Michael Carl e Moritz Jonas Schaeffer

Por que tradução é difícil: um estudo de corpus baseado na não literalidade da pós-edição e da tradução do zero

RESUMO: O estudo desenvolve uma definição de literalidade na tradução que é baseada na similaridade sintática e semântica entre os textos fonte e alvo. Nós providenciamos evidências teóricas e empíricas que traduções absolutamente literais são fáceis de produzir. Baseado em um corpus multilíngue de traduções alternativas, nós investigamos os efeitos de distância sintática e semântica no tempo de produção de uma tradução e descobrimos que a não literalidade faz com que a tradução do zero e a pós-edição sejam difíceis. Nós demonstramos que sistemas de tradução automática estatística encontram ainda mais dificuldades com a não literalidade.

PALAVRAS-CHAVE: tradução; pós-edição; não literalidade; tradução automática estatística.

1. Introdução

Tradução seria fácil se todas as palavras de uma língua fonte tivessem apenas uma tradução possível para a língua alvo (e vice-versa) e se a ordem das palavras fosse idêntica nos textos fonte e alvo. Nesse caso, traduzir seria reduzido ao ato de substituir palavras fonte para palavras alvo que poderiam ser enumeradas e pesquisadas em um dicionário. Se traduzir fosse fácil desse jeito, a tradução automática (TA) seria perfeita, não haveria necessidade de tradutores humanos nem de pós-edição de tradução automática (PEMT), pois, a simples e determinística substituição de léxicos produziria traduções perfeitas. Entretanto, sabemos que esse não é o caso. Traduções absolutamente literais como essas são raramente possíveis e dependem das similaridades e das possibilidades das línguas envolvidas. Polissemia e diferentes semânticas e representações conceituais na língua fonte e na língua alvo, assim como restrições sintáticas são algumas das razões pelas quais a tradução é difícil e não determinista.

Sun (2015, p. 31) argumenta que "a dificuldade ao traduzir pode ser vista como a que medida recursos cognitivos são usados em uma tradução para que o tradutor alcance os critérios de performance objetivos e subjetivos". Ele menciona um grande número de causas, efeitos e fatores para determinar a dificuldade de uma tradução. Ele sugere uma medida para a dificuldade da tradução em que "o número de diferentes traduções possíveis talvez não seja um indicador efetivo da dificuldade da tradução" (SUN, 2015, p. 42). Entretanto, como Koponen (2016, p. 24) destaca, o número de variáveis disponíveis e consideradas pelo tradutor ou pós-editor são mencionadas como indicadores do esforço cognitivo tanto por Krings (2001, p. 536-537) quanto por Dimitrova (2005, p. 26).

É o nosso objetivo encontrar um modo de descobrir o que torna uma tradução difícil e não determinística. Para fazer isso, nós sugerimos a definição de uma hipotética *tradução completamente literal*, que é sintática e semanticamente idêntica à fonte, e desenvolvemos um cálculo para medir a não literalidade de traduções reais. Nós fornecemos evidências empíricas e teóricas de que traduções completamente literais são mais fáceis (e mais rápidas) de produzir e mostramos que quanto mais a tradução se afasta do critério de literalidade, mais difícil e demorado será para realizar uma tradução do zero e uma PEMT. Nós mostramos que a pontuação de literalidade da tradução prevê medidas comportamentais dos toques de teclas e duração em tradução e em pós-edição. Também demonstramos que ambiguidades na tradução automática estatística e o esforço gasto na pós-edição automática podem apontar para o mesmo fenômeno de não literalidade.

Nessa dissertação, nós apresentamos uma abordagem empírica para medir a similaridade sintática e semântica usando uma abordagem de corpus. Baseamos nossa investigação em um corpus multilíngue que contém um grande número de traduções alternativas, ou seja, traduções do mesmo texto fonte produzidas por tradutores diferentes. Nós tomamos as variações de ordem das palavras e de escolhas lexicais nas

traduções como indicadores da literalidade sintática e semântica. Focamos no efeito de duração assumindo que quanto menor o tempo de produção, menor o esforço cognitivo (KRINGS, 2001).

Nossas investigações são baseadas em um subgrupo do TPR-DB, um corpus multilíngue de traduções "palavra-por-palavra" alternativas das quais induzimos a similaridade semântica e sintática dos textos fonte, assim como os traços de esforço cognitivo para produzir essas traduções.

Uma parte considerável dos estudos de bilinguismo investiga os mecanismos e as restrições da tradução humana fazendo uso de estudos de preparação. Estudos de preparação medem o impacto de estímulos previamente encontrados em processos de recuperação subsequentes para investigar os processos e representações mentais subjacentes. Portanto, a Seção 2 apresenta alguns conceitos de pesquisas sobre bilinguismo (estudos de preparação). Nós utilizamos um modelo de tradução psicolinguisticamente embasado (SCHAEFFER/CARL, 2013) que explica os resultados em um contexto de representações compartilhadas e um modelo repetitivo do processo de tradução. Os resultados, assim como o modelo, indicam que traduções literais são mais fáceis de serem produzidas do que as não literais. Na seção 3, nós preparamos o conceito de literalidade de tradução e desenvolvemos duas métricas que são usadas para pontuar as traduções como mais ou menos literais. Na seção 4, nós introduzimos o corpus de tradução e pós-edição de dados *multiLing*, que constitui a base empírica do nosso estudo.

Os resultados na seção 5 sugerem que escolhas sintáticas e lexicais estão fortemente ligadas na tradução e na PEMT e que o grau de similaridade semântica da fonte e do alvo (ou seja, a entropia da tradução das palavras) tem um impacto na duração da tradução e da PEMT. Entretanto, a seção 6 fornece evidências de que a PEMT leva a traduções mais literais e a menores variações lexicais na tradução e de que isso pode resultar em uma tradução de menor qualidade que uma tradução do zero. Na seção 7, nós apontamos escolhas durante a PEMT para a tradução automática estatística (TAE) que geraram os rascunhos das traduções que foram depois pós-editadas. Encontramos uma relação transitiva entre a complexidade no gráfico de busca do sistema TAE e a complexidade do rendimento da pós-edição. Isso mostra que se escolhas lexicais são difíceis (ou seja, muito) para um sistema TAE, então tem uma grande chance de elas também serem difíceis para os tradutores. Os resultados sugerem que os processos subjacentes em traduções do zero e na PEMT podem compartilhar algumas características em comum.

Michael Carl & Moritz Jonas Schaeffer

Why Translation Is Difficult: A Corpus-Based Study of Non-Literality in Post-Editing and From-Scratch Translation

Abstract The paper develops a definition of translation literality that is based on the syntactic and semantic similarity of the source and the target texts. We provide theoretical and empirical evidence that absolute literal translations are easy to produce. Based on a multilingual corpus of alternative translations we investigate the effects of cross-lingual syntactic and semantic distance on translation production times and find that non-literality makes from-scratch translation and post-editing difficult. We show that statistical machine translation systems encounter even more difficulties with nonliterality.

Keywords translation, post-editing, non-literality, statistical machine translation

1. Introduction

Translation would be easy if every word in the source language had only one possible translation in the target language (and vice versa) and the word order would be identical in the source and the target texts. In

that case, translation would be reduced to substituting source words for target words that could be enumerated and looked up in a dictionary. If translation was that easy, machine translation (MT) would work perfectly: there would be no need for human translators and not even post-editing of machine translation (PEMT) would be required, since simple, deterministic lexical substitution would produce perfect translations. However, we know that this is not the case. Such absolute literal translations are only exceptionally possible and depend on the similarities and possibilities of the languages involved. Polysemy, different semantic and conceptual representations in the source and the target languages, as well as different syntactic constraints are some of the reasons why translation is non-deterministic and difficult.

Sun (2015: 31) argues that "translation difficulty can be viewed as the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria." He mentions a large number of causes, effects and factors for assessing translation difficulty. He suggests a measurement for translation difficulty where the "number of different renditions may not be an effective indicator of translation difficulty" (Sun 2015: 42). However, as Koponen (2016: 24) highlights, the number of variants available to and considered by the translator or post-editor are mentioned as indicators of cognitive effort by both Krings (2001: 536-537) and Dimitrova (2005: 26).

It is our aim to find a way of describing what makes translation difficult and non-deterministic. In order to do so, we suggest a definition of a hypothetical *absolute literal translation* which is syntactically and semantically identical to the source and develop a computational framework to measure the non-literality of actual translations. We provide theoretical and empirical evidence that absolute literal translations are easier (faster) to produce and we show that the more the translation deviates from the literality criterion, the harder it is and the longer it takes to produce the translation in from-scratch translation and in PEMT. We show that the literality scores of the translation product predict behavioral measures of keystrokes and gaze times in translation and in post-editing. We also show that ambiguities in statistical machine translation and the effort spent in machine translation post-editing can be traced back to the same non-literality phenomena.

In this paper, we present an empirical approach to measure cross-lingual syntactic and semantic similarity using a corpus-driven approach. We base our investigation on a multilingual corpus which contains a large number of alternative translations, i.e., translations of the same source text produced by different translators. We take the variation in word order and the variation of lexical choice in the translations as indicators for syntactic and semantic literality. We focus on temporal effort under the assumption that shorter production times indicate less cognitive effort (see Krings 2001).

Our investigations are based on a subset of the TPR-DB,1 a multilingual corpus of alternative word-aligned translations from which we induce the semantic and syntactic similarity of the source texts, as well as the traces of cognitive effort to produce these translations.

A substantial body of work in bilingualism studies investigates the mechanisms and constraints of human translation making use of priming studies. Priming studies measure the impact of a previously encountered stimulus on subsequent retrieval processes in order to investigate the underlying mental representations and processes. Therefore, Section 2 presents some concepts in bilingualism research (priming studies). We draw on a psycho-linguistically grounded translation model (Schaeffer/Carl 2013) which explains the findings in the context of shared representations and a recursive model of the translation process. The findings, as well as the model, indicate that more literal translations are easier to produce than less literal translations. In section 3 we operationalise the concept of translation literality and develop two metrics which are used to grade translations as more or less literal. In section 4, we introduce the *multiLing* corpus of translation and post-editing data, which constitutes the empirical basis of our study.

Findings in section 5 suggest that syntactic and lexical choices are closely related in translation and PEMT and that the degree of semantic similarity of the source and the target (i.e. the word translation entropy) has an impact on translation and on PEMT duration. However, section 6 provides evidence that PEMT leads to more literal translations and less lexical variation in the translation product, and it may also result in lower translation quality than from-scratch translation. In section 7, we trace back choices during PEMT to the

internal representations of statistical machine translation (SMT) that generated the draft translations which were then post-edited. We find a transitive relation between the complexity in the search graph of the SMT system and the complexity of post-edited output. It shows that if lexical choices are difficult (i.e. plenty) for an SMT system then they are also likely to be difficult for translators. The results suggest that the underlying processes in from-scratch translation and PEMT might share common characteristics.