Avaliação de tradução automática: alguns conceitos e reflexões

Luís Sarmento, Anabela Barreiro, Belinda Maia e Diana Santos

1. Avaliação de tradução

A avaliação de tradução nunca foi um assunto pacífico. A tradução humana pode ser avaliada de vários pontos de vista, desde a correcção de elementos gramaticais e lexicais até à reformulação e adaptação do texto original para servir uma função diferente na cultura de chegada. É notória a falta de compreensão sobre a natureza da tradução por parte do público em geral. Por um lado, todos se acham no direito de se pronunciar sobre a qualidade de uma tradução; por outro, poucos têm ideia das dificuldades envolvidas quando é necessário efectuar com rigor e precisão uma análise linguística completa da integridade da tradução. A publicidade que se tem feito à tradução automática contribui bastante para a noção de que a tradução é uma actividade tão fácil que até uma máquina é capaz de a fazer.

Na realidade, a tradução é uma actividade muito complexa, que envolve a compreensão do texto da língua de origem a vários níveis. Em primeiro lugar é necessário considerar vários parâmetros lexicais, dado que o léxico é a área em que se encontram os exemplos mais flagrantes das diferenças entre as línguas, assim como ter em conta parâmetros sintácticos que envolvem estruturas e elementos semânticos nem sempre transferíveis de uma língua para a outra. É também fundamental entrar em linha de conta com questões relativas à construção do discurso, que segue convenções diferentes em cada língua, e questões relativas ao contexto concreto da produção do texto, o que exige conhecimentos específicos de pragmática. Assim, o processo de tradução envolve não só conhecimentos de sintaxe, semântica e léxico, mas também familiaridade com o assunto tratado, com o registo e género do texto, e até com experiências individuais diferentes do (mesmo?) mundo real.

O processo cognitivo humano é bem mais complexo do que o que a inteligência artificial actual consegue reproduzir, e a tradução humana processa-se de uma maneira essencialmente diferente da tradução automática. Isto ajuda-nos a entender a razão pela qual existe tanta variação nas traduções humanas. De facto, o mesmo texto terá tantas traduções como tradutores, não só no ambiente do ensino de tradução como também no ramo profissional.

É importante reconhecer que há diferentes razões e processos para avaliar a tradução, seja ela humana ou automática. Têm sido realizados vários tipos de experiências de avaliação de tradução automática (TA) que incluem o uso exclusivo de métodos de avaliação manuais (White & O'Connell, 1994), a comparação entre métodos de avaliação manual e métodos automáticos (Popescu-Belis, 2003), bem como estudos exclusivamente assentes em métodos automáticos de avaliação da TA tais como os métodos RED e BLEU entre outros (Akiba et al., 2001, Rajman & Hartley, 2002, Papineni et al., 2002). As experiências demonstram bastante variação nos resultados obtidos e nas opiniões expressas. Hovy et al. (2002: 52) consideram a diferença de métodos um factor essencial e positivo: "Em que parâmetros atentar, e que peso atribuir a cada um deles, compete ao avaliador decidir"[1].

Ao nível da tradução humana, embora haja ainda professores que façam da tradução muito literal um exercício de gramática, e clientes de tradução que achem que a tradução literal é a mais correcta, a experiência mostra-nos que os textos são construídos de uma maneira diferente em cada língua (Maia, 1997) e que o que é expresso nas traduções humanas difere significativamente do original (Santos, 2004). A teoria mais recente defendida pelos professores de tradução exige a formação de tradutores profissionais que valorizem facetas como a função do original e a função (por vezes diferente) da tradução. A avaliação da tradução humana é uma área de pesquisa onde poucos se aventuram por razões teóricas e práticas; destacamos, no entanto, os trabalhos de House (1977, 1997), que utiliza um método baseado na análise multi-

[1] "Which parameters to pay attention to, and how much weight to assign to each one, remains the prerogative of the evaluator"

nível de Halliday (1984). Para além da influência de House, há também quem trabalhe nesta área recorrendo a corpora construídos a partir de múltiplas traduções humanas dos mesmos textos (Popescu-Belis *et al.*, 2002).

Quanto à tradução automática, esta é talvez a área onde ao longo dos anos se despendeu mais esforço de avaliação, e de sistematização dessa avaliação. A variedade de avaliações em TA é significativa, desde o famoso relatório ALPAC (1966) a grandes projectos de avaliação conjunta, como os financiados pelo programa ARPA MT (White & O'Connell, 1994) e os presentes, levados a cabo pelo NIST com o apoio do LDC (projecto TIDES) desde 2001 (NIST, 2004; LDC, 2002). Sem tentar ser exaustivos, convém no entanto mencionar que no âmbito do projecto europeu EAGLES (1995) e do projecto japonês JEIDA (Nomura & Isahara, 1992) foram realizados vários trabalhos focando explicitamente a TA.

Na primeira metade dos anos 90 apareceram uma série de trabalhos de avaliação baseados no desenvolvimento de materiais de teste (*test suites*, Oepen *et al.*, 1988), entre os quais salientamos King & Falkedal (1990) e Balkan *et al.* (1994). Ainda no âmbito da avaliação, a estrutura de trabalho denominada FEMTI, desenvolvida no âmbito do projecto ISLE, é uma tentativa de organizar os vários métodos que são utilizados para avaliar os sistemas de TA e para os relacionar com a finalidade e o contexto dos sistemas. Em Hovy *et al.* (2002), os autores descrevem os princípios em que o FEMTI assenta, apresentando uma visão geral de trabalhos anteriores e oferecendo algumas perspectivas futuras. Várias propostas interessantes, nomeadamente as que dizem respeito à automatização da avaliação em TA, podem ser encontradas, entre outras, em Papineni *et al.* (2001), NIST (2002) e Popescu-Belis *et al.* (2002).

Não será, aliás, deslocado dizer que a tradução automática é a área de PLN em que as avaliações comparativas são mais frequentes, desde as avaliações conjuntas anteriormente mencionadas, a outras que recorrem a juízes humanos independentes, como em Richardson *et al.* (2001). Também a tradução automática foi uma das áreas de eleição pelos investigadores em processamento computacional do português, quando a Linguateca iniciou as actividades de divulgação e dinamização da avaliação conjunta que levaram à criação do grupo ARTUR. É nesse contexto que o trabalho do pólo do Porto surgiu e, como pretendemos demonstrar neste e nos capítulos seguintes (veja-se também Santos *et al.*, 2004), foi inovador em dois sentidos: por um lado, preocupou-se em compilar materiais, independentemente de um dado sistema, mas fulcralmente dependentes do par de línguas em causa (inglês-português); por outro lado, tentou fazer isso de uma forma cooperativa, inspirado e motivado pela iniciativa de avaliação conjunta da Linguateca.

No presente capítulo tentamos apresentar o muito material que, embora não directamente associado ao modelo da avaliação conjunta, é essencial para compreender a complexidade da tradução automática e a história da sua avaliação. Ainda que não seja possível esgotar o assunto, parece-nos essencial documentar que a abordagem da Linguateca não nasceu num vazio. Pelo contrário, tentou complementar um panorama vasto e complicado, do qual estávamos conscientes. Os próximos dois capítulos deste livro, (Sarmento e Maia & Barreiro, neste volume), surgem no encadeamento do presente e pretendem relatar o trabalho original que foi realizado na Linguateca no contexto da avaliação de TA, exemplificando muito do que aqui se irá abordar. Sarmento (neste volume) descreve as ferramentas desenvolvidas para observação do desempenho de sistemas de TA em paralelo e para a categorização dos erros e problemas de tradução encontrados (nomeadamente os siistemas METRA, Boomerangue e TrAva). Maia & Barreiro (neste volume), por seu turno, relata a experiência de recolha e classificação de exemplos de TA efectuada por um grupo de alunos de mestrado da Faculdade de Letras de Universidade do Porto utilizando o TrAva, apresentando alguns casos interessantes de problemas e de sucessos na TA.

2. Conceitos chave sobre avaliação de tradução automática

A área da avaliação de TA foi, desde há muito, alvo de um esforço de organização e formalização. O FEMTI sugere duas taxonomias que descrevem sistematicamente (i) os vários tipos de utilizadores de TA e as respectivas necessidades, e (ii) as várias características dos sistemas de TA (internas e externas) sobre as quais se poderá realizar uma actividade de

avaliação. Ambas as taxonomias desenvolvidas apresentam um grau de detalhe muito elevado, representando um importante marco na sistematização da área.

Contudo, numa primeira aproximação às actividades de tradução automática revela-se conveniente lidar com os vários conceitos associados de uma forma mais simplificada. Por essa razão, faremos de seguida uma breve descrição e análise de algumas metodologias possíveis para avaliação de TA, abordando algumas questões práticas relacionadas com estas metodologias. Esta análise, apesar de assumidamente incompleta e simplista, serve para enquadrar os nossos esforços de arranque das actividades de avaliação de TA, não pretendendo ser exaustiva nem lidar com todas as possibilidades envolvidas, mas documentar algumas questões com que tivemos de lidar.

2.1 Avaliação interna e avaliação externa

Uma questão essencial relacionada com a avaliação em geral, mas que assume contornos particulares no caso da tradução, prende-se com as diferentes formas de acesso e os possíveis pontos de observação do sistema em avaliação. A este nível podem distinguir-se dois tipos de metodologias de avaliação, *interna* e *externa*, dependendo das condições de acesso que o avaliador tem ao interior do sistema de tradução automática.

A *avaliação interna* (também conhecida como avaliação *glassbox* ou *caixa de vidro*) é realizada sobre os componentes internos do sistema de TA e, normalmente, é apenas passível de ser praticada pelos próprios grupos de desenvolvimento que implementam os sistemas, já que habitualmente são eles os únicos que podem aceder livremente ao interior dos ditos. O principal objectivo da avaliação interna é o de testar exaustivamente o desempenho de uma determinada etapa ou módulo específico do sistema de TA (léxico, regras sintáctico-semânticas, vários níveis de análise sintáctica, entre outros), para obtenção de indicações concretas relativamente ao seu nível de desenvolvimento. Por esta razão, neste tipo de avaliações são utilizados critérios muito concretos de medição de qualidade, tais como graus de cobertura e percentagens de acerto, já que o fenómeno a avaliar está perfeitamente localizado e pode ser quantificado com facilidade.

Outra perspectiva de avaliação consiste na *avaliação externa* (também conhecida como avaliação *blackbox* ou *caixa negra*), que é realizada tendo em conta exclusivamente a entrada e a saída do sistema, abstraindo totalmente dos seus detalhes de implementação. Este nível de abstracção sobre a tecnologia permite a realização de avaliações comparativas entre diferentes sistemas de TA, mesmo que estes utilizem tecnologias totalmente distintas. A avaliação comparativa pode ser feita através da introdução nos vários sistemas do mesmo texto para tradução e comparação das respectivas saídas.

A avaliação externa apresenta como principal dificuldade a necessidade de definir critérios de avaliação da qualidade sobre a própria tradução. Dependendo da definição dos parâmetros de comparação, o resultado deste tipo de avaliação pode fornecer pistas indirectas para o desenvolvimento de sistemas de TA, mas a sua principal utilidade consiste na obtenção de informação comparativa do ponto de vista do utilizador final. A avaliação externa pode, por exemplo, comparar usabilidade, factores humanos, funcionalidade, velocidade e qualidade linguística dos vários sistemas existentes.

Poderá ainda ser tida em consideração uma terceira metodologia de avaliação, que combina aspectos da avaliação interna e externa. É possível considerar que esta metodologia híbrida seja empregue em situações em que um determinado utilizador possui formas de alterar alguns dos parâmetros ou recursos associados ao sistema de TA, tais como dicionários e regras, e pretende avaliar o resultado dessas alterações no produto final da tradução. Cada sistema tem uma arquitectura própria que pode estar relacionada com os diferentes tipos de abordagem (modelos baseados em regras linguísticas, tradução baseada em exemplos ou tradução baseada em modelos estatísticos). A maior parte dos sistemas não permite a intervenção do utilizador no processo de TA, mas, em alguns sistemas comerciais, apesar de não ter possibilidade de observar ou influenciar directamente as etapas intermédias do sistema, o utilizador pode alterar, ainda que superficialmente, alguns dos parâmetros do seu processamento. Certos sistemas são além disso dotados de ferramentas interactivas que proporcionam uma maior intervenção do utilizador no processo de tradução. Estas ferramentas podem permitir seleccionar domínios específicos para efeitos de selecção de vocabulário mais apropriado a um tipo de tradução em

particular, associar informação sintáctica e semântica às entradas do dicionário, criar regras que protegem e/ou traduzem termos seleccionados, etc… Como o utilizador poderá apenas observar o impacto das alterações efectuadas sobre o produto final da tradução, podemos considerar-nos em presença de uma combinação de critérios de avaliação interna e de avaliação externa.

2.2 Avaliação manual e automática

Outra distinção possível de estabelecer é a entre metodologias de avaliação automática e de avaliação manual. De uma forma muito resumida, poderemos dizer que os métodos de avaliação manuais recorrem a avaliadores humanos para julgamento e identificação dos casos relevantes nos testes efectuados, enquanto que os métodos automáticos usam algoritmos para a execução dessas mesmas tarefas.

Embora seja possível aplicar a distinção de métodos automáticos e manuais sobre ambas as categorias definidas anteriormente (metodologias de avaliação externa e interna), na prática faz mais sentido aplicar esta distinção apenas às metodologias de avaliação externa. De facto, por razões de aplicabilidade já parcialmente explicadas, tem-se investido mais no estudo e desenvolvimento de novas metodologias de avaliação externa, quer manuais quer automáticas, e até ao final desta secção concentrar-nos-emos apenas nestas.

2.2.1 Avaliação manual externa

Tradicionalmente, as metodologias de avaliação manual têm sido as mais utilizadas. O funcionamento habitual destes métodos envolve um grupo de avaliadores humanos que executa julgamentos sobre vários pares texto-tradução em função de critérios e escalas pré-definidas.

A avaliação manual possui algumas particularidades muito interessantes, pois é, potencialmente, muito flexível. Outra vantagem destas metodologias, que advém do próprio julgamento humano, é a capacidade de avaliar a tradução em função de um determinado contexto, o que pode ser muito útil para testar o desempenho do sistema no tratamento de casos ambíguos. No entanto, as metodologias de avaliação manual apresentam algumas dificuldades, das quais destacamos:

1. **Custo e dificuldade na execução.** Recorrer a colaboradores para uma tarefa de avaliação deste género não é fácil, porque nem sempre se encontram disponíveis pessoas com os conhecimentos e a experiência necessários. Além disso, um teste em larga escala envolve custos que poderão ser muito altos em termos de recursos humanos e logísticos.
2. **Definição de sistemas de categorização da qualidade e construção do cenário de avaliação.** A definição de um sistema que permita categorizar as traduções em termos de qualidade é uma tarefa intrinsecamente difícil. Se se recorrer a critérios linguísticos, é necessário criar um sistema de categorias que seja sistemático, rigoroso e abrangente, mas ao mesmo tempo suficientemente simples para os avaliadores menos familiarizados com aspectos gramaticais complexos. Este compromisso entre o rigor e a cobertura linguística, por um lado, e a usabilidade, por outro, é extremamente difícil de alcançar (Niessen *et al.*, 2000). Para além disso, é necessário também relacionar os critérios linguísticos com parâmetros subjectivos de qualidade, para que os resultados da avaliação possam depois ser correlacionáveis com a experiência dos utilizadores. Se, em alternativa, se recorrer a sistemas e métricas de qualidade baseadas em tarefas práticas (por exemplo, ser capaz de executar uma tarefa descrita num manual de instruções traduzido automaticamente), torna-se necessário correlacionar correctamente os resultados obtidos nessas tarefas com a qualidade da tradução, dado que o sucesso da tarefa depende em grande parte do utilizador. O cenário para uma avaliação desse género terá de ser montado em função, quer do sistema de categorização, quer dos utilizadores, o que é uma tarefa extremamente difícil.
3. **Dificuldade em controlar e minimizar a subjectividade do avaliador.** Mesmo quando instruídos com indicações muito claras de avaliação, é inevitável que os julgamentos sejam influenciados por parâmetros subjectivos (alguns totalmente exteriores ao ambiente de avaliação) que acabam por introduzir ruído indesejável na avaliação.
4. **Possibilidade de flutuações de critérios ao longo do ensaio**. Os critérios dos avaliadores sofrem normalmente flutuações ao longo do ensaio, fenómeno esse chamado *efeito de*

treino. Por exemplo, depois de serem confrontados com algumas traduções de reduzida qualidade, os utilizadores têm tendência a baixar as suas expectativas de qualidade, fazendo com que a avaliação das traduções seguintes seja mais favorável do que se o utilizador não tivesse visto as más traduções anteriormente.

5. **Dificuldade em confirmar resultados das avaliações.** É muito provável que a repetição dos testes efectuados conduza a resultados diferentes dos obtidos anteriormente, já que as condições dos avaliadores se alteram permanentemente. Além de poder ser inviável reunir o mesmo grupo de avaliadores para uma nova tarefa de avaliação, não faz sentido repetir a avaliação dos mesmos textos pelas mesmas pessoas. Assim, a confirmação de resultados obtidos por avaliação humana é muito mais difícil do que a daqueles obtidos por avaliação automática.

2.2.2 Avaliação automática externa

Apesar das suas virtudes, a avaliação manual é um processo que é demasiado lento e dispendioso para ser realizado com frequência, o que reduz significativamente a capacidade das equipas de desenvolvimento de testarem frequentemente as alterações que vão introduzindo nos seus sistemas. A lentidão e o alto custo impedem também que os utilizadores finais realizem facilmente testes sobre o desempenho dos sistemas que pretendam adquirir. Por este motivo, têm vindo a ser desenvolvidos métodos automáticos de avaliação de TA que permitem a realização de testes de uma forma rápida e pouco dispendiosa (veja-se, por exemplo, Rajman & Hartley, 2002).

Um sistema de avaliação automático é composto por dois elementos básicos: um *programa de análise*, que analisa o texto original e a respectiva tradução (normalmente ao nível da frase), e um *sistema de métricas de qualidade*, que sustenta o cálculo do resultado da avaliação.

O programa de análise é responsável por identificar a ocorrência de certos padrões de erro previamente estudados, ou por efectuar uma comparação entre a tradução em teste e um conjunto de traduções de referência. No primeiro caso, as ocorrências (ou ausências) de determinados padrões problemáticos são quantificadas, sendo estes valores usados depois no julgamento da qualidade da tradução. No segundo caso, serão os valores resultantes da comparação entre a tradução em teste e uma ou várias traduções de referência que irão servir de base ao julgamento de qualidade. Essa comparação passa normalmente pelo cálculo de distâncias entre os segmentos em teste e as referidas traduções de referência.

O sistema de métricas de qualidade, por sua vez, executa o mapeamento dos resultados obtidos pelo programa de análise (número de ocorrências, distância para as traduções de referência) para valores que exprimam a qualidade da tradução. É fundamental que este mapeamento tenha a propriedade de gerar resultados que sejam (directamente) correlacionáveis com as avaliações produzidas sobre as mesmas traduções por avaliadores humanos.

As principais vantagens dos métodos automáticos estão relacionadas com a velocidade de teste e com o reduzido custo da sua utilização. Depois de desenvolvidos, os métodos automáticos podem ser aplicados em novos testes sem que para isso se incorra em custos significativos. Os testes terão um tempo de realização muito curto (eventualmente na ordem de alguns minutos) o que permite a sua repetição frequente.

Uma outra vantagem dos métodos automáticos prende-se com a sua imunidade à flutuação de critérios (por efeito de treino ou por influências externas), uma vez que são totalmente mecanizados. Ao contrário dos métodos manuais, não há qualquer influência de factores subjectivos durante a avaliação, nem alteração de critérios entre avaliações sucessivas, o que possibilita a repetição dos testes com variações mínimas para efeitos de confirmação de resultados.

Apesar destas vantagens interessantes, os métodos automáticos apresentam bastantes dificuldades de desenvolvimento e de utilização. Por um lado, são normalmente complexos de desenvolver, já que envolvem:

- o estudo sobre o conjunto de padrões que é necessário detectar e quantificar ou, em alternativa, a definição de formas de cálculo da distância entre a tradução a avaliar e as

traduções de referência. Ora, encontrar e classificar os padrões mais apropriados para descrever a qualidade da tradução, ou desenvolver as formas de cálculo de distâncias necessárias para a comparação é, na maior parte das vezes, um processo difícil, de natureza empírica e que, por sua vez, envolve uma análise humana dispendiosa que só pode ser feita por peritos.

- a criação de um sistema de métricas de qualidade que permita mapear coerentemente os resultados obtidos pelo programa de análise num julgamento de qualidade das traduções, que por sua vez seja consistente com as avaliações produzidas por seres humanos. O desenvolvimento destas métricas segue, muitas vezes, um processo de tentativa e erro, visto que é necessário ajustar um conjunto de dados obtidos heuristicamente a um conjunto de julgamentos de qualidade holísticos.

Outra limitação dos métodos automáticos é a de que estes são normalmente muito pouco portáveis para outros pares de línguas para além daquele para o qual estes métodos tenham sido originalmente desenvolvidos. Para cada língua, novos estudos terão de ser realizados relativamente aos padrões a detectar, às medidas de distância a usar e às métricas de qualidade empregues.

Além disso, no caso da avaliação ser feita com base na comparação entre o produto da TA e traduções de referência, surgem questões relacionadas com a obtenção destas mesmas traduções de referência. A produção de várias traduções de referência para cada um dos textos originais a ser testado é um processo moroso e difícil de realizar, podendo envolver custos significativos com a contratação de tradutores para o efeito.

Em suma, os métodos manuais e os métodos automáticos apresentam características quase complementares. Infelizmente, não parece trivial combinar as duas perspectivas em simultâneo extraindo as vantagens de ambos e reduzindo as suas limitações ao mínimo possível.

3. Algumas questões sobre a qualidade da TA

Pelo facto de ser exclusivamente direccionada para o produto final do processo de tradução, a avaliação de tradução automática segundo metodologias de avaliação externas envolve sempre a noção de qualidade de tradução. Mas o que é, de facto, uma boa tradução, humana ou automática? Esta é uma questão demasiado abrangente, pelo que iremos tentar abordá-la na perspectiva exclusiva da tradução automática, sobre a qual é possível impor algumas restrições.

A primeira restrição que colocaremos, e que advém da compreensão das próprias limitações da TA, será relativa ao tipo de texto cuja tradução iremos avaliar. Não se pode esperar avaliar o desempenho de um sistema de tradução automática baseado na tradução de um tipo de texto para o qual o sistema em causa não está preparado, ou para o qual nem sequer foi idealizado. Isso exclui, à partida, a utilização nos testes de textos com um elevado nível de subjectividade (texto literário, linguagem corrente, com ocorrência de coloquialismos e expressões idiomáticas), concentrando os esforços de avaliação sobre textos mais "objectivos" e, de preferência, com um nível de ambiguidade semântica reduzido, como é o caso dos textos técnicos, científicos e especializados. Aqui incluímos documentação técnica, manuais de instruções, documentos provenientes de sítios web informativos (por exemplo, portais temáticos), regulamentos, etc…[2] Nestas circunstâncias, e tendo em consideração o nível de tecnologia de que dispomos actualmente, os problemas de tradução que colocamos aos sistemas de TA tornam-se razoáveis, pelo que faz sentido avaliar o seu desempenho e discutir as noções de qualidade.

Iremos agora tentar formular algumas noções de qualidade de tradução automática e analisar o seu impacto nas actividades de avaliação. As noções de qualidade que formularemos evidenciam algumas das diferentes formas disponíveis para aferir a qualidade de uma tradução. Contudo, não são exclusivas e poderão ser combinadas na realização de uma dada actividade de

[2] Note-se que, quando se lida com domínios específicos, é necessário dotar o sistema de TA de conhecimento específico do domínio (por exemplo, glossários ou bases de dados terminológicas) e/ou do género de texto (expressões "feitas" e típicas da área, como no caso da linguagem legal utilizada nos textos da Comissão Europeia), para que o sistema consiga ter um desempenho compatível com as suas potencialidades.

avaliação de TA em concreto. Para uma cobertura mais extensiva das noções de qualidade de tradução e a sua aplicação em testes em concreto, veja-se Elliott (2002).

3.1 Qualidade em função do objectivo

Pode estabelecer-se uma noção simples de qualidade de TA assumindo que uma tradução será boa se permitir ao utilizador alcançar o objectivo prático que o levou a recorrer à TA. Por exemplo, se o objectivo do utilizador for o de encontrar a definição dum determinado termo numa página web escrita numa língua que lhe é inacessível, podemos considerar que a tradução terá qualidade se permitir ao utilizador encontrar e compreender a definição em causa.

Esta noção de qualidade poderá, no entanto, ser em grande parte dos casos insatisfatória, porque apresenta muitas limitações. Em primeiro lugar, não existe uma correlação simples entre a qualidade da tradução e o sucesso do utilizador, já que existe uma enorme dependência entre esse mesmo sucesso e as capacidades do próprio utilizador. Em última análise, um sistema que se limitasse a traduzir todo o texto palavra a palavra poderia, ainda assim, permitir ao utilizador alcançar o seu objectivo, sem que isso justificasse uma classificação positiva do sistema para outros propósitos. Em segundo lugar, e mesmo que desprezássemos a influência das capacidades do utilizador no sucesso da sua tarefa, não seria simples distinguir a qualidade de vários sistemas recorrendo a avaliações deste género, já que a avaliação do sistema é essencialmente qualitativa e depende apenas da obtenção ou não de sucesso na concretização de uma tarefa. É claro que poderiam ser efectuadas várias rondas de testes e depois comparar os resultados acumulados de cada um dos sistemas, ou introduzir mais parâmetros de avaliação do sucesso da tarefa (tal como o tempo de execução), mas seria sempre utilizado um processo indirecto e propenso à introdução de distorções.

3.2 Qualidade relativa à tradução humana

Uma outra noção de qualidade de TA resulta da obtida por comparação com a tradução realizada por um tradutor humano e que lhe servirá como padrão. Ou seja, uma tradução automática poderá ser considerada tanto melhor, quanto mais se aproximar da tradução que um ser humano executaria. Seguindo esta noção de qualidade, a avaliação de um sistema de TA passaria por comparar a tradução automática de um determinado texto com traduções humanas do mesmo texto. Se, do ponto de vista teórico, esta forma de avaliação parece ser capaz de levar a resultados correctos e inteligíveis, do ponto de vista prático é extremamente difícil de realizar, já que envolve o conceito de distância entre dois excertos de texto (por hipótese, frases) cujo cálculo não é de todo trivial. Além do mais, é muito raro o caso em que uma frase possua apenas uma tradução possível, pelo que existe um espaço de traduções possíveis composto por N frases (eventualmente nem todas conhecidas) perante o qual o conceito de distância ainda se torna mais complexo de tratar.

Em actividades de avaliação de TA, têm sido experimentadas várias métricas para a medição da distância entre frases, em particular em avaliações que usam métodos automáticos, como é o caso do padrão BLEU (Papinemi *et al.*, 2001) ou do trabalho de Callison-Burch & Flournoy (2001).

Além destes problemas de ordem conceptual, será necessário ter também em conta o possível custo económico associado à construção de um corpus de traduções de referência, já que envolve obrigatoriamente o recurso a vários tradutores. Pelo alto custo envolvido, é normalmente inviável a construção de corpora de grandes dimensões, limitando consequentemente a cobertura das avaliações neles baseadas.

3.3 Qualidade por correcção formal

Outra noção de qualidade de tradução encontra-se associada à correcção formal, do ponto de vista linguístico, da tradução obtida. Ou seja, uma tradução é considerada tanto melhor, quanto menor for o número e a gravidade dos erros lexicais, morfológicos e sintácticos que possuir, admitindo-se, por conveniência, a ausência de erros semânticos. A identificação dos erros é feita, normalmente, por um colaborador humano em função de um sistema de classificação de erros pré-definido (embora seja possível, para algumas tarefas, usar correctores automáticos). A medida de qualidade será calculada em função dos erros anteriormente identificados e da sua respectiva gravidade.

A principal vantagem desta aproximação é o facto de ser bastante objectiva e parametrizável, permitindo a eventual formulação de diagnósticos de alto nível sobre os erros de tradução que poderão fornecer algumas pistas para o desenvolvimento dos sistemas de TA. Além disso, esta noção de qualidade tem a vantagem de permitir uma comparação fundamentada entre vários sistemas de TA, quer a nível global, quer a nível específico, explorando apenas uma subsecção do sistema de classificação de erros.

Há, contudo, algumas dificuldades na utilização desta noção de qualidade e na correspondente implementação de testes práticos. Em primeiro lugar, é necessário estabelecer um sistema de classificação de erros de TA com um formalismo e cobertura adequados, o que não é trivial, apesar de poder parecê-lo à primeira vista. É fundamental que este sistema de classificação possua elevado rigor formal. Por outro lado, é também necessário que esteja bem adaptado à classificação de erros efectuados no âmbito da TA, o que não implica que seja apropriado para a classificação de erros efectuados num contexto mais geral. Isto obriga a um estudo prévio dos erros que realmente ocorrem nas traduções geradas por TA (ver Sarmento, neste volume) e a sua divisão em categorias distintas. Esta tarefa é mais complexa do que possa parecer a priori, porque os erros ocorrem tipicamente de uma forma não estanque. Para além disso, poderá também ser muito difícil estabelecer uma correcta correlação entre os erros cobertos pelo sistema de classificação e a sua gravidade subjectiva. Esta limitação complica o estabelecimento de uma relação entre o resultado da avaliação e a experiência de utilização por parte dos utilizadores. A dificuldade acentua-se quando se conciliam num mesmo sistema de classificação dois sistemas de análise linguística diferentes: o da língua de partida e o da língua de chegada, com todos os problemas inerentes associados aos fenómenos de cruzamento das línguas, tal como a interferência e o tradutês (*translationese*).

Uma outra dificuldade prática está relacionada com a necessidade de recorrer a colaboradores que possuam uma razoável formação em linguística e alguma experiência em tradução, humana e automática. No caso do português, pensamos que não existem actualmente recursos humanos em número suficiente que preencham estes requisitos em simultâneo, o que sugere que, para se realizar uma actividade de avaliação de TA nestes moldes, seria necessário investir inicialmente alguns recursos na formação de eventuais colaboradores.

Por último, verifica-se na prática que esta noção de qualidade só é viável em sistemas de TA que possuam à partida um desempenho razoável, ou seja, que não produzam demasiados erros na tradução. Para sistemas de TA com um desempenho muito fraco, a avaliação pode tornar-se impraticável, já que a abundância e dispersão de erros poderá inviabilizar a sua correcta identificação ou distinção.

3.4 Qualidade em função do esforço de pós-edição

Uma outra alternativa para medir a qualidade da TA está relacionada com o esforço de pós-edição que é necessário despender na tradução obtida para que esta atinja a qualidade desejada. Esta noção de qualidade não é nova e surge naturalmente com a utilização de sistemas de TA. Há muito tempo que os utilizadores de TA, em particular as organizações que recorrem com mais frequência à TA, estão habituados a tarefas de pós-edição que permitem elevar o nível das traduções obtidas ao patamar de qualidade desejado. Por exemplo, empresas como a Ford Motor Company e instituições como a Comissão Europeia utilizam sistemas de TA para a tradução de manuais técnicos e documentos legais, respectivamente, e estão dotadas de equipas de linguistas e tradutores especializados em TA que se encarregam de certificar-se de que todos os níveis de representação linguística e contextual estão traduzidos correctamente. Neste contexto, uma tradução será tanto melhor, quanto menor for o esforço de pós-edição necessário para atingir o nível de qualidade desejado.

Note-se que esta noção de qualidade baseia-se numa aproximação profundamente pragmática, centrada em critérios de adequação da tradução (até implicitamente económicos) e não apenas em critérios formais de ordem linguística. Uma característica inerente a esta noção de qualidade é a possibilidade de ser integrada naturalmente (pelo menos em teoria) num regime de avaliação interactivo, uma vez que se integra facilmente no processo de utilização da TA.

O cálculo do esforço de pós-edição poderá, no entanto, também não ser uma tarefa simples. Numa aproximação simplista, o cálculo do esforço de pós-edição poderia ser feito

recorrendo à ajuda dos próprios tradutores ou correctores que executam as tarefas de pós-edição. Estes colaboradores anotariam todas as operações de pós-edição por eles executadas, tornando depois possível um cálculo aproximado do esforço despendido na pós-edição. Contudo, esta solução não é prática, porque cria demasiado atrito no processo de pós-edição, e pode além disso depender do desempenho individual dos participantes.

A utilização de métodos automáticos para a computação do custo de pós-edição poderá ser uma alternativa viável (Su *et al.*, 1992). Os métodos automáticos identificam, a partir da tradução automática crua e do resultado da sua pós-edição, a possível sequência de operações de edição realizadas (inserção, remoção, substituição e reordenação) pelo pós-editor. Desta forma, não é necessário recorrer à anotação manual para contabilizar as operações de pós-edição. Após a atribuição de um custo a cada uma dessas operações, poderá ser facilmente calculado um custo global para o esforço de pós-edição.

Não é, no entanto, óbvio que o mesmo custo deva ser atribuído à substituição de uma palavra por outra sinónima, e à substituição por outra de sentido radicalmente diferente, mesmo que demore o mesmo tempo ou implique o mesmo esforço computacional, donde é preciso complementar este tipo de métricas com outras que levem em conta o significado.

Finalmente, e embora se possa imaginar uma avaliação conjunta em que a pós-edição poderá, teoricamente, ser realizada manualmente para todos os sistemas de TA em avaliação, por um conjunto de juízes, é preciso relembrar que a pós-edição não conduz necessariamente ao mesmo texto por tradutores diferentes, mesmo que instruídos com directivas rígidas. Isto introduz novamente o factor de múltiplas soluções, acrescido a que uma tal avaliação resultará consideravelmente dispendiosa.

4. A via do pólo do Porto da Linguateca

Na prática, são muitos os desafios que se colocam para que seja exequível uma avaliação de sistemas de TA em torno do português e, em particular, para a sua concretização na forma de uma avaliação conjunta. Embora como referência de partida já exista algum trabalho realizado no âmbito da avaliação de TA em torno do português (Oliveira *et al.*, 2000; Marrafa & Ribeiro, 2001) além dos trabalhos da década de 80 com materiais de teste (Santos, 1988), e nos tenham sido disponibilizadas[3] as grelhas de classificação usadas no projecto TRADAUT[4], estes trabalhos isolados não cobrem evidentemente ainda as imensas questões que a área coloca, tanto do ponto de vista teórico, como do ponto de vista prático.

Pensamos que a complexidade inerente a esta área e a necessidade de envolver equipas multidisciplinares tem dificultado o seu desenvolvimento entre os investigadores da língua portuguesa. De facto, no Porto não existia grande experiência na área específica da tradução automática, nem se encontrava disponível uma massa crítica suficiente para um estudo profundo sobre o assunto. Ainda assim, conseguimos reunir uma pequena equipa multidisciplinar para estudar a sua avaliação, equipa essa composta por três elementos com experiência e formação diferentes: uma especialista em tradução e no contraste entre as línguas inglesa e portuguesa (líder do projecto) com experiência na utilização de sistemas de tradução automática num contexto pedagógico; uma linguista com experiência no desenvolvimento e teste de sistemas práticos (comerciais) de tradução automática; e um engenheiro informático da área da inteligência artificial com experiência de desenvolvimento de sistemas integrando o utilizador. Esta equipa pôde ainda recorrer à colaboração dos alunos do Mestrado em Tradução e Terminologia da Faculdade de Letras da Universidade do Porto (FLUP), a maioria dos quais desempenhando funções profissionais ao nível da tradução, e que por isso eram bastante sensíveis aos problemas em causa. No entanto, o facto de estarem ainda em formação impôs algumas limitações ao tipo de tarefas que lhes puderam ser atribuídas, em particular quando necessitassem de conhecimentos mais profundos na área da linguística.

A nossa abordagem foi uma reaproximação à avaliação de TA. Na nossa opinião, esta reaproximação deverá primeiro observar atentamente o desempenho de vários sistemas de TA disponíveis, de forma a obter um visão concreta da capacidade actual da TA. Por outro lado,

[3] Por esse facto exprimimos a nossa gratidão a Maria Francisca Xavier e Graça Vicente.
[4] Veja-se http://terra.di.fct.unl.pt/.

urge saber que problemas se pretende resolver, ou seja, com que tipo de fenómenos se quer lidar: em suma, que tipo de traduções se pretende. Como veremos em Maia & Barreiro (neste volume), é frequente o emprego de conjuntos de teste na avaliação de TA. Estes conjuntos de teste são, sem dúvida, muito interessantes para tentar explorar exaustivamente um determinado fenómeno ou estrutura. Contudo, como referido por Whittaker & Stenton (1989), só é possível encontrar um conjunto de teste minimamente significativo se se conhecer os principais fenómenos do problema. Conjuntos de teste genéricos poderão não ser capazes de cobrir convenientemente o tipo de situações que são mais críticas na TA actual, já que estas não são verdadeiramente conhecidas na prática ou não estão convenientemente documentadas, pelo menos para o português. Além disso, a própria utilização de conjuntos de teste "tradicionais" poderá facilmente incorrer num erro de ingenuidade. De facto, como a maioria dos fenómenos cobertos tradicionalmente em conjuntos de teste partem de problemas linguísticos relativamente bem conhecidos e documentados, os próprios sistemas de TA encontram-se de alguma forma já satisfatoriamente preparados para lidar com essas situações. Verifica-se, na prática, que alguns sistemas apresentam um desempenho bastante bom em situações que à luz dos referidos conjuntos de teste revelariam ser de grande complexidade. Por outro lado, verifica-se também que os motores de TA se mostram muito pouco robustos em situações em que se esperariam menos problemas, como mencionado em Santos *et al*. (2004).

É neste contexto que surgem as actividades do pólo do Porto da Linguateca, que iremos descrever em detalhe nos próximos dois capítulos. O primeiro, sobre os mecanismos desenvolvidos para visualizar o desempenho actual da tradução automática com os sistemas disponíveis, descreve trabalho que permitiu desenvolver familiaridade com as capacidades actuais de TA e com as situações mais problemáticas; assim como documenta o desenvolvimento de um ambiente computacional para a recolha colectiva de exemplos de TA caracterizados pelo utilizador, o TrAva. O segundo capítulo, por seu turno, relata a experiência obtida na recolha de problemas de tradução utilizando o TrAva, descrevendo alguns dos casos mais sistematicamente focados pelo conjunto de utilizadores, e os benefícios pedagógicos de tal iniciativa.

Os esforços que desenvolvemos e que descrevemos nos próximos capítulos não serão, de forma alguma, a palavra final na avaliação de TA, nem certamente o único caminho interessante que conduzirá a uma eventual avaliação conjunta neste domínio. São, no entanto, pequenos passos desprovidos de ideias pré-concebidas acerca da tradução automática e cuja realização nos parece útil relatar à comunidade envolvida nesta fascinante empresa.