PER-FIDE: PROJETO DE COMPILAÇÃO DE UM *CORPUS* MULTILINGUE

José João Alameida, Sílvia Araújo, Idalete Dias e Ana Correia
UNIVERSIDADE DO MINHO

1. Caraterização do Corpus *Per-Fide*: línguas e domínios textuais

O projeto que aqui apresentamos resulta da colaboração entre o Departamento de Informática e o Instituto de Letras e Ciências Humanas da Universidade do Minho e tem por principal objetivo a compilação de corpora paralelos multilingues. Pretende-se, com este tipo de recurso, construir uma plataforma de investigação vocacionada para diferentes áreas, nomeadamente a linguística contrastiva, o ensino e o estudo de línguas e da tradução (Berber Sardinha, 2003), da lexicografia bilingue, entre outras.

O Corpus *Per-Fide* vem estabelecer uma rede de relações entre a língua portuguesa e outras seis línguas, nomeadamente Espanhol, Russo, Francês, Italiano, Alemão e Inglês. O corpus contempla diversas combinações linguísticas em que o Português (nas suas diferentes variantes: Português Europeu, Português do Brasil e de África) surge como língua pivô, podendo funcionar quer como língua de partida quer como língua de chegada.

Os corpora compilados contêm textos originais nas sete línguas a par das respetivas traduções no maior número possível de línguas. O leque de textos contemplados pelo projeto inclui romances contemporâneos e contos; textos religiosos (principalmente Encíclicas, Cartas e Angelus extraídos do website do Vaticano); artigos jornalísticos (*Le Monde Diplomatique*, *Le Courrier International*, *Presseurop*); textos jurídicos (direito comunitário e acordos internacionais) e textos de cariz técnico (textos e documentação técnica especializada

nos domínios da indústria automóvel, eletrónica, telecomunicações, informática, normalização, indústria farmacêutica e medicina).

1.1. *Contributo do Corpus* Per-Fide

A expressão da língua portuguesa nos corpora existentes é limitada. Um dos aspetos de base do Corpus *Per-Fide* consiste no papel preponderante atribuído ao português, nas suas diferentes variantes, como língua de partida ou de chegada. Foram desenvolvidos alguns projetos monolingues em português pela *Linguateca*[1] tais como o *CETEMPúblico* e o *CETENFolha*, que são corpora jornalísticos para o português europeu e português do Brasil, respetivamente. No âmbito da *Linguateca*, foi desenvolvido outro projeto de compilação de corpora jornalísticos em colaboração com as edições portuguesa e francesa do jornal *Le Monde Diplomatique*. Os textos fornecidos, que correspondem aos artigos publicados entre 1999 e 2002, foram alinhados ao nível da frase e o resultado está disponível em http://linguateca.di.uminho.pt/nat/nat.pl, selecionando o corpus LMD-PT-FR. Esta colaboração foi agora retomada a fim de integrar um maior número de textos, de edições anteriores e mais recentes, em francês e português, bem como alargar o trabalho de compilação às restantes cinco línguas do projeto. Da *Linguateca* fazem parte outros corpora de menores dimensões, incluindo corpora de literatura clássica, orais e políticos. Sob a tutela do Centro de Linguística da Universidade de Lisboa (CLUL), foi construído o *Corpus de Referência do Português Contemporâneo*. O *Corpus do Português*, composto por 45 milhões de palavras, foi criado por Mark Davies (Brigham Young University) em colaboração com Michael Ferreira (Georgetown University) e inclui textos em português europeu e português do Brasil do século XIV ao século XX. Embora esteja disponível para consulta na Internet, estes dois corpora que referimos não podem ser descarregados. Uma das preocupações da equipa do projeto *Per-Fide* consiste em disponibilizar para consulta e *download* todo o material compilado e produzido no âmbito do projeto, tornando o corpus acessível a toda a comunidade investigadora.

A ausência generalizada da língua portuguesa em corpora paralelos multilingues é um facto notório. Contudo, desde a adesão de Portugal à União Europeia em 1986, o número de corpora paralelos associados ao domínio jurídico aumentou de forma considerável. O *EuroParl* (Koehn, 2005) e o

1 Informação adicional está disponível para consulta em: http://www.linguateca.pt/

JRC-Acquis (Steinberger et al., 2006) são corpora de referência neste domínio, que contemplam a língua portuguesa. São compilados automaticamente para todas as línguas europeias com base na legislação produzida pela UE. Um dos problemas com que os utilizadores se deparam ao efetuar pesquisas neste tipo de corpora reside no facto de os formatos dos ficheiros que estão *on-line* necessitarem de software específico e de *scripts* de alinhamento para que possam ser consultados. Acresce ainda o facto de os textos conterem demasiado ruído, pelo que o nível de qualidade do alinhamento não é elevado. O padrão de qualidade do Corpus *Per-Fide* será garantido por meio de técnicas automatizadas, desenvolvidas para assegurar a qualidade da métrica do alinhamento. É de referir, ainda, que a pesquisa de concordâncias será efetuada com base em bitextos. Com efeito, alinhar um texto na totalidade das línguas contempladas pelo projeto iria necessariamente comprometer a qualidade do alinhamento, uma vez que a tradução pode implicar extração, junção ou adição de novas frases ao texto. Ainda assim, uma tarefa deste tipo seria difícil de cumprir, já que na maior parte dos casos não foi possível ter acesso ao mesmo texto em todas as sete línguas.

No domínio literário, a *Linguateca* desenvolveu o *COMPARA* (Garcia & Santos, 2003), que é um corpus bidirecional português-inglês. No *OSLO-Multilingual Corpus*, a língua portuguesa surge apenas como língua de chegada na combinação inglês-norueguês-português. Este sub-corpus trilingue contém um total de 15 textos sob a forma de extratos com 10 000 a 15 000 palavras. É nosso objetivo reunir mais textos literários paralelos, acrescentando, assim, um maior número de línguas que possam ser alinhadas com o português.

Em última análise, porém, é possível constatar que existe ainda uma necessidade premente de criar novos corpora que contemplem a língua portuguesa bem como outras áreas do conhecimento até à data não exploradas. A tipologia dos textos jurídicos dos corpora supramencionados pode ser alargada para incluir outros tipos de texto jurídico, tais como acordos internacionais, textos de jurisprudência e outros documentos potencialmente relevantes. Alguns destes documentos podem ser obtidos automaticamente a partir do website do *EUR-Lex*. Pretendemos, com este projeto, alargar o atual número de corpora paralelos jurídicos com base nos textos que estão disponíveis de forma gratuita na Internet, criando, assim, uma base de dados jurídica de maiores dimensões e mais atualizada.

2. Processo de compilação do corpus

A primeira etapa de compilação consistiu na seleção, classificação e digitalização de textos bem como na obtenção de direitos de autor, seguindo-se a anotação morfossintática com recurso a analisadores morfológicos específicos para cada língua. Durante a fase seguinte procedeu-se ao alinhamento dos textos ao nível frásico, tendo sido calculada a métrica do alinhamento para detetar segmentos mal alinhados. Uma vez concluído o alinhamento frásico, os corpora paralelos foram processados para extrair dicionários probabilísticos de tradução (Simões & Almeida, 2003), exemplos de tradução (Simões & Almeida, 2006) e terminologia bilingue (Guinovart & Simões, 2009). Posteriormente, foram disponibilizados para *download* em formato TEI (Erjavec, 1999), XCES (Ide & Romary, 2000) e TMX (Savourel, 2005) e para consulta através da interface do projeto em www.per.fide.ilch.uminho.pt/query. É de referir que estas etapas se encontram atualmente em diferentes estados de integração no *workflow*.

Devido à quantidade e variedade de textos coligidos, o processo de compilação do corpus colocou-nos vários desafios ao nível do armazenamento e classificação da informação. Nas seguintes subsecções, apresentamos os métodos selecionados para contornar estes desafios estruturais.

2.1. *Text Encoding Initiative – armazenamento de metadados*

O consórcio da Text Encoding Initiative (TEI) elaborou um conjunto de normas orientadoras para a representação dos textos em formato digital, fornecendo uma metodologia para a codificação dos textos em versão eletrónica. Estas normas revelaram-se particularmente úteis no âmbito das Humanidades, das Ciências Sociais e da Linguística. No projeto *Per-Fide*, utilizou-se a estrutura desenvolvida pela TEI para a anotação de meta-informação.

Na sua origem, a TEI teve como desígnio a reprodução de uma série de descrições e declarações que dotassem o documento eletrónico de um título de página equivalente ao de um texto de formato em papel. As informações descritas e declaradas no TEI *header* (i.e. cabeçalho) tanto abarcam o seu aspeto bibliográfico como o não-bibliográfico, pois na sua especificidade reside o facto de este permitir a anotação dos princípios da sua própria codificação. Ou seja, o TEI *header* permite a anotação do próprio texto eletrónico através de um modelo como sendo um banco de metadados (Giordano, 1995).

Neste projeto, todos os textos integrados no corpus são anotados com meta-informação e a sua descrição é organizada de acordo com as normas da TEI. A meta-informação é declarada e descrita num cabeçalho independente, que se encontra anexado ao documento eletrónico correspondente. Desta forma, o cabeçalho independente contém toda a informação relevante que identifica o documento eletrónico a que está acoplado. É de referir, contudo, que o nosso objetivo não consistiu em aplicar a norma TEI à totalidade do corpo de texto, mesmo porque não dispúnhamos dos meios necessários para levar a cabo uma tarefa de tal dimensão. Assim, centrámo-nos na elaboração de um documento-cabeçalho, anexado ao documento eletrónico correspondente, onde é possível armazenar toda a informação descritiva e todos os elementos declarados que funcionam como uma página de título eletrónica, definindo a estrutura do documento a que se refere. Com base no cabeçalho independente, é possível criar catálogos ou índices (e.g. índice dos "livros sagrados") por género, autor, língua, etc. Além disso, o cabeçalho permite aumentar o grau de precisão do motor de pesquisa do corpus.

Com intuito de melhor satisfazer as necessidades e os objetivos do projeto *Per-Fide*, criou-se uma versão adaptada do cabeçalho originalmente proposto pela TEI. A primeira secção diz respeito à descrição do ficheiro (*file description*), que contém uma descrição bibliográfica do ficheiro eletrónico e inclui um item denominado *source description* relativo à proveniência do documento. A segunda secção refere-se à descrição do projeto (*project description*) e contempla aspetos mais específicos do desenvolvimento do projeto *Per-Fide* bem como um elemento que diz diretamente respeito à classificação textual (*class declaration*). Por último, a secção *profile description* fornece informação sobre a criação do texto bem como a(s) língua(s) usada(s). Na página seguinte apresentamos um segmento ilustrativo do cabeçalho TEI.

Como pode ver-se, o cabeçalho TEI contempla diferentes tipos de informação; não só os usuais dados bibliográficos (e.g. título, autor, editora, ano de publicação, etc.) mas também dados não bibliográficos tais como direitos de autor, línguas e até um conjunto de classificadores (no item seguinte, iremos focar esta questão, explicando de que forma o projeto *Per-Fide* procedeu em termos de classificação de documentos). Com efeito, a quantidade de informação que pode ser anotada com base na TEI é digna de referência. Contudo, o cabeçalho TEI foi desenvolvido na qualidade de proposta, ou seja, trata-se de um modelo suscetível de ser adaptado: por um lado, para que o documento possa ser considerado como estando em conformidade com a norma TEI, apenas

```
<?xml version=”1.0” encoding=”UTF-8”?>
<TEI.2>
 <teiHeader id=“321”>                <!-- Sequential ID value -->

  <fileDesc>
   <titleStmt>...</titleStmt>          <!-- Document title information -->
   <extent>3540 words</extent>         <!-- Document size -->
   <publicationStmt>...</publicationStmt> <!-- Document copyright information -->
   <notesStmt>...</notesStmt>          <!-- Miscellaneous notes -->
   <sourceDesc>...</sourceDesc>        <!-- Document origin -->
  </fileDesc>
  <profileDesc>...</profileDesc>    <!-- language, document classification, etc. -->

 </teiHeader>
</TEI.2>
```

Figura 1. Segmento do cabeçalho TEI

alguns dos elementos originais são obrigatórios; por outro lado, é também possível alargar os elementos previstos na proposta original da TEI de acordo com a natureza de cada projeto. A descrição da estrutura completa da TEI não se enquadra nos objetivos do presente artigo, contudo convidamos o leitor a consultar a versão detalhada que está disponível na Internet[2].

O cabeçalho do *Per-Fide* inclui elementos que nem sempre podem ser identificados. No caso dos textos disponibilizados na página web do Vaticano, por exemplo, podemos constatar que não é possível determinar a língua de partida de cada texto. Nesses casos, a língua foi anotada com o termo “unknown” (desconhecido). Noutros casos, houve diferentes tipos de informação em falta, como a editora (para um documento do domínio público, por exemplo). Neste tipo de casos, ao invés de recorrer à anotação “unknown”, esse elemento foi simplesmente ignorado e, consequentemente, excluído do ficheiro XML final.

Os textos para integração no corpus foram reunidos pelos membros da equipa do projeto, sendo que alguns deles não estavam familiarizados com a

2 A estrutura completa da TEI está disponível em http://www.tei-c.org/Guidelines/P4/html/index.html e a descrição do cabeçalho TEI pode ser consultada em http://www.tei-c.org/Guidelines/P4/html/HD.html.

anotação TEI. Por esse motivo, era-lhes mais difícil preencher o cabeçalho TEI em formato XML. Assim, a fim de auxiliar o processo de criação do cabeçalho, desenvolveu-se um formulário *web* onde os membros do projeto preenchiam a informação relevante para cada um dos textos que pretendiam submeter para integração no corpus, procedendo, de seguida, ao *upload* do respetivo ficheiro. De seguida, o formulário gerava automaticamente o ficheiro do cabeçalho TEI e armazenava o documento juntamente com a respetiva meta-informação numa sistema de controlo de versões. Por último, é de referir que a estrutura deste formulário é idêntica à do cabeçalho TEI, o que significa que os campos a preencher correspondem aos elementos descritos no cabeçalho.

2.2. *Ontologia – classificação de textos*

A arrumação de textos em diferentes categorias não é uma tarefa fácil e a questão da terminologia a usar para a caraterização dessas mesmas categorias pode suscitar uma problemática que não seria aqui, de todo, objetivo nosso. No entanto, a distinção das categorias e a identificação dos textos estão dependentes da observação de determinadas caraterísticas. Deste modo, extrair as caraterísticas de determinado texto implica obter uma diversidade de elementos que poderia, em certo tipo de textos, levantar dúvidas quanto à sua pertença a determinada categoria. Este tipo de texto, a que chamamos de 'natureza híbrida', pode provocar diferentes pontos de vista relativamente à sua classificação.

Decorrente deste facto, e numa tentativa de classificação de textos nos vários domínios envolvidos, encontrámos, por exemplo[3], o último livro da obra de José Saramago, *Caim*[4] (2009). Se, por um lado, é considerado um romance[5], ou seja, um texto pertencente ao domínio literário, por outro,

---

3 Outros exemplos de obras que poderiam suscitar controvérsia na sua catalogação: *The Satanic Verses* (Salman Rushdie) ou *Conversations with God* (Neale Donald Walsch)

4 Jornal 'O Público': "(...) Segundo o Antigo Testamento da Bíblia, Caim terá sido o filho primogénito de Adão e Eva, que matou Abel, seu irmão mais novo, num acesso de ciúmes, após verificar que Deus mostrara preferência por este. (...) O autor salientou que não é, de forma nenhuma, um autor de livros religiosos, que é uma matéria que só lhe interessa porque está desde sempre muito presente na mente dos homens, e na história da Humanidade (...)" (http://www.publico.pt/Cultura/)

5 Pilar del Rio (esposa de Saramago) disse: "Este último romance de José Saramago, que não é muito extenso, nem poderia sê-lo porque necessitaríamos mais fôlego que o que temos para enfrentar-nos a ele, é literatura em estado puro". (http://www.josesaramago.org/)

podemos vê-lo categorizado pela *Wikipédia*[6] como "Livros críticos de religião". Daí, surgiu a necessidade de recorrer a um instrumento que pudesse registar uma informação que sugerisse essa inter-relação de conceitos[7] nas diversas áreas, mas que, ao mesmo tempo, ditasse uma espécie de padronização da terminologia a utilizar. Assim, entre os vários instrumentos usados para representar e estruturar os diferentes domínios envolvidos, optou-se por uma metodologia que fornece princípios para agrupar conceitos de uma mesma natureza em classes mais amplas, pois na sua hierarquia os termos[8] possuem entre eles um conceito de relação e de associação. A ontologia pareceu ser a opção mais adequada não só para a normalização de termos, representação de conceitos e indexação de textos, mas também porque permite ao utilizador encontrar informação num determinado domínio no processo de busca e seleção de documentos. O *thesaurus* proporciona uma abordagem semelhante à da ontologia. Contudo, a diferença entre ambos reside no facto de que a ontologia é um sistema extensível que permite modificar e/ou expandir as relações entre os termos de acordo com as necessidades que forem surgindo. Dada a imprevisibilidade do tipo de textos que iriam surgir durante a fase de compilação do corpus, foi essencial dispor de um sistema classificativo que pudesse ser retificado e aperfeiçoado ao longo do tempo conforme o tipo de textos submetidos para integração no corpus.

A organização hierárquica de uma lista de palavras contidas numa ontologia pode ser definida de diferentes formas. O método que geralmente se utiliza para definir a hierarquia das classes consiste em relações entre termos mais genéricos – *Broader Term* (BT) – e termos mais específicos – *Narrower Term* (NT).

A falta de ontologias padronizadas para fins classificativos levou-nos a encetar um projeto de pesquisa comparativa com base em esquemas internacionais de classificação de documentos, como a CDU (Classificação Decimal Universal) e em Bibliotecas Virtuais (thesaurus) nas áreas de Biblioteconomia e Ciências da Informação, como o *Thesaurus UNESCO* (Aitchison, 1983). A elaboração desta ontologia permitiu, assim, estruturar hierarquicamente

---

6 Cf. http://pt.wikipedia.org/wiki/Categoria:Livros_cr%C3%ADticos_de_religi%C3%A3o

7 Conceito: construções mentais que servem para classificar os objetos individuais do mundo exterior ou interior através de um processo de abstracção mais ou menos arbitrário. (ISO 704 - Terminology)

8 Termo: designação de um conceito definido numa língua especializada, por uma expressão linguística. (ISO 704 - Terminology)

domínios temáticos que posteriormente pudessem ser usados para a classificação das várias obras a incorporar no nosso corpus.

Como nota final, é importante explicar o funcionamento do regime de complementaridade entre a ontologia e a TEI. Conforme o que foi mencionado anteriormente, a ontologia é um sistema classificativo composto por uma rede de relações hierárquicas entre termos, com um grau de complexidade variável. Se considerarmos que a ontologia fornece uma estrutura e nomenclatura classificativas para preencher a secção que diz respeito à classificação de documentos patente no cabeçalho TEI, podemos facilmente concluir que a ontologia serve de base ao cabeçalho. Segundo a definição prevista pela norma ISO Thesaurus 2788, a identificação de cada termo na ontologia deve ser única, para que, ao analisá-la, seja mais fácil localizar um determinado termo e, consequentemente, identificar o respetivo contexto superior de classificação. Tendo em conta a possibilidade de reorganizar a estrutura ontológica, por meio da adição ou extração de níveis hierárquicos, optámos por declarar na TEI apenas o nível mais específico, e por conseguinte o mais baixo, da ontologia. No cabeçalho TEI, este elemento é declarado como <classdecl>. De seguida, apresentamos um esquema parcial que representa a ontologia e a sua inter-relação com o cabeçalho TEI:

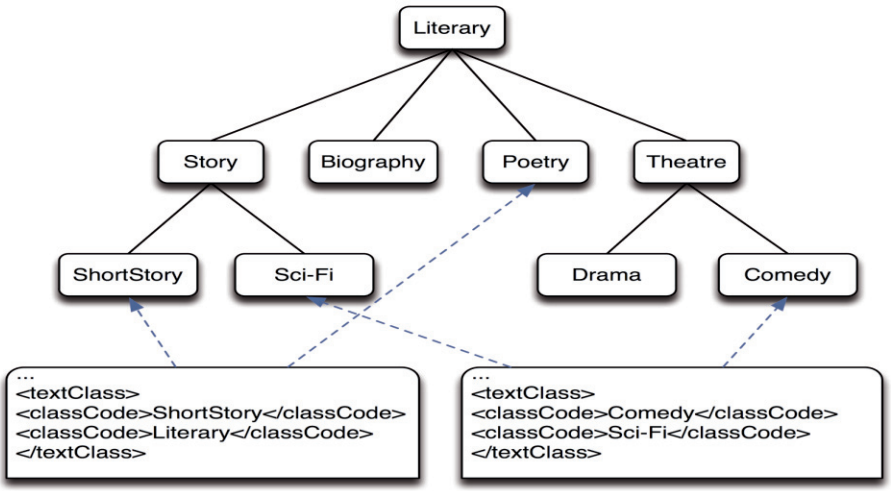

Figura 2. Ontologia e cabeçalho TEI

Ao utilizar a ontologia em colaboração com o cabeçalho TEI para fins de classificação de documentos, torna-se possível expandir ou limitar as relações conceptuais hierárquicas entre termos em qualquer etapa do processo de classificação sem que isso afete a estrutura do cabeçalho TEI.

3. Pesquisar o corpus

3.1. Pesquisa simples

A fim de realizar uma pesquisa no corpus *Per-Fide*, é necessário aceder à respetiva interface e selecionar o modo de pesquisa (monolingue/bilingue). De seguida, o utilizador deverá selecionar a(s) língua(s) e o(s) corpus da pesquisa, indicando por último o termo ou termos a pesquisar:

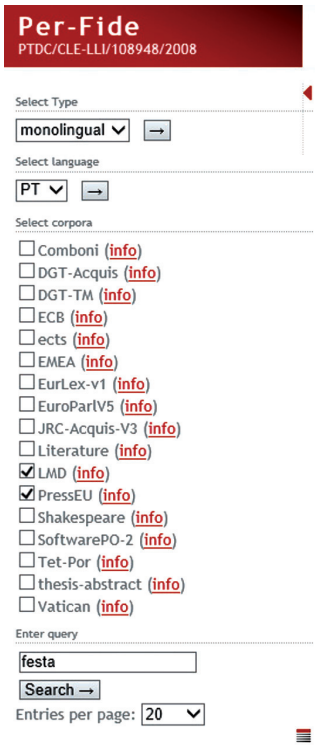

Figura 3. Pesquisa monolingue

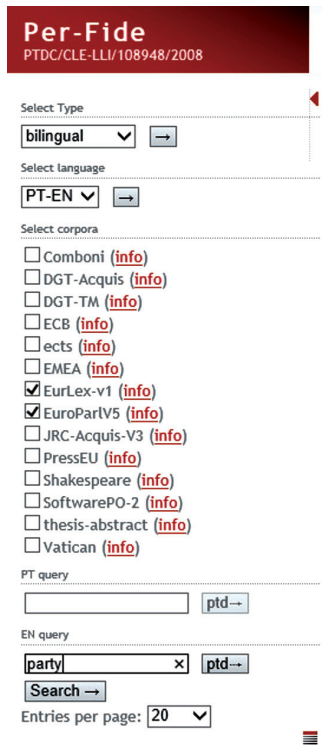

Figura 4. Pesquisa bilingue

Este tipo de pesquisa fornece ocorrências do termo pesquisado nos seus diferentes contextos – a esta lista de ocorrência atribui-se o nome de concordância. A concordância poderá ser monolingue:

Figura 5. Pesquisa monolingue do termo *festa* (resultado da pesquisa ilustrada na Figura 3)

ou bilingue:

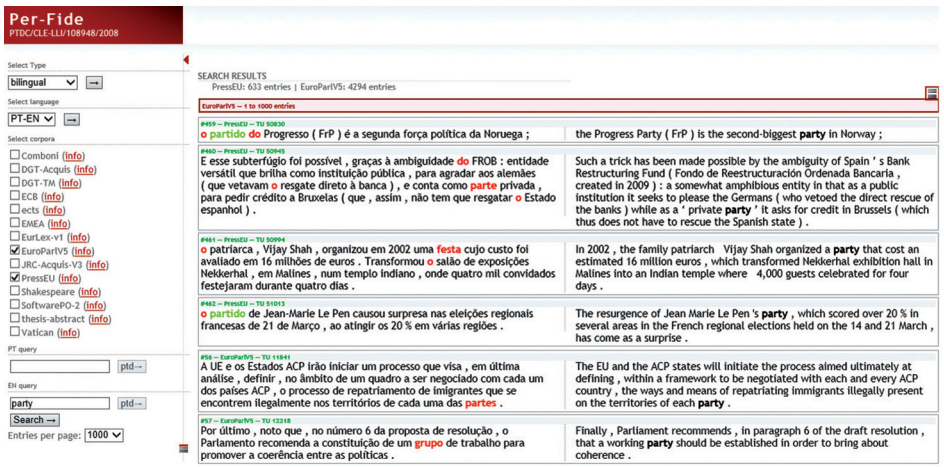

Figura 6 – Pesquisa bilingue de *party* (resultado da pesquisa ilustrada na Figura 4)

Como se pode ver acima, a palavra *party* surge traduzida de várias formas, o que se deve à variedade dos géneros textuais que compõem os corpora do *Per-Fide* e sobre os quais incide cada pesquisa. Daqui se conclui que não seria possível resgatar a polissemia dos termos sem o ecletismo textual que o Corpus *Per-Fide* apresenta. Além de dar conta da dimensão polissémica do léxico, este tipo de pesquisa também permite encontrar redes de sinónimos de um termo:

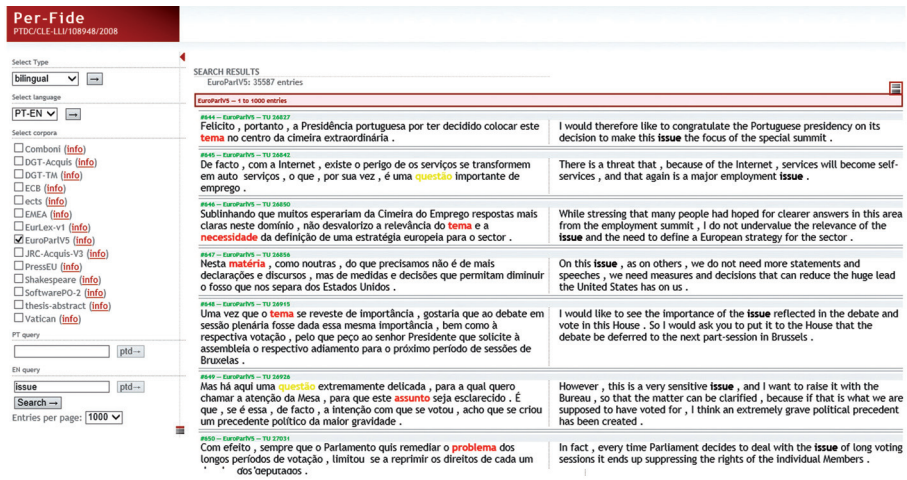

Figura 7. Rede sinonímica de *issue* através da pesquisa bilingue

3.2. *Pesquisa simultânea*

No modo bilingue, existe ainda a possibilidade de efetuar uma pesquisa nas duas línguas em simultâneo, o que poderá ser útil para confirmar a relação de equivalência entre dois termos:

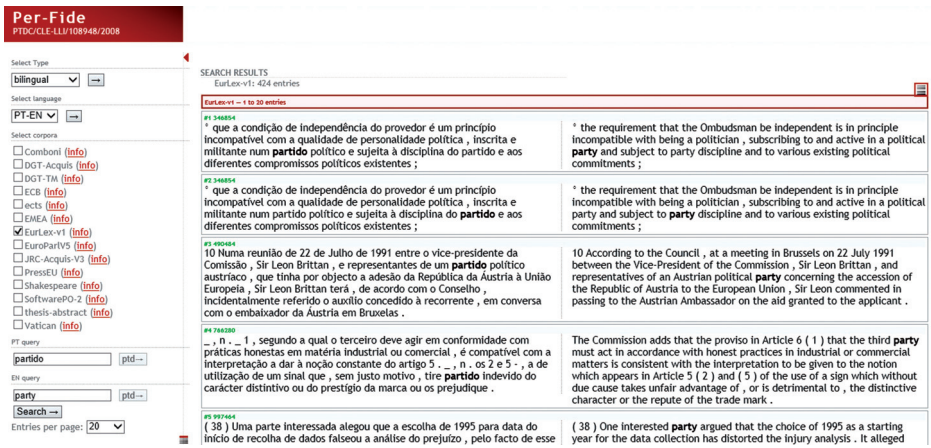

Figura 8. Pesquisa simultânea de *party* e *partido*

3.3. *Pesquisa multipalavra*

Quer no modo monolingue quer no modo bilingue, o utilizador poderá realizar pesquisas por termos compostos de uma ou mais palavras. Este tipo de pesquisa pode revelar-se de grande utilidade no caso das colocações, e.g. *third party*:

Figura 9. Pesquisa da colocação *third party*

No âmbito desta pesquisa sobre colocações, importa referir que o contexto mais imediato de uma palavra pode influenciar a sua tradução. No caso concreto de *party*, para além da sua associação ao numeral *third*, notámos também que a sua junção a adjetivos como *concerned/interested* e *contracting* pode resultar em traduções como *interessado* e *contratante*, respetivamente.

3.4 *Dicionários probabilísticos de tradução*

A simples pesquisa em modo bilingue pode ser complementada com o acesso a dicionários probabilísticos de tradução (PTD), que indicam o número de ocorrências de um palavra no(s) corpus selecionado(s) e oferecem sugestões de tradução, apresentando, estatisticamente, uma medida de confiança dessas traduções.

Figura 10. PTD de *party*

O PTD tem um caráter cíclico uma vez que, além de fornecer traduções para o termo pesquisado, também traduz as próprias sugestões de tradução,

voltando à língua em que se pesquisou inicialmente. O utilizador pode aceder às possíveis traduções de um qualquer termo constante do PTD, clicando nas quatro setas centrífugas à esquerda. As setas da direita, por sua vez, dão acesso às concordâncias das alternativas propostas pelo PTD.

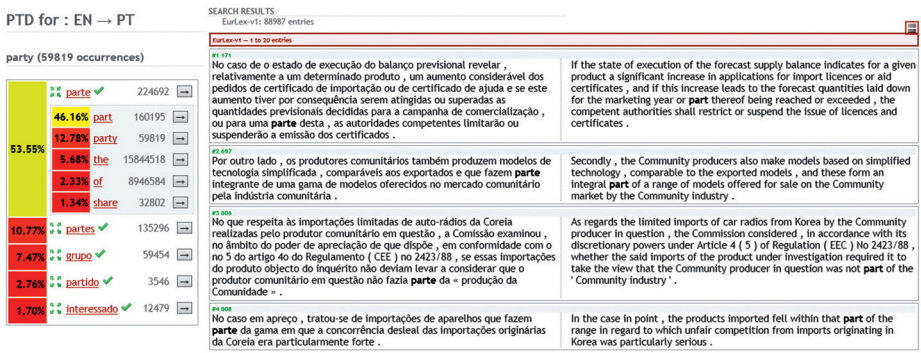

Figura 11. esquerda: PTD de *part* a partir das setas centrífugas; direita: lista de concordâncias de *part* a partir da seta que se encontra à direita do PTD.

Por motivos de impressão, o leitor não poderá aperceber-se através destas imagens da gradação cromática do PTD. De facto, este é percorrido por uma escala de cores, que representa em percentagem o grau de probabilidade de um termo ser traduzido por outro num determinado corpus. Nesta escala, a cor verde representa o grau máximo de equivalência entre os dois termos e a cor vermelha níveis inferiores de equivalência. Entre estas dois polos existem cores intermédias (e.g. laranja e amarelo).

4. Conclusão

Existe um interesse cada vez maior em corpora e nos recursos (e.g. PTD e terminologia bilingue) construídos a partir destes repositórios de dados organizados. Como pudemos ver, este tipo de ferramenta encerra um grande potencial no âmbito dos estudos linguísticos, e em particular da tradução (Berber Sardinha, 2003), uma vez que permite sistematizar e explicar, com base em dados empíricos, fenómenos inerentes ao funcionamento da língua que escapam a

um mero exercício de introspeção. A linguística de corpus tem mostrado quão inexata é a intuição humana no entendimento da linguagem (Sampson, 2001). Retomando o tema desta XIV edição do Colóquio de Outono, é legítimo afirmar que as Humanidades se deparam atualmente com inúmeros desafios tecnológicos que exigem de toda a comunidade científica a necessária abertura a novos paradigmas do conhecimento e da investigação, em que a transdisciplinaridade parece assumir um papel preponderante. Com este estudo, esperamos ter demonstrado de forma cabal os benefícios que resultam da sinergia entre as Humanidades e a Engenharia Informática.

Agradecimentos:

O *Per-Fide*, *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)*, é parcialmente financiado pelo projeto PTDC/CLE-LLI/108948/2008 da Fundação para a Ciência e Tecnologia.

Referências

Araújo, Sílvia, Alberto Simões, José João Almeida e Idalete Dias (2010), "Apresentação do projeto *Per-Fide*: Paralelizando o Português com seis outras línguas", *Linguamática*, vol. 2, nº 2, pp. 71-74.

Aitchison, Jean (1983), *Thesaurus de l'Unesco: liste structuré de descripteurs pour l'indexation et la recherche bibliographiques dans les domaines de l'éducation, de la science, des sciences sociales, de la culture et de la communication*, Paris : Unesco.

Berber Sardinha, Tony P. (2003), "Uso de corpora na formação de tradutores", DELTA, vol. 19, nº especial, pp.43-70.

Erjavec, Tomaz (1999), "A TEI encoding of aligned corpora as translation memories", in *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, pp. 49–60.

Garcia, Ana F. e Diana Santos (2003), "Introducing COMPARA, the Portuguese-English parallel translation corpus", in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translation Education*, Manchester: St. Jerome Publishing, pp. 71-87.

Giordano, Richard (1995), "The TEI header and the documentation of electronic texts", *Computers and the Humanities*, vol. 29, nº 1, pp. 75-84.

Guinovart, Xavier G. e Alberto Simões (2009), "Parallel corpus-based bilingual terminology extraction", in *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*.

Ide, Nancy, Patrice Bonhomme e Laurent Romary (2000), "XCES: an XML-based encoding standard for linguistic corpora", in *Proceedings of the Second International Language Resources and Evaluation Conference*.

Koehn, Philipp (2005), "EuroParl: A Parallel Corpus for Statistical Machine Translation", in *Proceedings of MT-Summit*, pp. 79-86.

Rocha, Paulo A. e Diana Santos, "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada,* pp. 131-140.

Sampson, G. (2001), *Empirical Linguistics*, Londres e Nove Iorque: Continuum.

Saramago, José (2009), *Caim*, Lisboa: Editorial Caminho.

Savourel, Yves (2005), "TMX 1.4b Specification", Relatório Técnico, disponível em http://www.gala-global.org/oscarStandards/tmx/, consultado em 20/03/2011.

Simões, Alberto e José João Almeida (2007), "Parallel Corpora based Translation Resources Extraction", in *Procesamiento del Lenguaje Natural*, vol. 39, pp. 265-272.

Simões, Alberto e José João Almeida (2003), "NATools – A Statistical Word Aligner Workbench", in *Procesamiento del Lenguaje Natural*, vol. 31, 217-224.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec Dan Tufiş e Dániel Varga (2006), "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", in *Proceedings of the 5th International Conference on Language Resources and Evaluation*.