DAILLE, B. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse de Doctorat. Université Paris VII, 1994.

FLOWERDEW, J. Definitions in Science Lectures. In: *Applied Linguistics*. Vol. 13 (2), 1992, p. 202-221.

JACQUELINE, C. ROYAUTE, J. Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework. In: *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on research and Development in Information Retrieval*. New York, Heidelberg: Springer-Verlag, 1994, p. 132-141.

NKWENTI-AZEH, B. *Positional and Combinational Characteristics of Satellite Communications Terms*. Final Report, Eurotra Project, UK-CCL-UMIST, 1992.

SINCLAIR, J. Corpus Typology: A Framework for Classification EAGLES document. 1-18. In: MELCHERS, G.; WARREN, B. *Studies in Anglistics*. Stockhom, Almquist and Wiksell International, 1995, p. 17-34.

TRIMBLE, L. *English for Science and Technology: A Discourse Approach*. Cambridge University Press, Cambridge, 1985.

CORPORA COMO PONTO DE PARTIDA PARA A EXTRAÇÃO DE DADOS TERMINOLÓGICOS[1]

Heribert Picht[2]
Tradução: Danatela Duarte[3] e Maria José Bocorny Finatto[4]
Revisão: Ulla Marisa Pedde Muss[5] e Maria José Bocorny Finatto

1-Introdução

Um *corpus* pode ser definido, de modo bastante genérico, como uma coleção de documentos que é compilada com base em critérios de seleção específicos, de tal maneira que conforma um conjunto empregável para uma ou mais finalidades. Uma definição como essa nada menciona sobre sua forma de apresentação e também não diz que um *corpus* deve ser apresentado de maneira que possa ser lido por computador. Mesmo antes da era dos computadores, no fundo, praticamente todas as pesquisas se basearam em coleções de textos – em *corpora* – de uma ou outra forma. Dito de outro modo: a idéia de *corpus*, em si, não é nova.

É incontestável que *corpora* passíveis de serem lidos por computador oferecem muitas vantagens. O tratamento de grandes quantidades de dados, o que permite um embasamento mais amplo para pesquisas, o processamento rápido e a utilização para diversas finalidades são, de modo geral, suas vantagens reconhecidas. As duas primeiras são de natureza quantitativa e podem ter uma influência direta ou indireta na qualidade da pesquisa. Entretanto, não podemos esquecer que esses elementos são, antes de tudo, de essência quantitativa, pois os programas de computador

[1] Traduzido, com a permissão do autor, a partir do texto em alemão "Korpora als Ausgangspunkt für die Extraktion von terminologischen Daten", publicado na revista SYNAPS 8(2001), p. 38-48.

[2] Norges Handelshøyskole, Institutt for språk, Norwegian School of Economics and Business Administration.

[3] Acadêmica do curso de Bacharelado – Tradução, UFRGS.

[4] Professora do Setor de Língua Portuguesa do Depto. de Letras Clássicas e Vernáculas, Instituto de Letras, UFRGS.

[5] Formanda do curso de Bacharelado – Tradução, UFRGS.

utilizados para processamento de um *corpus* (1) não têm propriedades analíticas cognitivas; e, (2) um *corpus* não pode fornecer mais do que aquilo que contém. Essa última limitação citada pode ser contrabalançada até um determinado ponto pela quantidade de dados com que se lida, mas, mesmo assim, há que se contar com alguns limites.

Neste trabalho não será colocado em discussão em que medida a Lingüística de *Corpus* poderia ou não ser considerada uma disciplina autônoma. Sobre essa questão, veja Hansen (1988). Hansen (*op.cit.*) examinou uma série de definições diferentes para o conceito de "*corpus*" e pôde constatar, a partir delas, as seguintes características em comum:

- um conjunto finito de dados lingüísticos sob a forma de textos ou partes de textos orais ou escritos;
- um recorte de uma área temática dada, que é compilado para uma finalidade específica;
- um subconjunto de dados tomado de um conjunto básico maior, de tal forma que todos os elementos desse conjunto básico estão contidos no subconjunto, com igual probabilidade estatística, de modo que esses elementos se distribuem de maneira igual ao longo de todo o conjunto básico;
- um subconjunto de dados que é representativo em relação a um conjunto básico maior.

A partir dessas características, o autor depreende cinco traços categoriais gerais de um *corpus*. São eles:

1. função, no que o autor distingue entre função de uso e função metodológica;
2. conteúdo;
3. extensão;
4. estrutura;
5. representatividade.

As quatro primeiras propriedades de um *corpus* são bastante auto-explicativas e não serão exploradas aqui. Porém, a questão da representatividade merece ser especialmente tratada, tendo-se em vista um aproveitamento terminológico de *corpora*. Neste sentido, Hansen distingue:

1. representatividade estatística, que conduz aos *corpora* estatisticamente representativos;
2. representatividade qualitativa, que produz *corpora* qualitativamente representativos;
3. representatividade temática, que determina a abrangência temática dos *corpora*.

Todas essas três formas de representatividade são relevantes para o trabalho terminológico baseado em *corpus*.

De acordo com finalidade e objetivo de uma investigação, outras classificações, com base em outros critérios, também são relevantes. Pode-se diferenciar, por exemplo, entre:

1. *corpora* polifuncionais e monofuncionais;
2. *corpora* de fragmentos de texto (*sample corpus*) e *corpora* de textos completos (*monitor corpus*);
3. *corpora* de linguagem comum e de linguagens técnico-científicas.

Na seqüência dessas observações gerais me ocuparei, daqui em diante, com questões terminológicas práticas, destacando os seguintes assuntos:

1. tipos de dados terminologicamente relevantes e suas formas de representação;
2. *corpora* para usos específicos;
3. Idade e atualidade de textos;
4. competência especializada do autor do texto – tipos de texto;
5. *corpora* para trabalhos terminológicos multilíngües;
6. estruturas de conhecimento em *corpora* e em produtos terminográficos.

2- *Corpora* e trabalho terminológico

2.1 Tipos de dados terminologicamente relevantes e sua forma de representação

Um *corpus* se compõe de uma quantidade finita de formas de representação explícitas e implícitas, que podem ser significantes para o trabalho terminológico – esse é um dado banal. Mas, quais formas de representação de dados terminológicos são relevantes?

A tabela a seguir sobre dados que são terminologicamente relevantes é derivada da norma ISO 12620 – *Computer applications in terminology*:

Formas de Representação		
	Lingüísticas	**Não-lingüísticas**
Objeto	nomes	figuras
	descrição de fatos	
Conceito	denominações e	
	informações adicionais	
	p.ex. de cunho gramatical	
	fórmulas	fórmulas
	paráfrases	símbolos
	unidades fraseológicas	representações gráficas
	definições	
	explicações	
Relações	expressas verbalmente	representadas graficametnte
		p.ex.
		sistemas de conceitos

Os *corpora*, de um modo usual, contêm apenas texto. *Corpora* que são compostos por texto e por figuras ainda são raros e a sua utilização se encontra em estágio experimental.

Mas, quais dados terminológicos, em função da sua forma de representação, podem ser extraídos de um *corpus*?

1. Denominações compostas por uma única palavra podem ser identificadas sem maiores problemas pela comparação entre signos, mas precisam ser trabalhadas sem apoio informatizado para que se possa verificar se são, realmente, denominações em sentido terminológico.
2. Denominações compostas por mais de uma palavra podem ser detectadas através de programas de concordância, mas não é raro que seja preciso estabelecer, sem auxílio do computador, os limites de início e fim dessas denominações.
3. Unidades fraseológicas podem ser identificadas por meio da busca de uma "denominação núcleo" e da sua periferia; mas também aqui é inevitável a relação homem - computador.
4. A paráfrase de um conceito só pode ser constatada por meio da observação de indicadores tais como "pode ser também ser compreendida como...", pois o seu acesso direto não é possível.
5. Definições e explicações também só podem ser reconhecidas através da observação de marcadores, em maioria verbos, tais como "definir", "significar", "explicar". Também aqui é impossível um acesso direto por via informatizada.
6. Exemplos se enquadram nas últimas duas categorias.
7. Representações gráficas e figuras geralmente só podem ser identificadas por meio de indicadores como "Fig.X" e são recuperáveis, na maioria das vezes, apenas nos documentos originais.

Em resumo, pode-se constatar que existem duas modalidades de extração de dados de um *corpus*: 1. através da comparação direta entre signos; e *2.* através da interação variada com o computador. Certamente já existe uma variedade de *softwares* que facilitam essa interação, mas, ainda assim, não podem substituí-la completamente. A respeito das possibilidades atuais e futuras para extração dos tipos de dados antes mencionados, veja Estopà (2000). Em qualquer caso, ainda assim, os instrumentos eletrônicos só poderão nos fornecer dados terminológicos *in bruto*.

2.2 Corpora para usos específicos

Se seguirmos a divisão antes referida, proposta por Hansen, e a apreciarmos em relação à utilização terminológica de *corpora*, teremos como resultante o seguinte quadro:

1. Corpora **estatisticamente representativos.** Obtidos especialmente do processamento de material didático. Aqui poderíamos mencionar, entre outros trabalhos, os vocabulários mínimos de termos de determinadas áreas de especialização de Hoffmann, como p.ex. o seu "*Vocabulário técnico da construção civil. Dicionário de Freqüência. Russo, Inglês, Francês*", material que tem o objetivo de tornar acessível ao usuário as denominações mais freqüentes da área e seus equivalentes. O grau de abrangência desses vocabulários é de cerca de 85%.

2. Corpora **qualitativamente representativos.** Um critério importante para todo e qualquer trabalho terminológico é o fato de que o produto terminográfico final deve ser tão completo quanto possível. Isto significa que o *corpus*, via de regra, deve ser abrangente, mas, além de uma pura abrangência temática, deve haver uma cobertura de tipologias de texto. Essa abrangência tipológica inclui qualidade técnica do texto quanto às suas dimensões pragmáticas, que se realizam, por exemplo, em diferentes graus de especialização (Hoffmann 1984:65) e respectivas situações de comunicação. Além disso, devem ser levadas em conta as oscilações terminológicas das grandes comunidades lingüísticas, como por exemplo as do espanhol na Espanha e do espanhol de diferentes países sul-americanos.

3. Corpora **tematicamente abrangentes.** Alguns autores usam como sinônimo dessa expressão a designação "*corpora* tematicamente relacionados". Esse uso me parece despropositado, visto que, exatamente na área terminológica, completude de abrangência é um objetivo que sempre se persegue. A expressão "tematicamente relacionados", ao contrário, é muito vaga, subestimando em muito a exigência de especificação, e diz apenas, na verdade, que um *corpus* tem uma relação temática com um determinado tema (de especialidade).

Nesse sentido, coloca-se também uma questão sobre a capacidade de utilização de *corpora* compostos por trechos de texto em um trabalho terminológico. Através da seleção de determinados segmentos de texto, é perdida, via de regra, uma quantidade de informação terminológica relevante, que não poderia ser desconsiderada. *Corpora* desse tipo são "tematicamente relacionados" mas não são propriamente "tematicamente abrangentes". E, dessa maneira não são de interesse para um trabalho terminológico qualitativo de alto nível. Disso resulta que somente um

corpus de textos completos pode ser apropriado para um trabalho terminológico. *Corpora* de textos completos podem apresentar-se de diferentes modos. A seguir, citamos alguns exemplos.

1. *Corpora de abrangência temática quantitativa e qualitativa*

A. Um *corpus* que é composto pela obra completa de um autor, p.ex. os escritos completos de Niels Bohr sobre a Física nuclear. Um *corpus* desse tipo, por exemplo, é apropriado para a realização de pesquisa diacrônica sobre o desenvolvimento de conceitos. (Ahmad, Jensen 1998:20 ss).

B. Um *corpus* de textos completos composto de:

1. **um** tipo de texto que abrange **uma** área de especialidade e **uma** região lingüística determinada. Poderíamos, por exemplo, organizar um *corpus* das normas DIN (Instituto de Normalização da Alemanha) sobre a área de especialidade "tratamento de efluentes";
2. **um** tipo de texto que abrange **uma** área de especialidade e **várias** regiões de uma comunidade lingüística. Um *corpus* desse tipo poderia ser composto, por exemplo, por todas as normas em língua espanhola da série ISO 9000 ou por todos os textos de leis de uma mesma região lingüística que versam sobre o tema "direitos humanos".
3. **vários** tipos de texto que abrangem **uma** área de especialidade e **várias** regiões lingüísticas. *Corpora* desse tipo podem ser compostos, por exemplo, por toda a produção escrita de uma dada escola filosófica ou pela totalidade da documentação de uma empresa relativa um produto específico.

2. *Corpora de abrangência temática estatística e qualitativa*

Um *corpus* de textos completos integrado por:

A. **vários** tipos de texto que abrangem **uma** área de especialidade e **uma** região lingüística; esse *corpus* poderia oferecer variadas formas de representação para um mesmo conceito e seria especialmente apropriado para o trabalho terminológico descritivo, visto que abrangeria, simultaneamente, um espectro também mais amplo de condições pragmáticas.

B. **vários** tipos de texto que abrangem **uma** área de especialização e **várias** regiões lingüísticas; esse *corpus* incrementaria o número de formas de representação para um conceito e contribuiria substancialmente para qualificar um trabalho terminológico contrastivo-descritivo intralingual.

2.3 Idade e atualidade de textos

Idade, no que diz respeito à atualidade de textos, é um conceito vago. Por isso, é melhor falar em termos de estágios de conhecimento que estão representados em um texto. Uma indicação apenas sobre idade não é um critério confiável para a aceitação ou a exclusão de um texto em um *corpus*. Muito mais decisivo é saber o quão rápido se dá o desenvolvimento de uma determinada área de especialidade e o quão alta é a velocidade de envelhecimento do estágio de conhecimento em um texto. Em outras palavras: a velocidade do progresso de uma área de especialidade é determinante para a seleção dos textos a serem aceitos em *corpora*.

Isso significa que, na composição de um *corpus*, os especialistas desempenham um papel decisivo na escolha de textos, uma vez que somente eles estão em posição de atestar a sua representatividade técnico-científica.

Quando se trata de aproveitar um *corpus* para um trabalho de atualização de dados terminológicos, um outro ponto importante é também a manter a atualização dos *corpora*. Um *corpus* especializado envelhecido pode ser comparado a um dicionário envelhecido. Todavia, na prática, vemos somente alguns poucos casos de atualização planejada de *corpus* especializado.

4.4 Competência especializada do autor do texto – tipologias de texto

O denominador comum entre os textos técnico-científicos destinados a formar um *corpus* é a sua função comunicativa e de mediação de conhecimento. Têm sido apresentados diferentes modelos, ma maioria dos casos muito simples, sobre os sujeitos envolvidos na comunicação técnico-científica. Em geral, esses modelos se limitam a apresentar combinações tais como:

Especialista – Especialista
Especialista – Leigo
Leigo – Especialista
Leigo – Leigo

Para uma prática baseada em *corpus*, esse esquema é muito rudimentar, não corresponde à realidade, assim como também não especifica uma escolha de textos. Como a transferência de conhecimento implica simultaneamente o conceito de "lacuna de conhecimento", tornam-se necessários modelos diferenciados (Picht 1999:29ss.). Tendo-se em mente a organização de *corpora* especializados, isso significa que textos que deixem a desejar tecnicamente necessitariam ser excluídos, pois não oferecem nenhuma garantia de correção técnica. Kaufmann (1992:67), com

razão, sublinha: "Quando um *corpus* especializado, numa relação especializada, serve como instrumento, precisa ser construído sobre um fundamento correto". Esse autor, ao observar um determinado tipo de texto, demonstrou a existência de erros científicos crassos em textos que são produzidos por leigos, por exemplo, por jornalistas, para leigos.

2.5 Corpora para trabalhos terminológicos multilíngües

Via de regra, um *corpus* é compilado apenas em uma língua. Como conseqüência do trabalho terminológico multilíngüe, são exigidos *corpora* em cada uma das línguas envolvidas. No que toca à comparabilidade de conteúdo desses *corpora*, parte-se freqüentemente do princípio de que ela seja relativamente alta nos *corpora* técnicos das ciências naturais. Entretanto, essa suposição não deve valer como um axioma, já que, dependendo da área de especialização, são constatadas diferenças consideráveis, tal como p.ex. diferentes tipos de arado em diferentes países ou diferentes formas de telhado e de seus respectivos materiais em diferentes regiões geográficas.

Para as áreas de especialização mais ligadas às ciências sociais, surgem ainda maiores dificuldades na criação de *corpora* com "conteúdos equivalentes". A comparação de dados terminológicos coletados é possível somente se for feita intelectualmente, isso por causa da forte vinculação desses dados com sistemas e estruturas sociais e nacionais. Daí porque seria necessário, p.ex., uma detalhada base de conhecimento de Direito comparado, o que os textos via de regra não contêm. Disso resulta que, embora *corpora* forneçam dados terminológicos monolíngües brutos, tanto a comparação terminológica quanto o trabalho com várias línguas precisará ser executado de forma intelectual e manual, de modo não automatizado.

2.6 Estruturas de conhecimento em corpora e em produtos terminográficos

Cada texto tem a sua própria estrutura de conhecimento em consonância com seu respectivo objetivo comunicacional. Essas estruturas de conhecimento quase nunca correspondem exatamente àquelas estruturas conceituais[6] que os produtos terminográficos tomam por base, mesmo quando alguns de seus passos estruturais possam ser verbalizados e diretamente empregados em produtos terminográficos.

[6] N. de T. O autor refere-se a esquemas de representação de estruturas de conceitos ou a mapas conceituais, o que é usualmente denominado em Terminologia "árvore de domínio".

Disso resulta que, na condição de dados terminológicos ou de partes de suas formas de representação, somente podem ser obtidos e extraídos de um corpus "fundamentos básicos do conhecimento". Os resultados obtidos, como já dissemos, são dados brutos que podem gerar, por meio de uma análise terminológica e da continuidade do processamento, uma estrutura de conhecimento terminológica. Essa estrutura precisará, porém, conter todos os dados terminológicos relevantes para que o usuário de um produto terminográfico possa, de um modo restaurado e correto, situá-los novamente em um texto dotado de um objetivo de comunicação determinado.

3- Conclusão

Pode ser constatado que, na compilação de *corpora* para usos terminológicos, merecem atenção os seguintes pontos:

1.Condição do *corpus*

- Na medida do possível, o *corpus* deve ser formado sempre por textos completos.
- É desejável o mais alto grau possível de representatividade estatística.
- O *corpus* deve ser tematicamente abrangente (em abrangência por ramo de especialidade).
- É imprescindível uma alta representatividade qualitativa.
- Uma atualização constante do *corpus* é inevitável para a conservação do seu valor.

2.Seleção dos textos

- A seleção deve ser feita sempre em conjunto com especialista da matéria em foco.
- A atualidade do conhecimento é fator preponderante; a idade, ao contrário, é subordinado.

3.Ferramentas para processamento do *corpus*

- Atualmente a limitação mais importante das ferramentas de processamento está na ausência de capacidade analítico-cognitiva.
- Ferramentas têm sido melhoradas consideravelmente nos últimos anos e facilitam, assim, substancialmente, a extração de dados. Esses, todavia, são apenas dados terminológicos brutos.
- A interface "ferramenta – homem" é condicionada analítica e cognitivamente.
- A interface está em contínuo movimento pela pesquisa em Lingüística Computacional. Entretanto, a meu ver, haverá limitação fundamental enquanto faltarem a esses instrumentos propriedades analíticas e cognitivas.

4.Rendimentos da utilização de *corpora*

- Conforme vejo, existe ainda uma grande distância entre a pesquisa e os rendimentos advindos da utilização de *corpora.*
- Até onde eu saiba, não foi realizada nenhuma análise de custo-benefício que pudesse comprovar o quão grande seria ou poderia ser a economia de trabalho pelo uso de *corpora* na produção de repertórios terminológicos de alta qualidade. Daí surge uma pergunta que é quase herética: seria economicamente vantajoso, por exemplo, para uma empresa ou uma instituição, tendo em vista um trabalho de natureza terminológica, organizar e preparar *corpora* para uma utilização em níveis ótimos e ainda sustentá-los?

Sem dúvida, alguém poderia contra-argumentar que os *corpora* são hoje um passo importante para a facilitação do trabalho terminológico. Eu concordo, mas saliento que, ainda assim, também é preciso reconhecer as suas limitações, as quais, apesar de todos os melhoramentos das ferramentas de processamento, ainda persistem.

Referências Bibliográficas:

AHMAD, K; JENSEN, L.(1998): *En lille smule om atomernes bygning: Niels Bohr Writing Atomic Structure.* In: Terminology Science & Research; vol. 9(1998), no. 2. Wien: International Network for Terminology (TermNet). 20 - 33.

ESTOPÀ, R.(2000): *Extracción de terminología: elementos para Ia construcción de un SEACUSE (Sistema de Extracción Automática de Candidatos a Unidades de Significación Especializada).* Tesis doctoral. Institut Universitari de Lingüística Aplicada. Universidad Pompeu Fabra Barcelona.

HANSEN, Steffen L. (1988): *Korpuslinguistik. Teori - metode - praksis.* LAMBDA nr. 5. Institut for Datalingvistik, Handelshojskolen i Kobenhavn.

ISO 12 620 *Computer applications in terminology - Data categories.* First edition 1999.

KAUFMANN, U. (1992): *Anvendelse af det danske genteknologiske tekstkorpus ved udarbejdelse af Genteknologisk Ordbog, med specielt henblik pa udvoelgelsen af eksempler.* In: Proceedings af Seminar om Korpuslingvistik i Fagsprogsforskning. Hindsgavl Slot, 26. og 27. Nov. 1992. Hindsgavl: [s.n.]. 56-68.

PICHT, H.(1999): *Die Begriffe 'Fachmann' und 'Laie'* in *der Fachkommunikation.* In: Internationale Wirtschaftsbeziehungen: Mehrsprachige Kommunikation von Fachwissen; W. Wieden, A. Weiss (Hrsg.). Göppingen: Kümmerle Verlag, 29 - 42.

STUMMANN, B. M. (1992).: *Anvendelsesmuligheder og faglig indhold af det danske genteknologiske tekstkorpus.* In: Proceedings af Seminar om Korpuslingvistik i Fagsprogsforskning. Hindsgavl Slot, 26. og 27. Nov. 1992. Hindsgavl: [s.n.]. 69 - 74.