corpus* é buscar *padrões de associação*, isto é, padrões lingüísticos que podem associar-se a outros aspectos lingüísticos ou não lingüísticos, ou seja, o pesquisador não somente faz uso da freqüência dos padrões

encontrados, mas também pode interpretá-los de acordo com objetivo da pesquisa. Nesse sentido, podemos pressupor que, nesta pesquisa, buscamos associar os *lexical bundles* recorrentes nos quinze exemplares investigados, com o *aboutness* (sobre o que se trata o texto) das nove seções do SMF.

Na visão de Gil (2002:43), uma pesquisa do tipo documental *vale-se de materiais que não recebem ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetos da pesquisa*. Já na pesquisa do tipo levantamento, Gil (:51) acredita que uma das principais vantagens é possibilitar uma análise estatística dos dados, pois *à medida que os levantamentos se valem de amostras probabilísticas, torna-se possível até mesmo conhecer a margem de erro dos resultados obtidos.*

Portanto, nesta *abordagem* probabilística de investigação da linguagem é possível compreender como os discursos são compartilhados nas diferentes esferas de comunicação, e descobrir como esses dados empíricos podem revelar variações de abrangência global e local, dependendo da especificidade do *corpus* a ser explorado.

Na seção seguinte, apresentamos as ferramentas computacionais utilizadas para esta análise.

2.3.2 Ferramentas de análise

Para esta pesquisa, utilizamos a linguagem de programação Perl (*Practical Extraction and Report Language*) e a ferramenta Cygwin. A linguagem de programação Perl foi criada por Larry Wall em 1987 e tem como principais características auxiliar o usuário na programação de pequenas tarefas ou *scripts*, como também facilitar a manipulação de textos e processos. Já a ferramenta Cygwin é um emulador do sistema operacional Unix para Windows®, ambiente para digitação de linhas de comando. Além desses instrumentos, com o suporte de um especialista da área de Ciências da Computação, foi desenvolvido um aplicativo denominado *Análise Lingüística*, e criado um Banco de Dados em SQL, para gerar as consultas dos *lexical bundles* (ver seção 2.4.2.1).

A escolha dessas ferramentas de análise deveu-se ao fato de ser o meio mais apropriado no processamento de um determinado volume de dados para se obter: a freqüência das palavras; as formações de *lexical bundles* recorrentes; e possibilitar o cruzamento de dados entre tipos de

textos diversos, ou textos de uma mesma especialidade, como, por exemplo, os quinze documentos SMF.

Segundo Berber Sardinha (2003:3), essa forma de investigação é uma das maneiras de aproximar a Lingüística de *Corpus* com foco no texto, e verificar sobre o que se trata o texto por meio da investigação das suas divisões internas. Segundo o autor:

> [...] O foco no texto permite ao analista de *corpus* lidar com questões de ordem textual, como a organização genérica, a léxico-gramática típica de gêneros específicos e a temática (*aboutness*) [...].

A respeito da temática ou *aboutness* das divisões internas, que compõem um gênero ou tipo de texto específico, neste estudo, o documento *Site Master File* (SMF) será analisado de forma a buscar elementos lexicais por meio da extração dos *bundles*, e estabelecer uma ligação semântica coesiva com os assuntos tratados em cada uma das nove seções do referido documento.

Na seção seguinte, apresentamos as etapas dos procedimentos para a apuração dos resultados desta análise.

2.4 Etapas dos Procedimentos

Esta pesquisa contemplou seis etapas de procedimentos para a realização da análise dos resultados, as quais foram descritas e organizadas nesta seção da seguinte maneira:

1ª Etapa: Identificar a estrutura organizacional dos sumários dos quinze documentos SMF;

2ª Etapa: Verificar os padrões léxico-gramaticais a partir da extração dos *lexical bundles* de três palavras;

3ª Etapa: Identificar os *lexical bundles* recorrentes entre as seções dos documentos SMF;

4ª Etapa: Comparar os *bundles* semelhantes do *corpus* de pesquisa com os *bundles* do *corpus* de referência BNC para medir a sua probabilidade de ocorrência.

5ª Etapa: Identificar os *lexical bundles* recorrentes em cada uma das quinze localidades do documento;

6ª Etapa: Estabelecer uma relação de coesão entre os *bundles* semelhantes e o tópico das seções do SMF.

De tal forma a explorar esses itens, a seguir detalhamos as seis etapas encarregadas dos procedimentos desta análise.

2.4.1 A estrutura organizacional do documento

A primeira etapa dos procedimentos para a análise deste estudo, a organização textual do documento *Site Master File*, foi realizada manualmente, já que o conjunto dos quinze documentos analisado, apresentaram no sumário (*Index*) uma estrutura organizacional explícita em sua própria forma composicional.

Nesse sumário, cada uma das nove seções que compõem o documento é nomeada por um título, acompanhada de seu respectivo número. Seguida do título, cada seção contém sub-seções e itens com indicações (por títulos, frases, etc.) dos assuntos que devem ser desenvolvidos ao longo da produção escrita. Para exemplificar, abaixo, apresentamos parte de uma das seções:

Seção 2. *Personnel*

1. *Organization chart*
2. *Qualification, experience and responsibilities of key personnel*
3. *Outline of arrangement for basic and in-service training*

3.1 *Identification of training needs*

3.2 *Training related to GMP requirements*

3.3 *Type of training*

3.4 *Efficacy of training*

3.5 *Identification of retraining need*s

3.6 *Training records*

Para estabelecer uma comparação entre os quinze documentos pesquisados, na busca de respostas para a nossa primeira pergunta de pesquisa - Quais são as semelhanças da estrutura

organizacional do *Site Master File* entre as quinze localidades em que o documento circula? - realizamos quatro procedimentos.

1º passo: separar os quinze documentos;
2º passo: destacar os sumários dos documentos;
3º passo: comparar as divisões dos sumários;
4º passo: listar as semelhanças e variações encontradas.

Seguindo essa ordem, primeiramente, separamos os quinze documentos SMF que serviriam de base para a comparação, conforme critérios discutidos na análise dos resultados. Em seguida, destacamos, manualmente, de cada documento SMF, o sumário ou *index*. Logo após, num terceiro passo, realizamos uma comparação dos nomes das seções, sub-seções e itens, a fim de identificar possíveis semelhanças ou variações na redação ou conteúdo dos quinze sumários dos documentos. Por último, representadas em tabelas, listamos as variações encontradas, e o conjunto de semelhanças entre os quinze sumários dos SMF, para então partirmos para a discussão (ver Capítulo 3).

Como podemos observar, ainda que a maioria das etapas de análise deste trabalho fez uso de instrumentos computacionais, nessa primeira etapa, em particular, não foi possível o uso do computador, haja vista que as peculiaridades das amostras dos sumários dos documentos poderiam ser melhores interpretadas manualmente.

2.4.2 Os elementos léxico-gramaticais

A segunda etapa dos procedimentos desta análise concentrou-se em verificar os padrões léxico-gramaticais da estrutura interna do documento SMF, visando responder a nossa segunda pergunta de pesquisa - Quais são os *lexical bundles* semelhantes entre as seções dos quinze documentos SMF?

Conforme apresentado na Fundamentação Teórica desta pesquisa, escolhemos a categoria *lexical bundles* sob a perspectiva de Douglas Biber et al. (1999), e das pesquisas de Michael Stubbs (2007), Cortes (2002, 2006, 2008), Berber Sardinha (2003), Scott e Tribble (2006), e Hyland (2008).

Conforme detalhado no Capítulo 1 desta pesquisa, segundo Biber et al. (1999), os *lexical bundle*s são seqüências de palavras que ocorrem naturalmente no discurso, ou seja, os *bundles* são formados por expressões recorrentes, independente de sua idiomaticidade ou de sua condição estrutural. Para Berber Sardinha (2003:6), *pacotes lexicais referem-se a uma seqüência de palavras de tamanho determinado, retirado do texto ou corpus por um programa de computador*. Stubbs (2007) apóia-se no conceito de "*multi-word sequence*" ou *n-grams* para referir-se aos estudos baseados na extração do conjunto de palavras ininterruptas recorrentes no texto por meio de programas computacionais.

Na visão de Berber Sardinha (2003), um 'pacote lexical' de 2 palavras pode conter apenas fragmentos que fazem parte de 'pacotes' maiores. Entretanto, segundo o autor, na medida em que os 'pacotes' aumentam de tamanho, a tendência de 'pacotes' fragmentados diminui, tornando-os mais informativos. Para Biber et al. (1999), os *lexical bundles* de três ou mais palavras são seqüências recomendáveis num processo de análise textual. No entanto, a formação de *bundles* de quatro, cinco, ou mais palavras são mais fraseológicos, portanto, menos comum de ocorrer.

O fato de o *corpus* desta pesquisa pertencer a um domínio específico nos conduziu a optar pela extração de *bundles* de três palavras, pois contávamos com a possibilidade de identificar unidades terminológicas num *corpus* de especialidade (Teixeira, 2007). Isto significa dizer, que este estudo implica uma possibilidade de refletir sobre uma possível "etiquetagem" [40] dos *bundles* de um *corpus* específico em futuras pesquisas, utilizando, como, por exemplo, as categorias estruturais e funcionais descritas por Biber et al. (1999). Dessa forma, contribuindo para o contexto de produção escrita de documentos específicos, como é o caso do SMF.

Para confirmar se de fato este *corpus* de pesquisa pertencia a um domínio específico, estabelecemos um critério de comparação dos *bundles* encontrados no SMF com os *bundles* processados a partir do *corpus* de língua inglesa BNC (*British National Corpus*).

Diante das asserções dos autores mencionados anteriormente, e do volume de extensão do *corpus* desta análise, optamos pela extração de *lexical bundles* de três palavras, porque também acreditamos ser um tamanho adequado, que nos possibilita extrair dados relevantes para o estudo

[40] Segundo Teixeira (2007:117-118), *etiquetagem é a inserção automática, semi-automática ou manual de qualquer tipo de informação em um corpus de estudo, com vistas a facilitar sua análise*. Para a autora, a etiquetagem pode ser de vários tipos, como por exemplo: Morfossintática – indica a classe gramatical de cada palavra do *corpus*; Sintática – analisa a sintaxe das frases e; Semântica – classifica as palavras de acordo com as características semânticas.

comparativo entre os padrões léxico-gramaticais contidos nos quinze documentos SMF da especialidade farmacêutica.

Na sub-seção seguinte, detalhamos os primeiros passos dessa segunda etapa dos procedimentos.

2.4.2.1 A extração dos *bundles* de três palavras

Dando prosseguimento à segunda etapa dos procedimentos a fim de extrair os *bundles* de três palavras dos quinze exemplares do documento SMF, percorremos três passos:

1º passo: gerar um diretório[41];
2º passo: realizar a limpeza nos arquivos, utilizando linhas de comandos do *shell*;
3º passo: realizar a importação dos dados, por meio do aplicativo *Análise Lingüística*.

De acordo com a ordem elencada acima, partimos para o primeiro passo gerando um diretório no Windows®, com a seguinte estrutura de pastas:

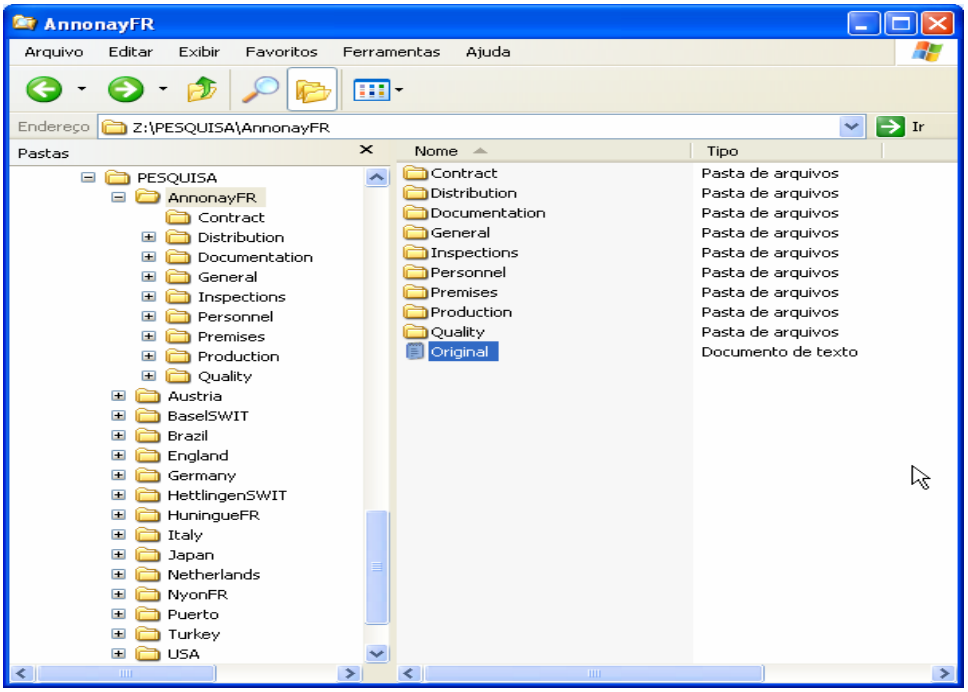
(i) uma pasta raiz denominada Pesquisa;
(ii) sub-pastas para cada uma das quinze localidades; e
(iii) uma pasta para cada uma das nove seções do documento, dentro de cada sub-pasta das quinze localidades.

Em seguida, convertemos os quinze documentos que estavam em formato original MS Word® (extensão.doc) para o formato texto simples (extensão.txt), gerando um arquivo para cada localidade, denominado Original.txt.

Conforme descrito, apresentamos na figura abaixo a estrutura de pastas para o armazenamento dos dados desta pesquisa.

[41] Estrutura hierárquica de arquivamento eletrônico.

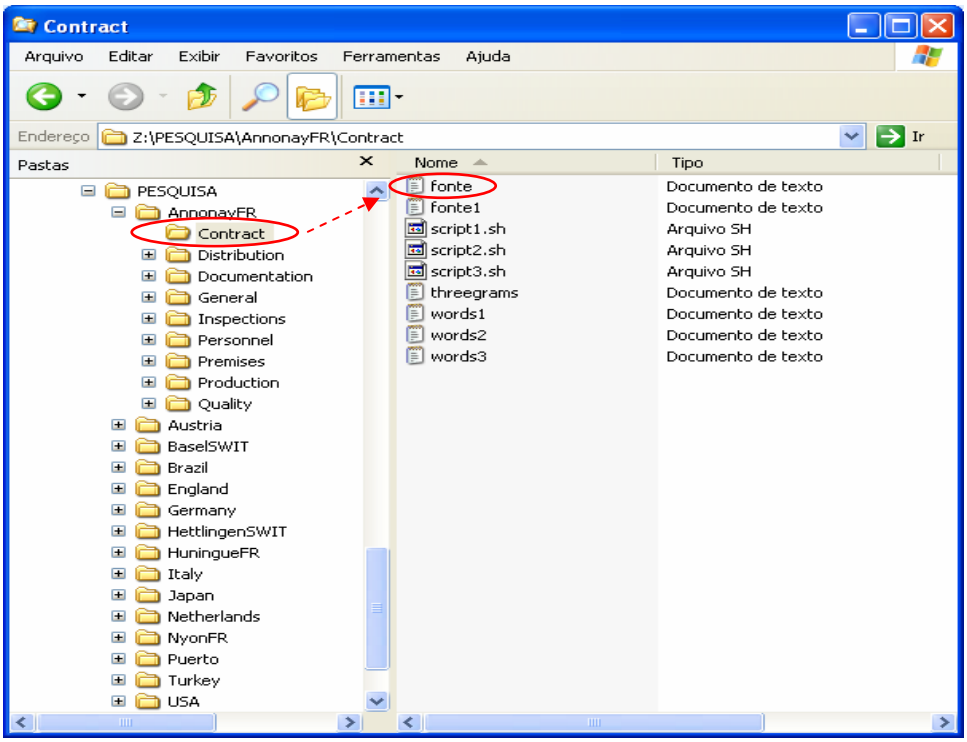
Figura 1. Estrutura do diretório para o armazenamento dos dados de pesquisa.



Após a organização e armazenamento dos dados no diretório, utilizando o aplicativo Notepad do ambiente Windows®, desmembramos o arquivo denominado Original.txt nas nove seções do documento, ou seja, para cada seção do documento SMF de cada localidade foi gerado um arquivo denominado fonte.txt. Esse arquivo fonte.txt foi gravado na pasta correspondente ao nome da seção do documento.

Na figura a seguir, demonstramos um exemplo da pasta seção-*Contract*, com seu arquivo fonte.txt.

Figura 2. Criação do arquivo fonte.txt nas pastas do diretório.

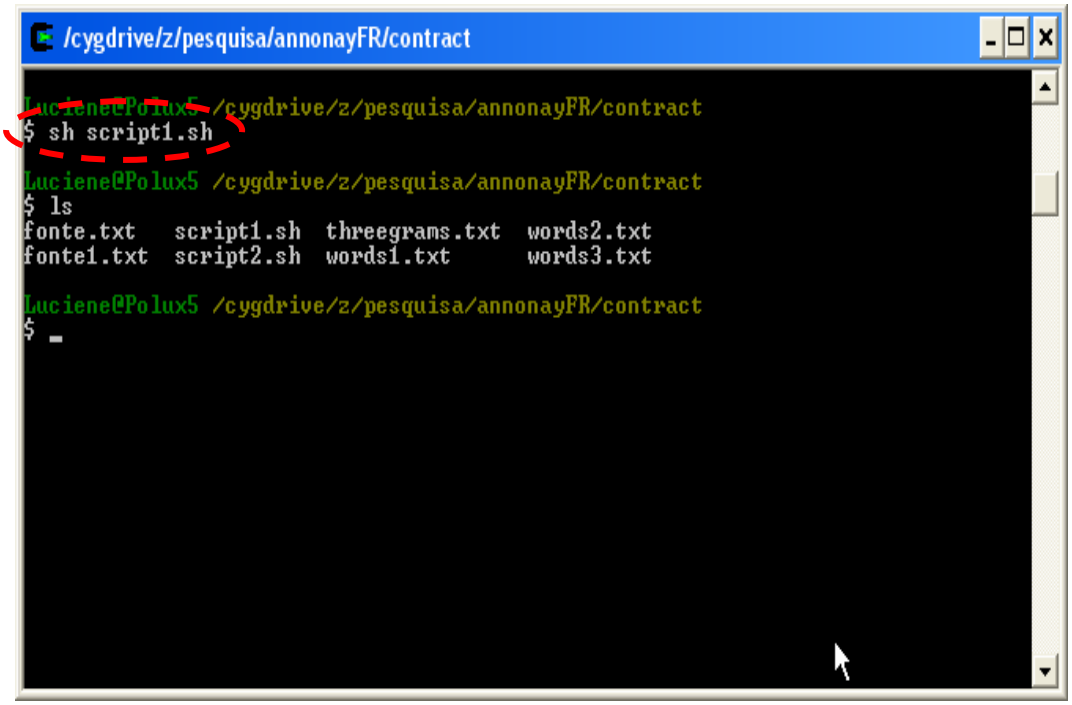


Encerrado o primeiro passo dos procedimentos, no segundo passo, utilizando linhas de comandos (*shell*) do ambiente C*ygwin*, um aplicativo emulador do ambiente Linux para Windows®, realizamos uma limpeza no arquivo fonte.txt. Os critérios de limpeza foram: espaços; substituições de todos os números por zero; substituições de nomes de pessoas e empresas por nomes fictícios; caracteres de pontuação e caracteres de controle (retorno de carro, tabulação e avanço de linhas)[42]. Essa limpeza foi realizada através do comando *sh* (um interpretador de *script* computacional), com a aplicação do arquivo script1.sh no arquivo fonte.txt, gerando o arquivo fonte1.txt.

Na figura a seguir, apresentamos o ambiente C*ygwin*, com a linha de comando *sh* em destaque.

[42] Comandos de edição de texto do Windows® conhecidos na sigla em inglês como CR (Carriage Return), LF (Line Feed) e TAB (Tabulação).

Figura 3. Ambiente *Cygwin* e linha de comando *sh*



Nessa fase dos procedimentos, por meio do aplicativo *Notepad* do ambiente Windows®, realizamos uma segunda limpeza no texto para complementar a limpeza anterior. Esse procedimento foi necessário devido ao arquivo script1.sh não ter conseguido eliminar todos os caracteres indesejáveis para uma formação adequada dos *lexical bundles*. Tais caracteres eliminados foram: marcadores de parágrafos; símbolos e espaços.

Seguindo com o tratamento dos dados, submetemos um outro arquivo de script computacional, denominado script2.sh, ao comando sh do *shell*, gerando os arquivos: words1.txt; words2.txt; e words3.txt. Esses arquivos contêm uma lista de palavras do texto fonte1.txt, e a combinação dos três arquivos gera a formação dos *lexical bundles* de três palavras, gravado em um arquivo denominado *threegrams*.txt. (ver Figura 3).

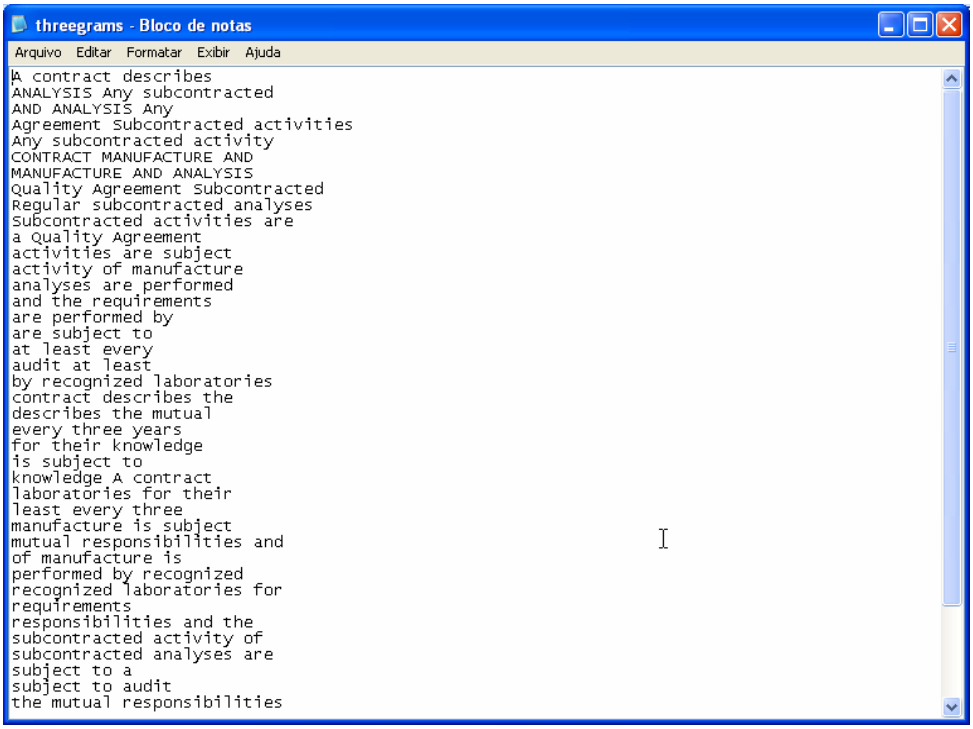
Exemplificando a lógica de formação dos *lexical bundles*, teremos, grosso modo, o seguinte: (quantidade de *bundles*) = (quantidade de palavras - 2), ou seja, se tomarmos como exemplo, um arquivo texto que contenha 100 palavras, a quantidade de *bundles* será 98, resultando no número de palavras menos 2 (100 palavras - 2 = 98 *bundles*). Porém, de acordo com o modelo de linguagem baseado em *n-grams* (conforme mencionado no Capítulo 1 seção 1.2), o modelo estatístico de *n-grams* é calculado em função do tamanho da seqüência de

palavras. Por exemplo: Total de *Bundles* (TB) = Total de Palavras (TP) do texto – (*n* – 1), onde *n* representa o tamanho da seqüência de palavras, e o número 1, a constante. No caso desta análise, utilizando o exemplo anterior, aplica-se a seguinte equação:

TB = TP – (*n* –1), ou seja, 98 = 100 – (3-1).

Na figura abaixo, apresentamos um exemplo do arquivo *threegrams*.txt referente à localidade AnnonayFR - seção *Contract*, gerado através do script2.sh.

Figura 4. Arquivo *threegrams*.txt gerado através do script2.sh.(seção *Contract* do SMF*AnnonayFR*)



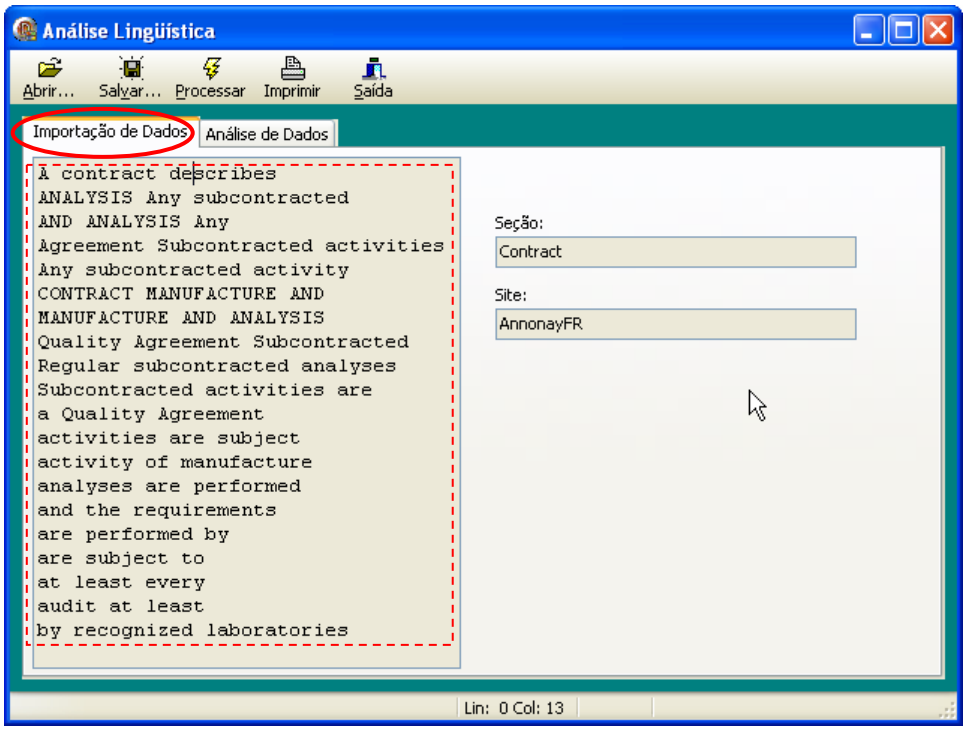
Os script1.*sh* e script2.*sh*, utilizados nessa fase, estarão disponíveis para consulta no Anexos 1 e 2 deste trabalho.

No terceiro e último passo desta etapa, utilizamos um aplicativo para Windows®, denominado *Análise Lingüística*. O aplicativo *Análise Lingüística* foi projetado para realizar duas operações básicas: (i) importar os dados do arquivo texto (os *lexical bundles* do arquivo *threegrams*.txt) para um banco de dados relacional (o SQL); e (ii) realizar as consultas lingüísticas para a análise dos dados. Em conjunto com a pesquisadora, esse aplicativo foi

desenvolvido por um especialista da área de Desenvolvimento de *Software*, e seu uso é específico para esta análise, podendo ser moldado ou estruturado de acordo com outras propostas de análise.

Ressaltamos que importar os dados significa, que os *lexical bundles*, processados através dos script1 e script2 (ver Figuras 3 e 4), foram transferidos com o auxílio da operação projetada no aplicativo. Para ilustrar, demonstramos o aplicativo *Análise Lingüística*, sinalizando a página *Importação de Dados* com a área de entrada dos dados, provenientes do arquivo *threegrams*.txt, tendo a localidade *AnnonayFR* – seção *Contract* como referência.

Figura 5. Importação dos *bundles* por meio do aplicativo *Análise Lingüística.*



Dessa forma, sucessivamente, a operação de importação dos dados foi realizada para cada arquivo *threegrams*.txt das nove seções dos quinze documentos SMFs.

Para uma breve noção a respeito de banco de dados relacional, trata-se de um depósito de dados gerenciado por um aplicativo servidor, ou seja, ele permite consultar, cruzar informações, fazer contagens, estabelecer freqüências, e outras funções necessárias à análise dos dados.

Logo, segundo Ramalho (1999:22), a linguagem SQL surgiu da empresa de tecnologia IBM, e é uma abreviação de *Structured Query Language*, padronizada pelo ANSI (*American*

National Standards Institute). De acordo com o autor, trata-se de uma linguagem muito *enxuta e especializada*, tendo como uma das suas principais vantagens não ser uma linguagem do tipo *procedural*, ou seja, onde o programador deve dizer, passo a passo, o que o computador deve fazer, ao contrário de uma linguagem SQL interativa (utilizada pelo especialista nesta análise), a qual opera diretamente um banco de dados, como o fim de produzir os resultados desejados. Conforme Ramalho (1999:23):

> [...] o usuário digita um comando *SQL* que é executado imediatamente e mostra os resultados após a execução dos comandos. A maioria dos bancos de dados possui uma ferramenta que permite a execução interativa da linguagem SQL, como é o caso do SQLTalk do SQLBase, o SQL Plus da Oracle ou o Query Analizer do SQL Server 7 da Microsoft [...]

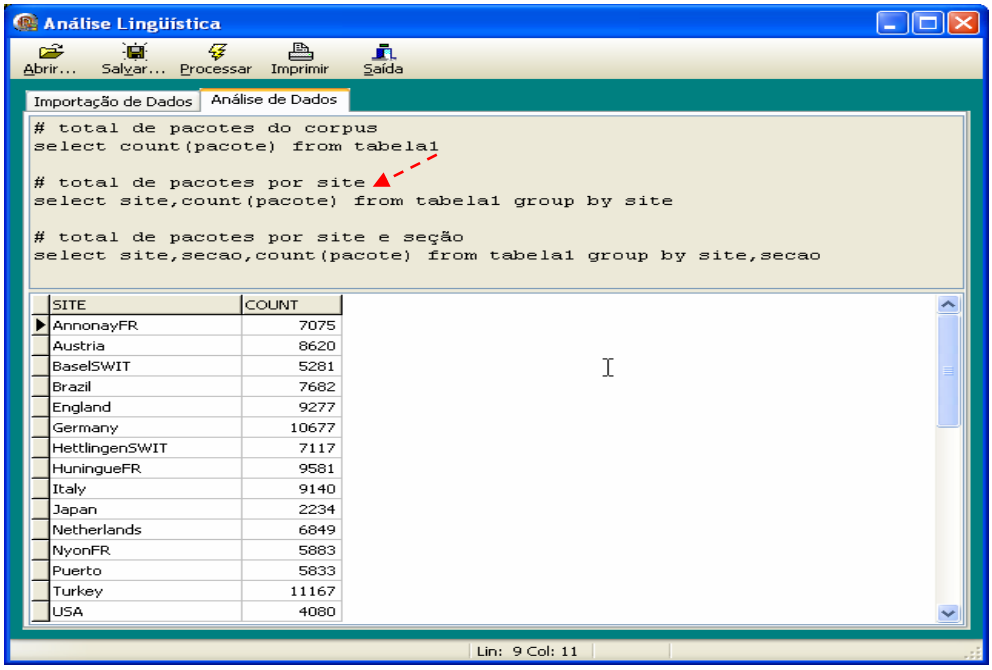
No entanto, o especialista desenvolvedor do aplicativo desta análise, criou uma ferramenta para realizar a interação com o Banco de Dados SQL (o aplicativo denominado *Análise Lingüística*), permitindo a interação entre o Banco de Dados e o usuário, neste caso a pesquisadora.

Outrossim, esse banco de dados pode ser hospedado em um *Datacenter*, que é uma central de operações em rede que oferece serviços de armazenamento de dados. Ainda, esse banco de dados pode estar disponível localmente, isto é, a base de dados torna-se disponível assim que o banco de dados é carregado (*Local Database*). Neste caso, o aplicativo *Análise Lingüística* foi instalado localmente na máquina da pesquisadora.

Nesse passo da etapa, apuramos a quantidade de *lexical bundles* contida em cada documento SMF, e a quantidade de *lexical bundles* contida em cada uma das nove seções do documento.

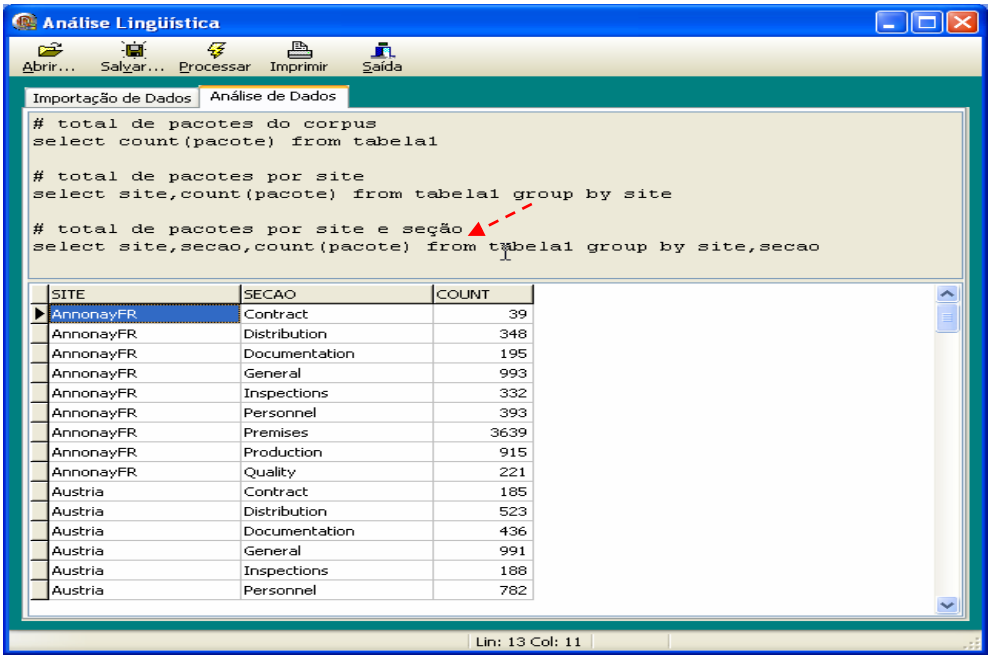
Na figura a seguir, demonstramos no aplicativo *Análise Lingüística*, a apuração da quantidade de *bundles* por localidade, sinalizando a linha de comando.

Figura 6. Contagem dos *bundles* por localidade por meio do aplicativo *Análise Lingüística*.



Quanto à quantidade dos *bundles* de cada seção dos SMFs das quinze diferentes localidades, os dados estão ilustrados na figura a seguir.

Figura 7. Contagem dos *bundles* de cada seção das quinze localidades.



Com o objetivo de sintetizar os números de *lexical bundles* por seção e por localidade, na tabela a seguir, apresentamos os números de *bundles* encontrados em cada uma das nove seções que compõem o documento SMF de cada uma das quinze localidades, como também o total geral dos *bundles* encontrados no *corpus*.

Tabela 1. Quantidade de *lexical bundles* por seção e por localidade.

		SITES (localidades)				
	Seção	Annonay (FR)	Austria	Basel	Brazil	England
1	General Information	993	991	1776	634	1409
2	Personnel	393	782	531	1810	2414
3	Premises	3639	4161	1503	2985	2977
4	Documentation	195	436	452	285	560
5	Production	915	799	602	802	822
6	Quality Control	221	555	249	276	206
7	Contract	348	523	56	290	520
8	Distribution	39	185	89	415	253
9	Self-Inspections	332	188	23	185	116
Total de pacotes por SMF		**7075**	**8620**	**5281**	**7682**	**9277**
	Seção	Germany	Hettlingen (SWIT)	Huningue (FR)	Italy	Japan
1	General Information	2219	1049	1934	1821	106
2	Personnel	1289	1804	1012	1138	818
3	Premises	4274	2167	3490	2976	472
4	Documentation	879	339	542	424	43
5	Production	718	817	1060	1569	335
6	Quality Control	334	302	403	280	136
7	Contract	605	434	717	504	270
8	Distribution	91	149	248	212	0
9	Self-Inspections	268	56	175	216	54
Total de pacotes por SMF		**10677**	**7117**	**9581**	**9140**	**2234**
	Seção	The Netherlands	Nyon (FR)	Puerto Rico	Turkey	USA
1	General Information	1676	1420	794	2257	858
2	Personnel	1487	591	1821	827	649
3	Premises	1129	1764	1276	4947	701
4	Documentation	638	309	388	501	274
5	Production	298	1000	332	1125	489
6	Quality Control	423	280	610	375	390
7	Contract	168	299	465	601	483
8	Distribution	333	106	15	145	111
9	Self-Inspections	697	114	132	389	125
Total de pacotes por SMF		**6849**	**5883**	**5833**	**11167**	**4080**
TOTAL GERAL		**110.496**				

Com o auxílio das ferramentas computacionais utilizadas nos passos desta etapa, e de posse dos resultados obtidos, na seção seguinte, passamos à terceira etapa dos procedimentos, com o intuito de buscar elementos para embasar a nossa segunda pergunta de pesquisa - Quais são os *lexical bundles* semelhantes entre as seções dos quinze documentos SMF?

2.4.3 Os *lexical bundles* recorrentes

Com a finalidade de responder a nossa segunda pergunta de pesquisa e fornecer subsídios à interpretação dos dados da nossa terceira pergunta de pesquisa – Como os *lexical bundles* semelhantes refletem o tópico das seções do documento SMF? - percorremos uma terceira etapa de análise.

Nesta terceira etapa, foi necessário definir alguns critérios para filtrar os dados relevantes para o propósito desta pesquisa, ou seja, estabelecer pontos de corte na extração dos *lexical bundles*. Além do mais, fizemos uso do aplicativo *Análise Lingüística*, exclusivamente, com a finalidade de cruzar as informações entre os textos, identificando as ocorrências entre os quinze documentos, resultando, dessa maneira, na listagem dos *bundles* semelhantes. Assim, percorremos dois passos, conforme seguem:

1º passo: listar a freqüência dos *bundles* recorrentes de acordo com o ponto de corte;
2º passo: identificar quais as localidades em que os *bundles* recorrentes ocorriam.

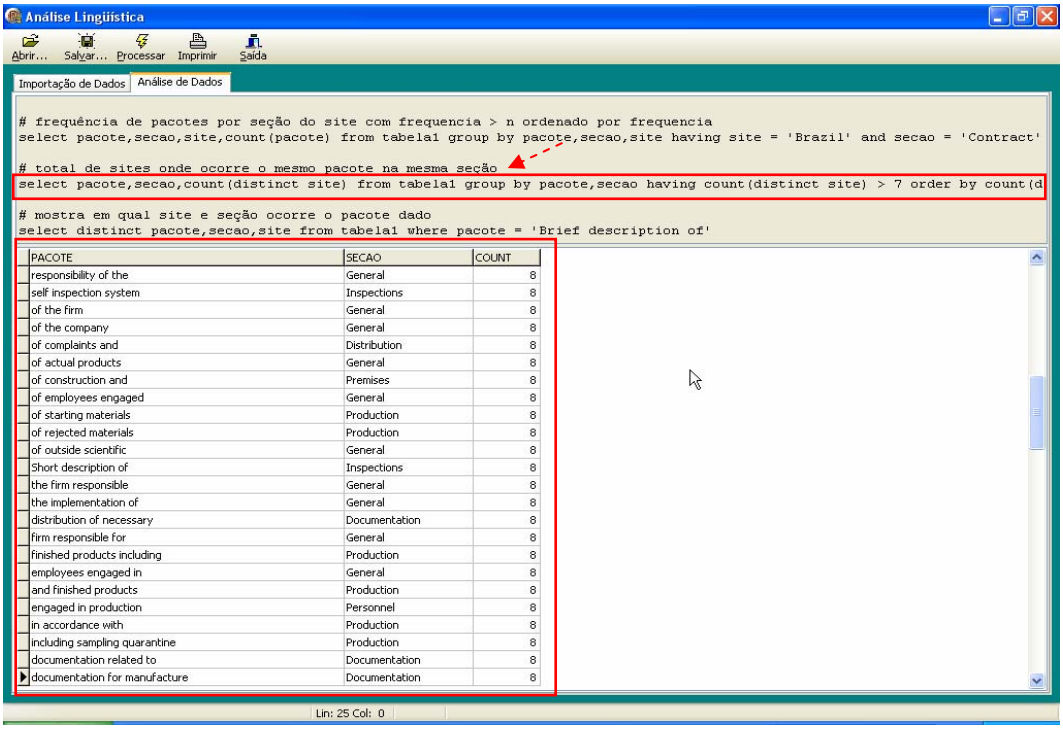
Conforme dito acima, no primeiro passo, executamos uma linha de comando no banco de dados a fim de listar a freqüência dos *bundles* recorrentes nas nove seções do documento. Devemos esclarecer, portanto, que o aplicativo *Análise Lingüística* possui um editor de comandos que possibilita a digitação da linha de instrução, obedecendo à sintaxe da linguagem SQL, isto é, por meio da digitação de comandos distintos, como, por exemplo: INSERT; UPDATE; DELETE; SELECT; FROM; COUNT; DISTINCT; entre outros, é possível realizar cada consulta ao Banco de Dados SQL, dos quais o resultado é exibido posteriormente na área designada no aplicativo.

Esses comandos do SQL são denominados DML (*Data Manipulation Language*) e DQL (*Data Query Language*)[43]. Devido à natureza técnica de tais procedimentos, referentes à sintaxe das linhas de comando, o especialista (desenvolvedor do aplicativo) forneceu as instruções necessárias para a digitação dos comandos desta análise, e em seguida a pesquisadora executou a busca pelas informações lingüísticas.

[43] Para maiores detalhes ver Ramalho (1999).

Na figura abaixo, apresentamos como o aplicativo *Análise Lingüística* respondeu a busca pela distribuição dos *bundles*. Para uma melhor compreensão, sinalizamos na figura a linha de comando, a área de digitação, e a área de resultados.

Figura 8. Total de localidades onde ocorreu o mesmo *bundle* na mesma seção.



Definimos o ponto de corte > 7, por representar 55% do total de quinze, ou seja, *bundles* que fossem recorrentes em no mínimo 8 localidades das 15 pesquisadas (ver detalhes no Capítulo 3). Conforme observado na Figura 8, obtivemos: na primeira coluna (da esquerda para a direita), o próprio *bundle*; na coluna do meio, a seção onde ocorre o *bundle*; e na terceira coluna, o número de localidades na qual esse *bundle* foi recorrente.

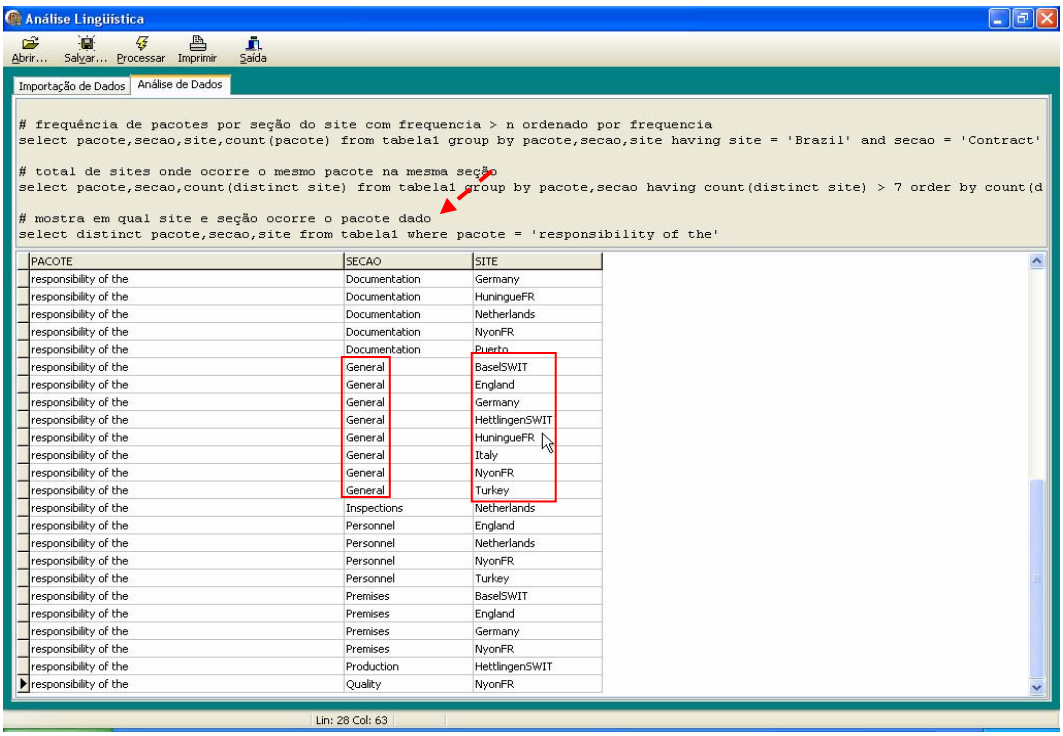
Só para ilustrar, o *bundle responsibility of the* ocorreu em 8 localidades na seção *General Information*, o *self inspection system* ocorreu em 8 localidades na seção *Inspections*, e assim por diante. Dessa maneira, o mesmo procedimento foi utilizado em todas as outras seções para a captura dos *bundles* semelhantes.

Na busca por *lexical bundles* semelhantes, o segundo passo foi identificar em que localidades esses *bundles* foram recorrentes. A opção por essa busca deveu-se a dois aspectos: (i)

possíveis perguntas que poderiam surgir durante a análise dos resultados, uma delas seria o porquê algumas localidades, como, por exemplo, o SMF *Japan* e outras, não apresentavam *bundles* semelhantes com relação às demais (ver Capítulo 3); e (ii) garantir a origem dos dados, ou seja, se os resultados apresentaram 8 localidades ou mais, quais seriam essas localidades?

Para essa fase, perguntamos ao banco de dados em qual localidade e seção ocorria o *bundle*. Essa busca foi realizada digitando na linha de comando SQL o nome do *bundle* identificado no passo anterior. Na figura abaixo, apresentamos um exemplo utilizando o *bundle responsibility of the*, e sinalizando a linha de comando e a área dos resultados.

Figura 9. Localidade e seção em que ocorreu o *bundle*.



Observando a Figura 9, não só podemos identificar a recorrência do *bundle responsibility of the* na seção *General Information*, como também em outras seções, como, por exemplo, as seções: *Documentation*; *Personnel*; *Premises*; entre outras. No entanto, dado o critério de corte estabelecido para essa fase (>7), não nos atentamos a esses achados. Encerrados os dois passos na identificação dos *bundles* semelhantes, e a recorrência nas suas respectivas localidades dos

quinze documentos SMFs, elaboramos tabelas e gráficos para uma melhor visualização e organização dos resultados (ver Capítulo 3).

Na próxima seção, apresentamos a quarta etapa dos procedimentos desta análise.

2.4.4 Os *bundles* do SMF *vs* os *bundles* do BNC

O propósito desta quarta etapa de análise, foi estabelecer um *baseline* com base nos *bundles* semelhantes encontrados no SMF e os *lexical bundles* de três palavras extraídos do BNC (*British National Corpus*)[44], ou seja, comparar os dois *corpus* com o intuito de verificar se a probabilidade de recorrência dos *lexical bundles* semelhantes do SMF era expressiva. Em outras palavras, se os *bundles* tinham relevância no domínio específico do documento SMF, inseridos no contexto farmacêutico, ou eram apenas um fenômeno comum de língua geral. Essa base permitiu-nos identificar quais *bundles* ou *key-bundles* (termo denominado pela pesquisadora) poderiam representar marcas características do documento SMF, e não de outros textos.

Retomando o que foi discutido no Capítulo 1 (seção 1.1.2), a respeito da questão da representatividade do *corpus*, e tomando como parâmetro alguns autores que debatem sobre o tema, como, por exemplo, Berber Sardinha (2000), que apesar das controvérsias, pressupõe que um *corpus* deve ser representativo da língua em uso, e de que a representatividade está relacionada à probabilidade e à extensão do *corpus*, propusemo-nos nesta etapa de análise a testar o *corpus* do SMF, embora o nosso *corpus* de estudo seja considerado de pequeno porte (em média 110 mil palavras, conforme mostra a Tabela 1).

Esse tipo de contraste, segundo Vasilévski (2007:59-60), é um recurso muito útil para pesquisas com dados empíricos, pois permite difundir a abrangência dos resultados obtidos entre *corpus* distintos. Ademais, segundo a autora, fazer uso de um *corpus* conceituado e de grande porte (por exemplo, o BNC), oferece maior credibilidade às pesquisas com *corpus*, contribuindo para *amenizar a tal questão da representatividade* (Vasilévski, 2007:60).

Com relação aos tipos de normalizações, utilizamos como parâmetro o que discute Rocha (2007) a respeito de uma base pré-determinada, isto é, um cálculo que estabeleça a probabilidade de um *lexical bundle*.

[44] Ver Capítulo 1 – seção 1.1 para detalhes sobre *corpora* eletrônicos da língua inglesa.

Assim, a partir dos pressupostos acerca da representatividade do *corpus*, engendramos a quarta etapa desta análise sob os seguintes passos:

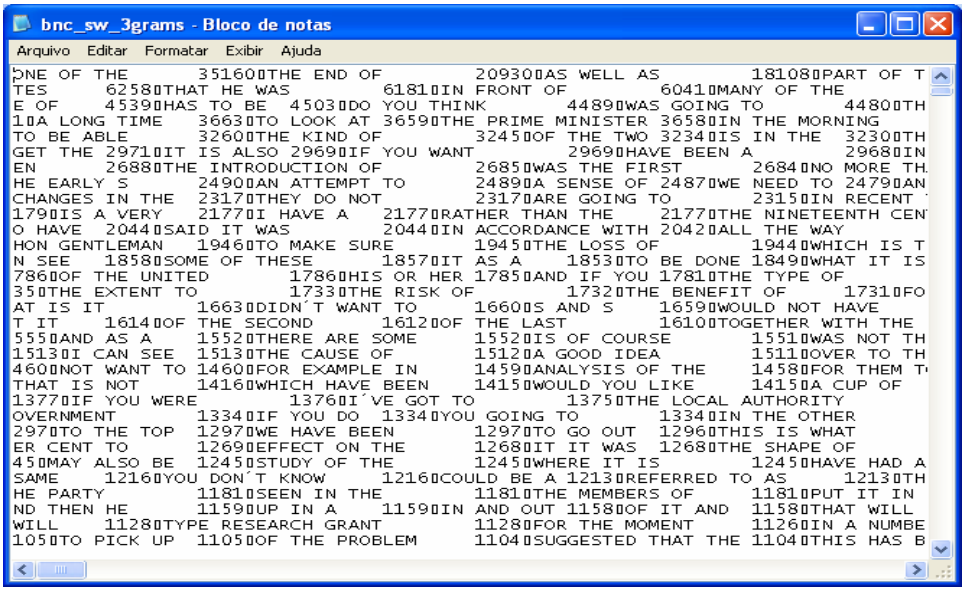
1º passo: importar os *bundles* de três palavras extraídos do *corpus* do BNC para o Banco de Dados SQL;

2º passo: identificar a freqüência de ocorrência dos *bundles* semelhantes do SMF no *corpus* do BNC;

3º passo: identificar a freqüência de ocorrência dos *bundles* semelhantes no *corpus* do SMF.

No primeiro passo, contamos com o apoio de um arquivo texto contendo a lista de 3grams[45] do BNC (escrito e falado), fornecido pelo doutorando em Lingüística Aplicada da PUC-SP, José Lopes Moreira Filho[46]. Segundo esse pesquisador, a geração dessa lista foi realizada com o auxílio de *scripts* computacionais, através de linhas de comando em *shell* (*sh*). Na figura abaixo, apresentamos o formato pelo qual recepcionamos a arquivo texto dos *bundles* de três palavras do BNC.

Figura 10. Formato do arquivo texto dos *bundles* de três palavras do BNC.



[45] Termo adotado pelo doutorando José Lopes Moreira Filho.
[46] Plataforma lattes: http://lattes.cnpq.br/4599251103040654

Os resultados processados a partir desse arquivo texto apresentaram os seguintes dados, conforme tabela abaixo:

Tabela 2. Quantidade de *lexical bundles* de três palavras no BNC.

Corpus do BNC	
Total de Palavras	Total de 3grams
99.613.200	4.423.944

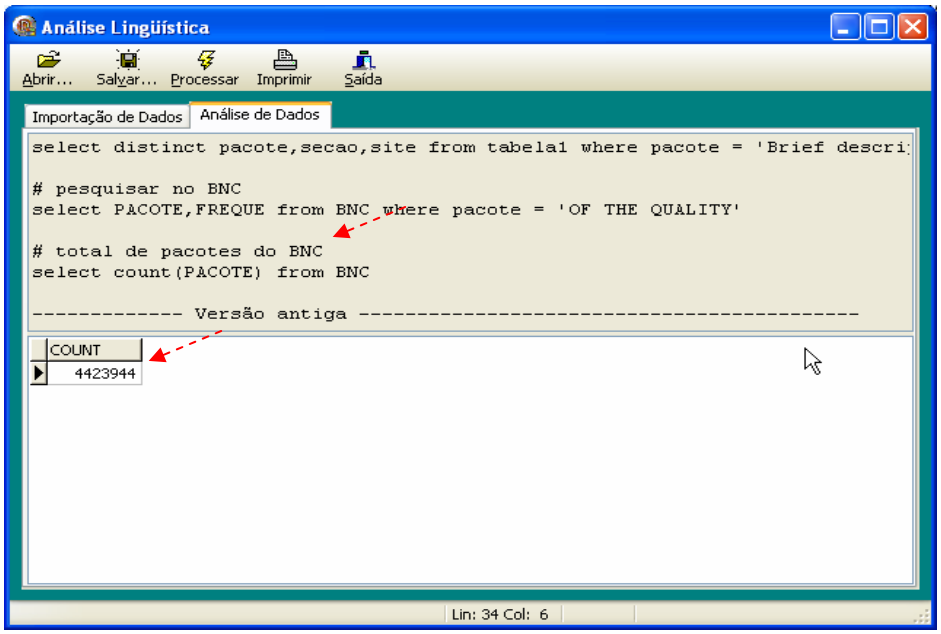
O critério de corte aplicado para a extração dos *bundles* do BNC foi o de > 1, diferentemente do critério adotado para a busca dos *bundles* semelhantes do SMF, conforme sub-seção 2.4.3. Desse modo, adotamos o ponto de corte > 1, pois dessa forma pudemos filtrar um maior número de recorrências possíveis, capaz de cobrir a abrangência dos resultados obtidos do *corpus* de referência (o BNC), denotando, assim, a representatividade do *corpus* de estudo a partir dos *bundles* semelhantes do SMF.

Assim, de posse da lista de *bundles* de três palavras do BNC (4.423.945), fornecida pelo pesquisador José Lopes Moreira Filho, realizamos a importação dos dados, igualmente como fizemos na segunda etapa dos procedimentos (ver sub-seção 2.4.2.1).

Antes de descrever esse passo, acreditamos ser necessário apresentar o resultado do número de *bundles* do BNC, exibido no aplicativo *Análise Lingüística*, para atestar se os valores estavam ou não equiparados à Tabela 2.

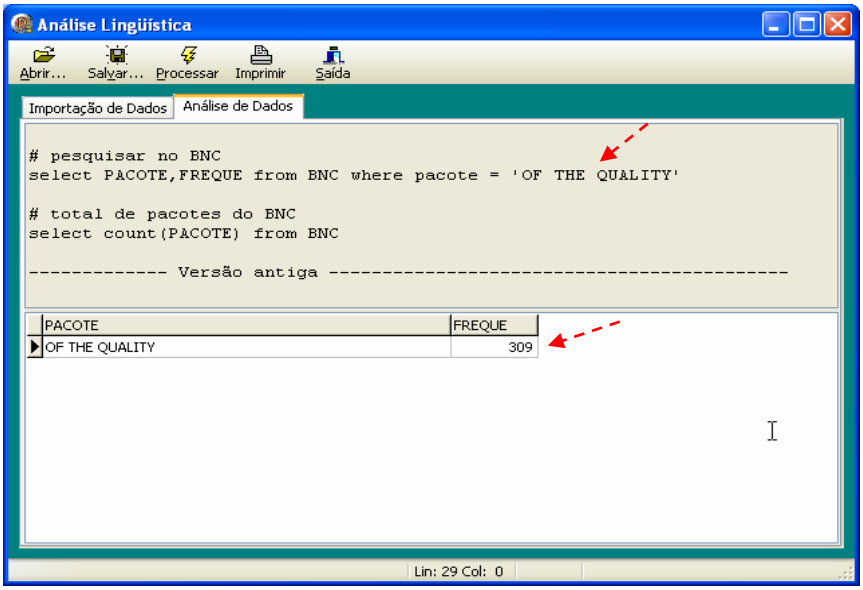
Na figura a seguir, apresentamos tais resultados, sinalizando a linha de comando e os valores esperados.

Figura 11. Resultado do total de *bundles* do BNC no aplicativo *Análise Lingüística*.



Comparado os valores do total de *bundles* do BNC, fornecidos no arquivo texto, com o processado no aplicativo, obtivemos a equiparação das informações em ambas as fontes, ou seja, 4.423.944 (quatro milhões, quatrocentos e vinte e três mil e novecentos e quarenta e quatro) *lexical bundles*. Em seguida, passamos à importação dos dados, digitando na área de interatividade do aplicativo, uma linha de comando com a devida instrução de obter a freqüência de cada *bundle* recorrente do SMF no banco do BNC. Na figura a seguir, demonstramos essa operação com um exemplo extraído da lista de *bundles* semelhantes do SMF, e sinalizamos a linha de comando e o valor apurado.

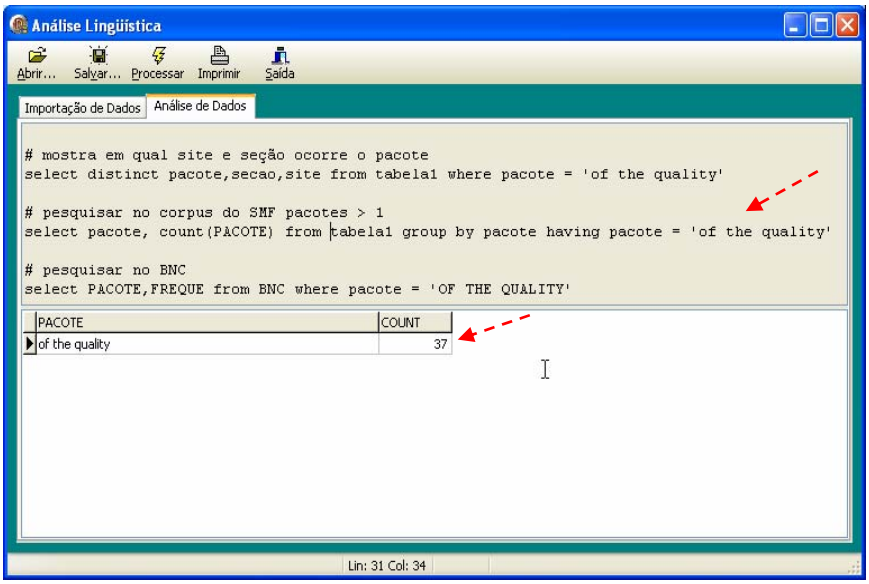
Figura 12. A freqüência de um *bundle* do SMF no BNC.



Conforme demonstrado na Figura 12, o exemplo *of the quality* ocorreu no banco do BNC 309 vezes. Num terceiro passo, para que pudéssemos medir a representatividade do *corpus* de estudo, realizamos a mesma operação para cada *bundle* do SMF, com o critério de corte >1, conforme mencionado anteriormente.

Abaixo, apresentamos uma figura para ilustrar essa operação no *corpus* do SMF.

Figura 13. A freqüência de um *bundle* no *corpus* do SMF.



De acordo com a Figura 13, a ocorrência do *bundle of the quality* em todo o *corpus* do SMF foi de 37 vezes. Assim sendo, esses procedimentos possibilitaram a coleta dos dados necessária para a elaboração de uma tabela, a qual apontasse a probabilidade de recorrência dos *bundles* semelhantes do *corpus* do SMF no *corpus* do BNC. Essa tabela foi elaborada no aplicativo Excel do Windows®, utilizando uma base de cálculo 1.000 (cf. Rocha, 2007 - Capítulo 1 – seção 1.1.2) para comparar os *bundles* semelhantes extraídos do *corpus* do SMF com o *corpus* do BNC, *corpus* de referência escolhido para essa fase de análise (ver resultados Capítulo 3 – seção 3.2.5).

Sobretudo, com relação aos critérios escolhidos para a apuração dos *lexical bundles* do SMF, devemos deixar claro para o leitor, que no momento do processamento eletrônico dos *bundles*, consideramos a seqüência de palavras assim como elas apareceram nos textos, ou seja, não atribuímos ao banco de dados que esse considerasse, quer letras maiúsculas, quer letras minúsculas, e sim que mantivesse a formatação das palavras como nos originais.

Tal procedimento é possível quando se opera com instrumentos de programação, como, por exemplo, o caso de *scripts* computacionais. Para exemplificar, observamos o *corpus* dos *bundles* do BNC, que foi programado para se apresentar em letras maiúsculas (ver Figura 10), diferente da análise de *lexical bundles* desta pesquisa, que foi *case-sensitive*, isto é, na programação foi feita a distinção entre palavras escritas em maiúscula e minúscula.

Essa opção, em manter a extração dos *lexical bundles* do SMF na forma em que essas seqüências se apresentaram no original, por certo, influenciou nos resultados com relação à quantidade dos dados apurados, pois, provavelmente, se tivéssemos optados na programação por *case-insensitive* (todas as letras em maiúsculas), teríamos obtidos uma quantidade maior de *bundles*. Por outro lado, o intuito em adotar esse critério foi o de preservar os nomes próprios característicos do SMF, como, por exemplo, os nomes típicos de departamentos, especificações, processos, entre outros.

Dando seqüência, apresentamos, na próxima sub-seção, a quinta etapa dos procedimentos.

2.4.5 Os *lexical bundles* locais

Para a quinta etapa, utilizamos um procedimento similar à segunda etapa de análise. No entanto, desta vez, buscamos por *bundles* recorrentes em cada localidade, independente das demais, ao contrário da segunda etapa, que buscou *bundles* recorrentes entre as quinze

localidades. O ponto de corte foi o de >1 para a extração dos *bundles* que ocorriam em cada seção de cada localidade.

Vale ressaltar que, a princípio, esse procedimento não foi incorporado à análise dos resultados, porque o objetivo principal foi o de cruzar e identificar os *bundles* semelhantes entre as quinze localidades do SMF. Porém, devido a algumas implicações que sugeririam responder possíveis diferenças encontradas entre os quinze documentos, tais como: as questões relacionadas à cultura organizacional local da empresa; a língua inglesa como idioma oficial (SMF*USA* e SMF*England*); ou ainda, a proximidade geográfica entre as localidades, por exemplo, a localidade *France-FR* (SMF*Annonay*, SMF*Nyon*, SMF*Huningue*), decidimos realizar uma busca por localidade.

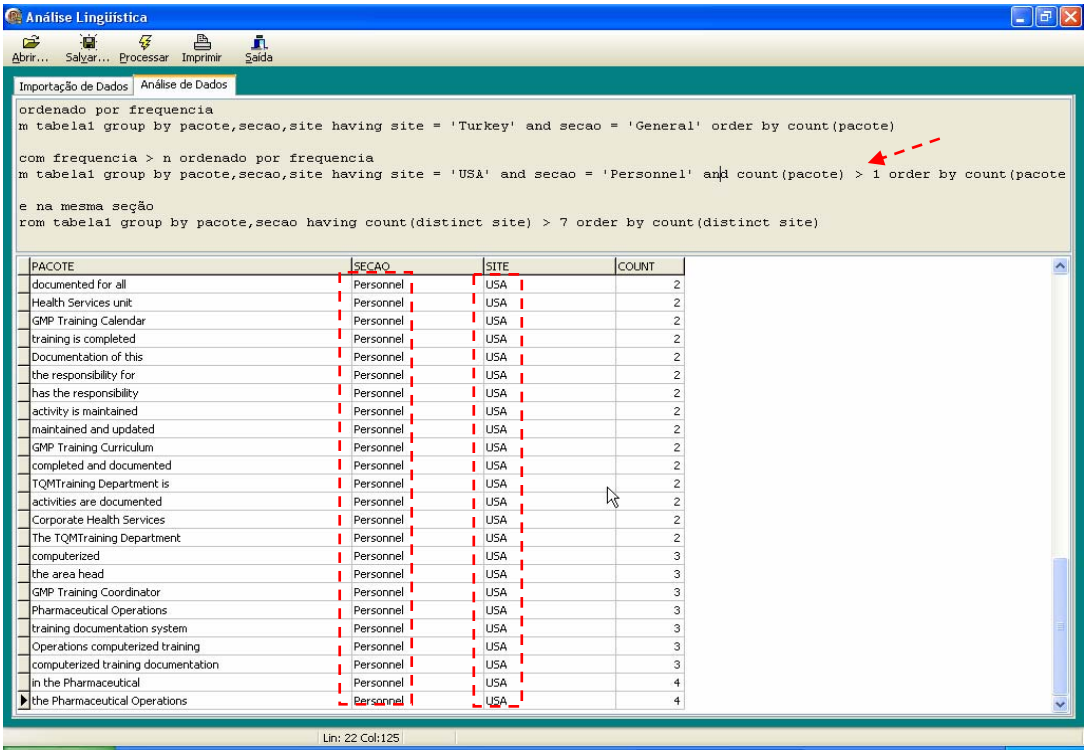
Para isso, percorremos um único passo na busca por *bundles* recorrentes em cada localidade:

1º passo: identificar os *bundles* recorrentes em cada seção de cada localidade.

Para executar esse passo, utilizamos o aplicativo *Análise Lingüística* e perguntamos ao banco de dados, por meio das linhas de comando, qual era a freqüência de *bundles* por seção associada à sua respectiva localidade. Assim, seguindo os critérios de corte adotados, essa freqüência foi ordenada pela recorrência de *bundles* >1.

Na figura a seguir, apresentamos a página *Análise de Dados* do aplicativo *Análise Lingüística*, com o exemplo da seção *Personnel* da localidade *USA*, sinalizando a área dos resultados.

Figura 14. Freqüência dos *bundles* por seção e por localidade.



De acordo com a Figura 14, observamos quatro colunas na área dos resultados: na primeira coluna (da esquerda para a direita), temos o *bundle*; na segunda coluna, o nome da seção; na terceira coluna, o nome da localidade; e na quarta e última coluna, o número de vezes que o *bundle* ocorreu. Observando que todos os *bundles* pertencem a uma mesma seção de um mesmo documento (seção *Personnel* – SMF*USA*). Os resultados desses procedimentos foram a título de contribuição, e de possíveis referências na discussão dos resultados.

2.4.6 As relações coesivas entre as nove seções do SMF

Nesta sexta e última etapa de análise, com a finalidade de responder a nossa terceira pergunta de pesquisa: Como os *lexical bundles* semelhantes refletem o tópico das seções do documento SMF? – passamos a investigar as relações de coesão no documento.

Conforme a exposição sucinta dos tipos de relações coesivas propostos pelos autores Hoey (1991) e Halliday e Hasan (1976, 1989), na visão de Berber Sardinha (1997) e Ramos (1997) (ver capítulo 1 – seção 1.3), o que pudemos abstrair *a priori* foi o tipo de coesão por repetição

simples, pois em nossos achados obtivemos muitos *bundles* que se repetiam igualmente, sem alterações, nos títulos e subtítulos das nove seções do documento.

A fim de analisarmos essa relação, percorremos os seguintes passos:

1º passo: examinar os títulos e subtítulos das seções do SMF;
2º passo: examinar a formação das palavras dos *bundles* semelhantes distribuídos entre as nove seções do documento;
3º passo: examinar a composição gramatical dos *bundles* semelhantes entre as seções;
4º passo: examinar a relação entre os *bundles* semelhantes e os títulos e sub-títulos das seções.

Para proceder aos passos descritos acima, dividimos o conteúdo das nove seções, e classificamos os *bundles* de acordo com as palavras encontradas na formação do *bundle,* ou seja, quais elementos da seqüência de palavras repetiam-se no título ou subtítulos (1 palavra, 2 palavras ou o conjunto das três palavras), indicando que o assunto fazia parte das seções. Em seguida, relacionamos a palavra, ou a palavra de conteúdo do *bundle* com os tópicos das seções, de maneira a estabelecer uma relação de coesão com o assunto tratado.

Tipicamente, a análise mostrou que os *bundles* continham ao menos uma palavra que ligava aos títulos das seções, como é o caso, por exemplo, da palavra *personnel* com o *bundle of key personnel*, encontrado na seção 2 *Personnel*, o qual estabelecia uma relação de coesão por repetição simples com o tópico. Um outro exemplo foi o *bundle short description of* que, em sua formação, exibiu parte do conteúdo de um dos subtítulos da seção 1 *General Information* (ver Capítulo 3 para detalhes).

Dessa forma, examinando a composição dos *bundles* semelhantes, observamos que as seguintes seqüências gramaticais ocorreram:

(i) grupos nominais + preposição, por exemplo: os *bundles*: *Brief description of; short description of*, entre outros;

(ii) preposição + grupo nominal, por exemplo, os *bundles*: *of the firm*; *on the site*; *of the quality*, entre outros;

(iii) verbo de ligação + verbo na passiva + preposição, por exemplo: os *bundles*: *is carried out; is responsible for*; *is described in*, entre outros;

(iv) grupos nominais, por exemplo: os *bundles*: *Personnel hygiene requirements*; *Qualtity Control Department*; *Standard Operating Procedures*, entre outros; e

(v) artigo *the* + grupo nominal, por exemplo: os *bundles*: *the quality management*; *the Quality Control*; *the quality assurance*, entre outros.

Analisando a relação de coesão lexical por repetição com o tópico das seções, e verificando a seqüência gramatical dos *bundles* semelhantes, encontrados nas quinze localidades, realizamos uma análise para responder se esses *bundles* semelhantes, de fato, poderiam dizer sobre o que se tratava o assunto das seções e sub-seções do SMF.

Portanto, apresentada as seis etapas dos procedimentos desta pesquisa, no capítulo a seguir, discutimos e apresentamos a análise dos resultados.

CAPÍTULO 3

APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

No capítulo anterior, apresentamos a Metodologia deste trabalho descrevendo o *corpus* de estudo e as etapas dos procedimentos adotados para a análise dos resultados. Neste capítulo, apresentamos e discutimos os resultados obtidos na apuração e extração dos dados, norteados pelas três seguintes perguntas de pesquisa: (1) Quais são as semelhanças da estrutura organizacional do *Site Master File* entre as quinze localidades em que o documento circula?; (2) Quais são os *lexical bundles* semelhantes que compõem as seções do documento *Site Master File*?; e (3) Como os *lexical bundles* semelhantes refletem os tópicos das seções do documento SMF?

Para discorrer essa análise, dividimos as etapas de apresentação e discussão dos resultados em três seções. Na primeira seção, descrevemos a estrutura organizacional do documento SMF, e discutimos as variações e semelhanças encontradas no sumário (*Index)* dos quinze exemplares. Na segunda seção, apresentamos os resultados apurados da extração dos *lexical bundles* semelhantes entre os quinze exemplares, e a probabilidade de ocorrência desses *bundles* no *corpus* do BNC. Na terceira e última seção, discutimos como os *lexical bundles* semelhantes, encontrados entre os quinze exemplares, estabelecem uma relação coesiva com as nove seções do SMF.

3.1. O sumário do SMF

Conforme mencionado anteriormente, esta primeira etapa de apresentação dos resultados discute e demonstra as semelhanças e as variações encontradas na estrutura organizacional do documento SMF a fim de responder a nossa primeira pergunta de pesquisa - Quais são as semelhanças da estrutura organizacional do *Site Master File* entre as quinze localidades em que o documento circula?

Reiterando o que já foi descrito na seção 2.2 do capítulo Metodologia desta pesquisa, o documento *Site Master File* é um conjunto de textos produzidos pelas indústrias do setor farmacêutico, que circulam em todas as unidades de negócios operacionais da empresa,

instaladas ao redor do mundo. Esse documento tem como objetivos principais: (i) atender às exigências de garantia e qualidade dos medicamentos e produtos fabricados e/ou importados; (ii) e obter a certificação internacional junto aos órgãos governamentais que regulamentam a vigilância sanitária, visando a segurança e saúde do consumidor. Ademais, com a finalidade de atender os procedimentos padrões de qualidade que demandam as inspeções nacionais e internacionais das indústrias farmacêuticas, esse documento deve seguir uma padronização instituída pelo GMP (*Good Manufacturing Practices*) ou BPF (Boas Práticas de Fabricação).

Seguindo as exigências de padronização das Boas Práticas de Fabricação, o documento SMF deve possuir no máximo entre 25 e 35 páginas, incluindo gráficos, tabelas, organogramas, e informações técnicas, distribuídas entre nove seções (ver Quadro 3) que descrevem os seguintes assuntos pertinentes a empresa: (i) informações gerais; (ii) pessoal; (iii) instalações e equipamentos; (iv) documentação; (v) produção; (vi) controle de qualidade; (vii) contratos de fabricação e análise; (viii) distribuição, reclamações ou *recall* do produto; e (ix) auto-inspeção.

Assim, buscando identificar essa padronização, manualmente, analisamos os sumários dos quinze exemplares para verificar se as partes que continham cada documento apresentavam semelhanças.

Além do mais, as localidades participantes do *corpus* desta pesquisa estão referenciadas neste trabalho como: Alemanha (*Germany*); Áustria (*Austria*); Brasil (*Brazil*); Estados Unidos (*USA*); França (*FR*) – com três localidades distintas, *AnnonayFR*, *NyonFR* e *HuningueFR*; Holanda (*The Netherlands*); Inglaterra (*England*); Itália (*Italy*); Japão (*Japan*); Porto Rico (*Puerto Rico*), Suíça (*SWIT*) – com duas localidades, *BaselSWIT* e *HettlingenSWIT*; e Turquia (*Turkey*). Nessa ordem, foram doze países participantes, tendo a França com três localidades e a Suíça com duas localidades farmacêuticas, totalizando quinze localidades.

Para exemplificar, elaboramos e apresentamos a seguir, um quadro com as informações, tais quais elas foram encontradas de forma semelhante nos quinze sumários do SMF, divididas em noves seções com seus títulos e sub-títulos, obedecendo a uma mesma seqüência numérica, conforme padronização exigida.

Quadro 5. Exemplo do sumário do documento *Site Master File*.

SEÇÃO	TÍTULO	SUB-TÍTULO
1	General Information	1. Brief information on the company; 2. Any other manufacturing activities carried on Site; 3. Name and exact address of Site; 4. Type of actual products manufactured on the Site and information about specifically toxic or hazardous substances handled, mentioning the way they are manufactured (in dedicated facilities or on a campaign basis); 5. Short Description of the Site; 6. Number of employees engaged in Quality Assurance, Quality Control, Production, Storage and Distribution; 7. Use of outside assistance for manufacturing; 8. Short description of the quality management system of the firm responsible for manufacture.
2	Personnel	1. Organization chart; 2. Qualification, experience and responsibilities of key personnel; 3. Outline of arrangement for basic and in-service training; 4. Health requirements for personnel engaged in Production; 5. Personnel hygiene requirements including clothing.
3	Premises and Equipment	1. Site Plan; 2. Nature of construction and finishes; 3. Brief Description of Ventilation systems; 4. Special areas for the handling of highly toxic, hazardous and sentisizing materials; 5. Brief description of the water system; 6. Maintenance; 7. Description of major production and control laboratory equipment; 8. Maintenance and Servicing; 9. Qualification, Validation and Calibration; 10. Sanitation.
4	Documentation	1. Arrangements for the preparation, revision and distribution of necessary documentation for manufacturing; 2. Other Documentation related to product quality.
5.	Production	1. Description of Production Operations; 2. Arrangements for Handling Starting Materials, Packaging Materials, Bulk and Finished products including Sampling, Quarantine, Release and Storage; 3. Arrangements for Reprocessing or Rework; 4. Arrangements for Handling Reject Materials and Products; 5. General Policy for Process Validation.
6	Quality Control	1. Activities of Quality Control Department.
7	Contract Manufacturing and Analysis	1. Details of Technical contract between the Contract Giver and Acceptor.
8	Distribution, Complaints and Product Recall	1. Description of Storage and Distribution Practices; 2. Records of Distribution.
9	Self Inspection	1. Short Description of the Self Inspection System.

Com base no Quadro 5, observamos que todos os sumários das quinze localidades pesquisadas, apresentaram a mesma estrutura organizacional, ou seja, nove seções com o mesmo título na mesma ordem seqüencial e numérica. Uma outra semelhança encontrada foi em relação às sub-seções dos sumários. Verificamos que todos os quinze sumários apresentaram os mesmos sub-títulos, com o mesmo conteúdo e a mesma ordem seqüencial e numérica. Portanto, de acordo com as exigências estabelecidas pelas Boas Práticas de Fabricação (BPF) na elaboração do SMF, e baseados nos sumários, podemos supor que todas as quinze localidades estavam padronizadas quanto ao conteúdo das informações.

Entretanto, a mesma semelhança não foi identificada nos itens que compõem as sub-seções dos sumários comparados. Nos quinze documentos analisados, foram encontradas as seguintes variações nas sub-seções dos sumários:

(i) inexistência de itens nas sub-seções;
(ii) diferentes quantidades de itens elencados;
(iii) itens com informações diferentes.

Com relação à inexistência de itens que compõem as sub-seções dos sumários, essa foi encontrada em cinco dos quinze documentos: SMF*England*; SMF*Italy*; SMF*Japan*; SMF*Netherlands*; e SMF*USA*, isto é, identificamos entre esses cinco documentos somente a estrutura exemplificada no Quadro 5.

Dessa forma, dos quinze documentos analisados, verificamos que dez dos SMFs apresentaram itens nas sub-seções dos seus respectivos sumários, os quais foram: o SMF*Puerto;* o SMF*Germany*; o SMF*AnnonayFR*; o SMF*HettlingenSWIT*; o SMF*Brazil;* o SMF*BaselSWIT;* o SMF*NyonFR;* o SMF*HuningueFR;* o SMF*Austria*; e o SMF*Turkey*.

Quanto à quantidade de itens elencados, partindo da comparação dos dez sumários SMFs que apresentaram itens nas sub-seções, observamos que tanto o número de itens, quanto o conteúdo, divergiam entre as localidades. Para ilustrar essas variações, utilizamos alguns dos exemplos encontrados nas localidades com a finalidade de elucidar a nossa discussão.

Por exemplo, a sub-seção 1 (*Brief information on the company*) da seção 1 (*General Information*) do SMF*Puerto*, continha quatro itens elencados (1.1 à 1.4). Enquanto que o SMF*Italy* continha três itens (1.1 à 1.3), e os itens de cada sub-seção continham informações diferentes, conforme apresentamos no quadro a seguir:

Quadro 6. Os itens elencados nas sub-seções dos sumários dos SMF*Puerto* e SMF*Italy*

Documento		Seção		Sub-Seção		Itens
SMF*Puerto*	1	General Information	1	Brief information on the company	1.1	In the world
					1.2	Pharmaceutical Products
					1.3	Consumer Health
					1.4	Animal Health
SMF*Italy*	1	General Information	1	Brief information on the company	1.1	Name and address
					1.2	Relation to other sites
					1.3	Any information relevant to understand the manufacturing operations.

Podemos observar no quadro 6, que a quantidade de itens encontrados na sub-seção do sumário SMF*Puerto* (4 itens) não foi a mesma com relação a quantidade encontrada no SMF*Italy* (3 itens). Ademais, cada uma das sub-seções apresentaram conteúdos diferentes, conforme ilustram os títulos.

Tomemos um outro exemplo, os SMF*Turkey* e SMF*HettlingenSWIT*, que além da quantidade dos itens apresentarem variações entre as localidades, alguns conteúdos apresentaram variações na maneira de expressar as informações. No quadro abaixo, apresentamos o exemplo da Seção 4 (*Documentation*) na sub-seção 2 (*Other Documentation related to product quality*).

Quadro 7. Os itens elencados nas sub-seções dos sumários dos SMF*Turkey* e SMF*HettlingenSWIT*

Documento	Seção		Sub-Seção		Itens	
SMF*Turkey*	4	Documentation	2	Other Documentation related to product quality	2.1	Equipment specifications
					2.2	Specifications for disposables, e.g. cleaning materials
					2.3	Standard Operating Procedures
					2.4	Quality Control Procedures
					2.5	Training procedures
					2.6	Computer program specifications
					2.7	Documentation control of process deviations
					2.8	Calibration and test documents
					2.9	Validation documents
					2.10	Reconciliation of batches of materials
					2.11	Additional Standard Documentation
Documento	Seção		Sub-Seção		2.1	Instructional Documents
SMF *HettlingenSWIT*	4	Documentation	2	Other Documentation related to product quality	2.2	Quality Modules Documents
					2.3	Standard Operating Procedures (SOP)
					2.4	Training Modules Documents
					2.5	Specifications Documents
					2.6	Validation Master Plan (VMP)
					2.7	Protocols
					2.8	Technical Agreements

De acordo com o quadro 7, a maioria dos itens apresentou conteúdos diferentes entre as localidades, e somente identificamos uma única semelhança entre eles - o item 2.3 (*Standard Operating Procedures*). As demais expressões apresentaram-se de modo diferente. Pudemos observar, assim, que em ambas as localidades, muitas das expressões carregavam a palavra *documents* (documentos) nos títulos, como, por exemplo, o SMF*Turkey*, com o item 2.8 (*Calibration and test documents*), e o item 2.9 (*Validation documents*). Ainda, tivemos o exemplo do SMF*HettlingenSWIT*, com os itens 2.1 (*Instructional Documents*), 2.2 (*Quality Modules Documents*), 2.3 (*Training Modules Documents*), entre outros, indicando que esses itens relacionavam-se ao assunto da seção, ou seja, a documentação da empresa (*Documentation*).

No entanto, mesmo tendo identificado essa relação, por meio da repetição da palavra *documents*, nos itens de ambas as localidades essas expressões não carregavam significados semelhantes. Portanto, pudemos concluir que as informações contidas nas duas localidades eram diferentes.

Além disso, verificamos itens com maior nível de detalhamento, isto é, um número maior de palavras para compor os títulos das sub-seções, como, por exemplo, o SMF*Turkey* com os itens: 2.2 (*Specifications for disposables, e.g. cleaning materials*); e 2.10 (*Reconciliation of batches of materials*). Também, tivemos o exemplo do SMF*Italy*, com o item 1.3 (*Any information relevant to understand the manufacturing operations*). Por conseguinte, notamos que esses detalhamentos não foram identificados no SMF*HettlingenSWIT* e no SMF*Puerto*.

3.1.2 Discussão

Em síntese, diante dos resultados apurados quanto à verificação das seções/ sub-seções e itens que compõem a estrutura organizacional do SMF, pudemos identificar as seguintes características encontradas nos sumários dos quinze exemplares investigados:

(1) As seções e sub-seções dos quinze documentos não se modificam. Todas elas possuem títulos e subtítulos iguais, com a mesma seqüência numérica, indicando que a padronização das informações exigidas pelos órgãos competentes na elaboração do documento é assegurada. Portanto, podemos dizer que a estrutura organizacional do documento é prescrita e obedecida.

(2) Os itens das sub-seções variam de acordo com a localidade em três principais aspectos:

(i) na inexistência de itens, pois cinco das localidades não apresentaram itens nas sub-seções;

(ii) na quantidade de itens elencados;

(iii) nas variações das escolhas lingüísticas que compõem os títulos e sub-títulos, ora as expressões são as mesmas, por exemplo, o item 2.3 (*Standard Operating Procedures*) da seção 4 (*Documentation*), que ocorreu nas localidades SMF*Turkey* e SMF *HettlingenSWIT*, ora as expressões apresentam um maior número de palavras para compor as informações, por exemplo, os itens 1.3 (*Any information relevant to understand the manufacturing operations*) da seção 1 (*General Information*) do SMF*Italy*, e 2.2 (*Specifications for disposables, e.g. cleaning materials)* da seção 4 (*Documentation*) do SMF*Turkey*.

Logo, podemos supor que as variações encontradas nos itens das sub-seções, demonstram as escolhas feitas por cada localidade, no sentido de organizar a sua escrita no documento, ou então, demonstram as necessidades locais onde a fábrica está instalada. Isso pode significar, que essas variações indicam a opção de determinadas localidades em expressar as informações de maneira mais detalhada na descrição de suas atividades, por exemplo, o item 2.11 (*Additional Standard Documentation*) da seção 4 (*Documentation*) do SMF*Turkey*. Por outro lado, outras localidades optaram em organizar a sua escrita de forma mais resumida, por exemplo, o item 2.5 (*Specifications Documents*) da seção 4 (*Documentation*) do SMF*HettlingenSWIT* .

Dessa forma, supomos que a indústria farmacêutica, objeto de pesquisa deste estudo, mesmo sendo uma multinacional, compartilhando a mesma missão junto aos seus clientes e fornecedores globais, e tendo que cumprir regulamentos para atender legislações locais e internacionais do setor, deve sofrer variações de cunho situacional, ou seja, provavelmente, em cada localidade, faz-se necessário uma adequação quanto aos seguintes aspectos: ao tipo de instalações ou equipamentos; ao tipo de produto fabricado; à logística de distribuição na importação ou exportação dos insumos farmacêuticos; à contratação ou terceirização de serviços especializados; entre outros.

Posto que todas essas informações, contidas no documento SMF, servirão de base para a certificação de qualidade de fabricação do produto, as variações apresentadas nos itens dos sumários podem indicar ainda, que determinadas localidades precisam apresentar informações mais complementares, ou não, devido a alguns fatores:

(i) às normas de legislação local;
(ii) ao tipo de produção, instalações ou equipamentos; e
(iii) à cultura organizacional, ou seja, a forma como o pessoal da organização se comporta no desempenho das suas funções com relação aos sistemas de valores e normas vigentes, que poderá distinguir uma localidade de outra.

Portanto, podemos dizer que o conteúdo apresentado no sumário do SMF segue o mesmo padrão, ou seja, expressa ou realiza os mesmo tópicos, "moldados" para atender os princípios globais, isto é, as diretrizes internacionais que regulamentam a padronização das operações industriais. Já os conteúdos encontrados nos itens das sub-seções são variáveis, porquanto podem estar submetidos a *priori* a condições locais, à legislação local, ou ao tipo de produção, o que os caracterizam com peculiaridades de escrita diferentes dos demais.

Enfim, acreditamos ter oferecido por meio desta discussão, e dos exemplos extraídos dos quinze exemplares coletados para o *corpus* desta pesquisa, elementos para responder a nossa primeira pergunta de pesquisa – Quais são as semelhanças da estrutura organizacional do *Site Master File* entre as quinze localidades em que o documento circula?

Na seção seguinte, passamos à segunda etapa de apresentação dos resultados.

3.2 Os padrões léxico-gramaticais

Nesta segunda etapa dos resultados, buscamos identificar padrões léxico-gramaticais recorrentes nas quinze diferentes localidades do SMF, de tal forma que pudéssemos responder a nossa segunda pergunta de pesquisa - Quais são os *lexical bundles* semelhantes entre as seções do documento *Site Master File*? Decidimos investigar a recorrência de *lexical bundles* entre as nove seções dos documentos para descobrir quais eram os *lexical bundles* comuns para as quinze localidades.

Lembramos que o *corpus* deste trabalho, diferentemente do que pudemos encontrar na literatura pesquisada sobre *lexical bundles* (exceto Berber Sardinha, 2003), foi compilado a

partir de um único tipo de texto (o SMF), originário de um domínio específico do contexto de negócio de uma indústria do setor farmacêutico.

A busca por esses padrões foi devida à hipótese de que: se o documento SMF apresentava-se semelhante em seus títulos e sub-títulos, dentro da sua estrutura organizacional, pressupomos, então, que essas semelhanças também poderiam ocorrer nas divisões internas do documento, no seu nível léxico-gramatical. Dessa maneira, partimos de dados quantitativos, processados por meio de ferramentas computacionais, na tentativa de confirmar a nossa hipótese.

Ressaltamos, conforme dito no capítulo introdutório desta pesquisa, que neste estudo não pretendemos descrever os *lexical bundles*. Fizemos uso dessa categoria com o intuito de apresentar como os quinze exemplares do documento assemelham-se entre as localidades em seu formato e conteúdo, ou seja, verificar qual grau de conformidade com elementos lingüísticos, um documento estabelecido e padronizado por instituições competentes no seu contexto de atuação, escrito por diferentes autores em diferentes partes do mundo pode atingir. No entanto, futuras pesquisas podem valer-se desses pressupostos para exploração de uma descrição gramatical em língua inglesa de *corpora* específicos.

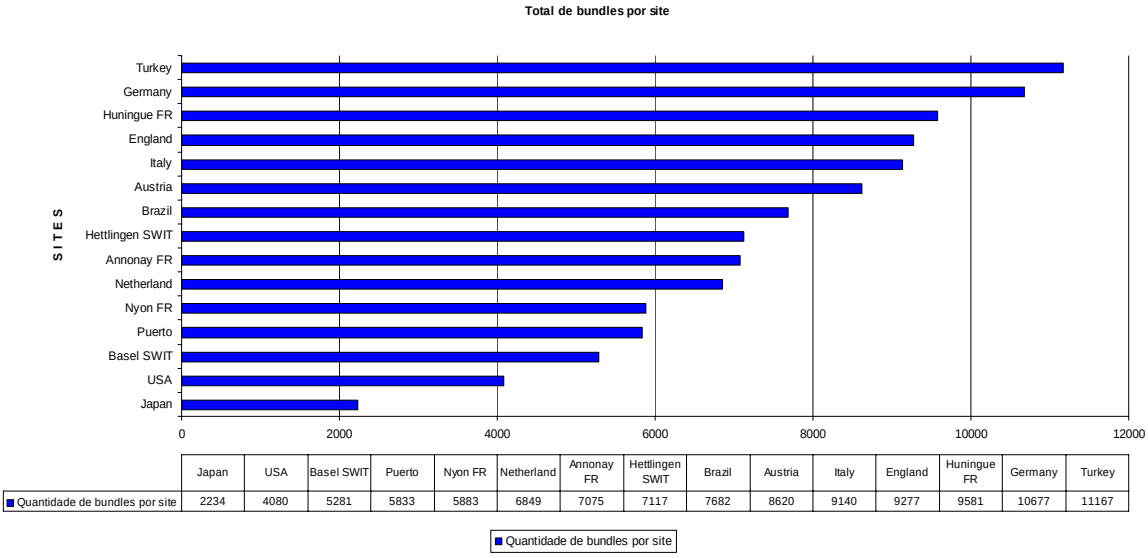
Desse modo, percorrendo quatro fases para atingir os resultados desta análise, conforme descritas nos procedimentos (ver Capítulo 2 – seção 2.4.2), apresentamos nas seções seguintes a discussão e detalhamento dos dados.

3.2.1 A contagem dos *bundles* por localidade

Nessa primeira fase, por meio de um estudo quantitativo, procuramos identificar a ocorrência de *lexical bundles* em cada uma das quinze localidades do SMF. O propósito nesta fase, conforme já dito, foi demonstrar a variação entre as localidades quanto à quantidade de *lexical bundles*.

A fim de apresentar o *ranking* referente à quantidade de *bundles* encontrados *versus* as localidades, no gráfico a seguir apresentamos os resultados.

Gráfico 1. Quantidade de *lexical bundles* por localidade.



	Japan	USA	Basel SWIT	Puerto	Nyon FR	Netherland	Annonay FR	Hettlingen SWIT	Brazil	Austria	Italy	England	Huningue FR	Germany	Turkey
Quantidade de bundles por site	2234	4080	5281	5833	5883	6849	7075	7117	7682	8620	9140	9277	9581	10677	11167

Observando o Gráfico 1, em ordem decrescente, a 1ª posição no *ranking* por maior quantidade de *bundles* encontrados foi dada ao SMF*Turkey*, seguida pelas demais localidades, tendo como 15ª e última posição, o SMF*Japan*. Podemos observar, abaixo das barras de progressão do gráfico, uma legenda apresentando o total de *bundles* de cada localidade. Por exemplo, da direita para a esquerda é apresentado o total de 11.167 *lexical bundles* para o SMF*Turkey*, e assim por diante.

Nesse mesmo *ranking*, por total de *bundles* encontrados, tivemos o seguinte: o SMF*Turkey* em 1ª posição; o SMF*Germany* em 2ª posição; o SMF*HuningueFR* em 3ª posição; o SMF*Japan* em 15ª posição, e assim de modo decrescente para os demais. No entanto, essa mesma ordem não ocorreu quando examinadas as nove seções, independentes uma das outras. Para exemplificar, elaboramos uma seqüência de nove gráficos, um para cada seção, e os comparamos com o *ranking* apresentado no Gráfico 1, com o fim de projetarmos as diferenças encontradas entre as seções das quinze localidades.

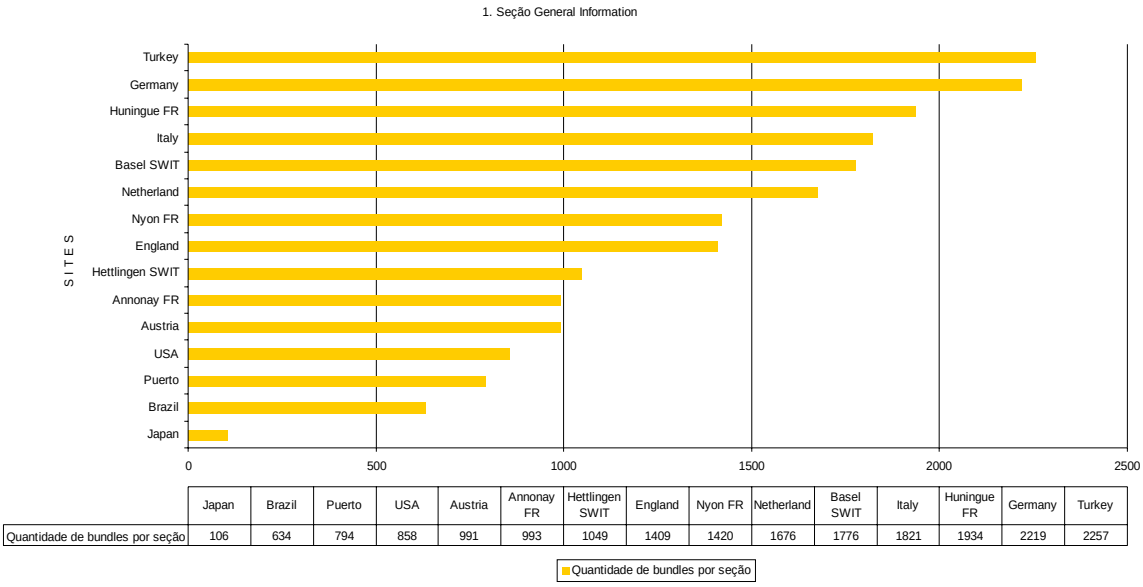
Na sub-seção seguinte, demonstramos os gráficos com a contagem dos *lexical bundles* encontrados em cada uma das seções dos SMFs.

3.2.2 A contagem dos *bundles* por seção

Na segunda fase desta etapa, desmembramos cada seção que compõe a estrutura organizacional do SMF (ver Quadro 3) para realizar a contagem do total de *bundles* por seção, e comparar os totais apurados entre as quinze localidades, com a finalidade de investigar

possíveis variações. Numa seqüência de nove gráficos, apresentamos, primeiramente, os valores encontrados na seção 1 *General Information.*

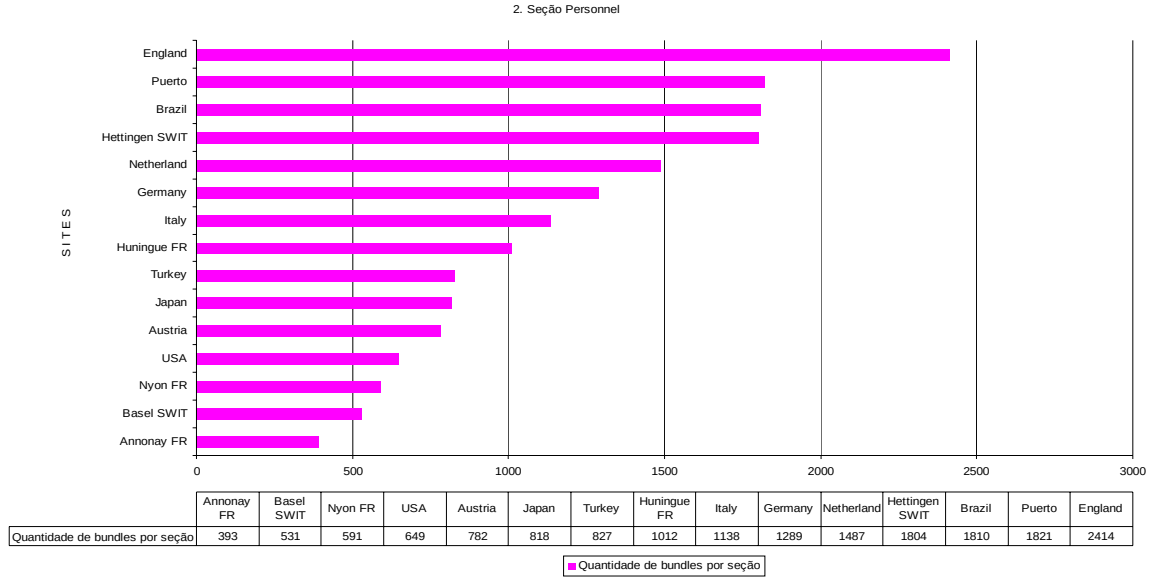
Gráfico 2: Quantidade de *lexical bundles* da seção 1 *General Information*.



	Japan	Brazil	Puerto	USA	Austria	Annonay FR	Hettlingen SWIT	England	Nyon FR	Netherland	Basel SWIT	Italy	Huningue FR	Germany	Turkey
Quantidade de bundles por seção	106	634	794	858	991	993	1049	1409	1420	1676	1776	1821	1934	2219	2257

Nos dados apresentados no Gráfico 2, observamos que quatro das localidades investigadas (SMF*Turkey*, SMF*Germany*, SMF*HuningueFR*, e SMF*Japan*) ocuparam a mesma posição, comparadas ao *ranking* representado no Gráfico 1 (1º, 2º, 3º e 15º lugares), já as seções das demais localidades, alteraram a ordem por maior número de *bundles* quando comparadas ao Gráfico 1. Na seqüência, apresentamos a seção 2 *Personnel.*

Gráfico 3: Quantidade de *lexical bundles* da seção 2 *Personnel*.

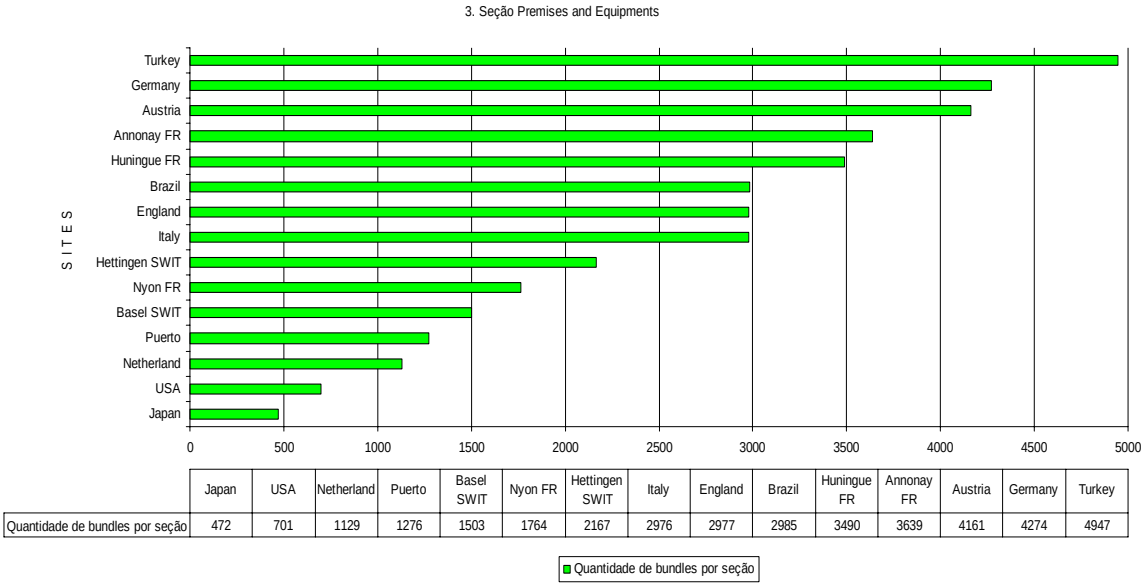


	Annonay FR	Basel SWIT	Nyon FR	USA	Austria	Japan	Turkey	Huningue FR	Italy	Germany	Netherland	Hettingen SWIT	Brazil	Puerto	England
Quantidade de bundles por seção	393	531	591	649	782	818	827	1012	1138	1289	1487	1804	1810	1821	2414

Verificamos na seção 2 *Personnel*, que a localidade que ocupou a 1ª posição em maior número de *bundles* foi o SMF*England*, seguida do SMF*PuertoRico* e SMF*Brazil*, ao passo que o SMF*Turkey* ocupou a 9ª posição, e o SMF*Japan* continuou em 15ª e última posição.

No próximo gráfico, apresentamos a seção 3 *Premises and Equipment.*

Gráfico 4. Quantidade de *lexical bundles* da seção 3 *Premises and Equipment*.

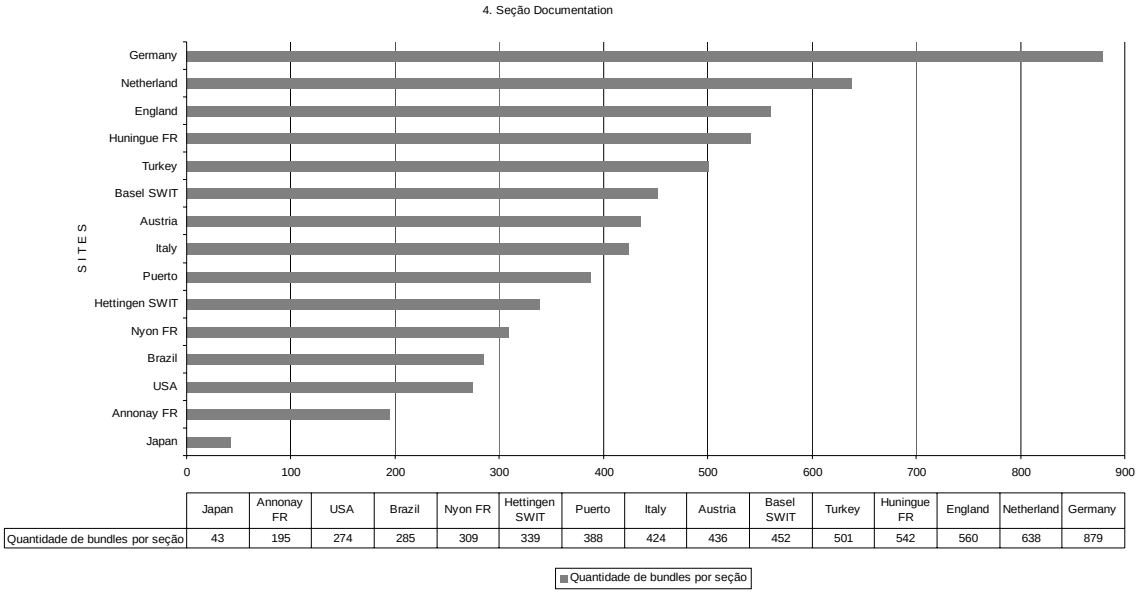


	Japan	USA	Netherland	Puerto	Basel SWIT	Nyon FR	Hettingen SWIT	Italy	England	Brazil	Huningue FR	Annonay FR	Austria	Germany	Turkey
Quantidade de bundles por seção	472	701	1129	1276	1503	1764	2167	2976	2977	2985	3490	3639	4161	4274	4947

Na seção 3 *Premises* and *Equipment*, os SMF*Turkey*, SMF*Germany* e SMF*Japan*, ocuparam as 1ª, 2ª e 15ª posições, da mesma forma como foi apresentado no *ranking* do Gráfico 1. Porém, as demais localidades sofreram variações em suas posições por maior número de *bundles* encontrados entre os quinze documentos.

Dando prosseguimento à apresentação dos gráficos, a seguir, ilustramos a seção 4 *Documentation*.

Gráfico 5. Quantidade de *lexical bundles* da seção 4 *Documentation*.

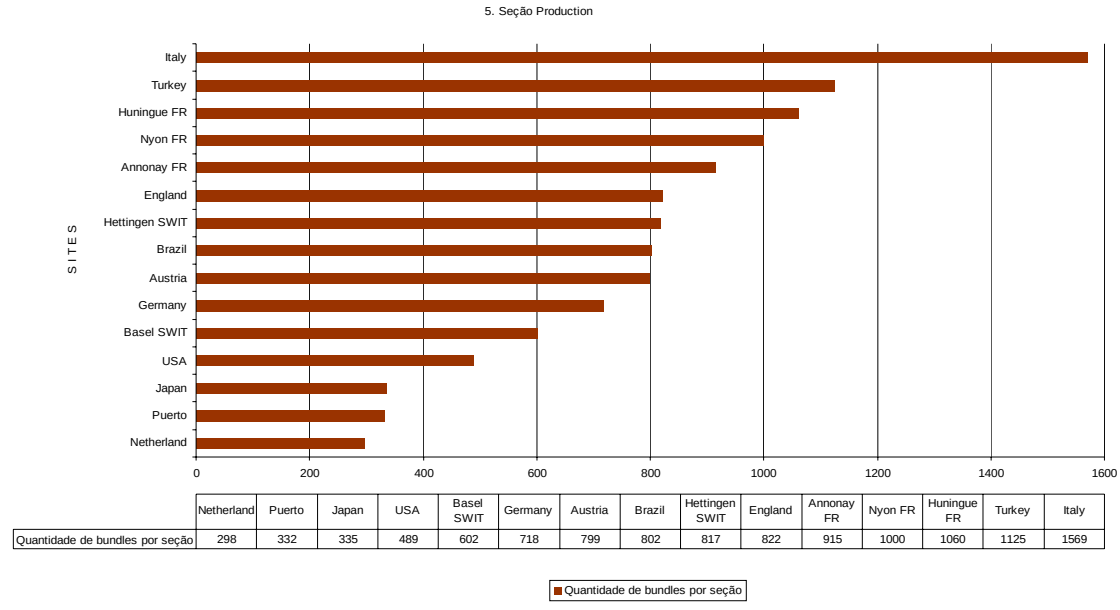


	Japan	Annonay FR	USA	Brazil	Nyon FR	Hettingen SWIT	Puerto	Italy	Austria	Basel SWIT	Turkey	Huningue FR	England	Netherland	Germany
Quantidade de bundles por seção	43	195	274	285	309	339	388	424	436	452	501	542	560	638	879

Na seção 4 *Documentation*, as localidades SMF*Germany*, SMF*Netherlands* e SMF*England*, ocuparam as três primeiras posições no *ranking*, de maneira diferente como ocorreu no *ranking* do total de *bundles* por localidade (Gráfico 1). Tivemos nessa seção o SMF*Turkey* em 5ª posição por total de *bundles*, e a localidade SMF*Japan* mantendo a 15ª e última posição no *ranking*.

A seguir, apresentamos o gráfico representando a seção 5 *Production*.

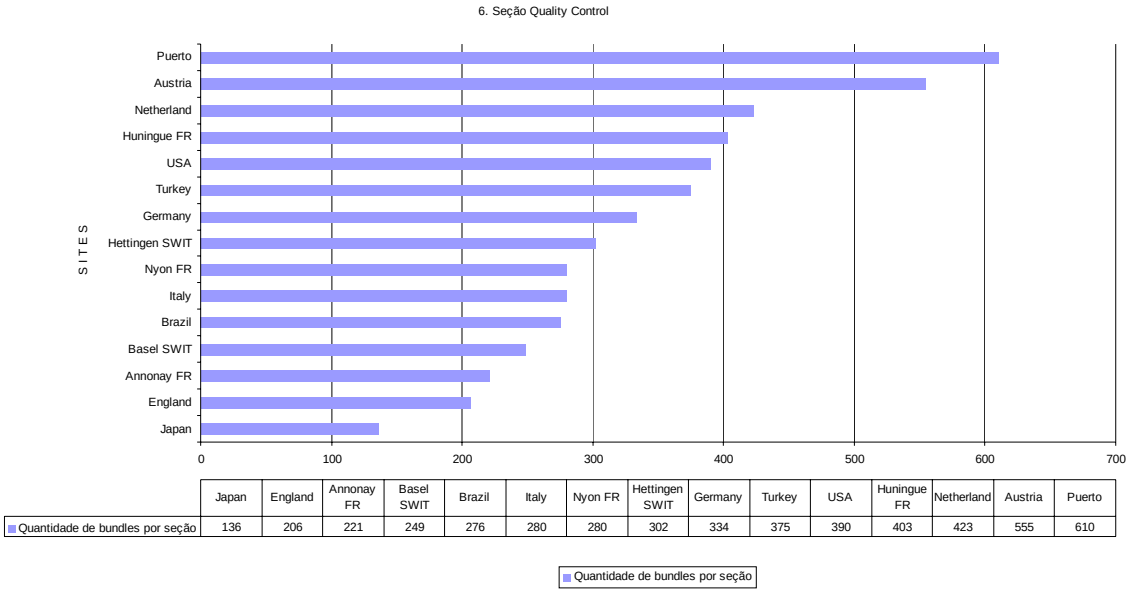
Gráfico 6. Quantidade de *lexical bundles* da seção 5 *Production.*



	Netherland	Puerto	Japan	USA	Basel SWIT	Germany	Austria	Brazil	Hettingen SWIT	England	Annonay FR	Nyon FR	Huningue FR	Turkey	Italy
Quantidade de bundles por seção	298	332	335	489	602	718	799	802	817	822	915	1000	1060	1125	1569

No Gráfico 6, as localidades SMF*Turkey* e SMF*HuningueFR* permaneceram entre as três primeiras posições, em comparação ao *ranking* do Gráfico 1. No entanto, nessa seção, o SMF*Turkey* posicionou-se em 2º lugar por maior número de *bundles*, e o SMF*Italy* encontrou-se posicionado na 1ª posição. Já o SMF*Japan* ocupou a 13ª posição, e o SMF*Netherlands* a 15ª posição. Na seqüência, apresentamos a seção 6 *Quality Control.*

Gráfico 7. Quantidade de *lexical bundles* da seção 6 *Quality Control.*

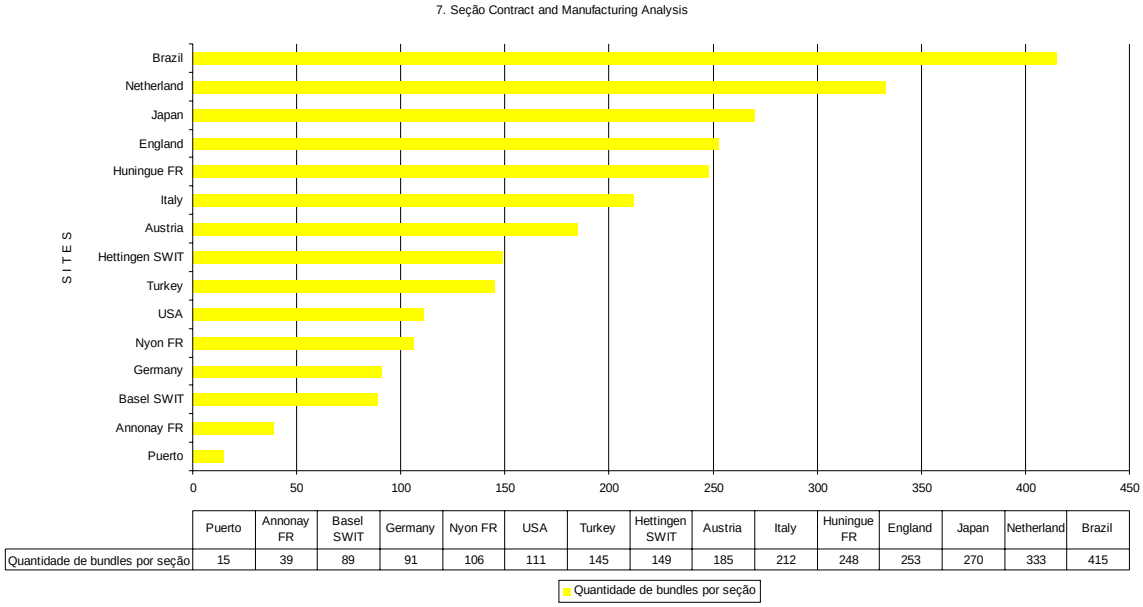


	Japan	England	Annonay FR	Basel SWIT	Brazil	Italy	Nyon FR	Hettingen SWIT	Germany	Turkey	USA	Huningue FR	Netherland	Austria	Puerto
Quantidade de bundles por seção	136	206	221	249	276	280	280	302	334	375	390	403	423	555	610

Na Seção 6 *Quality Control*, a localidade SMF*PuertoRico* posicionou-se em 1º lugar com maior número de *bundles* encontrados entre os quinze documentos SMF, seguido do SMF*Austria* e o do SMF*Netherlands*.

Porém, o *ranking* dessa seção, diferente do *ranking* apresentado no Gráfico 1, apresentou o SMF*Turkey* em 6ª posição, porém, permaneceu com o SMF*Japan* em 15ª e última posição. Em seguida, apresentamos o gráfico da seção 7 *Contract and Manufacturing Analysis.*

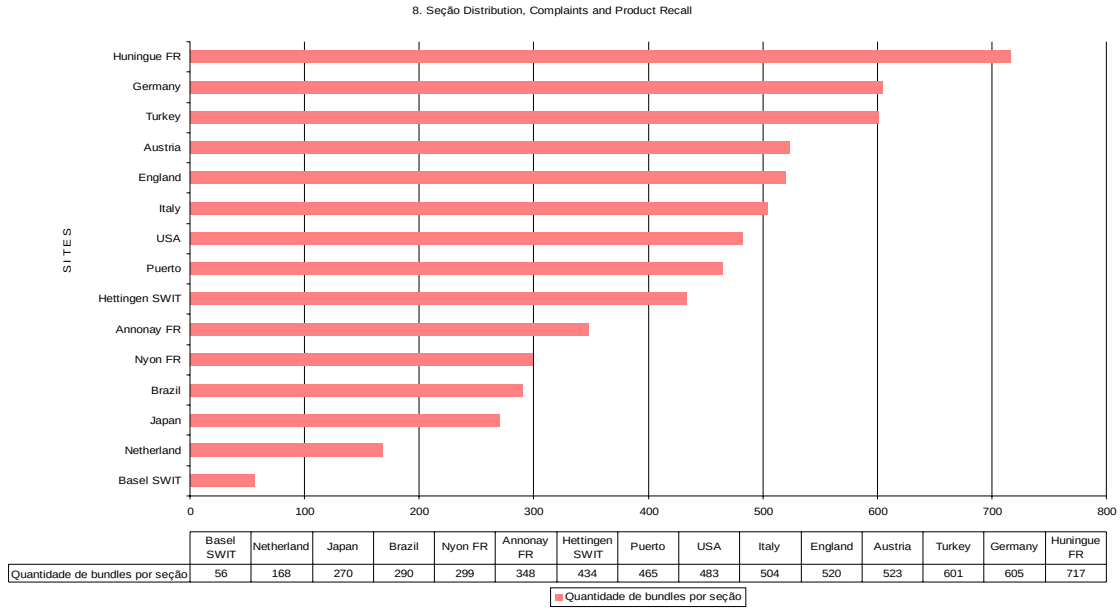
Gráfico 8. Quantidade de *lexical bundles* da seção 7 *Contract and Manufacturing Analysis.*



	Puerto	Annonay FR	Basel SWIT	Germany	Nyon FR	USA	Turkey	Hettingen SWIT	Austria	Italy	Huningue FR	England	Japan	Netherland	Brazil
Quantidade de bundles por seção	15	39	89	91	106	111	145	149	185	212	248	253	270	333	415

Na seção 7 *Contract and Manufacturing Analysis*, o SMF*Brazil* foi a localidade que apresentou o maior número de *bundles*, enquanto que *o* SMF*Netherlands* ocupou a 2ª posição, e a localidade SMF*Japan* a 3ª posição. Durante essa fase de comparação entre as seções das localidades, foi a primeira vez em que o SMF*Japan* apresentou um número maior de *bundles* com relação às seções anteriores. Logo, o SMF*Turkey* ficou em 9ª posição, da mesma forma que na seção *Personnel*. A seguir, apresentamos o gráfico da seção 8 *Distribution, Complaints and Product Recall.*

Gráfico 9. Quantidade de *lexical bundles* da seção 8 *Distribution, Complaints and Product Recall.*

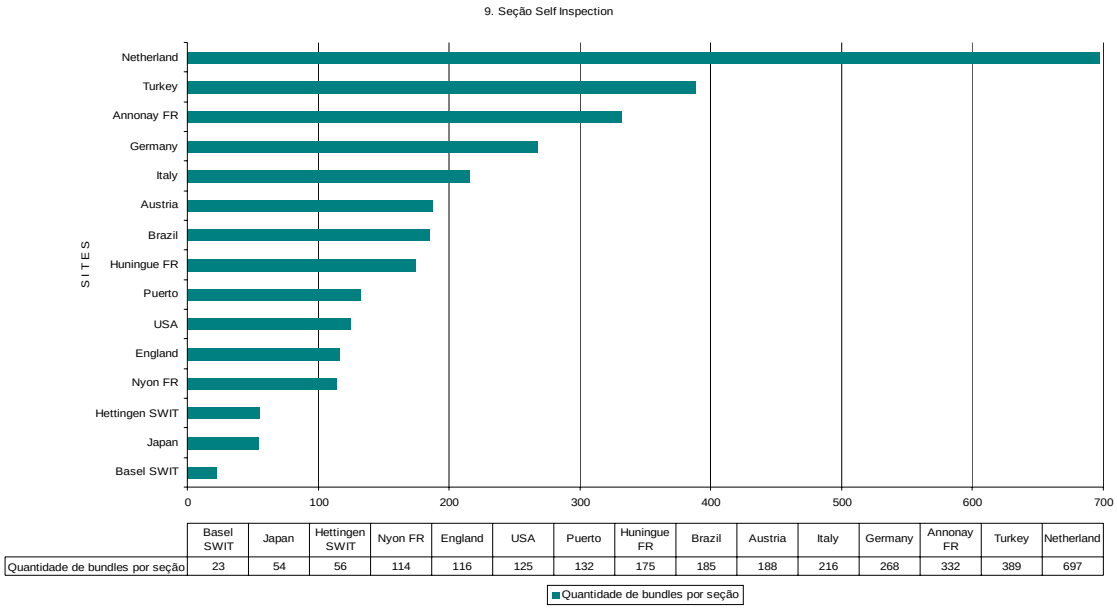


	Basel SWIT	Netherland	Japan	Brazil	Nyon FR	Annonay FR	Hettingen SWIT	Puerto	USA	Italy	England	Austria	Turkey	Germany	Huningue FR
Quantidade de bundles por seção	56	168	270	290	299	348	434	465	483	504	520	523	601	605	717

Na seção 8 *Distribution, Complaints and Product Recall*, os SMF*HuningueFR*, SMF*Germany*, e o SMF*Turkey* apresentaram-se como as três localidades com maior número de *bundles* nessa seção, do mesmo modo como ocorreu no *ranking* do Gráfico 1. Já o SMF*Japan* ocupou a 13ª posição, igualmente como ocorreu na seção 5 *Production*.

Para finalizar esta seqüência de apresentação dos nove gráficos correspondentes às seções do SMF, abaixo apresentamos a seção 9 *Self Inspection*.

Gráfico 10. Quantidade de *lexical bundles* da seção 9 *Self Inspection*.



	Basel SWIT	Japan	Hettingen SWIT	Nyon FR	England	USA	Puerto	Huningue FR	Brazil	Austria	Italy	Germany	Annonay FR	Turkey	Netherland
Quantidade de bundles por seção	23	54	56	114	116	125	132	175	185	188	216	268	332	389	697

Na seção 9 *Self Inspection*, as localidades: SMF*Netherlands* ocupou a 1ª posição; SMF*Turkey* a 2ª posição; SMF*AnnonayFR* a 3ª posição; e SMF*Japan* a 14ª posição, diferentes posições de *ranking*, quando comparadas ao Gráfico 1.

Em suma, demonstrados os *rankings* da quantidade de *bundles* por localidade e por seção, identificamos que, em total de *bundles* (ver Gráfico 1), o SMF*Turkey* foi o que apresentou o maior número de *bundles*, e o SMF*Japan* o menor número entre os quinze exemplares. Entretanto, na investigação da quantidade de *bundles* de cada localidade entre as nove seções (ver Gráficos 2 a 10), o SMF*Turkey* não foi a localidade que apresentou o maior número de *bundles*, bem como o SMF*Japan* não foi o que apresentou o menor número de *bundles*. Portanto, esse *ranking* representando as variações da quantidade de *bundles* entre as seções das quinze localidades, indicou que o volume de informações por seções se distribui diferentemente entre as localidades, reiterando os resultados da análise do sumário.

Assim, com base na contagem total dos *bundles* extraídos do *corpus* de pesquisa, na sub-seção seguinte discutimos os pressupostos dessas duas primeiras fases.

3.2.3 Discussão da contagem dos *bundles*

Conforme resultados apresentados, quanto à contagem dos *bundles* por localidade e por seção, concluímos que independentemente da quantidade total de *bundles* que determinada localidade apresente, a quantidade de *bundles* em cada uma das nove seções pode variar. Por exemplo, tomemos o SMF*Turkey*, embora essa localidade apresente uma quantidade total de *bundles* maior entre as demais (11.167 - ver Gráfico 1), se comparada às suas nove seções com as das outras localidades, o SMF*Turkey* não se apresenta sempre em primeiro lugar com maior número de *bundles*.

Exemplificando, observados os *rankings* da seqüência dos gráficos anteriores, tivemos o SMF*Turkey* nas seguintes posições:

1ª posição: seção 1 *General Information*;
2ª posição: seção 5 *Production* e seção 9 *Self Inspection*;
3ª posição: seção 8 *Distribution, Complaints* and *Product Recall*;
5ª posição: seção 4 *Documentation*;
6ª posição: seção 6 *Quality Control*;
9ª posição: seção 7 *Contract and Manufacturing Analysis*.

Esses dados quantitativos dos *rankings*, reforçam os resultados apresentados na comparação dos sumários, os quais mostraram variações nos itens das sub-seções entre os quinze exemplares analisados, indicando que o conteúdo das informações pode ser diferente, no sentido de conter informações mais expandidas, ou então reduzidas, no processo da escrita do SMF entre os diferentes escritores.

Ademais, acreditamos que essas variações na escrita do documento, levantadas por meio dos *lexical bundles*, podem estar associadas aos negócios da empresa, nos seguintes aspectos:

(i) ao tipo de produto fabricado na localidade, por exemplo: às formas de dosagem (comprimidos, drágeas, líquido, etc.); à matéria prima ou princípio ativo

utilizado na produção, seja de uso humano ou animal; às formas de embalagem; às áreas destinadas à armazenagem e distribuição do produto; entre outros. Todos esses componentes interferem, de certa maneira, no tipo de operação e produção farmacêutica e, conseqüentemente, demandam uma escrita diferente.

(ii) à organização administrativa interna, visto que, dependendo de alguns aspectos, tais como: o número de funcionários; a contratação de terceiros; os equipamentos e documentação, determinadas localidades, com relação às demais do grupo, demandam especificações e processos diferentes, interferindo na escrita do documento. Esses aspectos podem ocorrer quando, por exemplo: da contratação de pessoal especializado para realizar análises biológicas ou microbiológicas; na aquisição de equipamentos ou áreas específicas destinadas para a produção de determinado fármaco; na produção de documentos internos diversos; na periodicidade de inspeções de qualidade e de auditorias internas ou externas; nos controles de processos e de produção, entre outros;

(iii) às políticas locais (as agências reguladoras de vigilância sanitária), pois cada órgão governamental de cada localidade, conforme já dito na descrição do *corpus*, deve instituir normas de exigências de qualidade dos serviços prestados, e do produto direcionado a saúde do consumidor. Assim, é provável que exista uma política interna de leis sanitárias no país que demandam informações diferentes com relação aos outros países, nesse caso interferindo na escrita do documento entre as localidades.

Podemos concluir que os resultados da contagem dos *lexical bundles* por localidade e por seção, serviu para demonstrar que mesmo que o documento SMF apresente em sua estrutura organizacional as mesmas seções e sub-seções, elencadas na mesma ordem seqüencial e numérica, o volume de informações (a quantidade de *bundles*) distribuídas entre as nove seções do documento, podem sofrer variações entre as localidades da empresa. É provável que, ou essas diferenças sejam decorrentes dos três aspectos que apresentamos acima, ou estejam apontando para os autores que produziram os textos, resultando nas escolhas lingüísticas feitas por esses na composição do documento.

Ademais, atentamos ao fato de que essa é uma pesquisa lingüística focada na análise do conjunto de textos que compõem o SMF, portanto, as nossas conclusões emergem da investigação dos elementos textuais (os *lexical bundles*), de tal forma que outros aspectos a respeito das operações de fabricação, e dos processos administrativos da indústria farmacêutica participante desta pesquisa, estão além do escopo deste trabalho.

Por isso, linguisticamente falando, os resultados de algumas das variações encontradas nos itens das sub-seções, apontam que a organização formal do documento (o sumário), já pré-estabelecida pelas instituições competentes, não garante que a produção escrita no corpo do documento apresente semelhanças. Em outras palavras, como se escolhe dizer, ou seja, como escrever sobre o assunto é particular das pessoas e das situações em que se produzem os textos.

Portanto, dando prosseguimento a esta etapa de análise, na próxima sub-seção discutimos os *lexical bundles* semelhantes encontrados nos quinze exemplares SMF.

3.2.4 Os *bundles* semelhantes

Nessa terceira fase, buscando as semelhanças entre os documentos, realizamos consultas ao Banco de Dados SQL (ver capítulo Metodologia - seção 2.4.2.1) para obter os dados necessários à análise. Lembrando que nessas consultas identificamos a recorrência de *lexical bundles* em > 7 localidades, ou seja, o mesmo *bundle* ocorrendo em no mínimo 8 localidades, bem como identificamos a qual seção o *bundle* pertencia. O propósito dessa fase, conforme mencionado, foi verificar o nível de semelhança existente entre os documentos por meio dos *lexical bundles*.

Essa busca foi guiada pela seguinte hipótese: se o SMF é um documento padronizado e "ditado" por regulamentos internos corporativos e políticas governamentais, então, provavelmente, em todas as quinze localidades pesquisadas é necessário seguir os mesmos padrões de produção escrita, especialmente por tratar-se de um documento proveniente de um mesmo contexto (farmacêutico) e por ser escrito, oficialmente, em um idioma comum (a língua inglesa).

Por conseguinte, para essa busca percorremos três momentos a fim de apresentar:

(i) a distribuição dos *bundles* semelhantes que ocorreram em cada seção entre as quinze localidades;

(ii) a freqüência de ocorrência desses *bundles* entre as localidades, de acordo com um critério de corte estabelecido; e

(iii) a probabilidade de ocorrência dos *bundles* semelhantes do SMF em um *corpus* geral de língua inglesa (o BNC), com a finalidade de testar a representatividade do *corpus* desta pesquisa.

Porém, enfatizamos que nessa fase não nos ocupamos em descrever gramaticalmente os *lexical bundles*, e sim fizemos uso dos *bundles* para apresentar o nível de semelhança entre as localidades do SMF, bem como verificar a probabilidade dos *bundles* semelhantes pertencerem ao domínio específico do documento SMF no contexto farmacêutico.

Na seqüência, apresentamos, em primeiro lugar, os resultados obtidos quando da comparação dos *bundles* semelhantes deste *corpus* com os do BNC. Em seguida, apresentamos a distribuição e a freqüência dos *bundles* em cada uma das nove seções do SMF, distribuídos nas suas localidades correspondentes.

3.2.4.1 Os *bundles* do SMF no *corpus* do BNC

Nesta fase dos resultados abordamos questões a respeito da representatividade do *corpus*, conforme discutidas no capítulo Fundamentação Teórica (ver seção 1.1.2) e descritas no capítulo Metodologia (ver seção 2.4.4), para validar a hipótese de que os *bundles* encontrados no SMF são de uso específico do documento SMF de domínio farmacêutico. Para tanto, apresentamos na tabela a seguir os resultados dessa análise:

Tabela 3. A probabilidade de ocorrência dos *bundles* semelhantes do SMF no *corpus* do BNC.

	Lexical bundles	CORPUS			
		SMF		BNC	
		Total de *bundles*		Total de *bundles*	
		110.496		4.423.944	
		Freqüência	Probabilidade	Freqüência	Probabilidade
1	*according to the*	58	0,525	4640	1,049
2	*activities of the*	13	0,118	512	0,116
3	**actual products manufactured**	9	0,081	0	0
4	**and distribution of**	23	0,208	161	0,036
5	**and finished products**	28	0,253	8	0,002
6	**Arrangements for the**	37	0,335	238	0,054
7	*as well as*	51	0,462	18108	4,093
8	**Brief description of** (*)	43	0,389	67	0,015
9	*carried out by*	32	0,290	1229	0,278
10	**description of the** (*)	36	0,326	909	0,205
11	**distribution of necessary**	9	0,081	0	0
12	**documentation for manufacture**	9	0,081	0	0
13	**documentation related to**	9	0,081	0	0
14	**employees engaged in**	9	0,081	2	0
15	**engaged in production**	12	0,109	0	0
16	**experience and responsibilities**	12	0,109	0	0
17	**for clinical trials**	16	0,145	4	0,001
18	**for personnel engaged**	9	0,081	0	0
19	**for process validation**	11	0,100	0	0
20	**for the handling**	36	0,326	14	0,003
21	**for the preparation**	22	0,199	67	0,015
22	**general policy for**	10	0,091	4	0,001
23	**highly toxic hazardous**	14	0,127	0	0
24	**in accordance with**	64	0,579	2042	0,462
25	**is carried out**	47	0,425	411	0,093
26	**is described in**	27	0,244	189	0,043
27	**is responsible for** (*)	74	0,670	840	0,190
28	**management system of**	9	0,081	0	0
29	**manufactured on the**	11	0,100	0	0
30	**materials and products**	26	0,235	0	0
31	**Nature of construction**	10	0,091	0	0
32	**Number of employees**	12	0,109	76	0,017
33	**of actual products**	9	0,081	0	0
34	**of complaints and**	10	0,091	12	0,003
35	**of finished products**	22	0,199	4	0,001
36	**of key personnel**	15	0,136	3	0,001
37	**of necessary documentation**	10	0,091	0	0
38	**of rejected materials**	12	0,109	0	0
39	**of starting materials**	24	0,217	2	0
40	*of the company*	9	0,081	2013	0,455

Tabela 3. continuação

	Lexical bundles	CORPUS			
		SMF		BNC	
		Total de *bundles*		Total de *bundles*	
		110.496		4.423.944	
		Freqüência	Probabilidade	Freqüência	Probabilidade
41	*of the firm*	10	0,091	522	0,118
42	**of the quality** (*)	37	0,335	309	0,070
43	**of the site**	32	0,290	435	0,098
44	**on the site**	41	0,371	587	0,133
45	**or hazardous substances**	11	0,100	0	0
46	**other manufacturing activities**	17	0,154	0	0
47	*part of the*	48	0,434	17249	3,899
48	**personnel engaged in**	11	0,100	0	0
49	**Personnel hygiene requirements**	11	0,100	0	0
50	**products manufactured on**	11	0,100	0	0
51	**purified water is**	11	0,100	0	0
52	**Quality Control Department**	29	0,262	0	0
53	**quality management system**	17	0,154	8	0,002
54	**release and storage**	9	0,081	0	0
55	**requirements including clothing**	10	0,091	0	0
56	**responsibility of the**	37	0,335	415	0,094
57	**responsible for the**	56	0,507	2108	0,476
58	**revision and distribution**	17	0,154	0	0
59	**sampling quarantine release**	9	0,081	0	0
60	**self inspection system**	13	0,118	0	0
61	**Short description of** (*)	28	0,253	11	0,002
62	**Special areas for**	10	0,091	0	0
63	**Standard Operating Procedures**	22	0,199	5	0,001
64	**system of the**	12	0,109	210	0,047
65	**the firm responsible**	8	0,072	0	0
66	*the implementation of*	14	0,127	780	0,176
67	**the preparation revision**	17	0,154	0	0
68	**the quality assurance**	15	0,136	34	0,008
69	**the Quality Control**	49	0,443	16	0,004
70	**the quality management**	15	0,136	0	0
71	*the release of*	21	0,190	859	0,194
72	**the self inspection**	21	0,190	0	0
73	**toxic harzardous and**	13	0,118	0	0
74	**toxic or hazardous**	12	0,109	0	0
75	**Use of outside**	10	0,091	2	0
76	*used in the*	28	0,253	1745	0,394

(*) Esses *bundles* semelhantes repetem em mais de uma seção, portanto, são considerados uma única vez na contagem.

Para uma melhor visualização e compreensão dos achados apresentados na Tabela 3, em **negrito,** destacamos os *bundles* com maior freqüência encontrada no SMF, e em *itálico*, os *bundles* com maior freqüência encontrada no BNC. Desse modo, pudemos observar que o total de *bundles* distintos, apurados nessa fase de comparação das freqüências, foi de 76 *bundles*, sendo que em média 15% do total desses (10 *bundles*) apresentaram uma probabilidade de ocorrência maior em relação à probabilidade de ocorrência no *corpus* do SMF. Conforme mostra a Tabela 3, os *bundles* de maior freqüência no BNC foram os seguintes: *according to the*, *activities of the*; *as well as*; *carried out by*; *of the company*; *of the firm*; *part of the*; *the implementation of* ; *the release of* ; e *used in the*.

Isso significa dizer, que a maioria dos *bundles* semelhantes encontrados ocorre, proporcionalmente, com maior freqüência nos documentos do SMF do que no *corpus* do BNC. Assim, podemos concluir que os *bundles* do SMF (66 dos 76), são, tipicamente, mais expressivos e relevantes na representação de domínio específico do documento.

Outrossim, conforme apresentado no capítulo Fundamentação Teórica desta pesquisa, apesar de alguns estudiosos da Lingüística de *Corpus* afirmarem que é necessário trabalhar com uma grande quantidade de palavras para que a amostra seja representativa de uma determinada população, acreditamos que a partir do *corpus* do SMF (contendo 110.766 palavras - ver Quadro 4), pudemos oferecer a representativa dos *bundles* específicos do SMF. Ademais, o fato de o *corpus* de estudo ter sido comparado a um *corpus* de grande porte, como o BNC (*British National Corpus*), garantiu uma maior confiabilidade para os resultados desta análise.

Uma vez mostrada a representatividade desses *bundles* no *corpus* de domínio do SMF, passamos a apresentar os resultados da extração dos *bundles* semelhantes, encontrados em cada uma das nove seções do documento.

3.2.4.2 A distribuição e a freqüência dos *bundles*

Apurados os *bundles* semelhantes, por meio do uso de ferramentas computacionais, distribuímos cada um deles em suas respectivas localidades, obedecendo ao critério de corte estabelecido na metodologia deste trabalho, ou seja, freqüência > 7 localidades (ver capítulo 2 – seção 2.4.3). Dessa forma, organizamos os *bundles* semelhantes encontrados por seção e os apresentamos em tabelas, conforme seguem:

Seção 1 *General Information*

Tabela 4: Distribuição dos *bundles* semelhantes da seção 1 *General Information*

	Lexical bundles	Localidade															
		Freq	AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	of the quality	11	■	■	■	■	■	■	■	■		■	■	■			
2	quality management system	10	■	■	■	■	■		■	■		■	■	■			
3	the quality assurance	10	■	■	■	■	■	■		■		■	■	■			
4	short description of	10	■	■	■	■	■	■		■		■	■	■			
5	of the site	10	■	■		■	■	■	■	■	■	■	■				
6	description of the	10	■	■	■	■	■		■	■		■	■	■			
7	Number of employees	9	■	■	■	■		■		■		■	■	■			
8	Use of outside	9	■	■	■	■		■		■		■	■	■			
9	toxic or hazardous	9	■	■	■	■	■			■		■	■	■			
10	on the site	9	■	■	■	■				■	■	■	■	■			
11	the quality management	9	■		■	■	■		■	■		■	■	■			
12	manufactured on the	9	■	■		■		■		■	■	■	■	■			
13	system of the	8	■	■	■	■			■			■	■	■			
14	other manufacturing activities	8	■	■		■	■			■		■	■	■			
15	is responsible for	8		■	■	■		■				■	■			■	■
16	is described in	8		■	■	■	■		■	■		■			■		
17	products manufactured on	8	■	■		■				■	■	■	■	■			
18	or hazardous substances	8	■	■		■	■			■		■	■	■			
19	for clinical trials	8	■	■	■	■		■		■			■	■			
20	the implementation of	8		■		■	■	■		■		■		■			■
21	the firm responsible	8	■	■	■	■			■			■	■	■			
22	employees engaged in	8	■		■	■	■	■				■	■	■			
23	of actual products	8	■	■		■	■			■		■	■	■			
24	of the company	8		■	■	■				■			■	■		■	■
25	of the firm	8	■	■	■	■			■			■	■	■			
26	responsibility of the	8		■	■	■	■	■		■		■		■			
27	responsible for the	8		■		■		■	■			■	■			■	■
28	actual products manufactured	8	■	■		■	■			■		■	■	■			
29	management system of	8	■		■	■	■		■			■	■	■			

Observando a Tabela 4, na coluna da esquerda para a direita, encontramos os *bundles* seguidos do seu grau de freqüência, isto é, em quantas localidades esses *bundles* ocorreram. Logo após, na coluna da direita, encontramos assinalados os nomes das localidades para efeito de cruzamento dos dados.

Considerando o grau de freqüência, dentre os 29 *bundles* encontrados na seção 1 *General Information*, foi identificado um único *bundle* recorrente em 11 localidades (*of the quality*). Lembrando que não ocorreram *bundles* semelhantes nas quinze localidades. Com base na categoria dos *lexical bundles*, e observada a baixa freqüência de *bundles* no

SMF*Japan*, SMF*Brazil* e SMF*Puerto* (conforme Tabela 4), nessa seção, pressupomos haver diferenças quanto às escolhas lingüísticas feitas pelos escritores das quinze localidades da empresa. Vejamos a seguir os resultados da seção 2 *Personnel*.

Seção 2 *Personnel*

Tabela 5. Distribuição dos *bundles* semelhantes da seção 2 *Personnel*

	Lexical bundles	Localidade														
		Freq	AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	of key personnel	10	▨	▨	▨	▨	▨	▨		▨		▨	▨	▨			
2	experience and responsibilities	9	▨	▨	▨	▨	▨			▨		▨	▨	▨			
3	requirements including clothing	9	▨	▨	▨	▨	▨			▨		▨	▨	▨			
4	personnel engaged in	9	▨	▨	▨	▨	▨			▨		▨	▨	▨			
5	Personnel hygiene requirements	9	▨	▨	▨	▨	▨			▨		▨	▨	▨			
6	for personnel engaged	8	▨	▨	▨	▨				▨		▨	▨	▨			
7	engaged in production	8	▨	▨	▨	▨		▨		▨		▨	▨				

Na seção *Personnel*, encontramos 7 *bundles* recorrentes, e somente um único *bundle* ocorreu em 10 localidades (*of key personnel*). Em virtude desse resultado, notamos níveis de diferenças, como ocorrido na seção 1 *General Information*, visto que foi identificado a ausência de *bundles* semelhantes nos SMF*USA, SMF*Japan, SMF*Brazil*, e SMF*Puerto*, levando-nos a pressupor que as diferenças poderiam ser ainda maiores nessas localidades.

A seguir, apresentamos a tabela com os dados da seção 3 *Premises and Equipment.*

Seção 3 *Premises and Equipment*

Tabela 6. Distribuição dos *bundles* semelhantes da seção 3 *Premises and Equipment*

	Lexical bundles	Freq	Localidade														
			AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	according to the	12	■	■	■	■	■	■		■	■	■		■		■	■
2	for the handling	10	■	■	■	■	■			■	■	■	■	■			
3	highly toxic hazardous	10	■	■	■	■	■			■	■	■	■	■			
4	Brief description of	9	■	■	■	■	■			■		■	■	■			
5	Special areas for	9	■	■	■	■	■			■		■	■	■			
6	toxic harzardous and	9	■	■	■	■	■			■		■	■	■			
7	purified water is	8		■	■			■		■	■	■	■			■	
8	used in the	8			■	■			■	■	■	■		■			■
9	is carried out	8		■		■	■	■				■	■	■		■	
10	as well as	8		■	■		■	■				■	■	■		■	
11	part of the	8		■			■	■	■	■			■	■		■	
12	carried out by	8	■	■		■	■					■	■	■		■	
13	Nature of construction	8	■		■	■	■			■		■	■	■			

Como podemos verificar na Tabela 6, obtivemos 13 *bundles* semelhantes, com apenas um único *bundle* recorrente em 12 localidades (*according to the*). Além disso, nessa seção, verificamos, também, a ausência de *bundles* na localidade SMF*Japan*.

A seguir, apresentamos a seção 4 *Documentation*.

Seção 4 *Documentation*

Tabela 7. Distribuição dos *bundles* semelhantes da seção 4 *Documentation*

	Lexical bundles	Freq	Localidade														
			AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	and distribution of	11	■	■	■	■	■	■		■		■	■	■			■
2	the preparation revision	11	■	■	■	■	■	■		■		■	■	■			■
3	Arrangements for the	10	■	■		■	■	■	■	■		■	■	■			
4	for the preparation	10	■	■	■	■	■			■	■	■	■	■			
5	revision and distribution	10	■	■	■	■	■	■		■		■	■				■
6	of necessary documentation	9	■	■	■	■	■			■		■	■	■			
7	Standard Operating Procedures	8				■		■		■	■	■	■	■			■
8	documentation for manufacture	8	■	■	■	■	■			■		■	■				■
9	documentation related to	8	■	■		■				■		■	■			■	■
10	distribution of necessary	8	■	■	■	■	■			■		■	■				

Na seção *Documentation* foram encontrados 10 *bundles* semelhantes, sendo dois deles (*the preparation revision* e *and distribution of*) com freqüência em 11 localidades. Nessa seção, a localidade SMF*Japan* não apresentou nenhuma ocorrência.

Na seqüência, apresentamos a seção 5 *Production.*

Seção 5 *Production*

Tabela 8. Distribuição dos *bundles* semelhantes da seção 5 *Production*

	Lexical bundles	Localidade														
		Freq	AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	materials and products	12	▨	▨	▨	▨	▨		▨	▨	▨	▨	▨	▨		▨	
2	Brief description of	9	▨	▨	▨	▨	▨		▨	▨		▨	▨				
3	for process validation	8	▨		▨	▨	▨		▨	▨		▨	▨				
4	in accordance with	8		▨	▨		▨	▨	▨	▨	▨	▨					
5	and finished products	8	▨	▨	▨	▨	▨			▨		▨	▨				
6	of rejected materials	8	▨		▨	▨	▨			▨		▨	▨	▨			
7	of starting materials	8	▨	▨	▨	▨		▨		▨		▨	▨				
8	release and storage	8	▨	▨	▨	▨	▨			▨		▨	▨				
9	general policy for	8	▨	▨	▨	▨	▨			▨		▨	▨				
10	sampling quarantine release	8	▨	▨	▨	▨	▨			▨		▨	▨				

A seção *Production* apresentou 10 *bundles* recorrentes, tendo um único *bundle* com freqüência em 12 localidades (*materials and products*). Ademais, não identificamos ocorrências nas localidades SMF*Japan* e SMF*Puerto.*

Dando prosseguimento à apresentação dos resultados, a seguir, apresentamos os *bundles* semelhantes encontrados na seção 6 *Quality Control.*

Seção 6 *Quality Control*

Tabela 9. Distribuição dos *bundles* semelhantes da seção 6 *Quality Control*

	Lexical bundle	Localidade															
		Freq	AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	the Quality Control	12															
2	of the quality	11															
3	Quality Control Department	10															
4	the release of	10															
5	of finished products	9															
6	description of the	9															
7	is responsible for	8															
8	activities of the	8															

Na seção *Quality Control*, dos 8 *bundles* semelhantes, encontramos um único *bundle* com freqüência em 12 localidades (*the Quality Control*).

A seguir, apresentamos os resultados da seção 7 *Contract Manufacture and Analysis.*

Seção 7 *Contract Manufacture and Analysis*

A seção *Contract Manufacture and Analysis* não apresentou dados para essa fase. Portanto, para o objetivo desta fase da análise, não foi possível considerar essa seção do SMF, tendo ∅ de *bundles* semelhantes como resultado.

Na tabela seguinte, apresentamos os resultados da seção 8 *Distribution, Complaints and Product Recall.*

Seção 8 *Distribution, Complaints and Product Recall*

Tabela 10. Distribuição dos *bundles* semelhantes da seção 8 *Distribution, Complaints and Product Recall*

	Lexical bundle	Localidade															
		Freq	AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	of complaints and	8															

A Seção *Distribution, Complaints and Product Recall* apresentou apenas um único *bundle* semelhante (*of complaints and*), com freqüência em 8 localidades.

A seguir, para finalizar a apresentação da distribuição e freqüência dos *bundles* semelhantes, apresentamos a tabela com os dados da seção 9 *Self Inspection*.

Seção 9 *Self Inspection*

Tabela 11. Distribuição dos *bundles* semelhantes da seção 9 *Self Inspection*

	Lexical bundle	Freq	Localidade														
			AnnonayFR	HuningueFR	NyonFR	BaselSWIT	HettlingenSWIT	England	USA	Germany	The Netherlands	Italy	Austria	Turkey	Japan	Brazil	Puerto
1	the self inspection	10	▨	▨	▨	▨	▨			▨		▨	▨	▨	▨		
2	Short description of	8	▨	▨	▨	▨	▨			▨		▨	▨				
3	self inspection system	8	▨		▨	▨	▨			▨		▨	▨	▨			
4	description of the	8	▨	▨	▨	▨	▨			▨		▨	▨				

Nessa seção, obtivemos quatro *bundles* recorrentes, tendo um único *bundle* (*the self inspection*) com freqüência em 10 localidades. Além do mais, 4 das 15 localidades (SMF*England*, SMF*USA*, SMF*Brazil* e SMF*Puerto*) não apresentaram ocorrências.

3.2.4.3 Discussão dos *bundles* semelhantes

Antes de iniciarmos essa discussão, é importante salientar o seguinte questionamento que surgiu durante a apuração dos resultados, quanto à extração da seqüência de três palavras deste *corpus* de estudo: o porquê nenhum *lexical bundle* ocorreu em quinze localidades, considerando que os títulos das nove seções do SMF são semelhantes (ver Quadro 5) e, de acordo com a extração automática dos *bundles*, a seção 8 (*Distribution, Complaints and Product Recall*) formam 5 palavras, e a seção 7 (*Contract Manufacture and Analysis*) formam 4 palavras. Por conseqüência, esse conjunto de palavras deveria formar *bundles* do tipo: *Distribution Conplaints and*; *Complaints and Product*; *and Product Recall*, e assim da mesma forma para a seção 7. No entanto, nenhum desses conjuntos de palavras ocorreu durante a extração dos *bundles* semelhantes.

Para elucidar essa questão, explicamos que os sumários dos exemplares do SMF não foram considerados quando da coleta do *corpus* para o processamento eletrônico na formação

dos *bundles*. Assim sendo, a ausência de *bundles* semelhantes nas quinze localidades, com a seqüência de palavras formada a partir dos títulos das seções, provavelmente, foi devido ao fato de que no corpo do documento os títulos das seções e sub-seções, tal qual apareceram no sumário, não ocorreram da mesma forma escrita.

Com efeito, diante dos resultados apresentados nas tabelas precedentes da sub-seção anterior, pudemos observar o seguinte:

(i) em nenhuma das nove seções ocorreram *bundles* semelhantes nas quinze localidades;

(ii) o nível máximo da freqüência de *bundles* semelhantes atingido foi o de 12 localidades. Esses *bundles* mais freqüentes foram os seguintes: *according to the* (seção 3 *Premises and Equipments*); *materials and products* (seção 5 *Production)*; e *the quality control* (seção 6 *Quality Control*) - ver Tabelas 6, 8 e 9;

(iii) a freqüência de *bundles* semelhantes entre as seções foi muito variada, isto significa, que na seção 1 (*General Information*) encontramos 29 *bundles* semelhantes, e na seção 8 (*Distribution, Conplaints and Product Recall*) somente um único *bundle* semelhante.

(iv) na seção 7 (*Contract Manufacture and Analysis*) não ocorreu nenhum *bundle* semelhante.

A partir dos resultados apresentados nas tabelas, vejamos no quadro a seguir um panorama dos *bundles* semelhantes encontrados nas nove seções do SMF.

Quadro 8. Síntese da quantidade de *lexical bundles* semelhantes encontrados nas seções do SMF.

Seção		Quantidade de *bundles* semelhantes	Freqüência (>7 localidades)							
			8	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	General Information	29	17	6	5	1	∅	∅	∅	∅
2	Personnel	7	2	4	1	∅	∅	∅	∅	∅
3	Premises and Equipment	13	7	3	2	∅	1	∅	∅	∅
4	Documentation	10	4	1	3	2	∅	∅	∅	∅
5	Production	10	8	1	∅	∅	1	∅	∅	∅
6	Quality Control	8	2	2	2	1	1	∅	∅	∅
7	Contract Manufacture and Analysis	0	∅	∅	∅	∅	∅	∅	∅	∅
8	Distribution, Complaints and Product Recall	1	1	∅	∅	∅	∅	∅	∅	∅
9	Self Inspection	4	3	∅	1	∅	∅	∅	∅	∅

Para uma breve explanação a respeito do quadro 8 e a fim de justificar o item (i), mencionado anteriormente, verificamos que não houve ocorrência de *bundles* semelhantes nas quinze localidades do SMF. As ocorrências se estabeleceram em 11 ou 12 exemplares, por exemplo: na seção 1 (*General Information*) encontramos 29 *bundles* semelhantes, sendo que 17 ocorreram em 8 localidades, 6 em 9 localidades, 5 em 10 localidades, e um único *bundle* em 11 localidades. Essa variação, da freqüência por localidades, foi observada em todas as nove seções. Portanto, as semelhanças ocorreram entre 8 a 10 exemplares do documento.

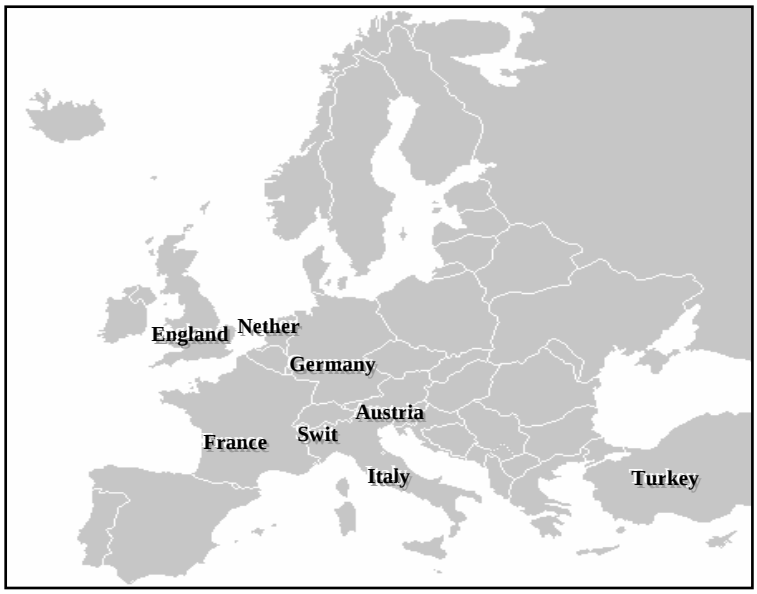
Logo, a nossa hipótese inicial: se o SMF é um documento padronizado e "ditado" por regras, então, provavelmente, em todas as localidades que a empresa esteja instalada é necessário seguir os mesmos padrões na produção escrita, especialmente por tratar-se da mesma especialidade e por ser escrito, oficialmente, em um idioma comum, não se confirmou. O SMF apresentou poucas semelhanças no cruzamento dos *bundles* entre as nove seções, além de apresentar ausência de *bundles* semelhantes em algumas localidades, como é o caso dos SMF*Japan*, SMF*Brazil*, SMF*Puerto*, e SMF*USA*. Ademais, na seção 7 (*Contract Manufacture and Analysis*) não houve nenhuma ocorrência, indicando que, possivelmente, essa seja a seção que mais apresente variações, tanto na escrita como no tipo de informações contidas nesse tópico do documento.

Acreditamos que as variações da freqüência, em maior ou menor número de *bundles* semelhantes, entre as seções das quinze localidades, podem sinalizar diferenças quanto às escolhas lingüísticas feitas por quem produziu o documento em cada localidade, ou que o

assunto é tratado linguisticamente de forma diferente, hipótese já levantada na discussão da contagem dos *bundles* (conforme seção 3.2.3).

Por um outro viés, observando os resultados dos *bundles* semelhantes, distribuídos na seqüência das tabelas anteriores (Tabelas 4-11), e considerando a proximidade geográfica das localidades, verificamos que a maior parte de ocorrência de *bundles* semelhantes estava concentrada em regiões da Europa, incluindo a localidade onde está situada a matriz da indústria farmacêutica em questão. Para ilustrar a nossa discussão, vejamos abaixo uma figura representando um recorte do mapa do continente europeu com as localidades européias do SMF.

Figura 15. A proximidade geográfica entre as localidades.



Diante disso, levantamos uma outra hipótese: de que os negócios da empresa, decorrente da proximidade geográfica entre as localidades (conforme explicado adiante), poderiam estar associados tanto à concentração de *bundles* semelhantes nessas regiões como à ausência de *bundle*s semelhantes nas demais localidades. Assim, ancoramos nossos pressupostos nos seguintes aspectos referentes aos interesses da empresa:

- à produção farmacêutica;
- ao órgão governamental de vigilância sanitária; e
- à distribuição e manipulação dos insumos e produtos farmacêuticos.

Analisando a ocorrência dos *bundles* semelhantes nas localidades próximas, como, por exemplo: *AnnonayFR*; *HuningueFR*; *NyonFR*; *BaselSWIT*; *HettilingenSWIT*; *England*; *Germany*; *Italy*; *The Netherlands*; *Italy*; *Austria*; e *Turkey*, e comparando-as com as localidades mais afastadas dessa região (Europa), como, por exemplo: *Japan*; *Brazil*; *Puerto Rico*; e *U.S.A.*, percebemos que entre essas localidades próximas, a ocorrência de *bundles* semelhantes foi maior do que entre as localidades mais afastadas (Japão e América). Assim, agrupamos e classificamos as quinze localidades em dois pólos: as localidades nucleares e as localidades periféricas.

No quadro abaixo, vejamos a disposição das localidades, baseadas nessa classificação:

Quadro 9. Os SMF agrupados em localidades nucleares e periféricas.

Nucleares	Periféricas
SMF*AnnonayFR* SMF*NyonFR* SMF*HuningueFR* SMF*BaselSWIT* SMF*HettlingenSWIT* SMF*England* SMF*Germany* SMF*Netherlands* SMF*Italy* SMF*Austria* SMF*Turkey*	SMF*Brazil* SMF*Japan* SMF*Puerto* SMF*USA*

De acordo com os pressupostos acima, e baseados no nosso conhecimento a respeito das operações da empresa e nos dados desta análise, consideramos o seguinte:

(i) as localidades nucleares apontam para as regiões onde estão alocadas as unidades de produção dos medicamentos, as quais possuem, entre outras, a missão de exportar os seus insumos a outros países. Portanto, é provável que a ausência de *bundles* semelhantes nas localidades periféricas é devida à empresa não empreender a fabricação de insumos farmacêuticos nessas regiões, isto é, essas localidades periféricas recebem, via importação, os produtos a serem distribuídos e comercializados, e a partir disso, talvez, surja a necessidade de se escrever sobre outros assuntos – descrever outras atividades;

(ii) as regiões nucleares participam de uma mesma convenção que regulamenta as drogas farmacêuticas na Europa - o EMEA (*European Medicines Agency*), ao contrário das regiões periféricas, que para cada localidade existe uma agência reguladora, como, por exemplo: o FDA (*Food and Drug Administration*) para o SMF*USA* e SMF*Puerto*; a ANVISA (Agência Nacional de Vigilância Sanitária) para o SMF*Brazil*; e o NPA (*National Policy Agency*) para o SMF*Japan*, o que pode gerar variações quanto ao conteúdo inserido nos textos com relação às exigências locais.

Apesar de todas as localidades seguirem as Boas Práticas de Fabricação Farmacêutica (as BPF), instituída por um tratado internacional para garantir a saúde do consumidor, pode haver tanto fatores locais, os quais diferem das políticas internas das agências de vigilância sanitária de outras localidades, como fatores relacionados à natureza dos negócios praticados em cada uma das localidades, os quais interferem na escrita do documento, ou seja, o assunto que se escreve pode ser particular de cada localidade.

Portanto, a nossa hipótese de que os negócios da empresa, decorrentes da proximidade geográfica entre as localidades, poderiam estar associados à concentração de *bundles* semelhantes nessas regiões, confirma-se na medida em que:

(i) a produção farmacêutica está na maior parte localizada em um território em comum (a Europa), facilitando o trâmite das informações e as negociações comerciais e logísticas entre elas;

(ii) a recepção (via importação) dos insumos farmacêuticos por países que não possuem unidades de fabricação, pode resultar na ausência de informações, ou informações diferentes que devem ser contempladas na escrita do documento;

(iii) as localidades como França (com três instalações: *Annonay*; *Huningue*; e *Nyon*) e Suíça (com duas instalações: *Basel* e *Hettlingen*), por possuirem mais de uma instalação de fabricação, refletem, por parte dos escritores dos SMF, um compartilhamento de recursos lingüísticos no uso de expressões e palavras semelhantes;

(iv) as agências que regulamentam a vigilância sanitária de cada país, no caso das localidades instaladas na Europa, estão sob fiscalização da mesma instituição governamental – o EMEA (*European Medicines Agency*), ao contrário das demais localidades, que estão submetidas à órgãos distintos, ou seja, participam de um outro conjunto de regras sanitárias, inerentes ao local no qual estão instaladas. Daí a necessidade de outras informações na escrita do documento.

É necessário esclarecer, que a indústria farmacêutica em estudo possui mais de 100 unidades espalhadas pelo mundo, sendo que o nosso *corpus* está limitado a quinze localidades. Dessa forma, seria interessante um estudo com um maior volume de dados (outras localidades) para reforçar a hipótese de que os *bundles* semelhantes podem estar associados ao tipo de operações realizadas na empresa (o que se faz), ou aos órgãos governamentais que regulamentam as atividades farmacêuticas de cada país, dependendo da proximidade geográfica em que essas localidades estejam instaladas.

Enfim, apresentados os resultados e a discussão dos *bundles* semelhantes, partimos para a última fase desta etapa dos resultados.

3.3 A coesão lexical entre o tópico das seções e sub-seções do SMF

Nesta terceira etapa dos resultados, buscamos relacionar os *bundles* semelhantes recorrentes entre as quinze localidades do SMF, com a finalidade de estabelecer uma relação de coesão com as seções e sub-seções do documento, e responder a nossa terceira pergunta de pesquisa: Como os *lexical bundles* semelhantes refletem os tópicos das seções do documento SMF?

Para discorrer sobre esta etapa, baseamo-nos na noção de coesão lexical apresentada nos trabalhos de Berber Sardinha (1997) e Ramos (1997), e identificamos que a maioria dos *bundles* semelhantes encontrado em cada seção, estabelecia com os títulos e sub-títulos, uma relação coesiva do tipo repetição simples ou complexa, isto é, quando um item é repetido utilizando a mesma forma ou formas gramaticais idênticas (ver Capítulo 1 – seção 1.3.2).

Para guiar a nossa discussão, recortamos cada seção com seus respectivos títulos e sub-títulos, seguidos dos *bundles* semelhantes que reiteravam o assunto tratado nessas seções. Vejamos a seguir a seção 1 *General Information*:

Quadro 10. Relação coesiva entre os *bundles* semelhantes e o títulos da seção 1 do SMF

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes	Outros *bundles* semelhantes
1	General Information	1. Brief information on the company;	- *of* **the company**	- <u>is described in</u> - <u>for clinical trials</u> - <u>the implementation of</u>
		2. Any other manufacturing activities carried on Site;	**- other manufacturing activities**	
		3. Name and exact address of Site;		
		4. Type of actual products manufactured on the Site and information about specifically toxic or hazardous substances handled, mentioning the way they are manufactured (in dedicated facilities or on a campaign basis);	**- toxic or hazardous** **- on the site** **- manufactured on the** **- of actual products** **- actual products manufactured** **- products manufactured on** **- or hazardous substances**	
		5. Short Description of the Site;	**- short description of** **- description of the** **- of the site**	
		6. Number of employees engaged in Quality Assurance, Quality Control, Production, Storage and Distribution;	- *the* **quality assurance** **- Number of employees** **- employees engaged in**	
		7. Use of outside assistance for manufacturing;	**- Use of outside**	
		8. Short description of the quality management system of the firm responsible for manufacture.	**- of the quality** **- quality management system** **- the quality management** **- system of the** - *is* **responsible for** - *the* **firm responsible** **- of the firm** **- responsible for** *the* **- management system of** **- responsible** *of the*	

No quadro 10, na coluna *Lexical bundles* semelhantes, em **negrito**, destacamos os *bundles* com as palavras que se repetiram igualmente nos tópicos das seções e, em *itálico*, as palavras que faziam parte do conteúdo dos *bundles* que não se repetiram. Na coluna Outros *bundles* semelhantes, em <u>sublinhado,</u> apresentamos os *bundles* que não mantiveram uma relação de coesão por repetição simples com os tópicos.

Podemos observar que dos 29 *bundles* semelhantes encontrados na seção 1 *General Information* (conforme Tabela 4), 23 deles se repetiram de forma idêntica no conteúdo dos sub-títulos apresentados no sumário do SMF. Por exemplo, no sub-título de nº 5 (*Short Description of the Site*), encontramos os *bundles*: *short description of; description of the; of*

the site, exatamente a seqüência de palavras que forma a sentença *Short Description of the Site*.

Em um outro exemplo, o sub-título de nº 8, com os *bundles*: *is reponsible for*; *responsible for the*; e *the firm responsible*, observamos que todos esses *bundles*, além de se repetirem de forma idêntica no texto, também apresentaram o tópico central dessa sub-seção - a gestão responsável pelo sistema de qualidade da empresa. No sub-título de nº 4, os *bundles*: *of actual products; actual products manufactured;* e *toxic or hazardous*, estabeleceram, por meio da palavra *products*, uma ligação com os produtos e tipo de produtos que, provavelmente, são produzidos nas localidades. Desse modo, a grande maioria dos *bundles* semelhantes encontrado na seção 1 (*General Information*) refletiram exatamente os tópicos tratados nessa seção, como se fossem blocos que edificam o texto e que estão firmemente ligados (*building blocks*).

Conforme apresentado no Quadro 10, 3 dos 29 *bundles*: *is described in*; *for clinical trials*; e *the implementation of* não estabeleceram uma relação de coesão por repetição simples em nenhum dos tópicos das sub-seções. É possível que esses três *bundles* não estabeleçam relação de forma coesiva por repetição com os tópicos, pois algumas diferenças em termos de informações nessa seção não se "amarraram" de forma semelhante entre as quinze localidades. Essas informações podem estar relacionadas, por exemplo: ao tipo de produção; aos equipamentos; ao pessoal; entre outras. Isso reitera o que já foi dito na discussão dos resultados anteriores, que as escolhas de como essas informações foram escritas parecem ser bem diferentes entre as localidades, apesar da obrigatoriedade organizacional do documento.

Quanto à composição gramatical dos *bundles* semelhantes, na seção 1 (*General Information*), encontramos *bundles* formados por grupos nominais, indicando lugares, processos ou eventos, como, por exemplo: os *bundles*: *on the site*; e *actual products manufactured*. Observamos, também, *bundles* formados de preposição + grupo nominal: como os *bundles*: *of the company*; *of actual products*; *of the site*; *on the site; of the firm*; *for clinical trials*, indicando a existência ou identificação de lugares da própria empresa, ou indicando as atividades da empresa. Nessa seção, pudemos identificar *bundles* formados de verbo de ligação + verbo na passiva, como, por exemplo, *is described in*, e verbo de ligação + predicativo, como o *bundle*: *is responsible for*, sinalizando como os agentes das ações foram referenciados.

Dando seqüência a apresentação e discussão dos resultados, a seguir temos a seção 2 *Personnel*.

Quadro 11. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 2 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes
2	Personnel	1. Organization chart;	--------
		2. Qualification, experience and responsibilities of key personnel;	**- experience and responsibilities** **- of key personnel;**
		3. Outline of arrangement for basic and in-service training;	-------
		4. Health requirements for personnel engaged in Production;	**- for personnel engaged** **- engaged in production** **- personnel engaged in**
		5. 5. Personnel hygiene requirements including clothing.	**- Personnel hygiene requirements** **- requirements including clothing**

Na seção 2 *Personnel*, da mesma forma que apresentamos anteriormente, os *bundles*, destacados em **negrito,** se repetiram igualmente nos tópicos da seção. Todos os 7 *lexical bundles* semelhantes encontrados (ver Tabela 5) possuíam um elo coesivo por repetição simples com os sub-títulos da seção. Por exemplo, o sub-título de nº 2 apresentou os *bundles*: *experience and responsibilities*; e *of key personnel*, repetindo exatamente a seqüência de palavras da sentença. Isso também ocorreu com os sub-títulos de nº 4 e 5, com os *bundles*: *for personnel engaged*; *personnel engaged in;* e *Personnel hygiene requirements*, todos com a palavra *personnel* em seu conteúdo.

Notamos que os *bundles* dessa seção do SMF, apontavam para uma relação coesiva com o assunto principal – o Pessoal (*Personnel*). Dessa forma, concluímos que os *bundles* relaciovam-se com as exigências de higiene do pessoal, das vestimentas, e das experiências e responsabilidades da equipe de profissionais, apresentando as informações que deveriam conter nessa parte do documento. Assim, pudemos verificar que esses *bundles* refletiam por completo o assunto tratado nessa seção. No entanto, a baixa quantidade de *bundles* apurada (7 *bundles*) demonstrou que, provavelmente, nessa seção as informações devem ser escritas de maneira diferente.

Na seção 2, encontramos composições gramaticais dos *bundles* semelhantes, tais quais: preposição + grupo nominal, por exemplo: o *bundle of key personnel*, ou somente grupos nominais, por exemplo: *Personnel hygiene requirements*, indicando as entidades tratadas na seção.

A seguir, apresentamos os resultados da seção 3 *Premises and Equipment*.

Quadro 12. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 3 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes	Outros *bundles* semelhantes
3	Premises and Equipment	1. Site Plan;		- used in the - is carried out - carried out by - according to the - as well as - part of the
		2. Nature of construction and finishes;	- **Nature of construction**	
		3. Brief Description of Ventilation systems;	- **Brief description of**	
		4. Special areas for the handling of highly toxic, hazardous and sentisizing materials;	- **for the handling** - **highly toxic hazardous** - **Special areas for** - **toxic hazardous and**	
		5. Brief description of the water system;	- *purified* **water** *is*	
		6. Maintenance;		
		7. Description of major production and control laboratory equipment;		
		8. Maintenance and Servicing;		
		9. Qualification, Validation and Calibration;		
		10. Sanitation.		

Na seção 3 do SMF, conforme demonstrado no quadro 12, dos 13 *bundles* semelhantes encontrados (ver Tabela 6), 6 deles (em **negrito**), elencados na coluna *Lexical bundles* semelhantes, relacionaram-se de forma coesiva por repetição simples com os sub-títulos, e retomaram a mesma seqüência de palavras.

Para exemplificar, tivemos o sub-título de nº 4 com os *bundles*: *for the handling; highly toxic hazardous; Special areas for; toxic hazardous and*, os quais juntos formaram a própria sentença do sub-título (*Special areas for the handling of highly toxic, hazardous and sentisizing materials*). Quanto ao aspecto de ligação semântica com o assunto da seção, tivemos o *bundle* do sub-título de nº 2: *Nature of construction*, que além de repetir a mesma formação como se apresentou na sentença (*Nature of construction and finishes*), também fez referência às instalações, supostamente, os prédios alocados nas áreas da empresa.

Já o *bundle purified water is* apresentou somente a ligação coesiva por meio de uma palavra da seqüência de três, a palavra *water*, e as outras palavras da seqüência, *purified* e *is* (em *itálico*), indicaram o tipo de água utilizada para a produção dos medicamentos.

O interessante a ser observado nessa seção, foram os 6 *bundles* que não estabeleceram relação coesiva por repetição simples com os tópicos, conforme apresentados na coluna Outros *bundles* semelhantes (em sublinhado), os quais foram os seguintes: *used in the*; *is carried out*; *carried out by*; *according to the*; *as well as*; *part of the*, sendo que os três

primeiros eram compostos por verbos na passiva, indicando ações diferentes, e não indicando os agentes dessas ações, e os outros três *bundles*, indicando expressões de relações lógicas e ou comparativas. Tais *bundles* foram os que apresentaram maior probabilidade no *corpus* do BNC quando comparados com o *corpus* do SMF. Talvez, o fato de esses 6 *bundles* terem obtido maior probabilidade no *corpus* do BNC, indiquem, que nessa seção, as informações podem ter sido tratadas ou escritas com menor especificidade, dada a representatividade do domínio específico encontrada no *corpus* deste estudo (conforme Tabela 3).

No quadro seguinte, apresentamos a seção 4 *Documentation* com os seus resultados.

Quadro 13. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 4 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes	*Outros bundles* semelhantes
4	Documentation	1. Arrangements for the preparation, revision and distribution of necessary documentation for manufacturing;	- **and distribution of** - **the preparation revision** - **Arrangements for th*e*** - **for the preparation** - **revision and distribution** - **of necessary documentation** - **documentation for** *manufacture* - **distribution of necessary**	- Standard Operating Procedures
		2. Other Documentation related to product quality.	- **documentation related to.**	

Na seção 4, conforme apresentamos no quadro 13, dos 10 *bundles* semelhantes encontrados (ver Tabela 7), 8, apresentados na coluna *Lexical bundles* semelhantes (em **negrito**) se repetiram no sub-título de nº 1, estabelecendo a relação coesiva com o assunto da seção. O 8 *bundles* recorrentes, em conjunto, formaram a sentença completa (*Arrangements for the preparation, revision and distribution of necessary documentation for manufacturing*), com exceção do *bundle documentation for manufacture*, que apresentou a palavra *manufacture* (em *itálico*) de forma gramatical diferente (*manufacturing*) do que ocorreu no sub-título, indicando nessa caso uma repetição complexa[47].

Esse número de concentração de *bundles* semelhantes que se associaram diretamente ao título por repetição, indicam que o conteúdo tratado nesse sub-item pode ser o mesmo em muitos dos documentos.

[47] Conforme Capítulo 1 Fundamentação Teórica – seção 1.3.2. A repetição complexa ocorre quando dois itens compartilham um mesmo morfema lexical, mas não possuem formas gramaticais idênticas.

No sub-título de nº 2, o *bundle documentation related to* se repetiu de forma idêntica na sub-seção. Além do mais, as palavras *documentation, preparation,* e *distribution,* perpassaram por quase todos os *bundles* semelhantes encontrados, indicando uma relação semântica com o assunto principal da seção: os trâmites de documentação da empresa.

Quanto às formas, tivemos novamente ocorrências, como, por exemplo: grupos nominais (o *bundle Standard Operating Procedures*); e artigo + grupo nominal (o *bundle the preparation revision),* designando as entidades do discurso (sobre o que se fala).

Prosseguindo com a análise, apresentamos no quadro abaixo a seção 5 *Production*.

Quadro 14. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 5 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes	Outros *bundles* semelhantes
5	Production	1. Description of Production Operations;	- *Brief* **description of**	- in accordance with
		2. Arrangements for Handling Starting Materials,Packaging Materials, Bulk and Finished products including Sampling, Quarantine, Release and Storage;	- **and finished products** - **release and storage** - **sampling quarantine release** - *of* **starting materials**	
		3. Arrangements for Reprocessing or Rework;		
		4. Arrangements for Handling Reject Materials and Products;	- **materials and products** - **of reject materials**	
		5. General Policy for Process Validation.	- **for process validation** - **general policy for**	

Na seção 5 *Production,* 7 dos 10 *bundles* semelhantes encontrados (ver Tabela 8) estabeleceram relação de coesividade por repetição simples com os textos dos sub-títulos, conforme apresentado na coluna *Lexical bundles* semelhantes (em **negrito**). A maioria deles associou-se ao sub-título de nº 2, com seqüência de palavras repetidas que reiteraram o assunto da sub-seção. Observamos que os *bundles* apresentaram em seu conteúdo palavras do tipo: *products*; *materials*; e *process,* sugerindo uma ligação semântica com o assunto produção. Ademais, outras palavras, por exemplo: *release; storage; sampling; quarantine;* e *reject,* apontaram para os contextos de fabricação, administração, e armazenagem do produto. Porém, o *bundle in accordance with* não se relacionou de forma coesiva com os tópicos das seções. Dessa forma, pressupomos que esse foi um outro caso tal qual ocorreu na seção 3

Premises and Equipment, ou seja, trata-se de uma seqüência de palavras de uso comum da língua.

Quanto à forma, nessa seção, notamos que os *bundles* semelhantes, em sua composição, apresentavam preposição + grupo nominal, por exemplo, os *bundles*: *for process validation*; e *of starting materials*, e formas de grupos nominais, por exemplo, o *bundle*: *sampling quarantine release*. Provavelmente, a composição desses *bundles* remete as atividades da empresa.

A seguir, apresentamos a seção 6 *Quality Control* com suas seções e *bundles* semelhantes.

Quadro 15. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 6 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes	Outros *bundles* semelhantes
6	Quality Control	1. Activities of Quality Control Department.	- *the* **Quality Control** - *of the* **quality** - **Quality Control Department** - **activities of** *the*	- the release of - of finished products - description of the - is responsible for

Na seção 6, dos 8 *bundles* semelhantes (ver Tabela 9), encontramos um único *bundle* (*Quality Control Department*), apresentado na coluna *Lexical bundles* semelhantes (em **negrito**), que se repetiu de forma idêntica no único sub-título da seção. Os outros 3 *bundles*, apresentaram palavras de conteúdo que se repetiram no título e no sub-título da seção. Além do mais, todos esses 4 *bundles*, apresentaram a expressão *Quality Control* incorporada, remetendo ao assunto (título) principal da seção.

Quanto aos *bundles* relacionados na coluna Outros *bundles* semelhantes (em sublinhado), os quais foram: *the release of*; *of finished products*; *is responsible for*; e *description of the*, pressupomos que esses *bundles* podem estar associados às atividades do departamento de Controle de Qualidade da empresa. O departamento de Controle de Qualidade é responsável por monitorar todo o processo de fabricação e liberação dos produtos fabricados. Assim, os *bundles* com as palavras de conteúdo *products* e *release*, apresentados na coluna Outros *bundles* semelhantes, estabeleceram uma relação de significado com o tópico da seção.

Na composição gramatical dos *bundles*, encontramos a formação de artigo + grupos nominais, ou somente grupos nominais, como, por exemplo, os *bundles*: *the Quality Control*; e *Quality Control Department*.

Na seção 7 *Contract Manufacturing and Analysis*, não obtivemos nenhum resultado de *bundles* semelhantes entre as quinze localidades. Portanto, não foi possível aplicar uma análise da relação coesiva com os tópicos. No entanto, podemos pressupor que, conforme já dito, é possível que essa seção apresente informações mais variadas entre as localidades do que as demais, por tratar-se de informações a respeito das condições e exigências locais, resultados já discutidos nas seções anteriores deste capítulo.

Na seqüência, apresentamos a seção seção 8 *Distribution, Complaints and Product Recall*.

Quadro 16. Relação coesiva entre os *bundles* semelhantes e os tópicos da seção 8 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes
8	Distribution, Complaints and Product Recall.	1. Description of Storage and Distribution Practices;	- *of* **complaints** *and*
		2. Records of Distribution.	

Na seção 8, conforme mostra o quadro 16, apenas ocorreu um único *bundle* semelhante (*of complaints and*), o qual foi possível estabelecer apenas a relação de coesão por repetição simples com o título da seção (*Distribution, Complaints and Product Recall*), por meio da palavra *complaints* (em **negrito**). As demais palavras de conteúdo do *bundle* não se repetiram de forma igual nos títulos e sub-títulos. Observamos nessa seção, que a quantidade de *bundles* semelhantes foi pequena e, assim como ocorreu na seção 7, apresentou o conteúdo das informações muito diferente entre as localidades.

Por fim, na seqüência, apresentamos os resultados da seção 9 *Self Inspection*.

Quadro 17. Relação coesiva entre os *bundles* semelhantes e os títulos da seção 9 do SMF.

SEÇÃO	TÍTULO	SUB-TÍTULO	*Lexical bundles* semelhantes
9	Self Inspection	1. Short Description of the Self Inspection System.	**- the self inspection** **- Short description of** **- self inspection system** **- description of the**

Na seção 9 *Self Inspection*, todos os 4 *bundles* semelhantes encontrados (ver Tabela 11) se repetiram de forma idêntica no sub-título da seção, formando o conjunto de palavras que resultou na sentença (*Short Description of the Self Inspection System*). Verificamos que as palavras de conteúdo, *self inspection*, se relacionaram coesivamente por repetição com o título

da seção, expressando o assunto a ser tratado, conforme o título da seção demonstra. Ainda, encontramos a mesma composição formal de grupos nominais, como, por exemplo, os *bundles*: *the self inspection*; e *self inspection system*.

Em suma, concluímos que a relação de coesão por repetição simples do conteúdo dos itens lexicais entre os *bundles* semelhantes, encontrados nas seções e sub-seções das localidades do documento SMF, mostraram que é possível identificar o assunto que é tratado nos textos, possibilitando até mesmo localizar em qual seção e sub-seção a ocorrência desses *bundles* são mais freqüentes. Ademais, indicaram que todos os *bundles* semelhantes, encontrados nessa busca por uma padronização no documento estavam, de fato, coesivos com os títulos e sub-títulos das nove seções do SMF.

A composição gramatical dos *bundles* pôde funcionar como indicadores de significado, ou seja, o *bundle* pode exercer a função de orientar sobre o que se fala no texto. Embora não tivéssemos obtidos uma quantidade maior de *lexical bundles*, em cada uma das nove seções, os resultados mostraram que todos os *bundles* semelhantes tratavam dos assuntos exigidos na escrita do SMF.

No sentido estrito da palavra, os *lexical bundles* foram semelhantes em suas composições, tanto lexicais como gramaticais. Portanto, os *bundles* se relacionaram coesivamente com o tópico (assunto) das seções e sub-seções do documento.

Enfim, encerramos a apresentação e discussão dos resultados deste trabalho e, a seguir, apresentamos as considerações finais.

CONSIDERAÇÕES FINAIS

Com a finalidade de oferecer uma visão de todo o trabalho, apresentamos um resumo dos resultados obtidos, e das principais conclusões deste estudo, bem como contribuições para futuras pesquisas.

O objetivo principal do trabalho foi examinar um tipo de documento do setor farmacêutico, denominado *Site Master File* (SMF), a partir dos *lexical bundles* recorrentes em quinze exemplares, verificando o que há de comum nas ocorrências lingüísticas entre as localidades em que o SMF circula, como também a relação de significado com os assuntos tratados nos textos. O documento SMF é produzido, oficialmente em língua inglesa, com a finalidade de atender à legislação local estabelecida pelas agências reguladoras de vigilância sanitária de cada país e deve, obrigatoriamente, conter nove seções, abordando todas as informações exigidas para a certificação de qualidade de fabricação.

O *corpus* de estudo foi formatado por meio de uma compilação eletrônica de quinze exemplares do documento SMF que pertencem a uma única empresa do setor farmacêutico, os quais estão distribuídos entre países da Europa, Ásia, América do Norte e América do Sul.

Utilizando ferramentas computacionais, tais como, linguagens de programação *Perl* e *Cygwin*, que tem como principais características auxiliar o usuário na programação de pequenas tarefas ou *scripts*, facilitando a manipulação de textos e processos, e utilizando um aplicativo denominado *Análise Lingüística*, desenvolvido por um especialista da área de Ciências da Computação, em conjunto com a pesquisadora, com a finalidade de gerar as consultas e o cruzamento dos *lexical bundles*, demos início à análise dos dados.

Engendramos seis etapas de procedimentos, a saber: (1) identificar a estrutura organizacional dos sumários do documento; (2) verificar os padrões léxico-gramaticais a partir dos *lexical bundles*; (3) identificar os *lexical bundles* recorrentes entre as seções do documento; (4) comparar os *bundles* semelhantes do *corpus* de estudo com um *corpus* de referência (o BNC); (5) identificar os *lexical bundles* recorrentes em cada localidade do documento; e (6) estabelecer uma relação de coesão entre os *bundles* semelhantes e o tópico das seções do SMF.

Desse modo, obtivemos os seguintes resultados dos procedimentos de análise adotados. Em primeiro lugar, a estrutura organizacional do documento, apresentada no sumário das quinze localidades investigadas, apresentou semelhanças entre as nove seções e suas

respectivas sub-seções. Todas as seções possuíam títulos e subtítulos iguais com a mesma seqüência numérica, assegurando a padronização das informações exigidas pelos órgãos competentes, na elaboração do documento. No entanto, nos itens das sub-seções de cada localidade, cinco das quinze localidades apresentaram variações quanto: (i) à inexistência de itens nas sub-seções; (ii) às diferentes quantidades de itens elencados; e (iii) aos itens com informações diferentes.

Analisando essas variações, identificamos que essas diferenças poderiam indicar que para determinadas localidades, seriam necessárias informações mais complementares, ou informações referentes às localidades. Assim, concluímos que os conteúdos das informações, apresentados nos sumários analisados, seguiam o mesmo padrão exigido e instituído pelo órgão governamental regulatório, isto é, possuíam os mesmos tópicos. Porém, a escrita se apresentou diferente quanto ao detalhamento das informações nas sub-seções.

Em segundo lugar, na investigação dos padrões léxico-gramaticais do documento a partir dos *lexical bundles*, realizamos uma contagem dos *bundles* por localidade e por seção, e descobrimos que independente da quantidade total de *bundles* que determinada localidade apresente, a quantidade de *bundles* de cada uma das nove seções pode variar, reforçando os resultados apresentados na comparação dos sumários. Isso significa, que o conteúdo das informações em cada localidade pode ser diferente, pois contêm informações mais expandidas ou reduzidas no processo de produção escrita. Desse modo, hipotetizamos que as variações encontradas poderiam estar associadas aos negócios da empresa da seguinte forma: (i) ao tipo de produto fabricado na localidade; (ii) à organização administrativa interna; e (iii) às políticas regulatórias de vigilância locais, bem como as escolhas lingüísticas feitas pelos escritores na composição do documento.

Em terceiro lugar, na busca por *lexical bundles* semelhantes, adotamos um critério de corte na metodologia para identificarmos os *bundles* semelhantes (mínimo de oito localidades), e verificarmos a qual seção os *bundles* pertenciam. Assim, obtivemos os seguintes resultados: (i) não houve ocorrência de *bundles* semelhantes no cruzamento das quinze localidades; (ii) em uma das nove seções do documento não ocorreu nenhum *bundle* semelhante; (iii) a freqüência de *bundles* semelhantes entre as seções foi muito variada. Na primeira seção, obtivemos 29 *bundles* semelhantes, enquanto que nas demais seções, obtivemos entre 1 a 10 *bundles* semelhantes. Esses resultados sinalizaram que as diferenças, ou estariam ligadas às escolhas lingüísticas feitas por quem produziu o SMF em cada localidade, ou ao fato de o assunto ser tratado linguisticamente diferente.

Sob um outro viés, considerando a proximidade geográfica das localidades, verificamos que a maior parte da ocorrência de *bundles* semelhantes estava concentrada em regiões da Europa. Já os demais *bundles*, em sua minoria, estavam concentrados em regiões mais afastadas, como países da América e Ásia. Concluímos, portanto, que essa proximidade poderia estar associada aos negócios da empresa e, por conseguinte, influenciar na distribuição dos *bundles* semelhantes entre as localidades.

Com relação à proximidade geográfica, em virtude dos negócios da empresa estar influenciando na produção escrita do documento, inferimos dois principais aspectos. O primeiro aspecto, diz respeito às localidades mais próximas apontarem para as regiões onde estão alocadas as unidades de produção dos medicamentos, as quais possuem, entre outras, a missão de exportar os seus insumos a outros países. Por outro lado, as localidades mais afastadas recebem, via importação, os produtos a serem distribuídos e comercializados. A partir disso, surge uma necessidade de se escrever sobre outros assuntos, descrever outras atividades. No segundo aspecto, verificamos que as regiões mais próximas participam de uma mesma convenção que regulamenta as drogas farmacêuticas na Europa, enquanto que as regiões mais afastadas estão submetidas a outros órgãos de vigilância sanitária. Portanto, um número maior, ou a ausência de *bundles* semelhantes, pode estar ligado a fatores locais, ou a fatores relacionados à natureza dos negócios da empresa, interferindo na escrita do documento.

Em quarto lugar, quanto à representatividade do *corpus* de estudo, identificamos que a maioria dos *bundles* semelhantes encontrados ocorre, proporcionalmente, com maior freqüência nos documentos do SMF do que no *corpus* do BNC, ou seja, pudemos concluir que os *bundles* do SMF são típicos do domínio específico do documento.

Enfim, na última etapa, quanto à relação coesiva entre os tópicos das seções do SMF, identificamos que a maioria dos *bundles* semelhantes mantém uma relação de coesão com o conteúdo das informações das nove seções do SMF. Essa coesão apresentou uma relação por repetição simples com os *lexical bundles* semelhantes encontrados nos títulos e sub-títulos das seções, confirmando uma relação semântica com o assunto tratado nos tópicos do SMF. Dessa forma, por meio dos *lexical bundles*, foi possível identificar o assunto tratado nas seções e sub-seções, bem como localizar onde os itens ocorrem

Em resumo, o documento *Site Master File*, embora tenha apresentado em sua organização estrutural uma mesma padronização com relação a todas as quinze localidades investigadas, não apresentou a mesma semelhança na sua estrutura interna. Foram apuradas

poucas semelhanças entre a escrita dos quinze exemplares, respondendo, assim, a inquietação inicial desta pesquisa, que foi tentar investigar qual era o grau de conformidade com elementos lingüísticos, um documento normatizado e regulamentado, escrito por diferentes autores em diferentes partes do mundo poderia atingir.

Dessa maneira, acreditamos que com os resultados e discussões da presente pesquisa, concluímos que esse tipo de documento apresentou diferentes escolhas lingüísticas, as quais podem ser exploradas de forma mais aprofundada, se fizermos uso de outros instrumentos de análise, além dos *lexical bundles*, como, por exemplo: palavras-chave (*keywords*); colocações (collocations), entre outros, tendo como propósito auxiliar os escritores na produção escrita do conjunto de textos que compõem o *Site Master File*.

Ainda, os *lexical bundles* semelhantes encontrados nesta pesquisa, contribuíram para identificar termos específicos que fazem parte da escrita do documento, apontando em qual seção do documento esses termos podem ocorrer, justificando o uso dessas seqüências de palavras na elaboração de um vocabulário que auxilie os escritores na escrita do SMF.

Uma outra contribuição que surgiu desta pesquisa é a de que os escritores do SMF necessitam adquirir, além dos conhecimentos das exigências globais para a produção escrita do documento, conhecimentos das condições locais. Porquanto, apesar dos fatores globais que governam o documento para atender a certificação de qualidade dos produtos farmacêuticos no mundo, há de se considerar os fatores locais que demandam informações diferentes, dependendo das exigências de cada localidade, em particular, informações referentes aos órgãos governamentais e ao tipo de negócio realizado nas diversas localidades.

Portanto, algumas das contribuições e implicações que este estudo oferece para as pesquisas acadêmicas estão relacionadas aos seguintes itens: (i) ao tipo de texto; (ii) à área específica; (iii) ao ensino de línguas estrangeiras (iv) às ferramentas computacionais; (v) às áreas da Lingüística de *Corpus* e Lingüística Computacional; e (vi) ao uso de seqüência de palavras nos estudos lingüísticos.

Quanto ao tipo de texto, não encontramos, na literatura pesquisada, textos originários do contexto de negócios, exceto Berber Sardinha (2003) que fez uso de um único exemplar de uma empresa do setor financeiro para analisar gênero textual, utilizando a abordagem de *lexical bundles*. A maioria dos pesquisadores analisou *lexical bundles* utilizando *corpora* de registros acadêmicos. Dessa forma, este estudo contribuiu no sentido de apresentar quinze exemplares do contexto empresarial, escritos por profissionais envolvidos no segmento farmacêutico.

Uma das implicações deste estudo refere-se à especificidade farmacêutica, pois essa área carece de estudos lingüísticos que possam auxiliar no ensino de língua inglesa para os profissionais que necessitam adquirir conhecimentos a fim de se comunicarem, tanto nas formas escrita como oral. Assim sendo, por meio das variações encontradas neste estudo, pudemos levantar algumas necessidades que a escrita do documento demandam, oferecendo meios para o entendimento dos textos escritos por diferentes autores de uma mesma comunidade específica, gerando uma alternativa pedagógica no ensino de línguas para fins específicos.

A criação de um aplicativo para consultas e análise de *lexical bundles* é uma outra importante contribuição deste estudo. O desenvolvimento de uma ferramenta computacional para uso desta pesquisa, além de agregar valores aos avanços tecnológicos na área de Lingüística de *Corpus* e Lingüística Computacional, também desperta a possibilidade de se construir outras modalidades a partir desse aplicativo, as quais poderão integrar um conjunto de instrumentos computacionais que atendam outros tipos de análises lingüísticas, como, por exemplo: consultas integradas via *web*; cruzamento de uma variedade de *corpus* para identificar e distinguir elementos lingüísticos de determinados gêneros e tipos textuais; entre outros.

O uso da seqüência de palavras (os *three-bundles)* contribuiu para evidenciar aspectos fraseológicos, localizando e identificando os assuntos tratados no texto, bem como implicando em futuros estudos para explorar terminologias que permeiam os textos de especialidades. Porém, outros trabalhos podem fazer uso de seqüências maiores, como quatro ou mais palavras, possibilitando abranger um espectro mais amplo nas relações entre as sentenças ou frases do texto.

Apesar de ainda não existir muitas pesquisas sobre *lexical bundles* no Brasil, acreditamos que com os resultados deste trabalho, esse seja um primeiro passo para fomentar o desenvolvimento de novos estudos relacionados às diversas áreas da linguagem.

REFERÊNCIAS BIBLIOGRÁFICAS

Aijmer, K. & Altenberg, B. (1991). *English Corpus Linguistics: studies in honour of Jan Svartvik*. London: Longman.

Barbara, L. & Berber Sardinha, T. (2008). Lingüística de Corpus e Análise de Discurso. In C. R. Caldas-Coulthard & L. Scliar-Cabral (orgs), *Desvendando discursos: conceitos básicos* (pp. 289-317). Florianópolis: Editora da UFSC.

Batista, M.E. (1998). *E-mails na troca de informações numa multinacional: o gênero e as escolhas léxico-gramaticais* (Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, 1998). Disponível em: <http://www.pucsp.br/pos/lael/lael-inf/def_teses.html>

Bazerman, C. (2006). *Gêneros Textuais, tipificação e interação* (2ª ed.). (Tradução e adaptação Hoffnagel, J. C). São Paulo: Cortez.

Berber Sardinha, T. (1994a). Lexical Patterns for 'Year' in Annual Business Report. Working Paper 9. Direct Papers. CEPRIL, PUC-SP, São Paulo, Brazil/ AELSU, English Department, Liverpool University, Liverpool, UK.

Berber Sardinha, T. (1994b). Collocations in a introduction to Annual Business Report. Working Paper 11. CEPRIL, PUC-SP, São Paulo, Brazil/ AELSU, English Department, Liverpool University, Liverpool, UK.

Berber Sardinha, T. (1997). *Automatic identification of segments in written texts*. Tese de Doutoramento. University of Liverpool, England, UK.

Berber Sardinha, T. (2000). Lingüística de Corpus: histórico e problemática. *D.E.L.T.A., 16* (2), 323-367. doi: 10.1590/S0102-44502000000200005.

Berber Sardinha, T. (2002). Tamanho de Corpus. *The ESPecialist, 23* (2), 103-122.

Berber Sardinha, T. (2003). Análise de gêneros e Lingüística de Corpus: Identificação das unidades internas do gênero por meio de padronização lexical. *DIRECT Papers, 51*, 1-30. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers51.pdf >

Berber Sardinha, T. (2004a). *Lingüística de Corpus*. São Paulo: Manole.

Berber Sardinha, T. (2004b). Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem – ReVEL, 3* (2), 1-5.

Berber Sardinha, T. (2004c). Informatividade, interatividade e narratividade na reunião de negócios – Análise Multidimensional e palavras-chave. *DIRECT Papers, 52*, 1-25. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers52.pdf >

Berber Sardinha, T. (no prelo). *Pesquisa em Lingüística de Corpus com WordSmith Tools*. São Paulo.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English.* London: Longman.

Biber, et al. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A.Wilson; P. Rayson e T. McEnery (eds). *Corpus Linguistics by the Lune* (pp. 71-93). Frankfurt: Peter Lang.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25* (3), 371-405.

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: Benjamins.

Bressane, T. B. R. (2000). *Construção de identidade numa empresa em transformação*. Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, Brasil.

Brown, G., & Yule, G. (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.

Carvalho, K. R. R. (2003). *Apresentações Orais de trabalhos científicos na especialidade de pneumologia* (Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, 2003). Disponível em: < http://www.pucsp.br/pos/lael/lael-inf/def_teses.html >

Cortes, V. (2002). *Lexical bundles in Published and Student Academic Writing in History and Biology*. Tese de Doutoramento, Northern Arizona University, Flagstaff, Arizona.

Cortes, V. (2006). Teaching bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17*, 391-406.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3 (1), 43-57. doi: 10.3366/E1749503208000063.

Crystal, D. (1997). *English as a Global Language*. Cambridge: Cambridge University Press.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3 (1), 59-80.

Eggins, S. (1994). *An Introduction to systemic functional linguistics*. London: Pinter.

Freire, M. (1995). Computer-mediated communication in the workplace. Working Paper 22. Direct Papers. CEPRIL, PUC-SP, São Paulo, Brazil/ AELSU, English Department, Liverpool University, Liverpool, UK.

Gil, A.C. (2002). *Como Elaborar Projetos de Pesquisa* (4ª ed). São Paulo: Atlas.

Graddol, D. (2006). *English Next* – Why global English may mean the end of ´English as a Foreign Language´. London: British Council. Disponível em <http://www.britishcouncil.org/learning-research-english-next.pdf>

Halliday, M.A.K. (1976). *Cohesion in English*. London: Longman.

Halliday, M.A.K. & Hasan, R. (1989). *Language, Context and Text: Aspects of Language in a Social – Semiotic Perspective*. Geelong, Victoria, Australia: Deakin University.

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In K.Aijmer & B.Altenberg (eds.). *English Corpus Linguistics: studies in honour of Jan Svartvik* (pp. 30-43). London: Longman.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hyland, K. (2007). Is there an ‘academic Vocabulary’? *TESOL Quarterly*, 41 (2), 235-253.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27* (1), 4-21. doi:10.1016/j.esp.2007.06.001.

Leech, G. (1991). The state of the art in corpus linguistics. In K.Aijmer & B.Altenberg (orgs.). *English Corpus Linguistics: studies in honour of Jan Svartvik* (pp. 8-29). London: Longman.

Levy, S. A. (2003). *Lexical Bundles in Professional and Student Writing*. Tese de Doutoramento, University of the Pacific, Stockton, California.

Lima-Lopes. R. (2001). *Estudos de transitividade em língua portuguesa: o perfil do gênero carta de vendas*. Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, Brasil.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martin, J.R. (1992). *Context: register, genre and ideology. English text – systems and structure*. Philadelphia/ Amsterdam: John Benjamins.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Partington, A. & Morley, J. (2004). From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In Lewandowska-Tomaszczyk (ed). *Practical Aplications in Language and Computers – PALC 2003*. Frankfurt: Peter Lang, pp. 179-92.

Phillips, M. (1989). *Lexical Structure of text* [Discourse Analysis Monographs 12]. Birmingham, University of Birmingham, Department of English.

Ramalho, J. A. (1999). *SQL: A Linguagem dos Bancos de Dados*. São Paulo: Berkeley.

Ramos, R. C. G. (1997). *Projeção de Imagem através de Escolhas Lingüísticas: um Estudo no Contexto Empresarial*. Tese de Doutoramento, Pontifícia Universidade Católica de São Paulo, Brasil.

Rocha, M. (2007). Métodos estatísticos comuns em Lingüística de Corpus: visão geral. In R. M. Gerber & V. Vasilévski (orgs.), *Um percurso para pesquisas com base em corpus* (pp. 194-221). Florianópolis: Editora da UFSC.

Rossini, A. M. Z. P. (2005). *A linguagem dos contratos bancários internacionais em inglês: um estudo descritivo baseado em Lingüística de Corpus* (Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, 2005). Disponível em: <http://www.pucsp.br/pos/lael/lael-inf/def_teses.html>

Santos, V. B. M. P. (1999). *Padrões interpessoais no gênero de cartas de negociação*. (Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, 1999). Disponível em: < http://www.pucsp.br/pos/lael/lael-inf/def_teses.html >

Santos, V. B. M. P. (2002). O perfil das comunicações internas escritas de uma empresa brasileira: um estudo de caso sobre o contexto de produção e as realizações discursivas em locais de trabalho. Tese de Doutoramento, Pontifícia Universidade Católica de São Paulo, Brasil.

Scott, M. (1997). *Wordsmith Tools Manual.* Oxford: Oxford University Press.

Scott, M. & Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1994). Trust the text. In M. Coulthard (ed.). *Advances in Written Text Analysis* (pp. 12-25). Londres: Routledge.

Sobhie, M. (2003). Stages in Business-to-Business Brochures. *DIRECT Papers*, *49*, 1-14. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers49.pdf >

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics., 7* (2), 215-244.

Stubbs, M. (2007). An example of frequent English phraseology: distribution, structures and functions. In R. Facchinetti (ed.). *Corpus Linguistics 25 Years on* (pp. 89-106). New York, Amsterdam: Rodopi.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tagnin, S. (2007). A Lingüística de Corpus na Universidade de São Paulo – o projeto COMET. In R. M. Gerber & V. Vasilévski (orgs.), *Um percurso para pesquisas com base em corpus* (pp. 166-173). Florianópolis: Editora da UFSC.

Teixeira, E. D. (2007). Etiquetagem em Lingüística de *Corpus*: Possibilidade de Aplicação. In R. M. Gerber & V. Vasilévski (orgs.), *Um percurso para pesquisas com base em corpus* (pp. 116-148). Florianópolis: Editora da UFSC.

Thompson, G. & Ramos, R. C. G. (1995). Ergativity in the analysis of business text. Working Paper 21. Direct Papers. CEPRIL, PUC-SP, São Paulo, Brazil/ AELSU, English Department, Liverpool University, Liverpool, UK.

Vasilévski, V. (2007). Aspectos histórico-teóricos da Lingüística de Corpus: Surgimento, abandono e uso. In R. M. Gerber & V. Vasilévski (orgs.), *Um percurso para pesquisas com base em corpus* (pp. 46-62). Florianópolis: Editora da UFSC.

Vian Jr., O. (1997). *Conceito de gênero e análise de textos de vídeos institucionais* (Dissertação de Mestrado, Pontifícia Universidade Católica de São Paulo, 1997). Disponível em: <http://www.pucsp.br/pos/lael/lael-inf/def_teses.html>

ANEXO 01

SCRIPT 1 (usando Shell do Cygwin)

```
#!/bin/bash
# Rotina para limpeza do texto
# Autor: Luciene Novais Mazza

# substitui os caracteres de controle por um espaço
tr '[:cntrl:]' ' ' < fonte.txt |

# deleta a pontuação ,;.:-
tr -d '[:punct:]' |

# substitui os digitos por zeros
tr '[:digit:]' '0'|

# resume os zeros
tr -s '0' |

# resume os espaços
tr -s ' ' |

# substituição dos nomes
sed 's/nomeempresa1/PHARMA/g' |
sed 's/nomeempresa2/PHARMA/g' |
sed 's/nomeempresa3/PHARMA/g' |
sed 's/nomeempresa4/PHARMA/g' |
sed 's/nomeempresa5/PHARMA/g' > fonte1.txt
```

ANEXO 02

SCRIPT 2 (usando o Shell do Cygwin)

```
#! \bin\bash
# Rotina para criar os Threegrams
# Autor: Luciene Novais Mazza

# coloca cada palavra do texto numa linha
tr ' ' '\n' < fonte1.txt > words1.txt

# desloca a sequência uma palavra para baixo
tail +2 words1.txt > words2.txt

# desloca a sequência duas palavras para baixo
tail +3 words1.txt > words3.txt

# concatena as três palavras e classifica os pacotes
paste words1.txt words2.txt words3.txt | sort |

# substitui o tab por um espaço
tr '\t' ' ' > threegrams.txt

# formata a saída
unix2dos threegrams.txt
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo