

- TEMMERMAN, R. (2000a). *Towards New Ways of Terminology Description. The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.
- TEMMERMAN, R. (2000b). 'Why Traditional Terminology Theory Impedes a Realistic Description of Categories and Terms in the Life Sciences'. *Terminology* (special issue), 73-87.
- WÜSTER, E. (1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. 3. Aufl. Bonn: Romanistischer Verlag.
- VINAY, J.P. & J. Darbelnet. 1968. *Stylistique Comparée du français et de l'anglais*. Paris: Didier.
- WINTERSON, J. (1994). *Art and Lies*. London: Jonathan Cape.

COMO TER ACESSO A ELEMENTOS DEFINITÓRIOS NOS TEXTOS ESPECIALIZADOS?¹

Jennifer PEARSON²

Tradução: Carolina Huang e Sandra Dias Loguercio³

Revisão: Luzia Araújo⁴

Resumo: Neste artigo, mostro a vantagem em se utilizar corpora de textos especializados num trabalho terminográfico (identificação de termos e elaboração de definições terminológicas). Tento identificar os tipos de textos especializados adequados a este trabalho e proponho alguns critérios para a seleção dos textos apropriados. Parto do princípio de que, em muitas situações de comunicação especializada, os autores irão, de maneira explícita ou implícita, explicar alguns dos termos que utilizam. Tomando por base três corpora distintos, proponho uma metodologia para se chegar a esses elementos definitórios.

Palavras-chave: corpora; extração de informação; definição; comunicação especializada.

1 INTRODUÇÃO

A terminografia exige grandes investimentos de tempo, dinheiro e energia. A disponibilidade de textos eletrônicos deveria permitir-nos construir corpora de textos relevantes que serviriam a uma extração

¹ Traduzido com a permissão da autora a partir do texto em francês "Comment accéder aux éléments définitoires dans les textes spécialisés?", publicado em *Terminologies Nouvelles* nº 19, mai 1999.

² Diretora do Bureau de Tradução, UNESCO, Paris.

³ Alunas de pós-graduação, Instituto de Letras, UFRGS.

⁴ Tradutora, Bolsista RD-CNPq e professora colaboradora, Instituto de Letras, UFRGS.

terminográfica semi-automática. Minha intenção é falar sobre a concepção e a construção de corpora voltados a essa atividade e propor uma metodologia para a extração de informações terminológicas.

O trabalho baseado em corpus não é uma atividade recente. Já há três décadas, pesquisadores constroem corpora cada vez maiores e os exploram com objetivos distintos: produzir dicionários, estudar a gramática de uma língua ou observar a sua evolução. O que caracteriza os grandes corpora mais conhecidos (pelo menos em língua inglesa) é o fato de conterem textos muito variados. Esses corpora foram construídos para serem representativos da língua tal como é usada no cotidiano e a intenção é criar um recurso que reflita os elementos comuns dessa língua. Se os corpora contêm textos especializados, estes representam apenas uma parte muito pequena de seu conjunto. Conseqüentemente, são menos interessantes para quem procura estudar a linguagem das áreas de especialidade. O trabalho baseado em corpora especializados é muito mais recente e, durante muito tempo, os poucos pesquisadores que desejavam explorar corpora de textos especializados eram obrigados a criar seu próprio corpus. Por diferentes razões, não era fácil construir esses corpora: quase não havia textos eletrônicos, era difícil obter as permissões necessárias e, por parte dos pesquisadores, não havia disposição para partilhar ou emprestar recursos já existentes. Por conseguinte, durante muito tempo, o número e o tamanho dos corpora de textos especializados eram relativamente pequenos. No entanto, essa situação evoluiu de maneira notável nesses últimos anos. Os editores mostram-se mais dispostos a dar sua permissão e o volume de textos disponíveis sob forma eletrônica não pára de crescer. Atualmente, quem quiser construir um corpus de textos especializados depara-se com a dificuldade das múltiplas opções, o que acarreta outros problemas sobre os quais retornarei mais adiante.

Minha proposta é levantar algumas questões que me parecem importantes para quem se interessa em trabalhar com corpora de textos especializados. Pergunto inicialmente o que é um corpus, para estabelecer o que distingue um corpus de outros conjuntos de textos, como arquivos, por exemplo. Tento também caracterizar aquilo que distingue uma abordagem baseada em corpus de uma abordagem mais tradicional. Depois, pretendo estabelecer o que para nós significa o termo *texto especializado*. Este termo é utilizado por comunidades diferentes para descrever um grande número de tipos de textos, sendo que nem todos são adequados à realização de um trabalho terminográfico. Procuo, portanto, na segunda parte de meu artigo, definir o termo *texto especializado* e caracterizar o tipo de texto especializado mais adequado para uma exploração terminográfica precisa, isto é, a extração de elementos

definitórios que servirão de base para a formulação de definições terminológicas. Assim que tivermos caracterizado o gênero de texto que nos interessa, restará decidir quais os critérios a serem aplicados a fim de selecionar os textos mais apropriados à nossa atividade. Na terceira parte, proponho, portanto, um conjunto de critérios que utilizo em meu próprio trabalho terminográfico.

A maior parte deste artigo consiste em descrever a metodologia adotada para realizar meu trabalho. Concebi duas abordagens diferentes: uma que consiste em explorar a presença de certas estruturas gramaticais e outra que consiste em utilizar os próprios termos como ponto de partida. A tônica deste artigo estará na primeira dessas abordagens. É importante notar que a proposta que faço não é uma solução perfeita; ela não leva a uma solução automática. O especialista terá somente os elementos com os quais poderá trabalhar depois.

2 O QUE É UM CORPUS?

Segundo Sinclair, um corpus é “uma coleção de fragmentos de linguagem que são selecionados e ordenados de acordo com critérios lingüísticos explícitos, usada como amostra dessa linguagem” (1994: 2). Segundo Atkins *et al.*, um corpus é “um subconjunto de uma BTE (biblioteca de textos eletrônicos) construído de acordo com critérios de projeto explícitos para um propósito específico, por ex. o Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), o Cobuild Corpus, o corpus Longman/Lancaster, o Oxford Pilot Corpus” (1992: 1). Uma vez que essas definições estão bem próximas daquelas propostas por outros lingüistas, podemos considerá-las como sendo representativas. Tais definições são reveladoras e nos permitem observar em que medida um corpus distingue-se de outras coleções de textos. Em primeiro lugar, os textos que encontramos num corpus não estão lá por acaso. Não se trata de tomar qualquer texto simplesmente porque está disponível. Cada texto é selecionado conforme critérios lingüisticamente precisos. Em segundo lugar, se criamos um corpus, o fazemos com um objetivo preciso; sabemos por que o criamos, sabemos com antecedência a que irá nos servir. Em terceiro lugar, os textos que encontramos num corpus são textos autênticos e naturais, isto é, não foram editados. Ainda que um corpus contenha textos que foram redigidos para serem lidos, ele não será jamais utilizado desta forma. Se nos ocorre de ler fragmentos, esses fragmentos serão provavelmente de textos distintos e, mesmo quando forem do mesmo texto, não virão em seqüência no texto de origem. Utiliza-se ferramentas eletrônicas (como *WordSmith Tools*) para se chegar a um corpus; essas

ferramentas nos permitem observar palavras em seu contexto, mesmo que, de uma certa maneira, possamos vê-las também fora dele. O fato de ler várias ocorrências ao mesmo tempo nos permite identificar tendências que não teríamos observado a partir de uma abordagem mais convencional. Em suma, um corpus é um produto artificial que contém textos autênticos, selecionados de acordo com critérios precisos para serem utilizados como amostra de uma linguagem. No que se refere a nosso trabalho, nossos corpora conterão textos especializados, o que nos leva à seguinte questão: o que caracteriza os textos especializados?

3 CARACTERIZAÇÃO DO TEXTO ESPECIALIZADO

Geralmente utilizamos o termo *texto especializado* para diferenciar os textos ditos *de linguagem geral* e os textos ditos *de linguagem de especialidade*. Podemos utilizar critérios diferentes para caracterizá-los e tais critérios dependerão, em grande parte, da razão pela qual se estuda esses textos. Alguns pesquisadores se interessam por aspectos lexicais, outros pelos gramaticais e outros ainda pelos de tipologia de texto. Aqueles que se interessam pelos aspectos lexicais, como os terminólogos, dirão que um texto especializado se caracteriza por seu vocabulário especializado, isto é, pela frequência de termos técnicos empregados no texto. Aqueles que se interessam pelos aspectos gramaticais, como os pesquisadores que se dedicam às sublinguagens, dirão que os textos especializados se caracterizam não apenas pelo vocabulário utilizado, mas também, e talvez mais, por suas estruturas gramaticais. Isto significa que, muitas vezes, a gramática dos textos especializados é diferente da gramática da linguagem geral, diferente porque é limitada no que se refere às estruturas utilizadas, ou porque as estruturas gramaticais que encontramos não obedecem às regras de gramática normais (uso de elipses, construção de termos compostos longos e complicados). Aqueles que se interessam pela tipologia dos textos se ocupam principalmente da função e da organização dos textos. Se, por exemplo, estudarem a linguagem da medicina, tentarão antes fazer um balanço das diferenças entre os diversos *tipos* de documentos, sejam artigos de pesquisa, bulas de remédio, relatórios de laboratório.

Quando queremos construir um corpus de textos especializados que sirva de base a um trabalho terminográfico, nenhuma das formas de caracterização dos textos citados acima, porém, mostra-se adequada às exigências de tal trabalho. Ao contrário do que se poderia pensar, nenhum texto especializado é adequado a uma extração terminográfica. Lembremos que nosso objetivo principal é selecionar textos que

contenham elementos definitórios, ou seja, textos nos quais alguns dos termos utilizados sejam definidos de forma implícita ou de forma explícita. Se utilizarmos unicamente o critério da frequência dos termos para selecionar nossos textos, iremos nos deparar com muitos textos que são pobres em elementos definitórios. O mesmo ocorre com o critério das estruturas gramaticais incomuns. Aparentemente, o critério de tipologia poderia servir melhor, mas ele também acarreta problemas. Parece-me que toda discussão sobre o que é um texto especializado deve considerar um elemento muito importante que é extra-textual, isto é, a relação entre o autor e o leitor.

Este aspecto é primordial quando queremos decidir se um texto deve fazer parte de nosso corpus. Procuramos construir um corpus que nos permita apontar não apenas termos, mas também elementos definitórios. Se podemos esperar encontrar um número considerável de termos técnicos em todo texto especializado, o mesmo não ocorre com os elementos definitórios. Da mesma forma que há muitas situações em que autores explicam de maneira explícita certos termos que utilizam, há também em mesma quantidade, se não mais, situações em que não o fazem. Por exemplo, não se espera encontrar definições em boletins de meteorologia, nem em receitas culinárias, nem em bulas de remédio. Em contrapartida, espera-se encontrar em livros de formação, em normas industriais, e até mesmo, às vezes, em artigos de pesquisa quando se trata de criar novas noções ou de redefinir noções existentes.

As relações entre o autor e o leitor determinarão a quantidade de explicações que serão fornecidas num dado texto e isto é válido para qualquer tipo de texto. Creio que existem três tipos de relações autor-leitor que podem nos interessar. O primeiro se refere à comunicação entre especialistas: os textos redigidos por especialistas para especialistas terão uma frequência muito alta de termos técnicos. Consideremos, por exemplo, artigos de pesquisa publicados em revistas especializadas. Nesses textos, encontraremos uma densidade muito alta de termos, mas provavelmente muito poucos elementos definitórios. A explicação é simples: supõe-se que o leitor conhece e entende os termos utilizados.

O segundo tipo de relação diz respeito à comunicação entre especialistas e pessoas que já têm uma certa competência na área em questão. Nesse caso, penso nas pessoas que trabalham na mesma área, mas que não possuem o mesmo nível de formação. Pode ser comunicação entre engenheiros e técnicos, médicos especializados e clínicos gerais. O que distingue esta situação da anterior é o nível de especialização. Os autores vão definir ou explicar alguns dos termos que utilizam quando acham que esses termos não são conhecidos por seus leitores.

O terceiro tipo de relação se refere à comunicação entre especialistas e pessoas que não têm formação nenhuma na área em questão, mas que necessitam conhecer e entender esta terminologia. Nessa situação de comunicação, a densidade de termos será nitidamente menos elevada do que nas duas categorias anteriores, porém, pode-se esperar encontrar uma grande densidade de elementos definitórios. Penso, por exemplo, em livros escolares ou livros de formação profissional. O que chama atenção aqui é que este tipo de texto raramente é considerado interessante por especialistas e estudiosos da linguagem especializada, simplesmente porque a frequência de termos não é elevada de modo suficiente. Como veremos mais adiante, são precisamente esses textos os mais ricos em elementos definitórios, sendo os da comunicação entre especialistas os mais pobres. O que é importante lembrar, no caso de cada uma dessas relações autores-leitores, é que o cenário comunicativo no qual encontraremos textos que correspondem a essas relações é profissional ou educativo.

Existe um quarto tipo de relação autor-leitor que poderíamos ter considerado, mas que se mostrou inadequado às nossas necessidades. É o caso das relações entre especialistas e leitores ocasionais; trata-se de comunicações especializadas que aparecem na literatura dita *de interesse popular*, em revistas como *Scientific American* nos EUA ou *New Scientist* na Inglaterra⁵, por exemplo. Nesse caso, o âmbito não é profissional nem educativo no sentido em que os leitores que lêem esses textos não o fazem no âmbito de seu trabalho ou formação.

4 CRITÉRIOS DE SELEÇÃO

Agora que já decidimos que nos interessa selecionar textos que correspondam a um dos três cenários comunicativos citados anteriormente, resta-nos estabelecer um conjunto de critérios para selecionar os textos que queremos incluir em nosso corpus. Não basta dizer que este ou aquele texto corresponde a um de nossos cenários comunicativos. Há outros critérios que entram em jogo, sendo os mais importantes:

a. Os textos devem ter sido publicados. Isto significa que, segundo a definição de Biber, “eles estão impressos em múltiplas cópias para distribuição, têm registro de direito autoral ou estão registrados em algum outro serviço de indexação” (1993: 245). Pensamos em analisar textos que foram redigidos para serem lidos.

⁵ N. de T.: *Superinteressante* no Brasil.

b. Os textos devem ser inteiros. Tradicionalmente, os corpora contêm exemplos de uma extensão precisa de textos diferentes.

c. O autor pode ser um indivíduo, um grupo de indivíduos, uma associação ou uma federação. Em qualquer desses casos, o autor deve ser competente para escrever no domínio em questão. Sua formação profissional deve corresponder ao assunto de que trata.

d. O âmbito deve ser profissional ou educativo. Os textos são destinados a serem utilizados num âmbito profissional ou num âmbito de ensino.

e. A data da publicação de um texto pode ser importante, sobretudo nas áreas em plena evolução, como a informática.

f. A função do texto deve ser informativa, didática ou normativa.

5 OS CORPORA

Na elaboração de minha metodologia, optei por utilizar três corpora diferentes. O primeiro, o corpus *Nature*, é uma coleção de artigos publicados na revista *Nature*. A *Nature* é uma revista inglesa especializada que contém artigos redigidos por especialistas para especialistas. O cenário comunicativo corresponde, portanto, à nossa primeira categoria, a comunicação entre especialistas. Optei por estudar este corpus porque minha intuição me sugeria que, por mais especializados que fossem estes textos, haveria grandes chances para que houvesse poucos elementos definitórios, o que se revelou ser o caso. O corpus *Nature* contém aproximadamente 230.000 palavras.

O segundo corpus, o corpus UIT, redigido em inglês, francês e espanhol, foi publicado pela União Internacional das Telecomunicações e está disponível em CD-Rom. O cenário comunicativo corresponde à nossa segunda categoria: trata-se de comunicação entre especialistas e pessoas que trabalham na mesma área, mas que não possuem necessariamente o mesmo nível de competência que os autores. O corpus UIT compreende textos normativos e contém aproximadamente 4,7 milhões de palavras.

O terceiro corpus contém livros didáticos para matérias do programa de estudos GCSE⁶ na Inglaterra. Os alunos passam no exame GCSE dois anos antes de passar no equivalente ao *baccalauréat* francês. O cenário comunicativo corresponde à nossa terceira categoria, a comunicação entre especialistas e pessoas que não têm muito, ou têm pouco ou nenhum conhecimento do domínio em questão, mas que

⁶ Exame nacional (geral, tecnológico ou profissional) prestado ao final do ensino secundário. (N. de T.)

necessitam conhecer e entender suas noções. O corpus compreende textos didáticos e contém aproximadamente um milhão de palavras.

6 ELEMENTOS DEFINITÓRIOS

Agora que definimos o que queremos dizer com *texto especializado* e, mais particularmente, o que entendemos *por um corpus especializado com finalidade terminográfica*, vou dirigir minha atenção à noção de elementos definitórios. Parto do princípio de que, nos âmbitos de comunicação mencionados anteriormente, os textos conterão elementos definitórios, isto é, os autores nos fornecerão explicações de alguns dos termos que utilizam.

Creio que existam vários meios de se ter acesso a esses elementos definitórios: por intermédio do termo, pelas estruturas gramaticais, etc. Proponho primeiramente a elaboração de uma descrição dos elementos definitórios do ponto de vista de sua estrutura gramatical. Partirei de estruturas muito formais para chegar em seguida a estruturas menos formais e, portanto, mais difíceis de serem identificadas. Proponho a utilização das descrições elaboradas por Trimble, em 1985, e depois desenvolvidas por Flowerdew, em 1992, dois pesquisadores que analisaram textos autênticos. Segundo esses pesquisadores, existem três categorias de definição: definição formal, definição semi-formal e definição não formal. Segundo Trimble, a definição formal é “claro, a já conhecida equação ‘Species = Genus + Differentia’, geralmente chamada formal devido à sua regidez de forma” (Trimble 1985: 75-76). A definição formal tem uma estrutura rígida que geralmente é representada pela seguinte fórmula: $X = Y + característica$. Nesta fórmula, x representa o termo, y , um hiperônimo, e a característica serve para distinguir x de todos os outros termos da mesma categoria. Eis alguns exemplos que encontramos em nossos corpora⁷:

- *A control circuit is a telephone-type circuit between the point of origin of the programme and the point where it terminates.* (UIT)
- *An ISDN connection is a connection established between ISDN reference points.* (UIT)
- *A spore is a single cell which by itself can grow into a new organism.* (GCSE)

Quanto à definição semi-formal, Trimble diz “por definição, uma definição semi-formal contém apenas dois dos três elementos definitórios

⁷ Estes e os demais exemplos foram mantidos em inglês por se tratar do corpus de trabalho da autora. (N. de T.)

básicos: o termo sendo definido e a exposição de diferenças” (Trimble, 1985: 77). A definição semi-formal é geralmente representada pela fórmula: $x = característica$, como nos seguintes exemplos:

- *Geometric elements are used to construct drawings of various types by a succession of overlay of points, straight lines, arcs, etc.* (UIT)
- *Photographic elements are used to render an image by the transmission and display of an array of picture elements (pixels).* (UIT)
- *Arachnids have four legs and their bodies are made of two parts.* (GCSE)
- *Amylase breaks up starch into sugar.* (GCSE)

A terceira é uma definição não formal, assim descrita por Trimble: “A função de uma definição não formal é definir num sentido geral para que um leitor possa perceber o elemento familiar em qualquer que seja o novo termo... A maioria das definições não formais é encontrada em forma de sinônimos” (1985: 78). Esse tipo de definição fornece ao leitor o nome correto do termo e uma outra unidade lexical que corresponde mais ou menos ao seu sentido, mas que não é obrigatoriamente um termo. Flowerdew diz que se trata de *uma definição por substituição* e sugere que tal definição pode ser expressa de três diferentes maneiras: por sinônimo, paráfrase ou derivação (1992: 211). Eis alguns exemplos do que encontramos em nossos corpora:

- *a signal of limited duration, known as a “measuring signal”* (UIT)
- *the aggregate of time during which the speech in question is present (called the active time)* (UIT)
- *A bulge called a pseudopodium...* (GCSE)
- *Peas have small shoots called tendrils* (GCSE)
- *...an extended region of homology called the POU domain* (Nature).

É interessante observar aqui que encontramos, além das paráfrases, não apenas relações de quase sinonímia, mas também relações de hiperônimo-hipônimo.

6.1 DEFINIÇÕES FORMAIS

6.1.1 Definições Formais Simples

Ao estudar as estruturas gramaticais utilizadas nos corpora, constatei que algumas delas correspondem geralmente aos diferentes tipos

de definições apresentadas anteriormente. Construí, então, um conjunto de condições que uma estrutura gramatical deve preencher para ser considerada um elemento definitório. Observemos, primeiramente, as condições que devem preencher as definições formais simples, isto é, aquelas que correspondem à estrutura $x = y + característica$ e que aparecem em uma única frase.

1. *X* deve ser um termo. Para ser considerado termo, deve ter uma estrutura que corresponda às estruturas terminológicas identificadas pelo corpus em questão. Se *x* aparece no início da frase, deve ser precedido ou do artigo indefinido ou de nenhum artigo. Isso é muito importante porque, em inglês, se o termo é precedido do artigo definido (quando está em início de frase), ele não tem necessariamente um estatuto genérico. Se o termo é precedido do artigo definido ou de um adjetivo demonstrativo, isto pode significar duas coisas: ou que a frase é a continuação de uma definição em parte já elaborada na frase precedente, o que é o caso das definições complexas que não nos interessam aqui, ou que a frase diz respeito a um caso específico que não pode ser considerado válido para o termo em geral. Em compensação, se *x* está no final da frase, ele pode vir precedido do artigo definido ou do indefinido, ou ainda sem nenhum artigo.
2. *Y* deve ser um termo ou um hiperônimo genérico, tal como técnica, método, processo, função, etc. A tabela 1 indica os hiperônimos genéricos em inglês que encontramos nos três corpora.

Hiperônimo Genérico	UIT	GCSE	Nature
Técnica	341	7	9
Método	2.374	149	32
Processo	1.143	201	63
Função	3.105	54	133
Propriedade	39	71	9
Sistema	7.396	712	135
Classe	1.598	416	81
Equipamento	746	9	4

Tabela 1: Lista de hiperônimos genéricos

3. Há um certo número de verbos ou de sintagmas verbais que podem ligar *x* e *y*: os chamados conectivos ou verbos de ligação, que servem de conexão entre os dois elementos da frase. Alguns exemplos de

conectivos em inglês: *is/are, is/are called, is/are known as, is/are defined as, denote(s)*. O verbo ou a locução verbal que serve de ligação entre *x* e *y* deve estar no presente do indicativo e não pode ser modificado por um verbo modal. O verbo não deve ser modificado por nenhuma partícula negativa, tal como *not, never*.

A frase definitória não deve ser modificada por um advérbio de tempo ou modo. Alguns exemplos desses advérbios em inglês: *commonly, frequently, usually, specifically, generally, mainly, primarily*. A tabela 2 indica os advérbios de tempo e modo encontrados nos três corpora.

Advérbio	UIT	GCSE	Nature
<i>chiefly</i>	5	9	1
<i>commonly</i>	91	19	4
<i>especially</i>	176	142	21
<i>exceptionally</i>	88	1	4
<i>exclusively</i>	61	3	6
<i>frequently</i>	93	59	10
<i>generally</i>	496	140	35
<i>mainly</i>	98	125	20
<i>mostly</i>	11	85	18
<i>occasionally</i>	18	13	1
<i>often</i>	213	574	37
<i>only</i>	4.504	1.457	385
<i>on the whole</i>	2	13	1
<i>predominantly</i>	12	5	6
<i>primarily</i>	118	4	19
<i>principally</i>	9	0	2
<i>purely</i>	61	10	3
<i>rarely</i>	22	31	1
<i>solely</i>	93	5	5
<i>sometimes</i>	111	252	11
<i>specifically</i>	175	2	27
<i>usually</i>	354	408	17

Tabela 2: Advérbios encontrados nos três corpora

Embora se possa pensar que esses advérbios sejam utilizados freqüentemente para acentuar a validade de uma frase definitória, muitas vezes apresentam o efeito inverso. Quando um autor diz que um termo é *freqüentemente* definido desta ou daquela forma, não se pode concluir que seja *sempre* o caso.

4. Não basta determinar quais palavras podem ocupar o lugar de x , y e $=$. É preciso olhar também a forma da frase e, simultaneamente, a situação do elemento definatório na frase em geral. Acima de tudo, a frase definatória deve constituir a proposição principal da frase e não deve vir precedida de frases subordinadas. Se a frase definatória for precedida de uma frase subordinada, esta pode atenuar a força e a validade daquela. Em contrapartida, a frase definatória pode vir seguida de uma outra frase com a condição de que estejam ligadas pela conjunção *and*.

5. Y deve vir seguido diretamente pela característica que distingue x de todos os outros membros da mesma categoria. A característica pode ser introduzida por uma proposição, um particípio passado ou um pronome relativo. Y não deve vir seguido imediatamente pela conjunção *and* porque isso geralmente marca o fim do elemento definatório.

Aqui estão alguns exemplos de definições formais simples em inglês:

- *A robot is a machine that tries to copy one or more human functions (GCSE).*
- *A videotex service centre is a computer used by the videotex service provider to authorize access to a videotex service (ITU).*
- *INTERLEUKIN-1 (IL-1) is a cytokine produced by mononuclear phagocytes (Nature).*

6.1.2 Definições Formais Complexas

Embora haja um número considerável de frases definatórias que correspondem a uma definição formal simples, isto é, a uma definição formal expressa em uma única frase, há muitas outras que são expressas em duas frases: são as que Trimble chama de *definições complexas*. Em geral, uma definição formal complexa é expressa conforme uma das duas formas a seguir: a) o termo é introduzido no final de uma frase e explicado no início da frase seguinte, ou b) o termo é citado no início de uma frase e explicado na frase precedente. Eis alguns exemplos do primeiro caso e, em seguida, do segundo:

Primeiro caso (o termo aparece no final da frase precedente):

- *...mean holding time. This is the total holding time divided by the total number of seizures and can be calculated on a circuit group basis or for switching equipment. (UIT)*
- *...digital speech interpolation (DSI). This is a technique whereby advantage can be taken of the inactive periods during a conversation, creating extra channel capacity. (UIT)*
- *...a ring mains system. This is a loop of cable that runs from the consumer unit round the house and back to the unit. (GCSE)*

- *...a device called a proboscis. This is a hollow tube which is normally tightly coiled beneath the head of the butterfly when it is not feeding. (GCSE)*
- *an exoskeleton. This is a hard outer protective covering made of chitin.*

Segundo caso (definição em uma primeira frase, termo citado na frase seguinte):

- *There are millions of compounds containing just hydrogen and carbon. These are called hydrocarbons.*
- *Some bacteria get their food from the dead bodies of plants and animals. These are called saprophytic bacteria.*

Essas definições não têm a mesma estrutura formal que as definições formais simples, mas todos os elementos desta última estão presentes. Tais definições são relativamente fáceis de ser localizadas. No caso (a), citado acima, basta lembrar que as palavras que se encontram no final da primeira frase devem corresponder a uma estrutura terminológica já estabelecida e que a frase definatória deve começar com a expressão *This is...* ou *These are...* A frase definatória deve preencher as mesmas condições estabelecidas para as definições formais simples, com exceção de x que é substituído por um pronome (*this* ou *these*). No caso (b), sabe-se que as palavras que seguem a expressão *These are* ou *This is* devem ser um termo. As definições complexas são mais difíceis de ser identificadas do que as simples, mas, particularmente no corpus UIT, elas são bem mais frequentes.

6.2 DEFINIÇÕES SEMI-FORMAIS

Ainda mais frequentes são as *definições semi-formais*. Enquadram-se aí as estruturas que correspondem à fórmula: $x = \text{característica}$. Em uma definição semi-formal, o hiperônimo está ausente da frase. Na realidade, quando se analisa as definições semi-formais encontradas no corpus, descobre-se que o hiperônimo está frequentemente marcado na frase precedente e que a definição semi-formal serve para fornecer informações suplementares. As condições para as definições semi-formais diferem um pouco das condições estabelecidas para as definições formais. X deve ser um termo que pode estar precedido do artigo definido ou do indefinido. O verbo que une x à característica deve estar no presente do indicativo; a frase definatória semi-formal deve constituir a proposição principal da frase. Veja alguns exemplos de nossos corpora:

- *Expanded polystyrene is made by blowing a gas (such as carbon dioxide) into the liquid polymer. (GCSE)*

- *One of the most important applications of neutralization is in making fertilizers like ammonium sulphate ((NH₄)₂SO₄) and ammonium nitrate (NH₄NO₃). Ammonium sulphate is manufactured by neutralizing sulphuric acid (H₂SO₄) with ammonia solution (NH₃(aq)). (GCSE)*
- *Whenever a change in the status of a signalling link, route or point occurs, the three different signalling network management functions (i.e., signalling traffic management, link management and route management) are activated. This signalling traffic management function is used to divert signalling traffic from a link or route to one or more different links or routes, to restart a signalling point, or to temporarily slow down signalling traffic in the case of congestion at a signalling point... (UIT).*

6. 3 DEFINIÇÕES NÃO FORMAIS

As definições não formais são talvez as mais interessantes de se estudar. Lembremos que, segundo Flowerdew, essas definições, chamadas por ele de *definição por substituição*, podem ser expressas de três maneiras diferentes: por um sinônimo, paráfrase ou derivação. Em caso de substituição, o autor pode fornecer um sinônimo para um termo (esse caso é bastante raro) ou introduzir um termo, utilizando uma palavra equivalente de língua comum já conhecida do leitor. Em caso de paráfrase, o autor pode explicar um termo utilizando uma paráfrase. Nesse caso, estará fornecendo os mesmos elementos que encontramos nas definições formais e semi-formais, mas empregará outros métodos para fazê-lo.

Há um certo número de indicadores que assinalam a presença de definições não formais. Compreendem, entre outros, a utilização de parênteses, de expressões tais como *i.e.*, *e.g.*, *called*, *known as*. Além das funções mencionadas anteriormente, esses indicadores podem também ser utilizados para indicar relações hiperônimo-hipônimo. Veremos, a seguir, alguns exemplos de cada um dos indicadores que revelam o quanto são multifuncionais. Não se pode prever qual indicador fornecerá tal informação.

- *...but all microphones contain a disc called a diaphragm. (GCSE)*
- *...threads of pure cellulose known as rayon. (GCSE)*
- *Modulation with two-phase conditions, called bi-phase modulation (2-PSK)... (UIT)*

- *...a signal of limited duration known as a "measuring signal"... (UIT)*
- *Row crops (e.g. strawberries) can be protected from... (GCSE)*
- *Cell types, e.g. root-hair cell, egg cell (ovum) sperm cell, muscle cell... (GCSE)*
- *Some are steady state impairments (e.g. loss, noise, quantization, distortion, phase jitter...) (UIT)*
- *But how can liquids change to solids (i.e. melt) or liquids to gases (i.e. boil) or evaporate? (GCSE)*
- *An idle signalling terminal, i.e. a signally terminal not connected to a signally data link (UIT).*

7 CONCLUSÃO

A metodologia aqui proposta é uma abordagem que procura explorar a utilização de certas estruturas em circunstâncias precisas. A identificação dessas estruturas nos permite localizar elementos definitórios que podem ser considerados na formulação de definições terminológicas. Pode-se também abordar o problema de uma outra maneira, tomando os próprios termos como ponto de partida. Nesse caso, produz-se correspondências para cada termo para ver se é possível delimitar elementos definitórios nas linhas de correspondência. Essa abordagem é muito parecida com a utilizada pela equipe lexicográfica do *Cobuild*. Embora seja também muito produtiva, tal abordagem requer uma análise manual mais exaustiva. Estou certa de que, associando-se uma abordagem gramatical, tal como a que foi descrita neste artigo, com uma abordagem que toma os termos como ponto de partida para a análise de textos especializados apropriados, conseguiremos reduzir os altos custos do trabalho terminográfico.

Referências Bibliográficas:

- ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus Design Criteria. In: *Literary and Linguistic Computing*. Oxford University Press, vol. 7, 1, Oxford, p. 1-16.
- BIBER, D. Representativeness in corpus design. In: *Literary and Linguistic Computing*. Oxford University Press, Oxford, vol. 8 (4), 1993, p. 243-257.

- 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*. Stockholm, Almqvist 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 TERMINOLÓGICOS¹

Heribert Picht²

Tradução: Danatela Duarte³ e Maria José Bocorny Finatto⁴

Revisão: Ulla Marisa Pedde Muss⁵ 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.