Possibilidades de aplicação e a aplicação atual de métodos estatísticos na pesquisa de linguagens especializadas[1]

Lothar Hoffmann[2]
Tradução: Leonardo Zilio[3]
Revisão: Erica Sofia Luisa Foerthmann Schultz, Maria José Bocorny Finatto e Elisandro José Migotto, Patrícia Chittoni Ramos Reuillard[4]

1. Estatística da Linguagem

1.1. A Estatística da Linguagem é uma disciplina integrante da Lingüística. Trata principalmente, de *aspectos quantitativos do uso da linguagem* e *do sistema lingüístico*[5]; para isso, utiliza procedimentos estatísticos. Em sua forma mais simples, complementa a descrição qualitativa da linguagem através de informações sobre a freqüência de ocorrências lingüísticas, o que é útil para certas aplicações, tais como a pesquisa de informações, o ensino de línguas estrangeiras etc. Sua pretensão teórica consiste em modelar a comunicação lingüística como um processo de probabilidades. Em ambos os casos, a Estatística da Linguagem chega a parâmetros objetivos da diferenciação lingüística e a como esses parâmetros se expressam em diferentes sublinguagens, linguagens especializadas, socioletos

[1] Traduzido com a permissão do autor para publicação nos *Cadernos de Tradução* do Instituto de Letras da UFRGS, a partir do texto original *Anwendungsmöglichkeiten und bisherige Anwendung von statistischen Methoden in der Fachsprachenforschung*. In: HOFFMANN, Lothar; KÄLVERKÄMPER, Hartwig; WIEGAND, Herbert Ernst (Orgs.). *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. Walter de Gruyter, Berlin, New York: p. 241-249, 1998.

[2] O Prof. *Em.* Dr. Lothar Hoffmann, nascido em 1928, é professor emérito da Universidade de Leipzig, está atualmente aposentado. É renomado pesquisador e tradutor das linguagens especializadas, tendo sido precursor de observações da linguagem com apoio informatizado, muitas com o intuito de subsidiar a tradução. Publicou inúmeros trabalhos e pesquisas que tratam, entre vários temas, de aspectos léxico-terminológicos, estatísticos, tradutológicos e aspectos textuais da comunicação científica e técnica.

[3] Bacharel em Letras pela UFRGS, habilitação Tradutor: Alemão-Português.

[4] Professoras do Instituto de Letras, UFRGS.

[5] As grafias em itálico, em determinados pontos do texto, obedecem o destaque no original.

ou estilos (ERMOLENKO, 1970; FRUMKINA, 1971 e 1974; GOLOVIN, 1971; GUIRAUD, 1954 e 1960; HERDAN, 1964; HOFFMANN, 1975; HOFFMANN & PIOTROWSKY, 1979; MULLER, 1968; NALIMOV, 1974; NÜBOLD, 1974; PIOTROVSKIJ & BEKTAEV & PIOTROVSKAJA, 1977; TULDAVA, 1987; e outros).

A Lingüística utiliza métodos estatísticos sobretudo naquelas situações em que se trata de compreender a língua em funcionamento, ou seja, em textos, e concluir, a partir de recortes (de amostras), quais são as propriedades de um todo (de um universo estatístico). Esse procedimento é algo que, em muitos casos, é impraticável realizar de outra forma devido à imensa abrangência e à grande variedade das comunicações lingüísticas.

1.2. Os *passos fundamentais da análise estatístico-lingüística* são: (1) definição dos elementos estatísticos, por exemplo, palavra, frase, oração; (2) planejamento dasamostras, ou seja, escolha e verificação da sua representatividade a partir de um determinado *corpus*, assim como a determinação de seu tamanho e de seu número; (3) averiguação da freqüência absoluta dos elementos em cada uma e no total das amostras; (4) cálculo da freqüência relativa e, com isso, da probabilidade em relação ao universo estatístico, por exemplo, sublinguagem, linguagem especializada, estilo funcional, obra de um autor; (5) teste da confiabilidade dos valores averiguados através do cálculo do desvio padrão, do erro relativo, dos limites de confiabilidade, ou com a ajuda de outros métodos de verificação; (6) apresentação dos resultados em listas, tabelas ou gráficos, por exemplo, do tipo circular, de áreas, histograma, poligonal, curva, diagrama de dispersão; (7) interpretação e generalização dos resultados visando à formulação de regularidades estatísticas. A apresentação de uma hipótese pode preceder esses passos, de forma que eles sirvam para identificar se a hipótese é verdadeira ou falsa.

1.3. A Estatística da Linguagem tratou essencialmente de 10 *grupos de problemas*: (1) princípios gerais e metodologia; (2) fonética; (3) métrica; (4) índices e concordâncias[6]; (5) distribuição e freqüência das palavras; (6) semântica; (7) morfologia; (8) sintaxe; (9) linguagem infantil; (10) problemas filológicos como, por exemplo, problemas estilísticos (GUIRAUD, 1960, 5). Ultimamente, a Estatística da Linguagem tem se voltado à descrição integrativa de unidades cada vez mais complexas devido às necessidades da Lingüística Textual e tem levado em consideração também elementos lingüísticos da coerência, marcas da sintaxe do texto e outros elementos.

[6] N.T: *concordâncias* são listas de contextos de uma determinada unidade que se estuda ou observa em um corpus.

1.4. Assim como a estilística funcional pode ser vista, em geral, como uma das precursoras da Lingüística das Linguagens Especializadas, também a *Estatística Estilística*[7] facilitou a aplicação de métodos estatísticos na pesquisa de linguagens especializadas (HOFFMANN & PIOTROWSKI, 1979, 148 – 156; PEREBIJNIS, 1967; GOLOVIN & PEREBEJNOS, 1974; entre outros). Seus princípios residem sobre o estudo dos estilos em um *corpus* textual de representatividade qualitativa e quantitativa e, ao invés de afirmar que uma ou mais ocorrências lingüísticas seriam características de um determinado estilo funcional[8], traz a indicação de sua provável freqüência de aparecimento no *corpus* textual pesquisado.

Nas pesquisas estilísticas de primeira geração, deparamo-nos, predominantemente, com valores absolutos e números percentuais para categorias gramaticais, formas e construções isoladas. O objetivo principal das contagens e fundamento para uma interpretação inicial é a comparação das freqüências nos diferentes estilos funcionais, como, por exemplo, linguagem coloquial, literatura artística e literatura técnico-científica, e, com isso, a comprovação de determinadas correntes de estilo, como impessoalidade, exatidão, seqüencialidade lógica etc. A segunda geração se reconhece pelo fato de seus representantes deixarem-se guiar pelos conhecimentos do plano e da teoria amostrais na determinação do *corpus* textual analisado. Calculam o tamanho e o número necessário de amostras e estimam a confiabilidade dos resultados. A terceira geração, por sua vez, busca, além dos elementos antes citados, um critério para a significância das diferenças entre as freqüências dos fenômenos lingüísticos nos diferentes estilos funcionais da linguagem. Na quarta geração, interpreta-se a atualização sistemática e completa das unidades de todos os níveis da linguagem em linguagens isoladas com seus respectivos estilos funcionais (ver PEREBIJNIS, 1967). Uma quinta geração pode ser percebida pela comparação de medidas estatísticas para estilos iguais em diferentes linguagens e "universais" estatístico-estilísticos.

A Estatística Estilística fornece, assim, uma descrição quantitativa exata de estilos isolados, que, freqüentemente, são base para considerações qualitativas. Uma grande variedade e distinção entre os textos, as quais são cunhadas, especialmente na comunicação especializada, em função de diferentes objetivos e de conteúdos diferenciados, são, contudo, minimizadas, por sua concentração, em um dado estilo funcional. Traços específicos dos gêneros[9] e variedades textuais

[7] N.T.: No original, *Stilstatistik*; expressão que equivale a um tipo de estatística lexical, pouco conhecida no Brasil, de caráter textual e gramatical, voltada para obtenção de dados sobre o estilo utilizado em determinado gênero ou grupo textual.

[8] N.T.: O termo *estilo funcional* identifica, principalmente, a estilística de perspectiva lingüística desenvolvida por lingüistas russos nos anos 60. Desse tipo de estudos de estilística, temos notícia, no Brasil, por algumas obras de Mattoso Câmara Júnior.

[9] N.T.: No texto original *Textsorten*. Em meio a uma extensa discussão teórica sobre gêneros ou tipologias textuais, de modo a obter uma melhor compreensão por parte do leitor brasileiro iniciante nos Estudos do Texto e da Linguagem, adotamos aqui o termo equivalen-

permanecem não sendo levados em conta, assim como também é ignorada a relação entre temática concreta e estilo. Além disso, as pesquisas são prejudicadas pela aplicação de categorias pré-estabelecidas, de influência tendenciosa e resultados determinados, tais como, por exemplo, exatidão, univocidade e seqüencialidade.

1.5. A *pesquisa estatística de sublinguagens* (ver Art. 15[10]) alcançou maiores sucessos onde a Estatística Lexical não encontrou um ponto seguro de enfoque: no nível do léxico. Seus resultados concretos são os dicionários de freqüência (ver Art. 195[11]) e as listas de freqüência de palavras das sublinguagens da ciência e da técnica, como, por exemplo, Medicina, Física, Química, Matemática, Eletrônica, Automação, Arquitetura e Engenharia civil, Processamento de petróleo e de gás natural, Ciências do solo, construção de máquinas agrícolas, Enologia, Ciências bélicas, Pedagogia, Política externa (HOFFMANN & PIOTROWSKI, 1979, 156 – 162). Esses produtos formam, por um lado, os fundamentos para o atendimento das necessidades práticas, como, por exemplo, na escolha de material para aulas de línguas estrangeiras ou na montagem de *tesauros*[12] para o processamento de informações. Por outro lado, oferecem uma motivação para responder questões teóricas. Pertencem à pesquisa estatística de sublinguagens, entre outros elementos, a comparação dos elementos lexicais utilizados em duas ou mais sublinguagens, observações sobre classes gramaticais, sobre formações de palavras, relações de empréstimo, campos semânticos, sobre potenciais de coincidência, e também sobre características estilísticas que permitem uma descrição qualitativa junto à descrição quantitativa.

O grande valor formador do léxico mais freqüente em um texto promete uma grande possibilidade de generalização dos resultados de observação em relação ao todo da sublinguagem. Também os parâmetros estatísticos do léxico poderiam ser comparados, como a representatividade na cobertura textual ou o número de palavras diferentes em amostras de duas ou mais sublinguagens. De forma parecida, as sublinguagens podem ser caracterizadas por meio de descrições sobre categorias (gênero, caso, número; pessoa, tempo, modo; nominalidade, verbalidade; predicabilidade, atributibilidade, causalidade, condicionalidade etc.), que representam as características da palavra, da ligação de palavras (em sintagmas) ou da oração.

te *gêneros textuais*. Na terminologia própria do autor, que justamente discute a diferença entre *tipus* e gênero, seria conceitualmente mais adequado o equivalente a *tipos textuais*.

[10] N.T.: Refere ao artigo número 15 da publicação original do artigo, *An International Handbook of Special-Language and Terminology Reserach*, volume I, de 1998. Edição de Walter de Gruyter em Berlim e Nova Iorque.

[11] N.T.: Idem à nota anterior.

[12] N.T.: *Tesauros* são repertórios de expressões que servem para a indexação de documentos e publicações por parte de bibliotecários e documentalistas.

Na ligação de palavras (em sintagmas), a análise estatística inclui o número de seus constituintes, suas classes de palavras, relações de dependência entre constituintes, entre outros itens. Além disso, a partir do ponto de vista sintático, a extensão da oração, o tipo de oração, a seqüência dos elementos oracionais e outras características da construção são fáceis de quantificar pela comparação entre sublinguagens. Do ponto de vista lingüístico-textual, é produtivo averiguar a freqüência de ocorrência de certas macroestruturas com seus contextos e a ocorrência de elementos para construção de coerência semântica e sintática, assim como também avaliar integrativa e estatisticamente os elementos lingüísticos enumerados anteriormente até o nível oracional.

2. Métodos

2.1. Como é praticamente impossível compreender a totalidade da comunicação especializada, mesmo que seja só para uma língua e uma área de conhecimento, a estatística de linguagens especializadas tem de se basear nas amostras representativas possíveis, ou seja, em textos especializados escritos ou orais. Por isso, cada análise estatístico-lingüística começa com a *escolha e preparação de um corpus material apropriado*. A primeira condição para isso é que se proporcione uma visão geral da literatura especializada da área de conhecimento em foco, que reproduza tanto o conhecimento básico quanto a situação atual e as tendências mais marcantes, mas que também reproduza a composição interna da área nas proporções certas. O *corpus* pode ser mais limitado nos casos em que há um estabelecimento de objetivo específico, no sentido da Lingüística Aplicada, como, por exemplo, na pesquisa para um glossário de aprendizes para aulas profissionalizantes de língua estrangeira ou para a montagem de um *tesauro* para informação e documentação empresarial interna.

Os trabalhos preparatórios para um dicionário trilíngüe de freqüência de Medicina podem servir como exemplo da preparação de um *corpus* textual apropriado (HOFFMANN, 1986): o número de amostras de mesmo tamanho para a Medicina Interna, 47, para a Cirurgia, 41, para a Ginecologia, 16, para a Fisiologia, 17, para a Química Fisiológica, 12, para a Patologia, 8, para a Neurologia, 8, para a Pediatria, 5, para a Saúde Pública, 6, para a História da Medicina, 1, para a Microbiologia, 2, para a Farmacologia, 1, para a Cirurgia Pediátrica, 3, para a Ortopedia, 3, para a Radiologia, 6 (HOFFMANN, 1975, 120). Junto às áreas de especialidade citadas também se prestou atenção ao tipo das publicações, ou seja, aos gêneros textuais.

Os livros didáticos de faculdades ou livros especializados de caráter geral são especialmente úteis para a pesquisa de um glossário técnico-científico básico. Essas obras oferecem mais facilmente a garantia de um registro sistemático, proporcional e completo do material e dos elementos lingüísticos necessários

para sua representação. Eles também são menos marcados por usos lingüísticos isolados que alguns outros gêneros textuais. Outro conjunto de materiais mais abrangente deveria ser extraído de revistas atualizadas, que também não têm um caráter muito especial. Por outro lado, trabalhos de consulta, páginas de apresentações, relatórios de pesquisa, manuais de uso e gêneros textuais dessa natureza são uma boa base de partida para a observação de especificidades das linguagens especializadas no nível da oração e do texto.

2.2. *O tamanho e o número de amostras* dependem principalmente do objeto de estudo, ou seja, do tipo e do conteúdo dos elementos estatísticos. Antes de qualquer cálculo de detalhes, existem os seguintes conhecimentos básicos: quanto maior é a amostra, mais certo é o resultado e menor é o erro relativo e o intervalo de confiança das freqüências encontradas. Quanto maior a freqüência relativa de um fenômeno lingüístico na amostra, tanto mais próximo está o seu valor da probabilidade de sua ocorrência no universo estatístico (linguagem, sublinguagem, linguagem especializada). Quanto mais freqüente a presença de uma unidade lexical e quanto menor o número de unidades diferentes, tanto menor pode ser o tamanho de amostra.

Isso significa que pesquisas no nível de *fonemas* ou *grafemas* podem ser realizadas com amostras de tamanho pequeno. Dessa forma, por exemplo, cinco estilos funcionais são suficientemente compreendidos em um total de 300.000 fonemas (PEREBIJNIS, 1967, 44). Recortes de textos com tamanho total de 30.000 grafemas são suficientes para verificar a especificidade de uma linguagem especializada.

Para a pesquisa das ocorrências mais importantes da *morfologia* e *sintaxe* em diferentes estilos de linguagens com sistema complexo de flexões, são recomendadas de 10 a 20 amostras para cada 500 palavras lexicais[13] (GOLOVIN, 1971, 58). Pesquisas de morfologia e de algumas categorias gramaticais fundamentais, como classe gramatical, tipo de declinação, classe verbal etc., em linguagens especializada,s são amplamente representativas em uma amostra com 20.000 palavras. Na sintaxe, é preciso fazer a distinção entre o registro de elementos sintáticos isolados e de sintagmas complexos inteiros: quanto mais elementos um sintagma complexo contém e quanto mais facilmente eles puderem trocar suas posições, maior tem de ser a amostra.

[13] N.T.: No original, *Autosemanticum*; o termo corresponde à característica de uma palavra que carrega em si um significado lexical relativamente independente, mesmo que não se encontre combinada com outras palavras. Como *autosemânticos* são entendidos os substantivos, verbos, adjetivos e, em parte, advérbios. (Fonte: cf. http://www.uni-leipzig.de/~fsrger/materialien/Texte/Lexikologie.pdf em 05/10/2005). Preposições, nesse sentido, não são autosemânticas, pois seu significado é muito mais dependente do contexto de uso. Ecoa aqui a oposição entre palavras lexicais e as palavras gramaticais.

Na pesquisa de linguagens especializadas, amostras com 2.000 orações têm levado até agora a bons resultados. As interpretações sobre o tamanho de amostra necessário nas pesquisas *lexicais* divergem amplamente. As amostras variam, para a linguagem comum, entre 1 milhão (GUIRAUD, 1960, 96) e 400.000 (FRUMKINA, 1963, 74 – 77) se for almejado o estabelecimento somente das 1.000 unidades lexicais mais freqüentes. Contagens de freqüência de palavras no nível das sublinguagens chegam a um texto corrido de 200.000 palavras (TULDAVA, 1987, 56). Mas já foram obtidos resultados bem úteis também com amostras de $N = 35.000$ (HOFFMANN, 1975, 25 – 42).

Ao se estabelecer o tamanho de amostra para pesquisas lingüísticas de textos de linguagem especializada, devem ser levadas em conta até as especificidades dos diferentes gêneros textuais. Os gêneros textuais chamados "pequenos", como crítica, relatório (resumo), entrada de um dicionário, são, em geral, compreendidos em sua totalidade, quando se trata da freqüência das marcas de estruturação e dos elementos da coerência ou da dominância de outras características textuais (HOFFMANN, 1987 b, 54 – 55). Para se encontrar diferenças significantes, são comparados entre si, no primeiro enfoque, de 10 a 20 exemplares textuais para cada gênero textual. No caso de gêneros textuais maiores, são utilizadas, comumente, amostras de mesmo comprimento retiradas do início, do meio e do fim.

Após a determinação do tamanho total da amostra (N), são importantes ainda os tamanhos das subamostras (n) e sua distribuição no *corpus* textual. Isso serve, principalmente, para recenseamentos de vocabulário. Se não se quer perder nenhuma parte essencial do vocabulário especializado a ser analisado, então se recomenda distribuir as subamostras pelo *corpus* com intervalos proporcionais, que não são nem tão pequenos, nem tão grandes. De outra forma, a classificação precedente por área e tipo de publicação não terá efeito. Valores bons são $200 \leq n \leq 500$ para as subamostras e $10 \leq S \leq 50$ para o espaço amostral.

2.3. O primeiro resultado da contagem das unidades lingüísticas é a *freqüência absoluta*. Ela indica quantas vezes cada ocorrência se encontra no texto analisado. Essa freqüência absoluta possui somente um valor diminuto para outras pesquisas, para a utilização prática dos resultados ou mesmo para afirmações generalizantes, pois ela depende diretamente do tamanho da amostra. A freqüência absoluta serve somente como valor de partida, por exemplo, para o cálculo da freqüência relativa.

2.4. A *freqüência relativa* é um valor percentual, que expressa a parcela ocupada pela unidade lingüística em relação ao universo de palavras do texto. A freqüência relativa resulta da divisão da freqüência absoluta pelo comprimento da amostra, por exemplo, para uma palavra com a freqüência 186 em uma amostra de $N = 50.000$ temos

$\frac{186}{50.000}$ ou 186 : 50.000 = 0,00372.

Dito em outras palavras: a freqüência relativa de uma forma lexical é a relação entre o número de suas ocorrências reais e o número de suas ocorrências (teoricamente) possíveis. Se a amostra for suficiente, ou seja, representativa para uma linguagem especializada, então a freqüência relativa pode ser equiparada à probabilidade do fenômeno lingüístico. Ela conta, então, para afirmações sobre a estrutura estatística da respectiva sublinguagem ou sobre a importância dos diferentes elementos para a constituição do texto.

2.5. Um passo especialmente importante na análise estatístico-lingüística é a *testagem da confiabilidade dos valores averiguados*. Existem vários métodos disponíveis para tal. Na estatística lexical e na estatística de linguagens especializadas, são calculados principalmente o desvio padrão, o erro relativo e os limites de confiabilidade. Para a comparação, utiliza-se também o teste chi-quadrado (χ^2).

O desvio padrão (desvio quadrático médio) é uma medida da variabilidade da freqüência averiguada de uma ocorrência lingüística nas amostras. Seu cálculo se dá através da seguinte fórmula:

$$\sigma = \sqrt{\frac{SQD}{n-1}}$$

(σ = desvio padrão; SQD = soma dos quadrados do desvio; n = número de amostras)

O *erro relativo* é usado principalmente para certas unidades lexicais em dicionários de freqüência de forma a determinar a confiabilidade desses dicionários. Isso é representado através da seguinte fórmula:

$$|Fr - p| \leq Z\rho \sqrt{\frac{p(1-p)}{N}}$$

(Fr = freqüência relativa; p = probabilidade; $Z\rho$ = coeficiente para o nível de confiança pré-estabelecido ρ; N = tamanho da amostra)

Em trabalhos estatístico-lingüísticos, são usadas variantes simplificadas dessa fórmula que, através da simplificação, resultam no fato de que no p minúsculo a diferença é 1 – p ~ 1. Uma variante corrente para averiguar o erro relativo é:

$$\delta = \frac{Z\rho}{\sqrt{NFr}} \text{ ou } \delta = \frac{Z\rho}{\sqrt{Fa}}$$

(δ = erro relativo; $Z\rho$ = coeficiente para o nível de confiança pré-estabelecido ρ; N = tamanho da amostra; Fr = freqüência relativa; Fa = freqüência absoluta)

(ALEKSEEV, 1975, 46)

O cálculo do *intervalo de confiança* é uma variante refinada do cálculo do erro relativo com a qual o limite inferior e o superior (p_1 e p_2) das variações em torno da freqüência média são averiguados. São encontradas para tal cálculo várias formas, como, por exemplo:

$$p_1 = \frac{FrN + \frac{1}{2}Z\rho^2 - Z\rho\sqrt{Fr(1-Fr)N + \frac{1}{4}Z\rho^2}}{N + Z\rho^2}$$

$$p_2 = \frac{FrN + \frac{1}{2}Z\rho^2 + Z\rho\sqrt{Fr(1-Fr)N + \frac{1}{4}Z\rho^2}}{N + Z\rho^2}$$

(ALEKSEEV, 1975, 47; HOFFMANN, 1975, 29)

Com a ajuda do teste de chi-quadrado (χ^2) verifica-se se as diferenças de freqüência, com as quais ocorrências lingüísticas aparecem em diferentes amostras, são significantes, ou se as amostras pertencem a um mesmo universo estatístico (estilo funcional, sublinguagem, linguagem especializada, gênero textual etc.). Na maioria das vezes, trata-se da testagem (verificação ou falsificação) de uma hipótese inicial (hipótese nula), por exemplo, a expectativa de que as classes gramaticais dos constituintes de um texto tenham aproximadamente a mesma função. A constante de teste χ^2 representa a soma dos quadrados da diferença entre as freqüências averiguadas e as esperadas em relação às freqüências esperadas para um determinado número de variáveis.

$$\chi^2 = \sum_{i=1}^{k} \frac{(F_{ei} - F_{ai})^2}{F_{ai}}$$

(k = número de variáveis; i = variável; F_{ei} = freqüência esperada das variáveis; F_{ai} = freqüência averiguada das variáveis) (HOFFMANN & PIOTROWSKI, 1979, 107-108). Ao se comparar amostras, a freqüência esperada F_{ai} é geralmente substituída pela freqüência média χ:

$$\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\bar{x}}$$ (GOLOVIN, 1971, 28-29)

(Exemplos se encontram em HOFFMANN & PIOTROWSKI, 1979, 105-126)

2.6 Na apresentação de resultados, a estatística de linguagens especializadas utiliza-se de diferentes listas, tabelas e gráficos (HOFFMANN & PIOTROWSKI, 1979, 126-148). Por meio de gráficos circulares e de gráficos de colunas, são

reproduzidos e comparados, principalmente, partes com valores percentuais. Para representar graficamente características quantitativas, como comprimentos de palavra e oração, são mais apropriados o histograma e o gráfico poligonal. Curvas com progressão mais ou menos típica vão além dessa simples agregação de freqüências e alcançam características qualitativas e quantitativas. Elas permitem perceber as relações entre as características e suas freqüências, e a própria freqüência de ocorrências lingüísticas pode se tornar uma característica, sendo marcada por outros valores.

Interdependências são mostradas, por exemplo, entre as freqüências de unidades lexicais e suas posições num dicionário de freqüência, entre a freqüência e a probabilidade no texto, entre a freqüência de um lexema e o número de seus sememas, entre freqüência e erro relativo, entre as posições em um dicionário de freqüência e o número cumulativo de unidades lexicais, entre as posições e a cobertura textual, entre a freqüência e a potencialidade de ligação, entre a freqüência e o grau de especialização do vocabulário especializado, entre o comprimento do texto e o tamanho do léxico etc. (HOFFMANN & PIOTROWSKI, 1979, 141-146).

2.7 *A interpretação e generalização* dos resultados de pesquisas estatísticas de linguagens especializadas ocorrem guiadas principalmente pelo ponto de vista de sua aplicação (veja a seção 4), mas podem também servir para embasar concepções teórico-lingüísticas. Nesse caso, interpretação e generalização têm como objeto, na maioria das situações, não o sistema lingüístico, mas sim textos de todos os tipos, e se inserem no âmbito da análise lingüística relacionada ao uso e orientada para a dimensão comunicativo-pragmática (FRUMKINA, 971; 1974; TULDAVA, 1987).

3. Resultados

3.1. O resultado mais importante da estatística de linguagens especializadas são os *dicionários de freqüência* das sublinguagens da ciência e da técnica (HOFFMANN, 1975, 25-42; HOFFMANN & PIOTROWSKI, 1979, 185-189; TULDAVA, 1987, 54-65). O dicionário de freqüência é um arquivo de conhecimentos, um livro de consultas, no qual as unidades lexicais são elencadas com sua freqüência de ocorrência, em que a freqüência pode se tornar um critério de ordenamento, de forma que surja uma lista categorial que começa com a palavra de maior ocorrência, mais freqüente, terminando com a de menor freqüência.

Os dicionários de freqüência são classificados conforme as seguintes características de forma e de conteúdo:
(1) De acordo com o ordenamento do material lexical. As unidades lexicais podem ser dispostas alfabeticamente ou de acordo com sua freqüência. Muitos dicionários de freqüência contêm tanto uma lista de freqüência quanto um índice

em ordem alfabética com informações sobre a freqüência.
(2) De acordo com a dimensão do material lexical. Existem dicionários de freqüência completos, nos quais estão compreendidas todas as unidades lexicais do *corpus* textual analisado, e dicionários de freqüência parciais, que só contêm apenas uma parte das palavras que o perfazem, normalmente as mais freqüentes.
(3) De acordo com a dimensão do *corpus* analisado. Distinguem-se dicionários de freqüência em grandes, médios e pequenos.
(4) De acordo com a mídia, o gênero, a temática ou o autor do respectivo *corpus* textual. Dicionários de freqüência são elaborados para a fala e para a escrita, para a prosa, poesia e drama, para temas gerais e especiais, assim como para diferentes publicações ou para a obra de determinados autores, mas também para sublinguagens inteiras e, conseqüentemente, para linguagens especializadas.
(5) De acordo com o tipo de unidades lexicais. Dicionários de freqüência podem registrar radicais de palavras, lexemas, formas de palavras ou grupos de palavras.
(6) De acordo com o tipo de freqüência indicada. São indicadas a freqüência absoluta e/ou a freqüência relativa. Uma terceira medida é o número de fontes, ou seja, de amostras, nas quais a palavra ocorreu (disposições da palavra no *corpus*). As freqüências relativas podem ser somadas até se chegar a uma freqüência cumulativa.
(7) De acordo com método de pesquisa aplicado. As unidades lexicais surgem ou de um *corpus* completo ou de uma seleção de amostras. O segundo caso, de amostras, é o mais freqüente, já que a maioria dos *corpora* não são passíveis de serem completamente abrangidos (ALEKSEEV, 1968, 61-62).

O número de unidades lexicais em dicionários de freqüência fica entre 40.000, para toda a linguagem, e de 1.000 a 1.200, para cada linguagem especializada. A estatística da linguagem garante uma cobertura textual de 60% com as 100 palavras mais freqüentes, 86% com as 1.000 mais freqüentes e 97,5 com as 4.000 mais freqüentes (GUIRAUD, 1960, 93-94).

Se for utilizada a classificação anterior, tratar-se-á, na maioria dos dicionários de freqüência, de um tipo combinado de um pequeno dicionário especial, que não dá importância à completude, que deve seu material ao processamento das amostras, e que contém indicações de categoria e de freqüência relativa ao lado das entradas em seu formato básico (por exemplo: HOFFMANN, 1970 *et seq.*).

3.2. Vários índices de freqüência para ocorrências gramaticais resultaram de pesquisas estatísticas de linguagens especializadas (HOFFMANN, 1987a, 96-124, 183-230). Foram compreendidas principalmente categorias classificatórias e suas características morfológicas, como, por exemplo, categoria gramatical; gênero, caso e número dos substantivos; graus dos adjetivos; pessoa, tempo, modo e gênero dos verbos; flexões; sintagmas e frases; tipos de orações e tipos dos respectivos constituintes. Enquanto nos dicionários de freqüência ficam claras as dife-

renças tanto entre as linguagens especializadas e outras sublinguagens quanto diferenças entre as próprias linguagens especializadas, o caráter específico da linguagem especializada se expressa no campo da gramática somente na freqüência ou escassez pontual de algumas ocorrências. No âmbito da linguagem especializada, tem-se uma diferenciação em relação aos gêneros textuais. Nessa diferenciação, certas peculiaridades gramaticais aparecem em certos gêneros textuais, abrangendo diferentes linguagens especializadas. Essa evidência é útil, por exemplo, para a condensação sintática automatizada que gera resumos de textos, para a informação sobre sintaxe de frases em entradas de dicionários, para os parênteses com notas de uso de palavras em livros didáticos, entre outros[14].

3.3. No campo da formação de palavras, e principalmente da *formação de termos*, foi estudada, com precisão, não a freqüência no texto, mas, sim, a freqüência no sistema, pois ela informa sobre a produtividade de tipos e meios de formação de palavras. Quanto ao que oferece a *derivação*, temos, na parte superior de uma listagem de freqüências a partir da linguagem especializada da Medicina os 11 seguintes sufixos:
-y (dispepsy), ***-ia*** (mammalgia), ***-sis*** (hemoptysis), ***-ion/-tion/-ation*** (infection, fixation), ***-a/-oma*** (malva, spheroma), ***-er*** (sterilizer), ***-ity*** (nervosity), ***-ness*** (acuteness), ***-ism*** (atropism), ***-ing*** (scalding), ***-itis*** (bronchitis).

Juntos, esses sufixos correspondem a 68,7% de todos os substantivos sufixados em um dicionário especializado da área. Outros sufixos, como ***-ment***, ***-ance/-ence***, ***-ure***, ***-ist***, ***-our***, ***-hood***, ***-age***, ocorrem somente em menos de 2% dos derivados.

Um *ranking* equivalente da linguagem especializada da Matemática mostra um quadro bastante diferente. Encontramos os seguintes 11 sufixos mais freqüentes:
-ion/-tion/-ation (addition, modification); ***-ity*** (probability; ***-ness*** (unrelatedness); ***-ance/-ence*** (expectence, ocorrence); ***-y*** (symetry); ***-ing*** (linking); ***-er*** (modifier); ***-ment*** (enlargement); ***-or*** (denominator); ***-ant/-ent*** (eliminant, component); ***ancy/-ency*** (discrepancy, frequency).

Juntas, tais terminações formam 69,6% de todos os substantivos sufixados em um dicionário especializado desse domínio. Outros sufixos, como, por exemplo, ***-ure***, ***-ism***, ***-age***, ***-osis***, ***-ship***, aparecem em menos de 2% dos derivados.

Pesquisas estatísticas de linguagens especializadas sobre a formação de palavras permitem até mesmo afirmações exatas sobre a utilização de determinados sufixos em um campo de comunicação ainda em formação, assim como também permitem comparações entre vários campos de comunicação próximos, os quais serão delimitados entre si por fatores como esses.

[14] N.T.: Fizemos aqui uma tradução livre, com significativa adaptação, a partir do segmento original. Isso deveu-se a uma determinada pressuposição de conhecimento de teoria sintática feita pelo autor. O trecho original é: *Das gilt z.B. für die syntatktische Kompression bei Referaten (Abstracts), den Typ III der aktuellen Satzgliederung bei Lexikonartikelns, die Parenthese in Lehrbüchern u.ä.m.*

Algo semelhante ocorreu na análise estatística de tipos de estruturas de *termos compostos por várias unidades lexicais*[15] (*termos formados por grupos de palavras*). Assim, na linguagem especializada russa da Construção Civil, esse tipo de termo corresponde a 8 configurações mais freqüentes em ralação a 85,2% de todas as configurações possíveis: adjetivo + substantivo (63,9%); substantivo + substantivo no genitivo (8,9%); adjetivo + adjetivo + substantivo (5,2%); substantivo + adjetivo + substantivo no genitivo (2,5%); substantivo + preposição + adjetivo + substantivo (1,8%); adjetivo + substantivo + substantivo no genitivo (1,3%); substantivo + preposição + substantivo (0,9%); substantivo + substantivo + substantivo no genitivo (0,7%).

Em outras terminologias especializadas, são maioria mais ou menos o mesmo tipo de estruturas, de forma que ficam evidentes sobretudo as características em comum entre as linguagens especializadas e as diferenças em relação às outras sublinguagens.

4. Aplicações

4.1. Através das pesquisas estatístico-lingüísticas, podem ser averiguados os *vocabulários básicos* de diferentes linguagens e de suas sublinguagens, e, por conseqüência, de linguagens especializadas, de forma mais exata que através de outros métodos.

O vocabulário básico tem sido definido de forma muito variada segundo o ponto de vista lingüístico: como léxico elementar relativamente estável desde a gênese das linguagens; como um núcleo para a formação de raízes lexicais produtivas; como base estrutural do léxico; como designações comumente usuais de coisas, ocorrências e ações de vital importância; como conjunto de palavras nacionais etc.

O que se vê de comum a todas essas definições é a concepção de um núcleo lexical produtivo, relativamente estável e amplamente propagado na comunidade lingüística. Centro, esse, no qual estão incluídos vocabulários periféricos, menos estáveis, limitados a um âmbito de ação e secundariamente derivados, como, por exemplo, os vocabulários especializados.

Todas as tentativas de reunir o vocabulário básico de uma língua em um

[15] N.T.: Denominados atualmente sintagmas terminológicos ou unidades poliléxicas.

índice de palavras mostram quão difícil é a delimitação em relação aos vocabulários periféricos. Também a dimensão desse vocabulário, que é designado como vocabulário básico por diferentes autores, oscila visivelmente (entre 200 e 20.000 unidades). Um critério confiável para o registro de unidades lexicais no vocabulário básico é a sua freqüência, pois a estatística da linguagem provou que as palavras mais freqüentes sobrevivem mais tempo às mudanças no decorrer do desenvolvimento histórico, que são constantemente utilizadas por toda a comunidade lingüística e são produtivas na formação de palavras. Junto à freqüência, também tem sido utilizado o grau de propagação (*alcance*) como critério de registro ou de exclusão em diferentes situações, âmbitos e textos.

4.2. Dicionários de freqüência também são uma importante ajuda para a elaboração de *tesauros* para a Informática. Esses *tesauros* podem contribuir, após a produção de uma ontologia, para a escolha dos descritores de uma área de conhecimento, quando aparecem sinônimos terminológicos em textos ou em dicionários especializados. Em *dicionários* tradicionais bi- ou trilíngües, a seqüência dos equivalentes de sentido após a entrada deveria ser determinada pela sua freqüência de uso, de forma a encurtar o processo de busca. Em dicionários pequenos, a freqüência pode auxiliar a decidir sobre o registro da entrada e sobre o número de equivalentes correspondentes que serão apresentados para o consulente.

4.3. Nas representações normativas da *gramática* (morfologia e sintaxe), que, compreensivelmente, têm partido em primeiro lugar da sistemática do objeto, faltam, infelizmente, até agora, indicações sobre a freqüência de ocorrência das diferentes categorias gramaticais e sua representação formal. Neste caso, oferecer-se-ia uma boa oportunidade para a união dos aspectos do sistema e do uso. A estatística de linguagens especializadas é capacitada para isso no que diz respeito aos âmbitos especializados e gêneros textuais pesquisados por ela. Seu principal interesse consiste, porém, em detalhar a representação da gramática no sentido da lingüística aplicada em relação às determinadas ocorrências relevantes para a comunicação especializada e, por conseqüência, colocar essas ocorrências em destaque.

4.4. Os resultados da estatística da linguagem são também amplamente considerados no *ensino de línguas estrangeiras*, principalmente no ensino de *linguagens especializadas*. Nesse âmbito, trata-se, sobretudo, da escolha e da delimitação do material na forma de *minima*. Um *minimum* é a quantidade de unidades lingüísticas necessárias para solucionar certas tarefas de comunicação e, por isso, constitui o cerne do ensino de línguas estrangeiras. Em outras palavras: um *minimum* deve conter as ocorrências lexicais e gramaticais mais úteis, com ajuda das quais o aluno pode aprender em menor tempo a quantidade máxima possível de conhecimentos, práticas e habilidades. *Minima* para o ensino de linguagens

especializadas se diferenciam substancialmente de *minima* para outras sublinguagens ou para a comunicação em geral devido à sua estrutura estatística específica.

O *Minimum lexical* passa por uma rigorosa seleção, com seus vários vocabulários especializados a partir do vocabulário total de centenas de milhares de palavras. A diísticas ou de certos gêneros textuais.

Bibliografia

ALEKSEEV, P. M. Èastotnye slovari i priemy ich sostavlenija. In: *Statistika reèi*. Leningrado: 1968, p. 61-63.

ALEKSEEV, P. M. *Statistièeskaja leksikografija*. Leningrado: 1975.

ERMOLENKO, G. V. *Lingvistièeskaja statistika. Kratkij oèerk i bibliografièeskij ukazatel'*. Alma-Ata: 1970.

FRUMKINA, R. M. Nekotorye praktièeskie rekomendacii po sostavleniju èastotnych slovarej. In: *Russkij jazyk v nacional'noj škole*, 5/1963. 1963, p. 74-77.

FRUMKINA, R. M. *Verojatnost' elementov teksta i rèevoe povedenie*. Moscou: 1971.

FRUMKINA, R. M. *Prognoz v reèevoj dejatel'nosti*. Moscou: 1974.

GOLOVIN, B. N. *Jazyk i statistika*. Moscou: 1971.

GOLOVIN, B. N.; PEREBEJNOS, V. I. *Voprosy statisti èeskoj stilistiki*. Kiev: 1974.

GUIRAUD, P. *Bibliographie critique de la statistique linguistique*. Utrecht/Anvers: 1954.

GUIRAUD, P. *Problemes et methodes de la statistique linguistique*. Paris : 1960.

HERDAN, G. *Quantitative linguistics*. London : 1964.

HOFFMANN, L. (Org.). *Reihe Fachwortschatz Russisch-Englisch-Französisch* (Medizin, Physik, Chemie, Mathematik, Bauwesen u. a.). Leipzig: 1970 *et seq.*

HOFFMANN, L. (Org.). *Fachsprachen und Sprachstatistik*. Berlin: 1975. (Sammlung Akademie-Verlag 41 Sprache)

HOFFMANN, L. (Org.). *Fachwortschatz Medizin*. 7. ed. Leipzig: 1986.

HOFFMANN, L. *Kommunikationsmittel Fachsprache*. 3. ed. Berlin: 1987a. (Sammlung Akademie-Verlag 44 Sprache)

HOFFMANN, L. (Org.). *Fachsprache - Instrument und Objekt*. Leipzig: 1987b. (Linguistische Studien)

HOFFMANN, L.; PIOTROWSKI, R. G. *Beiträge zur Sprachstatistik*. Leipzig : 1979. (Linguistische Studien)

MULLER, C. *Initiation à la statistique linguistique*. Paris: 1968.

NALIMOV, V. V. *Verojatnostnaja model' jazyka*. Moscou: 1974.

NÜBOLD, P. *Quantitative Methoden zur Stilanalyse literarischer Texte:* Ein Bericht über den Stand der Forschung in der Anglistik. Braunschweig: 1974. (Braunschweiger Anglistische Arbeitend)

PEREBIJNIS, V. S. *Statisti èni parametri stiliv*. Kiev: 1967.
PIOTROVSKIJ, R. G.; BEKTAEV, K. B.; PIOTROVSKAJA, A. A. *Matemati èeskaja lingvistika*. Moscou: 1977.
TULDAVA, J. *Problemy i metody kvantitativno-sistemnogo isslodovanija leksiki*. Tallin: 1987.

Abordagem cognitiva da tradução nas línguas de especialidade: para uma sistematização da descrição da conceituação metafórica[1]

Sylvie Vandaele e Leslie Lubin[2]
Tradução: Daniel Costa da Silva[3]
Revisão: Patrícia Chittoni Ramos Reuillard e Cleci Regina Bevilacqua[4]

Resumo: *A conceituação metafórica é um processo fundamental do pensamento, amplamente empregado no estabelecimento do modelo científico, incluindo a biomedicina. Para alcançar a compreensão dos textos científicos, o leitor precisa ser capaz de entender as metáforas conceituais da área de especialidade em questão. De acordo com nossa hipótese de trabalho, a conceituação metafórica não somente estabelece a especificidade de um domínio, mas também subentende, em grande parte, a terminologia e a fraseologia das línguas de especialidade. Saber identificar essas metáforas conceituais gera, portanto, uma poderosa ferramenta cognitiva, que permite fundamentar um grande número de decisões tradutórias. O presente artigo estabelece uma metodologia que possibilita a descrição sistemática de colocações motivadas por uma conceituação metafórica subjacente. Será inicialmente abordado o tema da identificação das unidades lexicais que denotam uma conceituação metafórica ("índices de conceituação"). Após, será proposta uma caracterização das colocações empregadas, em relação a essas unidades lexicais particulares. A descrição das colocações será considerada com base nas funções lexicais da Teoria Sentido Texto, as quais poderão ser usadas para facilitar a passagem de uma língua à outra.*

Palarvas-chave / Keywords: *Conceituação metafórica;Índice de conceituação; Tradução especializada; Funções lexicais;Colocações.*

---

[1] Traduzido com a permissão das autoras para publicação nos Cadernos de Tradução do IL, a partir do texto original sob o título Vandaele S. et Lubin, L. (2005) «Approche cognitive de la traduction dans les langues de spécialité : vers une systématisation de la description de la conceptualisation métaphorique» META, numéro spécial dirigé par H. Lee-Jahnke, vol. 50, nº 2, p. 415-431. Trabalho realizado com o apoio do CRSH – Conseil de Recherche en Sicences Humaines du Canada.

[2] Universidade de Montreal, Montreal, Canadá. As autoras agradecem às professoras Cleci Regina Bevilacqua, Maria da Graça Krieger e Patrícia C. R. Reuillard.

[3] Bacharel em Letras, habilitação francês-português, UFRGS.

[4] Professoras do Instituto de Letras, UFRGS.