Complexidade textual em tradução: análise de contos literários através de ferramentas computacionais

Este trabalho partiu da percepção de um fenômeno recorrente durante a nossa prática profissional de revisão de traduções literárias. Tal fenômeno não podia ser caracterizado como erro de tradução, tampouco como inconsistência de estilo, mas sim como uma dissonância entre o nível de complexidade do texto-fonte e o nível de complexidade do texto traduzido, tendo em mente níveis de proficiência de leitura.

O objetivo deste trabalho é investigar a dissonância aqui nomeada como complexidade textual (CT) e criar subsídios para ajudar o profissional do texto e o tradutor a perceber algumas das sutilezas coesivas envolvidas na construção do texto de chegada e a tomar decisões tradutórias compatíveis com o nível de proficiência de leitura dos leitores a quem as traduções se destinam. Além disso, pretende-se contribuir para a avaliação das ferramentas de Processamento de Língua Natural (PLN) utilizadas, com o intuito de qualificá-las.

Revisão de literatura

Neste trabalho, investigamos a hipótese de que traduções de literatura em língua inglesa produzidas no Brasil são mais complexas do que seus textos-fonte, tendo como parâmetro o leitor brasileiro médio, cuja proficiência de leitura situa-se em nível básico, tal como apontam os dados do Indicador de Alfabetismo Funcional (INAF, 2009).

As pesquisas sobre o tema da CT partem de pontos de vista diversos e envolvem áreas que investigam o tema pela ótica da leitura e do ensino de leitura. Na bibliografia internacional, as pesquisas sobre complexidade linguística partiram da necessidade de adequar materiais de leitura a públicos específicos e baseavam-se no pressuposto de que os problemas de leitura estão relacionados a traços textuais mensuráveis, os quais, após a sua identificação, são inseridos em fórmulas cujos resultados estimam a legibilidade de um texto. Os traços mais comumente mensurados, até hoje, são a dificuldade lexical, baseada na frequência e na extensão das palavras, e a dificuldade imposta pelo tamanho da sentença, a partir do cálculo do número de palavras por sentença (Dubay, 2004).

Davidson e Green (1988) criticam a superficialidade de fórmulas puramente lexicais – ou seja, fórmulas com base na medida de frequência e extensão de palavras e frases – e argumentam que a complexidade de um texto não é um traço que possa ser fisicamente isolado sem que se levem em conta a complexidade sintática, características discursivas, estrutura retórica e assim por diante. Além disso, para essas pesquisadoras, as fórmulas não consideram aspectos importantes relativos ao leitor, como motivação e interesse, objetivo de leitura, etc. As autoras afirmam que uma visão mais holística da questão da CT desenvolveu-se a partir desses estudos iniciais, e essas novas perspectivas são provenientes de três fontes: do leitor, das características intrínsecas do texto e de teorias linguísticas.

Para Hoey (1991), o texto escrito só é "ativado" nos níveis sintático, fonológico, semântico e pragmático por meio da leitura e da interação com um leitor real. O texto oferece conexões semânticas potenciais tanto no nível da palavra quanto no nível da oração, mas é preciso que o leitor ative esses recursos e selecione as conexões mais relevantes.

Segundo Graesser et al. (2004), as fórmulas de legibilidade e de avaliação de complexidade ignoram componentes linguísticos e discursivos que influenciam na dificuldade de compreensão textual. Os autores apontam para o fato de que, apesar de os parâmetros de tamanho das sentenças e das palavras terem alguma validade, tais parâmetros não revelam a complexidade de um texto. Assim, propõem uma análise da coesão e da coerência textual em múltiplos níveis.

No Brasil, pesquisadores de PLN também têm se interessado por fórmulas e medidas de CT, adaptando-as ao português. É o caso do índice Flesch (Martins et al., 1996) e da ferramenta Coh-Metrix-Port (Scarton et al., 2009), adaptados para o português brasileiro. O índice Flesch, criado pelo austríaco Rudolf Flesch na década de 1940, estima a complexidade de um texto em uma escala de 1 a 100, sendo 1 equivalente a muito difícil e 100 a muito fácil (DUBAY, 2004), e a faixa entre 50 a 60 indica textos de complexidade média.

De acordo com resultados do INAF (2009), o analfabetismo funcional no Brasil diminuiu de 39% para 27% entre 2001 e 2009, e o índice de alfabetismo funcional aumentou de 60% para 73% no mesmo período. A diferença mais notável é o aumento de indivíduos na faixa do alfabetismo básico: de 34% em 2001 para 46% em 2009, compondo a maioria da população entre 15 e 64 anos. Um dado alarmante do levantamento feito pelo INAF mostra que 60% dos indivíduos que têm da 5ª. à 8ª. série do Ensino Fundamental e 54% dos indivíduos que cursaram o Ensino Médio são considerados alfabetizados em nível básico. Assim, de acordo com os escores do índice Flesch, textos recomendados para o leitor brasileiro médio não deveriam apresentar um escore inferior a 60, índice que indica textos adequados a leitores com letramento básico.

Metodologia

Partindo da microperspectiva estrutural do texto, isto é, considerando sua costura coesiva, a pesquisa empreendida aqui é um estudo quantitativo e qualitativo sobre métricas (provenientes das ferramentas Coh-Metrix e Coh-Metrix-Port) para estimação de CT em um pequeno córpus:

- Bloco 1: originais em inglês e respectivas traduções para o português brasileiro, composto por 14 contos dos seguintes autores: Edgar Allan Poe (10); Nathaniel Hawthorne (01), O. Henry (01), Virginia Woolf (01) e James Joyce (01). Os tradutores dos contos de Edgar Allan Poe são: Marcelo Bueno (01), Oscar Mendes (04), Bernardo Carvalho (02), Celina Portocarrero (01), Rodrigo Breunig (01) e Dorothée de Bruchard (01). Os tradutores dos contos restantes são: Roberto Schmitt-Prym (01), Bianca Pasqualini (01) e Zaida Maldonado (01). Os textos originais têm uma média aproximada de 1.800 palavras (*tokens*) cada.
- Bloco 2: originais em português e respectivas traduções para o inglês, composto por 14 contos dos seguintes autores: Machado de Assis (06), Coelho Neto (02), Humberto de Campos (03) e Lima Barreto (03). Os tradutores são: Isaac Goldberg (02), Francis Johnson (10) e Gregory Rabassa (02). Os textos originais têm uma média aproximada de 1.600 palavras (*tokens*) cada.

Em primeiro lugar, é preciso lembrar que as ferramentas Coh-Metrix e Coh-Metrix-Port não foram criadas com a intenção de contrastar traduções. Isso gerou uma série de dificuldades que precisaram ser contornadas. Diferentemente do procedimento comum a trabalhos cuja metodologia segue os preceitos da Linguística de Corpus, os textos foram tratados individualmente. Uma particularidade da preparação dos textos foi a necessidade de corrigir marcas de parágrafo, letras maiúsculas e pontuação, uma vez que interferem no processamento textual das ferramentas e, assim, nos resultados. Desse modo, os textos foram salvos em arquivos individuais com extensão DOC, com um cabeçalho contendo informações como título, autor, fonte e número de caracteres. Após a divisão em blocos, cada um deles foi processado individualmente pelas ferramentas.

Das 48 métricas adaptadas do Coh-Metrix para o Coh-Metrix-Port, apenas 31 são comparáveis, pois há métricas próprias a cada recurso e métricas incompatíveis devido aos recursos utilizados, como a Wordnet, que é um recurso exclusivo do inglês. Inicialmente, selecionamos as métricas a serem analisadas, englobando todas as categorias de análise (lexicais, sintáticas e semânticas, tendo em vista que a categoria de medidas do tipo referencial ainda está em construção). O segundo passo foi realizar o teste *t-Student*, para cada métrica e entre os blocos de textos, para avaliar se as diferenças entre as médias dos resultados obtidos eram significativamente diferentes com confiança de 95% (*p-value* < 0,05). Assim, das 31 métricas comparáveis, 18 apresentaram resultados com diferenças estatisticamente significativas. Destas, seis foram eliminadas por problemas de ordem técnica. Por fim, os resultados das 12 métricas foram processados com a ferramenta Weka (conjunto de algoritmos de Aprendizagem de Máquina [AM]) a fim de aferir as métricas mais características, sob o ponto de vista estatístico, a cada bloco de textos. O algoritmo escolhido foi a implementação J48 do algoritmo de classificação C4.5 para construção de árvores de decisão. A árvore de decisão mostra quais são as relações discriminativas entre os atributos (no caso, as métricas das ferramentas) do total das instâncias (cada um dos textos analisados) em cada classe (ou seja, classe de textos originais e classe de textos traduzidos). Em outras palavras, a estrutura em árvore de decisão mostra visualmente quais são as métricas mais

distintivas em cada bloco estudado, de acordo com a sua natureza: (1) originais em inglês, (2) traduções para o português, (3) originais em português ou (4) traduções para o inglês.

Resultados

Na análise dos resultados das métricas, a confirmação da hipótese de que textos literários em inglês traduzidos para o português brasileiro tendem a ser mais complexos do que os seus textos-fonte ficou, de certa forma, diluída na comparação entre as medidas, pois nem todas indicaram maior complexidade das traduções para o português. Entretanto, na análise por AM, as métricas mais características a cada grupo textual foram determinadas, e então pudemos ter mais confiança nessa afirmação, uma vez que o índice Flesch revelou um nível de complexidade maior das traduções investigadas. Contudo, o índice Flesch é considerado um cálculo superficial. É preciso levar em conta, porém, que o índice Flesch foi mais um entre vários elementos nesta pesquisa; mesmo assim, destacou-se em todas as análises como altamente relevante na comparação dos níveis de CT do córpus estudado, tanto no contraste das métricas das ferramentas quanto na classificação das métricas por técnicas de AM. Desse modo, o índice Flesch, contextualizado e enriquecido pelo acréscimo de outros elementos textuais e de abordagens estatísticas de análise de CT, mostrou-se um indicador confiável de que, no que tange aos textos processados neste trabalho, as traduções para o português são mais complexas do que os seus textos-fonte e são também mais complexas do que os textos dos autores brasileiros.

Conclusão

A importância deste trabalho reside não só na descrição de índices lexicais, sintáticos e semânticos dos textos das línguas abordadas separadamente, mas também no levantamento de índices contrastivos entre o português e o inglês em tradução, mantendo a individualidade do texto dentro do córpus. A análise dos resultados, sob o ponto de vista linguístico, é relevante para a qualificação das ferramentas. Além disso, a análise contrastiva das métricas é um primeiro passo para uma possível automatização de índices de CT em traduções, ou seja, para a criação de uma ferramenta de processamento de CT de textos traduzidos do inglês para o português. Finalizamos afirmando que, por trás da proposta desta investigação, está a convicção de que a leitura, seja de textos traduzidos ou não, deve ser inclusiva, deve trazer o leitor para o texto, a fim de permitir que a produção artística e intelectual de todas as épocas seja compartilhada, e não compartimentalizada por aqueles que dela se apropriam e que lhe atribuem um significado acessível apenas a poucos.