View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Sapientia

Universidade do Algarve

Faculdade de Ciências Humanas e Sociais

Concepção de *Software* para Constituição e Gestão Semi-automática de *Corpora* de Especialidade

Mestrado em Linguística: Especialização em Terminologia

Carlos Alberto Mascarenhas Romualdo

FARO

2007

Universidade do Algarve

Faculdade de Ciências Humanas e Sociais

Concepção de *Software* para Constituição e Gestão Semi-automática de *Corpora* de Especialidade

Mestrado em Linguística: Especialização em Terminologia

Carlos Alberto Mascarenhas Romualdo

FARO

2007

NOME: Carlos Alberto Mascarenhas Romualdo

DEPARTAMENTO: Faculdade de Ciências Humanas e Sociais

ORIENTADOR: Professor Doutor Manuel Célio Conceição

DATA: 28 de Janeiro de 2008

TÍTULO DA DISSERTAÇÃO: Concepção de *Software* para Constituição e Gestão Semi-automática de *Corpora* de Especialidade

JÚRI:
Doutor Manuel Célio de Jesus da Conceição, Professor Associado da Faculdade de Ciências Humanas e Sociais da Universidade do Algarve;

Doutor Jorge Manuel Evangelista Baptista, Professor Associado da Faculdade de Ciências Humanas e Sociais da Universidade do Algarve;

Doutora Maria Rute Vilhena Costa, Professora Auxiliar da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa;

Doutora Zaida Maria Correia Lopes Pereira, Professora Auxiliar da Faculdade de Ciências Humanas e Sociais da Universidade do Algarve.

Resumo

Apresentamos um protótipo de *software* designado *e-Termite*, cujos principais objectivos passam pela constituição e pela gestão de *corpora* de especialidade. Estabelecendo-se uma separação de duas fases metodológicas distintas, ainda que ambas se complementem mutuamente, procuramos articular os distintos processos e funções que as constituem. A proposta de concepção surge no decurso da verificação de um défice no número de programas informáticos que auxiliem o terminólogo na constituição de um *corpus* de especialidade. Para tal, num primeiro momento, procuramos definir o enquadramento epistemológico e metodológico da Terminologia, abordando não só as alterações que nela ocorreram, nomeadamente, a influência que o seu carácter interdisciplinar teve na redefinição das práticas e concepções próprias da ciência terminológica, mas também o papel fundamental que tem vindo a desempenhar na definição de novas fronteiras do conhecimento especializado. Mostramos a importância do *corpus* e do texto na aquisição das unidades terminológicas, no quadro da Terminologia Textual, e a importância da definição de objectivos e critérios na constituição de um *corpus* adequado. Apresentamos um caso prático de funcionamento da aplicação, ainda que seja apenas uma representação do processo, dado que o estatuto prototípico do *software* não permite verificar de forma real o resultado dos procedimentos. Contudo, é possível compreender, a partir de *screenshots* e diagramas, o mecanismo de aplicação de critérios na constituição do *corpus*, observando-se as implicações da sua escolha. O objectivo principal do *software* é construir uma aplicação que permita a constituição de um *corpus* de especialidade de forma rápida e válida.

Abstract

The software prototype called *e-Termite* presented in this thesis focuses on building and managing *corpora* for specialized purposes. We establish two different phases in the compilation of *corpora*. Although these complement each other, we try to articulate their different processes and functions. The decision to create the prototype was made, when it was understood that there are not many software programs that help terminologists working on *corpora* building. Chapter 1 sets out the objectives of this thesis. Chapter 2 and 3 present the changes that have occured in Terminology and set the conceptual background for the *software* design. They also refer to the importance of *corpora* and text in the understanding of terminological units. The following chapter, chapter 4, describes *e-Termite* and its application through the presentation of a use case, showing the importance of the defining criteria. In the last chapter, we summarize the limitations of the program and foresee possible applications.

Conteúdo

Capítulo 1

Introdução

A Terminologia, considerada no âmbito dos estudos em ciências da linguagem, é uma ciência moderna que se tem vindo a afirmar definindo fronteiras epistemológicas e metodológicas próprias. O seu contributo actual para estudo das línguas de especialidade é o reflexo do processo de crescimento rápido que tem vindo a atravessar, tendo os contextos social e científico criado as condições ideais para que a evolução dos estudos terminológicos, mais do que uma necessidade, se transformasse numa inevitabilidade.

O nosso estudo pretende contribuir para o aprofundamento da metodologia de investigação em Terminologia e, simultaneamente, fornecer um instrumento informático de meta-análise epistemológica que permita ao investigador, conscientemente, efectuar escolhas e uma análise crítica dos procedimentos. Para tal, ao longo dos capítulos segundo e terceiro, apresentamos uma clara definição das bases teóricas que fundamentam a concepção do projecto informático e terminológico *e-Termite*, expondo os princípios que definem a Terminologia Textual e a Linguística de *Corpus*.

Para melhor compreender as abordagens teóricas múltiplas que têm caracterizado a Terminologia, procede-se a uma breve exposição sobre os fundamentos teóricos iniciais que emergiram da escola de Leste e os aspectos conceptuais e contextuais que espoletaram as reacções de contestação por várias perspectivas posteriores. Abordam-se, ainda, os fundamentos apresentados por cada uma das teorias e que

caracterizaram a multiplicidade de interacções com outras áreas do conhecimento. Considerou-se, também, essencial saber de que forma a noção de texto se veio a alterar e por que razões se transformou no centro da análise linguística nos nossos dias, precipitando o início dos estudos em Terminologia Textual.

Procura-se, de seguida, compreender a importância da Informática nas mudanças que afectaram os estudos terminológicos e, consequentemente, a concepção teórica da Terminologia. Tendo em consideração o progresso observado e previsto na área das tecnologias de informação, procede-se a uma tentativa de antever caminhos a partilhar pela Terminologia e Informática, nomeadamente no trabalho conjunto em áreas como a análise linguística e a transmissão de informação. Esta parceria pode trazer resultados consideráveis e deve ser encarada com muita seriedade, como um vector prioritário de investimento em recursos humanos, tecnológicos e intelectuais.

No terceiro capítulo, aborda-se a génese da Linguística de *Corpus* e as ligações que estabelece com a Terminologia, sendo essencial proceder a uma separação de duas fases metodológicas distintas, ainda que trabalhem sobre o mesmo objecto, o *corpus*. A fase da constituição está largamente dependente dos critérios definidos pelo investigador, inclusive o domínio e os objectivos, que determinam as características resultantes do produto final, enquanto a fase da gestão incide sobre o registo e a análise de informação sobre o *corpus* e sobre a actualização dos textos nele existentes. Assim, procura-se esclarecer os diferentes critérios e a significativa mudança que poderá surgir nos resultados devido às decisões que o investigador toma, principalmente no que diz respeito à recolha e anotação de informação, influenciando a longevidade do *corpus*.

Depois de estabelecida a base conceptual do projecto, apresentam-se os moldes em que a concepção da aplicação foi estruturada e quais as funções delineadas para cada um dos módulos. Através do recurso a um exemplo de uso, procede-se à descrição detalhada de todo o percurso do investigador, sendo possível, através de quadros e diagramas informativos, compreender os passos na perspectiva do utilizador e os resultados que se pretendem em cada fase.

Os objectivos do projecto *e-Termite* passam por conceber o protótipo de uma aplicação informática que consiga optimizar os processos de constituição e gestão de *corpora* de especialidade e, ao mesmo tempo, construir uma lista de termos anotados, de agora em diante, designada por dicionário, que possa representar o conhecimento da especialidade em estudo. Apresentam-se ainda, brevemente, as alternativas informáticas nacionais disponíveis para o mesmo tipo de tarefas e com objectivos próximos da nossa proposta.

Na conclusão, tentaremos abordar todas as propostas apresentadas na introdução, fazendo um sumário das ideias mais importantes discutidas ao longo da dissertação, dos problemas identificados, das limitações ao desenvolvimento do trabalho e das potencialidades que o projecto apresenta.

A bibliografia escolhida é representativa das áreas do conhecimento referidas na dissertação, tendo sido usada recorrentemente para complementar e justificar todas as opções metodológicas e escolhas teóricas efectuadas ao longo do discurso. Encontram-se também listadas obras não citadas no trabalho, mas que desempenharam um papel fundamental na formação científica sobre os mais diversos assuntos que correm transversalmente à Terminologia.

Capítulo 2

Terminologia textual

2.1 Introdução

Este capítulo começa por abordar a génese, o crescimento e a consolidação da Terminologia, enquanto disciplina ou teoria, cuja autonomia epistemológica ainda procura um consenso alargado por parte da comunidade científica. Segundo Cabré:

> «*Not all experts agree that terminology constitutes a separate discipline, nor do all consider it a theoretical subject.*» (Cabré, 1999:6)

Após o aparecimento formal da Terminologia na segunda metade do séc XX, observou-se uma certa estabilidade teórica. Nestes últimos trinta anos, desencadearam-se numerosas discussões à volta dos seus princípios e, consequentemente, foram surgindo alternativas ao nível das fundamentações teóricas.

> «*It is surprising that after many years of inactivity in terminological theory all of a sudden there has been a rush of critiques of established principles and suggestions proposing new alternatives to the traditional theory.*» (Cabré, 2003:163)

Estas alternativas provocaram modificações nos processos metodológicos, que foram acompanhando as divergentes correntes teóricas, e que importa conhecer para melhor se compreender as formas de trabalhar em Terminologia actualmente.

Para um enquadramento mais claro dessas alterações, referem-se sucintamente os mais importantes movimentos teóricos e a sua participação no desenvolvimento da Terminologia, enquanto ciência. Desde o movimento iniciado por Wüster, denominado Teoria Geral da Terminologia (TGT), passando pela Socioterminologia de Gaudin (1993), Teoria Comunicativa de Cabré (1999) e a abordagem Sociocognitiva de Temmerman (2000), apresenta-se uma pequena súmula dos principais vectores que orientam cada um destes movimentos e da influência que exerceram nas práticas terminológicas.

Sendo uma das principais características da Terminologia a sua interdisciplinaridade, este capítulo refere, ainda, a importância de alterações decorridas noutras áreas, destacando-se o desenvolvimento da Informática, que influenciaram directamente o percurso da Terminologia, como factor preponderante para compreender a revolução nas práticas terminológicas. L'Homme destaca até que ponto a informatização foi preponderante nesta revolução:

> «*What might appear as normal and standard in computational circles has had profound consequences for terminologists; this has led many to criticize traditional theoretical principles and some to propose new approaches [...] methods and practices have changed drastically due mostly to the extensive use of electronic corpora and computer applications.*» (L'Homme, 1998)

Não foi, no entanto, apenas a evolução da Informática a causar o rompimento tão abrupto com os princípios epistemológicos iniciais da Terminologia. A Linguística, que, com o surgir de novas perspectivas teóricas, atravessou uma fase de restruturação e redefinição conceptual, veio facilitar o aparecimento da Terminologia Textual. Tal como se pode ler na afirmação de Lino:

«*Nestes últimos anos, assistimos a uma rápida evolução da ciência terminológica, traduzida por uma definição de novos suportes teóricos e por uma abertura a novas perspectivas; relativamente a este novos modelos, destacamos a integração da pragmática, a perspectiva da socioterminologia e as metodologias em terminologia textual, o tratamento automático de corpora de especialidade.*» (Lino, 2000:26)

2.2 Teorias da Terminologia

Os estudos de carácter terminológico sempre se confundiram com os de natureza linguística e é apenas no século XX, com o confluir de diversos factores, que a sua importância e singularidade são reconhecidas. Auger (1988) chama-lhe a fase moderna da Terminologia, destacando-se, como principais motivos, o avançar da indústria e da ciência, a produção em massa, a estandardização de produtos e o aparecimento de uma sociedade onde a informação e a transmissão de conhecimento são fundamentais. Estes factores precipitam o aparecimento de novos conceitos e a necessidade acrescida de criar nomes para esses conceitos. A Terminologia começa, assim, por ser uma actividade desenvolvida pelos cientistas e especialistas, não por linguistas, no decurso do seu trabalho de investigação, à qual não era dado o devido crédito ou importância em termos formais ou sequer linguísticos, como aponta Cabré:

«*During the first half of the 20th century neither linguists nor social scientists paid special attention to terminology [...] It's no coincidence that the development of both theoretical and applied terminology in the second third of the 20th century occurred thanks to the interest of scientists and technicians.*» (Cabré, 1999:2)

Os estudos formais surgem quando Wüster, um engenheiro e professor universitário austríaco, decide apresentar a Terminologia como uma disciplina autónoma,

por acreditar que as línguas de especialidade, ou seja, todas aquelas que são usadas em áreas específicas do conhecimento, detêm vocabulário e estruturas com uso linguístico específico. Devem, por isso, ter metodologias de análise próprias e práticas diferentes das usadas para trabalhar as línguas gerais, como se pode atestar nesta afirmação de Wüster:

> «*Es wird angenommen dass sich die meisten der Leser des gegenständlichen Werkes dem Studium eines Zweiges der Sprachenwissenschaft gewidmet haben, genauer: einem Ausschnitt aus der Wissenschaft von der Gemeinsprache.*» (Wüster, 1985:1)

Esta decisão de autonomizar a Terminologia surge no sentido de a dotar de uma fundamentação teórica independente, tendo como objectivo principal atingir, dentro de áreas específicas do conhecimento, uma univocidade absoluta dos termos para que as comunicações dentro dessas áreas se pudessem efectuar de forma objectiva e inequívoca. Cabré apresenta resumidamente quais os objectivos de Wüster:

> «*It is fair to say that all Wüster's life was devoted to terminology. With his work he pursued a number of objectives, intended:*
>
> 1. *To eliminate ambiguity from technical languages by means of standardisation of terminology in order to make them efficient tools of communication.*
> 2. *To convince all users of technical languages of the benefits of standardised terminology.*
> 3. *To establish terminology as a discipline for all practical purposes and to give it the status of a science.*»
>
> (Cabré, 2003:165).

Entre os anos 30 e 60, Wüster publica uma série de trabalhos na área da Terminologia que acabam por culminar num dicionário, *The Machine Tool*, onde põe

em prática todas as suas ideias sobre o trabalho com terminologias. Ao conjunto de pressupostos por ele convencionados e seguidos pela Escola de Viena para tratamento do vocabulário especializado, dá-se o nome de Teoria Geral da Terminologia (TGT).

Até finais da primeira metade do século, os linguistas continuaram a não dar muito valor aos estudos ligados à Terminologia, considerando os termos unidades fixas e prescritivas sem interesse para o estudo das línguas naturais, e deixaram a cargo dos especialistas a construção das respectivas terminologias e critérios da sua elaboração. A Terminologia passou, então, por uma fase de estagnação ao nível das suas fundamentações teóricas, sem grandes contestações aos seus métodos de trabalho.

Entretanto, com o desenvolvimento dos meios de comunicação, a vulgarização e a fácil circulação de terminologias e de conhecimentos técnicos de especialidade fora das respectivas áreas tornam-se frequentes. Cada vez mais, o volume de informação especializada ao dispor de não-especialistas aumenta e chega a um número maior de falantes, notando-se um crescente cruzamento entre a língua geral e as línguas de especialidade. É cada vez mais habitual encontrarem-se termos técnicos integrados na língua geral e vice-versa.

A aproximação da língua de especialidade à língua geral e o seu reconhecimento, enquanto método de circulação de conhecimento, aumentam o grau de contacto com outras disciplinas ligadas ao estudo das línguas naturais e à aquisição e transmissão de conhecimentos, nomeadamente às ciências cognitivas, às ciências da comunicação e à Linguística, que se começaram a interessar pelo fenómeno terminológico. Todas estas áreas trouxeram perspectivas novas e provocaram uma dinâmica de instabilidade conceptual, conduzindo eventualmente a que as teorias tradicionais fossem questionadas e repensadas, principalmente a Teoria Geral da Terminologia, ainda muito centrada na estandardização de conceitos e termos, e surgissem novas propostas epistemológicas alternativas para a disciplina da Terminologia.

No âmbito das ciências cognitivas, afirmou-se ser necessário compreender os modelos cognitivos de aquisição da língua geral para chegar aos processos de formali-

zação do conhecimento especializado. Logo, separar os mecanismos de compreensão e estruturação do conhecimento especializado do conhecimento geral e tentar isolar as línguas de especialidade tornam-se tarefas muito complexas, senão impossíveis.

No âmbito das ciências da comunicação, por sua vez, encontra-se, nas línguas de especialidade, um campo de estudo muito importante para analisar as várias formas de apresentação e divulgação de conhecimento técnico e sua forma de disseminação em diferentes níveis de especialização, constatando-se que as línguas de especialidade, em contextos comunicativos diversos, apresentam alterações lexicais importantes e adaptações ao nível da estrutura sintáctica e textual, revelando uma flexibilidade típica da língua geral.

Em resultado desta integração de termos das línguas de especialidade na língua em geral e de palavras polissémicas e não especializadas nas terminologias, dá-se uma maior e inevitável aproximação entre a Linguística, os seus estudiosos e a Terminologia, como demonstra Rey:

> «*Comme la linguistique, la sémantique ou la sémiotique, la terminologie étudie des signes. Ces signes se manifestant au moyen des formes des langues naturelles (mots, etc.), leur rapport avec ces formes doit être précisé.*» (Rey, 1979:18-19)

Devido a esta interacção mais frequente e intensa, a Terminologia acaba por assimilar muitos dos métodos que a Linguística vai desenvolvendo e aplicando nos seus próprios estudos, afirma Cabré:

> «*The general scientific study of terminology is largely influenced by its relationship to applied linguistics, of which it is a branch.*» (Cabré, 1999:25)

Ao questionarem a divisão rígida entre a língua geral e de especialidade, como advogara Wüster, alguns linguistas consideram as unidades lexicais de especialidade

como portadoras de um significado específico dentro de um contexto técnico e especializado e não unidades lexicais com existência autónoma e descontextualizada da língua, funcionando, assim, como qualquer outra unidade lexical das línguas naturais, como se comprova na afirmação de Rey,

> «*Empiriquement, la définition terminologique est bien différent. [...] Comme elle est formée dans uns langue naturelle, elle véhicule toutes les ambiguïtés, polysémies, connotations (1), etc., des unités-mots de cette langue ;*» (Rey, 1979:43).

Dada a proximidade crescente entre a Linguística e a Terminologia, as profundas alterações verificadas na primeira, com o aparecimento das correntes funcionalista e discursiva, vão influenciar a forma como a Terminologia, estável e imune a influências até então, era concebida e provocar uma quebra epistemológica dentro da disciplina.

A Sociolinguística, que tem como um dos princípios fundamentais a valorização da análise do contexto social de produção das unidades lexicais, vai ser a primeira teoria a contestar os princípios da Terminologia nas suas práticas e concepções teóricas. Este movimento defende a importância do reconhecimento da variação terminológica e da sinonímia e polissemia nas línguas de especialidade e contextos especializados, todos eles recusados na Teoria Geral da Terminologia, e que

> «*c'est socialement que la référence des termes peut être construite de façon relativement étroite, par une action volontaire et concertée, donc toujours provisoire*»(Gaudin, 2003:46).

Com François Gaudin (1993), os estudos terminológicos são novamente transportados para o domínio do uso real da língua. Este novo modelo teórico ganha o nome de Socioterminologia, cuja utilização surge pela primeira vez com Boulanger (1981), e defende uma postura descritiva e de análise dos termos numa perspectiva

socioprofissional, tendo em conta os diferentes níveis de especialização. A Socioterminologia vem ainda questionar a possibilidade de existirem áreas de conhecimento estanques, introduzindo o conceito de «*noeuds de connaissances*» por sua vez ligados à ideia de que uma área de conhecimento é sempre o resultado de interacções interdisciplinares dos diversos «nós». Neste mesmo sentido, para que a perspectiva evolutiva, dinâmica e interdisciplinar da língua de especialidade possa ser considerada, recusa-se a visão sincrónica da teoria tradicional, considerada redutora e inflexível, e dá-se preponderância a análises diacrónicas.

Com Teresa Cabré (1999) desenvolve-se uma outra teoria, apelidada de Teoria Comunicativa, que, tal como a Socioterminologia, defende a variação terminológica, a interdisciplinaridade da Terminologia e a importância de usar as línguas reais como objecto de estudo, afirmando que

> «*Oral and written technical and scientific communication is the basic source material for extracting terms*» (Cabré, 1999:121).

A univocidade dos termos, segundo Cabré, só poderia ser atingida por um processo artificial e utópico que dificilmente conseguiria atingir o seu objectivo de unificação dos termos e conceitos. Além disso, as unidades terminológicas remetem para conceitos tecnológicos em constante evolução que reflectem uma sociedade permanentemente dinâmica, nunca podendo, por isso, ser estáticos. A unidade terminológica, objecto de estudo da Terminologia na Teoria Comunicativa, encerra em si as propriedades diversas resultantes dos pontos de vista variados que pode assumir, sendo um resultado da interdisciplinaridade da Terminologia e da hiperespecialização, também defendidas na Socioterminologia. Como demonstra Cabré,

> «*the ordering of thought and the conceptualization represent the cognitive side of terminology, the transfer of knowledge constitutes its communicative side*» (Cabré, 1999:45).

Nesta perspectiva, o ponto de vista que interessa ao terminólogo, por ser o que dita o uso da unidade terminológica em situações específicas, é o comunicativo,

tornando-se importante dar valor às condições de produção, de transmissão e de recepção. O campo da pragmática na língua surge, pois, em destaque, o que também já era observável na teoria da Socioterminologia.

Rita Temmerman (2000), partindo de muitos dos princípios já defendidos pela Socioterminologia e pela Teoria Comunicativa, apresenta a sua abordagem sociocognitiva e introduz a ideia de «*unit of understanding*» por oposição à de conceito defendida por Wüster. A unidade de compreensão ou percepção formaliza a importância da cognição e sua estruturação na aprendizagem e transmissão de conhecimento especializado. Para Temmerman, a ideia de conceito na perspectiva tradicional apresenta-se demasiado restritiva e rígida, raramente podendo ser um dado conceito apontado como pertencente a uma única e delimitada categoria. A conceptualização do mundo por parte do Homem é apenas uma visão da realidade, por isso, não pode ser considerada como objectiva e final, apesar de ser concebida como tal na teoria tradicional. Temmerman considera que a

> «*Modern Terminology could incorporate the idea that humans do not just perceive the objective world but have the faculty to create categories in mind*» (Temmerman, 2000:61).

A teoria tradicional é, assim, posta em causa, essencialmente, por se concentrar num objectivo absoluto de normalização, ignorando a observação e estudo da língua como necessidades fundamentais para poder descrevê-la. Segundo a abordagem sociocognitiva, o Homem capta o mundo a partir de modelos cognitivos idealizados (I.C.M.'s) nos quais as tais «*units of understanding*» se integram e se relacionam umas com as outras, sendo a compreensão e a aprendizagem apenas uma questão de organização e estruturação desses modelos cognitivos construídos pelo cérebro, processo esse que decorre ao longo do tempo, como se pode ler na afirmação de Temmerman,

> «*Sociocognitive Terminology believes understanding amounts to categorisation. Each category is understood as existing within cognitive models. Understanding is a structured event*» (Temmerman, 2000:225).

Esta forma de argumentação permite defender a funcionalidade não só da sinonímia e da polissemia, tal como se tinha apresentado na Socioterminologia e na Teoria Comunicativa, mas também todas as relações semânticas e ontológicas que se estabelecem entre os elementos da área de conhecimento, pois estes processos participam no desenvolvimento da compreensão e devem, por isso, ser descritos.

Um dos pontos comuns mais importantes a salientar nos três movimentos apresentados e que será amplamente abordado no próximo capítulo, é precisamente o facto de todos eles defenderem a preponderância da recolha e do uso de exemplos de língua real como objecto de estudo. A nossa concepção de software projecta uma análise no quadro do uso de produtos reais de língua, enquadrando-a, nesta perspectiva, com os objectivos de proceder à recolha das unidades que representam o conhecimento terminológico e de descrevê-las no contexto da especialidade.

2.3 A unidade terminológica

A unidade terminológica, vulgo termo, é o objecto de estudo da Terminologia. O termo tem vindo a sofrer alterações, como se pôde observar anteriormente, no quadro das reformulações epistemológicas que a ciência tem atravessado. O termo começou por ser, segundo Conceição,

> «*une dénomination qui étiquette un concept pré-existant, et il [avait] donc une statut proche de celui des unités des nomenclatures et des thesauri*» (Conceição, 2005:45).

A mudança de paradigma teórico da Terminologia trouxe alterações profundas ao conceito de termo. Assim, a unidade terminológica passa a apresentar diferentes

particularidades, devendo todas elas ser consideradas como traços característicos presentes. Assim, Conceição (2005) define o termo como uma unidade que apresenta uma complexidade conceptual, sendo simultaneamente uma:

- Unidade lexical
- Unidade de cognição e significação
- Unidade de referência
- Unidade de denominação
- Unidade de representação
- Unidade de conhecimento

Os termos pertencem a um conjunto conceptual e estão integrados em contextos múltiplos, tendo de ser analisados como pertencentes a um esquema complexo e organizado de informações. Como Bourigault, Jacquemin e L'Homme referem,

> «*terminological units can be further analyzed and organized into sophisticated networks that reflect the knowledge structure of a specialized field*» (Bourigault et al., 2001:VIII).

Para o nosso estudo, é importante compreender que o termo é uma unidade lexical de especialidade e, por isso, denota todas as características próprias das unidades lexicais. No entanto, possui, também, um sentido específico de ligação a um domínio do conhecimento e transporta, em si, informação múltipla e dinâmica, sendo fundamental aferir, no âmbito da concepção defendida na dissertação, o nível de proximidade que o termo apresenta em relação a esse domínio, dado que é possível conceber diferentes graus de representação do conhecimento de especialidade. Como Kageura e Umino referem,

> «*termhood refers to the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts*» (Kageura & Umino, 1996:11).

A questão de estabelecer o nível de proximidade não é, no entanto, tão simples quanto poderia parecer, como explicam os mesmos autores, mais adiante:

> «*most take a pragmatic standpoint, simply admitting noise or leaving the final decision to human evaluation*» (Kageura & Umino, 1996:11).

Este último ponto é a base da concepção de intervenção do investigador que defendemos e que nos leva a construir a aplicação informática numa perspectiva semi-automática, recaindo sempre a responsabilidade de proceder à classificação terminológica no terminólogo ou no especialista.

O estudo da unidade terminológica não pode ser dissociado do estudo da unidade lexical e do seu envolvente textual, sendo necessário compreender a evolução de cada um destes elementos no quadro da Linguística para compreender as alterações que se foram verificando e que passamos a apresentar.

2.4 A Linguística e o texto

Durante a fase inicial do estruturalismo, a busca por uma sistematização descritiva da língua, reforça o já importante papel da unidade lexical como ponto de referência na investigação linguística. A língua era vista como uma espécie de rede, na qual as palavras desempenhavam o papel de nós de ligação e a base de identificação e diferenciação residia na sua aparência. Uma unidade lexical considerava-se única por, na sua constituição, ser diferente de todas as outras existentes no léxico. Maher e Groves explicam que

> «*structuralists did not concern themselves too much with syntax. They were interested rather in making a detailed but compact taxonomy of all*

the elements which can be extracted from a corpus of data» (Maher & Groves, 1996:71).

Sob a influência de Chomsky e do generativismo, observa-se uma transição do eixo de análise da unidade lexical para a frase. Assim, os fenómenos sintácticos ganham cada vez mais importância, dado que, segundo Halliday e Teubert,

«*Chomsky's interest in the lexicon is, contrary to structuralists, only marginal*» (Halliday & Teubert, 2004:82).

As unidades lexicais, ainda que detentoras de um significado próprio e isolável do resto da comunicação, enquanto peças de um sistema linguístico, quando em conjunto, passam a ser interpretadas e contextualizadas numa dimensão mais ampla trazida pelas relações sintácticas, que lhes anexa informação sintáctica e semântica específica e decisiva para a sua correcta compreensão. Quando as unidades lexicais estão interligadas e dependentes de outros elementos frásicos, o seu significado é sempre resultado de variáveis exteriores a elas e procedentes da inserção em contexto da frase a que pertencem. A sintaxe passa a dominar a análise linguística e atribui à frase a função central de núcleo na estrutura das línguas naturais que, segundo Cook e Newson,

«*relies on the structural relationships in the sentence rather than on the sequence of words.*» (Cook & Newson, 1996:4)

A insatisfação de algumas correntes de investigação perante a proposta generativista está presente no discurso de Bakhtine, quando este afirma não ser ainda na frase que se encontra o nível de análise mais adequado ao trabalho de investigação nas línguas, como se pode confirmar no seguinte excerto:

«*La syntaxe des grandes masses verbales [...] attend encore d'être fondée ; jusq'à present, la linguistique n'a pas avancé scientifiquement au de-là de la phrase complexe : c'est le phénomène linguistique le plus long qui ait*

> *été scientifiquement exploré. On dirait que le langage méthodiquement pur de la linguistique s'arrête ici, et que au-delà commence aussitôt la science, la poésie, et ainsi de suite. Et cependant, on peut poursuivre plus loin l'analyse linguistique pure, si difficile que cela paraisse*» (Bakhtine, 1978:59).

Bakhtine refere-se ao nível de profundidade da análise linguística, pois há uma dimensão externa à frase que a sintaxe não consegue resolver. Como afirmam Biber, Conrad e Reppen, era necessário ascender ao nível discursivo:

> «*Discourse analysis focus on language characteristics that extend across clause boundaries. [...] such analyses are important for both descriptive and applied linguistics*» (Biber et al., 1998:106).

Na sua dimensão interna e formal, a gramática da frase não conseguia apresentar explicações para processos como a referência (anáforas e catáforas), a substituição, a elipse, a conjunção e a coesão lexical, tal como descritos por Halliday. Estes fenómenos referenciais não podem ser explicados à luz de blocos frásicos estanques e desligados do restante conjunto de frases que completa o contexto comunicacional, como explicam Halliday e Ruqaiya:

> «*we can interpret cohesion, in practice, as the set of semantic resources for linking a SENTENCE with what has gone before.*» (Halliday & Hassan, 1976:10)

A frase precisa de estar ligada não só com o que vem antes, mas também com o que virá depois e até com o que não é verbalizado, mas contextualiza toda a interacção e é fundamental para que o processo comunicativo se conclua com sucesso.

Assim, por acção da corrente funcionalista, o eixo das análises linguísticas sofre uma nova conversão. A focalização do objecto de estudo linguístico que residia na unidade lexical e na frase passa a ser dividida com o texto, devido às limitações que

as duas primeiras apresentam quando isoladas e desenquadradas do seu envolvente textual. A frase é uma estrutura importante na análise, mas Halliday relembra que

> «*the clause complex has certain inbuilt limitations, from the point of view of its contribution to the texture of a discourse*» (Halliday, 1994:309).

O texto já não é visto apenas como uma necessária soma de múltiplas palavras ou frases independentes, mas também poderá ser uma simples unidade que, ainda não possuindo uma identidade na gramática da língua, detém na sua globalidade um sentido próprio. Halliday e Ruqaiya afirmam que

> «*a text may be spoken or written, prose or verse, dialogue or monologue. It may be anything from a single proverb to a whole play, from a momentary cry for help to an all-day discussion on a committee.[...] It is not a grammatical unit, like a clause or a sentence; and it's not defined by its size [...] A text is a unit of language in use.*» (Halliday & Hassan, 1976:1)

Nasce, então, uma categoria na estruturação de análise linguística que se situa, em termos organizacionais, num patamar superior ao das frases, visto englobá-las como seus constituintes. Van Dijk (1977), ainda num quadro marcadamente estruturalista, chama-lhe superestrutura textual, tentando adaptar esta nova percepção às teorias dominantes. Este aparecimento do texto, enquanto unidade linguística de facto, analisável e categorizável, numa classe superior à frase em abrangência, conduz ao aparecimento de outras correntes linguísticas, fora do formalismo estruturalista, das quais se torna objecto central de estudo. Com este novo objecto identificado, torna-se necessário e fundamental, como em qualquer outra ciência, estabelecer fundamentações teóricas e proceder à sua categorização e caracterização, processos que se revelam complicados, devido a questões epistemológicas suscitadas por diferentes escolas de pensamento. As vertentes textual e discursiva entram numa discussão teórica em torno dos conceitos de discurso e de texto, tentando, através de propostas

consecutivas de diferentes autores, delimitá-los e separá-los, destacando-se, contudo, a proposta de Adam (1990):

> «*DISCOURS = Texte + Conditions de production*
>
> *TEXTE = Discours - Conditions de production*»
>
> (Adam, 1990:23)

Nesta configuração, o texto representa um qualquer registo escrito de uma produção comunicativa, seja ela efectuada numa situação oral (uma entrevista ou um diálogo) ou não (um anúncio escrito ou uma narrativa) e o discurso representa um qualquer evento comunicativo em contexto, ou seja, a sua interpretação. Segundo o critério de Nunan, a separação dos dois conceitos reside na interpretação que se faz do texto:

> «*text analysis and discourse analysis [...] deal with [...] linguistic analysis of texts and an interpretation of those texts*» (Nunan, 1993:7).

A proposta apresentada por Adam, porém, não é consensual e foi necessário aprofundar a diferenciação epistemológica de cada um dos conceitos, para que a distinção entre texto e discurso ficasse clara. Como refere Nunan,

> «*both text and discourse need to be defined in terms of meaning, and that coherent texts/pieces of discourse are those that form a meaningful whole.* »(Nunan, 1993:6).

Bronckart (1996) vai apresentar uma solução para a questão, partindo da sua perspectiva epistemológica de interaccionismo social e sociodiscursiva, na qual o texto é descrito como um produto da intervenção da sociedade na constituição do sujeito, que se reformula constantemente, e, por isso, é difícil de classificar de forma estável. O texto designa toda e qualquer «*unidade de produção verbal que transmita uma mensagem linguisticamente organizada e que produza no seu destinatário um*

efeito de coerência» (Bronckart, 1996:74), sendo considerado como «*l'unitée communicative de rang supérieur*» (Bronckart, 1996:74). Por outro lado, o discurso é apresentado como um segmento do texto, com uma forma fixa e tipificável pela suas características linguísticas estáveis.

Como atesta Coutinho (2003), «*essas unidades [«unidades globais» empiricamente atestadas] são textos – unidades diversas e empíricas de produção verbal oral e escrita, situada, acabada e auto-suficiente, que realizam uma função comunicativa*»(Coutinho, 2003:109), enquanto que «*os tipos de discurso correspondem a diferentes planos de enunciação, identificáveis através de "configurações" de unidades linguísticas*»(Coutinho, 2003:111).

Bronckart concretiza, assim, as fronteiras do texto, tornando-o num alvo objectivo e estudável, como Coutinho nos explica,

> «*os textos são tomados como realidades semióticas complexas que cumprem funções comunicativas concretas. Trata-se, portanto, de objectos empíricos, atestados*»(Coutinho, 2003:101).

Bakhtine alerta também para a multiplicidade de géneros em que os textos se podem enquadrar e os tipos de discurso que os caracterizam, mediante a esfera de utilização, como se pode ler na afirmação seguinte:

> «*Tout énoncée pris isolément est, bien entendu, individuel, mais chaque sphère d'utilization de la langue élabore ses types relativement stables d'énoncés, et c'est que nous appelons les genres du discours*» (Bakhtine, 1984:265).

O texto está, portanto, ligado à noção de géneros textuais que, segundo Coutinho, são «*"formas comunicativas" elaboradas pela actividade de gerações precedentes e sincronicamente disponíveis, em termos de intertexto*» (Coutinho, 2003:109) e, consequentemente, está, também, ligado a práticas sociais. Como Coutinho explica, os textos são

> «*produções linguísticas empíricas e atestadas, que realizam uma função comunicativa e se inserem numa prática social, correspondendo os géneros de texto às formas comunicativas relativamente instáveis (ou relativamente estabilizadas, num determinado período histórico, para uma sociedade ou grupo social), de que qualquer texto participa necessariamente (ainda que por divergência)*» (Coutinho, 2003:118-119)

A incapacidade da Linguística mais introspectiva - resultante das filosofias "chomskianas" e que domina os estudos na área das línguas - em lidar com a variação e multiplicidade da língua e o crescente afastamento da língua real em direcção a uma língua fabricada e moldada às necessidades da teoria fazem com que os linguistas se dividam e voltem a dedicar-se à investigação com amostras reais de língua natural. Jacques afirma que

> «*un parcours de la littérature permet de constater qu'elle [linguística introspectiva] est critiquée sur deux aspects principaux : le premier concerne le peu de fiabilité des jugements de grammaticalité, le second l'impuissance de la linguistique introspective à capter et rendre compte de faiçon satisfaisante de la variation*» (Jacques, 2005:22).

Não há uma reprovação da importância da linguística introspectiva, mas sim uma necessidade de completar as suas valias com as que o estudo baseado em elementos da língua real apresenta, sendo ambas as vertentes importantes, como Halliday e Teubert referem:

> «*The perspective of Chomskyan and cognitive linguistics represents a very different view of language [...] Both views are, of course, legitimate, and they are complementary. Corpus linguistics deals with meaning. Cognitive linguistics is concerned with understanding*» (Halliday & Teubert, 2004:98).

Ao acreditar que as escolhas linguísticas dos falantes não são despropositadas e contêm em si motivações contextuais, é essencial conhecer uma série de factores que determinam o acto linguístico, dos quais se destacam os participantes, objectivos, meios de comunicação, entre outros. Para poder contactar com o objecto de análise e todas as variáveis, ou seja, uma amostra contextualizada, é imprescindível que a Linguística recorra a exemplos reais de língua. Bronckart afirma que

> «*une langue naturelle n'est appréhendable qu'au travers des productions verbales effectives, et celles-ci prennent des allures très divers, notamment parce qu'elles sont articulées à des situations de communication différentes*» (Bronckart, 1996:71)

A linguística aplicada ao texto, vulgo Linguística Textual, surge como um ramo que privilegia o uso de amostras reais, tal como Bronckart aponta:

> «*ce sont ces formes de réalisation empiriques diverses que nous qualifions des textes*» (Bronckart, 1996:71).

No quadro da nossa proposta de concepção de software é essencial compreender a importância da unidade lexical, da frase e, sobretudo, do texto, enquanto unidade linguística emergente, contextualizadora dos dados terminológicos e que habita o *corpus*, como veremos mais adiante, dado que são elementos nucleares da análise terminológica.

O ascendente da Linguística no seio da epistemologia terminológica não é caso único, assistindo-se à continuada e à crescente articulação entre a Terminologia e outras disciplinas, das quais a Informática se destaca pelo relevo que foi conquistando, como se pode constatar de seguida.

2.5 A Informática

A renovada interacção com a Linguística leva também a Terminologia a ter um maior contacto com as novas metodologias de trabalho e com o ascendente da Informática na análise das línguas naturais. Ainda que a partir do séc. XX, com o aparecimento e desenvolvimento da tecnologia informática, a Terminologia, juntamente com outros campos ligados à Linguística, tenha aproveitado os recursos tecnológicos disponibilizados, nomeadamente ao nível da possibilidade de constituição de enormes bases de dados e sua consulta posterior, a tentativa de Wüster para autonomizar a Terminologia levou a que, durante décadas, a importância e o papel da Informática estagnassem e esta se resumisse a servir de catálogo digital para organizar, etiquetar e consultar os termos. Wüster estrutura as funções de um computador da seguinte forma:

> «*Els ordinadors executen dos tipus de funcions per a la documentació i la informació:*
>
> 1. *Un ordinador pot memoritzar grans quantitats d'informació formulada lingüísticament (dades). Aquesta informació pot ser o dades textuals o bé circumstàncies.*
> 2. *Les informacions memoritzades en un ordinador es poden retrobar amb una velocitat impressionant.*»
>
> (Wüster, 1996:194)

A partir de meados dos anos 80, com a revolução epistemológica e a reaproximação à Linguística, a Terminologia avalia e redefine a forma como a Informática deve ser utilizada. O enorme desenvolvimento dos recursos informáticos, que se verificou em poucos anos, permite executar processos e operações que já ultrapassam as tarefas de um simples instrumento de catalogação e consulta. Graças ao aumento da capacidade de armazenamento e da velocidade de processamento, torna-se exequível não só guardar elevadas quantidades de informação linguística e pesquisá-la, mas

também cruzar e relacionar toda essa imensa informação, situações que antes não eram viáveis ou simplesmente levavam mais tempo do que o considerado útil para atingir um resultado válido. Como salienta Kennedy,

> «*the computer [...] has introduced incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data*» (Kennedy, 1998:5).

Os progressos registados permitem entender que não só a Linguística, mas também a Terminologia, terão muito a beneficiar com a continuada interacção e investimento no diálogo com a Informática. A razão que fundamenta o desenvolvimento deste projecto no âmbito da computação e que motivou a apresentação de um protótipo de software está presente na afirmação de Kennedy, pois, no nosso entender, não é produtivo, ainda que seja possível, conceber uma investigação terminológica, no contexto epistemológico actual, que não passe pela utilização de ferramentas automáticas.

A existência de áreas na Informática, onde o conhecimento e a sua disseminação são factores determinantes, tais como a Inteligência Artificial e a *Information Retrieval*, fazem com que o interesse e o benefício seja recíproco e potenciem a disponibilidade e a vontade para trabalhar em projectos conjuntos na procura por resultados imediatos. Há, pois, uma conjuntura epistemológica e tecnológica favorável às metodologias de análise textual que surgem no quadro da Linguística e que vão propagar-se à Terminologia, como se relata de seguida.

2.6 Terminologia textual

A escola de Leste, seguidora do modelo wüsteriano, continuou, no entanto, a considerar que o trabalho terminológico tem como objectivo primordial reconhecer e recolher o vocabulário específico de uma dada especialidade e que tenha como resultado a construção de dicionários ou glossários técnicos, desvalorizando-se a im-

portância do contexto ou da variação semântica. Wüster afirma que, para a teoria tradicional,

> «*La concepció de la terminologia sobre l'estat de la llengua es caracteritza per tres aspects: prioritat i precisió dels conceptes, prioritat del lèxic davant la gramàtica, i prioritat del tractament sincrònic de la llengua*» (Wüster, 1996:159).

As divergências em relação aos métodos e propostas de Wüster para a Terminologia são concretizadas por Alain Rey (1979) que vem introduzir as variáveis contextual e semântica como fundamentais para a realização de uma investigação terminológica eficaz, defendendo que é através da análise do conjunto dos elementos que pertencem ao texto que se identifica e delimita o valor de um termo. Rey declara que

> «*Il s'agit ici de repérer dans un emsemble de textes (ou d'énoncés oraux) l'emsemble des unités linguistiques, mots et syntagmes, utilisés comme termes et d'inférer de l'analyse sémantique du discours la valeur de ces termes*» (Rey, 1979:92).

A Terminologia, acompanhando o percurso da Linguística, procura contrariar a tendência introspectiva e vira-se para os textos, como objectos reais e empíricos, portadores de informação linguística. Como Slodzian refere,

> «*La terminologie textuelle part des occurrences manifestées en texte, donc du syntagmatique*» (Slodzian, 2000:77).

Assim, o acesso aos conceitos, nesta perspectiva textual, já não é alcançado por processos mentais, mas pela consulta sistemática dos textos de especialidade que são, cada um deles, um retalho parcial da completude de um conceito. Jacquemin e Bourigault afirmam que

> «*the classical view assumes that experts in an area of knowledge have conceptual maps in their minds. This assumption is misleading and unproductive because experts cannot build a conceptual map from introspection. Terminologists constantly refer to textual data and analyze the lexical elements in order to acquire and validate a conceptual description*» (Bourigault & Jacquemin, 2003:2).

Ao modificar-se a análise linguística, afirmando-a dependente do seu contexto, tal como nas línguas não especializadas, afigura-se indispensável mudar o eixo principal da análise terminológica do espectro da frase e do domínio da sintaxe, ponto de referência da investigação linguística, para um espectro mais largo, o do texto, como já havia acontecido com alguns investigadores na Linguística, e é neste quadro que desponta a Terminologia Textual. Segundo Slodzian e Bourigault,

> «*L'activité de construction d'une terminologie est désormais essentiellement une tâche d'analyse de corpus textuels. Ils appellent du même coup à un renouvellement théorique de la terminologie : c'est dans le cadre d'une linguistique textuelle que doivent être posées les bases théoriques de la terminologie.*» (Bourigault & Slodzian, 1998:30)

Rey critica ainda a procura obsessiva pelo universalismo e rigidez estrutural dos conceitos na teoria wüsteriana e proclama uma cisão total com essa forma de abordar a conceptualização terminológica, como se pode ler:

> «*La terminologie se doit de critiquer vigoureusement cette optique mentaliste, selon laquelle les concepts permettent d'appréhender les propriétés caractéristiques objectives qui font q'une chose est ce q'elle est*» (Rey, 1979:32-33).

A análise linguística textual, tendo como ponto de partida os elementos encontrados nos textos, apresenta uma estrutura conceptual flexível em que as unidades

lexicais se redefinem constantemente e consoante o seu uso. Como explicam Béjoint e Thoiron,

> «*On a tendence à dire que le sens d'un terme se confond avec la conceptualisation de ce qu'il désigne, alors que le sens d'un mot (son "signifié" saussurien) serait basé sur l'usage qui est fait de ce mot et comprendrait d'autres composants, ne serait-ce que la connotation et tout ce qui est véhiculé para la forme linguistique particulière que la communauté utilise pour exprimer un sens*» (Béjoint & Thoiron, 2000:10).

Desta forma, é o sentido que se procura atingir e não o significado, pois é o primeiro, em última análise, que mais aproxima a unidade lexical de ser um termo de uma dada especialidade. Para melhor entendermos a diferença entre sentido e significado, é necessário passar pela abordagem de Depecker acerca de conceito e significado:

> «*le concept ne se résume pas au signifié. L'un et l'autre sont distinguables même s'ils ont tendance à être confondus dans la langue*» (Depecker, 2000:91).

A noção de conceito, em Terminologia, é apresentada como uma descrição em forma de definição linguística que está sujeita aos constrangimentos individuais e sociais, sendo, portanto, um ponto de vista cultural e contaminado. O significado é, assim, apresentado como um objecto semântico múltiplo, no qual se podem integrar vários sentidos que se materializam linguisticamente nos respectivos conceitos, continuando Depecker:

> «*Le signifié se decompose en sèmes, unités sémantiques différentielles de contenu [...]. [...] un sème connotatif détermine le sens d'un signe de façon relativement instable, virtuelle, voire individuelle : présent dans le signe, il est plus ou moins actualisable selon les contextes et les situations de communication*» (Depecker, 2000:95).

Segundo a perspectiva apresentada, o sentido é visto como «*le sens actualisé d'un signe*» (Depecker, 2000:111). O sema ganha, neste contexto, protagonismo sobre a perspectiva onomaseológica, também ela um traço distintivo da T.G.T., que é posta de parte, pois, o objectivo da análise terminológica não é atingir uma etiqueta a partir de um conceito delimitado, mas pegar nas unidades linguísticas reais e construir o conceito a cada realização textual que se encontre. Como Conceição aponta,

> «*nous avons donc adopté une méthodologie plutôt sémasiologique, tout comme la plupart de travaux faits avec des corpus, puisque nous partons des réalisations linguistiques textuelles (dans le cadre de la terminologie textuelle) pour en arriver à leurs correspondantes sémantiques et cognitives*» (Conceição, 2005:18).

A questão do sentido, que implica a polissemia e variação semântica, amplamente rejeitadas no quadro tradicionalista da terminologia, aparece agora, como um resultado natural das relações contextuais, ligada também ao conteúdo e à expressão. Como Slodzian destaca,

> «*par ailleurs, l'approche textuelle postulant une unité des plans du contenu et de l'expression, la fabrique du sens n'est plus envisagée sous l'angle unique de la lexicalization*» (Slodzian, 2000:76).

Tal como defende Hoffmann (1988), o texto especializado não difere muito do texto produzido em línguas naturais e que, por isso, os métodos de análise devem ser similares. Partindo-se do texto como base, no quadro da Linguística Textual, a tarefa de descrição lexical, segundo Bourigault e Slodzian,

> «*est un travail de fixation, stabilisation, homogénéisation d'une signification, dont le résultat est le terme. Il s'agit de construire un type (une signification stable) à partir des occurrences manifestées en texte.*» (Bourigault & Slodzian, 1998:30)

O processo de construção conceptual é o resultado de um método iterativo de consultas intra e intertextuais que aos poucos vão contribuindo para a construção desse conceito, conforme se pode confirmar pelo seguinte excerto de Conceição:

> *«Chaque allusion à une concept ne le délimite pas entièrement, d'autant plus que le discours (aus sens large) où cette désignation est utilisée contribue à la formation de le signification. Ceci revient à dire que pour atteindre le concept dans sa totalité, ou pour approcher cette totalité, il faut voir les actualizations des ses traits faites au long du discours. La signification verbalisée peut ainsi être envisagée en tant que phénomène transphrastique ou même transdiscursif»* (Conceição, 2005:14-15).

Chegar ao significado de um termo é, portanto, um procedimento complexo que obriga a respeitar o texto, como ponto de partida e de consulta permanente, para que o termo se reconstrua ao mesmo tempo que a análise do texto avança. Segundo Bourigault e Slodzian, há um método a respeitar,

> «*On va du texte vers le terme. [...] le terme est un construit. Il est le produit d'un travail d'analyse, mené par le linguiste terminologue*» (Bourigault & Slodzian, 1998:30).

Deste modo, a construção de produtos terminológicos (glossários, terminologias, bases de conhecimentos terminológicos e outros) deixa de ser apenas uma recolha prescritiva de unidades lexicais etiquetadas como termos e volta a perspectivar-se como uma abordagem descritiva. Como Bourigault e Slodzian descrevem,

> «*l'approche textuelle est descriptive (on analyse le fonctionnement d'unités lexicales en corpus) et non plus normative : les enjeux de la planification linguistique, si légitimes soient-ils, sont dissociés du travail terminologique proprement dit. L'objectif premier de la terminologie classique était la normalisation des langages techniques via la fixation apriori de la signification des mots*» (Bourigault & Slodzian, 1998:30).

Esta perspectiva flexível e aberta da interpretação textual transmite-se à natureza dos conceitos, passando a ser imprescindível consultar textos reais para proceder à descrição e normalização dos conceitos e termos. Seguindo Bourigault e Slodzian,

> «*Les textes réels qui prolifèrent et circulent en tous sens, bousculant les frontières de domaines, remettent en cause ce projet de mise en ordre des termes apriori. Un tel programme de régulation prescriptive est contredit par le caractère fondamentalement ouvert des textes et de leurs signes. Le constat de la plasticité du donné linguistique conduit à refonder une "bonne pratique terminologique" sur le descriptif*» (Bourigault & Slodzian, 1998:30).

Os dados textuais passam, assim, a desempenhar um papel fundamental nos trabalhos de aquisição terminológica, pois, os textos são o habitat natural dos termos. Neste sentido, a aplicação informática foi concebida para ter como objecto de análise principal o texto, considerando-o como ponto de partida para atingir os termos. Para proceder à recolha e análise dos dados terminológicos que vão definir a construção dos conceitos torna-se necessário compilar múltiplos textos que possam servir simultaneamente de objecto de análise e confrontação e permitam uma verdadeira normalização conceptual e terminológica. Conceição afirma que

> «*pour cerner un concept par le biais des expressions linguistiques qui le dénomment, il faut analyser ces expressions et les mettre en rapport les unes avec les autres*» (Conceição, 2005:14).

A necessidade de recorrer a *corpora* textuais vai, assim, aumentando, mas é fundamental que sejam recolhidos de acordo com os objectivos e necessidades em questão e para isso há que desenvolver métodos e ferramentas de trabalho adequados. Com afirmam Aussenac e Bouringault, citando Slodzian (2000),

> «*Depuis le milieu des années 90, un courant de recherche se développe autour de la terminologie textuelle, qui préconise la construction de ter-*

> *minologies à partir de textes, et qui sollicite le TAL [Traitement Automatique des Langues] pour des méthodes et outils d'analyse de corpus*» (Aussenac-Gilles & Bourigault, 2003:30).

A procura por um número elevado de textos que se concentrem na pertinência para a especialidade e na definição inicial de aplicação desejada pelo investigador, exige um protocolo de procedimentos e o respeito por uma série de critérios, abordados no capítulo seguinte, e que a Terminologia Textual vai buscar à Linguística de Corpus.

A Terminologia Textual propõe-se, pois, através de uma abordagem descritiva de textos produzidos numa dada área especializada, procurar os termos que representam o conhecimento nesse domínio. Como referem Bourigault e Slodzian,

> «*Les applications de la terminologie sont le plus souvent des applications textuelles (traduction, indexation, aide à la rédaction) ; la terminologie doit 'venir' des textes pour mieux y 'retourner'. C'est parce qu'elle n'est jamais déliée du texte qu'on parle de 'terminologie textuelle'*» (Bourigault & Slodzian, 1998:30).

Assim, para atingir o propósito e recorrendo a metodologias pertencentes à Linguística de Corpus, a metodologia da Terminologia Textual investe na recolha de grandes quantidades de textos reais e de especialidade para posterior análise. Como afirmam Bourigault e Slodzian,

> «*C'est dans les textes produits ou utilisés par une communauté d'experts, que sont exprimées, et donc accessibles, une bonne partie des connaissances partagées de cette communauté, c'est donc par là qu'il faut commencer l'analyse*» (Bourigault & Slodzian, 1998:30).

É no quadro da Terminologia Textual, portanto, que o projecto de protótipo se integra, valorizando a primazia do texto, tal como definido neste capítulo, enquanto

veículo e contentor privilegiado de dados terminológicos, no âmbito dos estudos em línguas de especialidade.

2.7 Síntese

A questão da essência epistemológica da Terminologia parece tornar-se cada vez mais debatida e, ainda, com incerteza no que diz respeito a soluções consensuais e finais. Talvez o consenso também não seja o caminho mais adequado, pois faz parte da natureza de qualquer ciência a instabilidade constante, no meio de uma estabilidade ilusória, tal como a aparente constância das línguas naturais. A Terminologia atravessou uma fase de estabilidade consentida, com a Teoria Geral da Terminologia de Wüster, mas, como compete a qualquer ciência, questionou-se com o aparecimento da textualidade e com o reavivar do *corpus* na Linguística, procurando agora definir novas bases que respondam aos recentes desafios metodológicos e conceptuais.

Há uma variedade de factores que impulsionam uma ruptura com a visão tradicionalista da Terminologia, principalmente as influências resultantes da interacção com outras ciências, das quais se destaca a Linguística. A mudança da perspectiva sobre a noção de conceito, o surgimento do texto como unidade de análise linguística, o postulado do sentido e do contexto e a evolução da Informática foram os maiores contributos para que as alternativas se apresentassem. Segundo Slodzian, há duas correntes maiores que se manifestam como dominantes no panorama da investigação terminológica actualmente,

> «*une terminologie conceptuelle qui se décompose en deux branches principales opposées sur la notion de concept*» e «*la terminologie textuelle, dont le refus du référentialisme est plus ou moins marqué selon les écoles, déplace la problématique de la terminologie aux relations entre signifiés et à la spécificité du fonctionnement des signifiés dans les textes à caractère technique et scientifique*» (Slodzian, 2006:2).

Portanto, uma terminologia conceptual e mais introspectiva e outra, textual, virada para a vertente mais pragmática, afirmando Slodzian que só o tempo e a prática poderão comprovar qual delas é a mais indicada.

No entanto, a procura que existe na sociedade pelos produtos terminológicos, consequência da progressiva evolução tecnológica e massificação da informação, forçou a Terminologia a procurar adaptar-se às necessidades e tornar-se mais prática e interventiva. Voltando-se para a análise dos objectos onde se afirma residir o conhecimento especializado, a Terminologia Textual parece querer ganhar um ascendente sobre os restantes movimentos concorrentes. Através da análise dos textos presentes nos *corpora* compilados, os terminólogos procuram descrever o conhecimento especializado, partindo das unidades terminológicas identificadas. Como L'Homme refere,

> «*terminologists will make decisions since they must interpret data and synthesize their findings, but these are based on the observation of interactions between lexical units that appear in corpora*» (L'Homme, 1998:6).

Apesar disso, a interdisciplinaridade intrínseca da Terminologia é um estímulo constante à renovação metodológica e epistemológica. A aparente vantagem de um movimento parece ser sempre uma inevitável transitoriedade até que surja uma melhor e mais adequada fundamentação teórica. As contínuas e crescentes evoluções nas áreas da Inteligência Artificial e da tecnologia informática, com as quais a Terminologia mantém um diálogo insistente, perspectivam avanços ainda mais sólidos, tanto mais que a Terminologia, pelo seu papel decisivo no estabelecimento de novas fronteiras no conhecimento especializado, estará sempre presente na vanguarda da descoberta, prestando o seu indispensável contributo para o progresso da ciência.

De acordo com o que se constatou ao longo do capítulo, a consecução de um projecto que tem como objectivo principal executar tarefas terminológicas de uma forma rápida e válida e que serve de referência para o desenvolvimento e fundamenta-

ção desta dissertação, implica, desde logo, dois pressupostos. O primeiro estabelece a formalização do carácter informático do protótipo apresentado que, por questões epistemológicas e logísticas, só se torna exequível com recurso a métodos automáticos. O segundo remete para a definição e implementação de estruturas metodológicas que partam das bases epistemológicos no quadro da Terminologia Textual, pois pensamos que, tal como foi exposto, os termos existem contextualizados nos textos, enquanto produtos comunicativos reais, e é a partir destes que a reconstrução conceptual deve ser efectuada. Deste modo, é importante proceder também à clarificação dos procedimentos para recolha e análise dos textos que passamos a apresentar no capítulo que se segue.

Capítulo 3

Definição de *Corpus*

3.1 Introdução

Qualquer investigador reconhece que a prática comum e necessária numa investigação científica em qualquer área, não sendo a Linguística excepção, obriga a reunir uma grande quantidade de materiais que possa complementar o estudo do tema em questão e, assim, provar uma qualquer teoria que pretenda ver validada.

Esses materiais de trabalho reunidos para investigação poderão ser divididos em dois tipos:

- Os que servem para formar o investigador e auxiliá-lo a adquirir e solidificar as suas competências e conhecimentos na área.
- Os que, depois de recolhidos e analisados, integram o conjunto de provas que ilustram e confirmam a tese apresentada.

Os primeiros serão os textos utilizados para construir o seu conhecimento sobre o assunto, enquanto os segundos, fornecedores de provas objectivas, como afirma Sinclair, recolhidos «*[. . .]to present the researcher with objective evidence*» (Sinclair, 1991:1), são o objecto de análise e constituem uma forma de tornar mais clara e inequívoca uma possível relação existente entre a hipótese levantada e a sua verificação, tornando-se, eventualmente, numa outra fase da investigação, a confirmação

ou rejeição dos pressupostos defendidos. Poder-se-á proceder a uma demonstração, ainda que muito simplificada, do funcionamento do processo argumentativo para se estabelecer um paralelo com uma qualquer investigação científica. Diz Perelman que

> «*a evocação de um certo número de exemplos da mesma natureza não pode deixar dúvida alguma no espírito do leitor: trata-se, decerto, de uma argumentação que visa passar do caso particular para uma generalização*» (Perelman, 1993:119).

No processo de construção de uma análise científica, é necessário chegar aos exemplos que contêm os dados relevantes e, a partir destes, preparar e desenvolver a investigação. Logo se compreende a importância que a recolha de materiais exemplificativos tem na validação de um qualquer processo de estudo científico, apresentando-se como um ponto de partida comum a todos eles. Como relembra Tognini-Bonelli,

> «*like all types of scientific enquiry, the starting point is actual authentic data*» (Tognini-Bonelli, 2001:2).

A investigação em língua de especialidade tem características próprias por se integrarem, no mesmo conjunto de materiais, os que servem para análise e os que servem para formação, dada a natureza linguística de ambos. O *corpus* será, assim, a colecção de materiais que reúne dados sobre uma determinada área a investigar, determinados pelo uso de um protocolo de critérios que definem as escolhas. Os processos de compilação e de análise de textos relacionados com especialidades, que são, de alguma forma, pequenas amostras linguísticas de conhecimento especializado, têm por objectivos atingir a representatividade da totalidade de textos relacionados com a área e, ao mesmo tempo, permitir a reconstrução conceptual da especialidade sob investigação. Logo, ao tentar aproximarmo-nos de uma generalização, seguindo o raciocínio de Perelman, estamos a procurar validar uma tese em questão. Qualquer argumento que se queira válido deverá, idealmente, ser acompanhado

de exemplos que permitam ilustrá-lo, pretendendo ser, qualquer um deles, provas inequívocas da relação entre a tese apresentada e a verificação prática dessa tese. O *corpus* apresenta-se, assim, como um elemento de comprovação e de constituição de conhecimento sobre uma determinada área. Esta flexibilidade permite conceber os *corpora* como «*multifunctional resources*» (McEnery, 2003:449).

Neste capítulo, abordam-se os aspectos epistemológicos e metodológicos que rodeiam a constituição e a gestão de um *corpus* de especialidade, com vista à melhoria das condições de execução em qualidade e em rapidez da tarefa de compilação de materiais para investigação linguística e terminológica numa área específica do conhecimento. O objectivo é compreender melhor todos os aspectos supra-referidos e enquadrá-los na fundamentação da concepção de protótipo de aplicação informática a apresentar no quarto capítulo, partindo-se do conceito de *corpus*, introduzido de seguida, até chegar às práticas de constituição e de gestão.

3.2 A Linguística e o *Corpus*

O primeiro significado da palavra *corpus* remonta ao latim[1], referindo-se ao conjunto das principais partes do corpo humano. Sobressaía já a ideia de que um *corpus* era uma totalidade, um objecto único, contudo divisível na sua origem pela separação de vários elementos. Ainda durante o período romano, no séc. I d.C., por ordem do Imperador Justiniano, surgiu uma compilação de leis denominada *Corpus Juris Civilis* que sistematizava e juntava num único volume todas as leis romanas. Ainda que a ideia de uma reunião de vários elementos que formam um todo se mantivesse associada à unidade lexical *corpus*, de notar que esta nova utilização do *corpus* deixava de ser exclusiva do domínio orgânico e anatómico e estendia-se até aos domínios do Direito, neste caso particular, representando a compilação, numa parte única e escrita, de todas as leis existentes.

[1]ETIM emprt. lat. *cŏrpus* nom. sing. de *corpus,ŏrporis* 'corpo' (Fonte: Dicionário Houaiss da Língua Portuguesa)

Há duas ideias essenciais a reter nesta actualização da unidade lexical *corpus*: em primeiro lugar, a preocupação em recolher o número máximo de elementos ligados a uma única área, neste caso o Direito, e, em segundo lugar, a denominação passar a reter em si a ideia de ser uma compilação de ideias, de afirmações ou de produções comunicativas em suporte escrito. Esta última ideia aproxima-se da defendida por Sinclair, no âmbito da Linguística, uns séculos mais tarde, quando afirma que

> «*a corpus is a collection of naturally-occuring language text*» (Sinclair, 1991:171).

A utilização do *corpus*, como base de estudo para investigação, acaba por se integrar na metodologia de trabalho nos estudos científicos e, em particular, nos linguísticos, ainda que seja apenas no séc. XIX, com Bopp (1787-1832), que pela primeira vez se aplica o termo linguista para denominar os estudiosos que se dedicam à investigação e confirmação de hipóteses nas línguas através de metodologias científicas. Como Law refere,

> «*It was Bopp who in many respects set the tone of mainstream linguistic research during the nineteenth century*» (Law, 2003:267).

No século seguinte, com o estruturalismo e influenciada pela busca de rigor científico, a compilação de *corpus* ganhou a sistematicidade e o método que não tinha atingido até então. Como confirma Lyons,

> «*a investigação da linguagem, tal como levada a efeito na Europa e nos Estados Unidos da América antes do século XIX, era subjetiva, especulativa e não-sistemática*» (Lyons, 1970:18).

Apoiando-se na recolha de amostras de língua, a partir das quais formulavam e confirmavam hipóteses investigativas, os linguistas desenvolviam os seus estudos de uma forma mais organizada e rigorosa. Como refere Teubert,

> «*[linguists] wanted to investigate the structure of language, based on analyses of texts, in order to understand the language system behind it*» (Halliday & Teubert, 2004:81).

Começa, assim, a formalizar-se metodologicamente uma linguística que tem, como ponto de partida e base de investigação, a recolha e análise dos elementos produzidos na própria língua. No entanto, não é de imediato que se vai impor, pois com Chomsky, nos anos 60, surge uma perspectiva conceptual de análise linguística que vai questionar a metodologia estruturalista. Ao afirmar que a língua surge de uma capacidade inata e genuinamente generativa do cérebro do homem, a corrente chomskiana põe num segundo plano a necessidade de recolher e analisar quaisquer dados relativos à produção linguística do ser humano. Chomsky afirma, citado por Cook e Newson, que

> «*true formalization is rarely a useful device in linguistics*» (Cook & Newson, 1996:36).

Na abordagem generativista, é menos relevante analisar o que já foi produzido pelo falante do que compreender os princípios que o levam a poder construir frases com combinações possíveis a roçar o infinito e testar esses princípios através da formulação de exemplos. Como referem Halliday e Teubert,

> «*Noam Chomsky and many of his followers have dismissed the corpus as a source of our linguistic knowledge. Language, they say, is productive*» (Halliday & Teubert, 2004:104).

Esta nova abordagem não foi consensual e houve contestação, por parte de linguistas que consideravam primordial partir das produções de língua, principalmente decorrente do uso de exemplos cuidadosamente escolhidos para cada caso pelos investigadores, pois, ao invés de se recolherem textos e trabalharem com casos reais, procedia-se à construção de frases que pudessem confirmar as condições apresentadas pelo investigador na defesa do seu modelo. Como demonstra Jacques,

> «*En définitive, autant les aléas du jugement de grammaticalité que le flou qui entoure la notion ont contribué à discréditer la linguistique introspective. [...] C'est donc une critique reéllement fondée car elle pose la question de ce qu'est cette langue que le linguiste décrit par inspection*» (Jacques, 2005:22).

O uso dos *corpora* não pode ser de todo abandonado, pois a tendência cognitiva e abstraccionista da teoria chomskiana, pilar da perspectiva universal da gramática, afasta-se em demasiado da realidade linguística, não conseguindo lidar com os aspectos contextuais, especialmente aqueles ligados ao sentido e à variação, pelo que tende a unificar a língua e a criá-la nos limites da capacidade introspectiva do investigador. A complementaridade da linguística introspectiva e de *corpus* é defendida por vários autores, entre os quais Halliday e Teubert, que afirmam o seguinte:

> «*the perspective of Chomskyan and cognitive linguistics represents a very different view of language [...] Both views are, of course, legitimate, and they are complementary. Corpus linguistics deals with meaning. Cognitive linguistics is concerned with understanding*» (Halliday & Teubert, 2004:98).

Com a incapacidade crescente de apresentar descrições adequadas para todos os casos, a teoria de Chomsky acaba por motivar indirectamente o regresso ao uso de *corpora*. A necessidade de textos com recolhas reais de língua veio demonstrar as incoerências da perspectiva unicamente generativa e tentar suprimir as lacunas que alguns casos haviam posto a descoberto. Como referem Halliday e Teubert, existiam

> «*certain features of the language insufficiently described [...] which could not be answered by introspection alone. Real language data were needed*» (Halliday & Teubert, 2004:107).

Por questões práticas e que se sobrepõem às considerações teóricas sobre o funcionamento da língua, no âmbito de grande parte dos estudos linguísticos, há

uma necessidade imediata de conseguir mais do que testar os limites da língua e de executar tarefas com objectivos imediatos que resolvam os problemas terminológicos do presente. A forma de conseguir trabalhar e atingir soluções passa por uma observação atenta e uma descrição detalhada da língua real, o que é possível e justificado através da utilização de *corpora*. Como Tognini-Bonelli descreve,

> «*most linguistic research demands evidence of language in use, and a corpus provides such evidence*» (Tognini-Bonelli, 2001:47).

A riqueza de uma amostra de língua proporciona uma oportunidade de extrair, não apenas informação linguística, mas uma diversidade múltipla de dados que, devidamente organizados e seleccionados, tornam possível uma reconstrução que extravasa o campo linguístico e atinge o domínio conceptual. Tal como refere Conceição,

> «*les informations véhiculés par les corpus sont de nature linguistique, sociolinguistique, pragmatique, diachronique et culturelle*» (Conceição, 2005:125).

Partindo das evidências apresentadas até este ponto, consideraremos, no âmbito deste trabalho, o *corpus* como um conjunto de textos em suporte electrónico recolhido em função da consecução de um objectivo pré-definido e com determinados critérios que delimitam as suas características e a tipologia em que se enquadram, formando uma base textual para análise posterior.

Segundo Tognini-Bonelli, podem encontrar-se, em geral, duas formas distintas de utilizar o *corpus* numa investigação linguística:

> «*a corpus can be used in different ways in order to validate, exemplify or build up a language theory. [...] Different terms are often used by different scholars, but all centre round one basic distinction. The terms that are frequently used are corpus-based, as against corpus-driven*» (Tognini-Bonelli, 2001:65).

Essencialmente, a diferença assenta na forma como o *corpus* participa no processo de análise linguística. Como Tognini-Bonelli explica,

> «*corpus-based is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora become available to inform language study*» (Tognini-Bonelli, 2001:65).

Este tipo de utilização é mais tradicional, sendo mais recorrente quando não existiam recursos para coligir grandes bases textuais. Por outro lado, continua Tognini-Bonelli, na

> «*corpus-driven approach [...] the theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus*» (Tognini-Bonelli, 2001:65).

A separação entre *corpus-based* e *corpus-driven* encontra-se, pois, na forma como é encaixado no processo de análise linguística o *corpus*, dado que se encontra presente em ambos. O método *corpus-based* usa os dados textuais como forma de comprovar as formulações linguísticas, aproximando-se mais de um uso possível na abordagem introspectiva. Já o método «*corpus-driven*» estende a utilização do *corpus* ao ponto de partida da análise e aplica-lhes critérios linguísticos e estatísticos que conduzam à formulação de hipóteses, tornando-se, assim, o *corpus* a base de toda a investigação.

Para que estas metodologias conseguissem singrar, principalmente a *corpus-driven*, foi essencial o contributo da Informática, que permitiu suportar tecnologicamente a logística necessária à recolha maciça de textos e a sua consulta e análise no espaço de tempo exigido para garantir a validade dos resultados das investigações. Assim, como afirma Kennedy, surge uma

> «*new scholarly enterprise known as corpus linguistics*» (Kennedy, 1998:1).

Recordando o que foi apresentado, podemos concluir que a Linguística de *Corpus* é uma derivação da Linguística que recorre ao *corpus* como objecto de análise para construir as suas formulações sobre a língua. A Linguística de *Corpus*, na sua especificidade, permite outro tipo de operações que não estão ao dispor da metodologia introspectiva, como enumera Jaques:

> «
>
> 1. *mettre en lumière des fonctionnemments linguistiques qui échappent à l'intuition*
> 2. *corriger les intuitions sur le fonctionnements de la langue*
> 3. *d'avoir des indications en terme e fréquence et établir des relations statistiques entre ensembles de faits*
> 4. *d'atteindre et rendre compte de la variation*»
>
> (Jacques, 2005:25-26).

Pelas razões apresentadas, a Terminologia, que procura reconstruir conceptualmente uma especialidade a partir de textos, pode tirar partido da metodologia usada na Linguística de *Corpus* e do carácter multifacetado do *corpus*. O *corpus* em Terminologia partilha as características apresentadas em Linguística, sendo alvo de aplicação de um protocolo de regras que define e delimita as suas características e tipologia da sua constituição, no âmbito de um objectivo, ainda que constrito a uma especialidade, como se pode verificar já de seguida.

3.3 Constituição de *corpora*

A fase de constituição do *corpus* tem a importância de poder conferir um alicerce sólido e produtivo para o processo de investigação. Não se pode confundir uma base textual, como é o caso de um conjunto de textos aleatórios ou mesmo a Internet, com um *corpus*. A primeira corresponde a um repositório de textos não

organizados e o segundo, como já vimos, é uma selecção criteriosa e organizada de textos, com um objectivo claramente definido. Por isso, quando se prepara a constituição de um *corpus*, é indispensável completar uma série de etapas que são antecedentes à compilação. Como refere Pearson,

> «*prior to compiling a corpus, compilers will have to address a number of issues*» (Pearson, 1998:50).

O trabalho de constituição do *corpus* é uma tarefa metódica, rigorosa e que obriga a uma preparação adequada para que seja realizada com sucesso, tal como se poderá observar de seguida, a partir da apresentação e da discussão de uma série de pressupostos e de etapas indispensáveis à consecução de um fim.

3.3.1 Definição de um objectivo

O passo que antecede imediatamente a compilação de um *corpus* e que serve de ponto de partida para tudo o que se desenvolve de seguida é a definição dos objectivos que se pretendem atingir. Só depois de delimitados os propósitos a alcançar com o *corpus* e de estar definido o domínio que servirá de referência, é que se parte para a recolha de textos. Como relembram Bowker e Pearson,

> «*the types of texts that you include in your corpus will depend on what you wish to study*» (Bowker & Pearson, 2002:51).

Portanto, é, ainda, sem quaisquer textos recolhidos, que se determina o domínio a estudar e se selecciona o tipo de *corpus* mais adequado ao desenvolvimento das actividades de investigação. Como nos demonstra Conceição,

> «*après la définition des objectifs de sa constitution et des fins pour lesquelles le corpus est constitué et aprés la délimitation du domaine du savoir qu'il est supposé représenter, intervient les choix des textes*» (Conceição, 2005:135).

Bowker acrescenta, também, ser muito importante conhecer o domínio para saber se há possibilidade de os objectivos pretendidos serem atingidos. A partir da aferição dos limites e da decomposição em partes mais pequenas, se necessário, ou até com o auxílio de um especialista, pode-se aceder a um melhor conhecimento do domínio. Por vezes, para atingir o grau de conhecimento necessário, será fundamental estruturar a investigação em etapas mais simples:

> «*As they familiarize themselves with the field, specialized lexicographers attempt to identify the boundaries of the subject field and to classify the field into major subdivisions*» (Bowker, 2003:161).

A procura pelo enquadramento eficaz entre o objectivo a atingir, que, como vimos, será o primeiro momento da investigação, o conhecimento do domínio, que discutimos de seguida, e o *corpus* adequado, que resultará da definição e aplicação de critérios também abordados mais à frente, permitirá que o estudo alcance resultados relevantes. Sem o estabelecimento destes pressupostos, de uma forma coerente, o protótipo de *software* informático pouco poderá fazer, no que diz respeito à validade dos resultados, pois, assenta completamente na sua correcta definição e aplicação.

3.3.2 Domínio

O domínio é um termo com diversas acepções, que nos cabe, desde logo, restringir, no âmbito deste trabalho, o seu uso aos estudos em Terminologia. Muitas vezes designado por tópico, assunto ou área de um texto, todos estes termos procuram representar o sistema conceptual restrito ao qual o texto e a análise se devem sujeitar e que decorre dos objectivos e necessidades da investigação terminológica. Dado que qualquer estrutura conceptual é sempre uma perspectiva que reflecte o ponto de vista de um grupo, a noção de domínio, tal como a de género textual, já apresentada no capítulo anterior, está indelevelmente ligada às práticas sociais e aos discursos produzidos pelos seus agentes, logo, como Gaudin explica,

> «*L'approche en termes de domaines, outre le fait qu'il s'agit d'une notion de sens commun, peut donc recevoir deux types de critiques. Le premiere concerne ce que nous savons du monde et peux être étayé par ce que nous disent les historiens et les éspistémologues. [...] Le second type de critiques relève de ce que nous savons du language. La notion de domaine propose à l'analyste un découpage qui n'est pertinent ni pour isoler des communautés de locuteurs [...]. En parlant des domaines, on ramasse trop large tout en séparant trop strictement des secteurs qui communiquent entre eux*» (Gaudin, 2003:51-52).

Pode-se, desta forma, no quadro da socioterminologia e no contexto desta dissertação, entender o domínio, na vertente de aplicação prática e com objectivos terminológicos, como uma esfera de actividade, intrínseca a uma área do saber, que confere uma interpretação específica aos textos produzidos e aos elementos, entre os quais os termos, neles presentes. Como refere Conceição,

> «*si l'on entend domaine comme sphére d'activité, et donc sphére du savoir, l'utilisation des concepts de domaine d'experience et d'application est justifiée*» (Conceição, 2005:137).

O domínio é, então, o conjunto de noções pré-estabelecidas, ainda que dinâmicas, que representa uma parte especializada do conhecimento e que serve de estrutura para o enquadramento linguístico e posterior selecção de textos para um *corpus*.

A aceitação da noção de domínio, enquanto sistema conceptual, no quadro da Terminologia Textual, é a garantia da identificação e classificação de unidades terminológicas, pois o domínio, neste quadro, é identificador de uma especialidade e serve como estrutura de enquadramento de um termo num grupo conceptual. No entanto, a impossibilidade que existe em delimitar fronteiras dos domínios, devido ao intercruzar constante dos conceitos que os compõem, faz com que, por vezes, a integração de alguns textos numa área conceptual específica seja complexa. Como demonstram Bowker e Pearson,

> «*Many specialized subjects are multidisciplinary (e.g. biochemistry), which means that it can be difficult to know where one subject field ends and the next begins.*» (Bowker & Pearson, 2002:50)

A decisão sobre o domínio a estudar tem implicações directas no *corpus* a recolher, quer pela disponibilidade, quer pelos critérios de aplicação, e terá de ser sempre uma opção bem avaliada, não sendo todos os domínios ideais para um qualquer estudo da língua de especialidade que se pretenda desenvolver. Torna-se fundamental, desde logo, nesses casos, adequar a selecção de domínio ao objectivo pretendido, podendo, inclusive, em determinados estudos, por se considerar a escolha de um domínio como pouco produtiva ou pouco influente, conferir uma preponderância superior a outros critérios. Como afirmam Bowker e Pearson,

> «*If your project sets out to study particular features of speacialized language, it may not be necessary for all your texts to be about the same subject.*» (Bowker & Pearson, 2002:50)

O avançar dos estudos no quadro da Terminologia Textual veio revelar que os dados terminológicos recolhidos eram insuficientes, principalmente para fins que não os linguísticos, e que era necessária uma melhor representação do conhecimento existente nos domínios de especialidade. Esta necessidade levou uma aproximação do conceito de ontologia no seio da Terminologia, destacando-se a importância da construção de ontologias que pudessem melhor objectivar o conjunto de conceitos, definições, relações e regras que capturam o conteúdo semântico de um domínio. Como Vossen refere,

> «*In general, an ontology can be described as an inventory of the objects, processes,etc. in a domain, as well as a specification of (some of) the relations that hold among them*» (Vossen, 2003:465).

Assim, como Meyer (1992) aponta, assistiu-se à conversão de grandes bases de dados terminológicas (BDTs[2]), que eram repositórios de dados textuais, gramaticais e terminológicos, mas muito limitados para as necessidades de representação conceptual de um domínio e para as possibilidades que a Informática disponibilizava. Meyer, Skuce, Bowker e Eck apresentam as BDTs da seguinte forma:

> «*A major weakness of TDBs is that they provide mainly linguistic information about terms (e.g. equivalents in other languages, morphological information, style labels); conceptual information is sparse (limited to definitions and sometimes contexts), unstructured, inconsistent and implicit. Given these problems, a growing number of terminology researchers are calling for the evolution of TDBs into a new generation of terminological repositories that are knowledge-based*» (Meyer et al., 1992:956).

As BDTs, tal como o excerto se refere, converteram-se em bases de conhecimento terminológico (BCTs[3]). Nestas bases de conhecimento, passa a constar, não só informação linguística, mas também a formalização ontológica do domínio em estudo, não se introduzindo na base apenas os termos, mas também classes, atributos, funções e relações que os conceitos representam dentro de um domínio. Meyer sintetiza, em três perspectivas, as principais alterações:

> «*The differences between a conventional TDB and a TKB can be examined from three points of view: 1) the information itself, 2) support for acquiring and systematizing the information and 3) facilities for retrieving the information.*» (Meyer et al., 1992:958)

As BCTs disponibilizam mais informação em quantidade e multiplicidade, permitindo uma melhor explicitação das relações conceptuais existentes nos domínios, melhoram a aquisição e a sistematização dessa informação, facilitando o processo

[2] *Terminological Data Bases* (TDBs)
[3] *Terminological Knowledge Bases* (TKBs)

de obtenção e organização dos dados, e dinamizam a busca de informação, multiplicando os vectores de cruzamento e de análise. A introdução das BCTs veio, desta forma, permitir a aquisição e a formalização do conhecimento nos domínios, que pela complexidade das relações conceptuais não conseguia ser representada nas BDTs.

Esta necessidade acrescida de melhorar a representação conceptual nos recursos terminológicos é explicada por Bourigault da seguinte forma:

> «*First, specialized texts in electronic form are easily accessible; [...] Secondly, several techniques borrowed from natural language processing, information retrieval, corpus linguistics, or artificial intelligence enable the extraction and representation of specialized knowledge in an efficient and often elegant manner. [...] In addition, specialists in natural language processing who were not primarily concerned with terminology have come to realize that the formalization of specialized texts is necessary in order to build useful applications.*» (Bourigault et al., 2001:VIII)

Os domínios de especialidade, tal qual foram mais atrás apresentados e definidos, constituem realidades dinâmicas e complexas, onde os conceitos que os integram e as suas relações são o reflexo dessa complexidade. No quadro da Terminologia Textual, considera-se o texto como o objecto ideal onde capturar essas realidades e o seu conteúdo conceptual para, de seguida, poder reproduzi-lo em sistemas onde essa informação é valiosa e reutilizá-lo através de processos automáticos.

No seguimento dos objectivos a que nos propusemos no início, pretendemos desenvolver uma ferramenta que faculte instrumentos para a formalização dessa complexidade conceptual, enquadrada nos pressupostos da epistemologia preconizada nos capítulos antecedentes e aqui defendidos.

3.3.3 Homogeneidade, representatividade e exaustividade

Para que o processo de constituição de *corpus* possa ser considerado válido, há questões fundamentais, no que diz respeito à metodologia da Linguística de *Corpus*,

que precisam de ser respeitadas. As noções de homogeneidade, de representatividade e de exaustividade aparecem, no início do séc. XX, associadas à boa prática de recolha de informação relacionada com uma área de investigação. Como refere Conceição,

> «*tout bon corpus devait obéir aux principes de l'homogénéité, de la representativité et de l'exhaustivité*» (Conceição, 2005:123).

Ainda que as regras da exaustividade e da representatividade se mantenham inalteradas, a natureza do princípio da homogeneidade, entretanto, modificou-se. O *corpus* homogéneo era o que apresentava as mesmas características e a mesma tipologia textual na sua constituição, porém, em determinadas situações, como afirma Conceição, «*par exemple, dans des recherches sur différents niveaux de langue*» (Conceição, 2005:135), é obrigatório recorrer à heterogenia tipológica textual para preencher os requisitos do objectivo definido. No âmbito da dissertação aqui apresentada, defende-se a presença do princípio da homogeneidade, não ao nível das características textuais, mas ao nível da aplicação dos critérios de selecção aos textos para integração no *corpus*. O *corpus* homogéneo é, assim, o produto final da definição e aplicação equivalente dos critérios de classificação que confere ao processo de constituição de *corpora* um indispensável rigor.

Quanto à representatividade e à exaustividade, são também garantias da validade e da coerência dos resultados obtidos, sendo necessário decidir a dimensão e o grau de completude de um *corpus*, mediante a ponderação da aplicação destes dois princípios sobre a disponibilidade de dados e o prazo de conclusão. A representatividade é um factor tão decisivo que Tognini-Bonelli define *corpus* como «*a collection of texts assumed to be representative of a given language*» (Tognini-Bonelli, 2001:2) e reitera a importância da representatividade, afirmando que

> «*a corpus which is taken to be representative is designed to be used as the basis for generalisations about the linguistic system*» (Tognini-Bonelli, 2001:79).

De acordo com este pressuposto, um *corpus* é considerado representativo quando, na sua constituição, se torna passível de ser uma fonte válida de identificação e de confirmação de regras linguísticas, pois, a representatividade permite validar os padrões detectados como extrapoláveis ou generalizáveis e considerá-los formalizações correctas da língua. No entanto, Pearson levanta um problema pertinente:

> «*but the question of how one determines the size of the representative subset is another unresolved issue*» (Pearson, 1998:59).

Não obstante o número de textos e amostras que se conseguem recolher dos vários *corpora* gigantescos já compilados, quando em comparação com os números totais existentes de produções de língua que se geram diariamente, será sempre difícil tornar qualquer amostra numericamente expressiva. Como refere Kennedy,

> «*it has to be stressed again that any corpus, however big, can never be more than a miniscule sample of all the speech or writing produced or received by all of the users of a major language on even a single day*» (Kennedy, 1998:66).

No entanto, poder-se-á argumentar, como já atrás foi referido, que a representatividade resulta da capacidade de apresentar e de justificar generalizações de hipóteses linguísticas e não obrigatoriamente do tamanho absoluto do *corpus*, ou seja, é possível ser representativo sem ser exaustivo, ainda que a aplicação da exaustividade na recolha dos textos aumente a probabilidade do *corpus* se tornar representativo. Sobretudo, é mais fácil ser representativo em línguas de especialidade do que na língua em geral, precisamente pela disparidade existente entre a totalidade de número de textos em cada. Como explica Pearson,

> «*if one wishes to carry out linguistic studies on a subset of the language, size may be less important but it will still be important for the corpus to*

be representative of the subset in question and, consequently, the larger it is, the more representative it is likely to be» (Pearson, 1998:51).

Deste modo, a exaustividade, dentro dos critérios definidos, é uma boa estratégia para se atingir a representatividade, como afirma Tognini-Bonelli, citando Leech:

«*Leech (1992:111), describing the paradigm of empirical research in corpus linguistics, states first of all that this will deal with observed evidence provided in the form of corpora, and secondly that this evidence will be used according to the principle of accountability, that is exhaustively; nothing will be selected in advance and nothing will be deliberately ignored as irrelevant*» (Tognini-Bonelli, 2001:71).

O *corpus* exaustivo é, assim, o que congrega o número máximo de textos que são possíveis recolher dentro dos critérios de classificação definidos. A importância da exaustividade das recolhas está patente na seguinte definição de Bilger, que apresenta o *corpus* como

«*recueils de textes rassemblent exhaustivement tous les documents disponibles pour certains champs d'étude*» (Bilger, 2000:11).

As ideias de representatividade e exaustividade não são bem aceites na abordagem generativa e introspectiva de Chomsky, pois violam flagrantemente a máxima da infinita competência linguística: nada pode ser representativo ou exaustivo o suficiente perante um número indefinível de produções possíveis.

A homogeneidade, a representatividade e a exaustividade, tal como foram apresentados neste trabalho, são, mais do que critérios, princípios nucleares da constituição do *corpus*, pois não são opcionais, condicionando a validade de qualquer estudo apresentado. Apresentam-se de seguida os critérios que definem as decisões para a constituição de *corpora* e que permitem delimitar as escolhas a efectuar na base textual de referência e orientar a pesquisa na direcção do objectivo definido no início.

Se a homogeneidade, a representatividade e a exaustividade são princípios cuja a prática de implementação precede a intervenção do protótipo informático na base textual e depende totalmente da intervenção do terminólogo, os critérios que se seguem são aplicáveis na massa global dos textos disponíveis e servem para, de forma automatizada, proceder a uma separação.

3.3.4 Critérios para classificação de *corpora* em Terminologia Textual

Na globalidade de textos disponíveis, todos eles candidatos ao *corpus*, a partir dos quais se podem recolher possíveis objectos de análise, é necessário aplicar regras pré-definidas de selecção, aqui designadas por filtros. A aplicação única e exclusiva dos princípios de homogeneidade, de representatividade e de exaustividade não garantem um bom *corpus* e é fundamental determinar critérios que filtrem o conjunto de candidatos. As exclusões ou inclusões de textos no *corpus* partem da definição e aplicação de um protocolo de critérios específicos ao universo de textos, o qual é indispensável para conferir rigor, método e validade à investigação. Aliás, como Bowker e Pearson alertam,

> «*corpora are not merely random collections of texts but, rather, they are collections that have been put together according to specific criteria*» (Bowker & Pearson, 2002:45).

Seja qual for a orientação da pesquisa, nunca se poderão ignorar as características dos textos a serem recolhidos. Estes são seleccionados a partir de particularidades existentes, que permitem separá-los e organizá-los de acordo com propriedades de escolha pré-definidas, integrando-os, assim, como parte do *corpus*. Bilger afirma que

> «*le terme de corpus désigne non pas simplement des collections de données de langage mais un choix organisé de ces données*» (Bilger, 2000:12).

A finalidade do *corpus* em Terminologia Textual é ser o objecto de um conjunto de análises linguísticas e estatísticas, apenas possíveis em produtos comunicativos reais, e facultar o acesso ao sistema conceptual da especialidade em análise a partir da descrição da língua, recorrendo-se, para isso, impreterivelmente, a critérios que permitam implementar os princípios fundamentais da validade, sem os quais a investigação não poderá ser bem sucedida. Como afirma Conceição,

> «*le corpus se veut représentatif du système linguistique, ils se doit d'être homogène et exhaustif*» (Conceição, 2005:126).

Os critérios pré-definidos e aplicados sobre os candidatos resultam num grupo limitado de textos que constituirão finalmente o *corpus*. O conjunto de critérios, como propõe Pearson, pode ser dividido «*essentially between non-linguistic (i.e. external) criteria and linguistic (i.e. internal) criteria*» (Pearson, 1998:52). Bowker e Pearson, não fazendo distinção entre critérios internos ou externos, apresentam uma lista de «*criteria required to design a useful special purpose corpus*» (Bowker & Pearson, 2002:45) que em grande parte coincide com os critérios de Pearson e também com os «*four aspects of corpus design*» em (Hunston, 2002:25). Como Tognini-Bonelli afirma, não é apenas importante que o linguista recolha textos, mas que observe o respeito por critérios que justifiquem a relevância linguística, pois

> «*a corpus cannot be equated with just a large collection of texts or citations, but needs to be justified in linguistic terms*» (Tognini-Bonelli, 2001:55).

A combinação de critérios internos, que dizem respeito ao conteúdo linguístico, e externos, que integram os elementos extra-linguísticos, produzirá o filtro para separar os textos que se adaptam aos seus objectivos. De entre os critérios externos, também designados socioculturais por Sinclair e Ball (1995:15), podem-se destacar o género (já discutido em 2.4), o modo, a forma, a data de publicação, o autor, a língua, a origem, os participantes, o enquadramento social e os objectivos dos textos

recolhidos. Normalmente, os critérios externos são os primeiros a ser definidos e aplicados para se proceder a uma triagem preliminar. Como Atkins, Clear e Ostler apontam,

> «*The initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily [...] A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation*» (Atkins et al., 2000:5).

Os critérios internos, tais como o registo, o estilo, o vocabulário e outras estruturas marcadamente linguísticas, definem a tipologia textual e o domínio a que o texto pode pertencer. A dicotomia dos critérios parece, nas palavras de Lee, encontrar no género e no registo uma forma adequada de comparar as duas perspectivas:

> «*I contend that it is useful to see the two terms genre and register as really two different angles or points of view, with register being used when we are talking about lexico-grammatical and discoursal-semantic patterns associated with situations (i.e., linguistic patterns), and genre being used when we are talking about memberships of culturally-recognisable categories*» (Lee, 2001:46).

Sinclair (2003) destaca a importância do relatório do EAGLES (*Expert Advisory Group for Language Engineering Standards*), que define vários critérios externos e internos, onde se afirma que

> «*The typology can be elaborated for the requirements of a particular application. Entries which were not made in the original establishment of the corpus can be added, and additional parameters can be introduced alongside those advocated here*» (EAGLES:1996).

Destacam-se, de entre os já referidos, alguns critérios importantes, apresentados agora de uma forma mais detalhada para que se possam esclarecer alguns pontos importantes.

3.3.4.1 Forma

Uma das considerações fundamentais que terá de se fazer, diz respeito à importância da oralidade e da escrita no estudo. Se uma delas apresentar uma maior relevância no domínio em estudo, a recolha terá de ser ajustada à forma mais adequada, ainda que a disponibilidade também deva ser tida em conta. Como afirma Bowker,

> «*The decision about whether you want to compile a written, a spoken or a mixed-medium corpus will again depend on what you want to study*» (Bowker & Pearson, 2002:50).

Mais questões sobre a forma do *corpus* são desenvolvidas mais aprofundadamente na tipologia de *corpus*.

3.3.4.2 Data de publicação

O período de tempo a cobrir pelo *corpus* também é um elemento importante, pois a decisão de optar por um estudo de carácter sincrónico ou diacrónico acarreta acções obrigatórias na selecção ou organização dos dados e resulta da adequação ao objectivo. Como afirmam Bowker e Pearson:

> «*the age of the texts that you include in your corpus will depend on what you hope to learn from your corpus*» (Bowker & Pearson, 2002:52).

A imposição de barreiras cronológicas nos textos vai forçar, por exemplo, o investigador a verificar as datas de publicação de cada um dos textos recolhidos e a ordená-los temporalmente. Num estudo baseado em critérios cronológicos, quando se inicia a fase da análise, podem-se estabelecer relações temporais de alteração, de inovação ou de desaparecimento, que o factor tempo permite identificar. Da mesma forma, numa investigação linguística que implicasse um levantamento terminológico de um qualquer domínio científico recente, seria desprovido de qualquer sentido

consultar textos na base textual cuja data de origem fosse anterior ao aparecimento desse domínio.

3.3.4.3 Autor

O critério da autoria dos textos poderá variar em grau de importância, estando dependente, no entanto, do objectivo que o estudo se propõe atingir. Se a investigação for sobre um determinado autor, o critério da autoria será um dos primeiros a ser aplicado na selecção. Noutros casos, poderá ser que, pela especificidade do domínio, o estudo obrigue ao uso de autores com créditos reconhecidos na área para que a validade seja reconhecida. Como refere Pearson:

> «*only "acknowledged" authors will be eligible for consideration*» (Pearson, 1998:60).

Ainda que a investigação não se centre no autor, a sua influência nos textos produzidos é, de tal modo, importante que nunca se poderá descartar a informação da autoria como irrelevante. No âmbito da Terminologia Textual, é essencial avaliar a questão da autoria, pois o texto, tal como apresentado em 2.4, é tido como um veículo de escolhas linguísticas que contém em si motivações contextuais. Assim, para melhor conhecer os factores que determinam o acto linguístico, é incontornável conhecer o autor.

3.3.4.4 Língua

Por defeito o *corpus* é compilado na mesma língua, mas é importante apresentar alguma situações em que o não é. A língua torna-se um critério com mais peso na compilação do *corpus* quando, por questões de tradução ou de análise interlinguística, é necessário proceder a uma mistura de textos em diferentes línguas. Por exemplo, no ensino de línguas para estrangeiros, também é normal encontrar situações de compilação de *corpora* com objectivos pedagógicos que passam pela comparação dos textos em momentos cronológicos distintos, ou seja, no caso, em

diferentes fases da aprendizagem. Ainda se podem destacar os casos óbvios, como no caso dos *corpora* paralelos, em que o objectivo passa por comparar as diferentes versões dos textos em várias línguas, ou ainda em *corpora* multilingues. O critério de análise dos textos não se limita, no entanto, a uma perspectiva interlinguística, trabalhando a Terminologia Textual também numa perspectiva intralinguística. Os estudos desenvolvidos para análise das circulações terminológicas dentro dos vários níveis das especialidades obrigam à observação e imposição de critérios linguísticos que possibilitem a investigação dos diferentes graus de especialização da informação que está nos textos. No quadro epistemológico de referência, já vimos que os conceitos são unidades instáveis, tal como os termos, podendo assumir dentro da mesma língua e atravessando os diversos níveis de formalização da transmissão de conhecimento, representações múltiplas.

As diferentes escolhas efectuadas a partir dos critérios aqui listados condicionam a imagem final dos *corpora*, que serão, na sua tipologia, o reflexo dos textos que neles se encontram. Desta forma, apresentam-se e descrevem-se, a seguir, alguns tipos de *corpora* que foram tipificados com vista à melhor organização e classificação dos conjuntos de textos recolhidos.

3.3.5 Tipos de *corpora*

A criação do protocolo de recolha de textos, onde ficam bem claras as regras que presidem ao trabalho terminológico a desenvolver, permite compilar o *corpus* com as características adequadas à investigação, limitando o número de textos seleccionáveis às especificidades dos objectivos da pesquisa. Este protocolo de recolha é um conjunto de premissas estabelecidas e cujo respeito é fundamental para garantir a validade da investigação. Segundo Bowker e Pearson, a multiplicidade e a liberdade de definição dos critérios é fundamental, pois,

> «*there are almost as many different types of corpora as there are types of investigations. Language is so diverse and dynamic that it would be hard*

> *to imagine a single corpus that could be used as a representative sample of all language*» (Bowker & Pearson, 2002:11).

Ainda que seja difícil classificar todos os tipos de *corpora* à disposição, pela diversidade de características que cada *corpus* produzido apresenta, podem estabelecer-se grupos mais genéricos e que revelam traços comuns. Os critérios, já explicitados no ponto anterior, permitem, assim, estruturar tipologias para os *corpora*. Como referem Bowker e Pearson,

> «*it is still possible to identify some broad categories that can be compiled on the basis of different criteria in order to meet different aims*» (Bowker & Pearson, 2002:11).

A proposta que se apresenta de seguida expõe uma divisão tipológica dos *corpora* alicerçada em propriedades comuns decorrentes dos possíveis critérios e que permite uma organização por conjuntos. A divisão tripartida apresentada dá destaque ao suporte em que são recolhidos os textos, o conteúdo que os caracteriza e a forma de apresentação dos *corpora*. Esta organização surge como um nível de pré-estruturação tipológica dos *corpora*, dado que cada critério poderia, como vimos, dar ele próprio origem à tipificação de um conjunto organizado de textos, logo a um *corpus*.

3.3.5.1 Suporte

O suporte designa, neste contexto, o meio de perpetuação do texto, ou seja, qualquer material, electrónico ou não, que retenha informação textual disponível para consulta e análise. Os suportes mais comuns para produtos linguísticos, podendo vir em forma escrita e/ou oral, podem transportar as mais variadas tipologias textuais, desde jornais, livros, blogues e correio electrónico, até programas de televisão ou de rádio e *podcasts*. Ainda que o suporte digital esteja cada vez mais implantado, principalmente por questões de facilidade e velocidade de comunicação,

simplicidade de armazenamento e poupança de espaço físico, o papel ainda é o suporte preferencial, mas nem sempre exclusivo, para documentos oficiais e técnicos, literatura e imprensa escrita, sendo, por isso, obrigatório contemplar no *software* métodos que prevejam o recurso a esse suporte. No entanto, já é raro encontrar textos manuscritos que não sejam apontamentos privados, pois mesmo o que está em papel, a maior parte das vezes, tem origem num processador de texto, tendo sido posteriormente impresso. Como Sinclair aponta,

> «*although still a lot of writing is originated with a pen or pencil and paper, very little of it survives unless it is transposed into a more formal mode, and there is hardly any representation of hand-written material in corpora*» (Sinclair, 2003:175).

A heterogeneidade dos suportes existentes dificulta o trabalho de harmonização do *corpus* para uma leitura informatizada. Na investigação linguística em Terminologia Textual, quando os textos não estão informatizados, é necessário convertê-los através da leitura óptica, a qual, ainda que demorada e susceptível ao erro, é, nos textos com dimensões consideráveis, mais rápida que a digitação integral. Depois de todos os elementos serem informatizados, o processamento automático e estatístico de cruzamento e consulta de dados textuais pode ser efectuado e produzir resultados de uma forma mais célere. O facto de todos os textos estarem disponíveis em suporte digital permite, ainda, que o processo de adição de novos textos ao *corpus* possa ser realizado com frequência e testado, sem que daí surja um acréscimo de trabalho significativo. O suporte electrónico revolucionou não só o processo de trabalho, mas também as condições de acesso ao *corpus*. A facilidade de circulação de informação em suporte informatizado permite a partilha de dados entre investigadores distanciados geograficamente uns dos outros e até a simplificação da deslocação física do centro de investigação, em caso de necessidade. A quase obrigatória informatização dos textos levou ao crescente aparecimento de tipos de *corpora* que estão ligados à Informática, destacando-se alguns de seguida. Ainda que partilhem

algumas das características, importa distinguir cada um deles por serem portadores de especificidades e obrigarem, por isso, à utilização de metodologias de recolha diferentes. Dado que o nosso projecto trabalha directamente com *corpora* em suporte electrónico, submete-se uma proposta de subdivisão tipológica que é fundamental executar, procedendo a uma descrição mais cuidada, para estabelecer as singularidades assumidas para cada um deles e a diferenciação dos métodos de abordagem à sua constituição.

- Os *corpora em suporte electrónico* designam o conjunto de textos que se encontra em suporte digital, quer tenham tido ou não origem informática, com o objectivo de facilitar o estudo da língua com base em estudos estatísticos e no cruzamento de informação textual. Neste contexto, incluem-se todos os textos que estão introduzidos no computador e em condições de serem processados para análise linguística.

- Os *e-corpora* designam os textos que têm origem no suporte digital, não necessitando, por isso, de sofrer um processo de conversão electrónica. Podem, no entanto, ser submetidos a um processo de harmonização informática com o objectivo de uniformizar as características digitais dos textos e facilitar o processamento electrónico. O recurso ao *e-corpus* facilita a recolha de textos na fase de informatização, por já se encontrar no suporte final necessário à análise linguística rápida e em larga escala. Podem enquadrar-se neste tipo de texto as dissertações, artigos ou publicações que são distribuídas já em formato digital e que evitam o processo de leitura óptica ou digitação.

- Os *corpora informatizados* são uma recolha de textos que, na sua origem, não se encontravam em suporte digital e que sofreram um processo de conversão para o suporte electrónico. Em muitos domínios, ainda é possível que grande

parte dos textos disponíveis para recolha se encontrem em suporte de papel, tornando-se inevitável, para poder recorrer a ferramentas de processamento automático, a conversão para suporte electrónico.

- Os *corpora web* são uma recolha de textos alojados na Internet e que, por isso, se encontram obrigatoriamente em suporte electrónico. O facto de serem recolhidos na Internet confere-lhes uma especificidade própria e uma atenção diferente por parte do investigador, devido ao facto de possivelmente serem textos já anotados e de ser necessário proceder à limpeza, processo este explicado mais adiante no subcapítulo Informatização. As razões de usar a Internet como fonte de textos para o *corpus* decorrem da vastidão de textos, da gratuitidade de utilização e da velocidade de recolha. O interesse crescente pelos *corpora* web desenvolve-se também graças a outros factores, como Baroni e Ueyama demonstram:

  > «*For these reasons (lack of resources in language of interest; data sparseness problems; need to study sub-languages or recent usages), researchers have been increasingly interested in the Web as a potential source of linguistic data*» (Baroni & Ueyama, 2006:1).

  A Internet, enquanto meio privilegiado de comunicação e de transmissão de informação, poderá, assim, servir como pólo congregador de recursos textuais de especialidade, criando aos poucos um repositório centralizado de fácil consulta que serve como base textual, mas que não é um *corpus* até serem aplicados os critérios necessários. Os *corpora* web, ainda que proporcionem um acesso fácil a uma fonte de recursos textuais enorme e actualizada, levantam, contudo, muitas reservas no que diz respeito à ausência de validação da informação textual neles contida.

3.3.5.2 Conteúdo

A definição de conteúdo apresenta alguma complexidade pelas várias acepções que podemos encontrar, tais como informação, dados ou conhecimento, ainda que apenas quando enquadrados num texto. O conteúdo, de acordo com Budin, diverge de todos os outros sentidos apresentados pela seguinte razão:

> «*When knowledge is then packaged as a product for a certain audience, presented in certain media presentation forms, then we can speak about content*» (Budin, 2002:57).

Budin relembra, ainda, que o conteúdo e a forma estão relacionados e que se influenciam mutuamente:

> «*the form of representing content and the medium chosen to do this is constitutive for distinguishing types of content*» (Budin, 2002:59).

Mesmo os textos digitais, quando não sujeitos a pré-tratamento, como, por exemplo, os existentes da Internet, são, em grande parte, muito difíceis de organizar e integrar num *corpus*. Há várias razões que propiciam este facto, como, por exemplo, a falta de regras de publicação, a não existência de limitações à sua edição, reprodução, adulteração e divulgação, a falta de critérios de selecção e de correcção linguística uniformizados ou a carência de imposições estruturais e temáticas. Os critérios que estão ligados ao conteúdo são os mais numerosos e, por isso, encontramos na respectiva categoria mais tipos de *corpora*. O motivo que poderá explicar tal situação encontra-se, provavelmente, no cariz linguístico que a recolha do *corpus* assume, sendo, assim, de esperar que os critérios linguísticos ou que influenciam a produção linguística predominem no *corpus*. De entre os vários tipos de *corpora*, destacam-se os mais comuns e apresentam-se sumariamente as características distintivas, os objectivos e as utilizações mais frequentes:

- Os *corpora gerais* são recolhas de textos de uma língua geral, que têm como objectivo a representação da língua corrente e são normalmente usados para

proceder a estudos no âmbito de dicionários e de gramáticas da língua. Partilham semelhanças com os *corpora de referência*, ainda que estes sejam representativos de um número inferior de variantes linguísticas, incluindo apenas as mais relevantes.

> «*Reference corpora contain the standard vocabulary of a language. They are a linguist's main resource to learn about meaning*» (Halliday & Teubert, 2004:118).

- Os *corpora de especialidade* são recolhas de textos filtradas por critérios conceptuais, concentram-se num domínio restrito e servem para proceder à análise linguística do sistema de noções de uma parte específica do conhecimento.

> «*It [corpus de especialidade] aims to be representative of a given type of text. It is used to investigate a particular type of language*» (Hunston, 2002:14).

- Os *corpora comparáveis* são recolhas de textos que se apresentam como semelhantes, numa perspectiva monolingue, com o objectivo de procurar marcas de igualdade ou de diferença entre si, podendo, inclusive, permitir a descoberta de informação estrutural e linguística comum, dependente do contexto de produção.

> «*Comparable corpora are corpora whose components are chosen to be similar samples of their respective languages in terms of external criteria*» (Tognini-Bonelli, 2001:7).

- Os *corpora paralelos* são recolhas de textos que integram traduções de si mesmos, bilingues ou multilingues, com o objectivo de procurar analogias linguísticas entre si e reconhecer padrões de utilização da língua.

> «*A parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original*» (EAGLES:1996).

- Os *corpora de tradução* são uma especificação dos *corpora* paralelos, pois, apresentam, também eles, textos comparáveis, mas que, neste caso, têm como objectivo estudar os processos de tradução. Este tipo de *corpora* proporciona a identificação de paralelismos estruturais ou linguísticos que, eventualmente, servem para criar memórias de tradução, úteis para reutilização futura pelos tradutores.

  «*Translation corpora are corpora of texts which stand in a translational relationship to each other*» (Tognini-Bonelli, 2001:6).

- Os *corpora de monitorização* são conjuntos de textos recolhidos e actualizados num determinado espaço de tempo, com objectivos de detecção, de acompanhamento e de análise de alterações linguísticas. Este tipo de *corpus* tem uma vertente diacrónica muito marcante, pois, com a crescente capacidade de armazenar digitalmente textos, tornou-se possível manter os textos mais antigos e monitorizar linguisticamente um determinado grupo, partindo de conjuntos de textos recolhidos com constituição semelhante, mas em tempos diferentes. Os *corpora* de monitorização partilham características semelhantes com os de referência por investigarem a língua corrente.

  «*[O corpus de monitorização usa-se] to track current changes in a language*» (Hunston, 2002:16).

3.3.5.3 Forma

A forma de apresentação designa o conjunto de sistemas gráficos ou sonoros capazes de reproduzir textos analisáveis linguisticamente. Na recolha de elementos linguísticos reais para o *corpus*, são usados dois formatos tradicionais: o escrito ou o oral. Como refere Hunston,

«*It [corpus] may include written or spoken language, or both*» (Hunston, 2002:13).

Apesar de ser mais fácil, hoje em dia, encontrar e recolher *corpora* orais, subsiste uma predominância de *corpora* escritos por os textos escritos estarem disponíveis em maior número e serem mais facilmente trabalhados para análise linguística em termos informáticos. O uso de *corpora* orais, no entanto, torna-se excepção obrigatória nos casos em que a oralidade é indispensável à investigação ou objectivo primeiro do trabalho. Para facilitar o processo de trabalho estatístico e de pesquisa com o *corpus* oral pode-se recorrer à sua transcrição, mantendo-se, no entanto, a gravação original para ser possível combinar os diferentes aspectos de análise. Nos *corpora* orais encontram-se aspectos prosódicos, como sejam as repetições, os bordões, as variações de pronúncia ou, até mesmo, as truncações ao nível do discurso, que podem ser relevantes para a compreensão de determinados fenómenos linguísticos. Tanto os *corpora* orais como os escritos são modernamente, por prática estabelecida, recolhidos ou convertidos para suporte digital de modo a facilitar o tratamento informático.

3.4 Gestão de *corpora*

Depois de recolhidos os textos que integram o *corpus*, é necessário passar a outra fase de equivalente importância. A gestão dos elementos existentes no *corpus* é uma multitarefa, pelo seu carácter plural, mas coeso, que se tem de realizar com o objectivo de rentabilizar os textos recolhidos.

3.4.1 Informatização

Na Linguística dos nossos dias consideramos impraticável a investigação que não recorra ao auxílio da Informática. Tanto a linguística introspectiva, como a linguística de *corpus*, que se apresentam como duas vertentes epistemológicas com percursos metodológicos diferentes, ainda que em fases diferentes e com objectivos diversos, projectam a investigação com recurso a procedimentos automáticos. A informatização, não só pela capacidade que introduz de processar muitos dados, mas

também pela possibilidade de reutilização ilimitada e imediata, permite trabalhar diferentes cenários com um esforço reduzido, principalmente se comparado com o contexto pré-informático. Como Kennedy refere,

> «*Corpus Linguistics is thus now inextricably linked to the computer, which has introduced incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data*» (Kennedy, 1998:5).

Mesmo quando o *corpus* é diminuto e pode ser trabalhado manualmente, o facto de se proceder à informatização permite reutilizá-lo e preservá-lo para que o material recolhido não se dê como perdido depois da conclusão da investigação. No entanto, a informatização dos textos nem sempre é um processo linear e de fácil concretização. Nas duas hipóteses possíveis de suporte para o texto, leia-se informatizado e não-informatizado, nenhuma delas exclui à partida a verificação das condições de informatização a que, respectivamente, já foi ou vai ser submetida. Se, na primeira, é obrigatório proceder a uma verificação do grau de limpeza do ficheiro que serve de suporte ao texto, ou seja, o nível de preparação para processamento automático, na segunda hipótese, existe a possibilidade de implementar critérios informáticos ainda durante o processo inicial de digitação ou leitura óptica. Entendemos por limpeza do ficheiro, o processo que decorre durante a fase da informatização dos textos e que remove todos os elementos supérfluos para a investigação e que complicam a leitura automática pelos sistemas informáticos, como sejam códigos de programação remanescentes, informação sobre a estruturação textual, gráficos, imagens, tabelas ou outros dados acessórios (pessoais e identificativos). Em alternativa, poderão acomodar-se todos esses elementos dispensáveis ao contexto da investigação, através de um processo de harmonização, anotando-os e tornando consciente ao sistema de que eles existem, mas devem ser ignorados neste contexto específico. A possibilidade de recolher textos de condições variadas origina que, por vezes, os textos dos *corpora* apresentem características bastante díspares ao nível da formatação

e codificação e seja necessário intervir, harmonizando-os igualmente. Como referem Habert, Nazarenko e Salem,

> «*La phase initiale de "nettoyage" et d'homogénéisation des textes collectés sous forme électronique est une étape souvent sous-estimée, alors qu'elle est cruciale*» (Habert et al., 1997:161).

Consideramos o processo de harmonização um procedimento de homogeneização das propriedades digitais dos textos, com o objectivo de possibilitar a automatização do trabalho estatístico e de análise que o sistema tem capacidade de executar. Como avança Kennedy,

> «*anyone compiling a corpus which consists of electronic versions of texts taken from many different sources soon learns that inconsistent methods of encoding the text and signposting the different parts of the text can cause confusion*» (Kennedy, 1998:82).

Sendo essencial proceder-se a uma harmonização digital do texto, como foi referido, é igualmente importante que se mantenham as propriedades textuais inicialmente encontradas pelo investigador. A questão torna-se mais relevante pelo exponencial crescimento de utilização da Internet e do seu repositório textual gigantesco, o qual é desprovido de mecanismos de controlo, na sua maioria, no que concerne a regras de produção, como base textual de arranque para a compilação de *corpus*. O computador e os programas informáticos são ferramentas indispensáveis à gestão avançada de conteúdos linguísticos, que se apresentam irregulares nas propriedades, demonstrando serem os únicos instrumentos com capacidade para acompanhar de forma equivalente a variabilidade textual.

Aceder à base textual da Internet, uma fonte de recursos crescente, múltipla e renovável, com o objectivo de recolher textos para um *corpus*, é fácil e rápido, ainda que a elevada disponibilidade seja, quase, proporcionalmente desorganizada. Mesmo sendo, na sua maioria, de difícil organização, os textos recolhidos da Internet contêm

algumas das características mais importantes para os investigadores que efectuam estudos com dados terminológicos, como sejam a actualidade e a disponibilidade. Como afirma Castagnoli,

> «*[...]it is possible to find on the Internet texts on virtually any specialized subject, written in a variety of genres and communicative settings, [...] new documents appear or are updated on the Web on a daily basis [...] Lastly, [...] the fact that Web access is becoming increasingly easier and inexpensive, and that it is constantly available*» (Castagnoli, 2006:160-161).

Kilgarriff acrescenta ainda que

> «*The initial-entry cost for this kind of research is zero. Given a computer and a web connection, you input the query and get a hit count. But if the work is to proceed beyond the anecdotal a range of issues must be addressed*» (Kilgarriff, 2007:1).

Como Kilgarriff explicita na parte final da anterior citação, para se conseguir uma investigação séria, não basta fazer uma simples pesquisa. A utilização e consulta da Internet, para que seja levada a cabo de uma forma rápida e representativa do acervo digital, depende de motores de busca sobre os quais pouco se conhece, no que diz respeito aos critérios efectivos de pesquisa. O exemplo mais recorrente é o do Google, que usa critérios de popularidade e algoritmos próprios de hierarquização dos conteúdos que não são explícitos, nem conhecidos pelo utilizador, nem exclusivamente linguísticos. Como afirmam Bourigault e Jacquemin,

> «*Les limites des moteurs de recherche sur le Web sont patentes : beaucoup des réponses retournées par ces systèmes sont jugées non intéressantes par l'utilisateur, alors que des réponses pertinentes ne sont pas proposées. Améliorer les performances de ces systèmes, en particulier pour des applications de veille technologique ou de recherche*

d'information dans des domaines spécialisés, devient un enjeu économique énorme» (Bourigault & Jacquemin, 2000).

A utilização do motor de busca, não sendo de todo consensual, é, contudo, a única forma de conseguir organizar e hierarquizar o que seria impossível de outra forma. Reconhecendo-se que o processo está contaminado à partida, pois o texto recolhido vem "pré-organizado", só existem duas formas de contornar esta questão: conhecer os critérios utilizados nessa escolha, para que se possam anular, ou construir um software que contorne os critérios pré-definidos e que constitua uma alternativa. A segunda das duas hipóteses foi uma das razões que conduziu à criação do protótipo de software que se apresenta na dissertação.

3.4.2 Classificação e anotação

Depois de ser informatizado e de passar pelo processo de homogeneização, o texto terá de ser classificado e anotado. Os processos de classificação e anotação são fundamentais para que, durante a análise, as consultas efectuadas ao *corpus* sejam fáceis e rápidas. A classificação é executada de acordo com os critérios definidos e serve para seleccionar os textos que encaixam no perfil do *corpus* e filtrá-los da base textual. O processo de classificação textual pode assentar num ou em vários critérios, como, por exemplo, a tipologia textual ou o domínio a que se presume pertencer o texto, estando decorrente essa classificação dos objectivos definidos inicialmente e que se pretendem atingir. A organização e classificação dos textos permite proceder à sua catalogação, seja por um ou vários dos critérios definidos, como, por exemplo, a autoria, a origem ou a data de origem. Como refere Kennedy,

«*In addition to the storage and cataloguing of texts and their electronic version on computer, it is normally essential to plan to collect and catalogue has much information as possible about the authorship or source of texts*» (Kennedy, 1998:76).

Quando se encontrar devidamente catalogado e inserido no respectivo grupo de classificação, poderá ainda proceder-se a uma organização interna do *corpus*, também mediante a aplicação de critérios, como seja, por exemplo, por número ou multiplicidade de termos do domínio. Esta organização permite que os textos mais ricos em contextos e informação terminológicas estejam mais rapidamente acessíveis.

Depois, procede-se à anotação dos textos, para que se possa manter informação descritiva necessária à concretização do estudo. A anotação inclui informação linguística, dependendo, mais uma vez, do objectivo a que se propõe o investigador. Como afirma Rute Costa,

> «*anotar um* corpus *significa associar informação linguística a segmentos de texto, recorrendo para o efeito a um conjunto de símbolos, as etiquetas, por forma a poder identificá-los, com vista aos seu tratamento automático. Esta operação é designada de etiquetagem, constituindo o produto final um* corpus *etiquetado*» (Costa, 2001:38).

Esta anotação, que usaremos como sinónimo de etiquetagem, permite anexar ao *corpus* também informação extralinguística que complementará os dados linguísticos anotados no texto. Por questões de reutilização, quanto maior for a quantidade e diversidade de informação anotada, maior é a probabilidade do *corpus* ser útil para outras investigações. Pode guardar-se, juntamente com o texto, qualquer informação que se considere relevante, nomeadamente etiquetas que conservem informação sobre as condições de produção ou de compilação do *corpus*, sobre os textos presentes no *corpus* ou sobre os dados presentes no texto. Aspectos linguísticos, sociolinguísticos, pragmáticos, diacrónicos e até culturais podem ficar registados, juntamente com o texto recolhido, para que depois se recorra aos instrumentos automáticos de análise. Como afirma Kennedy,

> «*The level of detail of markup has to be related to the potential use of the corpus*» (Kennedy, 1998:84).

Contudo, para que os sistemas informáticos conseguissem proceder à leitura dos dados da mesma forma que um humano, foi necessário criar formatos de gravação digital que permitissem o registo de anotações interpretáveis pelos computadores. Assim, foi estudado um sistema de anotação (*tagging*) e de processamento automático (*parsing*), com o propósito de anexar informação adicional ao texto e facilitar o intercâmbio de dados entre sistemas, que culminou com o aparecimento em 1986, de acordo com a norma ISO 8879, do formato SGML (Standard Generalized Markup Language). A falta de apoio em larga escala na divulgação e implementação do complexo SGML por parte da indústria informática, levou a maior parte das aplicações que necessitavam de utilizar a Internet a converter a informação para um formato mais compatível, o HTML (HyperText Markup Language). O HTML, apesar de mais simples, era limitado, tendo menos flexibilidade no tipo de informação que se podia anexar, apenas permitindo a adição de informação sobre a apresentação do texto. Mesmo sendo possível proceder a uma conversão de SGML para HTML, levantaram-se obstáculos à reutilização, à permuta de dados e ao processo de automatização, pois o código final em HTML não possuía as mesmas funcionalidades que o SGML original. Estas questões conduziram ao aparecimento de um novo formato em 1996 que conseguisse manter a mesma flexibilidade e capacidade do SGML e a facilidade de integração do HTML. É, assim, que surge o XML (Extensible Markup Language) derivado do SGML, ainda que numa versão simplificada, desenvolvido para guardar, transportar e trocar dados, num formato semelhante ao SGML, na Internet. A principal vantagem do XML reside em ser uma metalinguagem mais básica do que o SGML, podendo o utilizador construir a sua própria linguagem de anotação e descrever o conteúdo do texto. O SGML e o XML são os standards mais usados para anotação e partilha de dados, contribuindo para que a Internet se transforme no que o consórcio mundial de Internet W3C chama um «*universal medium for the exchange of data*».

O XML é composto por três partes lógicas: a primeira parte, o prólogo, subdivide-se na declaração XML, que define a versão XML usada, e em informa-

ções facultativas sobre instruções de tratamento para aplicações específicas, como, por exemplo, a codificação de caracteres usada no documento, a segunda parte, integra a definição do tipo de documento (DTD), e a terceira parte, é constituído pelo documento e os seus elementos.

As DTDs são formas de se descrever classes de documentos XML, tendo um objectivo semelhante ao das gramáticas nas línguas naturais. Segundo Rute Costa,

> «*Todas as regras que definem os tipos de elementos próprios a um documento estão contidas numa DTD. Assim, a DTD define as regras de balizagem de um documento ou de uma classe de documentos, permitindo a descrição da sua estrutura lógica hierarquizada*» (Costa, 2001:43).

De acordo com o TEI (*Text Encoding Iniative*), que é um consórcio composto por várias entidades e projectos espalhados pelo mundo, com o objectivo de criar um «*international and interdisciplinary standard that enables libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation*»[4], o XML passou a ser, a partir da versão *P4* das suas directivas lançadas em 2002, que tem como última versão a *P5*, mas que ainda se encontra em fase de desenvolvimento, recomendado como linguagem mais adequada para a troca e registo de dados em formato electrónico. Como se pode ler nas suas orientações,

> «*The Text Encoding Initiative (TEI) Guidelines are addressed to anyone who wants to interchange information stored in an electronic form. They emphasize the interchange of textual information, but other forms of information such as images and sound are also addressed. [...] The Guidelines provide a means of making explicit certain features of a text in such a way as to aid the processing of that text by computer programs running on different machines*» (Burnard & Sperberg-McQueen, 2002:i).

[4] `http://www.tei-c.org/index.xml`

O TEI apresenta vários "esquemas" que servem de modelo para serem utilizados em diversas situações e que proporcionam uma plataforma comum de anotação dos textos, permitindo a reutilização dos dados em diferentes contextos, como se pode ler na sua documentação:

> «*The scheme documented here can be used to encode a wide variety of commonly encountered textual features, in such a way as to maximize the usability of electronic transcriptions and to facilitate their interchange among scholars using different computer systems*» (Burnard & Sperberg-McQueen, 2002:1).

Da mesma forma, o XCES (XML *Corpus* Encoding Standard) tem vindo a ser desenvolvido a partir do CES (*Corpus* Encoding Standard), criado pelo *Expert Advisory Group on Language Engineering Standards* (EAGLES) em conformidade com as orientações do TEI, mas com o objectivo de criação de standards para a anotação dos *corpora*. Na conversão do CES para a norma do XCES, os "esquemas" mantêm-se iguais, em grande parte, sendo a substituição do SGML pelo XML a principal alteração a notar, pois, segundo se afirma, «*the XML framework provides us with means to go well beyond the capabilities of SGML*»[5].

Assim, através da utilização dos standards definidos pelo TEI para anotar listas, podemos apresentar uma demonstração simples de anotação realizada sobre o índice do terceiro capítulo desta dissertação:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Índice do terceiro capítulo -->
<list type="gloss">
 <head>Definição de <it>corpus</it></head>
  <label>1</label><item>Introdução</item>
  <label>2</label><item>A linguística e o <it>corpus</it></item>
  <label>3</label><item>Constituição de <it>corpora</it></item>
    <label>3.1</label><item>Definição de um objectivo</item>
    <label>3.2</label><item>Domínio</item>
    <label>3.3</label><item>Homogeneidade, representatividade e exaustividade</item>
```

[5] http://www.cs.vassar.edu/XCES/

```
    <label>3.4</label><item>Critérios para classificação de <it>corpora</it> em Termi
    nologia Textual</item>
      <label>3.4.1</label><item>Forma</item>
      <label>3.4.2</label><item>Data de publicação</item>
      <label>3.4.3</label><item>Autor</item>
      <label>3.4.4</label><item>Língua</item>
    <label>3.5</label><item>Tipos de <it>corpora</it></item>
      <label>3.5.1</label><item>Suporte</item>
      <label>3.5.2</label><item>Conteúdo</item>
      <label>3.5.3</label><item>Forma (de apresentação)</item>
  <label>4</label><item>Gestão de <it>corpora</it></item>
    <label>4.1</label><item>Informatização</item>
    <label>4.2</label><item>Classificação e anotação</item>
    <label>4.3</label><item>Actualização e reutilização</item>
  <label>5</label><item>Síntese</item>
</list>
```

A etiqueta <list>, segundo consta no capítulo 12 do manual do TEI, serve para marcar qualquer tipo de lista, definindo-se lista como «*a sequence of text items, which may be ordered, unordered, or a glossary list*». (Burnard & Sperberg-McQueen, 2002:27). Para que o código XML seja processado correctamente é obrigatório respeitar algumas regras, agora apresentadas:

- Todos os elementos XML têm de ter uma etiqueta de abertura e de fecho sinalizadas pelos parênteses angulares:

  ```
  <etiqueta>...</etiqueta>
  ```

- Os elementos XML nas etiquetas são *case sensitive*, ou seja, usar maiúsculas ou minúsculas na etiquetas faz diferença.

- Os elementos XML têm de estar correctamente dispostos:

  ```
  <etiqueta1><etiqueta2>...</etiqueta2></etiqueta1>
  ```

- Os elementos XML têm de ter sempre, pelo menos, uma etiqueta, que é o elemento raiz.

- Os atributos dos elementos XML têm de ser colocados entre aspas:

  ```
  <list type="gloss">...</list>
  ```

- O comentários em XML têm formatação especial:

```
<!-- Isto é um comentário -->
```

Podemos ainda usar o XML para criar a nossa própria informação, fora das definições do TEI, o que demonstra a flexibilidade do código, ainda que, ao escapar ao standard, corre-se o risco de perder compatibilidade com outro software, mesmo que o documento continue a ser um XML válido. Veja-se o exemplo apresentado de uma receita de sobremesa e de como a informação fica estruturada:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<receitas>
<sobremesas nome="bolos">
<titulo>Bolo de Coco</titulo>
<ingredientes>
<ingrediente quantidade="3">Ovos</ingrediente>
<ingrediente quantidade="200" unidade="gramas" tipo="ralado">Coco</ingrediente>
<ingrediente quantidade="150" unidade="gramas" tipo="granulado">Açúcar</ingrediente>
</ingredientes>
<instrucoes>
<passo>Colocam-se os três ingredientes num recipiente.</passo>
<passo>Mistura-se tudo com as mãos.</passo>
<passo>Distribui-se em pequenas formas de papel.</passo>
<passo>Leva-se ao forno pré-aquecido a 180º por 15 a 20 minutos</passo>
</instrucoes>
</sobremesas>
</receitas>
```

No processo de anotação, podem ser misturados os elementos standard do TEI com outros não previstos, não existindo qualquer tipo de incompatibilidade.

Os sistemas de anotação desenvolvidos conduziram à simplificação dos processos de contabilização morfológica e de identificação de padrões gramaticais, evoluindo depois para a área da representação conceptual com as BCTs. No entanto, ainda existe uma larga margem de progressão no tipo de interacção que pode estabelecer com o *corpus*, dado que, tal como aponta Kennedy, «*work in corpus analysis is still very much work in progress*» (Kennedy, 1998:206).

O XML foi adoptado como linguagem informática para anotação do protótipo de software que apresentamos pelos vários motivos já elencados. A sua adopção

pelo TEI como standard, o facto de ser transversal aos sistemas informáticos e a flexibilidade que demonstra conjugam-se como factores decisivos para a escolha do XML como linguagem eleita para proceder à formalização dos dados recolhidos. A flexibilidade característica permite também efectuar o processo de actualização dos dados terminológicos de uma forma mais simples e eficaz, promovendo a reutilização da informação, como podemos constatar de seguida.

3.4.3 Actualização e reutilização

A actualização e a reutilização do *corpus* são dois processos distintos, mas que estão interligados nos objectivos que pressupõem – a rentabilização e a extensão do período de validade do *corpus* e dos dados – e nos procedimentos necessários. A actualização diz respeito à renovação dos textos que compõem o *corpus* e, consequentemente, dos dados terminológicos disponibilizados, enquanto a reutilização aponta para a flexibilização do *corpus* para fins que não os determinados aquando da sua compilação.

A actualização é uma tarefa contínua, a partir do momento que se pretenda dar seguimento à investigação, exceptuando-se os casos óbvios em que as circunstâncias fixam um período de tempo para a recolha, não sendo, assim, possível, uma vez concluída, alterar a constituição do *corpus*. Em Terminologia, grande parte dos estudos é desenvolvida no âmbito da construção de bases de conhecimento terminológico (BCT), que, como já vimos anteriormente, procuram reproduzir uma formalização conceptual dos domínios, obrigando, por isso, à renovação constante dos textos e dados, já que as fronteiras dos domínios são muito flexíveis e estão constantemente a redefinir-se, tal como os conceitos que os integram. Desta forma, a actualização dos elementos que constituem o *corpus* é fundamental não só para manter a colecção de textos actualizada, mas também para renovar os dados terminológicos. Caso não se proceda à actualização dos textos, muito dificilmente se poderá dar outro uso ao *corpus*, uma vez que ele é concebido para um determinado objectivo e, as-

sim que se atinge a finalidade proposta, o *corpus* deixa de ser, potencialmente, útil, principalmente se for reduzido, específico e aplicado a um domínio de especialidade.

A possibilidade de reutilização do *corpus* deve ser considerada à partida no âmbito da investigação e fazer parte das decisões que se tomam antes da constituição, no entanto, depende, igualmente, do tipo e quantidade de anotações que forem efectuadas a posteriori. Se os dados recolhidos e anotados, linguísticos e não-linguísticos, ultrapassarem o estritamente necessário e requerido para a investigação, há um potencial de reutilização maior do que nos *corpora* que se limitam a ser suficientes para a investigação. Como Bowker e Pearson exemplificam,

> «*If you annotate the corpus, for example, by labelling each text with the date on which it was written, you could compare your early work against your most recent work*» (Bowker & Pearson, 2002:212).

Qualquer *corpus* tem um potencial de reutilização que não pode ser ignorado, mesmo que tenha sido constituído de forma restrita dentro de uma área de especialidade. Como Bourigault e Jacquemin afirmam,

> «*A priori, les ressources terminologiques sont construites pour un domaine donné et pour une application identifiée (section 9.2.2). Si cette prise en compte du caractère "ad hoc" d'une terminologie doit être complètement assumée, elle ne doit pas interdire une réflexion sur la généralité et la réutilisabilité. Il convient d'étudier dans quelle mesure l'acquisition de ressources terminologiques dans un contexte donné peut être facilitée par l'exploitation de ressources lexicales dites "générales", comme la base WordNet (JACQUEMIN, 1999) ou des dictionnaires de synonymes (HAMON, 1998), ou de données terminologiques élaborées pour des domaines proches ou pour d'autres applications sur le même domaine*» (Bourigault & Jacquemin, 2000).

A actualização e a reutilização do *corpus* são, assim, questões importantes a ponderar, pois a perspectiva do *corpus* que é descartável no fim da investigação

parece não fazer mais sentido. Os recursos informáticos disponíveis e a interpretação epistemológica da forma como os dados terminológicos se reconstroem nos textos, atravessando diferentes áreas do conhecimento, conduzem a que as estratégias de reciclagem da informação se sobreponham às de fixação de um produto terminado, logo cristalizado, e dificilmente produtivo daí em diante.

3.5 Síntese

A definição de *corpus* tem assumido contornos diferentes em períodos diversos. Se numa fase anterior anterior ao séc. XX apenas era possível compilar o *corpus* com dimensões reduzidas ou, em alternativa, com recurso a elevados meios humanos e num extenso período de tempo, a tradição viu-se alterada e a automatização de grande parte do processo mudou o conceito, no que diz respeito à dimensão possível e ao tempo de compilação. Com excepção de circunstâncias específicas, ditadas pelo objectivo da investigação, a compilação do *corpus* em Linguística é uma tarefa executada em larga escala e que envolve a consulta de muitos recursos textuais, principalmente nas investigações numa língua geral. Esta capacidade de armazenar e consultar uma quantidade de textos tão elevada influenciou directamente a forma de investigar em Linguística, que a dada altura, atravessando diferendos epistemológicos sérios e irreconciliáveis, considera rever os seus procedimentos e criar alternativas metodológicas condizentes. A introdução do *corpus*, enquanto requisito essencial para aproximar a Linguística da língua verdadeira, predispõe o aparecimento da Linguística de *Corpus* que, por oposição à vertente introspectiva, parte de uma perspectiva descritiva e procura atingir um modelo mais realista da língua com base em textos. No entanto, um banco ou uma colecção de textos não pode ser considerado um *corpus*, essencialmente, por questões de organização, como vimos. A aplicação de critérios é, assim, condição indispensável para distinguir um conjunto de textos de um produto elaborado, complexo e que visa a consecução de uma proposição bem definida, como é o *corpus*. Há todo um processo de preparação, antes da

compilação do *corpus*, onde é fundamental definir o objectivo a atingir pela investigação, para que, depois de escolhido o domínio que melhor se adapta, se seleccionem os critérios mais adequados para o cumprimento desse objectivo. A homogeneidade, representatividade e exaustividade são princípios que regem a compilação do *corpus* e que validam o estudo quanto à relevância e aplicabilidade dos resultados decorrentes. Depois de reunidos os textos do *corpus* é importante proceder a uma gestão dos mesmos, que passa por uma série de procedimentos informáticos de limpeza e harmonização textual. A concepção do protótipo de *software* foi desenvolvida de modo a que os processos de constituição e de gestão de *corpora* respeitem as fases descritas e possibilitem a aplicação dos critérios aqui apresentados, ainda que não caiba ao *software* o processo de decisão sobre a escolha dos mesmos. Não faz parte do objectivo desta dissertação defender a preponderância de algum dos critérios sobre outros, ainda que se estabeleçam quais os critérios a ter em consideração e algumas situações de utilização, pois cada investigação tem as suas particularidades e só o terminólogo, no processo inicial, poderá tomar essa decisão que marcará o cariz de todo o estudo. Um dos objectivos do *e-Termite* é facultar um instrumento de implementação e teste dos critérios para que se possam analisar os resultados obtidos em função do contexto de aplicação. No capítulo seguinte, poderemos confirmar a importância da definição desses critérios de acordo com a base epistemológica que agora terminamos de apresentar, apresentando-se a descrição da concepção do protótipo de *software* e exemplos de funcionamento.

Capítulo 4

Concepção de *Software*

4.1 Introdução

O desenvolvimento do conceito de um protótipo de aplicação informática com as características da que se vai apresentar surge com uma finalidade bem estabelecida: optimizar o processo de constituição e gestão do *corpus* de especialidade. Cada vez mais o número de *corpora* de referência disponíveis aumenta, principalmente com o crescimento contínuo das bases de dados textuais e com a fácil e rápida constituição de *e-corpora* acessíveis pela Internet, sem que, no entanto, surjam tão abundantemente recursos equivalentes para áreas de especialidade.

No âmbito dos estudos linguísticos, há grandes bases de dados textuais, nacionais e internacionais, que fornecem métodos gratuitos e fáceis de consulta. A questão que se levanta, no entanto, é se os *corpora* existentes e pré-compilados se ajustam à consecução dos objectivos pretendidos, tal como foi explicado no capítulo anterior. Apesar de existirem grandes bases de dados com *corpora* textuais, como, por exemplo, o CETEMPúblico[1], com cerca de 180 milhões de palavras extraídas dos diários PÚBLICO editados entre 1991 e 1998, é preciso que o estudo efectuado possa ser realizado com um *corpus* com estas características. Podemos facilmente aceitar que a grande maioria dos estudos linguísticos, no âmbito da língua geral e até

[1] `http://www.linguateca.pt/CETEMPublico/`

corrente, possam interessar-se em usar este *corpus* como referência, mas dificilmente poderá servir, por si próprio, por exemplo, para uma análise da neologia actual.

Tirando o facto de que alguns domínios têm mais procura do que outros, não nos parece adequado afirmar, só para conseguir justificar a escassez de bases de conhecimento terminológico disponíveis, que há uma fraca demanda de recursos terminológicos de especialidade, muito pelo contrário. Talvez pela menor disponibilidade de textos especializados e pelo número inferior de investigadores que se dedica ao trabalho em línguas de especialidade, em comparação com a língua geral, possamos entender, em parte, tamanha discrepância. Acima de tudo, talvez seja importante compreender que, pelas particularidades únicas de cada trabalho terminológico, torna-se mais complicado partilhar o produto da compilação ou reutilizá-lo numa língua técnica, dada a sua especificidade, do que na língua geral, que serve de base aos *corpora* de referência. O processo de renovação conceptual e terminológica nas áreas de especialidade é rápido, principalmente se for em áreas muito activas, decorrendo uma desactualização do *corpus* e dos termos. Se se utilizasse um *corpus* compilado com as datas que o CETEMPúblico apresenta, mas numa área de especialidade, a probabilidade de hoje se encontrar incompleto ou apenas desactualizado é grande, mesmo em áreas que não estejam ligadas às novas tecnologias.

Pelos motivos expostos, para além da optimização do processo de constituição, é importante apostar numa dinâmica de criação de recursos reutilizáveis, ou seja, aprofundar as técnicas de gestão dos *corpora* de especialidade, para que estes possam servir para mais do que um objectivo, como referia atrás Bourigault, ou então que esse objectivo se possa alargar a várias aplicações possíveis. Trata-se de uma questão simultaneamente metodológica e informática que tentaremos abordar no desenvolvimento da proposta de protótipo de *software*, procurando ir ao encontro do nosso objectivo principal, já apontado inicialmente.

As questões da desactualização e da desadequação dos *corpora* já compilados têm conduzido a uma tendência cada vez maior para se procurarem directamente os textos na Internet, o que, como já vimos anteriormente, também levanta alguns

problemas no que diz respeito à sua validação e datação, principalmente porque não se pode confundir uma base textual, como é a Internet, com um *corpus* devidamente constituído através da aplicação de critérios definidos e com um objectivo em vista. A maior disponibilidade de material de trabalho ao alcance do investigador é, *a priori*, um factor positivo, mas que, se não for bem gerido, pode levar a resultados errados por não se respeitarem os passos metodológicos essenciais para uma boa constituição de um *corpus*.

A informática teve o mérito de fomentar a globalização, com os recursos da Internet, e de promover a circulação de informação, inclusive a especializada, passando agora, novamente, pela informática, as soluções que ajudam à gestão da imensa informação que está disponível. A dificuldade em encontrar textos adequados para integrar nos *corpora* em certas áreas do conhecimento pode comportar, por vezes, consequências graves como, por exemplo, a desistência da constituição do *corpus* ou, então, uma decisão precoce de passar para a fase da análise, motivada pelo cansaço ou pela ideia errada de que o *corpus* atingiu uma completude satisfatória. Mesmo que não se desista, por teimosia ou inconsciência, o trabalho que seja desenvolvido em tais circunstâncias será facilmente questionável pela falta de representatividade e de exaustividade resultantes da conclusão precipitada da constituição de *corpus*.

A insuficiência dos métodos manuais, inadequados para a maioria dos trabalhos com *corpora*, é sobejamente conhecida e está em vias de se tornar obsoleta. No entanto, há tarefas que mesmo estando informatizadas não são rápidas o suficiente por não existir uma verdadeira optimização do processo de interacção entre o objectivo, o utilizador e o computador. Veja-se, por exemplo, a forma como os processos de consulta e de edição textual interagem com as bases de dados, sendo necessário proceder a uma série de passos informáticos, quando com uma acção única, como se exemplificará mais tarde, se pode concretizar o processo ou, ainda, quando o terminólogo recorre à Internet como base textual complementar e encontra dificuldades na selecção de textos adequados para o seu *corpus*, ressentindo-se a investigação, que se torna demorada e cansativa e conduz muitas vezes à ultrapassagem do tempo

útil de obtenção de resultados considerados válidos e pertinentes.

O trabalho desenvolvido em Terminologia implica uma consulta detalhada de textos relacionados com a prática exercida e com estudos desenvolvidos e publicados no domínio a que o trabalho diz respeito. É importante, pois, documentar-se ao máximo sobre a área que vai ser objecto de análise, enriquecendo o conhecimento sobre a mesma e orientando os caminhos de investigação de uma forma mais proveitosa e objectiva, ainda que seja sempre necessário a validação de um especialista. Se até há poucos anos seria fácil escolher as obras ou textos de referência na maioria das áreas, por serem muito divulgados ou por não ser fácil encontrar outras propostas, o conhecimento especializado avança agora de uma forma exponencial e o aparecimento de textos e obras complementares ou até alternativos sobre os mais variados assuntos processa-se de uma forma quase ininterrupta.

Não deixando de ser a compilação de *corpus* uma tarefa fundamental, ainda que apenas válida se levada a cabo de forma criteriosa e com resultados satisfatórios, a poupança do máximo de tempo possível na recolha dos textos pode compensar duplamente, encurtando não só o prazo para atingir a conclusão, mas também permitindo reaplicar esse tempo na investigação propriamente dita. Para um terminólogo, cuja tarefa de construção de glossários, dicionários ou qualquer outro tipo de listagem descritiva dos termos ligados à área do conhecimento a ser estudada é incontornável, o acesso aos textos mais relevantes é um factor crucial no desempenho do trabalho.

Como se poderá constatar mais adiante, a existência de vários programas informáticos que trabalham com terminologias e em Linguística de *Corpus* é uma realidade, mas também se poderá concluir que a maior parte dos programas se centra em determinadas etapas do processo de estudo, como, por exemplo, a anotação ou a extracção automática de termos do *corpus*, ficando outras, intermédias, que também são importantes, como a pesquisa e selecção de textos adequados para integrar no *corpus*, menos desenvolvidas. Foi nesse sentido que se procurou atingir um conceito de programa que privilegiasse a optimização dos processos de constituição e

gestão de *corpora* e que ambos se integrassem no sistema de forma a que beneficiassem mutuamente dos resultados que cada um deles obtém. Uma boa integração dos dois momentos, constituição e gestão, permite que a primeira antecipe a segunda e que o processo de actualização seja levado a cabo naturalmente.

Neste capítulo, apresenta-se a concepção de um protótipo de *software* de constituição e gestão semi-automática de *corpora* de especialidade denominado *e-Termite*, referem-se os objectivos que estiveram na sua idealização, descreve-se o perfil de utilizador que mais proveito retirará do manuseamento do *software* e formas de uso para cada um deles, listam-se as funções concebidas para facilitar a tarefa do terminólogo e descreve-se um procedimento de constituição e gestão de *corpus* em língua de especialidade, de acordo com a metodologia defendida, mas a título exemplificativo e simulado, dado que a aplicação não possui, neste momento, funcionamento real para testar esses processos.

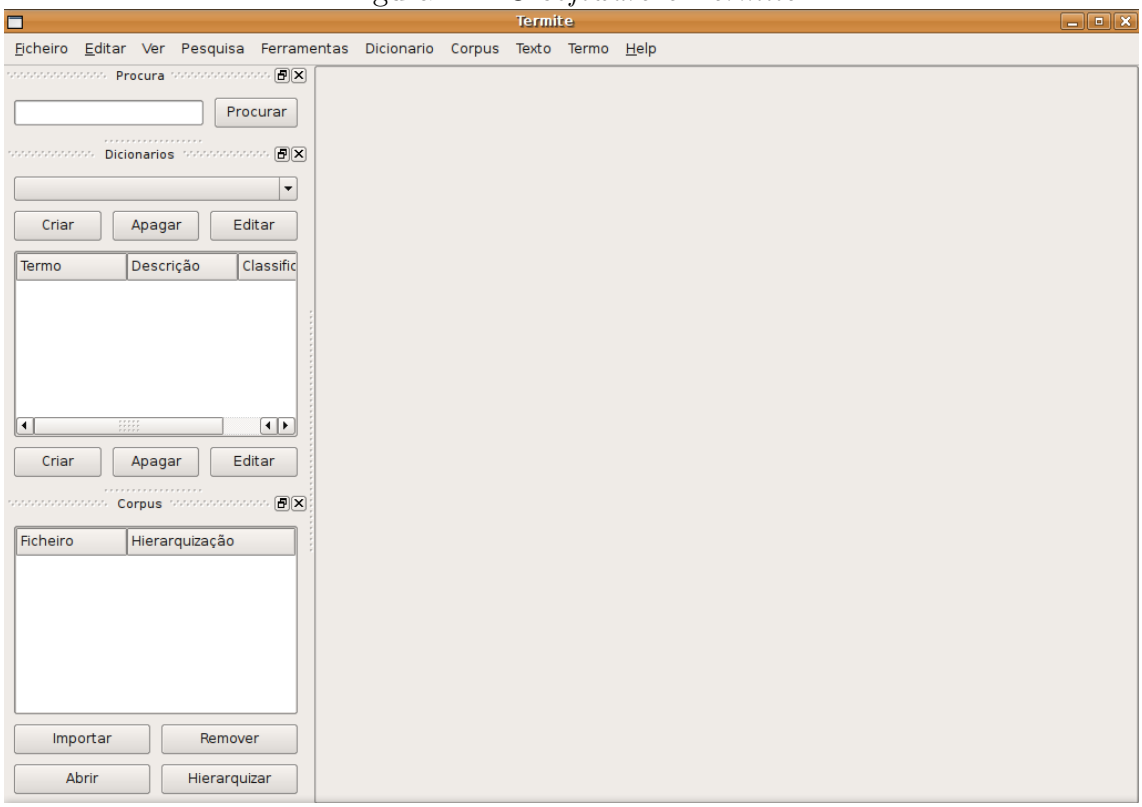
4.2 O protótipo *e-Termite*

O *e-Termite* é um protótipo em desenvolvimento de *software* de constituição e gestão semi-automática de *corpora* de especialidade, encontrando-se em fase de testes e com muitas das funções aqui descritas ainda em experimentação ou implementação. O nome surgiu por trabalhar exclusivamente com textos em suporte electrónico e servir para auxiliar em tarefas na área da Terminologia, ainda que possa ser usado para outro tipo de estudos estatísticos relacionados com os dados textuais. O programa foi desenvolvido de raiz, a partir de uma conceptualização sobre a optimização da constituição e gestão de *e-corpora*, recorrendo ao quadro teórico da Terminologia Textual, como suporte epistemológico, e às recentes tecnologias informáticas para dinamizar e autonomizar o processo ao máximo.

Este projecto conta com o apoio de alunos do curso de Engenharia Informática da Universidade do Algarve que têm vindo a implementar o código de modo a que o *e-Termite* exista digitalmente. Está a ser programado em C++, com livraria gráfica

QT e livraria SQLite para a base de dados, correndo actualmente apenas em sistemas Linux. O ambiente gráfico foi construído com janelas amplas, principalmente na parte de edição de texto, para que se possa trabalhar com visibilidade máxima sobre o texto. De qualquer forma, as janelas são movíveis e redimensionáveis ficando a decisão da melhor disposição ao critério de quem utiliza a aplicação. Podemos observar, na figura 4.1, o ambiente inicial da aplicação e a disposição das janelas, contando com quatro blocos de acesso rápido: a procura, os dicionários e termos, a lista de candidatos ao *corpus* e a, já acima mencionada, janela de edição de texto.

Figura 4.1: O *software e-Termite*



O conceito de funcionamento do *e-Termite* divide-se em módulos que equivalem, em grande parte, às funções nucleares do sistema, articulando-se entre si, nas diversas fases necessárias aos processos de constituição e gestão. Os módulos presentes são os de administração, de pesquisa, de importação, de edição, de classificação, de hierarquização, de anotação e de análise, sendo, cada um deles, explicado com maior detalhe de seguida e durante a exemplificação apresentada mais adiante. Para melhor se entenderem as diferentes partes e como se conjugam, vão ser, igualmente,

apresentados, ao longo da explicação sobre o funcionamento da aplicação, várias capturas de ecrã, tabelas informativas e diagramas de diferentes tipos elaborados em UML[2](Unified Modeling Language), que representam de uma forma simplificada cada um dos módulos e a sua articulação com os objectos (termos, dicionários, textos e *corpora*) e o agente (terminólogo). Expõe-se na figura 4.2 uma lista com símbolos UML que permite interpretar os diagramas que iremos apresentar mais tarde.

Figura 4.2: Legenda de símbolos utilizados na UML

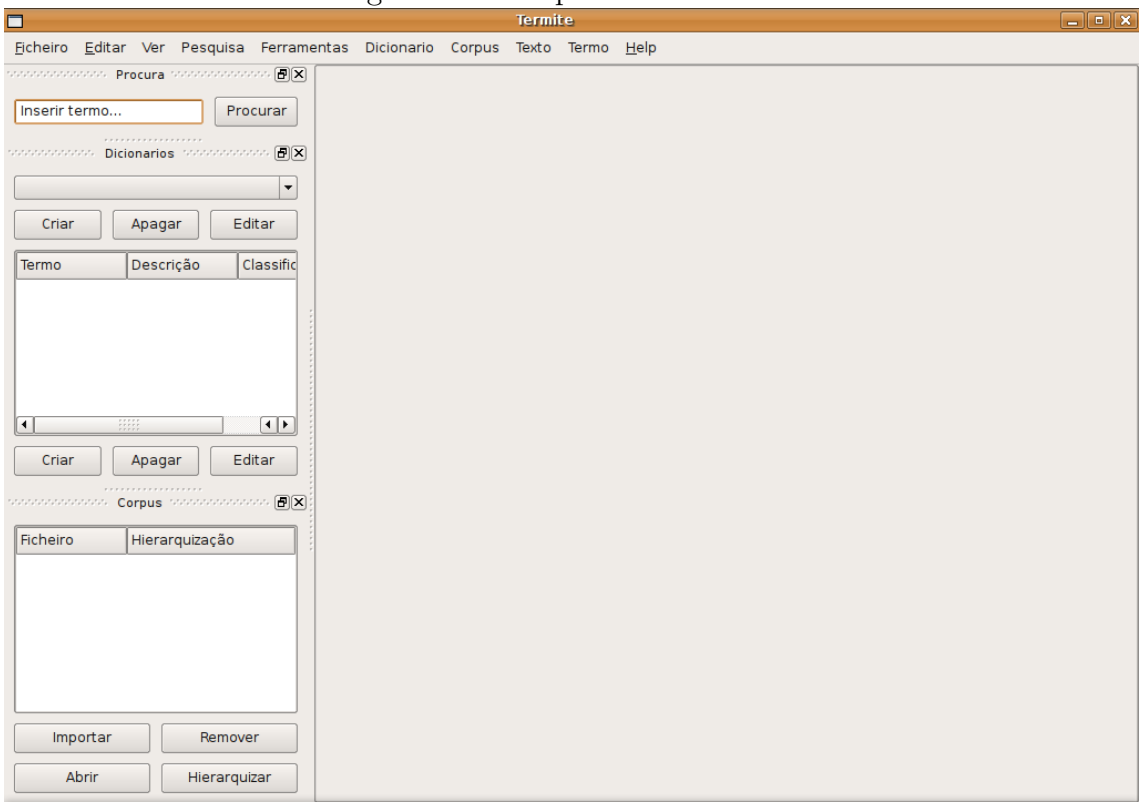
Antes de dar início à fase informatizada, que é aquela que mais nos interessa abordar neste capítulo, é necessário que se trate das questões preparatórias e incontornáveis, como sejam a definição prévia de objectivos e de critérios a respeitar, de acordo com os pressupostos epistemológicos da Linguística de *Corpus*. Estabelece-se também, desde logo, uma divisão na origem da base textual em três grupos, para que melhor se entenda algumas diferenças no *modus operandi* para cada um dos conjuntos: os textos em suporte papel, os textos informatizados e os textos na *web*. A importância de distingui-los surge, como já discutimos anteriormente, pelas características particulares, que obrigam a abordagens diferentes, como, por exemplo, a impossibilidade de pesquisar automaticamente informação em textos em suporte papel no contexto da aplicação informática, pois não existe nenhum módulo de leitura óptica para reconhecimento de texto, nem uma base de dados consultável sobre existências em papel, ficando, assim, excluídos do funcionamento directo na aplicação. No entanto, não significa que não se possam usar textos em suporte papel, apenas será necessário efectuar a conversão antes.

[2]`http://www.uml.org/`

Outra das situações que importam estabelecer é o estatuto da unidade terminológica e do texto no contexto da aplicação, enquanto objectos informáticos. Existem três fases distintas, no que diz respeito ao termo: a primeira fase é a do candidato a termo (CT), a segunda fase é a do candidato a termo classificado (CTC) e a terceira é a do termo classificado e validado (TCV). O CT é uma unidade lexical com potencial terminológico e que aguarda classificação e validação, antes de poder ser integrada no dicionário do domínio. O CTC não pode ser ainda confirmado como termo, ainda que detenha uma classificação provisória, pois aguarda validação pelo especialista. Esta decisão de permitir classificações provisórias tem como objectivo agilizar o trabalho do terminólogo que, assim, com base no conhecimento que tem do domínio poderá avançar classificações provisórias que serão mais tarde confirmadas ou não pelo especialista. O TCV, ou seja, simplesmente um termo, é uma unidade terminológica plena e validada com posição firmada no dicionário do domínio, reflectida pela sua classificação. Já o texto divide-se, no âmbito da aplicação, apenas em duas classes: o candidato ao *corpus* (CC) e o membro do *corpus* (MC), posições estas decorrentes da verificação ou não dos critérios definidos para a constituição do *corpus*.

Para dar início ao processo de constituição de *corpus*, pode-se partir de um termo ou textos com termos pertencentes ao domínio que se pretende estudar. Poderá ser usada como termo inicial a unidade lexical que designa o próprio domínio, pois tem uma probabilidade elevada de ser considerada um termo dentro do domínio que identifica, ou, caso se queira ser mais rigoroso, pode-se pedir a um especialista que valide inicialmente um pequeno grupo de termos num texto do domínio para serem utilizados como espoleta do processo de constituição. Na figura 4.3 podemos ver a caixa de busca do *e-Termite*, onde se dá início à pesquisa, bastando inserir o termo ou a pequena lista de termos (condição imposta pelo motor de busca que não suporta mais do que dez palavras em simultâneo), carregar em procurar e aguardar pelo resultado.

Figura 4.3: Pesquisa no *e-Termite*

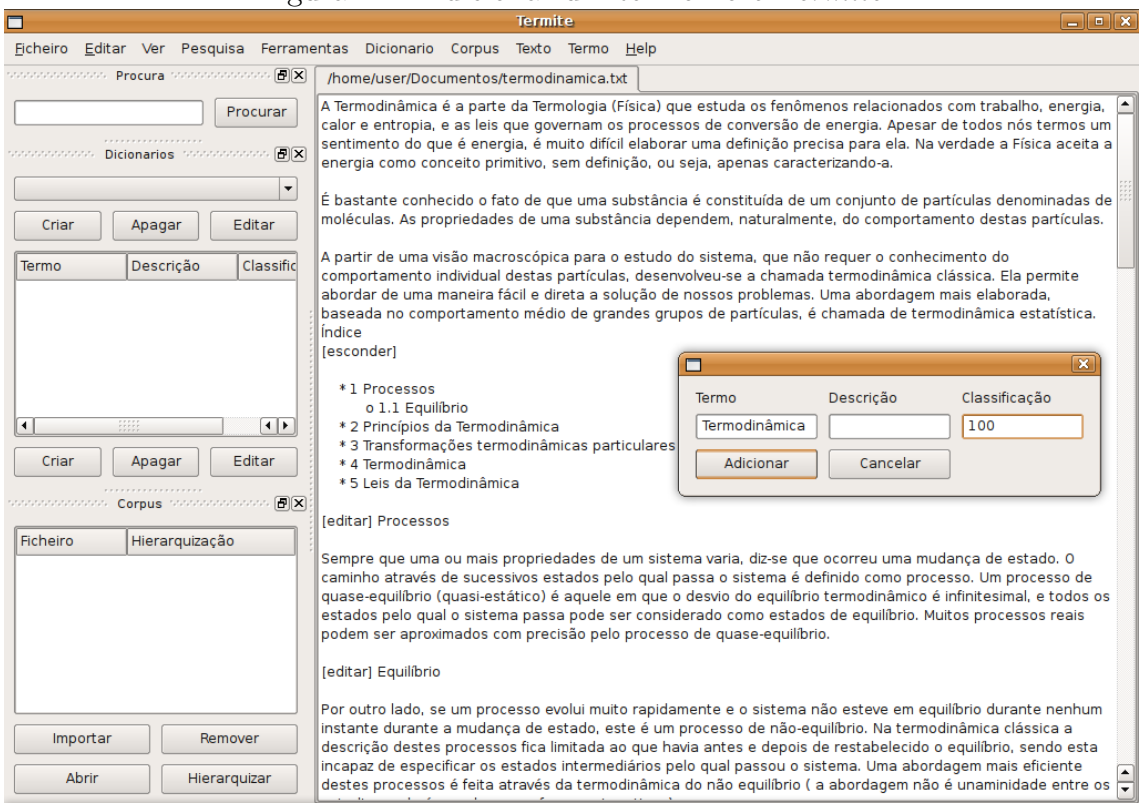


Durante a identificação e adição do termo ou lista de termos iniciais ao dicionário de domínio, procede-se a uma classificação quantitativa desses termos dentro de uma escala pré-definida pelo investigador. Essa escala é flexível e pode ser adaptada, sendo apenas importante que o valor mínimo e máximo possíveis, determinados para um termo, sejam sempre os mesmos, ou seja, podemos definir um valor mínimo de 1 ou 1000 e outro máximo de 100 ou 10000, desde que a escala aplicada a cada um dos termos identificados e classificados no contexto da constituição de *corpus* para um mesmo domínio seja igual, mantendo-se o princípio da homogeneidade presente, tal como defendido anteriormente. A amplitude da escala pode resultar do número de critérios a observar na classificação do termo, podendo cada um dos critérios ser o equivalente a um valor estipulado, assim, uma variedade maior de critérios levaria a um aumento de amplitude entre o valor mínimo e máximo possíveis, de forma a possibilitar o maior número de combinações e pontuações possíveis.

Como podemos observar na figura 4.4, apresenta-se um texto em modo de edição, selecciona-se o termo e, com o botão direito, acciona-se o menu de adição

de texto ao dicionário do domínio, permitindo em duas acções fazer introduções na lista de termos.

Figura 4.4: Adicionar um termo no *e-Termite*



A atribuição de pontuação aos termos será feita de acordo com os critérios definidos previamente, como foi referido, que poderão ser o grau de univocidade ou de relevância no âmbito do domínio em questão, tal como refere a noção de «*termhood*» (Kageura & Umino, 1996:11), já apresentada no segundo capítulo, ou outros que se considerem produtivos para o objectivo definido. No caso de se iniciar o processo com base em decisões pessoais de validação terminológica, será indispensável recorrer a um especialista para aferir sobre a validade e importância dos termos provisórios e confirmar se a pontuação atribuída se adequa ao valor do termo para aceitá-la ou corrigi-la.

Desta forma, repete-se o procedimento quantas vezes forem necessárias para cada um dos CTs, atribuindo-se-lhes uma pontuação, dentro da escala definida e de acordo com os critérios pré-determinados. As pontuações de cada um dos TCVs que existe no texto contribuem para uma soma final, que reflecte o número de

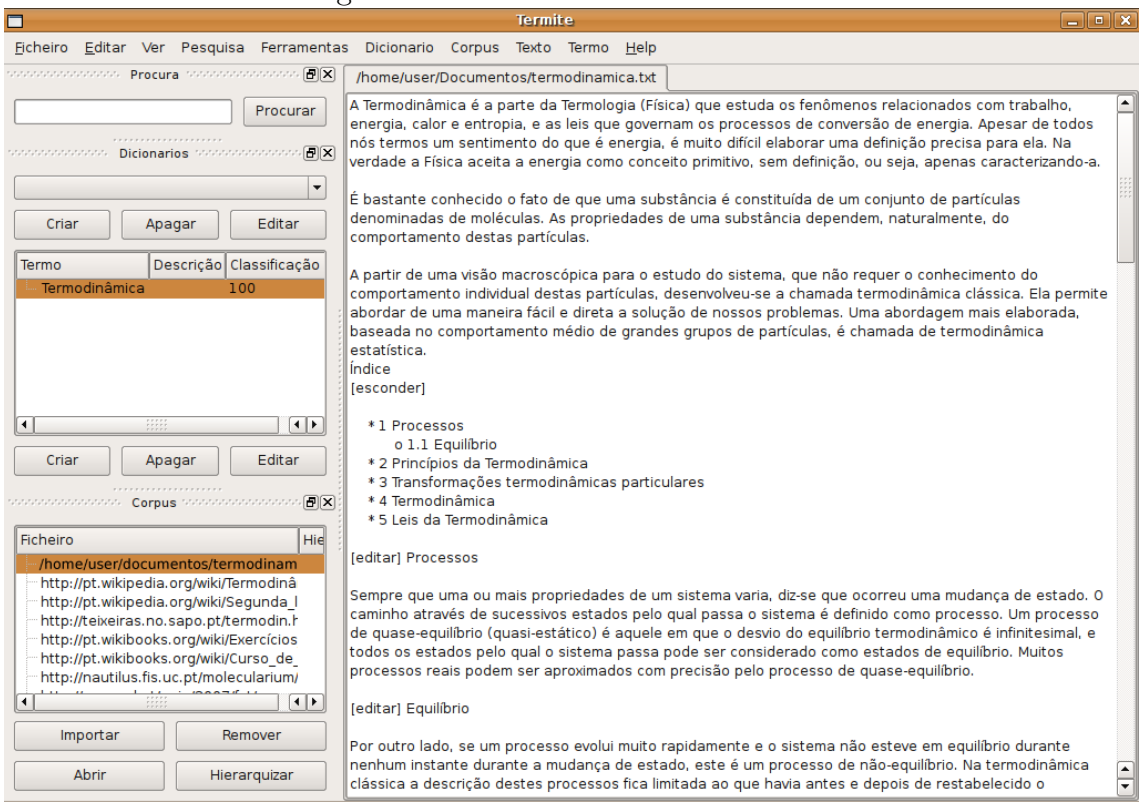
TCVs identificados e o valor que lhe atribuímos. A esta soma corresponde um valor real, mas sempre temporário até à conclusão do trabalho, dado que qualquer nova introdução de um termo no dicionário do domínio e que conste de um dos textos alterará a soma obtida previamente.

Numa fase mais avançada e já com vários termos classificados e validados, vai construir-se aos poucos uma lista que podemos usar para processar automaticamente qualquer texto e obter uma soma. Essa lista de termos classificados será designada sempre por dicionário do domínio, de ora em diante, para facilitar a explicação. Pegando num grupo de textos e aplicando-lhe o dicionário do domínio, será possível atingir uma lista de textos classificados e ordenados mediante o resultado numérico que atingirem. A hierarquia resultante tem como base, no início, as poucas existências terminológicas no dicionário do domínio. Como se pode observar na figura 4.5, na lista de candidatos ao *corpus* constam as ligações para os textos que serão analisados, tanto as ligações locais, provenientes de documentos existentes em dispositivos de armazenamento ou de rede interna, como as ligações externas, originárias da Internet.

Os textos analisados que obtenham maior pontuação final ocuparão um lugar superior na lista ordenada, por neles existirem mais e «melhores» ocorrências terminológicas, ainda que numa avaliação sujeita ao dicionário do domínio. Partindo-se do princípio de que os termos de um determinado domínio têm tendência para coexistir em textos da mesma área, por questões de coesão lexical, semântica, sintáctica, conceptual e textual, procede-se à análise do texto na íntegra, acrescentando ao dicionário do domínio todos os termos que ainda não constam e classificando-os, segundo os mesmo critérios.

Este é um processo iterativo que resulta do crescimento e aperfeiçoamento constante do dicionário do domínio pela introdução e classificação de novos termos e que permite também uma reavaliação dos candidatos ao *corpus* existentes na base textual e elaboração de uma lista de textos hierarquizada, com os textos mais pontuados de acordo com o dicionário de domínio no topo dessa lista. Assim, é mais

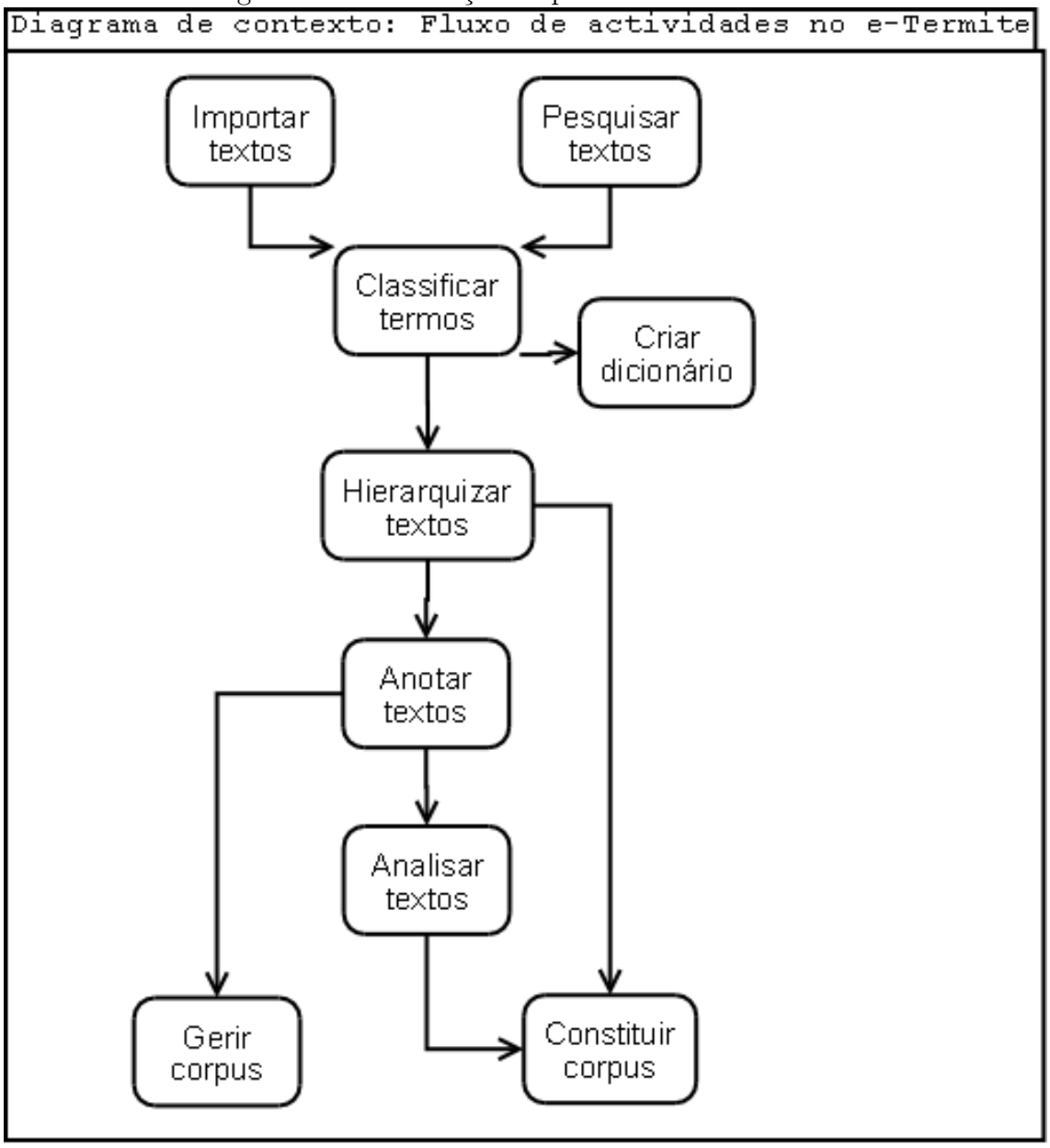
Figura 4.5: Lista de CCs no *e-Termite*



fácil aceder aos textos que estão mais próximos do domínio e retirar os termos ainda não introduzidos ou classificados no dicionário de domínio. Podemos observar o diagrama de actividades na figura 4.6, que retrata sumariamente os processos no *e-Termite*, sendo toda esta sequência de processos apresentada em maior detalhe mais adiante e ainda exemplificada pela apresentação de uma situação de constituição de um *corpus*.

A questão sobre quem poderá retirar benefícios do uso do programa remete para uma resposta mais abrangente do que se for feita sobre a quem é que se destina o *e-Termite*. Sendo os principais alvos de utilização deste *software* os terminólogos ou qualquer investigador a executar tarefas terminológicas, pois trata-se de uma ferramenta de construção e análise de *corpora* de especialidade, pelas múltiplas funções de pesquisa, anotação e análise que a aplicação disponibiliza, é a estes que se destina a concepção. No entanto, pode-se conceber uma possível generalização da sua utilização às mais diversas tarefas relacionadas com o uso de *corpora* e com a pesquisa de informação, alargando-se, pois, a sua utilidade, possivelmente, a qualquer

Figura 4.6: Articulação de processos no *e-Termite*



investigador que trabalhe com línguas. A própria simplicidade com que se executam as tarefas mais básicas, como seja a pesquisa por informação textual, possibilita que qualquer utilizador comum que tenha competências informáticas possa explorar e utilizar o programa sem dificuldades, ainda que não tenha sido desenhado para esse segmento de utilizadores. O interface foi concebido para ser intuitivo e conduzir o utilizador progressivamente no processo de pesquisa de informação textual, de constituição de *corpus*, na construção de dicionários do domínio e na análise de texto.

Antes de apresentarmos o caso de utilização, é importante relembrar e desenvolver mais aprofundadamente os objectivos de concepção do *e-Termite*, pois são estes que estruturam a concepção do protótipo apresentado na dissertação.

4.2.1 Objectivos

Quando se definiram os objectivos a atingir pela metodologia desenhada para a aplicação informática, na base das questões mais importantes a ter em conta, prevaleceram os recursos informáticos e tecnologias disponíveis, a fundamentação teórica da Linguística de Corpus e da Terminologia Textual, bem como o impacto na prática terminológica. Assim, os objectivos que o *e-Termite* procura atingir enquadram-se em três categorias diferentes:

- Tecnológicos
- Epistemológicos
- Práticos

Dentro dos objectivos tecnológicos, a aplicação *e-Termite*, apesar de ainda se encontrar numa fase de desenvolvimento alfa[3], tem como finalidade usar as mais actualizadas tecnologias de pesquisa e catalogação de informação do momento. O *e-Termite* pode efectuar buscas recorrendo ao uso do motor de pesquisa Google, pretendendo-se numa versão posterior autonomizar o programa, dotando-o de capacidades ao nível de *web crawling*[4] e, assim, evitar filtros desconhecidos e implementados por terceiros alheios à investigação.

O sistema de menus privilegia as funções nucleares do processo, como sejam a pesquisa, a classificação de termos e a interacção com o dicionário de domínio. Os dados recolhidos são armazenados num servidor de dados com tecnologia SQL, podendo residir no próprio computador, onde se desenvolve a investigação, ou ser

[3] Jargão informático para versão de teste não disponibilizada ao público.

[4] `http://en.wikipedia.org/wiki/Web_crawler`

consultados através da Internet, facilitando o trabalho remoto ou em grupo. A base de dados foi desenhada para permitir o ajuste do número de campos que se sintam necessários para o trabalho que está a ser desenvolvido, podendo guardar-se informação linguística, não-linguística e, até, as relações conceptuais identificadas entre cada um dos campos, consoante as necessidades do projecto. O *e-Termite* apresenta-se, assim, como uma ferramenta moldável e virada para os objectivos do utilizador, com as potencialidades do estudo estatístico desenvolvidas também para ir ao encontro das necessidades da investigação.

A estrutura do programa foi concebida de forma a permitir uma integração modular de outras funcionalidades, tendo-se dividido em fases mais importantes dos processos de constituição e gestão. A flexibilidade das partes estruturantes e margem de progresso do programa preconizam o objectivo de garantir a possibilidade de actualização do *software* ao nível dos conteúdos e das funcionalidades. A modernização constante da vertente tecnológica é fundamental para impulsionar os objectivos epistemológicos e práticos, tal como vimos nos capítulos anteriores, pois o facto de existirem mais possibilidades técnicas motivam outros procedimentos, podendo daí advir também outras concepções teóricas que sejam aperfeiçoamentos das anteriores.

Os objectivos epistemológicos apontam para dois vectores principais. O primeiro é a capacidade do *software* acompanhar os desenvolvimentos mais modernos da teoria em Terminologia para que possam ser aplicadas e testadas metodologias inovadoras na área da pesquisa e investigação em línguas de especialidade. Se este objectivo não for conseguido, qualquer trabalho levado a cabo corre o sério risco de ser questionado por desactualização de procedimentos ou desfasamento das bases teóricas. O segundo objectivo epistemológico é contribuir para o desenvolvimento conceptual da Terminologia, facultando ao linguista ferramentas que lhe permitam, além da execução do trabalho prático, proceder a tarefas de análise e meta-análise linguística, ainda que não seja o objectivo principal da aplicação. Deste modo, o programa consegue aliar ao pragmatismo da funcionalidade dos métodos correntes, a potencialidade de investigar e descobrir novas formas de trabalhar em Terminologia.

De todos os objectivos, aqueles que dizem respeito à prática terminológica foram considerados os prioritários, pois o *e-Termite* é uma ferramenta virada para a execução de tarefas terminológicas. No entanto, não se pode retirar valor aos objectivos tecnológicos e epistemológicos, pois contribuem para que a prática seja validada e melhorada constantemente. A procura pela optimização do tempo dispendido na compilação de um *corpus* de especialidade adequado constitui o mais importante objectivo desta aplicação informática.

O tempo gasto na recolha de materiais para integrar o *corpus* depende sempre da disponibilidade de materiais e da própria dimensão do domínio, mas pode ser sujeito a uma significativa redução, se o processo de pesquisa for devidamente efectuado. A maior parte das vezes, mesmo que a busca seja realizada num ambiente informático, quando a dispersão de textos relacionados com um domínio é muito grande, encontrar materiais adequados é uma tarefa lenta e na qual se tira pouco proveito do facto de a maior parte dos textos já existirem em suporte informático. Assim, poder-se-ia restituir aos terminólogos tempo efectivo de investigação que agora é dispendido de forma escusada na pesquisa e recolha de textos.

Para conseguir concretizar uma redução do tempo de constituição do *corpus* é obrigatório atingir outros objectivos intermédios, como, por exemplo, a simplificação do interface de trabalho para tarefas de análise de texto e a criação de critérios e filtros linguísticos e estatísticos que permitam disponibilizar, o mais rápido possível, para análise terminológica os textos relevantes para a descrição conceptual dos domínios, evitando perder tempo em processos de decisão e selecção textual. Outro dos objectivos práticos do *e-Termite* é constituir um dicionário do domínio com os termos identificados e verificados como tal a partir da análise dos textos.

Assim, conhecidos e apresentados os objectivos que determinam o funcionamento da aplicação, passamos de seguida a descrever as funções implementadas para a consecução desses objectivos.

4.2.2 Funções

Tendo em conta que o *e-Termite* é uma concepção de protótipo de *software* voltada para a prática terminológica, as funcionalidades à disposição do investigador reflectem procedimentos considerados essenciais para constituir e para gerir na prática um *corpus* de especialidade. Desta forma, as funções implementadas, que serão apresentadas já de seguida de forma mais detalhada e justificada, são uma resposta aos requisitos identificados como prioritários para atingir um bom nível de desempenho da aplicação no cumprimento dos objectivos delineados. Para melhor se compreenderem as funções, decidimos separá-las em dois conjuntos representativos das funções, as gerais, que remetem para questões não relacionadas directamente com o trabalho a realizar nos *corpora*, e as específicas, que intervêm no procedimento de constituição e gestão de *corpora*. As funções gerais são poucas e resumem-se a administrar, a partilhar e a disponibilizar. Já as funções específicas são mais e podem subdividir-se em tarefas de constituição (pesquisar, importar, editar, classificar e hierarquizar) e tarefas de gestão (anotar, analisar e armazenar). Apresenta-se, então, uma descrição de cada uma das funções concebidas para o protótipo de *software e-Termite*, complementando-se com capturas de ecrã ou diagramas sempre que a complexidade da função descrita o exigir.

4.2.2.1 Administrar

As funções de administração do *software* dizem respeito a questões técnicas que são essenciais para manter o programa em bom funcionamento, como seja proceder a actualizações ou instalação de componentes adicionais (módulos, impressoras, scanners, etc), e, por vezes, necessárias ao desenvolvimento de tarefas. No entanto, a administração e suas tarefas não podem ser muito complicadas ou demoradas, pois o objectivo é libertar ao máximo o utilizador para a investigação, não podendo ser uma aplicação que exija demasiada atenção neste sector.

4.2.2.2 Partilhar

Tal como vimos anteriormente, o valor de poder partilhar a informação é fulcral porque permite não só uma potencial reutilização dos dados, mas também proporciona trabalhar em conjunto e em simultâneo, situação muito frequente na prática terminológica, onde a presença de um especialista que auxilie o terminólogo é frequente e necessária. A utilização de uma base de dados simultaneamente interna e externa, que se sincroniza entre si sempre que a situação requer, permite a troca de informações e tarefas com a frequência necessária.

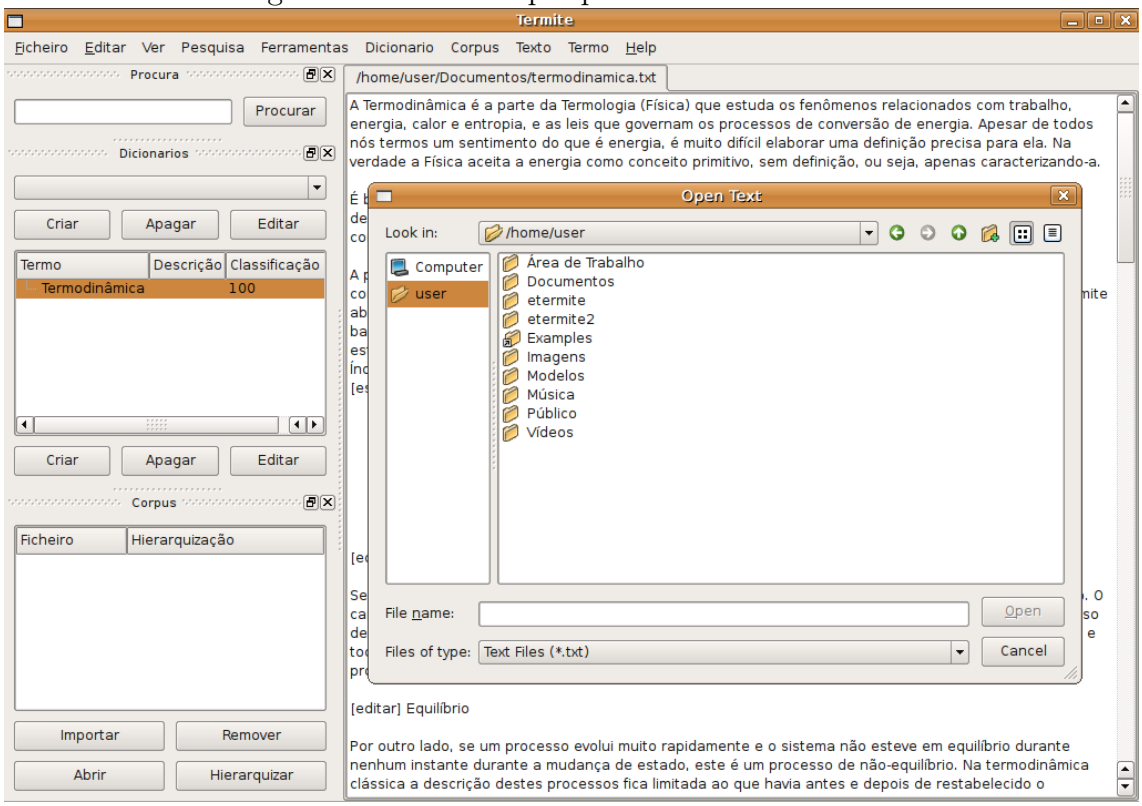
4.2.2.3 Disponibilizar

A existência de um ambiente *web* que sirva de suporte ao alojamento da estrutura de dados é também uma prioridade, dada a importância que reveste a disponibilidade máxima de acesso aos dados terminológicos. Assim, o trabalho poderá ser executado em contextos diversos e a partilha de informação, tal como apresentada no ponto anterior, é, mais uma vez, facilitada.

4.2.2.4 Pesquisar

A pesquisa serve para encontrar textos que sejam candidatos a integrar o *corpus* em constituição. A busca pode incidir sobre dois contextos informáticos diferentes, as redes interna e externa e os dispositivos de dados, sendo que consideramos ambos funcionalmente equivalentes, pois representam, em qualquer dos casos, um repositório com uma base textual de incidência. Como se pode observar na figura 4.7, para iniciar uma pesquisa local, basta seleccionar o menu respectivo, esperar pela caixa de diálogo e introduzir a pasta onde os textos estão guardados. O sistema encarrega-se de ler a pasta e subpastas, se definido, puxando para a lista de candidatos ao *corpus* as ligações para os ficheiros que se enquadram nas extensões conhecidas como ficheiros de texto.

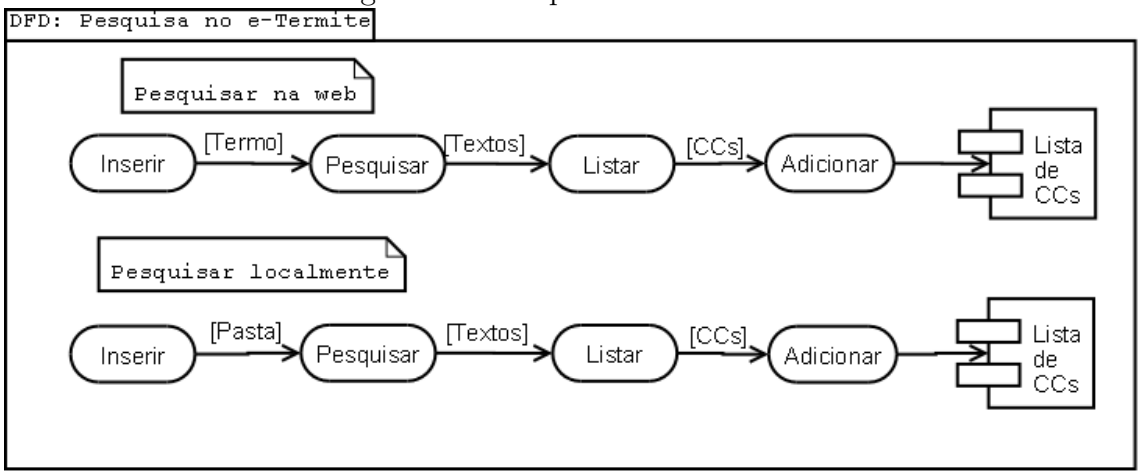
Figura 4.7: Início de pesquisa local no *e-Termite*



A pesquisa tem como objectivo principal povoar o *software* com ligações para textos candidatos ao *corpus* (CCs), recorrendo, nesta fase de desenvolvimento do *e-Termite*, ao motor de busca Google para obter os resultados a partir da Internet. Basta introduzir-se um termo na caixa de busca (pode-se inserir a designação do domínio que se está a estudar ou um pequeno conjunto de termos validados) e obtém-se uma lista de ligações ou atalhos, como, por vezes, são referidos, para possíveis textos relacionados com a designação do domínio ou com o conjunto de termos utilizados. Esta lista de resultados está organizada mediante os critérios aplicados pelo motor de busca e a ordem será respeitada até que o dicionário de termos possa ser aplicado e se executar uma hierarquização. No caso da pesquisa em dispositivos de armazenamento de dados locais, se não existir um dicionário do domínio para processar os textos encontrados, o *e-Termite* vai limitar-se a produzir uma lista de ficheiros organizada por ordem de leitura, até que exista um dicionário de termos do domínio que possa hierarquizar esses textos. Podemos observar, na figura 4.8, o funcionamento da pesquisa através dos diagramas de fluxo de dados

que representam a forma como as pesquisas funcionam em contextos diferentes, mas com procedimentos e passos quase idênticos.

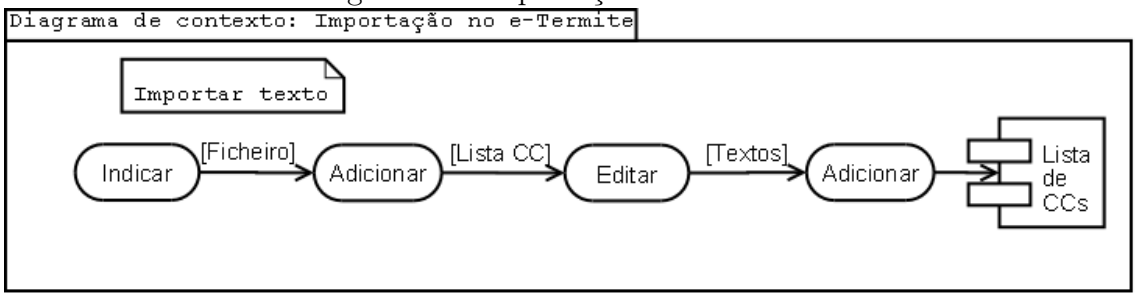
Figura 4.8: Pesquisas no *e-Termite*



4.2.2.5 Importar

A importação é a função que permite abrir directamente para edição e classificação no programa um ficheiro de texto, omitindo-se, nestes casos, a fase da pesquisa no âmbito da constituição com o *software*. Deste modo, é possível integrar, na investigação desenvolvida com o *e-Termite*, os textos que não existam na Internet ou que possam ser recolhidos de formas diferentes, pois a *web* não é a única fonte de recursos textuais disponível. Qualquer texto em suporte de papel poderá ser digitalizado e introduzido na lista de candidatos ao *corpus* desta forma. A importação pode ser efectuada apenas para um ficheiro ou para todos os ficheiros que existam numa determinada pasta, sendo necessário proceder com cautela, pois, esta operação procede à abertura automática para edição dos ficheiros, para que se proceda à sua análise. A importação deverá suportar os seguintes formatos, que são os mais utilizados: XML, SGML, PDF, DOC, ODT, RTF, PS, HTML e TXT. Na figura 4.9, podemos observar o diagrama que apresenta o processo de importação e comparar os diferentes passos em relação à figura anterior (4.8), principalmente onde contorna o processo de pesquisa e abre directamente na aplicação o ficheiro para modo de edição.

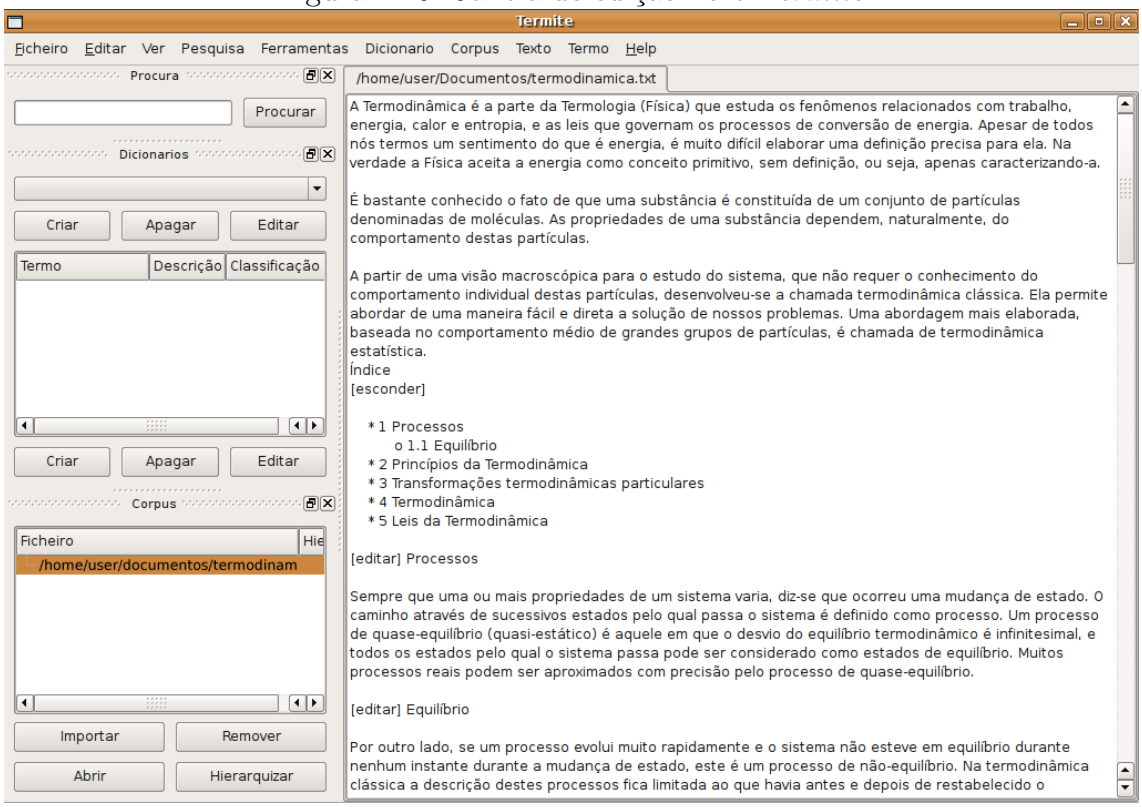
Figura 4.9: Importação no *e-Termite*



4.2.2.6 Editar

A edição é a operação que permite alterar o conteúdo de um texto e realizar tarefas directamente relacionadas com o dicionário do domínio, tornando-se, assim, tais tarefas possíveis sempre que um texto esteja disponível na janela de edição do *e-Termite*, tal como é visível na figura 4.10.

Figura 4.10: Janela de edição no *e-Termite*

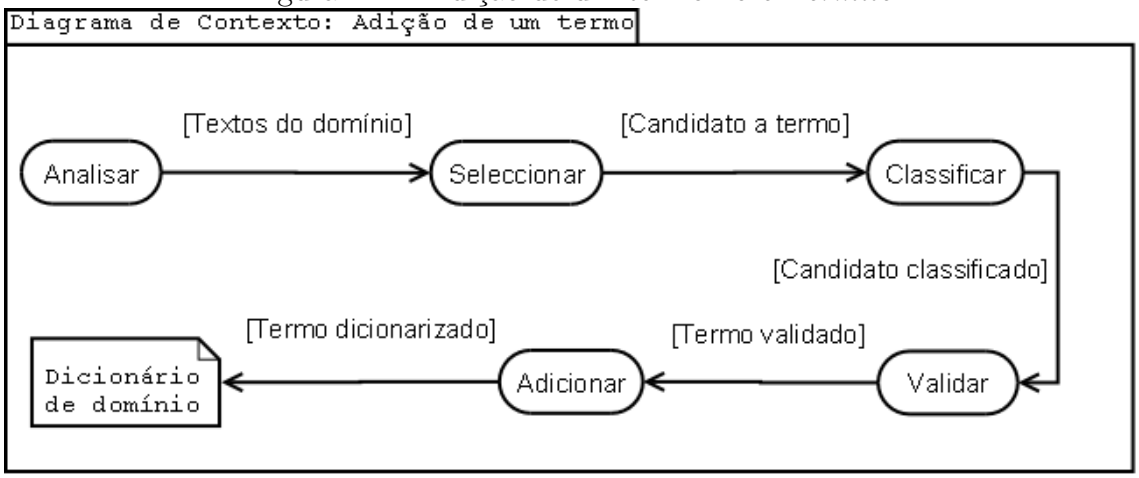


Para o efeito, basta abrir um ficheiro que esteja presente na lista de ligações para textos, que resulta do processo de pesquisa, ou importar o texto directamente e esperar que o processo de conversão esteja concluído. No editor de texto, pode-

mos executar tarefas como a correcção de gralhas ou a inserção de forma rápida e prática dos CTs identificados no texto, recorrendo a um conjunto mínimo de acções. Através de uma operação com o botão direito do rato em cima do CT, selecciona-se "Adicionar termo" e introduz-se informação mínima, pedida pelo programa, indispensável para criar a entrada no dicionário de termos, como sejam o domínio a que se pretende associar o termo e a classificação correspondente nesse domínio. No sistema de edição, está, ainda, disponível um modo especial que apresenta o texto com o código XML existente visível e que permite fazer acertos manuais.

Na figura 4.11, podemos observar o processo de adição de um termo ao dicionário de domínio e acompanhar as alterações que decorrem no termo ao longo da introdução.

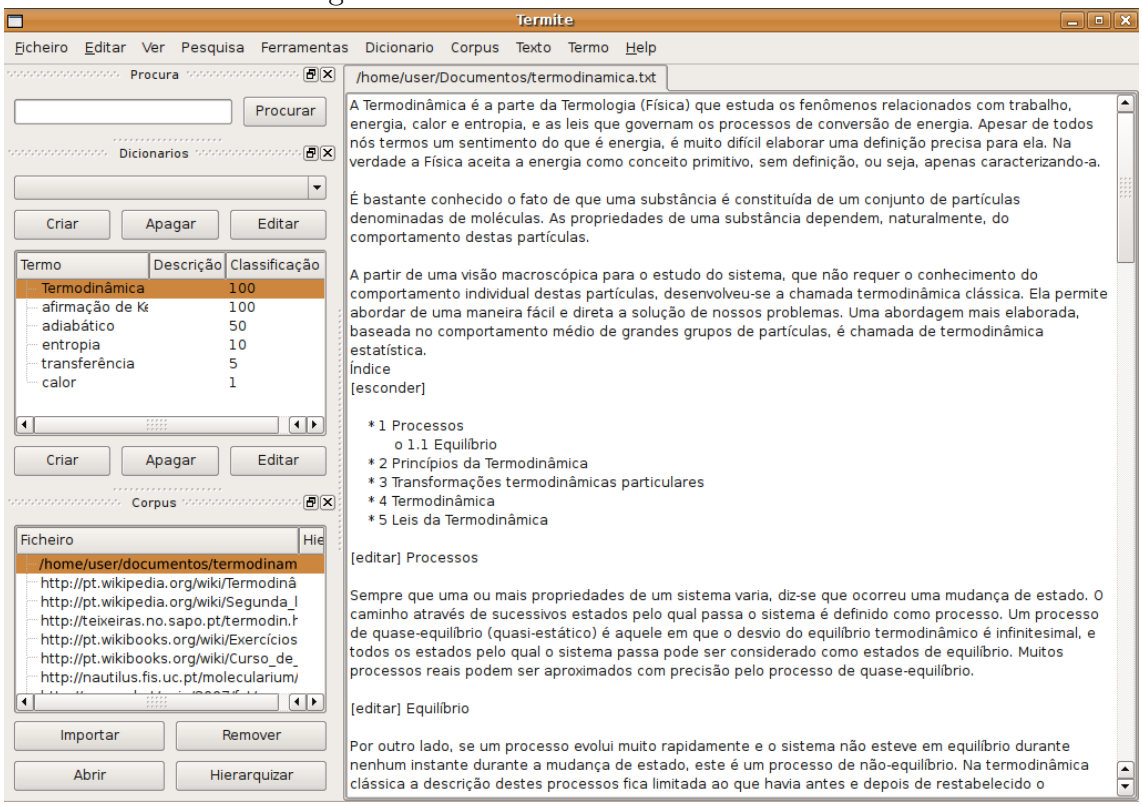
Figura 4.11: Adição de um termo no *e-Termite*



4.2.2.7 Classificar

O processo de classificação resume-se à atribuição de um valor numérico a cada CT que é identificado, para posteriormente ser validado, como termo do domínio em estudo. Já observámos, na figura 4.4, a adição de um termo e a atribuição de um valor, apresentando-se agora, na figura 4.12, um exemplo de como a lista de termos se organiza nas suas classificações. As classificações que se mostram na imagem serão explicadas mais adiante na tabela 4.1 e representam alguns termos do domínio da Termodinâmica.

Figura 4.12: Lista de CTs no *e-Termite*

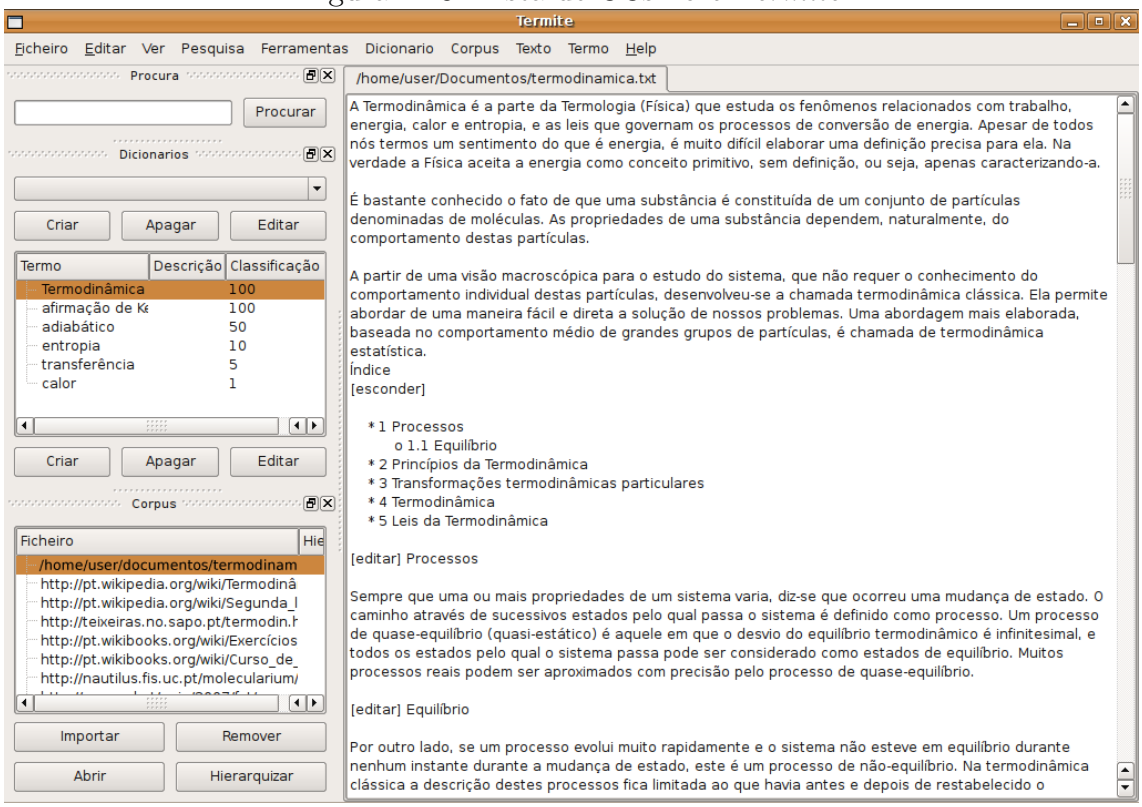


A classificação dos termos é o eixo de toda a constituição de *corpus* no *e-Termite* por duas razões: primeiro porque todas as outras funções, de uma maneira ou de outra, estão dependentes ou trabalham para o processo de classificação e segundo porque, sendo uma parte tão crítica do sistema, é a que depende por inteiro das decisões e das intervenções do terminólogo. De qualquer forma, não se pode encarar este ponto como negativo, pois a concepção do sistema não só prevê a situação, como também considera que é a solução mais adequada para o processo de classificação e de validação de termos. A classificação totalmente automática dos termos, que se encontra presente no extractores automáticos de termos, nunca é totalmente independente, pois também obriga à intervenção do terminólogo e do especialista que em conjunto têm de proceder à verificação e validação dos termos extraídos automaticamente.

4.2.2.8 Hierarquizar

A hierarquização é um processo de reclassificação dos textos que tem como objectivo principal reordenar frequentemente a base textual que existe no programa para que possamos analisar primeiro os textos com maior relevância para o estudo do domínio. Pode-se observar, na figura 4.13, a lista de candidatos ao *corpus* já constituída e bastando, para a reclassificação ter lugar, carregar no botão "Hierarquizar".

Figura 4.13: Lista de CCs no *e-Termite*



A relevância decorre de uma reclassificação dos textos feita com base no dicionário de termos do domínio, que vai sendo construído pelo processo de classificação de termos. O processo de hierarquização deve ser efectuado sempre que se conclui a adição e classificação de novos termos de um texto ao dicionário do domínio para que a lista de textos candidatos ao *corpus* possa estar actualizada.

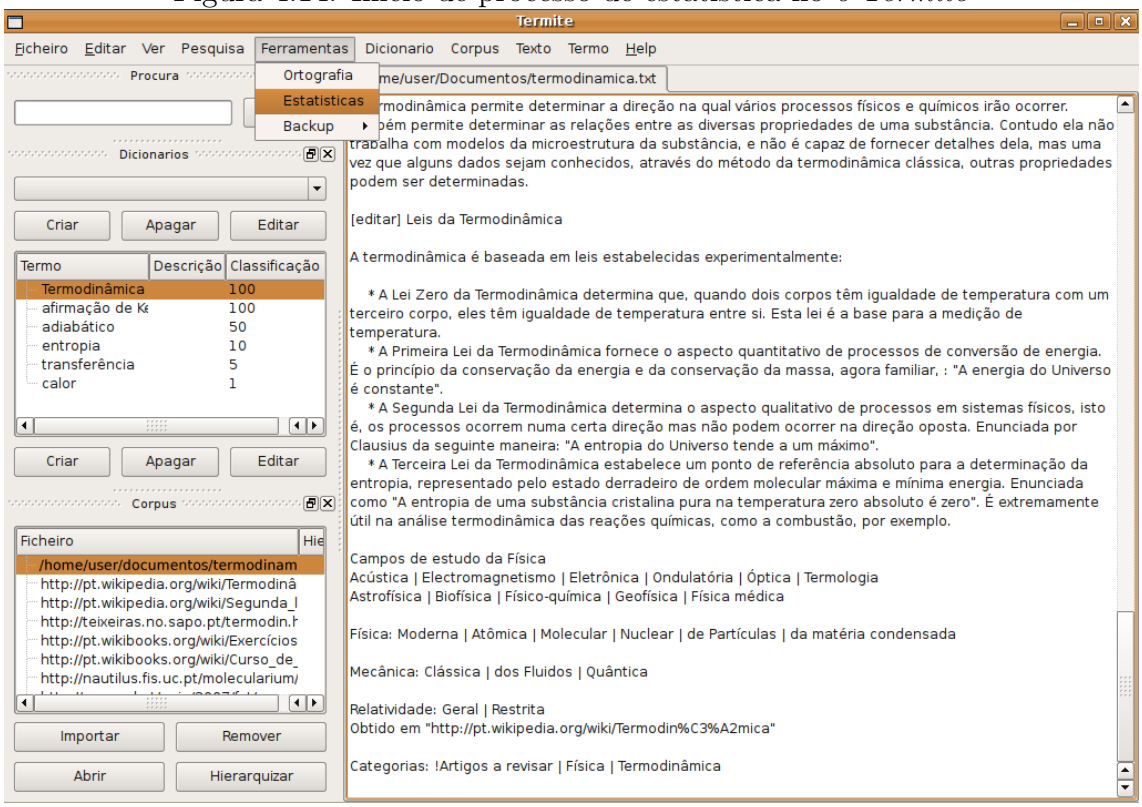
4.2.2.9 Anotar

A anotação de texto é um processo que permite anexar informação extra ao texto, como já vimos anteriormente. É um dos processos mais complexos de executar, sendo também, no entanto, a fase que mais poderá expandir o potencial de aplicação e reutilização do *corpus*, por potencialmente aumentar a quantidade de informação que transporta e criar as condições favoráveis à sistematização e consulta. O *e-Termite* foi desenhado para ser flexível o suficiente e permitir que a informação relativa à investigação possa ser anexada a quatro entidades diferentes, quando materializadas em objectos na base de dados, a saber: o *corpus*, o texto, o dicionário e o termo. As quatro entidades correspondem a dois níveis diferentes de profundidade na análise linguística: o nível dos *corpora* e dos textos (textual) e ao nível dos dicionários e dos termos (lexical). Para cada um dos objectos registados na base de dados, podem ser criados campos de registo diferentes, correspondendo cada um deles a categorias que o investigador considere importantes e decida preencher com informação. Essas categorias servirão para melhor definir cada um dos objectos que for gravado na base de dados. O programa já traz categorias pré-definidas para algumas entidades, como o dicionário, onde, por exemplo, os campos para o termo, a descrição e a classificação, são obrigatórios para a introdução de um termo, pois não faria sentido se fosse possível introduzir uma entrada no dicionário sem a existência do termo e da sua classificação. No entanto, as categorias obrigatórias podem e devem ser complementadas com outras categorias, como, por exemplo, a data de produção, o autor ou a classe morfossintáctica que cada termo transporta no texto. Essas decisões são facultativas e decorrem da iniciativa do terminólogo, que julgará, no âmbito dos objectivos da investigação, quais as informações a reter na base de dados. Este tipo de anotação flexível permite uma descrição mais aprofundada e variada, com a reutilização da informação a ser privilegiada. As vantagens deste tipo de anotação já foram discutidas no capítulo anterior e não iremos entrar em detalhe mais uma vez.

4.2.2.10 Analisar

O processo de análise engloba um conjunto de tarefas que podem ser executadas sobre um *corpus* anotado e permitem formular hipóteses e verificar conjecturas com base nos dados recolhidos. Apesar das ferramentas estatísticas ainda não estarem implementadas, a estrutura de menus já prevê o seu desenvolvimento, como se pode observar na figura 4.14.

Figura 4.14: Início de processo de estatística no *e-Termite*



A inclusão de dados que ficam anexados ao dicionário, ao *corpus*, ao texto e aos termos não só permitem a reutilização futura desses mesmos dados, mas também proceder a análises estatísticas e linguísticas sob várias perspectivas. No entanto, passa sempre por uma decisão do investigador anotar devidamente os termos e os textos com a informação necessária à prossecução dos seus objectivos para que o domínio e o suporte informático não sejam os únicos critérios a serem preservados. O processo de análise vira-se, assim, para o processo de anotação e depende muito da forma como o segundo é conduzido. Como já foi afirmado antes, a concepção de protótipo de *software* prevê uma flexibilidade muito significativa na informação

que se pode anexar, a qual acaba por ser complementada pela possibilidade de ser toda ela cruzada e estatisticamente analisada. Por exemplo, ao criarmos o campo de "data de produção", no objecto texto, ele propaga-se automaticamente a todos os termos que são recolhidos desse texto para o dicionário, poupando-se muito tempo no preenchimento de campos da base de dados. É certo que, em situações pontuais, se pode proceder à edição do campo da data por termo ou evitar a propagação automática a todos os termos. Com o campo da data, juntamente com a adição de um campo de "autor", permitiria que se pudesse traçar uma evolução cronológica de um termo quanto à frequência em determinado autor.

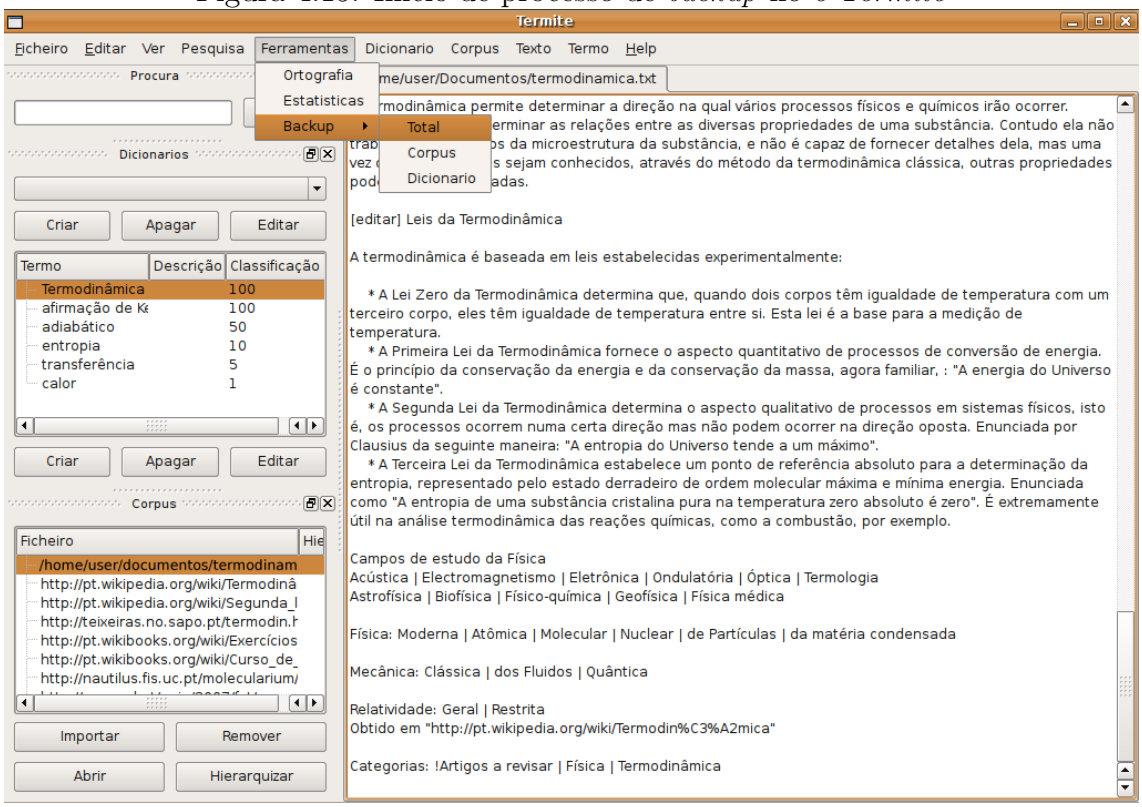
Os objectivos mais importantes do *software*, como já foi referido, são a constituição e a gestão do *corpus*, existindo processos de análise que contribuem de forma relevante para a consecução de ambos. A análise, que não foi implementada ainda nesta fase, depende em grande parte da informação guardada durante a anotação. Ainda que não seja um objectivo prioritário, pois até já existem vários conceitos e vectores de análise desenvolvidos por outros programas informáticos, a presença impreterível da função de análise condiciona a concepção do programa e por isso não pode deixar de ser referenciada. O *e-Termite* prevê os tipos de análise mais comuns ao nível da investigação linguística, como, por exemplo, buscas por palavras-chave, expressões regulares, colocações e concordâncias. O conceito de palavra-chave implementa a ideia de procurar palavras que estatisticamente se destaquem no conjunto de textos que integra o *corpus*. As expressões regulares, no contexto da Informática, definem padrões lexicais, frásicos ou textuais que também são usados para procurar por palavras, grupos de palavras ou porções de texto. As colocações são expressões usuais que, pelo seu uso reiterado e grau de fixidez acentuado, acabam por ser consideradas como uma palavra única, não aceitando transformações morfossintácticas nos seus constituintes. Ainda se antevê a possibilidade da elaboração de concordâncias, com as quais se torna possível estudar as ocorrências vizinhas das unidades terminológicas existentes no *corpus*. O processo de análise só deverá ser efectuado na fase final da investigação, depois da constituição e da gestão conferirem ao *corpus* al-

guma estabilidade. Mesmo que depois se proceda a alguma actualização, só quando o *corpus* for considerado anotado e completo para o objectivo que o investigador se propôs atingir é que se deve proceder à análise.

4.2.2.11 Armazenar

A função de armazenar surge pela necessidade de organizar toda a informação que se prevê aumentar com o tempo. Dado que se defende um conceito de reutilização, não será desejável descartar os dados, mesmo depois do final da investigação.

Figura 4.15: Início de processo de *backup* no *e-Termite*



A discussão sobre o armazenamento levanta duas questões que são importantes na gestão dos *corpora*. A primeira diz respeito ao espaço que a base de dados ocupa e que depende, obviamente, do número e do tamanho de dicionários, de *corpus*, de textos e de termos, juntamente com a informação anexa, que cada programa albergar. A segunda questão é relativa à forma como os dados são armazenados de forma a permitir a sua reutilização. Para tornar as situações de falta de espaço de armazenamento para os dados facilmente solucionáveis, antecipou-se a criação de uma

ferramenta de exportação por entidade (*corpus*, texto, dicionário e termo) e global que permite salvaguardar todos os dados armazenados e repô-los noutro computador. Esta função permite, acima de tudo, minimizar a questão das imprevisíveis falhas de *hardware* (falha electrónica ou mecânica de componentes de computador), problemas de *software* (vírus, corrupção de ficheiros, utilização indevida ou outras), permitindo a reutilização física dos dados. As funções de armazenamento enquadram-se nas funções de gestão de *corpus* e representam o conjunto de processos que está ao dispor para fazer a manutenção dos elementos produzidos/armazenados durante os projectos de investigação.

Esta funções foram entendidas como as indispensáveis para constituir o *corpus*, proceder à análise e armazenar para reutilização recorrente, estando prevista a adição de outras, que se acharem necessárias, ao longo do desenvolvimento da aplicação. Nos subcapítulos seguintes, apresenta-se a implementação dessas funções, ainda que de uma forma meramente teórica, dado que a aplicação ainda está em desenvolvimento, como foi referido, não sendo possível trabalhar com o *e-Termite*, para já, a maioria das funções.

4.2.3 Constituição de *corpus*

A investigação pressupõe duas etapas principais no trabalho com o *corpus*: a constituição e a gestão. Cada uma delas tem fases próprias, ainda que tenham sido desenhadas para trabalhar em conjunto e mutuamente se complementarem. A constituição do *corpus* subdivide-se numa série de pequenos procedimentos que, se implementados correctamente, conduzirá a resultados práticos significativos de uma forma rápida e válida. Cada uma das etapas corresponde a um módulo que desempenha funções diferentes no programa. Segue-se um exemplo prático de utilização da aplicação informática *e-Termite* e todos os passos necessários para constituir e gerir um *e-corpus* de especialidade.

4.2.3.1 Preparação

Antes de iniciar o trabalho prático com o *software* é necessário definir os objectivos, o domínio de aplicação e os critérios que regem a compilação do *corpus*. O objectivo, já parcialmente abordado antes, passa por exemplificar a constituição de um *corpus* de especialidade com cerca de 200000 unidades lexicais, o mais rapidamente possível, de forma a demonstrar os processos idealizados para o *e-Termite*. Sendo este um procedimento de carácter terminológico e apesar de estarem a ser testados os processos metodológicos teóricos, é necessário proceder à escolha de um domínio que sirva de base para o trabalho de análise de léxico de uma especialidade. Após ponderação sobre possíveis domínios a utilizar como exemplo, escolheu-se o campo da Termodinâmica, como objecto de análise. A opção pela Termodinâmica deu-se por vários motivos de carácter teórico e prático que passo a apresentar:

- Existência de uma instabilidade conceptual ligada ao domínio da Ecologia e das práticas de poupança energética, que estão na ordem do dia.
- *Corpus* em português inexistente.
- Domínio que é uma subdivisão de especialidade, de tal forma que é possível remeter o âmbito do estudo a uma esfera com fronteiras relativamente reduzidas e delimitadas para que, numa fase posterior, se possa partir para a verificação de procedimentos a um nível mais geral.
- Existência de múltiplos termos em português.
- Lacuna nos estudos terminológicos da área (inexistência de dicionários ou glossários do domínio.
- Área com uma vertente muito técnica e prática que facilita o aparecimento de termos de especialidade.
- Assume como preponderante na sua evolução e desenvolvimento o diálogo entre os vários especialistas do próprio domínio e de outros com que interage, de-

sempenhando a língua um papel fundamental na comunicação de experiências e perspectivas do domínio que impulsionam a sua constante transformação.

- Disponibilidade de um especialista para validar termos.

Os critérios para classificação de *corpora*, dado que o objectivo é recolher qualquer texto informatizado pertencente a um domínio, estão reduzidos a integrarem textos em suporte electrónico e que pertençam à área da Termodinâmica, explicando-se já de seguida o processo que permite identificar os textos como válidos para integrar o *corpus*.

4.2.3.2 Critérios de classificação dos termos

Para se poder iniciar a busca por textos candidatos ao *corpus* (CCs), é necessário estabelecer os critérios a usar para a classificação dos termos, uma vez que são eles que vão produzir o resultado que, no final, cada texto somará e decidir, assim, se o texto integrará ou não o nosso *corpus*, como se explicará de seguida. Embora não faça parte do âmbito deste trabalho identificar ou validar os melhores critérios e pontuações para classificação dos termos, é imprescindível usar critérios pelos motivos que já foram apresentados anteriormente.

Assim, apresenta-se uma proposta de funcionamento, baseada na noção de «*termhood*» defendida por Kageura e Umino, a qual já foi explicada no subcapítulo 2.3, cujo o pressuposto é criar um *corpus* do domínio com base nos termos que existem nos textos. Relembramos muito brevemente que a noção de «*termhood*» sustenta a ideia da existência de diferentes níveis de proximidade das unidades terminológicas no domínio.

Desta forma, poderá ser observado o estatuto de univocidade, que remete para o grau de maior proximidade, se o termo apresentar apenas um significado com um único sentido, sendo exclusiva a interpretação e utilização no contexto do domínio de aplicação, de outros domínios e da língua em geral. Inversamente, se o termo apresentar sentidos e interpretações várias, consoante a multiplicidade de significados

e de utilizações, a exclusividade perde-se e o grau de proximidade também decai, tornando-se o termo menos pertencente ao domínio, podendo mesmo questionar-se o estatuto terminológico.

Como foi apontado, sendo o programa concebido num paradigma semi-automático, cabe ao terminólogo, recorrendo ao auxílio dos especialistas para confirmar a validade do termo, determinar o nível de pertença ao domínio, tal como foi acima descrito. Para poder enquadrar o termo numa quantificação, é necessário proceder a tarefas preliminares, como, por exemplo, criar uma escala de pontuação, que servirá de referência na classificação que cada termo recebe, com base nos critérios que melhor se adequam à investigação.

O objectivo inicial é constituir um *corpus* de especialidade cujo o critério base é, unicamente, nesta fase, pertencer ou não ao domínio, sendo esse o primeiro passo para reduzir a base de incidência onde os demais critérios possam ser aplicados de seguida. Terminado esse objectivo, a base textual já se pode considerar um *corpus*, ainda que com um critério muito abrangente e será necessário, caso seja esse o objectivo final, aplicar os restantes critérios que moldem o *corpus* à nossa investigação. A aplicação dos outros critérios decorre, assim, do recurso a processos estatísticos e aos processos de anotação para delimitar ainda mais o *corpus*, como seja, por exemplo, para seleccionar textos que apresentem uma determinada dimensão, uma determinada variação linguística (PT ou BR) ou um determinado nível de língua. É neste segundo nível de definição do *corpus* que os critérios definidos poderão ser aplicados, consoante as necessidades da investigação.

Para já, simulamos o funcionamento do primeiro nível de constituição do *corpus*, conforme apresentado, que separará os textos do domínio da base textual de referência, criando a primeira versão do *corpus*. Assim, elaborou-se um sistema de classificação que tem como critério único a proximidade do termo em relação ao domínio, já referido anteriormente como «*termhood*». A escala de classificação usada é simples e seriam necessários ainda muitos testes com a aplicação informática a funcionar em pleno para que se pudesse chegar a conclusões definitivas sobre a pro-

dutividade dos resultados obtidos. Relembro, no entanto, que o mais importante não é a escala usada, que, neste exemplo, é composta por 5 níveis, pois essa pode ser sempre redefinida, uma vez que é extrínseca ao programa e serve apenas de referência para os pontos a conferir a cada termo. A pontuação escolhida para representar cada nível da escala (desde 1 até 5), que de ora em diante referiremos também como classe, ainda que elaborada para efeito de exemplo, seguiu critérios que procuram tornar o funcionamento do processo de classificação dos textos candidatos ao *corpus* mais evidente.

Um dos critérios foi uma distribuição ponderada dos pontos com base na importância de cada classe. Como se pode verificar na tabela 4.1, há uma discrepância nos intervalos de pontuação entre as cinco classes, que se baseia na decisão de não se uniformizar a diferença de importância entre cada um dos níveis, como passaremos a explicar. A classe de nível 1 vale o dobro da pontuação do nível 2 (100 e 50 pontos respectivamente) e se a distribuição de pontos na escala fosse uniforme, a classe de nível 2 valeria o dobro da classe de nível 3 (50 e 25 pontos respectivamente) e assim sucessivamente, terminando na classe nível 5 com metade da pontuação da classe de nível 4, ou seja, com um valor por unidade de 6,25 pontos. No entanto, dado que se considera o peso de 1 termo unívoco muito superior ao peso de 16 unidades terminológicas de nível 5 (100 pontos, que é a pontuação por uma unidade de nível 1, seria equivalente a 16 unidades de nível 5 (16x6,25=100)), que são utilizadas com múltiplos sentidos e não possuem uma interpretação única para o domínio em questão.

Quanto aos termos unívocos, consideramos serem compostos essencialmente por dois tipos de unidades terminológicas: as que Depecker designa como «*entités scientifiques, molécules ou étoiles essentiellement, designées par des codes en raison de leur trop grand nombre*» (Depecker, 2000:107) e as que, por ainda não terem tido contacto suficiente com outros domínios, se conservam temporariamente unívocas. Dado que a interacção com outros domínios é frequente e múltipla, estas unidades não conservam muito tempo o seu estatuto de univocidade, pois há uma tendência

para a assimilação e integração dos termos pelos domínios de contacto, o que conduz a que muitos desses termos deixem de ser unívocos rapidamente. Assim, o termo unívoco, classificado com nível 1, integrará um texto do domínio com uma probabilidade cem vezes superior à de um termo de nível 5. Esse texto terá igualmente grandes probabilidades de ser relevante para o domínio e, consequentemente, para observação pelo terminólogo. Veja-se, então, na tabela 4.1, os níveis e pontuações utilizados para definir o sistema de classificação, neste exemplo.

Tabela 4.1: Tabela de classes e pontuações de termos

Classes de Termos			
Nível	Descrição	Exemplo	Pontos
1	Referência absoluta para o domínio (termo técnico unívoco)	Afirmação de Kelvin-Planck	100
2	Referência partilhada com domínios próximos (termo técnico partilhado)	Adiabático	50
3	Referência partilhada com domínios afastados (termo técnico generalizado)	Entropia	10
4	Referência comum, mas com uso próprio no domínio (palavra comum, mas com aplicação técnica na área)	Transferência	5
5	Referência com significado comum partilhado (palavra comum, com aplicação técnica partilhada)	Calor	1

Além destas pontuações mais directas, há outros processos possíveis com influência na variação dos resultados que poderão ser tidos em conta, por melhor ajustarem o critério de valor relacionado com a proximidade, ainda que não podendo ser testados, será difícil verificar até que ponto podem ser ou não úteis ao processo de classificação dos termos e de pesquisa de textos. Apresentam-se de seguida alguns possíveis critérios de pontuação exemplificativos:

- Se uma palavra do dicionário do domínio ocorrer no título ou no resumo poderá ser atribuída uma bonificação, pois são contextos especiais de ocorrência

em que as unidades lexicais são escolhidas com o objectivo de representar a globalidade do assunto do texto.

- Ocorrência múltipla de um termo no mesmo texto (um termo validado que ocorre várias vezes no mesmo texto deve receber um bónus gradual por cada existência, pois indica um texto com probabilidade superior de pertencer ao domínio que integra).

- Bonificação diferenciada para ocorrências múltiplas (um termo com um nível superior, dentro da escala definida, deve ser mais bonificado por ocorrer mais vezes no mesmo texto, pois a sua presença é um indicador forte de que o texto poderá pertencer ao domínio que integra).

- Reconhecimento flexível (permitir alguma flexibilidade no reconhecimento de termos, não se limitando a identificar a sequência de caracteres introduzida no dicionário e tornando, também, possível a detecção de gralhas gráficas e de correspondentes semânticos). Observe-se uma pequena lista de fenómenos a ter em conta:

  * Aproximação (correcção automática de possíveis gralhas ortográficas).

  * Capitalização (reconhecimento sensível a maiúsculas e minúsculas).

  * Sinónimos (designações variadas que referem um mesmo conceito.

  * Variantes morfossintácticas (variações em género, número, etc).

  * Reformulações (expressões ou unidades lexicais que contribuem para uma actualização de um conceito).

Tendo definido para esta primeira fase, os critérios de classificação de termos e as suas pontuações pode-se avançar para a pesquisa por textos relacionados com o domínio da Termodinâmica.

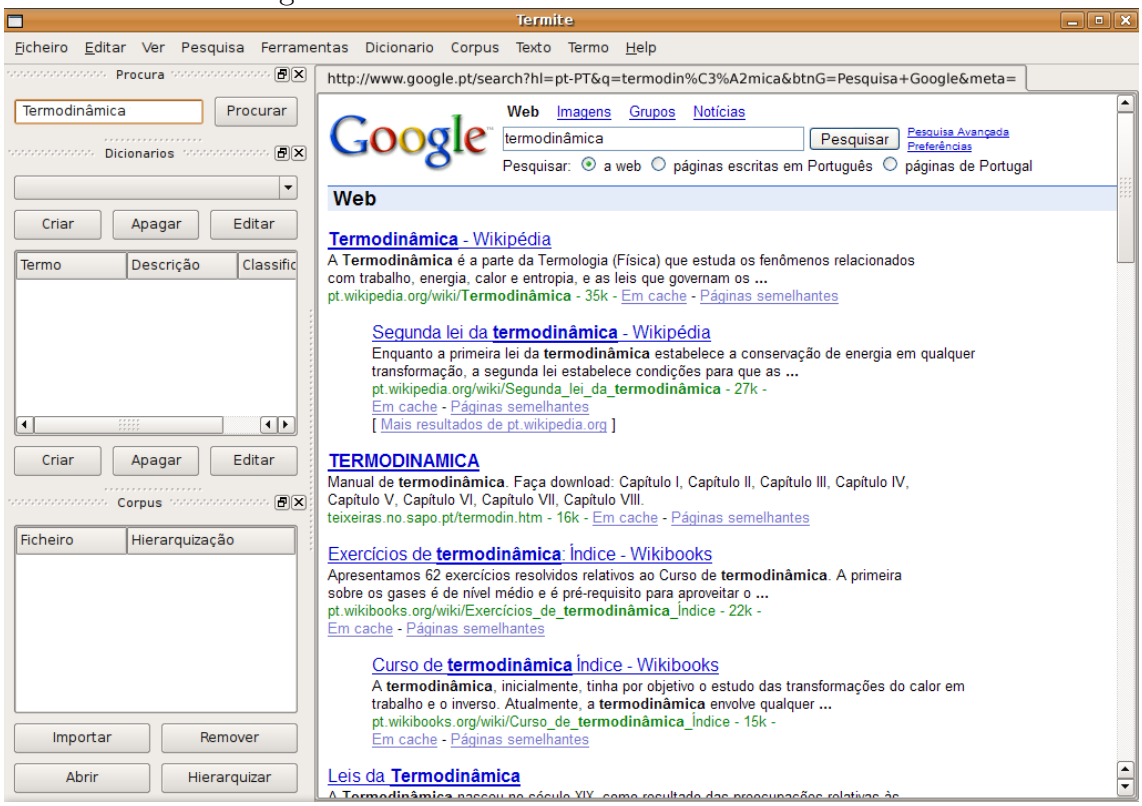
4.2.3.3 Pesquisa

Depois de uma obrigatória e incontornável fundamentação teórica para que se conheçam os pressupostos que constituem a base da investigação, passa-se necessariamente para uma parte de implementação prática do trabalho. O módulo de pesquisa é um dos principais da aplicação e serve para estabelecer a ligação entre o *software* e a Internet, não sendo, no entanto, obrigatório, se os textos já se encontrarem gravados em suporte electrónico e forem abertos manualmente. Relembramos que a Internet poderá servir apenas de complemento aos textos já recolhidos noutras fontes e introduzidos na constituição por outros processos. Para facilitar a tarefa de introdução de vários textos candidatos ao *corpus* na lista de candidatos a analisar, é possível recorrer à função de pesquisa numa pasta local ou num dispositivo de armazenamento, onde estejam guardados os textos, num sítio específico da Internet, através da introdução do endereço na *web*, ou a partir de um motor de busca que usa termos para procurar directamente na Internet, como já analisámos.

No exemplo que apresentamos, recorre-se ao uso de uma pasta local e combinam-se os textos nela existentes com o resultado de uma pesquisa na Internet, usando a palavra-chave, que designa e representa o domínio, «Termodinâmica». Recorrendo ao motor de busca do Google, a palavra «Termodinâmica» apresenta cerca 1.260.000 resultados para ocorrências textuais registadas na base de dados do Google. Deste número de textos, apenas uma parte pertence ao domínio da «Termodinâmica», pois há textos nos quais a unidade lexical é mencionada, mas que não se relacionam obrigatoriamente com o domínio, e ainda haverá outros que, pelo contrário, não contendo a palavra «Termodinâmica», também não aparecem listados, mas que pertencem ao domínio. Podemos observar a lista de resultados na figura 4.16, que dá uma ideia de como a pesquisa funciona na prática.

Por agora, vamos aceitar e utilizar a hierarquização do Google, adicionando a lista dos primeiros vinte endereços que nos são fornecidos e servirão de ponto de partida para o início da recolha de termos. Começamos por abrir o primeiro

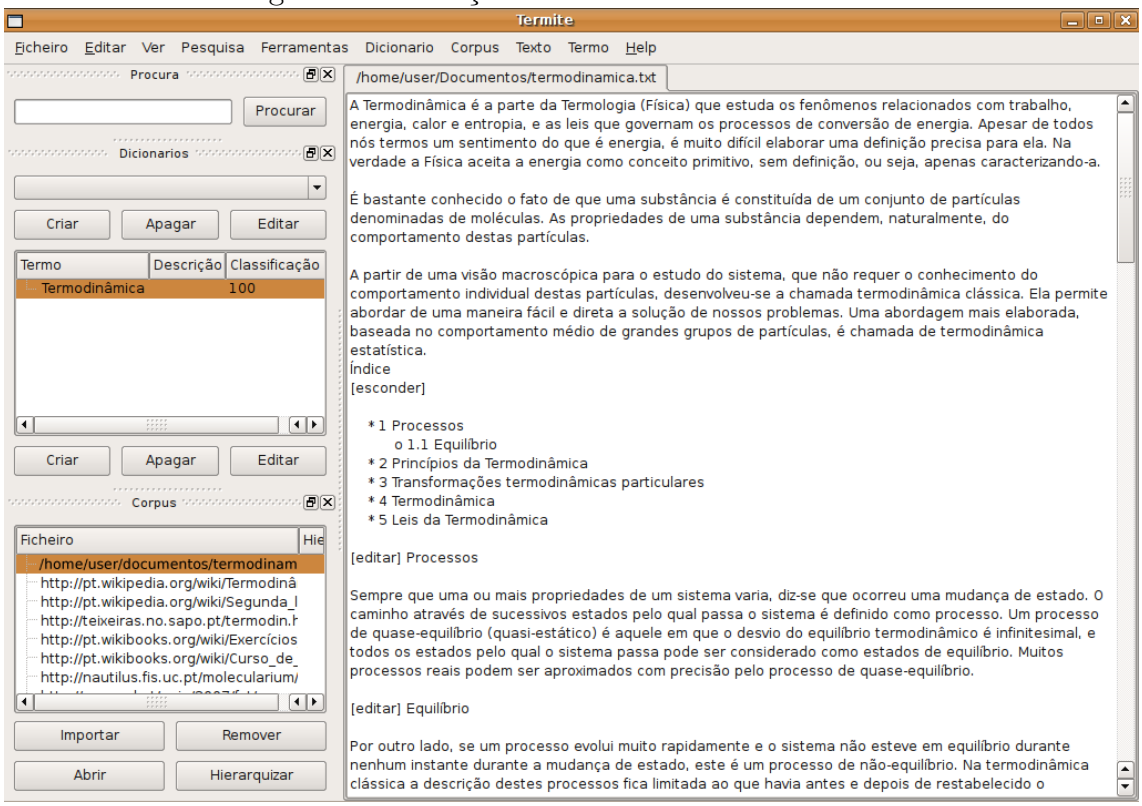
Figura 4.16: Resultados de busca no *e-Termite*



endereço na janela de edição da aplicação, onde é feito um processo automático de limpeza de todo o conteúdo que não seja língua natural (tabelas, imagens e código de programação). Existem, assim, vinte possíveis candidatos recolhidos da Internet, mais um existente numa pasta, a integrar o *corpus* sobre Termodinâmica e à espera de serem pontuados. Como se pode verificar pela captura de imagem, na figura 4.17, os textos, candidatos ao *corpus*, estão em pé de igualdade pois não foi aplicada nenhuma reclassificação.

Depois de terminado o processo de classificação de todos os termos identificados nos textos, que iremos observar mais adiante, é necessário ir buscar mais textos. Para isso, reinicia-se o procedimento indo buscar à Internet mais vinte ficheiros ou ligações para serem analisados. O programa dá a cada um dos ficheiros uma assinatura única que prevenirá a duplicação de análises e ainda marcará todos os endereços já visitados para que sejam excluídos a cada nova importação.

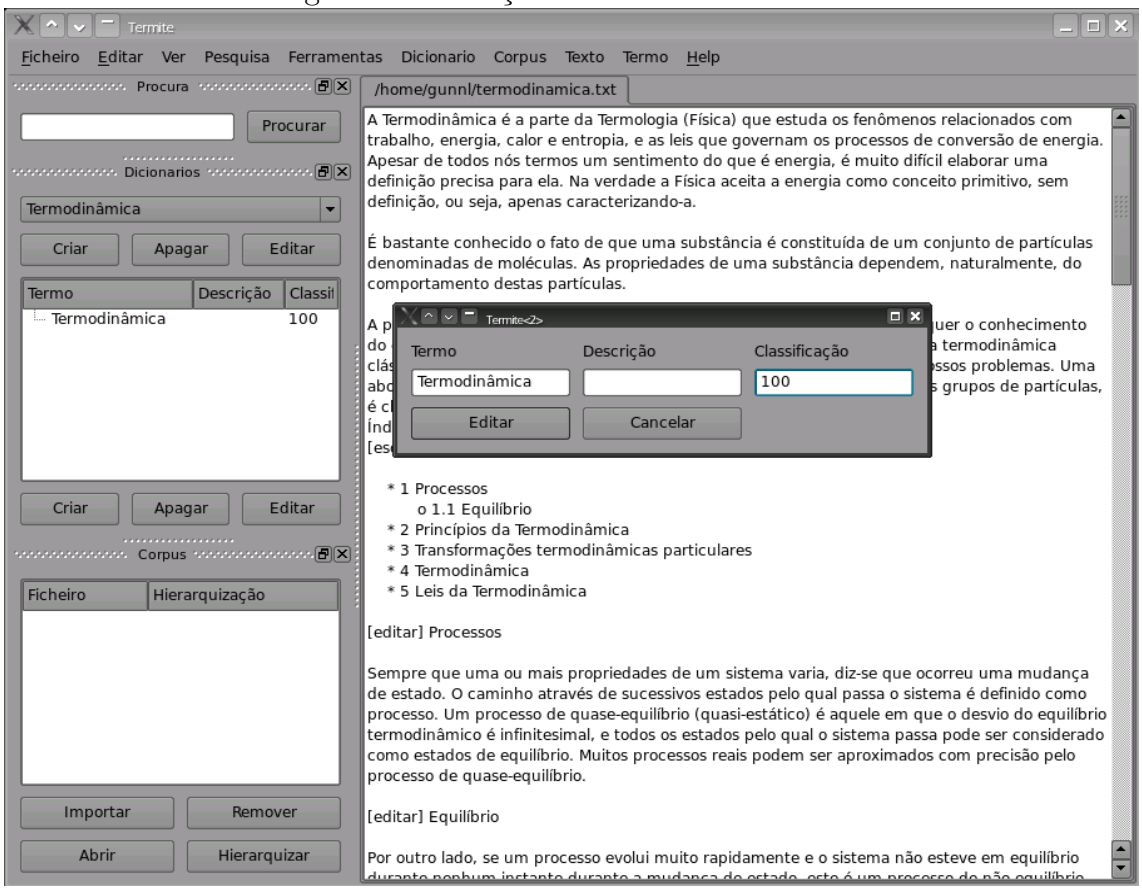
Figura 4.17: Adição de CCs à lista do *e-Termite*



4.2.3.4 Classificação

A classificação dos termos é um processo muito simples de efectuar, enquanto processo de execução informática, até porque terá de ser executado manualmente e diversas vezes. Graças à facilidade de utilização que o interface disponibilizado pelo *e-Termite* apresenta, o processo realiza-se em poucas acções. Para dar início à classificação de termos, é apenas necessário adicionar um termo pela primeira vez e a aplicação pede para criar um dicionário, caso não tenha sido criado e escolhido nenhum, pedindo apenas um nome para o mesmo (neste caso, designa o domínio (Termodinâmica)). Visto ter sido o termo que desencadeou a pesquisa, começa-se por adicionar a própria palavra «Termodinâmica» que, de acordo com as nossas classes é considerado um termo de nível 2, pois também existe com outras utilizações que não a de ciência (por exemplo, a eficiência termodinâmica de algo), mas por designar o domínio, terá uma importância acrescida e será integrada no nível 1, atribuindo-se-lhe, assim, a pontuação máxima de 100 pontos, como se pode observar na figura 4.18, onde se apresenta a entrada mais simples do dicionário.

Figura 4.18: Edição do termo no *e-Termite*



Desta forma, o texto passaria a somar, pelo menos, 100 pontos, ainda que se existissem mais ocorrências da palavra dicionarizada, elas seriam contabilizadas e efectuada a multiplicação necessária por todos as existências no texto. Sucessivamente, classificam-se os termos identificados e presentes no primeiro texto e introduzem-se no dicionário para que, de seguida, se possa reorganizar a lista de textos. A classificação pode ser efectuada num texto apenas ou na totalidade de textos importados para o programa, ainda que seja uma prática importante, sempre que se introduz um número considerável de termos no dicionário, proceder a uma nova contabilização e reordenação da hierarquia dos textos, carregando no botão "*Hierarquizar*".

4.2.3.5 Hierarquização

O processo de hierarquização permite uma renovação da ordem dos textos e procura contribuir para dois objectivos essenciais na investigação. O primeiro é actualizar a lista ordenada de candidatos ao *corpus* que pertencem ao domínio por nível de proximidade, tornando a lista mais actualizada e de acordo com os critérios definidos. O segundo é forçar a reorganização dos textos e disponibilizar, no início da lista, para análise seguinte, aqueles com maior potencial terminológico. Para efectuar uma reordenação, cada vez que se introduzir um termo ou um conjunto de termos no dicionário do programa, é necessário dar-lhe essa indicação, pressionando o botão «*Hierarquizar*» do menu respectivo. Nesse momento, o programa vai voltar a pontuar todos os ficheiros na sua base, de acordo com os termos e suas classificações no dicionário. A actualização regular da classificação dos textos é importante porque, ao colocar no topo da nossa lista os ficheiros com maior pontuação, permite ao investigador analisar sempre os textos mais bem pontuados do conjunto, logo os mais próximos do domínio. Fica ao critério do investigador visualizar ou não, nessa lista, os ficheiros já totalmente analisados, através da opção «*Listar analisados*», para que não se misturem visualmente com os textos à espera de análise.

Quando a lista de candidatos ao *corpus* (CCs) chegar aos 10 textos analisados, prevê-se que, de acordo com uma estimativa média não comprovada, em cerca de 250 unidades lexicais por candidato, se adicionem em média 15 CTs e se atinja o número aproximado de 150 termos no dicionário do domínio. Esta média será sempre inferior no início dado que o filtro ainda não funciona em pleno por ter poucas entradas, mas, à medida que os termos vão sendo adicionados, haverá uma tendência crescente para que os textos que são melhores, durante a fase da hierarquização, sejam "puxados" para o topo da lista. Quando se atingir um número de termos considerado relevante, a aplicação do filtro estará na sua fase de maior impacto na hierarquização dos CCs e deverá ser aplicado à parte da base textual que se achar conveniente (consideremos, por exemplo, cerca de 2000 ligações). De todos os textos processados pelo filtro, apenas 10 textos foram realmente analisados e os 1990 restantes serão processados

pelo filtro, sendo de esperar que alguns deles possam subir na lista por integrarem mais e "melhores" termos do que os 10 inicialmente processados.

Poderemos declarar a primeira fase de constituição do *corpus* como concluída de várias formas: estabelecer um limite mínimo de pontos para que um candidato possa integrar o *corpus*, por exemplo, 5000 pontos, e terminar a constituição quando há um número suficiente de textos que ultrapassem essa margem; seleccionar um número de candidatos a recolher, por exemplo, os 500 mais pontuados, quando se atingir um número pré-definido de textos processados, por exemplo, 5000, ou, simplesmente, usar todos os candidatos ao *corpus*, tornando-os MCs para que, de seguida, sejam anotados de acordo com os objectivos e se possa iniciar a segunda fase da constituição.

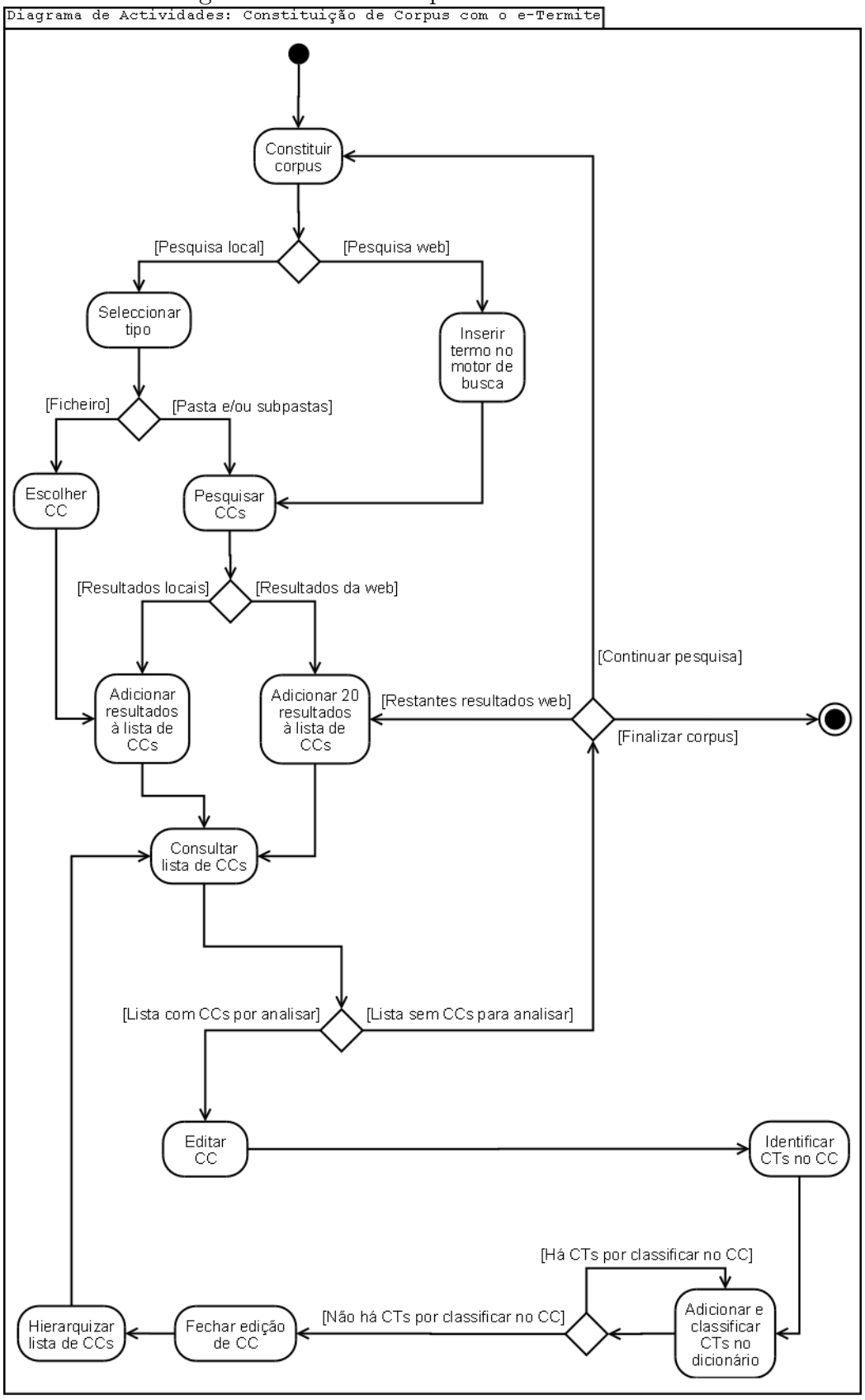
Observemos um diagrama de actividades na figura 4.19 que sintetiza num panorama mais alargado a articulação das funções básicas presentes na primeira fase de constituição de *corpus*.

Nem todos os trabalhos exigem que a constituição passe à segunda fase, sendo um dos objectivos primários da aplicação compilar um *corpus* mais geral, ainda que restrito ao domínio, numa primeira fase, para que se permita um grau de reutilização maior e só depois se apliquem as técnicas de anotação que permitem filtrar candidatos que correspondam a um número mais alargado de critérios.

4.2.4 Gestão de *corpus*

A etapa de gestão sucede à constituição do *corpus* e introduz procedimentos ao nível do tratamento e anotação dos textos, ao armazenamento e preservação dos dicionários, dos *corpora*, dos textos e dos termos e ao controlo das actualizações necessárias, tanto dos procedimentos do *software* como dos conteúdos. Uma boa gestão de um *corpus* pode torná-lo reutilizável em múltiplas investigações, em diferentes áreas e durante um largo período de tempo. Os processos de gestão podem influenciar a constituição do *corpus*, numa segunda fase, caso seja necessário,

Figura 4.19: Fluxo de processos no *e-Termite*



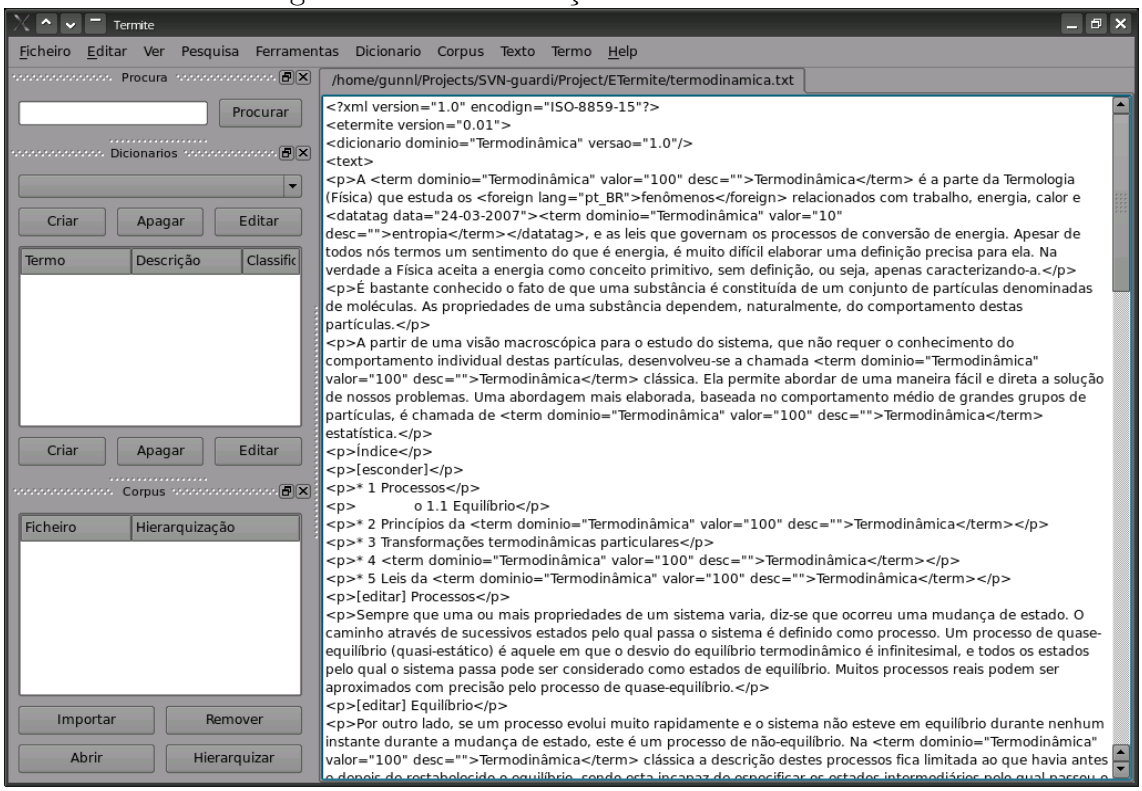
para a consecução do objectivo, proceder a uma delimitação das características do *corpus*, aplicando-se critérios complementares. Têm, também, uma importância fundamental no que o *corpus* pode revelar sobre uma série de aspectos, linguísticos e extra-linguísticos, que são detectados durante a aplicação das ferramentas de análise, após a anotação dos textos.

4.2.4.1 Anotação

Neste nosso exemplo, procederemos a uma anotação simples para que se perceba, até que ponto pode ser útil esta fase. O processo de anotação, como já foi discutido anteriormente, consiste em anexar informação adicional ao texto que permita proceder a análises estatísticas simples, como, por exemplo, saber o número de ocorrências de determinada unidade lexical, ou mais complexas, como quantificar as vezes que um determinado termo ocorre junto de preposições e recolher esses contextos para uma análise mais cuidada.

Para além da anotação automática, que será efectuada pela aplicação cada vez que um termo existente no dicionário de domínio com uma determinada classificação constar no texto em edição, etiquetando-o com essa mesma informação, existem ainda mais dois tipos de anotação, uma semi-automática e outra manual. A anotação semi-automática é executada pelo *software* cada vez que se cria um campo na base de dados e se atribui um termo como ligado a esse campo. Por exemplo, se criarmos um campo designado "autor" na base de dados, será possível acrescentar a informação de que um determinado termo foi criado por certo autor. Além disso, podemos criar etiquetas manualmente para acrescentar informação adicional, como, por exemplo, a data em que etiquetamos o termo, aquilo que se poderia considerar meta-etiquetagem, sem recorrer à base de dados. Reparemos na figura 4.20 para diferenciarmos os tipos de anotação aqui apresentados. A etiqueta "foreign" refere-se à variante do português do Brasil, enquanto a etiqueta "datatag" permite reter a informação sobre a data em que se procedeu à etiquetagem do termo.

Figura 4.20: Modo edição XML do *e-Termite*



Quando se termina o processo de anotação e toda a informação necessária está anexada ao texto, podemos prosseguir para a análise.

4.2.4.2 Análise

O processo de análise serve para encontrar nos textos ou no *corpus* informação relevante para a análise linguística. Já apresentámos anteriormente as ferramentas que se consideram essenciais, no quadro epistemológico de referência, para efectuar uma análise textual e terminológica. Dado que essas ferramentas continuam por implementar e não são o objectivo primordial da aplicação, não serão objecto de análise aprofundada nesta dissertação.

4.3 *Software* para *corpora*

Uma busca na Internet por ferramentas de trabalho para *corpora* rapidamente se depara com uma enorme quantidade de *software* disponível e disperso pelas mais

variadas categorias. Desde *software* para efectuar concordâncias, até análises sintácticas e anotações, múltiplas são as funções para as quais podemos encontrar programas informáticos. Para facilitar a nossa breve análise, que não pretende efectuar um exercício descritivo, nem avaliativo do *software* existente, poderemos dividir genericamente os programas que encontrámos em dois grupos principais, distribuindo-os por ferramentas para constituição e para gestão de *corpora*.

Pela consulta de páginas com listas de *software* para *corpora* ou a partir da leitura de artigos que referem ferramentas de análise disponíveis, é fácil verificar que grande parte do *software* é direccionado para a gestão e não para a constituição. Há razões que podem justificar a diferença de valores, entre as quais, o superior número de investigadores da língua geral que já dispõem de vários e enormes *corpora* pré-compilados que podem ser acedidos gratuitamente *online* ou descarregados para o computador. Os linguistas de especialidade, no entanto, não sentem essa facilidade e, a maior parte das vezes, por especificidades inerentes ao trabalho, vêem-se forçados a constituir o seu próprio *corpus*.

A possibilidade de construir um *corpus* manualmente quase sempre esteve vedada a grande parte dos estudiosos da língua que não possuíam, na maioria das ocasiões, recursos para tal. Além de ser difícil e moroso recolher textos suficientes, o *hardware* capaz de executar esse tipo de tarefas não estava acessível a qualquer um. Contudo, as evoluções tecnológicas ao nível dos computadores e a facilidade de informatização dos recursos linguísticos ou de consulta aos que já estão informatizados ditam novas leis no trabalho com *corpora*. Como Maia refere:

> «*I particularly feel for those who find that all the blood, sweat and tears they put into building up a corpus manually, or with prehistoric forms of hardware and software, are now rendered obsolete by modern IT. I know how they feel. To use a now popular phrase - "Been there - done that!" Modern technology, however, is making the making of small specialised corpora much easier.*» (Maia, 1997)

Não é, por isso, surpreendente que vão surgindo, aos poucos, recursos que apontem para a execução de um trabalho completo por parte do terminólogo, cada vez mais auxiliado pelo computador, na constituição e na gestão de *corpora*. Também em Portugal, se encontram iniciativas que incentivam o trabalho do terminólogo a ser realizado na íntegra por si próprio e com recurso a ferramentas automáticas, de entre as quais destacaríamos o *Corpógrafo*.

O *Corpógrafo*[5] é um *software* que oferece recursos para a construção de *corpora* a partir da *web*. Este projecto desenvolveu-se aos poucos e atingiu uma dimensão considerável, quer ao nível de utilizadores, quer pelo número de funções e recursos que disponibiliza. O *Corpógrafo* foi desenvolvido na Faculdade de Letras da Universidade do Porto e construído a partir do conceito de «*Do-It-Yourself Corpora*» de Belinda Maia Maia (1997), apresentando como seus objectivos principais:

- Ajudar o utilizador na pesquisa e não substituir o utilizador.
- Acelerar o processo manual de constituição de *corpus*.

Durante uma conferência sobre o *Corpógrafo* no Rio de Janeiro em Maio de 2006, Luís Sarmento, um dos investigadores envolvidos na criação do projecto, sublinhou os pontos fortes: simples para o utilizador, técnicas simples de PLN, aplicações práticas de *corpora*, aprendizagem colectiva, criação de comunidade, criação de recursos de terminologia e o apoio pedagógico. E apontou também alguns pontos fracos: complexidade técnica do sistema, dependências externas, dificuldade em suportar todos os utilizadores, dificuldade de instalação de novos servidores e uma ainda pouca aplicação dos recursos produzidos.

O *Corpógrafo* encontra-se na versão 3 e permite que gratuitamente se proceda à inscrição e se possa fazer uso das suas ferramentas *online*, sendo um caso paradigmático de aproximação do *software* às necessidades do terminólogo e de uma redistribuição do peso da constituição e da gestão, enquanto fases igualmente importantes na investigações terminológica e linguística.

[5] http://www.linguateca.pt/Corpografo/

4.4 Síntese

Tendo em conta a utilização cada vez maior da Internet como ponto de partida para a constituição de *corpora*, dado que o número de textos disponíveis para consulta instantânea é significativo, há uma necessidade emergente de criar ferramentas que procedam a uma selecção criteriosa dos textos para que o terminólogo não se sinta ultrapassado pela quantidade de informação ao seu dispor, nem perca tempo desnecessário a consultar textos inadequados para a sua pesquisa. A concepção de uma aplicação informática semi-automática que permita facilitar a forma como o terminólogo procede à preparação do *corpus* sobre a área de estudo é indispensável para transformar a constituição e a gestão de *corpora* em processos rápidos e eficazes.

De entre as prioridades funcionais que uma aplicação na área deve ter, destacam-se a capacidade de lidar com grandes quantidades de informação e proceder a uma selecção textual pertinente para os objectivos, permitindo a aplicação de critérios terminológicos, fraseológicos, lexicais, sintácticos, semânticos e pragmáticos. Apesar de ser concebido idealmente para terminólogos, o resultado deste trabalho poderá ser de interesse para outras áreas, dado que a Terminologia se cruza com muitas outras áreas do conhecimento. De acordo com a concepção idealizada, o produto final apresentará duas formas: a de um *corpus* com textos hierarquizados por ordem de relevância no domínio de estudo e um dicionário, composto por termos classificados de acordo com a maior exclusividade de utilização no domínio e anotados com informação, linguística ou extra-linguística, que se considere importante. Para que o funcionamento dos processos atinja os objectivos a que se propõe é necessário que se entenda a construção do dicionário do domínio como gradual e decorrente dos textos que são analisados, sendo, por isso, importante chegar aos textos mais adequados, como se demostrou no capítulo anterior. Assim, o dicionário vai tornar-se num melhor filtro e hierarquizar os textos de modo a que os mais próximos de domínio surjam para análise em primeiro lugar da lista de candidatos ao *corpus*.

A cada termo introduzido, os textos ficam mais bem ordenados e, assim, torna-se também mais fácil encontrar termos do domínio. Esta dependência mútua é o ponto forte da análise, que quanto mais é trabalhada, mais apurada e significativa se torna. Não existindo um filtro que exclui, a não ser no primeiro momento, mas filtros que ordenam, pois cada termo é um filtro, não há textos excluídos, mas sim textos ordenados. Espera-se, assim, atingir uma taxa de cem por cento de precisão na busca de candidatos, ainda que provavelmente o ruído gerado seja considerável. Com a utilização progressiva do dicionário do domínio para filtrar os textos espera-se que ocorra uma selecção e coloque no final da lista os textos menos importantes, ou seja, os que mais provavelmente serão ruído. No entanto, tendo em conta que o objectivo é a redução de tempo, terá sempre de ser efectuada uma comparação entre as duas formas de constituir *corpora*, a tradicional e a do *e-Termite*, que ainda não é possível dado ao estatuto prototípico do *software*, analisando o número final de textos em cada um, o número de termos identificados e o tempo dispendido na transição de base textual até ser um *corpus*.

Capítulo 5

Conclusões e limitações do estudo

Apresentámos nesta dissertação um projecto de concepção de *software* para utilização em Terminologia, tentando sintetizar, nas conclusões, o percurso efectuado e explicitando cada uma das ideias mais importantes que foram sendo apresentadas e discutidas ao longo do trabalho. Por se tratar de uma concepção de um protótipo de *software*, existem limitações inerentes à verificação prática que impedem a constatação de determinados processos como inequivocamente certos ou errados, mas há, contudo, pistas que sugerem reflexão e necessidade de resolver questões que, por motivos epistemológicos ou metodológicos, devem ser alvo de referência e consideração, como iremos ver mais adiante.

Este é um projecto que procurou o desenvolvimento de um conceito metodológico e que culmina com a apresentação formal de um modelo de concepção para constituição e gestão semi-automática de *corpora* de especialidade. Será necessário reavaliar este modelo e compreender se o projecto *e-Termite*, agora que termina a sua apresentação, se adequa ao objectivo pressuposto inicialmente e se tem suporte para atingir os seus propósitos, pois ainda que esteja numa fase alfa de desenvolvimento informático, tem um conceito visível e aplicável que permite aferir da sua exequibilidade.

No início deste trabalho, procurámos compreender a instabilidade teórica que a Terminologia, durante algum tempo, atravessou e analisámos a incessante busca

por bases epistemológicas sólidas dos vários movimentos, na tentativa de definir uma prática metodológica, que tem sido, ela própria, um dos principais factores de renovação. Comprovámos a importância da interdisciplinaridade da Terminologia, cuja prática tem sido alvo de interesse constante por parte de outras áreas que nela procuram uma ferramenta de reconstrução conceptual através da investigação terminológica. Discutimos a importância crescente da análise textual e a sua integração nos estudos terminológicos, na senda do que já havia sucedido na Linguística, a definição de um novo quadro teórico e uma remodelação efectiva dos métodos utilizados, observando-se uma transição da análise frásica para a textual. Paralelamente, apontámos o crescimento de importância do produto comunicativo real, que surgiu como reacção ao movimento linguista introspectivo e que promoveu, indirectamente, o aparecimento da Linguística de *Corpus*.

Procurámos demonstrar a relevância da Informática, produto das grandes evoluções tecnológicas e científicas, que proporcionou à Terminologia Textual e à Linguística de *Corpus* a estrutura e suporte técnicos necessários para a implementação dos seus complexos e pesados processos. A partir dos resultados atingidos actualmente, observámos que o diálogo entre estas duas áreas tem sido produtivo, mas está longe de ser concluído, prevendo-se que um desenvolvimento maior da interacção entre a Linguística e a Informática venha a provar-se uma aposta de sucesso.

Nesta fase em que o *corpus* continua a assumir um destaque na metodologia de análise linguística, torna-se necessário aprofundar o estudo dos critérios de classificação, de modo a que a investigação possa ser efectuada com melhores resultados e mais rapidamente e tire verdadeiro partido dos recursos informáticos. Neste sentido, apresentámos, assim, uma concepção de protótipo de *software* que procura simplificar e tornar mais eficazes os processos de constituição e gestão de *corpora* de especialidade, optimizando a tarefa dos terminólogos.

O *e-Termite* preconiza um modelo que privilegia a prática da investigação terminológica, defendendo a flexibilidade como paradigma estrutural para que a sua acção não se limite a um grupo restrito. Expusemos uma descrição detalhada da

aplicação informática e apresentámos os objectivos e cada uma das funções mais importantes, descrevendo um exemplo de uso com todos os passos detalhados para que se possam compreender melhor os mecanismos de constituição e gestão defendidos.

Contudo, alguns dos pontos base que servem de fundamentação ao *e-Termite* continuam em discussão actual e, portanto, estão sujeitos a apresentação de uma argumentação contrária. Parte-se, por exemplo, do pressuposto que o conceito de «termhood» é funcional, ainda que necessite de ser mais clarificado no critérios que o fundamentam, na definição do relacionamento de proximidade entre o termo e a especialidade que integra e na transparência das variáveis que o compõem para que possa ser melhor compreendido e formalizado. Assume-se ainda que os termos têm tendência para coexistir nos mesmos textos, no entanto, sem fundamentação crítica que suporte essa opção, sendo que esta premissa é basilar para o conceito de funcionamento do *e-Termite*.

Há, também, questões epistemológicas quanto à definição do conceito de termo e o próprio estatuto da unidade terminológica, que condicionam a forma como a identificação no texto pode ser efectuada. A aceitação de que a unidade terminológica é uma unidade lexical, mas com uma actualização particular dentro do contexto de uma utilização especial, ainda que inquestionável, no quadro que defendemos, levanta sérios problemas na definição de métodos seguros, eficazes e adequados para delimitar os termos e que proporcionem uma identificação clara e rápida em contexto de especialidade.

Podem, ainda, ser levantadas outras questões que dizem respeito à definição de um conceito de texto que seja funcional, principalmente, no contexto da Internet, pois apresentam-se muitos obstáculos à identificação delimitada e formalizada, por questões de pontuação e de dimensão, para que a conversão dos textos em unidades homogéneas permita um trabalho rigoroso e eficaz. Além de que há situações, como a possível quebra dos direitos de autor sobre os textos que existem na Internet e que são recolhidos sem permissão, que não estão completamente resolvidas.

A juntar a algumas das limitações epistemológicas, podemos encontrar também

situações procedimentais que devem ser analisadas e revistas, sendo a principal o facto de ainda não existir uma versão funcional do *software* para que se possa testar efectivamente todo o processo. Há, no entanto, outras questões como, por exemplo, a necessidade de formalização do método que determina a escolha da unidade terminológica que dá início a todo o processo de constituição de *corpus*, a importância da criação de uma estrutura base que permita formalizar ontologias ou até a questão da inclusão de um sistema que permita integrar outras categorias definidas pelo investigador como válidas para pontuar os candidatos ao *corpus*.

Consideramos, no entanto, os resultados obtidos encorajadores e, acima de tudo, indicadores de que a concepção idealizada irá na prática ser concretizada e atingir os objectivos a que se propõe, diminuindo efectivamente o tempo que decorre no processo de constituição e de gestão de um *corpus*. Existe, contudo, um largo percurso a efectuar, ficando em aberto a continuação e desenvolvimento de mais actividades no âmbito do projecto *e-Termite*.

Bibliografia

Adam, J.-M. (1990), *Éléments de Linguistique Textuelle*, Mardaga, Liège.

Adam, J.-M. (1999), *Linguistique Textuelle - Des Genres de Discours aux Textes*, Nathan, Paris.

Antia, B. E. (2000), *Terminology and Language Planning*, John Benjamins, Amsterdam.

Aston, G., Bernardini, S. & Stewart, D. (2004), *Corpora and Language Learners*, John Benjamins, Amsterdam.

Atkins, S., Clear, J. & Ostler, N. (2000), 'Corpus Design Criteria', in *G. Dixon, ed., Journal of Literary and Linguistic Computing*, Oxford University Press, Oxford, pp. 1–16.

Auger, P. (1998), 'La terminologie au Quebec et dans le monde, de la naissance à la maturité', in *Actes du sixiéme colloque OLF-STQ de terminologie. L'ère nouvelle de la terminologie.*, Quebec: Gouvernment du Quebec, pp. 27–59.

Aussenac-Gilles, N. & Bourigault, D. (2003), 'Construction d'ontologies à partir de textes', in *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, Batz-sur-Mer, pp. 27–50.

Bakhtine, M. (1978), *Esthétique et Théorie du Roman*, Gallimard, France.

Bakhtine, M. (1984), *Esthétique de la Création Verbale*, Gallimard, France.

Baptista, J. (2000), *Sintaxe dos predicados nominais construídos com o verbo-suporte SER DE*, Tese de Doutoramento, Universidade do Algarve, Faro.

Baroni, M. & Ueyama, M. (2006), 'Building general and special purpose corpora by web crawling', *Language Corpora: Their Compilation and Application* pp. 31–40. **URL:** *http://tokuteicorpus.jp/result/pdf/2006_004.pdf [09-11-2007]*

Bergenholtz, H. & Tarp, S., eds (1995), *Manual of Specialized Lexicography*, John Benjamins, Amsterdam.

Biber, D., Conrad, S. & Reppen, R. (1998), *Corpus Linguistics: investigating language structure and use*, Cambridge Univesity Press, Cambridge.

Bilger, M. (2000), *Corpus, Méthodologie et applications linguistiques*, Champion, Paris.

Boulanger, J.-C. (1995), 'Comptes rendus', *Meta* **Vol.40(1)**, 133–137. **URL:** *http://www.erudit.org/revue/meta/1995/v40/n1/002116ar.pdf [09-11-2007]*

Bourigault, D. & Jacquemin, C. (2000), 'Construction de ressources terminologiques', in *J.-M. Pierrel, ed., Ingénierie des langues*, Hermès, pp. 215–233.

Bourigault, D. & Jacquemin, C. (2003), 'Term Extraction and Automatic Indexing', in *R. Miktov, ed., The Oxford Handbook of Computational Linguistics*, Oxford University Press, London, pp. 599–615.

Bourigault, D., Jacquemin, C. & L'Homme, M.-C. (2001), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam.

Bourigault, D. & Slodzian, M. (1998), 'Pour une terminologie textuelle', *Terminologies nouvelles* **19**, 29–32. **URL:** *http://www.cfwb.be/franca/termin/charger/rint19.pdf [09-11-2007]*

Bowker, L. (2003), 'Specialized lexicography and specialized dictionaries', in *P. van Sterkenburg, ed., A Practical Guide to Lexicography*, John Benjamins, Amsterdam, pp. 154–164.

Bowker, L. & Pearson, J. (2002), *Working with Specialized Language: A practical guide to using corpora*, Routledge, London.

Bronckart, J.-P. (1996), *Activité Langagière, Textes et Discours*, Delachaux et Niestlé, Paris.

Budin, G. (2002), 'Global Content Management - Challenges and Opportunities for Creating and Using Digital Translation Resources', in *Proceedings of the Workshop "Language Resources in Translation Work and Research" a pre-conference workshop to LREC 2002: Third International Conference on Language Resources and Evaluation.*, pp. 57–61.
URL: *http://www.ifi.unizh.ch/cl/yuste/postworkshop/repository/proceedings.pdf [09-11-2007]*

Burnard, L. & Sperberg-McQueen, C. M. (2002), 'Eagles', TEI Lite: An Introduction to Text Encoding for Interchange.
URL: *http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_en.pdf [09-11-2007]*

Béjoint, H. & Thoiron, P. (2000), 'Les sens des termes', in *H. Béjoint & P. Thoiron, eds, Le Sens en Terminologie*, Presses Universitaires de Lyon, Lyon, pp. 5–19.

Cabré, M. T. (1999), *Terminology - Theory, methods and applications*, John Benjamins, Amsterdam.

Cabré, M. T. (2003), 'Theories of terminology: Their description, prescription and explanation', *Terminology* **9**, 163–199.
URL: *http://www.hf.uib.no/forskerskole/cabre.pdf [09-11-2007]*

Castagnoli, S. (2006), 'Using the web as a source of LSP corpora in the terminology classroom', in *M. Baroni & S. Bernardini, eds, Wacky! Working papers on the Web as Corpus*, Gedit, Bologna, pp. 159–172.

Chien, L.-F. & Chen, C.-L. (2001), 'Incremental extraction of domain-specific terms from online text resources', in *D. Bourigault, C. Jacquemin & M.-C. L'Homme, eds, Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, pp. 89–109.

Conceição, M. C. (2001), *Termes et Reformulations*, Tese de Doutoramento, Universidade Nova de Lisboa, Lisboa.

Conceição, M. C. (2005), *Concepts termes et reformulations*, Presses Universitaires de Lyon, Lyon.

Cook, V. J. & Newson, M. (1996), *Chomsky's Universal Grammar*, Blackwell, Oxford.

Costa, M. R. (2001), *Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas*, Tese de Doutoramento, Universidade Nova de Lisboa, Lisboa.

Coutinho, M. A. (2003), *Texto(s) e Competência Textual*, Fundação Calouste Gulbenkian, Lisboa.

de Bessé, B. (2000), 'Le domaine', in *H. Béjoint & P. Thoiron, eds, Le Sens en Terminologie*, Presses Universitaires de Lyon, Lyon, pp. 182–197.

Depecker, L. (2000), 'Le signe entre signifié et concept', in *H. Béjoint & P. Thoiron, eds, Le Sens en Terminologie*, Presses Universitaires de Lyon, Lyon, pp. 86–126.

Depecker, L. (2003), *Entre signe et concept: Éléments de terminologie générale*, Presses Sorbonne Nouvelle, Paris.

EAGLES (1996), 'Eagles'.

URL: `http://www.ilc.cnr.it/EAGLES96/browse.html` *[09-11-2007]*

Frantzi, K., Ananiadou, S. & Tsujii, J. (1999), 'Classifying Technical Terms', in *J. W. T. Smith, A. Ardo & P. Linde, eds, Redefining the Information Chain - New Ways and Voices*, ICCC Press, pp. 144–155.
URL: *http://elpub.scix.net/data/works/att/9915.content.00351.pdf [09-11-2007]*

Frey, C. & Latin, D. (1997), *Le corpus lexicographique: Méthodes de constitution et de gestion*, Duculot.

Gaudin, F. (2003), *Socioterminologie - Une approche sociolinguistique de la terminologie*, Duculot, Bruxelles.

Grishman, R. (2003), 'Information Extraction', in *R. Miktov, ed., The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, pp. 545–559.

Habert, B., Nazarenko, A. & Salem, A. (1997), *Les linguistiques de corpus*, Armand Colin, Paris.

Halliday, M. A. K. (1994), *An introduction to functional grammar*, Arnold, London.

Halliday, M. A. K. & Hassan, R. (1976), *Cohesion in english*, Longman, London.

Halliday, M. A. K. & Teubert, W. (2004), *Lexicology and Corpus Linguistics: An Introduction*, Continuum Intl Pub Group, London.

Honeste, M. L. (2003), 'Polysémie et référence', in *S. Rémi-Giraud & L. Panier, eds, La polysémie ou l'empire de sens*, Presses Universitaires de Lyon, Lyon, pp. 149–156.

Hunston, S. (2002), *Corpora in Applied Linguistics*, Cambridge University Press, Cambridge.

Jacques, M.-P. (2005), 'Pourquoi une Linguistique de *Corpus*?', in *G. Williams, ed., La Linguistique de Corpus*, Press Universitaire de Rennes, Rennes, pp. 21–30.

Jones, R. & Ghani, R. (2000), 'Automatically Building a Corpus for a Minority Language from the Web', *Poster paper in proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

URL: *http://citeseer.ist.psu.edu/316887.html [09-11-2007]*

Kageura, K. & Umino, B. (1996), 'Methods of automatic term recognition: a review', *Terminology* **3**(2), 259–289.

URL: *http://citeseer.ist.psu.edu/kageura96methods.html [09-11-2007]*

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*, Longman, London.

Kilgarriff, A. (2007), 'Googleology is Bad Science', *Computational Linguistics* **Vol.33(1)**, 147–151.

URL: *http://www.kilgarriff.co.uk/Publications/2007-K-CL-Googleology.dvi [09-11-2007]*

Kilgarriff, A. & Greffenstette, G. (2003), 'Web as corpus', *Computational Linguistics* **Vol.29(3)**.

URL: *http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf [09-11-2007]*

Kit, C. (2002), 'Corpus tools for retrieving and deriving termhood evidence', *The 5th East Asia Forum of Terminology* pp. 69–80.

URL: *http://personal.cityu.edu.hk/~ctckit/papers/termhood.pdf [09-11-2007]*

Law, V. (2003), *The History of Linguistics in Europe*, Cambridge, London.

Lee, D. Y. W. (2001), 'Genres, Registers, Text Types, Domains, and Styles: clarifying the concepts and navigating a path through the BNC jungle', *Language Learning & Technology* **Vol.5(3)**, 37–72.

URL: *http://llt.msu.edu/vol5num3/lee/ [09-11-2007]*

L'Homme, M.-C. (1998), 'A lexico-semantic approach to the structuring of terminology', *Computerm 2004* pp. 7–14.

Lino, M. T. (2000), 'Terminologia e Indústrias das Línguas', in *M. Correia, ed., Terminologia e Indústrias das Línguas*, ILTEC, Lisboa, pp. 25–40.

Lyons, J. (1970), *As Idéias de Chomsky*, Cultrix, São Paulo.

Maher, J. & Groves, J. (1996), *Chomsky for Beginners*, Icon Books, Cambridge.

Maia, B. (1997), 'Do-it-yourself corpora ... with a little bit of help from your friends!', *PALC '97 Practical Applications in Language Corpora* pp. 403–410.
URL: *http://web.letras.up.pt/bhsmaia/belinda/pubs/PALC-1997.DOC [09-11-2007]*

Maia, B. (2002), 'Do-it-yourself, disposable, specialized mini corpora - where next? reflections on teaching translation and terminology through corpora', *Cadernos de Tradução* **No.IX**.
URL: *http://web.letras.up.pt/bhsmaia/belinda/pubs/PALC-1997.DOC [09-11-2007]*

Malmkjaer, K. (2004), *The Linguistics Encyclopedia*, Routledge, London.

Mamede, N. J., Baptista, J., Trancoso, I. & das Graças Volpe Nunes, M., eds (2003), *Computational Processing of the Portuguese Language*, Springer, Berlin.

McEnery, T. (2003), '*Corpus* Linguistics', in *R. Miktov, ed., The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, pp. 448–463.

Meyer, I. (2001), 'Extracting knowledge-rich contexts for terminography', in *D. Bourigault, C. Jacquemin & M.-C. L'Homme, eds, Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, pp. 279–302.

Meyer, I., Skuce, D., Bowker, L. & Eck, K. (1992), 'Towards a new generation of terminological resources: An experiment in building a terminological knowledge

base', *Proceedings of the 14th International Conference on Computational Linguistics* **Vol.40(1)**, 956–960.
URL: *http://acl.ldc.upenn.edu/C/C92/C92-3146.pdf [09-11-2007]*

Mitkov, R., ed. (2003), *The History of Linguistics in Europe*, Oxford, London.

Nakagawa, H. (2001), 'Experimental evaluation of ranking and selection methods in term extraction', in *D. Bourigault, C. Jacquemin & M.-C. L'Homme, eds, Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, pp. 303–325.

Nunan, D. (1993), *Introducing Discourse Analysis*, Penguin, London.

Oakes, M. P. (1998), *Statistics for corpus Linguistics*, Edinburgh University Press, Edinburgh.

Ooi, V. B. Y. (1998), *Computer Corpus Lexicography*, Edinburgh University Press, Edinburgh.

Pearson, J. (1998), *Terms in Context*, John Benjamins, Amsterdam.

Perelman, C. (1993), *Império retórico / retórica e argumentação*, Edições Asa, Porto.

Phillips, L. & Jorgensen, M. W. (2002), *Discourse Analysis as Theory and Method*, SAGE, London.

Rastier, F. (2001), *Arts et Sciences du Texte*, Presses Universitaires de France, Paris.

Rastier, F. (2003), 'Les valeurs et l'evolution des classes lexicales', in *S. Rémi-Giraud & L. Panier, eds, La polysémie ou l'empire de sens*, Presses Universitaires de Lyon, Lyon, pp. 39–56.

Rastier, F. (2005), 'Enjeux épistémologiques de la Linguistique de *Corpus*', in *G. Williams, ed., La Linguistique de Corpus*, Press Universitaire de Rennes, Rennes, pp. 31–45.

Rey, A. (1979), *La Terminologie - Noms et notions*, Presses Universitaires de France, Paris.

Russell, S. J. & Norvig, P. (1995), *Artificial intelligence: A Modern Approach*, Prentice Hall, New Jersey.

Ruwet, N. & Chomsky, N. (1979), *A Gramática Generativa*, Edições 70, Lisboa.

Sarmento, L. (2006), 'Corpógrafo - um ambiente livre para o ensino e desenvolvimento de terminologia'.
URL: *http://www.linguateca.pt/documentos/corpografo_maio_2006.pdf [09-11-2007]*

Sharoff, S. (2006), 'Creating general-purpose corpora using automated search engine queries', in *M. Baroni & S. Bernardini, eds, Wacky! Working papers on the Web as Corpus*, Gedit, Bologna, pp. 159–172.

Sinclair, J. (1991), *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Sinclair, J. (2003), '*Corpora* for lexicography', in *P. van Sterkenburg, ed., A Practical Guide to Lexicography*, John Benjamins, Amsterdam, pp. 167–178.

Slodzian, M. (2000), 'L'Émergence d'une Terminologie Textuelle et le Retour du Sens', in *H. Béjoint & P. Thoiron, eds, Le Sens en Terminologie*, Presses Universitaires de Lyon, Lyon, pp. 61–85.

Slodzian, M. (2006), 'La terminologie, historique et orientations'.
URL: *http://www.sdc2006.org/cdrom/contributions/Slodzian_SDC2006.pdf [09-11-2007]*

Sommers, H. (1996), *Terminology, LSP and Translation - Studies in language engineering in honour of Juan .C. Sager*, John Benjamins, Amsterdam.

Stubbs, M. (2001), *Words and Phrases: Corpus Studies in Lexical Semantics*, Blackwell, Oxford.

Temmerman, R. (2000), *Towards new ways of terminology description - The sociocognitive-approach*, John Benjamin, Amsterdam.

Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*, John Benjamins, Amsterdam.

Trask, R. L. & Mayblin, B. (2000), *Introducing Linguistics*, Icon Books.

Trask, R. L. & Mayblin, B. (2001), *The Handbook of Linguistics*, Blackwell, Oxford.

Tzoukermann, E., Klavans, J. L. & Strzalkowski, T. (2003), 'Information Retrieval', in *R. Miktov, ed., The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, pp. 529–544.

van Dijk, T. A. (1977), *Text and Context / Explorations in the Semantics and Pragmatics of Discourse*, Delachaux et Niestlé, London.

van Sterkenburg, P. (2003), *A Practical Guide to Terminology*, John Benjamins, Amsterdam.

Vossen, P. (2003), 'Ontologies', in *R. Miktov, ed., The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, pp. 464–482.

Wright, S. E. & Budin, G. (1997*a*), *Handbook of Terminology Management*, Vol. Volume 1, John Benjamins, Amsterdam.

Wright, S. E. & Budin, G. (1997*b*), *Handbook of Terminology Management*, Vol. Volume 2, John Benjamins, Amsterdam.

Wüster, E. (1985), *Einführung in die allgemeine terminologielehre und terminologische lexikographie*, Handelshochschule Kopenhagen, Kopenhagen.

Wüster, E. (1996), 'La teoria general de la terminologia: una zona fronterera entre la lingüistica, la lògica, l'ontologia, la informàtica i les ciènces especialitzades', in *M. T. Cabré, ed., Terminologia: Selecció de textos d'E. Wüster*, Servei de Llengua Catalana, Barcelona, pp. 153–204.

Wüster, E. (1998), *Introducción a la teoría general de la terminología y a la lexicografía terminológica*, IULA, Barcelona.

Lista de Tabelas

Lista de Figuras