OSLa Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 207–222. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

TRADUÇÃO AUTOMÁTICA, MA NON TROPPO

ANABELA BARREIRO

ABSTRACT

This paper describes two machine translation tasks that require language expertise: (1) paraphrasing as a technique to prepare texts for translation and a method for linguistic quality assurance, and (2) the evaluation of translation produced by machine translation systems. These tasks will be exemplified through support verb constructions, a subtype of multiword units that machine translation systems have difficulty translating. The paper raises awareness of the need to integrate enhanced linguistic knowledge in machine translation systems and the need to place the human factor as a core value in order to ensure translation quality.

[1] INTRODUÇÃO

A tecnologia de tradução automática chegou à vida do comum dos mortais com o advento da Internet. Apesar de a qualidade da tradução ainda ficar aquém das expectativas dos utilizadores linguisticamente mais exigentes, esta é uma ferramenta imprescindível para resolver as necessidades diárias de tradução de milhões de internautas. Por esse motivo, os investigadores e programadores de sistemas de tradução automática sentem a necessidade de criar sistemas linguisticamente mais robustos e capazes de produzir traduções com qualidade comparável às que são produzidas por tradutores humanos. Décadas de investigação nesta área resultaram na invenção e aperfeiçoamento de métodos estatísticos que aceleram o processo de tradução, e no desenvolvimento de ferramentas e de recursos linguísticos de melhor qualidade, em maior quantidade e para mais línguas. Os avanços alcançados em diferentes aproximações e técnicas de tradução automática tornam-se um campo fértil para o desenvolvimento de uma nova geração de sistemas linguisticamente mais avançados, os sistemas híbridos, que combinam regras de análise e aquisição de conhecimento linguístico típicas dos sistemas de regras (Scott 2003) com métodos estatísticos característicos dos sistemas baseados em dados (Koehn 2005). A combinação de sistemas conduz geralmente a uma melhoria da qualidade da tradução, na medida em que sistemas de natureza diferente abordam desafios de tradução diferentes, completando-se na resposta às dificuldades. No entanto, embora os sistemas híbridos representem uma linha de investigação promissora, a tradução automática é um problema longe de estar resolvido. Uma hibridação bem sucedida requer uma compreensão profunda das diferentes abordagens, dos seus pontos fracos e fortes, tema que tem sido discu-

tido apenas marginalmente na investigação em tradução automática. A integração ainda mal explorada dos recursos linguísticos em sistemas essencialmente estatísticos é, em muito, responsável pelos erros crassos que as traduções produzidas pelos sistemas de tradução automática online apresentam, impedindo que estas sejam usadas para fins comerciais na ausência de um esforço significativo de pós-edição. No caso dos sistemas de base gramatical, a falta de recursos linguísticos para alimentar as bases de dados destes sistemas também cria graves lacunas de origem maioritariamente lexical. Não se sabe ainda que aproximação híbrida será a mais eficaz a longo prazo e conduzirá a uma qualidade de tradução superior.

Enquanto os investigadores procuram avançar o estado da arte e melhorar a tecnologia através da criação e desenvolvimento de sistemas que traduzem cada vez melhor, a tradução automática representa uma realidade que já não pode ser ignorada também no universo da tradução profissional, fazendo parte da formação e currículo dos tradutores (Maia 2005). Apesar dos resultados ainda pouco fidedignos, a tradução automática começa a integrar o quotidiano de um número crescente de clientes e mercados, que colmatam as suas deficiências através do treino de sistemas em domínios específicos usando corpora baseados em textos traduzidos profissionalmente para esses domínios (Bick & Barreiro 2015) e através do uso de ferramentas automáticas de pós-edição dos textos traduzidos automaticamente (Vieira & Specia 2011). Por conseguinte, na esfera da tradução profissional, a intervenção humana é essencial no processo de correção e certificação do controlo de qualidade linguística da tradução automática. Outra forma de intervenção e que tem sido menos explorada do ponto de vista do processamento da linguagem natural é a do parafraseamento usado como técnica de pré-edição do texto da língua-fonte, por vezes conduzindo a uma linguagem controlada usada em textos técnicos e científicos. Queremos aqui reforçar que uma tradução automática de qualidade não será alcançável sem o fator humano, nomeadamente sem a intervenção de especialistas das línguas envolvidas na tradução e a sua participação nas tarefas que visam a qualidade do texto a traduzir e do texto traduzido.

Este artigo apresenta duas importantes tarefas da tradução automática que requerem a participação de peritos com conhecimentos linguísticos profundos das línguas de tradução. A primeira tarefa consiste no parafraseamento como método de preparação do texto na língua-fonte, de modo a garantir uma melhor qualidade de tradução desse texto. A segunda tarefa corresponde à avaliação da tradução produzida pelos sistemas de tradução automática. As duas tarefas serão exemplificadas através das construções com verbos-suporte, um tipo de unidade lexical multipalavra que os sistemas de tradução automática em vigor não conseguem traduzir com qualidade.

[2] CONSTRUÇÕES COM VERBOS-SUPORTE EM TRADUÇÃO AUTOMÁTICA

As construções com verbo-suporte são um tipo de unidade lexical multipalavra que se caracteriza pela ocorrência de verbos semanticamente fracos, designados verbos-suporte, e predicados nominais (*fazer um esforço* (*por*)), adjetivais (*ser útil* (*para*)) ou adverbiais (*ficar aquém* (*de*)). Estas construções desempenham um papel de destaque na comunicação em muitas línguas, incluindo o português. Num estudo anteriormente realizado por Barreiro (2009), num corpus de 500 frases, 64.2% da ocorrência dos verbos *dar*, *tomar*, *pôr*, *fazer* e *ter*, realiza-se em construções com verbos-suporte; i.e., em apenas 33.8% dos casos esses verbos ocorrem como verbos plenos. A léxico-gramática, proposta por Gross (1984) e estabelecida no quadro da gramática transformacional harrissiana, explora uma metodologia sistemática para o processamento automático das construções com verbos-suporte, contemplando trabalhos para o português (Ranchhod 1983, 1990; Baptista 2005; Chacoto 2005) e estudos contrastivos entre o inglês e o francês (Salkoff 1999).

As unidades lexicais multipalavra, nas quais as construções com verbos-suporte se incluem, ocorrem frequentemente quer em textos genéricos (Gross & Senellart 1998) quer em textos técnicos ou de domínios específicos (Ramisch et al. 2010) em muitas línguas. A integração eficaz das unidades lexicais multipalavra em modelos de tradução automática tem sido assinalada como um fator de impacto na obtenção de tradução de qualidade (Chiang 2005; Marcu et al. 2006; Zollmann & Venugopal 2006). A reforçar esta posição, a avaliação da tradução automática de construções com verbos-suporte descrita em Barreiro et al. (2013) comprovou que estas unidades multipalavra *constituem um osso duro de roer* para o processamento de linguagem natural, especialmente para a tradução automática. A maioria dos sistemas não consegue apresentar uma solução eficaz para o problema da não composicionalidade das construções com verbos-suporte que, quando processadas de forma incorreta, provocam um impacto negativo na compreensibilidade e qualidade das traduções.

A ambiguidade dos verbos-suporte, por um lado, e sua leveza semântica, por outro, representam fatores adversos à tradução (humana e automática) das construções com verbos-suporte, que impedem que estas sejam, em muitos casos, traduzidas literalmente. As suas traduções, por vezes idiomáticas e pouco previsíveis, devem-se ao facto de nem sempre existir uma expressão equivalente na língua-alvo ou, no caso de existir, essa expressão assumir uma forma distinta da forma da língua-fonte. Por exemplo, mesmo com proximidade estrutural entre as duas línguas, a expressão em português *dar um passeio* traduz-se em inglês por *take a walk* ou *go for a walk*. Uma tradução literal da expressão por **give a walk* teria um efeito nefasto para a qualidade da tradução.

As propriedades morfossintáticas das construções com verbos-suporte permitem um certo número de variações formais com a possibilidade de dependências

entre os elementos mesmo quando estão distantes entre si na frase. Por exemplo, *deu* [*muitos e longos*] *passeios pela* [N] ou *não fez* [*absolutamente nenhum*] *comentário sobre* [N] representam construções com verbos-suporte não adjacentes que mantêm inserções entre os verbos-suporte *dar* e *fazer* e os predicados não verbais *passeios* e *comentário*, respetivamente. Uma inserção é qualquer palavra que se encontre entre dois elementos da unidade lexical multipalavra, exceto se essa palavra for um artigo definido ou indefinido antes de um nome predicativo. Em geral, quanto mais inserções e variabilidade morfossintática existir numa construção com verbo-suporte, mais difícil é a sua tradução automática. Os estudos já referenciados mencionam também a variedade linguística apresentada pelas variantes estilísticas ou parafrásticas (*fazer um estudo* = *realizar/efetuar/desenvolver um estudo* ou *fazer um trabalho* = *elaborar um trabalho*, entre outras), que utilizam verbos-suporte não elementares (Ranchhod 1990). Essas variantes estilísticas podem apresentar diferentes graus de variabilidade, indo desde as construções que permitem um número consideravelmente extenso de inserções entre o verbo-suporte e o predicado nominal, até as expressões idiomáticas semi- ou totalmente fixas (*dar o braço a torcer* = *ceder*)[1]. Construções com verbos-suporte não adjacentes são difíceis de processar, alinhar e traduzir, permanecendo um dos maiores desafios contrastivos para os sistemas de tradução automática.

[3] FACTOR HUMANO NO CONTROLO DA QUALIDADE LINGUÍSTICA

Desde que os sistemas de tradução automática estatística começaram a ser treinados com grandes quantidades de dados, nomeadamente com milhões e milhões de corpora paralelos disponíveis na internet, que o efeito de erro gramatical se começou a diluir e a ter um impacto gradualmente menor em traduções cada vez mais robustas do ponto de vista lexical. Ao nível da tradução comercial, os menores custos envolvidos na tarefa da pós-edição justificam o uso da tradução automática e um papel relevante desempenhado pelos tradutores tem consistido na correção dos erros gramaticais nos textos traduzidos automaticamente. No entanto, muitos dos problemas linguísticos das traduções automáticas têm na sua base a falta de qualidade do texto na língua-fonte. Em geral, o controlo da qualidade linguística dos textos da língua-fonte tem sido relegado para segundo plano, não havendo ferramentas robustas de auxílio à edição e revisão de texto que envolvam parafraseamento. Neste sentido, em trabalho anteriormente realizado, apresentámos uma abordagem científica baseada no parafraseamento que tem como objetivo melhorar a tradução automática (Barreiro 2009), acentuando a necessidade

[1] Como expressões idiomáticas entendem-se expressões não transparentes, não entendidas/traduzidas literalmente, em que o significado da expressão é diferente do significado individual das palavras que a constituem. Podemos considerar a existência de uma 'gradação' da idiomaticidade, que pode variar entre o ligeiramente não literal e o muito obscuro. Algumas expressões idiomáticas assumem um valor figurativo que se conhece apenas através do uso comum, outras acabam por fossilizar-se com o passar do tempo.

de uma aproximação parafrástica para a resolução de problemas que levantam no campo da tradução, tal como descrevemos na secção [3.1]. O controlo da qualidade linguística também não prescinde de uma avaliação sistemática, feita por humanos, que permite verificar onde e como é que os sistemas falharam. De facto, a avaliação é uma fase importante no desenvolvimento de um sistema de tradução automática e é dela que depende a qualidade das traduções obtidas pelos tradutores automáticos. Também nesta área, apresentámos anteriormente um exercício de avaliação humana sistemática de construções com verbos-suporte que contempla a sua tradução do inglês para várias línguas por dois sistemas de tradução automática conhecidos (Barreiro et al. 2014). A avaliação realizada prova que os atuais sistemas de tradução automática não conseguem traduzir com qualidade os fenómenos linguísticos representados por estas construções. É com base nos estudos referidos e face aos desafios que as construções com verbos-suporte apresentam para a tradução automática que reforçamos e ilustramos a necessidade de envolver especialistas linguísticos nas diferentes tarefas da tradução automática.

[3.1] *Parafraseamento como Técnica de Pré-Edição*

Um dado importante no estudo das construções com verbos-suporte consiste em estas lhes terem geralmente associadas um verbo semanticamente forte, morfossintaticamente relacionado, que constitui o seu sinónimo. Por exemplo, a construção com verbo-suporte *fazer uma apresentação* (*de*) é morfossintática e semanticamente equivalente ao verbo *apresentar*. Uma das abordagens por nós realizada anteriormente (Barreiro 2009) consiste no parafraseamento de construções com verbos-suporte de modo a melhorar a qualidade da tradução automática. Para além da criação de uma ferramenta de parafraseamento, o desafio dessa investigação consistiu em parafrasear expressões nominais predicativas por construções verbais (*fazer uma análise* = *analisar*), tirando partido das potencialidades parafrásticas da língua. Em casos particulares, o parafraseamento consistiu em substituir o verbo-suporte da construção nominal, semanticamente fraco, por uma variante lexical ou estilística (*realizar uma análise* ou *efetuar uma análise*), entre outras. Quando as construções com verbos-suporte foram identificadas e substituídas por verbos lexicais ou expressões verbais semanticamente equivalentes ou próximas, numa fase de pré-processamento do texto, obteve-se aproximadamente 21% de melhoria na qualidade dos resultados avaliados da tradução automática do português para o inglês e aproximadamente 31% na dos resultados avaliados da tradução automática do inglês para o português. A investigação baseou-se numa análise linguística contrastiva, em que as construções com verbos-suporte foram organizadas em subclasses sintático-semânticas de acordo com os princípios teóricos e metodológicos da Léxico-Gramática. Esse estudo incidiu sobre as construções com verbos-suporte, mas seria interessante aplicá-lo a outros tipos de unidades lexicais multipalavra, nomeadamente a expressões idiomáticas, mas também a

construções sintáticas livres, tais como a coordenação de sintagmas nominais e a passiva, entre outras. A informação linguística relevante para a construção das paráfrase que foram geradas (como resultado dessa investigação) foi formalizada em dicionários e gramáticas desenvolvidos no ambiente linguístico NooJ e utilizados em várias tarefas de processamento de língua natural, sob o ponto de vista monolingue e bilingue. Os recursos bilingues português-inglês do Port4NooJ, disponível em domínio público[2], integram a ontologia SAL do modelo OpenLogos e foram construídos como o alicerce desse estudo. O seguimento desse trabalho deu origem aos sistemas ReEscreve, ReWriter, ParaMT e eSPERTo apresentados em (Barreiro 2008, 2009, 2011; Barreiro & Cabral 2009; Barreiro et al. 2011). O eSPERTo é um Sistema de Parafraseamento para Edição e Revisão de Texto, atualmente em fase de desenvolvimento no âmbito de um projeto com o mesmo nome[3]. Este projeto tem como objetivo o desenvolvimento de uma plataforma web para geração de paráfrases linguisticamente complexas. As paráfrases serão geradas a partir da aplicação de uma técnica híbrida de aquisição de conhecimento linguístico baseada em estatística e regras gramaticais. A integração de conhecimento frásico e de unidades lexicais multipalavra no sistema permitirá um mapeamento otimizado de construções, estruturas e frases semanticamente equivalentes, que servirá de auxílio no ensino de escrita e na produção e revisão de textos em português. Este conhecimento linguístico poderá ser também usado em pré-edição para a tradução automática, de modo a garantir uma maior qualidade dos textos a traduzir e da qualidade da tradução desses textos.

[3.2] *Avaliação da Qualidade da Tradução Automática*

A tarefa de avaliação da qualidade da tradução automática para o português ganhou força no início da década de 2000, com os primeiros esforços direcionados para uma avaliação conjunta no âmbito do projeto Linguateca. Nessa época, criou-se um grupo de interesse na área, o ARTUR, integrado no AVALON 2003, que deu origem a diversos trabalhos sobre avaliação da tradução automática, nomeadamente o desenvolvimento de uma ferramenta automática de geração de baterias de teste e de um programa de categorização de erros, realizados na Universidade do Porto (Maia et al. 2003, 2004; Maia & Barreiro 2007; Sarmento et al. 2007). A avaliação desta área permitiu identificar problemas relacionados com a preservação de significado no processo de tradução, em particular em no que respeita a usos não literais, envolvendo construções idiomáticas, coloquialismos, usos metafóricos, entre outros. Nesta linha de ação, e na tentativa de criar um modelo híbrido de tradução automática melhorando a tecnologia atualmente existente, uma análise humana sistemática do desempenho de diferentes modelos pareceu-nos um passo importante a dar. Muito do trabalho de avaliação que se tem feito nos últi-

[2] http://www.linguateca.pt/Repositorio/Port4NooJ/
[3] http://esperto.l2f.inesc-id.pt/

mos anos incide essencialmente sobre a tarefa da pós-edição e contempla aspetos relacionados com a definição de métricas de medição do esforço humano e tempo usados na correção de erros gerados pelos sistemas, tais como contar a quantidade de teclagem utilizada pelos revisores (Aziz et al. 2012). Foi com base nas lacunas verificadas ao nível da avaliação qualitativa dos fenómenos linguísticos em sistemas de tradução automática com abordagens diferentes, que propusémos, em Barreiro et al. (2014), uma avaliação humana conjunta dos erros de tradução de construções com verbos-suporte pelo OpenLogos e pelo Google Translate.

O OpenLogos (Scott 2003; Barreiro et al. 2011) é a cópia em código aberto do sistema comercial Logos, um sistema pioneiro de tradução automática (1970-2001). Baseia-se em regras que contemplam a morfologia, a sintaxe e a semântica, mas assemelha-se em espírito à aproximação estatística na medida em que as regras são aplicadas a padrões em contexto. O sistema tem analisadores sintáticos (*parsers*) robustos, conjuntos de regras semântico-sintáticas, terminologia e ferramentas variadas, tais como um construtor automático de termos (TermBuilder) e uma ferramenta de aquisição automática de regras semânticas (Semantha), entre outras. Devido à sua ênfase na semântica, é considerado um sistema de alta qualidade, que se baseia na análise da língua de forma a que esta seja "entendida" pelo sistema computacional. O "motor" que faz girar o sistema consiste numa linguagem de representação intermédia (SAL) que é usada para codificar toda a informação linguística e processar texto. O conhecimento linguístico representado nesta linguagem permite aliviar o problema da escassez de dados e colmatar falhas apresentadas pelos métodos estatísticos, contribuindo para um aumento da qualidade das traduções. Devido ao grande investimento de tempo e recursos humanos no desenvolvimento do sistema OpenLogos, as suas bases de dados de conhecimento linguístico já não são atualizadas desde 2001.

O Google Translate é um dos sistemas de tradução online mais usados na atualidade. É um sistema de base estatística que beneficia de grandes volumes de corpora paralelos existentes na internet. O Google Translate traduz mais de 80 pares de línguas, mas a qualidade da tradução varia muito do par de línguas envolvido, produzindo melhores resultados para pares de línguas mais próximas (português e espanhol) e línguas para as quais existam grandes quantidades de corpora paralelos. A qualidade dos dados também é pertinente para a tradução, pelo que quanto melhor for a qualidade dos corpora de um par de línguas, melhor será a qualidade dos textos traduzidos para essas línguas. As traduções podem variar de qualidade dependendo do domínio do texto e dos corpora (ou outros recursos) que foram utilizados para treinar o sistema nesse domínio. O Google Translate é um sistema comercial não se sabendo como funciona, e muito menos se tem algum módulo de 'compreensão semântica'.

A avaliação do desempenho dos sistemas OpenLogos e Google Translate relativamente às traduções de construções com verbos-suporte, para além de nos

ter dado a possibilidade de contrastar um sistema de regras baseadas em padrões com um sistema estatístico, permitiu-nos diagnosticar e avaliar qualitativamente erros de tradução em fenómenos linguísticos muito específicos.

Corpus e Metodologia de Avaliação

O corpus usado na avaliação contém 100 construções com verbos-suporte que ocorrem em frases recolhidas de notícias e da internet (textos genéricos, de nenhum domínio específico). Cada construção com verbo-suporte foi anotada no contexto frásico em que se encontra e classificada de acordo com a tipologia apresentada na Tabela 1. Seguidamente, o corpus foi traduzido para alemão, espanhol, francês, italiano e português pelos sistemas de tradução automática OpenLogos e Google Translate. Nenhum dos sistemas foi previamente treinado para esta tarefa de avaliação. Linguistas falantes nativos das línguas de chegada avaliaram a qualidade da tradução das construções com verbos-suporte para as suas línguas (um avaliador por língua) e classificaram as traduções de acordo com uma métrica binária: `OK` para as traduções corretas e `ERR` para as traduções erradas. Nas classificações marcadas como `ERR`, respeitantes a traduções semanticamente incorretas ou com problemas sintáticos dentro da construção, os linguistas identificaram erros de concordância (`AGREE`) e erros de outro tipo (`OTHER`) para distinguir erros relacionados com a morfologia da palavra ou outros problemas, tais como o uso incorreto de preposições, ordem de palavras incorreta dentro da construção, etc. Por último, os linguistas também apresentaram uma avaliação mais detalhada onde descreveram os problemas mais relevantes nas traduções que avaliaram de acordo com os diferentes tipos de construção.

Primeiros Resultados

O objetivo principal da avaliação realizada foi identificar a raiz dos problemas na tradução das construções com verbos-suporte tendo em conta cinco pares de línguas e indicar que direção é em que a avaliação qualitativa deve avançar para que estes desafios linguísticos à tradução de qualidade sejam vencidos. Fizemos isso, tendo em conta dois sistemas de natureza diferente (o OpenLogos e o Google Translate) para podermos verificar, em relação a este fenómeno linguístico em particular, até que ponto o fracasso de um sistema pode ser colmatado pelo sucesso do outro. Nesse sentido, verificámos que o desempenho de ambos os sistemas relativamente às construções com verbos-suporte foi globalmente mau por razões que se prendem à natureza intrínseca de cada um destes sistemas. Os problemas de tradução apresentados pelo Google Translate são, em geral, de natureza mais estrutural (cf. exemplo (viii)), enquanto que os problemas de tradução do sistema OpenLogos são de natureza mais lexical (cf. exemplo (i)). A avaliação humana sistemática das traduções das construções com verbos-suporte obtidas através destes sistemas mostrou que, à excepção do par de línguas inglês-alemão,

Construção com verbo-suporte nominal
make a presentation
Nominal não adjacente
have [ADV+ADJ-particularly good] links
Nominal preposicional
give an illustration of
Nominal preposicional não adjacente
be the [ADJ-immediate] cause of
Nominal idiomática
set in motion, place at risk, go on strike
Nominal preposicional idiomática
earn an income of
Nominal idiomática não adjacente
hold [NP-the option] in place
be of [ADJ-practical] value
Nominal preposicional idiomática não adjacente
give [PRO-us] a [bird's-eye] view of
be [ADV-clearly] at odds with
open talks [May 14] with
Construção com verbo-suporte adjectival
be meaningful
Adjetival não adjacente
be [ADV-extremely] selective
Adjetival preposicional
be known as; be involved in
Adjetival preposicional não adjacente
fall [ADV-so far] short of

TABELA 1: Categorias principais de construções com verbos-suporte no corpus

o Google Translate traduziu corretamente mais construções com verbos-suporte do que o OpenLogos, devido à larga dimensão da sua base de dados lexical.

Em relação ao par inglês-alemão, o OpenLogos traduziu corretamente 60 construções, enquanto que o Google Translate traduziu corretamente apenas 40. Os erros, tanto do OpenLogos como do Google Translate dizem respeito à escolha incorreta de palavras, à ordem incorreta das palavras dentro da construção, à escolha incorreta da forma da palavra (morfologia) e à falta de palavras. Os maiores problemas apresentados pelo Google Translate foram a falta de cobertura lexical em relação às construções adjacentes e a dificuldade em traduzir bem a separação do verbo.

No caso da tradução do inglês para as línguas românicas, o desempenho do Google Translate foi superior ao do OpenLogos. A maior parte dos erros de tradução dos dois sistemas correspondem a uma escolha lexical incorreta para alguns dos elementos da construção (por vezes, não existe tradução de algumas palavras, outras vezes, a tradução é literal), erro de concordância (entre o sujeito e o verbo, ou entre o sujeito e o adjetivo predicativo), e construções não adjacentes e idiomáticas. No caso das construções menos idiomáticas, há preposições erradas, tradução

literal do verbo-suporte e escolha lexical errada para o nome predicativo, preposições e determinantes. Estes problemas requerem um esforço pequeno de pós-edição, já que se tratam de palavras muito curtas. Os resultados quantitativos, os exemplos ilustrativos, e as avaliações qualitativas detalhadas para todos os pares de línguas podem ser consultados em Barreiro et al. (2014). Passaremos a apresentar com especial pormenor a descrição dos erros de tradução de construções com verbos-suporte do par inglês–português, apenas superficialmente referidos no trabalho anterior.

Análise Linguística dos Erros de Inglês-Português

Distribuímos os erros de tradução das construções com verbos-suporte do par inglês-português entre erros lexicais e erros estruturais. Os erros relacionados com a falta ou uso incorreto de palavras dentro da construção são caracterizados como erros de cobertura lexical, incluindo a escolha de verbo-suporte, de predicado não verbal, de preposição ou de qualquer outro elemento inserido. Estes erros não afetam a estrutura geral da frase. Por outro lado, os erros relacionados com a ordem incorreta das palavras, com a morfologia e com a concordância são caracterizados como erros estruturais. Os erros de ordem das palavras dizem respeito à inversão da posição das palavras dentro da construção. Os erros morfológicos dizem respeito a problemas relacionados com a forma incorreta das palavras, como o tempo verbal errado. Finalmente, os erros de concordância dizem respeito à falta de concordância entre os elementos do interior da construção com verbo-suporte ou entre um ou mais elementos dentro da construção e os elementos exteriores, tal como o sujeito da frase. Os erros estruturais ocorrem no interior da construção com verbo-suporte ou na relação entre esta e outros elementos da frase e afetam a sua gramaticalidade. Por exemplo, a falta de concordância entre o sujeito da frase e a construção com verbo-suporte é um erro que, embora esteja relacionado com a construção com verbo-suporte, ultrapassa as suas fronteiras.

A grande maioria dos erros de tradução, tanto por parte do Google Translate como por parte do OpenLogos, diz respeito à escolha lexical das palavras dentro da construção com verbo-suporte. Muitos dos erros dizem respeito a uma tradução direta de construções com verbos-suporte idiomáticas, que tornam o significado destas incompreensível. Ambos os sistemas falharam em *come to a rest*, *open talks*, *put in place*, *fall short of* e *have a spotty record*. O Google Translate apresentou erros nas traduções das construções *hold in place*, *be in charge of*, *be on guard*. O OpenLogos apresentou erros nas traduções das construções *come into the picture*, *place at risk*, *put under the microscope*, *be on strike*, *be at odds with*, *earn an income*. O exemplo (i) ilustra a tradução literal de *give a bird's-eye view of*.

(i) $_{EN}$ - It **gives us a bird's-eye view of** the economy.
$_{PT}$ - **Dá-nos uma *vista de olho de pássaro da** economia.

Em alguns casos, ambos os sistemas apresentaram erros na tradução do nome predicativo com consequências ao nível da tradução da preposição por este selecionada. Por exemplo, em (ii), o nome predicativo *insight* foi traduzido como *visão* em vez de *perspetiva* com um consequente erro no uso da preposição. A preposição *into* é selecionada pelo nome predicativo *insight* em inglês, mas o nome predicativo *perspetiva* em português seleciona a preposição *de* e não *para*.

(ii) $_{EN}$ - These specifications **gave insight into** the space of possible case-based systems, and elucidated human interaction properties.
$_{PT}$ - Estas especificações **deu uma *visão *para** o espaço de possíveis sistemas baseados em casos, e elucidou Propriedades interação humana.

Nos casos de construções menos idiomáticas, os erros afetam geralmente apenas um ou dois elementos da construção, como o verbo-suporte ou a preposição. Por exemplo, em (iii) o verbo-suporte *makes* foi traduzido literalmente por *faz* em vez de *torna*. Em (iv), a preposição *for* foi traduzida por *para* em vez de *por*. Em (v), a preposição *to* foi traduzida pela preposição *para* em vez de *a*.

(iii) $_{EN}$ - On the one hand, such a rich grammatical theory **makes it possible** to write grammars that contain very rich linguistic knowledge.
$_{PT}$ - Por um lado, uma teoria tal gramatical rica ***faz possível** escrever gramáticas que contêm o conhecimento linguístico muito rico.

(iv) $_{EN}$ - Schafer testified he believed his bureau chief in Beirut, Lester Coleman, **was responsible for** his photo appearing as part of the Pan Am affidavit.
$_{PT}$ - Schafer atestou que ele acreditou no seu chefe de escritório em Beirut, Lester Coleman, **foi responsável *para** sua fotografia que aparece enquanto a parte da panela é declaração.

(v) $_{EN}$ - The new Government which came to power in April 1984 has expressed a desire **to give priority to** agriculture development and to remove past obstacles.
$_{PT}$ - O governo novo que assumir poder em Abril 1984 exprimiu um desejo de **dar *a prioridade *para** o desenvolvimento de agricultura e de retirar-se por obstáculos.

O sistema Google Translate apresenta vários erros de concordância em construções que o sistema OpenLogos consegue traduzir corretamente. Esses erros podem ser entre o sujeito da frase e o verbo-suporte (vi), ou entre o sujeito da frase e o adjetivo predicativo da construção com verbo-suporte ((vii) e (viii)).

(vi) $_{EN}$ - the protests will **have no effect on** negotiations
$_{PT}$ - os protestos não ***terá nenhum efeito sobre** as negociações

(vii) $_{EN}$ - Descriptive economics and economic theory **are both concerned with** facts
$_{PT}$ - Economia descritiva ea teoria econômica ***são *tanto *preocupado com** os fatos

(viii) $_{EN}$ - **To be meaningful**, facts must be systematically arranged, interpreted, and generalized upon.
$_{PT}$ - **Para *ser *significativa**, os fatos devem ser sistematicamente organizados, interpretados e generalizada sobre.

Tarefas de Avaliação Futuras

As traduções produzidas por sistemas de tradução automática vastamente utilizados ainda mostram erros lamentáveis que requerem um esforço significativo de pós-edição. As construções com verbos-suporte, entre outras unidades lexicais multipalavra, são responsáveis por muitos desses erros de tradução. As atuais métricas de avaliação da qualidade, concentradas na medição do tempo e esforço de pós-edição, não contemplam este e outro tipo de unidades linguísticas, mostrando-se ineficazes e insuficientes para avaliar a verdadeira qualidade dos sistemas e incapazes de identificar problemas que possam ajudar a melhorar a estrutura sintática e o significado na tradução. O trabalho de avaliação de sistemas de tradução automática apenas deu os seus primeiros passos. Há um trabalho ainda muito grande a fazer para colmatar as deficiências na avaliação qualitativa atual. Não existem publicações sobre uma avaliação linguística conjunta que tenha como objetivo comparar os pontos fortes e fracos de diferentes abordagens de tradução automática, com o objetivo de melhorar a qualidade da tradução. Os investigadores precisam de desenvolver métricas para a avaliação periódica sistemática da qualidade linguística da tradução automática, independentemente da natureza de cada sistema. A avaliação deve incluir tarefas de categorização de erros onde fenómenos linguísticos específicos possam ser avaliados individualmente por linguistas especializados em tradução automática. Esta avaliação deve ser elaborada por fases, em que cada fase corresponda à avaliação de um fenómeno linguístico particular (por exemplo, para as unidades lexicais multipalavra, avaliar individualmente as construções com verbos-suporte, as unidades compostas, os *phrasal verbs* do inglês, etc.). A categorização de erros em unidades menores do que a frase pode contribuir para tarefas de avaliação mais controladas e sistemáticas. A avaliação tem de ser dirigida a grupos de erros linguísticos e identificar que sistemas têm mais dificuldades em traduzir cada tipo de desafio linguístico (avaliação paradigmática). Para além disso, devem ser construídos corpora específicos ou coletâneas de frases que serão usadas para avaliar construções relativas, passivas, pronomes, determinantes, preposições locativas, etc. Essas métricas de avaliação qualitativa deverão ser desenvolvidas e validadas por especialistas linguísticos que trabalham na área da tradução automática. Estamos convencidos de que um método eficaz para o avanço da investigação em tradução automática é comparar os resultados de diferentes abordagens e medir que módulos requerem melhoramento. Uma hibridização eficaz só terá lugar quando o desempenho de sistemas com abordagens diferentes for linguisticamente testado. Acreditamos que tal avaliação qualitativa conjunta possa ser valorizada pela comunidade científica.

[4] CONCLUSÃO E TRABALHO FUTURO

Estudos realizados anteriormente revelam lacunas importantes ao nível da anotação, identificação, representação, reconhecimento, processamento e avaliação das construções com verbos-suporte. Os atuais sistemas de tradução automática não conseguem traduzir com qualidade os fenómenos linguísticos apresentados pelas construções com verbos-suporte. Uma tarefa importante que pode conduzir a uma melhor tradução das construções com verbos-suporte é a do seu parafraseamento. Um sistema que permita mapear construções com verbos-suporte com os seus equivalentes semânticos, sejam eles variantes estilísticas, variantes parafrásticas ou verbos, constitui uma mais valia para a tradução (humana e automática). Entre outros aspetos positivos, o parafraseamento tem a vantagem de servir como ferramenta de auxílio na transformação estilística de textos, permitindo a conversão de um texto "palavroso" num texto semanticamente equivalente, mas utilizando um menor número de palavras e uma linguagem mais controlada, e por conseguinte, mais fácil de traduzir por uma máquina.

Outra tarefa de grande relevo para o aperfeiçoamento dos sistemas de tradução é a da avaliação da tradução das construções com verbos-suporte. Os erros refletidos na tradução destas construções por dois importantes sistemas de tradução automática, o OpenLogos e o Google Translate, permitem concluir que as unidades lexicais multipalavra continuam a ser um problema em aberto na área da tradução automática, independentemente do tipo de aproximação adotada pelo sistema. Os erros encontrados no interior das construções traduzidas poderiam ser minimizados se as unidades lexicais multipalavra fossem tratadas como unidades indissociáveis. A falta de composicionalidade das unidades lexicais multipalavra, nomeadamente a das construções com verbos-suporte, fica também comprometida com a falta de intervenção humana qualificada na tarefa de alinhamento de segmentos bilingues ou multilingues usados para treinar sistemas de aprendizagem automática. Apesar da grande pertinência da qualidade dos alinhamentos dos vários elementos da frase nos sistemas estatísticos, este tema está ainda pouco explorado do ponto de vista linguístico e computacional, motivo pelo qual optámos por não o incluir neste artigo. No entanto, não podemos deixar de referir que a impossibilidade de os sistemas de tradução automática estatísticos permitirem alinhar unidades lexicais multipalavra cujos elementos que as compõem se encontrarem em situações de não adjacência, constitui uma das razões do fracasso dos sistemas de tradução automática. Também nesta tarefa, o envolvimento de fator humano especializado ou a especializar-se em tradução será determinante para o processo de aprendizagem automática de conhecimento linguístico que conduzirá à qualidade da tradução destas expressões, tema que merece ser abordado com a devida atenção em trabalho futuro.

AGRADECIMENTOS

Agradeço a Diana Santos e a Stella Tagnin os comentários pertinentes, que permitiram melhorar este artigo. Este trabalho foi parcialmente financiado pela FCT através de uma bolsa de pós-doutoramento (SFRH/BPD/91446/2012).

REFERÊNCIAS

Aziz, Wilker, Sheila Castilho Monteiro de Sousa & Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. Em *Eighth International Conference on Language Resources and Evaluation (LREC2012)*, 3982–3987.

Baptista, Jorge. 2005. *Sintaxe dos nomes predicativos com verbo-suporte SER DE*. Fundação para a Ciência e a Tecnologia/Fundação Calouste Gulbenkian.

Barreiro, Anabela. 2008. ParaMT: A paraphraser for Machine Translation. Em *Computational Processing of the Portuguese Language, 8th International Conference, (PROPOR 2008)*, 202–211.

Barreiro, Anabela. 2009. *Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation*: Universidade do Porto. Tese de Doutoramento.

Barreiro, Anabela. 2011. SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing. Em Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 6609 Lecture Notes in Computer Science, 365–376. Springer.

Barreiro, Anabela & Luís Miguel Cabral. 2009. ReEscreve: a translator-friendly multi-purpose paraphrasing software tool. Em Marie-Josée Goulet, Christiane Melançon, Alain Désilets & Elliott Macklovitch (eds.), *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*, 1–8.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac & Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation. Em *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*, 26–33.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista & Isabel Trancoso. 2014. Linguistic Evaluation of Support Verb Constructions by OpenLogos and Google Translate. Em Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 35–40. ELRA.

Barreiro, Anabela, Bernard Scott, Walter Kasper & Bernd Kiefer. 2011. OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. *Machine Translation* 25(2). 107–126.

Bick, Eckhard & Anabela Barreiro. 2015. Automatic anonymisation of a new Portuguese-English parallel corpus in the legal-financial domain. Neste volume.

Chacoto, Lucília. 2005. *O Verbo Fazer em Construções Nominais Predicativas*: Universidade do Algarve. Tese de Doutoramento.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05*, 263–270. Association for Computational Linguistics.

Gross, Maurice. 1984. Lexicon-grammar and the syntactic analysis of French. Em *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, Proceedings of COLING* , 275–282.

Gross, Maurice & Jean Senellart. 1998. Nouvelles bases pour une approche statistique. Em *Actes du colloque international JADT-98*, .

Koehn, Philipp. 2005. EuroParl: A Parallel Corpus for Statistical Machine Translation. Em *Conference Proceedings: the tenth Machine Translation Summit*, 79–86. AAMT.

Maia, Belinda. 2005. Machine Translation and Human Translation: using machine translation engines and parallel corpora for teaching and research. Em *International Contrastive Linguistics Conference*, 123–145.

Maia, Belinda & Anabela Barreiro. 2007. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português. Em Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, 205–216. IST Press.

Maia, Belinda, Anabela Barreiro & Luís Sarmento. 2003. EVAL - Evaluation of Machine Translation at FLUP. *Apresentação em AvalON'2003*. http://www.linguateca.pt/documentos/MaiaBarreiroSarmentoEVALAvalon2003.pdf.

Maia, Belinda, Diana Santos, Luís Sarmento & Anabela Barreiro. 2004. TrAva - a tool for evaluating Machine Translation - pedagogical and research possibilities. Apresentação na *ABRAPT*. http://web.letras.up.pt/bhsmaia/belinda/pres/abrapt-trava.ppt.

Marcu, Daniel, Wei Wang, Abdessamad Echihabi & Kevin Knight. 2006. SPMT: statistical machine translation with syntactified target language phrases. Em *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06*, 44–52. Association for Computational Linguistics.

Ramisch, Carlos, Aline Villavicencio & Christian Boitet. 2010. Multiword Expressions in the wild? The mwetoolkit comes in handy. Em *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 57–60.

Ranchhod, Elisabete. 1983. On the Support Verbs *Ser* and *Estar* in Portuguese. *LingvisticæInvestigationes* Volume 7. 317 –353.

Ranchhod, Elisabete. 1990. *Sintaxe dos Predicados Nominais com Estar*. Instituto Nacional de Investigação Científica.

Salkoff, M. 1999. *A French-English Grammar: A Contrastive Grammar on Translational Principles* Linguisticae investigationes. J. Benjamins.

Sarmento, Luís, Anabela Barreiro, Belinda Maia & Diana Santos. 2007. Avaliação de Tradução Automática: alguns conceitos e reflexões. Em Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, 181–190. IST Press.

Scott, Bernard (Bud). 2003. The Logos Model: An Historical Perspective. *Machine Translation* 18(1). 1–72.

Vieira, Lucas & Lucia Specia. 2011. A review of translation tools from a post-editing perspective. Em *3rd joint EM+/CNGL Workshop bringing MT to the user: Research meets translators (JEC)*, 33–42.

Zollmann, Andreas & Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. Em *Proceedings of the Workshop on Statistical Machine Translation, StatMT '06*, 138–141. Association for Computational Linguistics.

CONTACTOS

Anabela Barreiro
INESC-ID
anabela.barreiro@inesc-id.pt