Calidoscópio
E-ISSN: 2177-6202
calidoscopio@unisinos.br
Universidade do Vale do Rio dos Sinos
Brasil

Novais Mazza, Luciene
Processamento linguístico-computacional de pacotes lexicais: um estudo de corpus na área de Regulamentação Farmacêutica
Calidoscópio, vol. 13, núm. 3, septiembre-diciembre, 2015, pp. 424-439
Universidade do Vale do Rio dos Sinos

Calidoscópio
Vol. 13, n. 3, p. 424-439, set/dez 2015
Unisinos - doi: 10.4013/cld.2015.133.13

Luciene Novais Mazza
lucienenovais@uol.com.br
lucienemazza@unip.br

Processamento linguístico-computacional de pacotes lexicais: um estudo de *corpus* na área de Regulamentação Farmacêutica

Computational-linguistic processing of lexical bundles: A corpus-based study in the area of Pharmaceutical Regulation

RESUMO – Este trabalho tem por objetivo demonstrar um aplicativo computacional desenvolvido para a extração de pacotes lexicais de três palavras e apresentar por meio deste as unidades lexicais recorrentes entre documentos de especialidade. O método quantitativo aplicado, em princípio, explora um tipo de texto produzido pelas indústrias do setor farmacêutico, o qual está diretamente relacionado a assuntos regulatórios no âmbito das agências internacionais de vigilância sanitária. No entanto, os procedimentos de análise podem ser adotados para investigar outros aspectos linguísticos dentre a variedade de gêneros e tipos textuais, como também possibilita a identificação de termos. O estudo tem como principal enfoque a frequência de ocorrência dos padrões lexicais em *corpus* autêntico da língua em uso por meio de ferramentas linguístico-computacionais, em particular nas pesquisas voltadas ao estudo da linguagem em contextos empresariais, e busca multiplicar os trabalhos de Douglas Biber com base na combinação de palavras recorrentes em *corpora* específicos. O referencial teórico-metodológico baseia-se na Linguística de *Corpus*, que é capaz de dialogar, especificamente, com a Linguística Computacional e oferecer meios para o desenvolvimento do aplicativo e ao processamento dos pacotes lexicais. O *corpus* coletado reúne quinze exemplares do documento escrito na língua inglesa, totalizando cerca de 110 mil palavras, cuja delimitação contempla diferentes localidades do mundo, envolvendo vários autores. Os resultados desvelam a possibilidade de investigação nas divisões internas dos textos mediante o cruzamento entre documentos de uma mesma especialidade.

Palavras-chave: pacotes lexicais, *corpus* de especialidade, ferramenta linguístico-computacional.

ABSTRACT – The present paper aims to demonstrate a computational tool developed to extract three-word lexical bundles and show – by working through this – the automatic recognition of recurring lexical items among regulatory documents. In this quantitative analysis a specific document prepared by pharmaceutical industries (in which the matter is directed related to the public health protection agencies) is generally examined. Nonetheless, the quantitative data collection methods can also be used to search any other linguistics features within a variety of genres and specific type of documents and it allows the linguistics researcher to easily identify which terms fall under a domain of specific texts. The study focus their main concern on investigating lexical pattern frequency of language use, particularly across the current context of business, and it seeks to spread Douglas Biber works based on recurrent word combinations that makes use of tools and techniques developed in corpus-based linguistics. As the theoretical framework for this study we primarily draw upon Corpus Linguistics, a theory that is able to connect its concepts over the computational assumptions and design tools for end users and extract the lexical bundles as well. The collected corpus gathers documents in English from fifteen different manufacturing sites of a multinational Pharmaceutical company, totaling about 110,000 words, whose limits include different writers among different geographic parts of the world. The investigation shows that it is possible to search text-internal features by the extraction of lexical bundle between data across the same specific-domain document.

Keywords: lexical bundles, domain-specific corpus, linguistic-computational tool.

Introdução

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 socioeconômicos e culturais e exigindo um sistema de comunicação rápido e padronizado, ou seja, uma linguagem universal que ultrapassasse fronteiras.

No entanto, esse processo acelerado de difusão dos mercados globais transcende 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 as condições locais, principalmente as situações que envolvem pessoas, isto é, a mão-de-obra especializada contratada para escrever, pensar e agir de acordo com as políticas estabelecidas pela cultura organizacional da empresa matriz. Com efeito, essas condições podem desencadear diferenças no modo de

redigir os diversos tipos de texto de uma dada especialidade em uma língua globalizada, melhor dizendo, pode haver interferências quer decorrentes de uma comunicação intercultural quer da diversidade nas relações interempresariais.

Nas ciências da saúde, em particular no campo das ciências farmacêuticas do segmento industrial, existem muitos documentos técnicos, como, por exemplo: laudos, métodos analíticos, guias direcionados às práticas inerentes à produção, análise e controle de drogas farmacêuticas, entre outros. Esses documentos servem de base aos especialistas que atuam em laboratórios e controle de qualidade a fim de realizar as suas atividades profissionais. Também, outros documentos de natureza jurídica e regulamentados por normas e procedimentos governamentais do setor devem ser produzidos pelos profissionais que desempenham funções administrativas na empresa. Todavia, percebe-se que esses profissionais não possuem habilidades linguísticas adequadas para atender às atividades que demandam uma produção escrita; em virtude, talvez, da formação acadêmica e/ou profissional específica, tais como: farmácia, química, biologia, medicina, engenharia, e áreas afins.

Assim, inserido nesse cenário e dentre a variedade de textos de especialidade produzidos em língua inglesa e veiculados nas indústrias farmacêuticas, temos o documento *Site Master File* (doravante SMF)[1]. O SMF é um dos principais documentos que uma empresa fabricante de medicamentos deve portar. O documento trata de um conjunto de textos produzido especificamente para cada empresa fabricante e tem como objetivo principal certificar a garantia e qualidade dos produtos fabricados. Essa certificação é guiada por princípios estabelecidos mundialmente pelos órgãos governamentais de vigilância sanitária.

Para cada unidade de negócios (na sigla em inglês BU-*Business Unit)* deve-se elaborar e providenciar o trâmite do seu próprio SMF e, em seguida, apresentá-lo ao organismo público de proteção à saúde de cada país em que a unidade estiver instalada. Os colaboradores designados à elaboração do documento estão distribuídos entre os diversos departamentos da empresa, e a equipe responsável pela aprovação e expedição do SMF está alocada no departamento de Garantia da Qualidade – departamento que converge todas as informações referentes ao atestado de qualidade do produto e ao controle das operações de produção. Para a elaboração do SMF, é necessário um conhecimento linguístico na área de especialidade farmacêutica que possibilite aos profissionais dos diversos departamentos redigirem de forma coesa e coerente os documentos que circundam o contexto no qual estão inseridos.

Entre as razões apontadas acima, emerge a motivação para investigar o documento SMF dando ênfase a três aspectos: em primeiro lugar, a necessidade profissional da autora em traduzir documentos do setor farmacêutico; em segundo lugar, as contribuições para a área de processamento de língua natural na criação e aplicação de ferramentas computacionais para os estudos linguístico-discursivos baseados em *corpora* de especialidades produzidos por profissionais de negócios; e por último, o fomento à construção de um banco de dados terminológico bilíngue, nos pares de línguas inglês e português, que compreende a área da Farmácia, subárea de Gestão Industrial Farmacêutica – Assuntos Regulatórios de Medicamentos, com o propósito de atender aos interesses de tradutores técnicos, colaboradores/profissionais e pesquisadores e estudantes da área ou subárea em questão.

Em vista disso, este trabalho busca examinar um tipo de documento do setor farmacêutico com base na investigação de pacotes lexicais recorrentes entre quinze exemplares, partindo da hipótese de que é possível haver uma regularidade nos padrões léxico-gramaticais, uma vez que o SMF apresenta uma estrutura organizacional semelhante, com seções e tópicos em uma mesma ordem sequencial e temática. Acrescentamos ainda que, na literatura acadêmico-científica pesquisada, não encontramos registros de trabalhos voltados ao estudo da linguagem no contexto empresarial farmacêutico. No entanto, foram encontradas pesquisas desenvolvidas por pesquisadores do Projeto DIRECT (*Development of International Research in English for Commerce and Technology*)[2] da Pontifícia Universidade Católica de São Paulo (PUC-SP). Tais pesquisas concentram-se em examinar a organização retórica que define como os documentos nas empresas se organizam, e a léxico-gramática de textos de especialidade.

Dentro da abordagem da Linguística de *Corpus* (doravante LC) (Biber *et al*., 1998; Berber Sardinha, 2004) com relação à categoria dos pacotes lexicais, encontramos 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 Linguística Aplicada) a fim de extrair pacotes lexicais baseados numa análise contrastiva das suas formas e funções; (ii) Scott e Tribble (2006), que realizaram, a partir de textos escritos extraídos do BNC, um levantamento do conjunto ou agrupamento de

[1] A criação do SMF é devida à necessidade de uma sistematização para a forma composicional do conteúdo pertinente à validação das regras internacionais de Boas Práticas de Fabricação Farmacêutica (BPF). Daí, a deliberação, no ano de 1995 pelo programa PIC/S (*Pharmaceutical Convention Co-Operation Scheme*), de uma nota explicativa (*Explanatory Notes for Pharmaceutical Manufacturers on the Preparation of a Site Master File*) elencando detalhadamente as etapas de construção desse documento. Consultar Picscheme (s.d.).

[2] O DIRECT é um projeto iniciado em 1989 e desenvolvido pelo Programa de Pós-Graduação em Linguística 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 às habilidades de comunicações específicas no uso das línguas inglês e português para propósitos de negócios. Tais trabalhos estão disponíveis eletronicamente em Lael (s.d.).

palavras (*cluster*) para verificar a variação estatística e a função de determinadas palavras na formação dos pacotes; (iii) Cortes (2006), que investigou o uso das combinações recorrentes de palavras em registros escritos do contexto da disciplina de História, visando à apreensão de pacotes lexicais recorrentes por parte dos alunos universitários para a escrita de trabalhos acadêmicos; (iv) Berber Sardinha (2003), que estudou os pacotes lexicais recorrentes nas unidades internas de um documento da área financeira com o objetivo de estabelecer os elos (*links*) de coesão com as sentenças do texto e; (v) Levy (2003), que se baseou na gramática de Biber *et al*. (1999) para pesquisar a co-ocorrência de pacotes lexicais dentre uma variedade de registros das áreas profissionais e acadêmicas, visando estabelecer um parâmetro de proficiência de alunos nativos e não nativos na língua inglesa, alunos esses provenientes de uma universidade norte-americana.

Assim, no percurso deste artigo, primeiramente apresentamos os dois eixos teórico-metodológicos norteadores do presente estudo, que são: os pressupostos da Linguística de *Corpus*, sob o enfoque dos *corpora* computadorizados, e o conceito de pacote lexical. Em um segundo momento, demonstramos 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. Logo após, discutimos a análise dos resultados apurados na extração dos pacotes lexicais semelhantes e a comparação desses pacotes com um *corpus* de referência (o BNC – *British National Corpus*), estabelecendo os parâmetros de conformidade apurados.

Para concluir, sustentamos a tese de que o mercado de trabalho corporativo necessita aperfeiçoar e tomar conhecimento da produção escrita de gêneros e tipos de textos específicos redigidos em língua inglesa ou em qualquer outra língua, de tal forma a suprir as necessidades de comunicação internacional nas transações entre seus clientes e parceiros comerciais.

A LC e os *corpora* computadorizados

Segundo Berber Sardinha (2004), a Linguística 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 (Berber Sardinha, 2004, p. 3).

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

Embora o uso de textos autênticos com base em *corpora* venha sendo explorado desde a antiguidade, houve, segundo Aijmer e Altenberg (1991, p. 1), uma grande expansão da LC nas últimas décadas. Segundo os autores, essa expansão é devida a dois importantes eventos ocorridos no início da década de 1960. O primeiro evento foi o lançamento de um *corpora* do inglês britânico – o *The Survey of English Usage* (SEU), contendo 1.000.000 de palavras de textos de língua escrita e falada, coordenado por Randolph Quirk e Greenbaum (Svartvik e Quirk, 1980). O segundo evento foi o advento da computação que viabilizou o armazenamento e o processamento de um grande volume de dados linguísticos.

Paralelo a esses acontecimentos, nessa mesma década surgiram duas referências à prospecção dos *corpora* na língua inglesa: o *The Brown Corpus*, um *corpora* de 1.000.000 de palavras do inglês americano, compilado por Henry Kučera e Nelson Francis (Kučera e Francis, 1967); e o *The Lancaster-Oslo/Bergen Corpus* (LOB Corpus) (Hofland e Johansson, 1982), um *corpora* no mesmo formato do *Brown Corpus* capaz de comparar as variantes da língua e contendo 1.000.000 de palavras do inglês britânico – compilado e computadorizado pelos então pesquisadores Geoffrey Leech (*Lancaster University*), Stig Johansson (*University of Oslo*), e Knut Hofland (*University of Bergen*).

Além do pioneiro *Brown Corpus*, outros dois principais *corpora* eletrônicos na língua inglesa são: o *British National Corpus* (BNC) (Garside *et al.*, 1997), contendo 100 milhões de palavras; e o *The Bank of English* (Sinclair, 1995), contendo em média 4,5 bilhões de palavras. Conforme Berber Sardinha (2004, p. 9), o *Brown Corpus* (1967) e o BNC (1995) são *corpora* de amostragem, planejados e fechados, enquanto que o *The Bank of English* (1995) é um *corpus* dinâmico, ou seja, em expansão. De acordo com Tagnin (2007, p. 160-162), o uso dos diferentes tipos de *corpora* existentes depende do objetivo da pesquisa. A autora classifica os *corpora* de quatro formas, conforme segue:

(i) Fechados e abertos: o BNC e o *Bank of English* (em torno de mais de 500 milhões de palavras, e em constante atualização), constituído de textos em inglês britânico, americano, australiano e canadense.

(ii) Monolíngues ou multilíngues: em uma só língua, como o BNC, e em duas ou mais línguas, como o COMPARA[3].

[3] O COMPARA é um *corpus* paralelo contendo trechos de obras literárias em variantes da língua portuguesa e língua inglesa com suas respectivas traduções.

(iii) Comparáveis ou paralelos: os comparáveis são entendidos como àqueles 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.
(iv) De língua geral ou de língua de especialidade: os de língua geral são compostos por diversos gêneros para assegurar a representatividade da língua usada no cotidiano, e os de especialidade são compilados para fins específicos, sendo bastante restritos.

Conforme a classificação apontada acima, é possível atribuir ao nosso *corpus* de estudo a categoria de *corpora* de especialidade, compilados para fins específicos com acesso restrito aos usuários do segmento farmacêutico. Isso significa dizer que o documento somente pode ser manuseado e veiculado entre e pelos departamentos da empresa, ou, então, por instituições governamentais vinculadas à saúde da população. Portanto, estamos tratando de uma língua de especialidade representativa do seu domínio de atuação.

Referente às características dos *corpora* existentes para as pesquisas linguísticas, é relevante 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 textos. Por conseguinte, as novas tecnologias para o Processamento de Linguagem Natural (PLN)[4] trouxeram modernos *softwares* com ferramentas capazes de processar uma grande massa de textos em um curto espaço de tempo.

Outra premissa que serve de base às pesquisas ancoradas na LC é a frequência de uso das palavras. De acordo com o britânico Michael Halliday (1991), o "sistema" linguístico foi sempre inerentemente probabilístico e, por consequência, há frequência nos textos e na representação de probabilidades na gramática. Em outras palavras, na perspectiva do autor, o estudo de *corpus* ocupa uma posição central na investigação da linguagem e serve como um dos muitos caminhos a serem explorados na descoberta de evidências relativas às frequências encontradas na gramática. Halliday (1991, p. 30), citando Svartvik, acrescenta que o uso de *corpus* pode estabelecer uma taxionomia de classes gramaticais passível de se calcular a proporção das categorias encontradas no texto, oferecendo a possibilidade de relacionar essas categorias a uma extensão de registros variados. Essa proposta de taxionomia de classes foi explorada por Biber *et al*. (1999) em seus trabalhos referentes às formas estruturais e funcionais dos elementos linguísticos, por exemplo, os pacotes lexicais que discutimos na próxima seção.

O conceito de pacotes lexicais

Um dos principais fundamentos propostos para este trabalho baseia-se nos estudos das combinações de palavras recorrentes no texto com base em *corpus* específico. Essa combinação de palavras foi nomeada, por Biber *et al*. (1999), *lexical bundles* e traduzida na língua portuguesa para pacotes lexicais[5]. Assim sendo, a fim de engendrarmos essa discussão, destacamos algumas definições empreendidas por linguistas aplicados cujas pesquisas são baseadas em *corpora* para a extração de pacotes lexicais.

Biber *et al*. (1999) definem pacotes lexicais como "sequências de palavras que ocorrem naturalmente no discurso" (Biber *et al*., 1999, p. 990)[6]. De outro modo, podemos dizer que são pacotes formados por expressões recorrentes, independente de sua idiomaticidade ou de sua condição estrutural. Alguns dos exemplos em língua inglesa considerados pelos autores como pacotes lexicais são: *the end of the*, *in addition to the*, *the point of view of*, entre outros.

Cortes (2006, p. 392) define pacote lexical como sequências de três ou mais palavras identificadas empiricamente em um *corpus* de língua natural. Todavia, segundo a autora, a aquisição e o uso apropriado dessas sequências ininterruptas não é um processo tão natural, dada a importância em considerar o significado que essas expressões apresentam em determinadas disciplinas.

Stubbs (2007) apóia-se no conceito de sequência múltipla de palavras (na língua inglesa *multi-word sequence*) para referir-se aos estudos baseados na extração dos conjuntos de palavras ininterruptas recorrentes no texto por meio de programas computacionais. Stubbs (2007, p. 90) assevera que uma das possíveis definições para essa sequência ininterrupta de palavras está fundamentada nos modelos de linguagem *n-grams,* os quais fornecem, com o auxílio de instrumentos computacionais, um determinado número de palavras que podem ser ordenadas de forma alfabética ou por frequência. Para o autor, não existe um termo padrão para essa sequência de palavras, pois esses grupos de palavras podem ser nomeados diferentemente, por exemplo, em língua inglesa temos os seguintes termos: *statistical phrases*, *recurrent word-combinations, lexical bundles*, *cluster*, *chains*, *multi-word sequences*, ou mesmo *n-grams*[7].

Os modelos de linguagem *n-grams* foram desenvolvidos no âmbito da Estatística pelo matemático

[4] O PLN é uma disciplina ligada à Ciência da Computação que compartilha assuntos com a LC, mas ambas as áreas mantêm-se independentes.
[5] Pacote lexical é a tradução de *lexical bundle* para a língua portuguesa (termo em língua inglesa originariamente denominado por Biber *et al*., 1999). A tradução foi atribuída por Berber Sardinha (2003) e já está consagrada e adotada pelos pesquisadores brasileiros.
[6] No original: "[…] sequences of words that commonly go together in natural discourse".
[7] Sobre o assunto/aplicação do termo *n-grams*, ver, por exemplo, dentre outros, o artigo disponível em Lopes *et al.* (2009).

russo Andrey Markov (1856-1922) com a finalidade de reconhecer padrões estatísticos do uso da língua baseados em cadeias, conhecidas como cadeias de Markov (*Markov chains*)[8]. Manning e Schütze (1999, p. 192-193) declaram 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[9].

Em Biber *et al*. (1999), as combinações de palavras por sequências 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 as combinações sejam consideradas pacotes lexicais, elas devem recorrer frequentemente entre uma larga extensão de textos, entre cinco ou mais textos distribuídos entre registros variados, evitando tendências idiossincráticas por parte do usuário da língua. Na metodologia adotada na pesquisa com pacotes lexicais apresentada na gramática descritiva de língua inglesa desenvolvida por Biber *et al*. (1999), os autores estabelecem uma frequência de extração de pacotes com critério de corte de no mínimo dez vezes por milhão de palavras. Isso significa que o pacote deve co-ocorrer pelo menos dez vezes em cada milhão de palavras em um determinado *corpus* selecionado pelo pesquisador, a fim de que esse pacote seja representativo num dado contexto[10].

Contudo, em Biber *et al*. (2004, p. 376), esse critério de corte é relativamente arbitrário, em razão de essa escolha depender dos objetivos apontados na pesquisa e das questões levantadas pelo pesquisador e, principalmente, do tamanho dos *corpora* ou do *corpus* explorado.

Biber *et al*. (1999), em suas pesquisas com registros acadêmicos, consideram que os pacotes de três palavras co-ocorrem com maior frequência por serem um tipo estendido. Para elucidar, um tipo estendido indica que se extrairmos um pacote lexical de três palavras numa sequência de encadeamento, por exemplo, o pacote *of the quality* do SMF, conseguiremos um pacote lexical de quatro palavras, o *of the quality management*, posteriormente um pacote de cinco palavras, o *of the quality management system*, e, assim, de modo contínuo. Por esse motivo é que Biber *et al*. (1999, p. 992) afirmam que pacotes lexicais mais longos como os de cinco ou seis palavras são mais fraseológicos, em virtude de encapsularem pacotes de três palavras em sua formação, ou seja, baseados nos exemplos do SMF mencionados acima, a partir do conjunto de três palavras dá-se início à formação de sequências maiores, conforme ilustramos na Figura 1.

Com efeito, à medida que a sequência de palavras aumenta, a probabilidade de co-ocorrência de pacotes diminui e, dessa forma, o ponto de corte pode variar conforme a extensão do pacote que o pesquisador determinar e o tamanho do *corpus* disponível. A extração do tamanho do pacote fica a critério do pesquisador, sendo que tanto tamanhos menores como tamanhos maiores de pacotes contribuem para a análise. No caso deste trabalho, o fato de o *corpus* de estudo ser de pequeno porte leva-nos a optar pela extração de pacotes de três palavras, porque esse critério favorece extrair uma quantidade maior de conjunto de palavras. Por outro lado, se optássemos por pacotes de quatro ou cinco palavras teríamos uma quantidade menor para o estudo. Além do mais, um pacote lexical de duas palavras (um *bigram*) pode conter apenas fragmentos que são partes de pacotes maiores, e trabalhar com pacotes com sequências maiores (três ou mais palavras) torna-os mais informativos (Berber Sardinha, 2003).

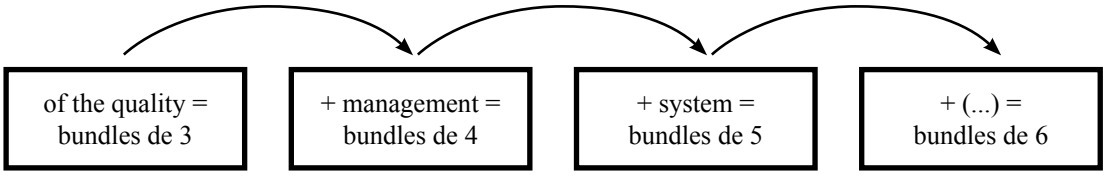


Figura 1. Sequência de encadeamento na formação de pacotes lexicais.
Figure 1. Recurrent word combinations that form lexical bundles.

[8] Conforme Manning e Schütze (1999, p. 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, os pesquisadores dizem que, 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, resultando, nos casos de *n-grams* (2, 3, 4 ou demais palavras) nas denominações do tipo: modelos *bigram, trigram, four-gram* ou mesmo *quadrigram* – esse último com prefixo em latim.

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

[10] Esse é um procedimento estatístico para ajustar as frequências 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 pacotes lexicais. Esse procedimento é denominado "normalização". Consultar Biber *et al*. (1999) e Rocha (2007).

Sobre o documento regulatório SMF

Há décadas o mundo ocidental obrigou-se a um processo acelerado de harmonização regulatória aplicado às indústrias farmacêuticas. O ponto de referência desse processo de harmonização está na criação dos blocos econômicos que resultou na formação da CEE (Comunidade Econômica Europeia). Dessa formação, estabeleceu-se o primeiro ato que é considerado referência em nossos dias, trata-se da Diretriz CEE 65/65, de 26 de janeiro de 1965[11], que estabelece os critérios básicos de harmonização entre os países membros da comunidade do setor industrial farmacêutico.

A indústria farmacêutica, considerada o setor industrial sujeito ao mais elevado grau de regulamentação e controle por parte das autoridades públicas, segundo o conjunto de regras de produção farmacêutica, tem por primazia facilitar a consulta aos trabalhos empreendidos pelas instituições comunitárias do setor. Para isso, elegeu-se em 1975, por meio do Decreto 75/320/EEC[12], um comitê europeu de especialidades farmacêuticas, denominado *Pharmaceutical Committee* (Comitê de Especialidades Farmacêuticas – CEF), que após duas décadas reuniu todos os documentos relevantes em uma única coletânea contendo cinco volumes que foram publicados no ano de 1991 e disponibilizados ao público interessado[13].

No volume 4 dessa coletânea, encontra-se um guia comunitário de *Boas Práticas de Fabricação* (BPF) ou *Good Manufacturing Practices* (GMP, em língua inglesa). Esse guia, em particular, é o principal instrumento que norteia o documento investigado nesta pesquisa, o documento SMF, uma vez que determina a inspeção farmacêutica quanto aos requisitos necessários para o cumprimento das Boas Práticas de Fabricação.

As BPF são um conjunto de normas reconhecido e regulamentado pelos órgãos de saúde pública mundial, com a finalidade de certificar a qualidade dos produtos farmacêuticos. Essas regras e normas estabelecem os procedimentos e as práticas que visam à padronização quanto aos métodos de fabricação, às condições de instalações da empresa, aos equipamentos e manutenções, aos critérios de segurança, às matérias-primas e embalagens, às condições de estocagem, e aos aspectos relacionados ao meio ambiente. Cada localidade (*Site*) pode adequar os seus procedimentos para dar cumprimento às exigências das BPF ou GMP. O objetivo principal 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 SMF é escrito na língua local em que a empresa está situada e, em seguida, traduzido para a língua inglesa por tradutores técnicos especializados e homologados na área farmacêutica. Essa prática tradutória deve-se ao esquema de certificação da qualidade de produtos farmacêuticos como objeto de comércio internacional, e tem como propósito atender às resoluções legais para o monitoramento das operações farmacêuticas no mundo. As indústrias farmacêuticas devem seguir o mesmo padrão na elaboração escrita do SMF, isto é, todas as seções/capítulos 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, ou seja, quando da instalação da fábrica e início das operações na localidade, salvo as revisões que devem ser realizadas decorrentes de desvios na garantia da qualidade. Quanto à organização estrutural do SMF, o documento deve conter nove seções que devem ser enumeradas de acordo com os seus respectivos assuntos, conforme Tabela 1.

Podemos observar que as informações técnicas contidas nas nove seções elencadas acima convergem às diretrizes envoltas no sistema da qualidade, nas operações de produção farmacêutica e nas condições de instalações da planta ou *site*. Portanto, encontramos aqui um documento regulatório que circula em todas as unidades de negócios da empresa espalhadas pelo mundo, norteado pelas políticas internacionais de saneamento de saúde pública e formulado a atender às exigências de garantia e qualidade dos medicamentos.

Aspectos metodológicos e procedimentos de análise

Para esta pesquisa utilizamos a linguagem de programação Perl (acrônimo para *Practical Extraction and Report Language*) e a ferramenta *Cygwin* (*Cygnus Solution*, 1995)[14]. A linguagem de programação Perl foi criada por Larry Wall em 1987 e tem como principais

[11] Diretiva do Conselho da Comunidade Econômica Europeia relativa à aproximação das disposições legislativas, regulamentares e administrativas respeitantes às especialidades farmacêuticas. A Diretriz 65/65 de 1965 é composta por regras que regulamentam os produtos farmacêuticos na Comunidade Europeia e estabelece que, antes que os produtos farmacêuticos sejam disponibilizados no mercado para consumo, se comprovem que os medicamentos apresentem boa qualidade e sejam seguros e eficazes para os pacientes em termos das indicações terapêuticas propostas. Também, a fabricação dos produtos farmacêuticos, a sua rotulagem e o fornecimento de informações sobre os medicamentos, tanto para o paciente como para o médico, devem estar sujeitos a controles estritos.

[12] Conselho de decisão de 20 de maio de 1975 para estabelecer um comitê farmacêutico entre os membros da comunidade europeia.

[13] Os cinco volumes estão disponíveis na internet e traduzidos em diferentes idiomas.

[14] Um emulador Unix para Windows que contém uma coleção de ferramentas desenvolvidas pela Cygnus Solution. Atualmente, a empresa americana Red Hat detém a licença da biblioteca *Cygwin*, que permite o link a *software* livres. Para detalhes do funcionamento e um breve histórico, ver artigos disponíveis em Linux (s.d.) e Cygwin Project (s.d.).

Tabela 1. As nove seções do SMF.
Table 1. The nine sections of SMF.

Número da seção	Nome da seção
1	*General Information* (Informações Gerais)
2	*Personnel* (Pessoal)
3	*Premises and Equipment* (Premissas e Equipamentos)
4	*Documentation* (Documentação)
5	*Production* (Produção)
6	*Quality Control* (Controle de Qualidade)
7	*Contract Manufacturing and Analysis* (Análise e Contrato de Fabricação)
8	*Distribution, Complaints and Product Recall* (Distribuição, Reclamação e *Recall* do Produto)
9	*Self Inspections* (Auto-inspeção)

características auxiliar o usuário na programação de pequenas tarefas ou *scripts*[15], como também facilita 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 Ciência da Computação, foi desenvolvido um aplicativo denominado *Análise Linguística*, e utilizada a linguagem SQL (acrônimo para *Structured Query Language*)[16] para gerar as consultas dos pacotes lexicais de três palavras.

O *corpus* deste trabalho é formado pela compilação de quinze exemplares do SMF que pertencem a uma única empresa. Os exemplares estão distribuídos entre países da Europa, Ásia, América do Norte, América Central e América Latina, e são datados entre o período de 2004 a 2006, sendo a coleta dos dados realizada no ano de 2007. Na Tabela 2, apresentamos uma síntese das quinze localidades às quais os documentos analisados pertencem, contendo os números de páginas e palavras de cada SMF.

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, por exemplo: o nome da empresa recebeu o pseudônimo de *PharmaCo*, e as pessoas responsáveis receberam uma numeração conforme a quantidade de funcionários envolvidos nos processos de cada localidade. Feita a coleta dos dados e a devida limpeza nos textos, foram extraídas dos documentos as figuras que não seriam necessárias ao tipo de análise linguística efetuada neste estudo, como: 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. Dessa forma, após a constituição do *corpus* de estudo, os documentos foram gravados em arquivos texto (extensão.txt) para então viabilizar o processamento dos dados. Assim, a extração dos pacotes lexicais de três palavras foi organizada nas seguintes etapas:

1ª etapa:geração de um diretório[17];
2ª etapa:realização de limpeza nos arquivos, utilizando linhas de comandos do *shell*;
3ª etapa:importação dos dados por meio do aplicativo *Análise Linguística*.

De acordo com a ordem elencada acima, partimos para a primeira etapa 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".

[15] Para a área da computação, *script* é uma série de instruções para que a máquina execute determinadas tarefas segundo a necessidade de programação. Essa linguagem de programação é baseada em linhas de código. Por exemplo, podemos utilizar um *script* para contar quantos visitantes entram num *site* diariamente, ou saber os lugares de onde acessam o conteúdo (ver Duarte, 2013).
[16] O SQL é uma linguagem de pesquisa declarativa para Banco de Dados Relacionais (base de dados relacional). É uma linguagem desenvolvida no início dos anos 70 pela IBM. Normalmente, é utilizada para consultar, adicionar, atualizar ou remover informações de um banco de dados. Consultar: Ben-Gan (2010).
[17] Estrutura hierárquica de arquivamento eletrônico.

Tabela 2. Número de páginas e palavras por localidade do documento SMF.
Table 2. Number of pages and words of each SMF.

Documento	Empresa	Localidade	Nº Palavras	Nº Páginas
SITE MASTER FILE (SMF)	PHARMACo.	*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

Após a organização e o armazenamento dos dados no diretório, utilizando o aplicativo Notepad do ambiente Windows®, desmembramos o arquivo denominado "Original.txt" entre as nove seções do documento, ou seja, para cada seção do 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 2, demonstramos um exemplo da pasta referente à seção *Contract* com o seu arquivo fonte.txt.

Encerrada a primeira etapa dos procedimentos (geração de um diretório), na segunda etapa, utilizando o *shell* (*prompt* de comando do Unix)[18] no ambiente *Cygwin*, realizamos uma limpeza no arquivo fonte.txt. Os critérios de limpeza foram os seguintes: (i) eliminação de espaços em branco; (ii) substituições de todos os números por zero para garantir o sigilo das informações contidas no SMF; (iii) substituições de nomes de pessoas e empresas por nomes fictícios; (iv) caracteres de pontuação; e (v) caracteres de controle (retorno de carro, tabulação e avanço de linhas)[19]. Essa limpeza foi realizada através do *shell script*[20] *sh* (linha de comando do *shell*) com a aplicação do arquivo script1.sh no arquivo fonte.txt, gerando o arquivo fonte1.txt. Na Figura 3, ilustramos o ambiente *Cygwin* com a linha de comando *sh* em destaque.

Ainda nessa fase dos procedimentos, por meio do aplicativo Notepad, realizamos uma segunda limpeza no texto para complementar a limpeza anterior. Esse procedimento foi necessário devido ao procedimento anterior de limpeza não ter conseguido eliminar todos os caracteres indesejáveis para a formação dos pacotes lexicais. Assim sendo, os caracteres eliminados foram: (i) os marcadores de parágrafos; e (ii) os símbolos e espaços. Seguindo com o tratamento dos dados, submetemos outro arquivo de *script* computacional, denominado "script2.sh", ao comando *sh* do *shell*, gerando os arquivos: (i) words1.txt; (ii) words2.txt e; (iii) words3.txt. Esses arquivos contêm uma lista de palavras do texto fonte1.txt, e a combinação dos três arquivos gera a formação dos pacotes lexicais de três palavras que passa a ser gravado no arquivo denominado *threegrams*.txt (conforme observamos na Figura 3).

Para explicar a formação dos pacotes lexicais de três palavras, adotamos o modelo de linguagem baseado em *n-grams* relativo à *Cadeia de Markov*, na direção da seguinte linha de raciocínio: a fórmula estatística de *n-grams* é calculada em função do tamanho da sequência de palavras, por exemplo: total de pacotes (TPac) = total de palavras do texto (TP) – (*n*-1), onde *n* representa o tamanho da sequência de palavras, e o número 1 a

[18] *Shell* pode ser definido como um interpretador de instruções e comandos do Linux. Quando um usuário ou sistema executa qualquer comando, o *shell* é responsável pela correta interpretação. Consultar Viva o Linux (s.d.).
[19] Comandos de edição de texto do Windows® conhecidos na sigla em inglês como CR (Carriage Return), LF (Line Feed) e TAB (Tabulação).
[20] Um *shell script* é um arquivo que guarda vários comandos e pode ser executado sempre que preciso.

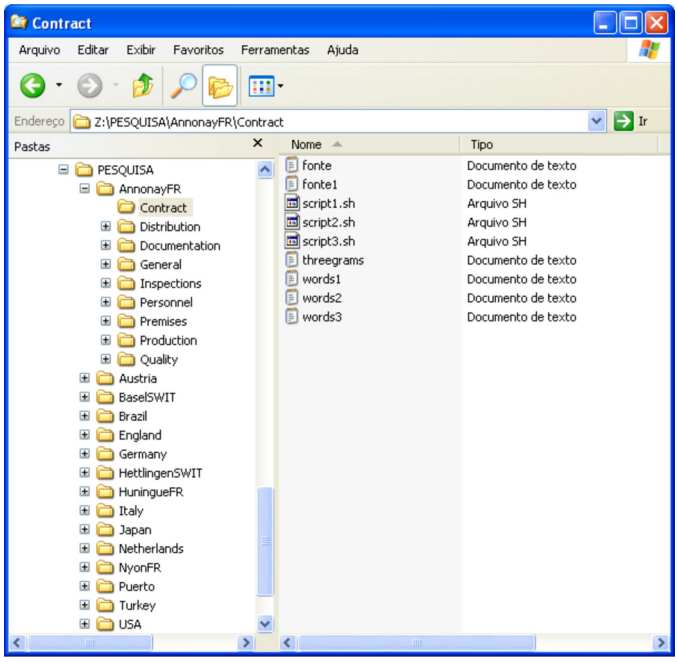


Figura 2. Criação do arquivo fonte.txt nas pastas do diretório.
Figure 2. Creation file "fonte.txt" into directory folders.

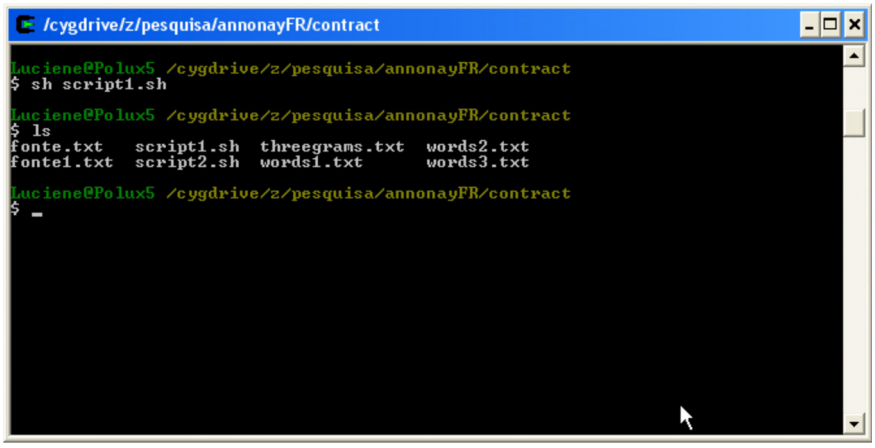


Figura 3. Ambiente *Cygwin* e linha de comando *sh*.
Figure 3. Cygwin environment and command-line sh.

constante. Isso significa dizer que, se tomarmos como exemplo um arquivo texto que contenha 100 palavras (TP), a quantidade de pacotes será 98 (TPac), resultando no número de palavras menos 2 (100 palavras - 2 = 98 pacotes). Vejamos uma possível equação para a explicação de pacotes de três palavras:

TPac = TP – (*n* –1), ou seja, 98 pacotes = 100 – (3-1)

Para efeito de ilustração, na Figura 4, apresentamos um exemplo do arquivo *threegrams*.txt referente à localidade AnnonayFR – seção *Contract* gerado através do script2.sh.

Na terceira e última etapa, utilizamos o aplicativo *Análise Linguística*. O *Análise Linguística* é um aplicativo para Windows® e foi projetado para realizar duas operações básicas: (i) importar os dados do arquivo texto (os pacotes lexicais do arquivo *threegrams*.txt) utilizando a linguagem SQL; e (ii) realizar as consultas linguísticas para a análise dos dados. Lembrando que esse aplicativo foi desenvolvido por um Desenvolvedor de *Software* em conjunto com a autora, e seu uso é exclusivo para esta análise. Ressaltamos, ainda, que importar os dados significa que os pacotes de três palavras foram processados através dos script1 e script2 e transferidos para o aplicativo.

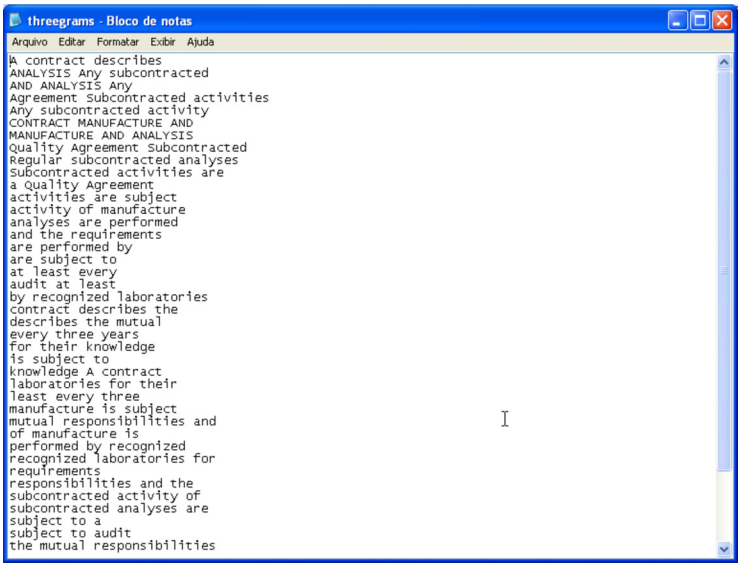


Figura 4. Arquivo *threegrams*.txt gerado através do script2.sh. – Seção *Contract* – SMF*AnnonayFR*.
Figure 4. Threegrams.txt file processed through script2.sh – Contract Section of SMFAnnonayFR.

A criação desse aplicativo é devida à relação que se estabelece com o banco de dados gerado a partir da estrutura SQL, permitindo a interação entre os dados coletados e o usuário, neste caso o pesquisador linguista. Somado a isso, esse banco de dados relacional pode ser hospedado em um *Datacenter*, que é uma central de operações em rede que oferece serviços de armazenamento de dados, ou, então, pode estar disponível localmente, isto é, a base de dados se torna disponível assim que os dados são carregados (um *Local Database*).

Para uma breve noção sobre a linguagem SQL e as vantagens de utilização para o presente estudo, podemos dizer que se trata de uma linguagem muito enxuta e especializada, tendo como uma das principais características não ser uma linguagem do tipo procedural, na qual o programador deve dizer passo a passo o que o computador deve fazer; ao contrário, é uma linguagem interativa que opera diretamente com um banco de dados a fim de produzir os resultados desejados. Conforme Ramalho (1999):

> [...] 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 (Ramalho, 1999, p. 23).

Continuando a apresentação dessa terceira etapa, apuramos a quantidade de pacotes lexicais contida em cada SMF e em cada uma das nove seções do documento para quantificar os números de pacotes em cada uma das nove seções que compõem o SMF de cada uma das quinze localidades, como também o total geral dos pacotes lexicais encontrados no *corpus* (ver Tabela 3).

Na busca por pacotes lexicais semelhantes entre as seções dos quinze exemplares do SMF, com o auxílio das ferramentas computacionais projetadas nessas três etapas dos procedimentos de análise, identificamos a co-ocorrência de pacotes lexicais em > 7 localidades, ou seja, o mesmo pacote co-ocorrendo em no mínimo 8 localidades, bem como identificamos a seção à qual o pacote lexical pertencia. 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 porque o documento pertence a um mesmo domínio e submete-se à obrigatoriedade de ser produzido em uma língua franca (a língua inglesa). Baseados nessa hipótese, percorremos três fases, a saber:

(i) a distribuição dos pacotes semelhantes que co-ocorrem em cada seção entre as quinze localidades;
(ii) a frequência da co-ocorrência desses pacotes entre as localidades, de acordo com o critério de corte estabelecido e;
(iii) a probabilidade de ocorrência dos pacotes semelhantes do SMF em um *corpus* geral de língua inglesa – o BNC – a fim de testar a sua representatividade.

Para demonstrar a primeira dessas três fases, a título de exemplo, apresentamos na Tabela 4 a distribuição e a frequência dos pacotes lexicais encontrados na seção *General Information* do SMF.

Tabela 3. Quantidade de pacotes lexicais por seção e por localidade.
Table 3. Number of lexical bundles by Section and Site.

		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**				

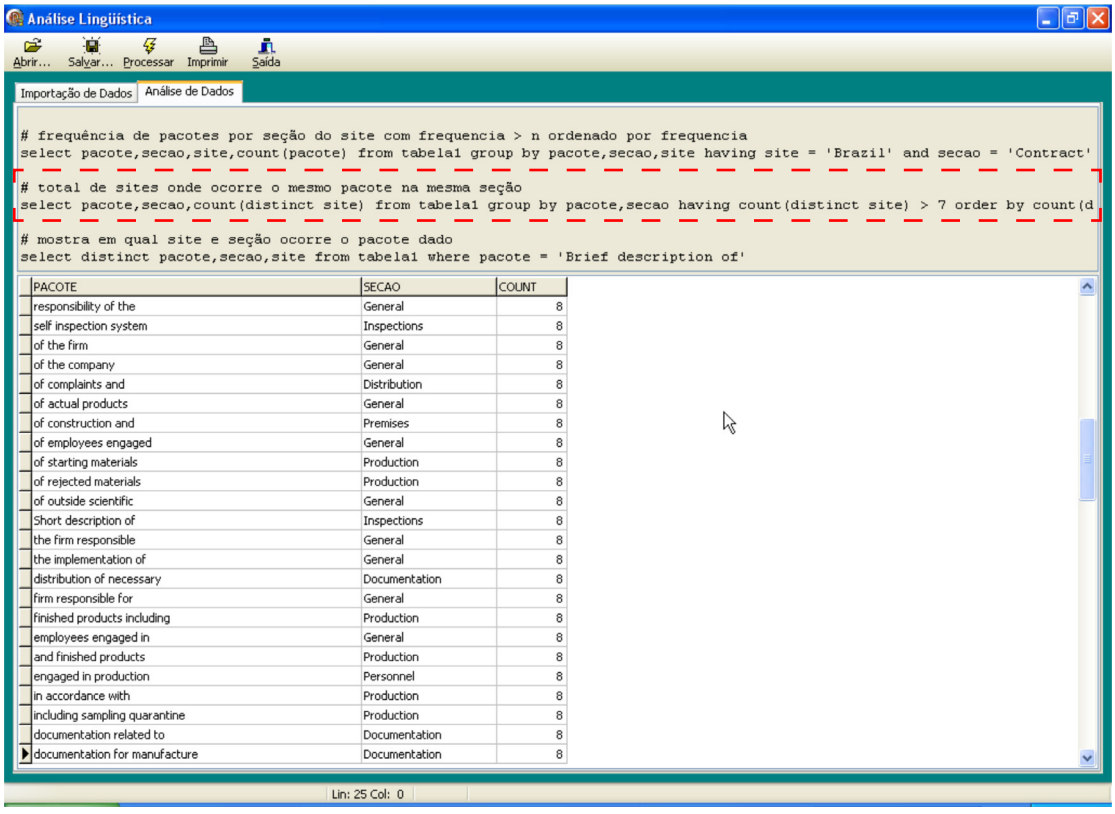
Tabela 4. Distribuição dos pacotes semelhantes da seção 1 *General Information*.
Table 4. Distribution of the lexical bundles found in General Information Section.

	Pacote lexical	**Freq.**	**Localidade**														
			AmonayFR	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	■		■	■	■		■			■	■	■			

Na segunda fase, foi necessário definir alguns critérios para filtrar os dados e estabelecer os pontos de corte na extração dos pacotes de três palavras. Desse modo, fizemos uso do aplicativo *Análise Linguística*, exclusivamente, para cruzar as informações entre os textos, identificando as co-ocorrências entre os quinze documentos, resultando na listagem dos pacotes semelhantes. Para detalhar esse procedimento, executamos uma linha de comando no banco de dados, a fim de listar a frequência dos pacotes recorrentes nas nove seções do SMF. Esclarecemos que o aplicativo *Análise Linguística* 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, por exemplo INSERT, UPDATE, DELETE, SELECT, FROM, COUNT, DISTINCT, entre outros, é possível realizar cada consulta ao Banco de Dados SQL e obter, posteriormente, o resultado dos pacotes, que serão exibidos na área designada no aplicativo. Esses comandos do SQL são denominados DML (*Data Manipulation Language*) e DQL (*Data Query Language*)[21]. 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, e em seguida a autora executou a busca pelas informações.

Na Figura 5, apresentamos como o aplicativo *Análise Linguística* respondeu à busca pela distribuição dos pacotes, sinalizando a linha de comando, a área de digitação e a área de resultados.

Para o ponto de corte, definimos > 7, por representar 55% do total de quinze, ou seja, pacotes que fossem recorrentes em no mínimo 8 localidades das 15 pesquisadas. Conforme podemos verificar na Figura 5, temos, portanto, na primeira coluna (da esquerda para a direita), o nome do pacote; na coluna do meio, a seção onde ocorre o pacote; e, na terceira coluna, o número de localidades na qual esse pacote co-ocorre. Exemplificando, os pacotes: *responsibility of the* co-ocorre na seção *General Information* em 8 localidades; o *self inspection system* co-ocorre na seção *Inspections* também em 8 localidades; e assim por diante. Dessa maneira, o mesmo procedimento foi utilizado em todas as outras seções para a captura dos pacotes lexicais semelhantes.



PACOTE	SECAO	COUNT
responsibility of the	General	8
self inspection system	Inspections	8
of the firm	General	8
of the company	General	8
of complaints and	Distribution	8
of actual products	General	8
of construction and	Premises	8
of employees engaged	General	8
of starting materials	Production	8
of rejected materials	Production	8
of outside scientific	General	8
Short description of	Inspections	8
the firm responsible	General	8
the implementation of	General	8
distribution of necessary	Documentation	8
firm responsible for	General	8
finished products including	Production	8
employees engaged in	General	8
and finished products	Production	8
engaged in production	Personnel	8
in accordance with	Production	8
including sampling quarantine	Production	8
documentation related to	Documentation	8
documentation for manufacture	Documentation	8

Figura 5. Total de localidades onde co-ocorreu o mesmo pacote na mesma seção.
Figure 5. Total of Site when co-occurred the same lexical bundles within the same Section.

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

O propósito da terceira fase foi estabelecer um *baseline* entre os pacotes semelhantes encontrados no SMF e os pacotes lexicais de três palavras extraídos do BNC, ou seja, comparar os dois *corpus* com o intuito de verificar se a probabilidade de co-ocorrência dos pacotes lexicais semelhantes do SMF é expressivo; em outras palavras, se os pacotes têm relevância no domínio específico da área de Regulamentação Farmacêutica, ou se esses são apenas um fenômeno comum de língua geral. Também, essa base permitiu-nos identificar quais pacotes, ou pacotes-chave (termo denominado pela autora), podem representar marcas características do documento SMF e não de outros textos. Esse tipo de contraste, segundo Vasilévski (2007, p. 59-60), é um recurso muito útil para pesquisas com dados empíricos, pois permite difundir a abrangência dos resultados obtidos entre *corpus* distintos. Conforme a autora discute, 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.

Nessa trajetória, procurando identificar a co-ocorrência de pacotes lexicais em cada uma das quinze localidades do SMF, concluímos que a quantidade de pacotes lexicais encontrados em cada uma das nove seções investigadas sofre variações, independente da quantidade total de pacotes que determinada localidade apresenta. Por exemplo, no caso do SMF*Turkey*, a quantidade total de pacotes é maior entre as demais localidades, mas, se comparadas as suas seções com as seções das outras localidades, o SMF*Turkey* não se apresenta sempre em primeiro lugar com maior número de pacotes. Acreditamos que as variações aqui apuradas por meio desse método quantitativo de análise, podem estar associadas aos seguintes aspectos relacionados aos negócios da empresa:

(i) ao tipo de produto fabricado na localidade, ou seja, as formas de dosagem (comprimidos, drágeas, líquido, etc.), à matéria-prima ou princípio ativo utilizada na produção, seja de uso humano ou animal, às formas de embalagem, às áreas destinadas à armazenagem e distribuição do produto, entre outras. Todos esses componentes interferem, de certa maneira, no tipo de operação e produção farmacêutica e, consequentemente, demandam uma escrita diferente.

(ii) à organização administrativa interna, visto que, dependendo de alguns aspectos como número de funcionários, contratação de terceiros, equipamentos e documentação, determinadas localidades com relação às demais do grupo podem demandar especificações e processos diferentes que interferem na escrita do documento. Esses aspectos podem ocorrer quando 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 auditorias, 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 país deve instituir normas de exigências de qualidade dos serviços prestados e do produto direcionadas à saúde do consumidor. Assim, é provável que exista uma política interna de leis sanitárias no país que demanda informações diferentes com relação aos outros países, nesse caso interferindo na escrita do SMF entre as localidades.

Portanto, os resultados da contagem e distribuição dos pacotes lexicais por localidade e por seção fornecem subsídios e auxiliam o pesquisador a evidenciar que mesmo que o documento SMF, aparentemente, apresente em sua estrutura organizacional as mesmas seções e sub-seções, elencadas na mesma ordem sequencial e numérica, o volume de informações distribuídas entre as nove seções do documento pode sofrer variações entre as localidades da empresa. É provável que, ou essas diferenças sejam decorrentes dos três aspectos mostrados acima, ou estejam apontando para os autores que produzem os textos, por conseguinte, as escolhas linguísticas feitas pelos autores. Contudo, devemos atentar ao fato de que essa é uma pesquisa linguística focada na análise do conjunto de textos que constitui o SMF. Em vista disso, as nossas conclusões indicam somente a investigação dos elementos textuais nas divisões internas do documento a partir da extração dos pacotes lexicais de três palavras. Outros aspectos a respeito das operações de fabricação e dos processos administrativos da indústria farmacêutica em questão estão além do escopo deste trabalho. Por isso, linguisticamente falando, os resultados das variações encontradas apontam que a organização textual do documento, já pré-estabelecida pelas instituições competentes, não garante que a produção escrita no corpo do documento apresente semelhanças. Assim, inferimos que as escolhas linguísticas feitas pelos diferentes autores do SMF são particulares das pessoas e das situações em que se produzem os textos.

Considerações finais

Com a metodologia construída para extrair pacotes lexicais do documento regulatório SMF, foi possível apurar poucas semelhanças entre a escrita dos exemplares coletados, respondendo a inquietação inicial da conformidade entre os elementos linguísticos de um documento

normatizado, regulamentado e escrito por diferentes autores em diferentes partes do mundo. No entanto, os pacotes semelhantes encontrados contribuem de forma a identificar termos específicos que fazem parte da escrita do documento, justificando o uso desses termos para a construção de um banco terminológico que auxilie os profissionais na escrita e/ou tradução do SMF. Outra contribuição que surge desta análise quantitativa é a de que os escritores do SMF necessitam adquirir, além do conhecimento das exigências globais para a produção escrita do documento, conhecimento das condições locais. Porquanto, apesar de os fatores globais que governam o documento sejam os de atender à 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, as informações pertinentes aos órgãos governamentais e aquelas referentes às atividades de negócio realizadas em diferentes localidades.

A criação de um aplicativo para a extração, consulta e posterior análise dos pacotes lexicais é a proposta basilar deste estudo. O desenvolvimento dessa ferramenta computacional, além de agregar valores aos avanços tecnológicos na área da Linguística de *Corpus* e Linguística Computacional, também desperta a possibilidade de se construir outras modalidades a partir desse projeto, viabilizando integrar um conjunto de instrumentos computacionais que atendam outros tipos de análises, por exemplo: as consultas integradas (via *web*), o cruzamento de uma variedade de *corpus* para identificar e distinguir elementos linguísticos de determinados gêneros e tipos textuais, entre outros. Além disso, o uso da sequência de três palavras revela aspectos fraseológicos no texto, localizando e identificando os assuntos tratados, bem como contribuindo para a extração de terminologias que perpassam os textos de especialidades. Todavia, futuros estudos podem fazer uso de sequências de quatro ou mais palavras, possibilitando abranger um espectro mais amplo nas relações entre as sentenças ou frases do texto.

Quanto à área farmacêutica, o setor carece de estudos linguísticos que possam auxiliar os profissionais na comunicação, tanto na forma escrita como na oral. Por ora, as variações encontradas neste estudo denotam algumas necessidades que a escrita do documento demanda, uma delas é o reconhecimento das diferenças na escrita regulamentada do SMF por diferentes autores de uma mesma comunidade específica. Por fim, o tipo específico de documento escolhido pode, também, fazer uso de outros instrumentos de análise além da categoria dos pacotes lexicais na exploração das diferentes escolhas linguísticas feitas pelos autores do SMF.

Referências

AIJMER, K.; ALTENBERG, B. 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London, Longman, 338 p.

BEN-GAN, I. 2010. Microsoft SQLServer 2008 – Fundamentos em T-SQL. Porto Alegre, Bookman, 416 p.

BERBER SARDINHA, T. 2004. *Lingüística de Corpus*. São Paulo, Manole, 410 p.

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. Acesso em: 01/08/2014.

BIBER, D.; CONRAD, S.; CORTES, V. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, **25**(3):371-405. http://dx.doi.org/10.1093/applin/25.3.371

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRADO, S.; FINEGAN, E. 1999. *The Longman Grammar of Spoken and Written English*. London, Longman, 1204 p.

BIBER, D.; CONRAD, S.; REPPEN, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, Cambridge University Press, 312 p. http://dx.doi.org/10.1017/CBO9780511804489

CORTES, V. 2006. Teaching bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, **17**(4):391-406. http://dx.doi.org/10.1016/j.linged.2007.02.001

CYGWIN PROJECT. [s.d.]. A brief history of the Cygwin project. Chapter 1. Cygwin Overview. Disponível em: https://www.cygwin.com/cygwin-ug-net/brief-history.html. Acesso em: 21/12/2015.

DUARTE, H. 2013. O que são scripts; entenda para o que servem. *Techtudo,* 18 dez. Disponível em: http://www.techtudo.com.br/noticias/noticia/2013/12/o-que-sao-scripts-entenda-para-o-que-servem.html. Acesso em: 13/01/2016.

GARSIDE, R.; LEECH, G.; McENERY, T. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London, Longman, 281 p.

HALLIDAY, M.A.K. 1991. Corpus studies and probabilistic grammar. *In:* K. AIJMER; B. ALTENBERG (orgs.), *English Corpus Linguistics: studies in honour of Jan Svartvik*. London, Longman, p. 30-43.

HOFLAND, K.; JOHANSSON, S. 1982. *Word Frequencies in British and American English*. Bergen/London, Norwegian Computing Centre for the Humanities/Longman, 560 p.

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

KUCERA, H.; FRANCIS, W.N. 1967. *Computational Analysis of Present Day American English*. Providence, Brown University Press, 424 p.

LEVY, S.A. 2003. *Lexical Bundles in Professional and Student Writing*. Stockton, Califórnia. Tese de Doutoramento. University of the Pacific, 172 p.

LINUX. [s.d.]. Linux Magazine Online. Disponível em: http://www.linux-magazine.com.br/images/uploads/pdf_aberto/LM_81_68_71_04_tut-%20entredoismundos.pdf. Acesso em: 21/12/2015.

LOPES, L.; FINATTO, M.J.; ZANETTE, A.; VIEIRA, R.; MARTINS, D.; RIBEIRO JR., L.C. 2009. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. *Reciis,* **3**(1). http://dx.doi.org/10.3395/reciis.v3i1.244pt

MANNING, C.D.; SCHUTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press, 680 p.

PICSCHEME. [s.d.]. The Pharmaceutical Inspection Co-operation Scheme. Disponível em: http://www.picscheme.org/publication.php?id=15. Acesso em: 21/12/2015.

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

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*. Florianópolis, Editora da UFSC, p. 194-221.

SCOTT, M.; TRIBBLE, C. 2006. *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam, John Benjamins, 203 p. http://dx.doi.org/10.1075/scl.22

STUBBS, M. 2007. An example of frequent English phraseology: distribution, structures and functions. *In:* R. FACCHINETTI (org.), *Corpus Linguistics 25 Years on*. New York/Amsterdam, Rodopi, p. 89-106. http://dx.doi.org/10.1163/9789401204347_007

SVARTVIK, J.; QUIRK, R. (orgs.). 1980. *A Corpus of English Conversation*. Lund Studies in English 56, Lund, CWK Glerup, 893 p.

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*. Florianópolis, Editora da UFSC, p. 166-173.

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*. Florianópolis, Editora da UFSC, p. 46-62.

VIVA O LINUX. [s.d.]. Home. Disponível em: http://vivaolinux.com.br. Acesso em: 21/12/2015.

Submetido: 02/08/2015
Aceito: 07/12/2015

Luciene Novais Mazza
Universidade Paulista
Av. Comendador Enzo Ferrari, Swift
13043-900, Campinas, SP, Brasil