CORPUS PARALELO E CORPUS PARALELO ALINHADO: PROPRIEDADES E APLICAÇÕES

Helena de Medeiros CASELI (PG – USP)

Textos paralelos – textos acompanhados de sua tradução em uma ou várias línguas – 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, ou seja, possuírem marcas que identifiquem os pontos de correspondência entre o texto original e sua tradução, o conhecimento daí derivado (as equivalências de tradução) assume importância capital em inúmeras aplicações de Processamento de Língua Natural (PLN). No alinhamento sentencial, são determinadas as correspondências entre as sentenças do texto original (texto fonte) e de sua tradução (texto alvo), ou seja, quais sentenças alvo são traduções das sentenças fonte e vice-versa. Este artigo apresenta o processo de construção de um corpus paralelo e sua versão sentencialmente alinhada, bem como algumas aplicações que se beneficiam destes recursos. O corpus construído é composto por resumos e *abstracts* de trabalhos acadêmicos desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo em São Carlos. Os textos paralelos deste corpus foram alinhados sentencial e automaticamente por métodos computacionais e uma versão manualmente alinhada também foi gerada para servir de referência na comparação com os alinhamentos produzidos pelos métodos. Esses corpora paralelo e alinhado podem ser utilizados em diversas aplicações de PLN como tradução automática, recuperação de informações por meio da troca de dados entre línguas diferentes e aprendizado de idiomas.