CORPUS PARALELO E CORPUS PARALELO ALINHADO: PROPRIEDADES E APLICAÇÕES
(PARALLEL CORPUS AND ALIGNED PARALLEL CORPUS: FEATURES AND APPLICATIONS)

Helena de Medeiros CASELI (PG – Universidade de São Paulo – São Carlos)
Maria das Graças Volpe NUNES (Universidade de São Paulo – São Carlos)

ABSTRACT: *Parallel texts – texts in one language and their translation in other – and aligned parallel texts – with identification of translation correspondences – are very important in many applications such as machine translation. In this paper we describe four Brazilian Portuguese and English parallel corpora, their sentence aligned version and some applications.*

KEYWORDS: *parallel texts; aligned parallel texts; sentence alignment of parallel texts; machine translation.*

0. Introdução

Textos paralelos, segundo a terminologia estabelecida pela comunidade de lingüística computacional, são textos acompanhados de sua tradução em uma ou várias línguas. São considerados distintos dos textos sobre um mesmo tópico, escritos em línguas diferentes, mas que não são, necessariamente, traduções mútuas: os textos comparáveis.

Os textos paralelos – também chamados de bitextos quando apenas duas línguas estão envolvidas – são fontes ricas de conhecimento lingüístico, isso porque a tradução de um texto para uma outra língua pode ser entendida como uma anotação detalhada do significado do texto original. Se, além de paralelos, os textos forem alinhados, isto é, possuírem marcas que identifiquem os pontos de correspondência entre o texto original (texto fonte) e sua tradução (texto alvo), o conhecimento daí derivado (as equivalências de tradução) assume importância capital em muitas aplicações, como as traduções humana e automática e a recuperação de informação entre línguas diferentes.

Tais pontos de correspondência podem ser os textos completos ou suas partes constituintes: capítulos, seções, parágrafos, sentenças, palavras e até mesmo caracteres. Os alinhamentos mais estudados na atualidade são o de sentenças e o de palavras, também chamados de alinhamento sentencial e lexical, respectivamente. Este artigo faz menção apenas ao alinhamento sentencial de textos paralelos.

O processo automático de alinhamento sentencial, resumidamente, pode ser entendido como a "busca", no texto alvo (em inglês, nesse caso), de uma ou mais sentenças que correspondam à tradução de uma dada sentença no texto fonte (em PB). Na Figura 1 têm-se um exemplo de um bitexto alinhado sentencialmente na qual o texto fonte (em PB) é apresentado à esquerda, enquanto o texto alvo (em Inglês), à direita.

Nesta figura, as sentenças estão separadas de acordo com um alinhamento sentencial estabelecido entre elas. Assim, a primeira sentença do texto fonte corresponde à primeira sentença do texto alvo e a segunda sentença do texto fonte às duas últimas sentenças do texto alvo.

Texto Fonte (em PB)	**Texto Alvo (em inglês)**
Este trabalho apresenta requisitos funcionais identificados no processo de Engenharia Reversa de Software que possam ser suportados por um Sistema Hipertexto.	This paper discusses the functional requirements identified in the software reverse engineering process which can be supported by a hypertext system.
Por meio da modelagem conceitual e navegacional do domínio de informações relativas ao método de engenharia reversa Fusion-RE/I, foram estabelecidos os requisitos funcionais de um aplicativo hipermídia de suporte ao método, de forma a nortear o engenheiro de software responsável pelo processo de engenharia reversa e possibilitar o acompanhamento da evolução desse processo.	By means of a conceptual and navigational modeling of information related to the reverse engineering method Fusion-RE/I, we established the functional requirements of a hypermedia application to support the method.
	Our purpose is to offer guidelines to the software engineer in charge of the reverse engineering process and to make possible to follow the evolution of this process.

Figura 1 – Par de textos paralelos alinhados sentencialmente.

Nesse contexto, este artigo descreve quatro corpora paralelos PB-inglês construídos como base para a extração de dados destinados a diferentes tipos de investigação em Processamento de Língua Natural (PLN). Inicialmente, os corpora foram construídos para servir de base para o estudo de técnicas e metodologias de alinhamento sentencial de textos paralelos [1].

Além da motivação decorrente da grande utilidade dos textos paralelos e, principalmente, dos textos paralelos alinhados, a iniciativa de construção desses corpora teve ainda uma motivação especial em relação às línguas envolvidas. Embora diversos trabalhos na área de alinhamento de textos paralelos utilizem corpora de textos escritos em português europeu (Santos e Oksefjell, 2000; Ribeiro et al., 2000) não havia, no momento da construção desses corpora, o conhecimento de publicações envolvendo o português brasileiro.

A próxima seção (Seção 1) traz uma descrição dos corpora paralelos; a Seção 2 cita algumas aplicações que se utilizam dos corpora paralelos e paralelos alinhados, e a última Seção (3), traz algumas conclusões deste artigo.

1. Os corpora paralelos

Ao construir um corpus de textos, procura-se fazer uma seleção de dados representativa, isto é, que constitua um corpo de evidências lingüísticas que possa suportar generalizações e contra as quais se possam testar hipóteses.

[1] Pesquisa desenvolvida no NILC (Núcleo Interinstitucional de Lingüística Computacional), no âmbito do projeto PESA (*Portuguese-English Sentence Alignment*). Página do projeto: http://www.nilc.icmc.usp.br/nilc/projects/pesa.htm.

Além disso, a restrição quanto ao domínio do conhecimento de onde proviria cada um dos corpora, foi resultado da opção pela seleção de textos de um domínio específico. De acordo com a literatura consultada[2], essa restrição de domínio é um critério a ser adotado quando a construção do corpus representa um estágio do desenvolvimento de um produto ou da busca de um objetivo de pesquisa, precisamente o caso do CorpusPE.

Assim, três corpora de textos paralelos PB-inglês, de gêneros diferentes em domínios específicos, foram construídos: o CorpusPE (científico), o CorpusALCA (jurídico) e o CorpusNYT (jornalístico).

O CorpusPE é composto por 65 pares de resumos e *abstracts* de trabalhos acadêmicos na área de computação desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) em São Carlos[3], apresentados na forma de artigos publicados em revistas especializadas, monografias de qualificação de mestrado, dissertações de mestrado e teses de doutorado e pertencentes a subdomínios variados da Computação[4].

Esse conjunto inicial, na verdade, foi dividido em dois: o corpus autêntico (CAT) e o corpus pré-editado (CPT). O primeiro é formado pelos 65 bitextos na forma em que foram originalmente redigidos, sem nenhuma alteração em sua forma ou em seu conteúdo. O pré-editado, por sua vez, também é formado pelos mesmos 65 pares de textos, porém com correções, alterações e marcações feitas por um tradutor humano para a eliminação de ambigüidades, equívocos e erros de gramática e/ou tradução para o inglês[5].

A delimitação desse corpus baseou-se no fato de se saber que, quando um corpus está relativamente isento de ruídos de comunicação e provém de domínios técnicos, nos quais as traduções literais são esperadas, o alinhamento automático de sentenças torna-se mais eficaz (Gaussier et al., 2000). Dessa forma, a subdivisão do corpus em autêntico e pré-editado permitiu uma análise dos resultados apresentados pelos métodos de alinhamento sentencial de textos paralelos quando aplicados a corpus com e sem ruídos, enriquecendo ainda mais a pesquisa nessa área.

O CAT é constituído por 855 sentenças e 21.432 palavras enquanto o CPT contém 849 sentenças e 21.492 palavras. Sendo que, em ambos os corpora, há uma média de 7 sentenças por texto e e25 palavras por sentença.

O CorpusALCA, por sua vez, é composto por textos de domínio jurídico. São 4 textos paralelos extraídos da documentação oficial da ALCA (Área de Livre Comércio das Américas)[6] num total de 725 sentenças e 22069 palavras. Cada texto tem, em média, 91 sentenças e cada sentença, 30 palavras.

---

[2] Vide (Renouf, 1987) e (Sinclair, 1991).

[3] A identificação e aquisição dos textos foram feitas por Feltrim (Feltrim et al., 2001).

[4] Banco de dados, computação de alto desempenho, computação gráfica e processamento de imagens, engenharia de software, hipermídia, inteligência computacional, matemática computacional, sistemas digitais, sistemas distribuídos e programação concorrente.

[5] Detalhes sobre o processo de coleta e pré-edição dos textos dos corpora autêntico e pré-editado podem ser obtidos em (Martins et al., 2001).

[6] Disponíveis no site oficial da ALCA (http://www.ftaa-alca.org/alca_e.asp).

Finalmente, o último corpus construído para o projeto PESA foi o CorpusNYT. Esse corpus é composto por 7 pares de textos paralelos que são artigos do jornal "The New York Times"[7]. O CorpusNYT possui 422 sentenças e 10595 palavras, sendo que cada texto tem, em média, 30 sentenças e cada sentença, em média, 25 palavras.

A Tabela 1 detalha o número de palavras em cada corpus citado e em cada um dos idiomas envolvidos (PB e inglês).

Tabela 1. Número de palavras nos corpora em cada um dos idiomas envolvidos (PB e inglês).

Palavras	CAT	CPT	CorpusALCA	CorpusNYT
PB	11349	11306	11217	5410
Inglês	10083	10186	10852	5185
Total	**21432**	**21492**	**22069**	**10595**

Esses quatro corpora paralelos (CAT, CPT, CorpusALCA e CorpusNYT) foram alinhados sentencialmente por meio de um processo semi-automático para servirem de referência na comparação dos alinhamentos produzidos automaticamente pelos métodos de alinhamento sentencial avaliados no projeto PESA. Assim, além dos quatro corpora paralelos outros quatro alinhados sentencialmente foram obtidos.

Cinco métodos de alinhamento sentencial de textos paralelos baseados em diferentes critérios de alinhamento - palavras cognatas, listas de palavras âncoras[8], tamanho das sentenças, entre outros - foram avaliados no projeto PESA. Constatou-se que a maioria desses métodos, quando avaliados com textos PB-inglês, manteve uma precisão de acordo com a relatada na literatura para outras línguas: acima de 95%. Apenas nos corpora CAT e CorpusNYT a precisão foi menor; o primeiro por ser um corpus com ruídos, e o segundo, por possuir traduções mais livres, já que se trata de um corpus formado por textos jornalísticos. Além disso, confirmou-se o que já havia sido relatado em (Gaussier et al., 2000): o desempenho de todos os métodos foi melhor em corpora sem ruídos (CPT) do que em corpora com ruídos (CAT).

Tanto os corpora paralelos quanto os paralelos sentencialmente alinhados são muito importantes para diversas aplicações, como mostra a próxima Seção (2).

2. Aplicações

Como mencionado anteriormente, os textos paralelos são fontes ricas de conhecimento lingüístico e esta Seção cita algumas das aplicações que se beneficiam dos textos paralelos e dos textos paralelos alinhados como: tradução automática; a recuperação de informações por meio da troca de dados entre línguas diferentes; a construção de léxicos bilíngües; a extração de terminologia de textos técnicos; o

[7] Disponíveis na *web* em inglês (http://www.nytimes.com) e em PB (http://ultimosegundo.ig.com.br/useg/nytimes).

[8] Listas de palavras (ou expressões multi-palavras) na língua fonte (PB, por exemplo) e suas traduções na língua alvo (inglês, por exemplo).

esclarecimento de ambigüidade; e o aprendizado de idiomas. Cada uma dessas aplicações será brevemente introduzida a seguir.

Atualmente, uma das técnicas de tradução automática mais estudada é a Tradução Automática Baseada em Exemplos (ou Example Based Machine Translation, EBMT) na qual uma base de exemplos bilíngües é utilizada como recurso de tradução. Esta técnica é utilizada nas memórias de tradução (ou Memory-based Machine Translation, MBMT): ferramentas computacionais que tentam evitar a tradução desnecessária de segmentos de texto previamente traduzidos através da consulta e recuperação automática das informações contidas nesses segmentos em um banco de dados que, muitas vezes, são construídos com base em textos paralelos alinhados. Alguns exemplos de memórias de tradução são: Trados[9], Transit, Déja Vu[10], XL8, Eurolang, Catalyst, SDLX e Word Fast[11], desenvolvidas por diversas empresas de tradução americanas e européias.

Além das memórias de tradução, a tradução automática se beneficia dos corpora paralelos no que diz respeito à aquisição automática de conhecimento, por exemplo, dicionários e regras de tradução. Neste último caso, regras são induzidas a partir de corpora paralelos alinhados sentencialmente tomando-se como base, por exemplo, a estrutura sintática das sentenças fonte e alvo.

A recuperação de informações através da troca de dados entre línguas diferentes é uma aplicação que ganhou muita importância com a disseminação do uso da Internet e a crescente quantidade de documentos, escritos em várias línguas, disponíveis eletronicamente. Esta aplicação está presente, principalmente, em buscas na *web* quando a consulta é feita em uma determinada língua e o resultado é apresentado em outra ou diversas outras.

A construção de léxicos bilíngües, bem como a extração de terminologia de textos técnicos, são aplicações bastante favorecidas pelo alinhamento de textos paralelos. Ambas são consideradas tarefas difíceis e que demandam muito tempo para serem concluídas, mas podem se tornar menos onerosas com o auxílio de corpora paralelos e, principalmente, corpora paralelos alinhados. Além disso, o tipo de conhecimento que gerampode serutilizado portodas as outras aplicações já citadas.

O esclarecimento de ambigüidade se refere a um tipo comum de ruído gerado na tradução que, na maioria dos casos, está presente em apenas uma das línguas envolvidas na comunicação. Utilizando-se o alinhamento de textos paralelos, é possível recuperar informações sobre a língua que esclareçam a ambigüidade resultante da tradução.

O aprendizado de idiomas é outra aplicação que se beneficia do alinhamento de textos paralelos, pois os bancos de dados gerados a partir dos resultados desse alinhamento são valiosos para o aprendiz de uma língua estrangeira.

3. Conclusões

[9] Em: http://www.trados.com.
[10] Em: http://www.atril.com
[11] Em http://www.champollion.net.

Textos paralelos e textos paralelos alinhados são fontes ricas de conhecimento lingüístico e, por isso, são de grande importância para diversas aplicações de PLN. Este artigo apresentou quatro corpora paralelos PB-inglês construídos, inicialmente, para o projeto PESA, e as versões alinhadas sentencialmente obtidas como um dos resultados desse projeto.

Outros corpora paralelos e paralelos alinhados, incluindo outros gêneros e domínios de textos, principalmente, nos idiomas PB, inglês e espanhol, serão construídos como base para outras pesquisas em PLN. Uma dessas pesquisas, atualmente em desenvolvimento no NILC, visa utilizar os textos paralelos alinhados sentencialmente para extrair, de modo automático, algumas regras de tradução. Essa pesquisa é um exemplo de aquisição automática de conhecimento para a tradução automática, citada anteriormente (na Seção 2).

RESUMO: Textos paralelos – textos acompanhados de sua tradução em uma ou várias línguas – e textos paralelos alinhados – com as identificações das traduções – são fontes ricas de conhecimento lingüístico para muitas aplicações, como tradução automática. Neste artigo, são descritos quatro corpora paralelos, suas versões alinhadas sentencialmente, bem como algumas aplicações.

PALAVRAS-CHAVE: textos paralelos; textos paralelos alinhados; alinhamento sentencial de textos paralelos; tradução automática.

REFERÊNCIAS BIBLIOGRÁFICAS

FELTRIM, V.D.; NUNES, M.G.V.; ALUÍSIO, S.M. *Um corpus de textos científicos em português para a análise da estrutura esquemática*. Série de Relatórios do NILC. NILC-TR-01-4. 2001.

GAUSSIER, E., HULL, D., AÏT-MOKTHAR, S. Term alignment in use: Machine-aided human translation. In: VÉRONIS, J. (org.). *Parallel text processing*. s.l.: Kluwer Academic Publishers, p. 253-74, 2000.

MARTINS, M.S., CASELI, H.M., NUNES, M.G.V. *A construção de um corpus de textos paralelos inglês-português*. Série de Relatórios do NILC. NILC-TR-01-5. 2001.

RENOUF, A. Corpus development. In: SINCLAIR, J.M. (org.). *Looking up: An account of the COBUILD Project in lexical computing*. Londres/Glasgow: Collins, p. 1-22, 1987.

RIBEIRO, A.; LOPES, G.; MEXIA, J. Using confidence bands for parallel texts alignment. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, China, p. 432-439, 2000.

SANTOS, D.; OKSEFJELL, S. An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese. In: *Proceedings of the 12th "Nordisk datalingvistikkdager"*, Trondheim, Departmento de Lingüística, NTNU, p. 191-205, 2000.

SINCLAIR, J.M. *Corpus, concordance, collocation*. Oxford: Oxford University Press, p. 13-26, 1991.