CONTRIBUIÇÃO PARA O ESTUDO DA CONSTRUÇÃO E UTILIZAÇÃO DE *CORPORA* NO PROCESSO DE TERMINOLOGISAÇÃO

Manuel Moreira da Silva

A última década do século XX permitiu, graças à evolução tecnológica, a concretização das aspirações de muitas áreas do saber, quer em termos da obtenção de resultados que comprovassem as suas teorias, quer pelo reconhecimento do seu valor científico e contribuição no surgimento de novos paradigmas, ou de novas abordagens aos já existentes. A Linguística foi uma das áreas que mais beneficiou com o desenvolvimento das novas tecnologias, que permitiram, finalmente, analisar um sem número de fenómenos relacionados com as línguas naturais e o seu uso, descrevê-los, quantificá-los e construir, a partir dos dados pesquisados, novas certezas, ao mesmo tempo que se abriram novos caminhos para a reflexão e para a pesquisa científica.

A disciplina que se dedica à análise empírica dos dados obtidos a partir do estudo das línguas naturais é a Linguística de Corpus, cujo papel, potenciado pelas novas tecnologias, passou a ser central na análise de factos linguísticos e na sua explicitação ou (re)conhecimento. Um dos factores essenciais ao desenvolvimento desta disciplina, e, paralelamente, uma das suas principais consequências, foi o incremento, por parte de terminólogos e lexicógrafos, da pesquisa terminológica de línguas de especialidade, tarefa à qual a comunidade linguística tem vindo a atribuir maior atenção, dadas as necessidades comunicativas próprias de cada área específica do saber.

O acentuado desenvolvimento técnico e científico, e a sua necessariamente rápida e eficiente divulgação em ambientes unilingues e multilingues, obriga ao estabelecimento das respectivas terminologias nas várias línguas de especialidade, o que aumentou a necessidade de proceder à sua normalização, a fim de evitar a proliferação descontrolada de termos. A Terminologia, recorrendo a uma abordagem que concentra a sua atenção num determinado conjunto de dados linguísticos representativos de uma área científica, compilados num *corpus* ou em *corpora,* e utilizando uma metodologia de análise que considera factores quantitativos e qualitativos, internos e externos, procura determinar com exactidão a relação entre os termos de uma determinada língua

de especialidade e o seu conceito, de forma a delimitar e harmonizar o seu uso. Ao mesmo tempo constroem-se bases de dados terminológicas dinâmicas, resultado da confluência de competências e capacidades específicas e da complementaridade do trabalho inter-disciplinar.

Procuraremos neste artigo analisar os factores a considerar *a priori* no desenho e na construção de um *corpus* ou *corpora*, bem como as potencialidades que estes oferecem na sua utilização, sobretudo para a extracção de terminologia. Partindo de uma introdução geral sobre o papel da Linguística de Corpus e a problemática que rodeia a sua abordagem do estudo da língua, concentraremos a nossa atenção na delimitação dos princípios e das metodologias necessárias à construção de um *corpus* textual representativo dum ramo concreto de uma área do saber, de forma a que possa servir como base de dados utilizável na análise dessa língua de especialidade, ou em processos de extracção e normalização terminológica, apresentando, em seguida, metodologias que se orientam no sentido de tornar o processo de terminologisação quasi automático.

1. A LINGUÍSTICA DE CORPUS: RECENTRAR DA ABORDAGEM TEÓRICA

Na última década, uma das metodologias que viu renascer o interesse em torno da sua abordagem do estudo da língua foi a Linguística de Corpus. Beneficiando do incremento das chamadas Indústrias da Língua ou da Engenharia da Linguagem, de estudos desenvolvidos nas décadas anteriores, e do trabalho de recolha de *corpora* efectuado pelas mais diversas instituições, esta metodologia deixou de ser o parente pobre e esquecido da Linguística, papel que ocupou durante cerca de 30 anos, e tornou-se numa área central que fervilha de novidades e novas perspectivas. Segundo Mateus (1994: 12), o desenvolvimento de núcleos de investigação linguística na Europa e na América, enquadrados por uma visão estruturalista, conduziu a que as preocupações científicas começassem a ser "dominadas pela necessidade de criar "corpora" analisáveis e de estabelecer sistemas e subsistemas descritivos que evidenciassem a organização dos dados das línguas. O objectivo de descrever extensivamente o particular estimulou a construção de métodos e técnicas de análise [...]".

Esta construção, tal como a história da Linguística de Corpus, esteve, e está, condicionada pela tecnologia, que permite não somente o armazenamento

massivo de dados em *corpora*, mas também o seu processamento com uma rapidez e eficácia impressionantes. Daí que o percurso desta área esteja directamente relacionado com a disponibilização de ferramentas computacionais para a análise de *corpus*. A existência de uma colecção de dados linguísticos naturais, legíveis por computador, é um meio fundamental de pesquisa, sendo que, para formarem um *corpus*, estes dados devem estar armazenados de acordo com critérios e formalismos pré-estabelecidos, segundo um desenho explícito e objectivos específicos, devendo o *corpus* ser desenvolvido de forma sistemática para poder ser analisado, no seu conjunto ou em partes específicas, através de técnicas automáticas e/ou interactivas.

Em função do exposto, um *corpus* pode ser perspectivado como um

> conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira a que sejam representativos da totalidade do uso linguístico ou de algum dos seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (Sanchez, citado por Sardinha, 2000: 338).

Esta perspectiva de Sanchez está de acordo com o ponto de vista de Sinclair (1994: 2), que descreve um *corpus* como sendo "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of language". Um *corpus* resulta, assim, de uma colecção de textos naturais (não fabricados pelo computador), escolhidos e organizados de forma criteriosa (atendendo a condições de naturalidade e autenticidade), e serve para caracterizar um estado ou variedade de linguagem, constituindo amostras da linguagem (em número e género representativos), utilizáveis como base para o desenvolvimento da pesquisa linguística e do teste das suas hipóteses. Esta pesquisa é levada a cabo pela Linguística de Corpus, pela Terminologia e pela Lexicografia, através da exploração de evidências empíricas, formalizadas, extraídas e analisáveis por meio de computador.

Esta metodologia empírica fez com que o ressurgimento desta abordagem linguística não acontecesse livre de polémicas, quanto ao valor das suas pesquisas, descobertas e generalizações, uma vez que, numa primeira observação, não só punha em causa os fundamentos soberanos da gramática generativa, como contrariava na quase totalidade a sua metodologia. De facto, a discussão entre os defensores da Linguística Generativa e da Linguística de

Corpus já se prolonga há algumas décadas, centrando-se a discussão no valor relativo da intuição e da introspecção por um lado, e no valor da evidência factual ou empírica, por outro, opondo-se a competência do falante nativo, na terminologia Chomskiana, ao uso factual da língua por uma comunidade de falantes, representada num *corpus*.

Assim, os gramáticos generativistas afirmam que um corpus é "a sample of performance only and that one still needs a means of projecting beyond the corpus of the language as a whole (Cristal, 1980). (…). So the overall potential or competence of a language cannot be examined by the corpus". (Dash e Chaudhuri, 2001: 195s). Entretanto, os linguistas de corpus defendem-se, afirmando que: "Um dos aspectos mais interessantes no uso de um corpus para levar a cabo uma tarefa lexicográfica é o imediato confronto com a impossibilidade, baseada na evidência, de utilizar qualquer tipo de descrição que repouse numa fronteira estanque entre o que é admissível e o que não é" e que "só uma análise detalhada e cuidadosa de dados provenientes de corpora pode constituir uma base sólida para uma abordagem realista à construção do léxico", opondo-se, no entanto, à utilização simplista dos dados provenientes de *corpora*, que necessitam de ser modelados e estruturados à luz de uma dada hipótese teórica e de regras previamente aceites pela comunidade linguística.

Parece-nos que nenhuma das partes tem totalmente razão e que, como afirmam McEnery e Wilson (1996: 170), "As time goes on, it must be assumed that this sharp, and false, distinction between one type of language study and another, will be replaced by a synthesis of both approaches". Esta complementaridade previsível pode vir a assumir o sentido a que Kjellmer atribui a designação de "syntheticist", segundo o qual:

> Intuition and introspection have an important place in the investigation of human language. It is only by the use of those faculties or procedures that we know the framework of our own language and it is through them that we become aware of phenomena in it that need to be investigated, described, and, hence, better understood. But likewise, when our attention has been directed to such phenomenon, *corpus* work will access its place in the world of language as is used. (Kjellmer, 2001: 130).

A simples recolha e posterior análise de milhões de palavras, sem uma orientação criteriosa e uma introspecção prévia sobre o fenómeno linguístico a observar, conduziriam, assim, a resultados pouco interessantes e, até, falaciosos. Da mesma forma, a simples introspecção não teria o mesmo valor sem a

validação de um *corpus* que permitisse visualizar a frequência do fenómeno, reconhecer os seus utilizadores e o contexto de utilização, estabelecer ligações com outros fenómenos linguísticos, desambiguar sentidos, bem como obter todo um conjunto de informações relevantes e exactas conducentes à sua compreensão cabal. É assim que, num momento anterior à construção do *corpus*, o pesquisador tem que delinear, com base em princípios teóricos sólidos, uma orientação criteriosa e rigorosa para a recolha das amostras que o virão a constituir.

2. DESENHO E CONSTRUÇÃO DE UM *CORPUS* LINGUÍSTICO: TIPOLOGIA, REPRESENTATIVIDADE E EXTENSÃO

Um *corpus* textual serve de base a um sem número de pesquisas e pode prosseguir, na perspectiva de Teresa Lino (1996: 30), vários objectivos, tais como a selecção e observação do comportamento de unidades terminológicas e neónimos, o estudo de aspectos conceptuais e linguísticos associados ao aparecimento de uma noção ou conceito, a selecção de contextos, a delimitação de definições estabilizadas, e a observação de colocações, entre outros. Para que se atinjam esses objectivos com um alto grau de fiabilidade, a escolha dos textos deve ser feita cuidadosamente e em compatibilidade com os objectivos da pesquisa, de forma a que se recolha e seleccione apenas o material necessário, e se constitua a amostra desejável, sendo que um *corpus* textual, dependendo da função a que se destina, pode conter um ou mais tipos de texto, entre os quais textos científicos e técnicos, textos de "semi-vulgarização", textos de "banalização", textos de língua corrente e textos literários, todos eles produzidos com propósitos e em contextos comunicativos diferentes que interessa ponderar.

O desenvolvimento de um *corpus* depende ainda de uma série de factores a considerar *a priori* como o período de abrangência; a escolha dos documentos (livros, jornais, revistas, etc.); a escolha das páginas (aleatória, regular, selectiva); problemas com a digitalização da informação (omissão de palavras estrangeiras, citações, dialectos, símbolos matemáticos, poemas, gráficos e figuras, etc.); a forma de digitalização; a correcção dos dados recolhidos; o tamanho do *corpus*; a gestão dos ficheiros dos *corpus*, e outros.

Para a sua caracterização, contribuem ainda factores como a tipologia, a representatividade e a extensão. Quanto aos tipos de *corpus* que encontramos,

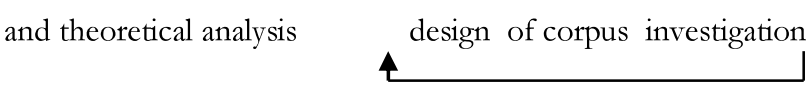
podemos dizer que a sua variedade é quase tão numerosa quanto os seus propósitos, pelo que recorreremos à seguinte tabela, criada a partir da proposta de agrupamento de Sardinha (2000: 339-341), para exemplificar a nossa perspectiva.

Tipologia para a classificação de *Corpora*

Modo	Tempo	Selecção	Conteúdo	Autoria	Dispo-sição interna	Fina-lidade	Outros meios classificação
Falado Escrito	Sincró-nico Diacró-nico Con-tempo-râneo Histó-rico	Amos-tragem (sample *corpus*) Monitor Dinâmico ou orgânico Estático Equili-brado (balan-ced)	Espe-cializado Regional ou dialectal Multi-lingue	De aprendiz De língua nativa	Paralelo Alinhado	Estudo Refe-rência Treino ou teste	Pluralidade de autorias Origem de autoria Meio Integralidade Especificidade Dialecto Equilíbrio Fechado vs aberto Renovação Temporalidade Plurilinguismo

Esta tabela, em que se descreve a tipologia empregue na definição de conteúdos e propósitos dos *corpora*, permite-nos perspectivar a panóplia de factores externos que presidem ao seu desenho e construção, para que se torne representativo de uma linguagem, de um idioma ou de uma variedade dele. Esta representatividade não é meramente uma questão de tamanho do *corpus*, como afirma Gerhard Leitner, mas do "full range of contexts of language use and the whole range of registers and genres" (2001: 151), ao que acrescenta, em acordo com Biber, que "stratified sampling" é preferível a "proportional corpus sampling", sendo que esta é uma metodologia essencialmente quantitativa, à qual se podem atribuir omissões, enquanto que a primeira se esforça por recolher amostras que incluam todas as variedades linguísticas. Biber (1993: 256) defende ainda que:

> The bottom line in *corpus* design, however, is that parameters of a fully representative corpus cannot be determined at the outset. Rather, corpus work proceeds in a cyclical fashion that can be schematically represented as follows:
>
> Pilot empirical investigation → Corpus → Compile portion → Empirical



Uma metodologia cíclica surge, para este linguista, como a única forma de estabelecer o grau de representatividade e adequação do *corpus*. Sardinha, por seu lado, liga a questão da representatividade à questão da probabilidade. Este autor está em sintonia com Haliday (1996: 30), que afirma "It had always seemed to me that the linguistic system was inherently probabilistic, and that frequency in text was the instantiation of probability in the grammar". Sardinha atribui à linguagem um carácter probabilístico e, "sendo assim, há a possibilidade de estabelecer uma relação entre traços que são mais comuns e menos comuns em determinado contexto" (Sardinha, 2000: 341). A questão da representatividade interliga-se com a da extensão do *corpus*, sendo perceptível que um *corpus* maior é, em geral, mais representativo do que um menor, estando, por isso, mais próximo da população de que deriva, ao mesmo tempo que apresenta, com maior grau de probabilidade, as palavras e as polissemias de menor frequência. Daí que tenhamos de considerar, no que se refere à extensão de um *corpus*, o número de palavras, de textos, e o número de registos e de tipos textuais, sendo que cada *corpus*, na sua especificidade e nas suas necessidades representativas, deve procurar recolher uma amostra o mais vasta e variada possível, de forma a abranger a totalidade do espectro da língua que pretende analisar, a fim de poder generalizar os resultados obtidos com segurança.

Os factores que apresentámos até aqui, e que antecedem e condicionam a construção de um *corpus,* são essencialmente externos e extra-linguísticos. No entanto, Sinclair (1995: 47) defende que tanto os critérios externos, como os critérios internos, essencialmente linguísticos, devem ser considerados aquando da classificação textual. Cita, a este propósito, Atkins *et al.* (1992), que acreditam que "é impossível equilibrar um corpus baseando-nos, apenas, nas suas características extra-linguísticas", ao que acrescentam que "um *corpus* inteiramente seleccionado com base em critérios internos não ofereceria informações sobre a relação da linguagem com o respectivo contexto situacional". Daí que na Tipologia Textual EAGLES, elaborada sob a orientação de Sinclair, as categorias aceites para a classificação textual sejam o género literário, o tema, o meio de difusão, o ser ficção ou não-ficção, o estilo e outros factores, tais como o facto de se tratar de uma tradução ou de um manual.

Estas categorias externas e internas definidas por Sinclair, apesar da sua abrangência, ignoram alguns tipos de textos, como os técnicos e científicos,

fulcrais para o processo de estandardização e normalização de termos e respectivos conceitos, bem como de extracção de terminologia especializada, sem esquecer o papel que assumem na descrição de uma língua de especialidade. Este processo de normalização terminológica surgiu como uma necessidade imperativa da sociedade moderna de controlar a denominação a atribuir às constantes inovações e desenvolvimentos que irrompem no tecido cognoscível e nas mais diversas esferas profissionais, onde o saber se impõe e divulga através da verbalização que, recorrendo a um manancial linguístico existente, renova, ora pela forma ora pelo conteúdo, a língua assumida na sua globalidade, mas agora sujeita a um uso concreto específico. A língua de especialidade assume, assim, um carácter específico e funcional, sendo usada como objecto de comunicação científica e técnica, muitas vezes em domínios interactivos, com uma forte correlação entre si, que importa compreender e delimitar.

3. AS LÍNGUAS DE ESPECIALIDADE E A EXTRACÇÃO TERMINOLÓGICA

O estudo e a análise de uma língua de especialidade, enquanto meio de expressão e comunicação no seio de um grupo específico e entre este e a sociedade no seu todo, colocam-nos, à partida, perante duas evidências: a linguagem que o grupo usa tem um valor e uma significação própria, verbaliza um saber concreto, mas é, ao mesmo tempo, composta por elementos linguísticos de um tronco comum, mais geral, ao qual foi beber o seu significado. No entanto, embora se fundamente na linguagem comum, apresenta aspectos distintos, tais como:

> um sistema conceptual mais diferenciado e mais exacto, o alargamento e especialização crescentes, a nível lexical; a nominalização, isto é, a predominância de substantivos, que constituem grande parte das terminologias (resultando essa preponderância do facto de o significado funcional de uma unidade material já estar contido na própria designação do objecto). (Bernardo, 1996: 2)

Da mesma maneira, os diversos tipos de texto científico e técnico, apesar de diferentes entre si, apresentam, no seu conjunto, alguns traços comuns, "caracterizando-se genericamente pelo primado do conteúdo sobre a forma, pelo uso de uma linguagem específica espelhada na terminologia, pela predominância da função informativa (e apelativa), pela sua universalidade... e ainda pela precisão, objectividade e clareza da informação nela veiculada".

Bernardo (1996: 1). Pensando em comunicação profissional especializada, sujeita a restrições pragmáticas, semântico-cognitivas e linguísticas, e no ideal da univocidade pretendido pela teoria terminológica tradicional, segundo a qual a cada termo deve corresponder um conceito, somos levados a imaginar que este tipo de textos e o seu conteúdo não seja contaminado por duplos sentidos, imprecisões, ambiguidades e nuances da língua comum, ao mesmo tempo que contêm um grande número de termos, dado que se referem a assuntos de elevada especificidade de uma forma previamente convencionada e aceite por uma dada comunidade, num contexto comunicativo determinado.

As características anteriormente evidenciadas permitem-nos perceber que os textos de especialidade, criteriosamente compilados em *corpora*, contêm todas as potencialidades inerentes à actividade terminológica e ao consequente processo de terminologisação. É ao analisar o uso dessa linguagem que o terminólogo espera encontrar os *metalanguage patterns* que a tornam específica ao constituírem "a common feature of certain types of specialised texts and frequently offer clues to the meaning of the terms to which they refer". (Pearson, 1998: 1). Estas marcas de metalinguagem estarão tanto mais presentes quanto os contextos comunicativos (*communicative settings),* que orientaram a produção textual e a sua recolha para o *corpus,* forem tidos em atenção na escolha e recolha dos textos.

Ao basear a sua pesquisa em *corpora* de especialidade, a Terminologia poderá desenvolver, com um elevado grau de fiabilidade, a sua actividade de compilação, descrição, processamento e apresentação de termos, bem como a da análise teórica das relações que se estabelecem entre conceitos e termos. Esta última componente corresponde ao que Mateus (1989: 179) denominou de *base ontológica da terminologia*, a qual consiste "na delimitação dos conceitos produtivos de um campo específico, sendo certo que cada termo só pode definir-se como tal quando corresponde a um único conceito, por ele transmitido com concisão e precisão". Esta percepção do termo enquadra-se na visão mais consensual de linguistas e terminólogos, como Bernardo (1996: 3), que o define como sendo "toda a palavra ou expressão limitada a um significado e a uma área específica, cujo uso está claramente estabelecido, e que se insere no contexto de uma determinada terminologia, que assegura a sua univocidade". Para Pearson (1998: 1), os termos são uma "separate class which operate as labels and appear to work in much the same way as a system of

proper names works in general language". Estes termos "mais restritos e termos mais amplos, termos convenientes e sinónimos levam à criação de uma rede ordenada do conhecimento que facilita a abstracção e a condensação da informação sobre múltiplos assuntos". (Mateus 1989: 182), o que coloca as terminologias como um instrumento adequado e facilmente manejável para dar resposta à criação e transferência de novos saberes, numa dinâmica de renovação e metamorfose contínua do léxico de uma língua.

Para tal, a terminologia desenvolveu metodologias de extracção de termos, actividade na qual os terminólogos diferem ainda bastante de opinião, sobretudo no que respeita à forma de extracção e manipulação da linguagem específica, de forma a ser usada na formulação das definições terminológicas. Como refere Pearson, alguns linguistas, como Dalile e Yang, propõem uma análise baseada na frequência da ocorrência e distribuição dos termos para decidir sobre a sua elegibilidade. No entanto, a baixa frequência de alguns dos termos, as restrições ao seu uso e os problemas semânticos que causa a sua análise são indicativos de que um estudo meramente quantitativo ou estatístico da frequência de certos lexemas não é suficiente para que se tirem conclusões quanto ao seu uso e valor linguístico e terminológico. Um outro problema que se põe prende-se com as diferenças entre *corpus*, que tornam a afirmação "quanto mais frequente maior a possibilidade de ser um termo" verdadeira nuns casos e falsa noutros. Nestes, os lexemas, apesar de recorrentes, podem pertencer ao léxico geral, o que obriga à construção de modelos de extracção que atendam às características particulares de cada *corpus*.

Uma outra metodologia de extracção de termos elabora a análise do *corpus* a partir de artigos, verbos de definição, marcadores de reformulação, reformuladores discursivos, abreviaturas e outras expressões de busca como: *o, a, ou seja, i.e., assim, corresponde, descreve, significa, este processo, este método*, *conhecido como*, *denominado*, etc. Este meio "manual" corresponde normalmente a um primeiro momento, que precede a marcação automática do *corpus*, recorrendo a aplicações informáticas, dado que a pesquisa terminológica que aborda os modelos da formação de termos em *corpora*, fá-lo muitas vezes no sentido de conceber processos de marcação dos *corpora* em geral e dos termos em especial. Este tipo de actividade recorre, normalmente, à identificação de sequências gramaticais como:

-det+adj+nome
+det+verbo+nome+nome(s)

cuja existência no texto indicaria a presença de termos, ou, para reconhecer definições, à identificação de sequências sintácticas, como:

> *X = Y + distinguishing characteristic, whereby X is subordinate to Y*
> *Y + distinguishing characteristic = X, whereby X is subordinate to Y*
> (Pearson, 1998: 136,137)

Um outro método, recorrente em termos de pesquisa lexicográfica, é a análise do *corpus* a partir da concordância de termos e da observação do(s) contexto(s) que os envolvem, método que, embora dependente da dimensão do *corpus* de especialidade e da frequência dos termos, é cada vez mais usual na pesquisa terminológica, por fornecer, para além de dados sobre o significado, informação quanto ao uso do termo e à presença de termos relacionados.

Naturalmente que qualquer das práticas aqui referenciadas, de análise estatística, morfo-sintáctica ou outra, pode ser desenvolvida em conjunto, ou em complementaridade com outro tipo de análises ou abordagens linguísticas, para que o processo de manuseamento dos dados e de extracção de termos a partir de *corpora* se torne o mais eficiente, profícuo e completo possível, garantindo a prossecução dos objectivos do terminólogo.

4. CONCLUSÃO

Muito mais haveria a dizer ainda sobre este tema, e com um maior grau de profundidade, já que a actividade terminológica e todos os processos que com ela se prendem nos colocam outras questões que, dado o âmbito deste artigo, não nos foi possível abranger. Nesta nossa contribuição, procuramos descrever a problemática complexa que envolve a construção de *corpora* e as potencialidades que encerra para a pesquisa linguística nas mais diversas vertentes. Para além de representarem uma fonte inestimável de informação, são uma ferramenta com um sem número de aplicações, quase tantas quantas as necessidades de pesquisa linguística ou de elaboração de documentos (dicionários, glossários terminológicos, etc.), ainda que se reconheçam, naturalmente, limitações.

Os *corpora* textuais específicos, construídos a partir de excertos de línguas de especialidade, são peças centrais para o processo terminológico de extracção e normalização de termos, processo que se tornou uma necessidade imperiosa

para impedir que o desentendimento linguístico tolde a comunicação entre grupos com saberes específicos, e entre estes e a sociedade em geral. Aqueles *corpora*, graças às características lexicais, semântico-cognitivas e linguísticas dos textos que os compõem, bem como aos contextos comunicativos que orientaram a sua produção, contêm muitos dos factores que potenciam a actividade de construção e estabelecimento de terminologias, apesar de não existir ainda um consenso quanto à metodologia mais eficaz para a sua análise, elegibilidade respectiva e extracção dos termos e dos conceitos adjacentes.

Daí que seja necessário, a nosso ver, proceder-se à solidificação e sedimentação da argumentação teórica da Linguística de Corpus, bem como ao estabelecimento de metodologias e princípios comuns e de aceitação generalizada pela comunidade científica, ao mesmo tempo que seria importante que as perspectivas teóricas e metodológicas actuais, que divergem nos seus pontos de vista teórico-práticos, convergissem no sentido da complementaridade, de forma a permitir que introspecção e empirismo refundam sinergias, consagrando a adopção de uma perspectiva holística quanto às potencialidades que cada uma encerra e disponibiliza para a verificação de hipóteses e resultados da outra.

Todo este processo complexo e faseado só pode ser fruto de um trabalho rigoroso e altamente especializado, que tem de ser levado a cabo por equipas interdisciplinares, compostas por terminólogos, técnicos de informática e peritos dos vários domínios, o que torna previsível, de futuro, a disseminação e a diversificação do uso e dos utilizadores de *corpora*, sejam eles lexicógrafos, linguistas, professores de língua, tradutores ou cientistas interessados no desenvolvimento da inteligência artificial e no *diálogo* Homem-Máquina.

Uma última nota para as terminologias e *corpora* existentes em língua portuguesa, que são, a nosso ver, insuficientes e de difícil acesso. O *Corpus* do Cetempúblico surge como a melhor amostra de trabalho inter-disciplinar e inter-organizações, cujo exemplo deve ser reforçado e seguido por outras instituições, tais como editoras, que se dediquem à produção de dicionários, de forma a que mais produtos desta natureza sejam disponibilizados a todos os que se interessam pelo estudo da língua. Finalizamos o nosso artigo com as palavras de Mira Mateus, que já em 1989 afirmava na obra *O Português: Caminhos da Investigação*:

Pela sua própria natureza, a terminologia é uma obra aberta e sempre actualizável, acompanhando dialecticamente o pulsar das novas teorias e das novíssimas aplicações.

BIBLIOGRAFIA

AIJMER, Karin e Altenberg, Bengt (1996) *English Corpus Linguistics*. London: Longman.

BERNARDO, Ana Maria (1996) "A criação de neologismos científicos e técnicos. Algumas considerações metodológicas". in *Runa 26*. Porto, FLUP.

CABRÉ, M. Teresa (1998) *Terminology – Theory, methods and apllications*. Amsterdam/Philadelphia: John Benjamins.

CALZOLARI, Nicoletta (1995) "Observação e Generalização: Análise Linguística de Verbos Declarativos Italianos com Base em *Corpora* Linguísticos". *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (Vol. I)*. Lisboa: APL.

DASH, N. S. e Chaudhuri, B. B. (2001) "The Process of Designing a Multidisciplinary *Corpus*". *International Journal of Corpus Linguistics*, Vol. 5, Nº 2. Amsterdam/Philadelphia: John Benjamins.

HALIDAY, M.A.K. (1996) "*Corpus* Studies and Probabilistic Grammar". *English Corpus Linguistics*. London: Longman.

KJELLMER, Göran (2001) "No Work Will Spoil a Child". *International Journal of Corpus Linguistics*, Vol. 5, Nº 2. Amsterdam/Philadelphia: John Benjamins.

LEITNER, Gerhard (2001) "Lexical Frequencies in *Corpus* of Australian Newspapers". *International Journal of Corpus Linguistics*, Vol. 5, Nº 2. Amsterdam/ Philadelphia: John Benjamins.

LINO, Maria Teresa Rijo da Fonseca (1994) "Neologia, Terminologia e Novas Tecnologias da Informação". *Actas do Congresso Internacional sobre o Português*. Vol. I. Lisboa: Universidade de Lisboa.

MATEUS, Maria Helena Mira (1994) "O Português: Caminhos da Investigação". in *Actas do Congresso Internacional sobre o Português*. Vol. I. Lisboa: Universidade de Lisboa.

_____ (1989) "A Criação de Terminologias: Algumas Reflexões". in *Boletim CNALP*. Lisboa: Presidência do Conselho de Ministros.

MCENERY, Tony e Wilson, Andrew (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

PEARSON, Jennifer (1995) *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

POINTER Project (1999) *PointerFinal Report*. European Community.

REY, Alain (1995*) Essays on Terminology*. Amsterdam/Philadelphia: John Benjamins.

SAGER, Juan C. (1990) *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

SARDINHA, Tony Berber (2000) "Lingüística de *Corpus*: Histórico e Problemática". *DELTA – Revista de Documentação de Estudos em Lingüística Teórica Aplicada*, Vol. 16 – Nº 2. São Paulo: ABRALIN.

SINCLAIR, John (1995) "Tipologia Textual EAGLES". *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (Vol. I)*. Lisboa: APL.

WRIGHT, Sue Ellen e Budin, Gerhard (1997) *Handbook of Terminology Management – Volume 1: Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins.