Apresentação do projecto Per-Fide: Paralelizando o Português com seis outras línguas

Sílvia Araújo[1], José João Almeida[2], Idalete Dias[1], Alberto Simões[3]
[1]Instituto de Letras e Ciências Humanas, Universidade do Minho
[2]Departamento de Informática, Universidade do Minho
[3]Escola Superior de Estudos Industriais e de Gestão, Instituto Politécnico do Porto
{saraujo@ilch,jj@di,idalete@ilch}.uminho.pt, alberto.simoes@eu.ipp.pt

Resumo

Neste documento apresentamos o projecto Per-Fide que tem como principal objectivo a criação de recursos bilingues entre a língua portuguesa e seis outras línguas: espanhol, russo, francês, italiano, alemão e inglês. Este processo iniciar-se-á com a compilação de corpora paralelos em diferentes áreas, nomeadamente a literatura, religião e política (legislativa e jurídica) e técnico-científica. Os corpora serão alinhados à frase e à palavra e serão objecto de extracção automática de dicionários e terminologia bilingues. O documento irá focar inicialmente os principais objectivos do projecto, seguindo-se uma visão geral sobre as tarefas propostas e os resultados que se esperam obter.

1 Introdução

Nos últimos anos a quantidade de recursos paralelos para o processamento de linguagem natural tem vindo a aumentar. Infelizmente este aumento tem sido significativo especialmente para outras línguas que não a portuguesa e, o aumento para esta língua, tem-se verificado essencialmente na área política ou legislativa (graças à adesão de Portugal à União Europeia).

O projecto Per-Fide[1] tem como principal objectivo a compilação de corpora paralelos entre a língua portuguesa e seis outras línguas: espanhol, russo, francês, italiano, alemão e inglês.

A escolha das línguas foi dirigida pela relevância global das línguas em causa, mas também da relevância das línguas dentro do Centro de Estudos Humanísticos da Universidade do Minho (CEH/UM), onde o projecto está sediado.

A construção destes corpora será realizada em diferentes áreas do conhecimento que podem ser divididas em dois grandes blocos:

- **ficção:** neste primeiro bloco inclue-se essencialmente a produção literária (portuguesa e estrangeira) e textos religiosos[2]
- **não ficção:** este segundo bloco focará textos jornalísticos, político ou legislativo, e técnico-científico.

Tentar-se-á, também, a recolha de textos tendo como língua de origem cada uma das sete línguas, incluindo as diferentes variantes da língua portuguesa: português europeu, português brasileiro e português africano (nos seus diferentes dialectos).

O Per-Fide não culminará com a construção dos corpora. Esse será apenas o primeiro passo. O Per-Fide tem um conjunto de objectivos mais vasto. Destes, o mais importante, e ortogonal a tudo o resto, é a disponibilização aberta de todos os recursos recolhidos e produzidos durante o projecto.

2 Per-Fide: linha de produção

Esta secção detalha as principais etapas previstas no projecto, relacionando-as com recursos produzidos, projectos semelhantes já existentes e com trabalho desenvolvido previamente pela equipa Per-Fide. Estas etapas estão esquematizadas na figura 1 que destaca o lado aberto do projecto com a disponibilização de todos os recursos produzidos.

2.1 Limpeza, Tratamento e Anotação

No processo previsto pelo Per-Fide, a primeira etapa corresponde à compilação dos recursos (nomeadamente de textos). Esta tarefa vai ser especialmente custosa devido à necessidade de negociação de direitos de autor, o que nem sempre será possível. Assim que negociado o modo

[1]O nome do projecto tem a sua génese nos nomes das sete língua envolvidas: Português, English, Russian, Français, Italiano, Deutch, Español.

[2]A classificação dos textos religiosos não é consensual. A nossa escolha não quer de forma alguma levantar essa discussão já que é irrelevante para o contexto em causa.

Figura 1: Linha de Produção Per-Fide

de disponibilização dos documentos, seguir-se-á a sua limpeza e tratamento. Estas duas tarefas incluem a conversão entre formatos, a remoção de ruído obtido durante as conversões efectuadas, a detecção de documentos duplicados e a avaliação da qualidade dos recursos em causa.

Além disso, os documentos serão classificados e anotados com meta-informação como sejam a língua (e respectiva variante), o género do documento, o título e autor (se aplicável), informação sobre a tradução (caso não esteja na sua língua original[3]), etc. Toda esta informação será armazenada em formato textual usando as estruturas já disponibilizadas pelo TEI — Text Encoding Initiative (Vanhoutte, 2004) — para esta finalidade.

Embora esta etapa não seja completamente automatizável, existe um conjunto de ferramentas já existentes que serão imprescindíveis:

- diversos conversores de formato, como o `antiword` ou o `pdftotext`, que permitem uma conversão de qualidade razoável entre alguns formatos proprietários e documentos de texto;
- identificadores de língua, como seja o módulo Perl `Lingua::Identify`;
- identificadores de variante da língua (ou da grafia), e comparadores de grafia (Almeida, Santos e Simões, 2010);

Nesta etapa serão disponibilizados os corpora monolingues para pesquisa na rede ou, sempre que as licenças o permitam, para serem descarregados e processados localmente.

2.2 Etiquetação Morfosintáctica

No Per-Fide não temos como objectivo desenvolver um sistema de etiquetação morfosintáctica, especialmente dado que o teríamos de fazer para as sete línguas envolvidas. Embora neste momento ainda não se tenha procedido à aquisição de nenhum etiquetador morfo-sintáctico está previsto o uso do Palavras (Bick, 2000). Outras hipóteses já se encontram delineadas como seja o uso do TnT (Brants, 2000) ou do TreeTagger (Schmid, 1994). Uma outra alternativa será a etiquetação ambígua usando um simples analisador morfológico como o jSpell (Almeida e Pinto, 1994) ou o FreeLing (Atserias et al., 2006).

Existem outros projectos que disponibilizam corpora etiquetados, como por exemplo a Floresta Sintá(c)tica para a língua portuguesa (Afonso et al., 2001), ou o Penn Treebank para a língua inglesa (Marcus, Marcinkiewicz e Santorini, 1993). No caso concreto do Per-Fide o objectivo não é tanto a disponibilização dos corpora monolingues etiquetados, mas a sua disponibilização como parte integrante de um corpus paralelo.

Os corpora etiquetados serão disponibilizados nos formatos convencionais, e também disponibilizados para pesquisa na rede.

2.3 Alinhamento ao Nível da Frase

Existem várias alternativas de ferramentas para o alinhamento ao nível da frase, desde as mais simples como o Vanilla Aligner (Gale e Church, 1991) aos mais usados recentemente como o easy-align, parte do IMS Corpus Workbench (Christ et al., 1999), o HunAlign (Varga et al., 2005) ou o Clue Aligner do PLUG (Tiedemann, 2003).

Em relação à disponibilização de corpora paralelos alinhados para consulta na rede ou para processamento local existem já vários projectos, dos quais salientamos o COMPARA, um corpus paralelo português-inglês literário (Frankenberg-Garcia e Santos, 2003), e o projecto OPUS que disponibiliza vários corpora paralelos de diferentes áreas para todas as línguas europeias (Tiedemann e Nygard, 2004).

Os corpora alinhados serão disponibilizados

[3] Alguns géneros linguísticos, como jornalístico ou religioso, não permitirão uma fácil classificação de língua de origem. Neste caso nenhuma das línguas levará essa classificação.

em TEI[4] (Erjavec, 1999), XCES (Ide, Bonhomme e Romary, 2000) ou TMX (Savourel, 2005)[5].

2.4 Extracção de Dicionários de Tradução

Existem duas abordagens diferentes no alinhamento de corpora ao nível da palavra. Uma, que tem como ferramenta preferencial para a sua extracção o Giza++ (Och e Ney, 2003), pretende associar cada instância de uma palavra com a instância correspondente na sua tradução. A outra abordagem, que pode ser obtida usando o NATools (Simões e Almeida, 2003), pretende extrair um dicionário probabilístico de tradução que associa a cada palavra tipo as suas possíveis traduções, devidamente pesadas probabilisticamente.

No Per-Fide serão aplicadas as duas abordagens, sendo que a extracção de dicionários probabilísticos de tradução será imprescindível para a abordagem planeada na extracção de terminologia bilingue. Os dicionários serão também úteis para auxiliar a pesquisa bilingue de concordâncias (Simões, 2008).

2.5 Extracção de Terminologia Bilingue

Para além da extracção de dicionários de tradução, o Per-Fide também tem como objectivo a compilação de terminologia bilingue que, sendo originária de corpora devidamente etiquetados, poderá ser facilmente classificada por área de conhecimento. Para esta extracção tenciona-se usar uma abordagem baseada em padrões bilingues (Simões e Almeida, 2008; Guinovart e Simões, 2009). Estes padrões especificam regras morfológicas dos contituintes de cada termo, permitindo a sua extracção automática. Ao usar-se um corpus etiquetado morfo-sintacticamente e não um analisador morfológico levará a um aumento da precisão de extracção.

2.6 Disponibilização na Rede

Como foi sendo referido nos itens anteriores, o projecto Per-Fide tem como ponto fulcral a disponibilização dos recursos construídos. Esta disponibilização será feita à medida que os recursos sejam desenvolvidos e não apenas no final do projecto. Deste modo espera-se que o retorno efectuado pelos utilizadores permita melhorar o material disponibilizado.

Além disso, tentar-se-á que os recursos sejam disponibilizados de forma integrada, de modo que o mesmo interface permita a consulta nos corpora (paralelos ou bilingues, com ou sem anotação), nos dicionários de tradução e na terminologia bilingue. Esta integração permitirá também que os recursos extraídos possam ser utilizados para enriquecer as consultas de outros recursos. Por exemplo, a apresentação de concordâncias bilingues poderá beneficiar dos dicionários probabilísticos de tradução para alinhar (ou sublinhar) o resultado da pesquisa em ambas as línguas.

3 Conclusões

O projecto está agora a iniciar e os seus membros têm a perfeita noção de que os objectivos são ousados. Um dos principais problemas será a questão de direitos de autor, essencialmente no que respeita à obtenção de textos literários. Neste sentido está a ser desenvolvido um manual de negociação que permitirá, sucessivamente, aumentar a probabilidade de sucesso.

Um outro problema terá que ver com a língua Russa, e o facto de esta usar um sistema de codificação diferente. Embora tecnologicamente se possa falar do Unicode como solução global, a verdade é que a maioria das ferramentas que prevemos utilizar nunca foi testada com outro tipo de codificação que não seja o ISO-8859-1.

Neste momento o projecto pode ser acompanhado em `http://natura.di.uminho.pt/per-fide`.

Agradecimentos

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

Referências

Afonso, Susana, Eckhard Bick, Renato Haber, e Diana Santos. 2001. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editores, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, pp. 533–545, Lisboa, Portugal, 2-4 de Outubro, 2001. APL.

Almeida, José João e Ulisses Pinto. 1994. Jspell – um módulo para análise léxica genérica de linguagem natural. Em *Actas do X Encontro da Associação Portuguesa de Linguística*, pp. 1–15, Évora.

[4]Embora o TEI não seja desenhado para a codificação de corpora paralelos existem alguns projectos que o usam para indicar alinhamentos.

[5]Neste ponto ainda não se decidiu por qual (ou quais) optar. Sendo possível proceder à automatização na conversão de formatos, serão todos disponibilizados.

Almeida, José João, André Santos, e Alberto Simões. 2010. Bigorna – a toolkit for orthography migration challenges. Em *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, May, 2010. forthcomming.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, e Muntsa Padró. 2006. FreeLing 1.3: syntactic and semantic services in an open-source NLP library. Em *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 48–55.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, Aarhus University.

Brants, Thorsten. 2000. TnT – a statistical part-of-speech tagger. Em *6th Applied NLP Conference, ANLP-2000*, Seattle, WA, April 29, 2000.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann, e Esther König, 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stuttgart, March, 1999. `http://www.ims.uni\_stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`.

Erjavec, Tomaz. 1999. A tei encoding of aligned corpora as translation memories. Em *In Proceedings of the EACL Workshop on Linguistically Interpreted Corpora '99*, pp. 49–60.

Frankenberg-Garcia, Ana e Diana Santos. 2003. Introducing COMPARA, the portuguese-english parallel translation corpus. Em Silvia Bernardini Federico Zanettin e Dominic Stewart, editores, *Corpora in Translation Education*. Manchester: St. Jerome Publishing, pp. 71–87.

Gale, William A. e Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. Em *Meeting of the Association for Computational Linguistics*, pp. 177–184.

Guinovart, Xavier Gomez e Alberto Simões. 2009. Parallel corpus-based bilingual terminology extraction. Em *8th International Conference on Terminology and Artificial Intelligence*, Toulouse, France, November, 18–20, 2009.

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

Marcus, Mitchell P., Mary Ann Marcinkiewicz, e Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Och, Franz Josef e Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Savourel, Yves, editor. 2005. *TMX 1.4b Specification*. Localisation Industry Standards Association (LISA), April, 2005. `http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm`.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. Em *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.

Simões, Alberto e José João Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, 41:281–288, September, 2008.

Simões, Alberto M. e J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September, 2003.

Simões, Alberto Manuel Brandão. 2008. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento, Escola de Engenharia, Universidade do Minho, Braga, 19 May, 2008.

Tiedemann, Jörg e Lars Nygard. 2004. The OPUS corpus - parallel and free. Em *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, Portugal, May, 2004.

Tiedemann, Jörg. 2003. Combining clues for word alignment. Em *10th Conference of the European Chapter of the ACL (EACL03)*, Budapest, Hungary, April 12–17, 2003.

Vanhoutte, Edward. 2004. An introduction to the tei and the tei consortium. *Lit Linguist Computing*, 19(1):9–16, April, 2004.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, e V. Nagy. 2005. Parallel corpora for medium density languages. Em *Recent Advances in Natural Language Processing (RANLP 2005)*, pp. 590–596.