Universidade de Lisboa

Faculdade de Letras

Tradução Automática e Linguagens Controladas:

Contributos para um Português Controlado

Dissertação de Mestrado em Tradução

Ana Lucrécia Madeira Gomes

2010

Universidade de Lisboa

Faculdade de Letras

Tradução Automática e Linguagens Controladas:

Contributos para um Português Controlado

Ana Lucrécia Madeira Gomes

Dissertação apresentada à Faculdade de Letras da Universidade de Lisboa para obtenção do grau de Mestre em Tradução.

Orientadora:

Professora Doutora Palmira Marrafa.

Agradecimentos

Realizar este trabalho foi para mim a mais dura prova da minha vida. É com enorme alegria que chego ao fim e escrevo estas palavras.

Gostaria de agradecer à minha colega de Mestrado, Náheda Ibrahimo, pelo refúgio que eu encontrei em sua casa. Agradeço-lhe o apoio, a orientação científica, a leitura crítica do meu trabalho. Agradeço-lhe tudo o que me ensinou sobre a fé e o poder que temos em nós.

Gostaria de agradecer à minha mãe, Elisabete, e à minha irmã, Célia. As minhas companheiras de vida, no melhor e no pior.

Gostaria de agradecer aos meus tios Helena e Neca e à minha madrinha Aldára. Agradeço-lhes por me permitirem viver.

Gostaria de agradecer à Professora Palmira Marrafa, pela orientação tão paciente, por me ter ensinado a pensar. Agradeço também à Sara Mendes e à Raquel Amaro pela forma como nos receberam no Centro de Linguística da Universidade de Lisboa.

Gostaria de agradecer aos meus amigos que estiveram comigo nestes três anos de Mestrado. Aos amigos da Residência, aos amigos que tenho nos quatro cantos do mundo e nos quatro cantos de Portugal. À Sam, pela revigorante viagem a Praga, ao Maciek pela ajuda informática, à Sónia C. com quem partilhei a escrita de partes deste trabalho pelo altifalante do telemóvel.

À Tica…

Gostaria de agradecer ao Ivo.

Não tenho palavras para te dedicar, nada do que eu possa dizer está à altura de quanto significas para mim. Mas agradeço-te o amor, a força, o apoio; agradeço-te o quanto lutaste por mim, as vezes que me ralhaste para eu não desistir.

Índice

Resumo

A linguagem natural ainda coloca vários problemas tanto ao seu processamento automático, em geral, como à tradução automática, em particular, devido a fenómenos como a ambiguidade, os processos de inferência e, no caso dos sistemas que envolvem a fala, certas características da oralidade. As linguagens controladas constituem uma estratégia para adaptação dos textos a serem traduzidos por sistemas de tradução automática, de forma a obterem-se *outputs* aceitáveis.

A presente dissertação justifica-se, na medida em que se desconhecem propostas de elaboração de uma linguagem controlada para o português europeu. Neste trabalho analisa-se a utilidade das linguagens controladas no âmbito da tradução automática e oferece-se um contributo para a construção de um fragmento de um Português Controlado.

A partir da tradução automática de frases que integram determinados fenómenos linguísticos, entre os quais expressões verbais complexas, sujeito nulo, alternâncias verbais, modalidade e uso de determinantes, sistematizam-se formas de controlar estes fenómenos no sentido acima referido.

Os resultados a que se chegou no âmbito do presente trabalho demonstram que um possível Português Controlado há que encontrar formas de essencialmente veicular o sentido expresso por alguns enunciados ambíguos ou de difícil processamento automático.

O presente trabalho assume uma importância particular, na medida em que contribui para a investigação no âmbito das linguagens controladas em língua portuguesa, que constituem ainda um domínio de investigação pouco explorado.

Palavras-chave: tradução automática, linguagens controladas, *input*, *output*, Português Controlado

Abstract

Natural language still poses some challenges to computational processing, as well as to machine translation. Among the features particular to natural language and that can be an obstacle to computational processing are ambiguity, inference and, in the case of spoken-language machine translation systems, some features of oral speech.

This dissertation is relevant for it gives a contribution to the construction of a controlled language for European Portuguese, having in mind that one ignores such a proposal. This study provides an analysis on the utility of controlled languages for machine translation.

The analysis in this dissertation begins with the machine translation of some linguistic phenomena such as complex predicates, null subject, verbal alternations, modality and the use of determiners in English. The machine translation of such phenomena aimed then at finding possible ways of controlling them in order to achieve an acceptable output.

The results of this study show that a future controlled Portuguese has to essentially convey the meaning expressed by some linguistic structures that are difficult for computational processing.

This dissertation is of particular importance, as it contributes to the research on controlled languages in Portuguese, which is an unexplored field of investigation in Portugal.

Key-words: machine translation, controlled languages, input, output, controlled Portuguese

1. Introdução

1.1 Objecto de estudo

Esta dissertação enquadra-se numa área em crescimento, e cuja importância é cada vez mais reconhecida: as linguagens controladas.

O desenvolvimento de linguagens controladas (ou simplificadas) para o português não tem merecido a devida atenção, tendo em conta a escassa literatura sobre o assunto em Portugal e sobretudo o facto de não se conhecerem propostas de uma forma de linguagem controlada para o português europeu.

As linguagens controladas são uma forma de linguagem natural submetida a uma série de restrições de forma a eliminar todas as características da linguagem natural que possam constituir um entrave ao processamento computacional de um texto destinado a ser traduzido automaticamente, ou que possam colocar dificuldades à compreensão por parte do falante. Neste seguimento, distinguem-se duas formas de linguagens controladas. As que se destinam à pré-edição de determinado texto de forma a ser mais facilmente processado pela máquina (*machine-oriented controlled languages*) e que, consequentemente, têm eficácia comprovada na tradução automática. E as que se destinam ao utilizador humano (*human-oriented controlled languages*) e têm como fim a resolução de problemas de ambiguidade, por exemplo, que dificultem a compreensão do texto

A história das linguagens controladas remonta à década de 30 do século passado, quando o *Simple English* foi criado. Este era uma forma simplificada da linguagem natural destinada a facilitar a compreensão do inglês pelos falantes não nativos e por todos os que aprendiam inglês. Ao longo dos anos, esta área tem crescido e a ela se tem dedicado muito esforço de investigação, de tal forma que a aposta no estudo e desenvolvimento de linguagens controladas é cada vez maior.

A aplicação das linguagens controladas tem-se cingido praticamente ao texto técnico, mais especificamente à documentação técnica, nas empresas de grande dimensão. No entanto, o seu uso começou a alargar-se a outras áreas técnicas,

nomeadamente à Aeronáutica, onde, por razões de segurança, a ambiguidade inerente à linguagem natural deve ser totalmente eliminada.

Actualmente, a aplicação das linguagens controladas estende-se a outras áreas, como a Medicina e o Direito, as comunicações de emergência, a recuperação de informação em bases de dados em linha, a *Web* Semântica, etc.

No âmbito da tradução automática, a importância das linguagens controladas deriva do facto de estas permitirem adaptar o texto às características dos sistemas, o que resulta num *output* de maior qualidade relativamente ao de um texto não controlado, tornando quase desnecessária a intervenção humana na pós-edição. Claro que o grau de intervenção humana na pós-edição de determinado texto depende da finalidade da tradução. Por exemplo, a pós-edição pode ser nula se a tradução se destinar simplesmente à captação grosseira do sentido, ou pode ser detalhada se a tradução se destinar à publicação.

Em termos gerais, as linguagens controladas existentes são usadas nas empresas transnacionais, na redacção de documentação técnica, e destinam-se ao utilizador humano. É o caso do Simplified Technical English. Esta, como outras linguagens controladas, cobre o léxico e a gramática. As regras nestas áreas visam fundamentalmente alcançar um nível de compreensão bastante elevado para o utilizador humano. Para tal, por exemplo, a nível lexical, há uma selecção de palavras permitidas de forma a reduzir ou evitar casos de ambiguidade. A nível gramatical, simplifica-se, por exemplo, o uso de tempos verbais que se deve reduzir ao Infinitivo, Imperativo, Presente do Indicativo, Pretérito Perfeito Simples e Futuro.

1.2 Objectivos e motivação

O presente trabalho pretende dar um contributo para a construção de uma linguagem controlada para o português europeu orientada para a melhoria da qualidade dos *outputs* em tradução automática. Nesse sentido, analisam-se construções linguísticas que induzem resultados inadequados, problematizando-as, de forma a apontar possíveis soluções.

Os principais problemas que se colocam à tradução automática têm a ver com as ambiguidades da linguagem natural mas também com a complexidade sintáctica. Tomem-se, a título de exemplo, as construções que se seguem:

(1) […] um relatório com documentos e provas que permitem identificar […]
(2) […] chegou a ser vetado […]
(3) Se se comprar a casa rapidamente, vou ficar contente.
(4) A Ana anda nervosa porque não consegue acabar o trabalho.

Todas as construções resultaram num *output* problemático na tradução automática, como se pode constatar nas respectivas traduções produzidas pelo sistema de tradução automática Systran:

(1) (a) […] a report with documents and you prove that […]

(2) (a) […] arrived to be vetoed […]

(3) (a) If the house to be bought quickly, goes to stay glad.

(4) (a) The Dwarf walks nervous because it does not manage to finish the work.

Os problemas que estas traduções exibem são vários. Em (1) (a), por exemplo, o substantivo *provas*, sendo formalmente idêntico à 2ª pessoa do singular do Presente do Indicativo do verbo *provar – tu provas*- e, por conseguinte, ambíguo para a máquina, dá origem a um resultado inadequado, em que o substantivo é traduzido como sendo a forma verbal, a que o sistema atribui sujeito.

O exemplo em (2) (a) é um caso de um predicado complexo traduzido palavra a palavra, gerando, consequentemente, um *output* inadequado. Para além disso, o verbo *chegar* é ambíguo, podendo ser interpretado como verbo principal e como operador modal. No entanto, o sistema selecciona inadequadamente o primeiro sentido, traduzido-o por *arrive*.

Em (3) (a) os problemas evidenciados prendem-se com a omissão do sujeito na oração principal.

À semelhança do que acontecia em (2), um dos problemas que a frase em (4) apresenta tem a ver sobretudo com a ambiguidade lexical do verbo *andar*. Um outro problema prende-se com a inadequada realização do sujeito na oração subordinada.

São problemas desta natureza que serão contemplados no presente trabalho, com o objectivo de criar regras que se constituam como um modo eficaz de transmitir o mesmo conteúdo, resultando num *output* melhorado na tradução pela máquina.

Esta proposta de construção de uma linguagem controlada para o português europeu atentará somente na sua utilidade para a tradução automática. Não será abordada uma simplificação da linguagem natural com o fim de tornar determinado tipo de texto mais inteligível para o utilizador humano.

Uma das regras que se estabelece, por exemplo, diz respeito ao sujeito nulo, possível em português, como se pode observar em (4). Em inglês, língua de chegada neste trabalho, não são possíveis sujeitos nulos, pelo que o sistema atribui, na maior parte dos casos analisados, um sujeito expletivo. Desta forma, para evitar este problema, pode estabelecer-se que o sujeito seja realizado sempre que possível, em português.

Os problemas que a linguagem natural coloca à tradução automática são vários e as regras em linguagem controlada ajudam a contorná-los e a adaptar os textos às características dos sistemas.

As motivações que orientam a ideia da presente dissertação são várias. Por um lado, na exaustiva pesquisa bibliográfica efectuada comprovaram-se as vantagens e a utilidade das linguagens controladas nas diferentes áreas de aplicação. Por outro, justifica-se o tema com o facto de se desconhecer uma proposta de uma linguagem controlada para o português europeu. Mas, acima de tudo, a motivação maior para o presente trabalho está na criação de uma linguagem controlada para o português com o objectivo de obter um *output* mais aceitável na tradução automática do português para o inglês: o Português Controlado.

1.3 Metodologia e obtenção de dados

Para a elaboração desta dissertação, numa primeira fase, recolheu-se e analisou-se um número de textos jornalísticos nas páginas electrónicas de vários jornais portugueses, entre eles O Público, Diário de Notícias, páginas electrónicas da TSF e da revista Visão. A motivação para a escolha de textos jornalísticos prende-se com facto de este tipo de escrita ter algumas semelhanças com texto dito técnico, na medida em que, de uma forma geral, tal como o texto técnico, o texto jornalístico envolve vocabulário e estruturas muito próprios e recorrentes. A análise centra-se em propriedades da linguagem natural que colocam problemas à tradução automática.

Os textos recolhidos foram traduzidos em sistemas de tradução automática disponíveis na Internet. Os sistemas utilizados foram o Systran e o Google. A partir da análise dos *outputs*, seleccionaram-se alguns fenómenos linguísticos mais recorrentes e que colocaram mais problemas ao processamento computacional. Na demonstração dos problemas analisados há também exemplos de frases que foram retiradas da *Gramática da Língua Portuguesa* de Mateus *et al.* e outros resultantes de introspecção.

1.4 Organização da dissertação

Esta dissertação está organizada em cinco capítulos. No primeiro, o capítulo introdutório, dão-se a conhecer as noções fundamentais sobre tradução automática e sobre linguagens controladas que serão debatidas ao longo do trabalho.

O segundo capítulo aborda a tradução automática, em termos da sua história, dos diferentes tipos de sistema, da qualidade dos *outputs* e das limitações dos sistemas. Partindo da análise de *outputs* de sistemas de tradução automática disponíveis em linha, apresentam-se exemplos de produções linguísticas ainda problemáticas para a tradução automática.

O terceiro capítulo aborda o tema das linguagens controladas. Fala-se da sua história, dos tipos e dos fins a que se destinam. Discutem -se ainda as suas vantagens e o impacto na tradução automática.

No quarto capítulo esboça-se uma forma preliminar de uma linguagem controlada para o português europeu, direccionada especificamente para a tradução automática. Neste capítulo, sistematizam-se questões lexicais e sintácticas problemáticas para a tradução automática, a partir da tradução de textos em sistemas como o Google e o Systran. A partir desta sistematização, criam-se algumas regras para a escrita em linguagem controlada.

Por fim, o quinto capítulo constitui uma breve conclusão, onde se analisa o trabalho desenvolvido e se comenta o contributo desta dissertação para a criação de um português europeu controlado.

2. Tradução Automática

Neste capítulo, aborda-se o tema da tradução automática sob diversas vertentes: histórica (breve panorâmica), tipos de sistemas, qualidade dos *outputs* e problemas em tradução automática.

Em 2.1, traça-se uma breve história da tradução automática que pretende dar conta dos seus marcos mais importantes, evidenciando a aposta que tem sido feita na investigação nesta área.

Em 2.2, apresentam-se os diferentes tipos de sistemas de tradução automática e analisam-se os seus *outputs*. De seguida, procede-se a uma sistematização de alguns problemas que ainda se colocam à tradução automática.

2.1 Breve Panorâmica Histórica

Em traços muitos gerais, pode dizer-se que a tradução automática é a tradução de um texto de uma língua para outra através de computador.

Em 1933, surge a ideia de utilizar o computador para efectuar traduções entre línguas. Os precursores desta ideia foram Artsrouni e Troyanskii. No entanto, esta proposta não teve atenção por parte da comunidade científica de então, pelo que ainda não é nesta altura que nasce a tradução realizada por computador.

Na década de 40, surgem as primeiras investigações em tradução automática e consequentemente surgem os primeiros sistemas. Inicialmente, acreditava-se que a tradução automática podia fazer-se apenas ao nível lexical, bastando estabelecer equivalências entre palavras, como se pode ler, por exemplo, em Hutchins (1999):

> "For many years, the systems were based primarily on direct translations via bilingual dictionaries, with relatively little detailed analysis of syntactic structures."

A estrutura sintáctica e a composicionalidade semântica não eram de início sequer consideradas. No entanto, logo se percebeu que uma abordagem em que as expressões linguísticas são tratadas como sequências lineares de elementos, tem resultados muito pobres e está destinada ao fracasso (cf. Marrafa (1993), entre outros). Surgiu então a necessidade de criar sistemas capazes de dar conta das propriedades da linguagem natural.

A investigação em Linguística, embora um tanto divorciada do processamento automático da linguagem natural, virá mais tarde a influenciar de algum modo a tradução automática. Chomsky publica em 1957 *Syntactic Structures*, um dos grandes marcos da chamada Gramática Generativa. Por essa altura, surgem os sistemas de 1ª geração e, mais tarde, os sistemas de 2ª geração. Chomsky contribui, com a sua obra *Aspects of the Theory of Syntax* (1965), para o desenvolvimento dos sistemas de 2ª geração. A importância do contributo de Chomsky é tanto mais importante quanto traz à luz uma maior «compreensão da língua natural» (Marrafa, 1993: 284), com consequências, tardias embora, no processamento automático da linguagem natural e, neste seguimento, no desenvolvimento da tradução automática.

Mais tarde, no ano de 1966, nos Estados Unidos, o ALPAC (Automatic Language Processing Advisory Commitee) publica um relatório onde avalia e dá conta do estado da tradução automática e conclui que os resultados são ainda muito insatisfatórios para se apostar na investigação nesta área. No referido relatório aponta-se ainda no sentido de se promoverem as ferramentas de apoio à tradução, tal como se pode ler em Hutchins e Somers (1992:7):

> "In its influential 1966 report it concluded that MT was slower, less accurate and twice as expensive as human translation and stated that "there is no immediate or predictable prospect of useful Machine Translation". It saw no need for further investment in MT research; it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support of basic research in computational linguistics."

Como consequência deste relatório, deu-se o descrédito na tradução automática. No rescaldo do referido relatório, surgem os métodos de tradução baseada em regras, isto é, os sistemas de transferência e interlíngua. O EUROTRA é um exemplo desse tipo de sistemas. Trata-se de um projecto multilingue de tradução automática desenvolvido

desde finais dos anos 70 até aos primeiros anos da década de 90, pelos países da então Comunidade Económica Europeia.

Anos mais tarde, na década de 80, o entusiasmo renasceu (Mateus, 1995:119). Nesta década, a tradução automática era utilizada pelos governos de vários países, no comércio e na indústria. Por esta razão, a investigação nesta área tornou-se, nessa época, uma aposta de empresas que desenvolviam os seus próprios sistemas (Slocum, 1985).

Nos anos 80, a investigação centrou-se em novos tipos de sistemas de tradução automática e, neste seguimento, surgiram, já nos anos 90, os sistemas baseados em *corpora* que englobam os sistemas baseados em exemplos e os sistemas estatísticos.

Ao longo do tempo, a investigação em tradução automática tem seguido por vários caminhos. Recentemente tem-se centrado também no desenvolvimento de sistemas de tradução automática da linguagem oral. Estes sistemas colocam novas questões à investigação nesta área, pelo facto de terem de lidar com as características inerentes à linguagem oral, tais como hesitações, incorrecções, auto-correcções, diferentes pronúncias (Somers, 2003:521).

Actualmente, a tradução automática impõe-se como uma ferramenta de relevo para diversos propósitos, tais como a comunicação interlinguística e intercultural e a disseminação de informação. A utilização da tradução automática na Internet é uma realidade. As ferramentas de tradução automática são disponibilizadas por alguns *browsers*. Desta forma, muitas páginas HTML dispõem dessas ferramentas para a tradução dos conteúdos.

Na Internet há ainda páginas electrónicas que disponibilizam sistemas de tradução automática. A história dos sistemas de tradução automática, em linha e gratuitos, remonta a 1997. Neste ano, uma cooperação entre a Systran Software Inc. e o motor de busca Altavista deu origem ao Babelfish, o primeiro sistema de tradução automática em linha ao dispor dos utilizadores de forma gratuita. Outros sistemas são o Systran, Google, Free Translation Online, SDL, entre outros. Nas páginas electrónicas que disponibilizam estes sistemas, os utilizadores dispõem de uma janela de tradução onde podem colar o texto ou o URL da página electrónica que desejam traduzir.

A tradução automática na Internet está também ao dispor do correio electrónico e de *chat rooms* (Hutchins (1999), Sommers (2003:523)). O recurso à tradução

automática na Internet é uma aposta em crescimento, há cada vez mais empresas a desenvolverem sistemas e a apostarem na investigação em tradução automática.

Em suma, num mundo globalizado, numa era dominada pela velocidade na transmissão de informação, a tradução automática é uma forma de permitir a comunicação entre pessoas separadas pela barreira linguística e de tornar o processo rápido, eficaz e barato.

2.2.1 Tipos de Sistemas

A área da investigação em tradução automática é vasta e tem resultado em sistemas de características diferenciadas que visam tratar os problemas que ainda se põem à tradução automática das mais diversas formas.

Para uma breve classificação dos sistemas de tradução automática, quanto ao número de línguas que estão envolvidas no processo de tradução, pode dizer-se que estes se dividem entre sistemas bilingues e sistemas multilingues. Isto é, sistemas que traduzem entre duas línguas e sistemas que traduzem entre mais do que duas línguas.

A tradução automática pode também estabelecer-se num só sentido, ou pode ser reversível. Isto é, a tradução pode ocorrer num único sentido, entre uma língua de partida e uma língua de chegada, ou pode ocorrer nos dois sentidos (Hutchins e Somers, 1992:4).

Os sistemas podem ainda classificar-se quanto ao método em que se enquadram: tradução directa, transferência, interlíngua ou métodos empíricos (baseados em *corpora*, estatísticos e sistemas híbridos).

O Triângulo Vauquois, proposto na década de 70 por Vauquois, ilustra os níveis de profundidade do conhecimento dos diferentes métodos não empíricos.

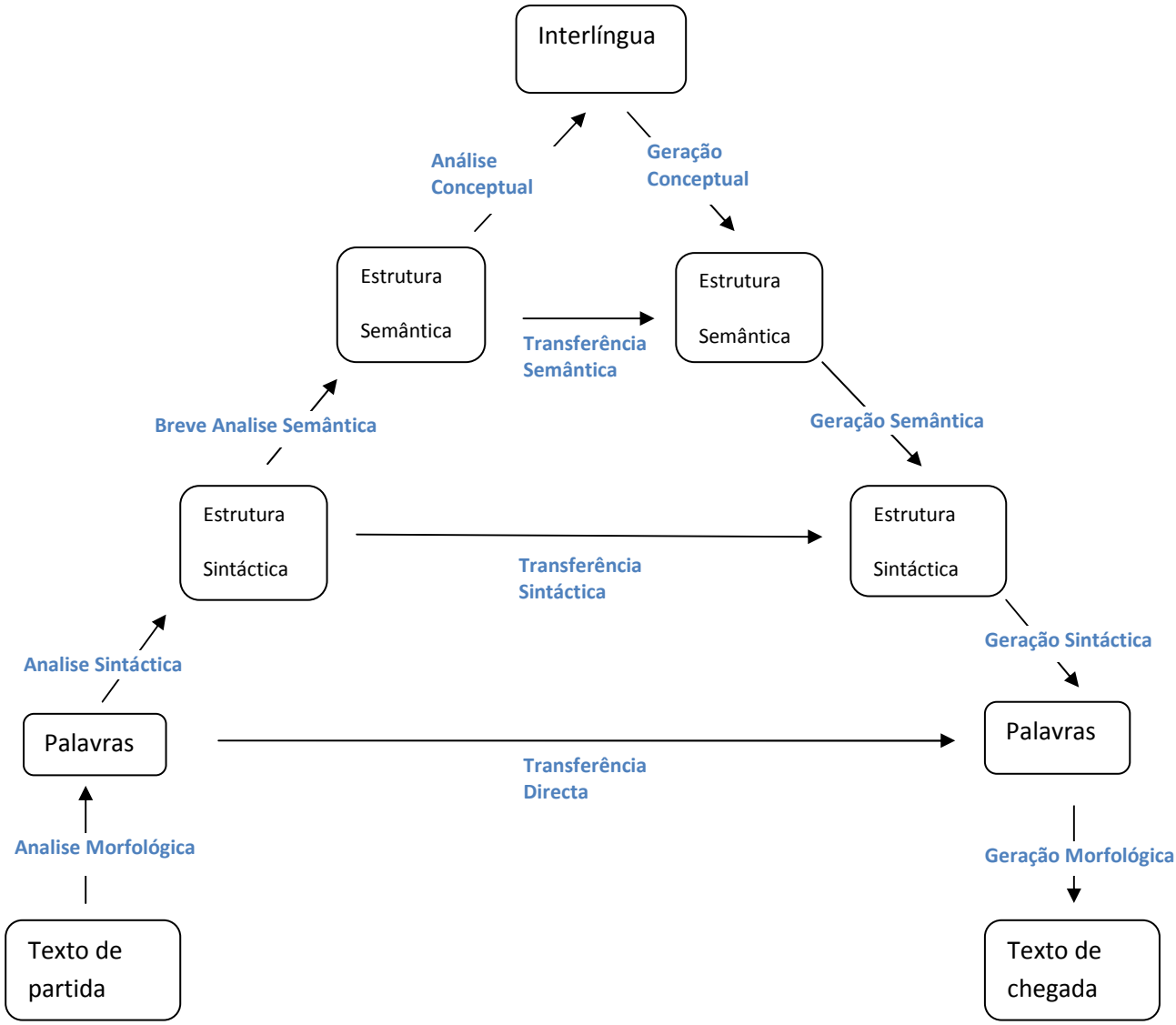

Figura 1 Triângulo de Vauquois

A estrutura piramidal mostra que, ao nível da tradução directa, a necessidade de conhecimento limita-se apenas ao conhecimento necessário para a transferência directa do significado básico de palavras individuais. À medida que se vai subindo no triângulo, aumenta a profundidade da transferência semântica necessária. Desta forma, a um nível intermédio, a tradução automática por transferência já requer uma análise sintáctica e até semântica. No topo do triângulo, a tradução automática por interlíngua é a que mais se afasta da tradução directa. Esta, sendo levada a cabo por meio de uma representação

do sentido (a interlíngua), chega a ser independente, tanto da língua de partida como da língua de chegada. Vejam-se estes métodos ao pormenor.

A tradução directa estabelece-se entre palavras das línguas envolvidas no processo de tradução, sem qualquer representação intermediária. Numa primeira fase da tradução automática, o método de tradução directa tinha resultados muito pobres, pois traduzia quase só ao nível do léxico. Qualquer operação de análise sintáctica ou morfológica era bastante elementar e resumia-se, segundo Hutchins (2003:503), à resolução de ambiguidades, à identificação correcta das expressões da língua de partida e à especificação da ordem das palavras na língua de chegada. Os sistemas directos dispõem apenas de um dicionário bilingue e de um único programa para analisar e gerar o texto de partida.

Numa primeira fase do processo de tradução directa, um *parser* analisa o texto e o resultado desta análise é um conjunto de palavras com a indicação de categoria e função. Segue-se o confronto deste conjunto de palavras com regras que assentam num dicionário bilingue. Essas regras englobam indicações para organizar a ordem das palavras na frase e regras morfológicas, para a geração dos traços de pessoa, número ou de sufixação/prefixação. Este processo vai resultar na geração do texto de chegada.

O método de tradução indirecta parte de conhecimento linguístico, ou seja, de modelos gramaticais que permitem uma análise abstracta e relativamente profunda tanto do texto/língua de partida como do texto/língua de chegada. O método de tradução indirecta engloba duas abordagens: a tradução por transferência e a tradução por interlíngua.

A tradução por interlíngua consiste num processo por fases: uma primeira fase em que o texto de partida é convertido numa representação independente das línguas, a interlíngua, que por sua vez, serve de *input* para a geração do *output* na língua de chegada, numa segunda fase.

A interlíngua é uma representação do «significado», independente das línguas envolvidas no processo de tradução. O nível de análise que se consegue com este método é pormenorizado e profundo, ao ponto de permitir a independência da interlíngua em relação às línguas de partida e de chegada. O conceito de interlíngua não

é fácil de assimilar quando se fala apenas de uma representação semântica. Hutchins (2003:503) dá uma definição bastante esclarecedora:

> "[…] the complexity of the Interlingua itself is greatly increased. Interlinguas may be based on a 'logical' auxiliary language (such as Esperanto), a set of semantic primitives common to all languages, or a supposedly 'universal' vocabulary."

O método por transferência, por sua vez, tem três fases. Numa primeira etapa, o texto de partida é analisado a vários níveis (lexical, morfológico, sintáctico, semântico), com recurso a um *parser* e a uma gramática do texto de partida. Desta análise surge uma forma de representação do texto de chegada ("an abstract target text representation" (Hutchins e Somers, 1992:75). Numa segunda fase, a transferência, a representação a que se chegou na fase da análise é transferida para uma representação na língua de chegada. Posteriormente, na fase de geração ou síntese, gera-se o *output* da tradução, passando da representação na língua de chegada para o *output* da tradução, usando uma gramática da língua de chegada.

Os métodos empíricos, tais como os sistemas de tradução automática baseada em estatística e de tradução automática baseada em exemplos, estão excluídos do Triângulo Vauquois. Estes sistemas têm um conhecimento linguístico reduzido ou nulo e baseiam-se sobretudo em estatística e em exemplos. Os sistemas baseados neste método constituem novas apostas na área da tradução automática, tal como, por exemplo, os sistemas de tradução automática baseada em diálogo ou de tradução automática baseada em redes neuronais.

No contexto dos paradigmas empíricos, os sistemas estatísticos (*Statistical MT*) baseiam-se num vasto *corpus* bilingue de exemplos de traduções. Este tipo de sistemas coloca lado a lado sintagmas, grupos de palavras e palavras (Somers, 2003:516) presentes em textos de *corpora* paralelos[1]. O passo seguinte no processo de tradução deste tipo de sistema é calcular a probabilidade de determinada palavra e/ou sintagma do *input* corresponder à palavra e/ou sintagma presente no texto paralelo.

Já os sistemas baseados em exemplos (*Example-based MT*) dispõem de exemplos de traduções armazenadas que são recuperadas e utilizadas durante o processo de tradução. Este tipo de sistema dispõe de um *corpus* bilingue de pares de traduções

[1] Entende-se por *corpora* paralelos um vasto conjunto de textos aos quais se faz corresponder a respectiva tradução.

que são então usados na tradução. O processo de tradução neste tipo de sistemas desenrola-se em três fases.

Na primeira fase, *matching*, o sistema procura exemplos com base na semelhança com o *input* da tradução (Somers, 2003:515). Isto é, tendo como *input*, por exemplo, uma frase A, o sistema vai procurar, na sua base de dados, os exemplos de traduções que mais se assemelham à frase do *input* (frase A).

Na segunda fase, *alignment*, o sistema identifica vários exemplos de traduções da sua base de dados que devem conter porções que traduzam partes da frase que serve de *input*.

Numa última fase, *recombination*, essas porções de frases recolhidas dos exemplos de traduções são recombinados de forma a construir uma frase aceitável na língua de chegada.

Importa aqui fazer um contraponto com as memórias de tradução (*Translation Memories*). Numa primeira análise, pode dizer-se que as memórias de tradução funcionam da mesma forma que os sistemas baseados em exemplos. No entanto, as memórias de tradução distinguem-se por serem interactivas, isto é, o tradutor tem a hipótese de escolher aceitar ou rejeitar a tradução proposta.

As ferramentas de memória de tradução procuram, na sua base de dados, unidades de tradução (segmentos), que são geralmente frases. Estes segmentos caracterizam-se por serem idênticos ou similares ao segmento que serve de *input*. Durante o processo de tradução, os segmentos que a memória de tradução recupera da sua base de dados são apresentados ao tradutor. Cabe a este aceitar ou recusar a proposta de tradução. [2]

Nos sistemas de tradução automática baseados em exemplos, por sua vez, não há hipótese de escolha, o processo de tradução decorre sem interrupção. O sistema vai comparando os pares de traduções do *corpus* de forma a encontrar a tradução que mais se aproxima do texto de partida.

[2] Atente-se na descrição do funcionamento de um sistema de memória de tradução (Freigang, 2001).

Face às limitações que estes tipos de sistemas ainda apresentam, surgiram os sistemas híbridos (*Hybrid MT*). São híbridos os sistemas que englobam diferentes componentes dos acima descritos. Por exemplo, há sistemas híbridos que englobam a tradução baseada em regras e a tradução baseada em exemplos, de forma a colmatar os problemas que estes sistemas colocam individualmente. Há ainda aqueles que englobam diferentes fases de diferentes tipos de sistema (Somers, 2003: 518):

> "Other hybrid systems combine rule-based analysis and generation with example-based transfer. "

Estes sistemas híbridos conjugam as vantagens apresentadas pelos sistemas que dispõem de conhecimento linguístico com as dos sistemas estatísticos.

2.2.2 Fases do processo de tradução automática

Segundo Hutchins (2003:506), as ferramentas ao dispor dos sistemas de tradução automática são importadas de outras áreas do processamento automático da linguagem natural:

> «The tools available are familiar from other fields of computational linguistics: the provision of dictionaries with lexical, grammatical, and translational information; the use of morphological and syntactic analysis to resolve monolingual ambiguities and to derive structural representations, the use of contextual information, of semantic features, of case markers, and of non-linguistic ('real world') information to resolve semantic ambiguities.»

Ora, tendo em conta a arquitectura dos sistemas de tradução automática por transferência, o processo de tradução decorre em três fases em comum: a análise, a transferência e a geração ou síntese.

O processo de tradução automática apoia-se numa componente fundamental dos sistemas: os dicionários. Pode dizer-se que os dicionários são um módulo fundamental nos sistemas de tradução automática e a informação que contêm é importante para cada uma das fases do processo de tradução. Segundo Marrafa (2004) «é no Léxico que se codifica toda a informação relativa aos contextos sintácticos e semânticos em que os diversos itens podem ocorrer.» A opção pela designação Léxico prende-se com o facto de a componente dos sistemas que reúne as palavras não ser um dicionário no sentido

tradicional. Contrariamente aos dicionários tradicionais, esta componente pode conter para além de toda a informação presente nos dicionários tradicionais, outro tipo de informação especificamente necessária para o processamento computacional do *input* dos sistemas de tradução automática.

É necessário distinguir entre as características inerentes a uma palavra e as restrições de co-ocorrência que esta impõe. As características inerentes à palavra incluem informações sobre género e número, por exemplo. No entanto, este tipo de informação é bastante elementar e não se aplica a todos os casos, pelo que se torna necessário encontrar uma maneira mais produtiva de organizar a informação referente a uma dada palavra. Neste seguimento, Pustejovsky (1995) propõe três níveis de estruturas: argumental, eventiva e Qualia. De uma forma breve, a estrutura argumental envolve uma palavra predicativa e os complementos que ela selecciona; a estrutura eventiva diz respeito ao tipo de evento relacionado com uma dada palavra; por fim, a estrutura Qualia respeita os atributos semânticos de uma palavra. A título de exemplo, a interacção destes três níveis de estruturas, nomeadamente entre a estrutura argumental e a estrutura Qualia, permite a resolução de questões como a polissemia (Pustejovsky (1995), *apud* Silva (2008)).

A informação gramatical que dá conta das restrições que as palavras impõem pode dividir-se em dois tipos: informação de subcategorização, que diz respeito ao contexto sintáctico em que a palavra pode estar integrada; e as restrições de selecção, que estão dependentes das propriedades semânticas da palavra.

Um exemplo de informação sobre subcategorização é a que indica que, por exemplo, um verbo selecciona um sintagma nominal como objecto directo. Outros itens lexicais exercem selecção de complementos, tais como os nomes e os adjectivos, por exemplo.

No que respeita a restrições de selecção semântica, o objecto directo em *O João abotoou o casaco*, ou seja, *o casaco*, tem a função de TEMA, na medida em que veicula a entidade que é afectada pelo evento denotado pelo verbo. Por sua vez, a semântica de *abotoar* apenas é compatível com um objecto directo [-*humano*].

Nos sistemas de tradução indirecta, os módulos de análise e geração dispõem de dicionários individuais monolingues para a língua de partida e para a língua de chegada. Nestes sistemas, há também léxicos bilingues para a transferência entre as línguas envolvidas no processo de tradução. Na fase de análise, o léxico monolingue da língua de partida dispõe de informação que dá conta das categorias gramaticais, dos traços semânticos, etc., necessária para a análise e desambiguação estruturais. O léxico bilingue usado numa fase intermédia do processo de tradução automática, na transferência entre a língua de partida e a língua de chegada, centra-se sobretudo nas correspondências lexicais entre as línguas envolvidas no processo de tradução.

Por fim, o léxico de que os sistemas de tradução automática indirecta se servem na fase de geração é mais básico do que os léxicos das outras fases do processo de tradução, uma vez que a informação essencial para o processamento computacional do *input* já está presente nos léxicos das fases que precedem a fase de geração.

Vejam-se agora os diferentes níveis que a fase da análise pode englobar: nível morfológico, nível sintáctico e nível semântico.

Na fase da análise morfológica, os sistemas centram-se na identificação do lexema das formas flexionadas, tanto regulares como irregulares, na identificação de formas derivadas (por exemplo, *ligar/desligar*).

Refira-se que a inclusão de um módulo de análise morfológica nos sistemas de tradução automática, torna possível a redução do tamanho dos dicionários e, desta forma, todo o esforço dispendido na sua elaboração e os custos com a melhoria da eficiência do sistema.

Na fase da análise sintáctica, os sistemas dispõem de *parsers* para identificar sequências de categorias gramaticais (nome, determinante, verbo, adjectivo, advérbio, etc.), para reconhecer grupos de categorias (sintagmas, orações, e frases) e para identificar as relações de dependência que se estabelecem entre categorias: por exemplo, a questão da regência verbal ou a questão da modificação.

Ainda na fase da análise sintáctica, resolvem-se problemas que se prendem com o sentido das palavras. Por exemplo, é nesta fase que se descodifica o sentido pretendido para a palavra *light* em inglês: *leve* ou *claro*. De igual forma, é nesta etapa que se resolvem ambiguidades estruturais e diferenças lexicais entre línguas (por exemplo, a questão do verbo *to know* e da sua transferência para o português como *saber* e *ter conhecimento* ou o verbo *to learn* que pode ser traduzido em português como *aprender* e *ficar a saber*). São duas as estratégias frequentemente utilizadas para lidar com estas questões na fase da análise semântica. Por um lado, é pela identificação de traços semânticos que se estabelecem os contextos de ocorrência das palavras. O tratamento destas questões vai servir de *input* para a fase de análise semântica.

Na fase da análise semântica, um *parser* semântico tem a função de dar o sentido global das expressões. Neste seguimento, a análise semântica tem por base o princípio da composicionalidade, pelo qual uma expressão adquire significado pelo sentido das partes que a constituem, assim como as relações que se estabelecem entre elas.

2.2.3 Qualidade dos *outputs* e limitações dos sistemas

Apesar de existirem vários sistemas, com diferentes características e baseados em diferentes paradigmas de tradução, a verdade é que, um *output* que seja inteiramente aceitável e que dispense completamente a pós-edição é ainda uma meta a atingir e não uma realidade.

O presente trabalho centra o seu estudo na análise do desempenho de dois sistemas disponíveis gratuitamente na Internet: um sistema actualmente híbrido, mas que inicialmente se baseava exclusivamente em regras, o Systran, e um sistema estatístico, o Google.

O desenho destes dois sistemas baseia-se em diferentes arquitecturas. O Systran é um sistema que traduz a partir da representação de conhecimento linguístico, em conjugação com uma abordagem estatística. Ou seja, o Systran é um sistema híbrido, que alia a tradução baseada em regras à tradução baseada em estatística (Melero i Nogués, 2006)

Retomando o que já foi aqui descrito em pormenor, na secção 2.2.1 (Tipos de Sistemas), para os sistemas baseados em estatística, saliente-se que o Google é um sistema estatístico, que integra também memórias de tradução. Trata-se de um sistema totalmente orientado para os dados, que, consequentemente, não integra qualquer tipo de regras linguísticas.

A eficácia destes dois sistemas é bastante discutível. Tomem-se para análise os seguintes exemplos de vários tipos de frases – uma frase simples em (6), uma frase complexa em (7), uma frase estruturalmente ambígua em (8) e uma frase com uma anáfora em (9):

(6) O João comeu o bolo.
(7) A Maria pensa que está gorda.
(8) O Pedro trouxe o livro da escola.
(9) A Maria tirou uma fotografia a si própria.

Vejam-se as traduções das frases de (6) a (9) nos dois sistemas, Systran (de (6) (a) a (9) (a)) e Google (de (6) (b) a (9) (b)):

(6) (a) The João ate the cake.

(7) (a) The Maria thinks that she is fat.

(8) (a) Pedro brought a book of school.

(9) (a) The Maria took a photograph proper itself.

(6) (b) The John ate the cake.

(7) (b) Maria thinks it is fat.

(8) (b) The Peter brought a book from school.

(9) (b) Maria took a photograph of herself.

A primeira frase, uma frase simples, sem alterações à ordem básica das palavras, não coloca problemas significativos ao Systran, pelo que o sistema consegue traduzi-la sem dificuldades - veja-se (6) (a). O único problema digno de referência que a tradução

do sistema Systran exibe é a inserção do determinante artigo definido *the* antes do nome próprio *João*. Visto este sistema, alguns casos, ainda apresentar traduções palavra a palavra, o determinante é inserido antes do nome próprio. O que acontece com os restantes *outputs* do Systran, com excepção da tradução em (8) (a). O Google, por sua vez, insere o determinante artigo definido antes dos nomes próprios, em duas frases da quatro frases traduzidas, em (6) (b) e (8) (b) (cf. acima: Google (de (6) (a) a (9) (a) e Systran (de (6) (b) a (9) (b)).

Em (7) (a), temos uma frase complexa, isto é, uma frase que integra uma subordinada completiva. Em português, o sujeito não realizado da oração subordinada é interpretado como co-referente do sujeito da oração subordinante. O sistema Systran, em (7) (a), consegue recuperar o sujeito da frase subordinada. Já o Google, em (7) (b), não resolve esta questão de forma aceitável, isto é, o sistema não consegue resolver essa interpretação e preenche o lugar do sujeito nulo com o expletivo *it* (cf. acima: Google (de (6) (a) a (9) (a) e Systran (de (6) (b) a (9) (b))

Em (8), explora-se um caso de ambiguidade estrutural. A frase permite duas interpretações:

(A) o Pedro veio da escola e trouxe consigo um livro.

(B) o Pedro trouxe um livro que pertence à escola.

Na primeira acepção, temos na frase um verbo de três lugares que selecciona um argumento externo sujeito, um argumento interno com a função de objecto directo e um argumento interno preposicionado com função de locativo de origem. A preposição *de* marca, por conseguinte um locativo de origem/proveniência. Na segunda acepção, a preposição *de* marca um complemento que exprime pertença. As traduções obtidas em ambos os sistemas são diferentes: em (8) (a), o Systran traduz a frase na segunda acepção, ou seja, traduz a preposição *de* por *of*. Por sua vez, o Google, (8) (b), traduz a preposição *de* pela preposição de proveniência/origem - *from*. Embora cada um dos sistemas opte por uma das interpretações e nenhum se revele capaz de produzir ambas as interpretações em inglês, pode considerar-se que o seu desempenho é aceitável na medida em que as frases são aceitáveis e ambas as interpretações são possíveis.

Em (9), propõe-se para análise uma frase com uma anáfora. A expressão *a si própria* refere-se ao sujeito da frase, *A Maria*. O Google, em (9) (b), consegue produzir um *output* aceitável. O sistema reconhece que a expressão anafórica se refere ao sujeito, traduzindo-a por *herself*. Já o Systran, em (9) (a), tem um resultado inaceitável, e os problemas prendem-se com a expressão anafórica *a si própria*. O sistema traduz separadamente *própria* e *si*. *Própria*, a forma de reforço anafórico, é traduzida pelo adjectivo *proper,* e a forma *si* por *itself*. No exemplo em análise, a forma *si* refere-se ao sujeito da frase, um nome próprio feminino, logo a tradução aceitável seria *herself*.

2.2.4 Problemas em Tradução Automática

A tradução realizada por um tradutor humano não é uma tarefa fácil. Os desafios que se colocam são vários e prendem-se, por exemplo, com questões culturais. Há certos aspectos culturais próprios de uma língua que não encontram equivalência directa numa outra língua. Um bom exemplo desses aspectos culturais são as tradições (*Latada dos Estudantes*, por exemplo, para a língua portuguesa). Sendo esta uma tradição que provavelmente não existe noutro país, torna-se difícil conseguir uma tradução, que não possua qualquer tipo de explicação. A gastronomia e até sistemas jurídicos diferentes constituem também desafios de tradução para o tradutor humano. Todos estes aspectos dificultam o processo de tradução, ainda que não seja impossível expressá-los noutra língua, o que pode acontecer por meio de uma explicação mais extensa ou por meio de notas de rodapé, por exemplo. O tradutor humano pode também contar com expressões próprias de uma língua, inexistentes na outra língua e que só podem ser traduzidas por meio de paráfrases ou explicações extensas; exemplos destas expressões podem ser *pain in the ass* ou *cair o Carmo e a Trindade*.

Por sua vez, à tradução automática colocam-se problemas de ordem linguística, como é o caso da resolução de ambiguidades, dependências de longa distância, expressão do tempo, do aspecto e da modalidade. Questões como é o caso da ambiguidade, não constituindo um problema de maior para a tradução humana, são um problema para a tradução automática.

Em muitos casos de ambiguidade, para o tradutor humano, o contexto e o conhecimento do mundo ajudam a captar o sentido. Ora, os sistemas de tradução

automática não incluem informação suficiente para resolver ambiguidades a partir do contexto. Mas, os problemas que se colocam à tradução automática não se esgotam na resolução de ambiguidades. A seguir, analisam-se alguns dos fenómenos mais problemáticos.

Ambiguidade. Um dos principais problemas que se colocam ao processamento automático da linguagem natural é, como já referido, a questão da ambiguidade, seja ela lexical ou estrutural. Muito genericamente, a ambiguidade lexical prende-se com o facto de uma palavra ter mais de um sentido. A ambiguidade estrutural resulta do facto de serem licenciadas relações diversas entre constituintes de uma dada expressão.

No que diz respeito à ambiguidade lexical, torna-se importante precisar os seguintes conceitos: polissemia, homonímia e vaguez.

Entende-se por polissemia a propriedade de algumas palavras terem mais do que um significado e de esses significados estarem relacionados. Um exemplo de uma palavra polissémica é *atingir*.

Quanto ao conceito de homonímia, respeita a associação de dois ou mais sentidos não relacionados entre si a uma dada forma lexical. São exemplos de palavras homónimas os seguintes itens lexicais: *cura* (pároco) e *cura* (acto de *curar*, forma do verbo *curar*) e *dó* (nota musical) e *dó* (comiseração). A forma *cedo* é também um caso de homonímia, que, tal como *cura*, envolve ambiguidade categorial. A forma *cedo* tanto pode ser um advérbio, numa frase como:

(6) A Maria chegou cedo a casa.

cedo pode ser uma forma verbal, da 1ª pessoa do singular do Presente do Indicativo do verbo *ceder*, como na seguinte frase:

(7) Eu cedo sempre às exigências da Maria porque ela é muito teimosa.

Por seu lado, o conceito de vaguez envolve falta de informação contextual. Um exemplo de uma palavra vaga é a palavra *tia*, na medida em que, individualmente, este item lexical não encerra em si a informação necessária para saber se se trata da irmã da

mãe ou da irmã do pai. Isto é, a leitura que se faz da palavra *tia* é uma leitura geral numa frase como:

(8) Ontem fui visitar a minha tia.

Segundo Cruse (1986:51), que apresenta como exemplo a palavra *cousin* para o inglês:

> "We shall say that the word form *cousin* is general with respect to the distinction "male cousin"/"female cousin", […]".

Hutchins (1995:87) refere uma forma de ambiguidade lexical que se prende com a questão da transferência entre línguas – *transfer ambiguity*. Este tipo de ambiguidade acontece quando uma palavra numa língua de partida pode ter mais do que uma tradução equivalente na língua de chegada, como é o caso da palavra *light* em inglês, que pode ter as seguintes traduções para o português: *leve* e *claro*.

Em Arnold *et al.* (1994:110), apresenta-se um exemplo de não correspondência lexical entre o japonês e o inglês no que diz respeito ao termo utilizado para indicar o evento de vestir. A língua inglesa dispõe de dois lexemas (*wear, put on*) para designar o acto/evento de colocar peças de vestuário e calçado no corpo, ao passo que o Japonês dispõe de vários lexemas. Os lexemas japoneses têm no seu significado especificações muito concretas quanto ao objecto usado no corpo. O verbo *kakeru* é usado preferencialmente para óculos, *haku* para sapatos, *kaburu* para chapéus, *hameru* para luvas e *haoru* para casacos. Paralelamente, a língua portuguesa dispõe de itens lexicais para o evento em questão. São eles *vestir*, *usar*, *pôr* e *calçar*.

Esta divergência entre as línguas coloca problemas aos sistemas testados, que não são capazes de reconhecer, por exemplo, que em português, o verbo *calçar* designa apenas o acto de colocar peças de vestuário nas mãos (*luvas*) ou o acto de colocar calçado nos pés (*sapatos*).

Outro tipo de ambiguidade que causa problemas na tradução automática é ambiguidade sintáctica ou estrutural. Este tipo de ambiguidade acontece quando uma frase pode ter duas ou mais interpretações, em virtude da possibilidade de diferentes relações entre os seus constituintes. Veja-se o exemplo, retirado de Chomsky (1965:21):

(9) Flying planes can be dangerous.

Esta frase é ambígua na língua inglesa, pois pode ter duas interpretações:

(A) Pilotar aviões pode ser perigoso.
(B) Aviões a voar podem ser perigosos.

O resultado da tradução desta frase ambígua nos sistemas de tradução automática ilustra o problema que se coloca.

(15) (a) Os planos do vôo podem ser perigosos.

(15) (b) Aviões voando pode ser perigoso.

O Systran, em (15) (a), traduz a frase em (15) com um sentido diferente dos indicados acima. O Google, em (15) (b), traduz a frase no sentido referido em (B). Apesar de os *outputs* serem aceitáveis, os resultados das traduções servem para demonstrar os problemas que este tipo de ambiguidade coloca.

Anáfora/Catáfora As estruturas anafóricas causam também problemas em tradução automática. Uma anáfora é um processo pelo qual uma determinada palavra ou expressão, ou até um elemento elíptico são interpretados em relação a uma entidade anteriormente referida na frase. A catáfora envolve igualmente uma dependência referencial, sendo que o chamado antecedente (expressão que fixa a referência) ocorre em posição posterior à do elemento referencialmente dependente. Veja-se a abordagem a estes conceitos na *Gramática da Língua Portuguesa* (Mateus *et al.*, 2003: 802):

> «Na linguística moderna, o conceito de anáfora não é uniforme, tendendo, de qualquer modo, a ser visto como o processo que consiste em utilizar uma forma linguística ou um vazio para remeter para algo que foi dito anteriormente (o antecedente); nesta visão alargada, a anáfora distingue-se da catáfora, que consiste em remeter para algo que é dito no discurso posterior.»

Refere-se aqui forma linguística na medida em que a relação anafórica entre constituintes de uma frase pode ser feita por meio de pronomes, determinantes, etc., e

vazio, porque «pode ser uma elipse ou um vestígio de um constituinte deslocado» (Mateus *et al*, 2003:802). Distingue-se então entre:

Anáfora nominal[3].

A relação anafórica é feita por uma expressão nominal, como em *A doutora Rita não dá consultas hoje. A doutora está doente.*

Anáfora pronominal.

A retoma é feita através de pronomes (pessoais, possessivos, reflexos, recíprocos) ou vazios e «que, por isso, não têm "referência virtual" em si mesmos» (Mateus *et al*., 2003:803).

Anáfora através de quantificador *todos*, *tudo* e outras expressões de síntese.

É o caso da seguinte frase *Mulheres, crianças, velhos, todos são atingidos pelas minas.*[4]

Anáfora através dos demonstrativos invariáveis *–o* ou *isso*.

Em frases como *As pessoas que fogem aos impostos fazem isso deliberadamente*. e *A Rita quer ser pianista e a amiga quer sê-lo também*.[7]

[3] Dentro da anáfora nominal, distingue-se ainda entre anáfora fiel, quando esta é feita através da repetição do nome, e anáfora infiel, quando não se opera a repetição do nome e se utiliza uma expressão nominal de diferente significado mas com a mesma referência, como em *A Maria faz muita birra. A garota está mal-habituada*.

[4] Exemplos retirados de Mateus *et al.* (2003:803)

Anáfora através de elipse.

Numa frase como *Gosto da voz da Tebaldi mas prefiro a da Callas,* o elemento não realizado na oração subordinada é facilmente recuperado em relação com o seu antecedente na oração subordinante, ou seja, o vazio pode ser preenchido pelo nome *voz*.

Anáfora temporal. O tempo Pretérito Mais-que-Perfeito é exemplo de um tempo anafórico «na medida em que necessita, para a sua localização temporal no passado, de um outro ponto de referência, isto é, um Ponto de Perspectiva Temporal, também passado, que habitualmente se encontra expresso no quadro de uma frase complexa ou de um texto, mas que também se pode reconstruir ou inferir.» (Mateus *et al*., 2003: 161)

A seguir apresentam-se alguns exemplos, em que se ilustra a questão da anáfora:

(10) O João magoou-se com a faca.
(11) O João partiu o vaso da mãe e ela gostava muito dele.
(12) O meu vestido é bonito mas o teu é lindo.

As frases dos exemplos (15) e (16) contêm elementos anafóricos. Em (15), o elemento anafórico é o pronome reflexo *se*, que está em relação anafórica com o sujeito da frase, *O João*. Em (16), há uma estrutura de coordenação. Na frase que constitui o segundo termo da estrutura de coordenação, há dois elementos anafóricos, o pronome *ela* e o pronome *ele* (inserido num sintagma preposicional, com a preposição *de*, que constitui a regência do verbo *gostar*). Estas expressões pronominais estão em relação anafórica com *a mãe* e *o vaso* respectivamente.

Em (17), a relação anafórica estabelece-se entre uma expressão nominal (*vestido*), na frase que constitui o primeiro termo da estrutura de coordenação, e um elemento que não está realizado, na oração coordenada. Esse elemento não realizado é co-referente a *vestido* na oração principal.

Os dois primeiros exemplos ilustram um tipo de relação anafórica, a anáfora pronominal, que é particularmente difícil de processar pelos sistemas de tradução automática. Veja-se então, a seguir, as respectivas traduções dos sistemas para as frases acima enunciadas. No Systran, em (15) (a) e (16) (a), e no Google, em (15) (b) e (16) (b):

(15) (a) *The João was hurted with the knife.

(15) (b) *Joao hurt with the knife.

(16) (a) *The João broke the vase of the mother and it liked it very.

(16) (b) *The John broke the vase of the mother and she loved him.

Em (15) (a) e (15) (b), a tradução do reflexo *se*, em relação anafórica com o sujeito da frase, põe problemas aos dois sistemas. Na tradução do Systran (15) (a), a ambiguidade da forma *se* resulta numa tradução inaceitável, uma vez que o sistema interpreta a forma *se* como sendo a partícula apassivante, numa passiva de *–se*. Na tradução do Google (15) (b), o sistema não traduz o pronome reflexo *se*, o que resulta numa frase sem objecto directo.

Em (16) (a), o Systran tem um desempenho inadequado no estabelecimento da relação entre o pronome (*ela*) e o seu antecedente (*mãe*), traduzindo o pronome pelo expletivo *it*. No que diz respeito à interpretação do pronome *ele* (*dele*), co-referente com *vaso*, o sistema, ao contrário do Google, tem um desempenho aceitável.

Em (16) (b), o Google obtém um resultado inaceitável na tradução do segundo pronome, contraído com a preposição *de*, traduzindo-o por *him*, quando este se refere a um objecto. Este desempenho deve-se à ambiguidade na interpretação do pronome, uma vez que este tanto pode ter como antecedente o sujeito da frase como o objecto directo da primeira oração, como pode não ter antecedente discursivo. Para um tradutor humano, a questão da ambiguidade desta frase não se coloca, devido ao conhecimento do mundo que vai permitir descodificar com sucesso a entidade a que o pronome pessoal *ele* se refere, uma vez que esse conhecimento permite perceber que *ele* tem preferencialmente como antecedente *vaso*. No caso da segunda cadeia anafórica no

mesmo exemplo, o Google, contrariamente ao Systran, resolve de forma aceitável a relação de anafórica entre *a mãe* e *ela*.

Na frase em (17), a questão fundamental relaciona-se com o pronome possessivo *teu*. Na frase portuguesa, o pronome, na segunda frase da estrutura de coordenação, remete para o substantivo *vestido* da primeira frase da estrutura de coordenação. Tome-se para consideração as traduções automáticas da frase em (17):

(17) (a) *My dress is pretty but yours he is pretty,

(17) (b) *My dress is beautiful but your is beautiful.

O pronome possessivo *your*, a que o sistema recorre em (17) (b), concorda em género e número com o sintagma nominal que representa, ou seja, num exemplo como (17) este pronome não é o adequado[5].

Expressões multipalavra e Expressões Idiomáticas. Outra dificuldade que se coloca à tradução automática relaciona-se com as expressões idiomáticas. De uma forma muito breve, entende-se por expressão idiomática um grupo de palavras que sofreram um processo de cristalização e que devem ser tomadas como unidades de sentido, não tendo valor composicional pois não podem ser entendidas tendo em conta o significado individual das suas partes. O quadro seguinte reúne algumas expressões idiomáticas em várias línguas:

Português	**Inglês**	**Italiano**	**Francês**	**Alemão**
Pôr a carroça à frente dos bois	Ants in one's pants	Ragazza acqua e sapone	Parler à un mur	Äpfel mit Birnen vergleichen
Agarrar com unhas e dentes	Any Tom, Dick or Harry	Prendere un ganchio	Être comme pain et beurre	Haare auf den Zähnen haben

Quadro (1) Diferentes expressões idiomáticas em várias línguas.

[5] Ao invés, o uso do pronome possessivo *yours* é muito mais adequado, uma vez que remete para o objecto possuído.

A tradução automática de expressões idiomáticas é muito difícil e os resultados são quase sempre inaceitáveis, principalmente nos casos em que a expressão idiomática não é partilhada pelas línguas envolvidas no processo de tradução. Há casos em que as línguas partilham a expressão idiomática, como é o caso de *vender gato por lebre*, que em inglês encontra correspondência em *sell cat for hare*. Nos casos em que as expressões idiomáticas de ambas as línguas divergem, a tradução automática falha, pelo facto de as expressões serem objecto de decomposição. Tomemos para análise os seguintes exemplos:

Expressão idiomática	**Tradução do Systran**	**Tradução do Google**
She has ants in one's pants	*Tem formigas em suas calças.	*Ela tem de formigas em uma calça.
Mon frère et moi, nous sommes comme pain et beurre.	*O meu irmão e mim estão como pão e manteiga.	*Eu e meu irmão são como pão e manteiga.

Quadro (2) Expressões idiomáticas em inglês e francês e as respectivas traduções nos sistemas de tradução automática.

As traduções do quadro (2) mostram que os sistemas traduzem as expressões idiomáticas pelo sentido de cada uma das palavras. Tanto assim é que os sistemas não traduziram as expressões idiomáticas do francês e do inglês para as expressões idiomáticas equivalentes em português: *ter bichos-carpinteiros no rabo*, para a primeira, e *ser como unha e carne*, para a segunda. E se este é um problema fácil de resolver, na medida em que as expressões idiomáticas são codificadas como tal no dicionário, se tivermos em conta o quadro (2), fica claro que os sistemas não possuem, nos seus dicionários, entradas para grande parte das expressões idiomáticas das línguas que integram. Desta forma, as expressões idiomáticas constituíram, na maior parte dos casos, um problema para a tradução automática nos sistemas usados.

Colocações. O conceito de colocação respeita a unidades linguísticas que envolvem a co-ocorrência preferencial de certas expressões. São exemplos:

(18) tirar uma fotografia

(19) fumador compulsivo

(20) throw a party

O mesmo será dizer que a possibilidade de o nome *fotografia*, por exemplo em (18), ocorrer com o verbo *tirar* é bastante mais elevada na língua portuguesa, do que a possibilidade de ocorrer com *produzir* ou *fazer.* O mesmo acontece com os elementos que formam os restantes exemplos nas respectivas línguas.

Veja-se o resultado da tradução destes exemplos no Systran. A frase (18) foi traduzida para francês, a frase (19) para inglês e a frase (20) para português, de forma a ilustrar a questão no âmbito da tradução entre mais do que duas línguas. A seguir apresentam-se os resultados:

(13) (a) * enlever une photografie
(14) (a) * compulsory smoker
(15) (a) * para jogar um partido

As traduções, de (18) a (20), no sistema de tradução automática Systran mostram claramente que os sistemas exibem um desempenho inaceitável na tradução de combinatórias lexicais. O Systran traduziu as combinatórias sem ter em conta que constituem grupos fixos de palavras na língua de chegada. Por exemplo, *fumador compulsivo* em português é *heavy smoker* em inglês.

3. Linguagens Controladas

Neste capítulo, aborda-se o tema das linguagens controladas, em termos da sua história, tipos e fins a que se destinam, discutindo-se o impacto que estas podem ter na tradução automática.

Em 3.1, traça-se uma breve panorâmica das linguagens controladas que engloba a sua definição, estado da arte, aplicações e história.

Em 3.2, abordam-se as virtualidades das linguagens controladas em termos das suas vantagens e desvantagens e do seu impacto na tradução automática.

3.1 Breve panorâmica e estado da arte

Em Arnold *et al.* (1994: 147), uma linguagem controlada é apresentada como:

«a form of language usage restricted by grammar and vocabulary rules.»

Uma linguagem controlada não é, pois, uma linguagem artificial, mas uma forma controlada/simplificada da linguagem natural por meio de regras gramaticais e de um vocabulário reduzido e normalizado. Distinguem-se aqui as linguagens controladas das línguas da especialidade. Estas últimas restringem-se a uma área do saber - como por exemplo, a Medicina, o Direito, a Linguística, entre outras -, tendo, por conseguinte, um vocabulário particular.

Para uma classificação das linguagens controladas quanto à sua utilização, distinguem-se aquelas que se destinam ao utilizador humano (*human-oriented controlled languages*), quer seja um tradutor ou outro tipo de técnico; e aquelas que se destinam a ser processadas pela máquina (*machine-oriented controlled languages*). As primeiras visam esbater problemas de compreensão que a linguagem natural pode colocar ao utilizador humano, enquanto as segundas visam eliminar o mais possível os obstáculos que um texto em linguagem natural pode colocar aquando do seu processamento computacional, em particular na tradução automática.

Quanto à sua aplicação prática, as linguagens controladas têm muita importância na redacção da documentação técnica de empresas transnacionais. Entende-se por documentação técnica toda a documentação usada no interior de uma empresa pelos trabalhadores (textos com informações, procedimentos e instruções) e toda a documentação exterior à empresa e que se destina ao utilizador dos produtos (manuais de instrução).

Rascu (2006) coloca desta forma as vantagens da escrita de documentação técnica em linguagem controlada:

> "Document production, in general, and technical writing in particular are inevitably linked to the concept of text optimisation. From a linguistic point of view, standard language, consistent use of terminology, unambiguous linguistic structures are required to increase the readability, comprehensibility and translatability of technical documentation."

Por um lado, recorre-se às linguagens controladas na Indústria de qualquer tipo. No âmbito de uma empresa transnacional, estas permitem a uniformização da documentação e da terminologia, o que resulta em textos mais claros e de qualidade mais elevada, que vão poder ser reutilizados, o que resulta na optimização da documentação técnica que Rascu (2006) refere na citação acima apresentada. A escrita em linguagem controlada permite que os textos disponíveis para o grande público (manuais de instrução) se tornam mais claros e menos equívocos, o que, muito provavelmente, vai resultar num nível de satisfação muito mais elevado por parte do consumidor.

As linguagens controladas têm influência comprovada na tradução automática da documentação do tipo de empresas referido. Segundo Guimarães (2008), a necessidade de tradução prende-se com a globalização dos mercados, ou seja, é cada vez maior a quantidade de documentos técnicos a serem traduzidos nas diferentes línguas dos diferentes mercados onde os produtos são comercializados. A autora refere ainda que a redacção de documentação na língua materna do público-alvo é especialmente importante tendo em conta as estratégias de *marketing*, pois os consumidores aceitam melhor os produtos que melhor entendem. A tradução dos produtos para a língua materna de cada mercado está intimamente relacionada com o conceito de *product*

liability[6]. A localização de um produto ou serviço é um bom exemplo da necessidade de tradução no mundo globalizado actual. A localização pressupõe o processo de adaptação de um produto ou serviço ao mercado onde será comercializado, tendo em conta a língua e as especificidades culturais do mercado alvo. Esta necessidade de tradução conjuga-se com as vantagens que a tradução automática oferece: rapidez e baixos custos.

Um exemplo concreto de uma aplicação prática das linguagens controladas é a Aeronáutica, mais concretamente na redacção da documentação referente à manutenção de motores e equipamentos e na redacção de procedimentos. Neste caso, o seu uso está intimamente relacionado com o critério da segurança, na medida em que a redução de ambiguidade e, consequentemente, a clareza obtida na escrita em linguagem controlada, permite que se evitem mal-entendidos e erros, que estão, muitas vezes, na origem de acidentes. Um exemplo de como a ambiguidade na linguagem natural pôs em causa a segurança no âmbito da Aeronáutica é descrito por Allen (2009, *apud* Banjar, 2001). O caso ocorreu durante um voo em que o avião se deslocava numa zona de visibilidade reduzida. Perante esta situação, o controlador de tráfego aéreo deu indicações ao piloto para que virasse à esquerda, proferindo em inglês a ordem 'vira à esquerda' (*turn left)*. O piloto acatou a indicação e repetiu a informação que lhe foi dada – 'estou a virar à esquerda' (*I am turning left*). O controlador de tráfego aéreo proferiu então a expressão de ênfase 'Certo' (*right*). Ora, a ambiguidade entre a expressão de ênfase *right* e o substantivo *right, que* significa 'direita' (o oposto de esquerda) esteve na origem de um acidente, uma vez que o piloto pensou que devia virar à direita (*right*).

A Airbus, o maior fabricante de aviões comerciais, desenvolveu um projecto, em 1998, que visava a criação de uma linguagem controlada para a redacção de mensagens de alarme (Spaggiari, 2005). A par da aplicação das linguagens controladas, na área da Aeronáutica, para a redacção de mensagens de alarme, redigem-se também manuais de manutenção e manuais de procedimentos. O ASD Simplified Technical English (antes designado por AECMA Simplified English) é a linguagem controlada em que se redigem todos os documentos de referência de manutenção na indústria aeroespacial. O

[6] The responsibility of a manufacturer or vendor of goods to compensate for injury caused by defective merchandise that it has provided for sale. (The Free Dictionary Online)

Français Rationalisé, uma forma de linguagem controlada para a língua francesa, foi criado para facilitar a tradução de textos do francês para o AECMA Simplified English, bem como para melhorar a facilidade de leitura para os falantes do francês.

Também na área do Direito se recorre às linguagens controladas, o que se justifica pelas características do texto jurídico, isto é, pelo facto de este tipo de texto conter uma linguagem muito recorrente. Na área do Direito das Obrigações, a própria estrutura das frases de determinado tipo de contrato, assim como a estrutura do próprio texto mantém-se de contrato para contrato. Desta forma, tendo em conta o carácter repetitivo da linguagem jurídica, torna-se viável a criação de uma linguagem controlada para o texto jurídico, de forma a simplificar este tipo de texto e a torná-lo passível de ser traduzido automaticamente. Por outro lado, a Globalização assume peso no Direito. O Direito Internacional pressupõe a troca de informação jurídica a nível global, o que implica a necessidade de uma linguagem mais simplificada, para facilitar a comunicação entre pessoas que possuam diferentes línguas maternas.

Pace (2009) descreve uma aplicação de uma forma de linguagem controlada no domínio do Direito, nomeadamente na redacção de contratos, pelas razões anteriormente enunciadas. Na introdução do artigo, justifica-se da seguinte forma o recurso à linguagem controlada:

> "Natural language contracts specifying obligations between contracting parties, are so complex that we are generally forced to rely upon costly legal specialists for their formulation and analyzing their implications. This motivates a controlled language for contracts, rich enough to be useful in a range of real applications, simple enough for ordinary people to understand, and precise enough to be amenable to automated methods of reasoning."

Na Medicina, a criação de terminologias controladas facilita a comunicação entre as pessoas (médicos, enfermeiros, etc.) em, por exemplo, situações de crise em zonas de guerra onde médicos de várias nacionalidades trabalham em conjunto e precisam de comunicar entre si de forma rápida e eficaz (Castilla *et al.*, 2005). Neste campo há projectos para desenvolver uma aplicação de uma linguagem controlada – Attempto Controlled English - para a redacção de protocolos de prática clínica - clinical practice guidelines - (Shiffman, 2009). Outro exemplo da aplicação de linguagens

controladas ao campo da Medicina é a criação de um *thesaurus*[7] pelo Center for Disease Control com o seguinte objectivo (Bunker, 2005):

> «to support the automated indexing and retrieval of public health content on the CDC web site.»

A par das áreas referidas até aqui, verifica-se ainda o recurso às linguagens controladas, por exemplo, nas comunicações marítimas. Neste âmbito, Campa Portela e Valle (2006) descrevem a necessidade de se recorrer a uma linguagem simplificada, visando obter um nível de segurança mais elevado:

> "El conocimiento de la lengua inglesa en el ámbito marítimo, además de ser fundamental para sumarse y contribuir a los avances y al desarrollo de la tecnología, revierte en cuestiones de seguridad de tal forma que un conocimiento insuficiente de inglés marítimo entre el personal de a bordo y el personal en tierra tiene implicaciones sobre la seguridad del buque y sobre la organización de los negocios marítimos. […] La Organización Marítima Internacional (OMI) es consciente de esta situación, y reconociendo su responsabilidad para con el mantenimiento de una navegación y un comercio marítimo seguros, ha intentado desde hace muchos años mejorar las comunicaciones verbales en este ámbito apostando por una normalización de uso a nivel internacional, finalmente materializada en las FNCM."

Também no âmbito dos sistemas de informação geográfica se desenvolvem as linguagens controladas (Mador-Haim, 2008).

O sistema TAUM[8]-METEO foi criado em 1974-5 pela Universidade de Montreal e é um sistema de tradução automática especializado em boletins meteorológicos que traduz do inglês para o francês. Este sistema obtém resultados muito satisfatórios, que podem ser explicados pelo facto de os boletins meteorológicos apresentarem um «vocabulário simples e reduzido e uma sintaxe elementar, quase telegráfica» (Schwitter, 1998:54-55) e, desta forma, se assemelharem a uma linguagem controlada.

No mundo actual, a comunicação à escala global pode ser dificultada pela diversidade de línguas. As linguagens controladas na Internet facilitam a troca de informação em grande escala, pois tornam determinadas línguas muito mais acessíveis a

[7] Um *thesaurus* é uma listagem controlada de conceitos claramente especificados, em termos das relações que se estabelecem entre si, tendo em vista a padronização dos referidos conceitos.
[8] Traduction Automatique de L'Université de Montréal

falantes não nativos. Em consequência, têm vindo a ganhar uma importância cada vez maior, estando a sua aplicação na Internet intimamente relacionada com a globalização. O seu uso verifica-se especificamente no comércio em linha, em catálogos de produtos (Kaji, 1999). Outra aplicação das linguagens controladas no domínio da Internet é a Web Semântica (*semantic Web)*, uma extensão da Internet convencional, que permite uma utilização mais produtiva da Internet. A Web semântica permite uma «estruturação de informação na descrição da mesma através de metadados ou no recurso a agentes inteligentes para apoiar o utilizador na realização de determinadas tarefas.» (Gonçalves, 2009). Desta forma, a procura de informação em motores de busca tem resultados mais eficientes e úteis. A informação na Web Semântica está expressa para facilitar a sua compreensão por humanos e computadores (Pool, 2006).

Lehtola (1998), em *Controlled Language Technology in Multilingual User Interfaces*, apresenta uma ferramenta que alia as linguagens controladas à Tradução Automática na Internet. O Webtran é caracterizado como:

> "a generic CL [Controlled Language] translation software intended for building multilingual services on Internet"

Este *software* está disponível em linha para duas aplicações. Uma primeira, na descrição dos produtos numa página electrónica de venda de roupa por catálogo. Neste âmbito, Lehtola (1998) afirma:

> "with the translation software, the original catalogue needs to be maintained in only one language. Accurate translations can be provided in real time for customers with other languages"

Uma segunda aplicação do *software* Webtran está na recuperação de informação [*Information Retrieval*] no motor de busca disponível em linha. Nesta aplicação, o utilizador tem a possibilidade de efectuar a sua busca numa língua A tendo em vista obter os resultados dessa busca em bases de dados em diferentes línguas.

Em termos históricos, surge nos anos 30 o BASIC[9] English, a primeira forma de linguagem controlada, criada por Charles K. Odgen, para ser a linguagem internacional do comércio e da ciência. A criação desta linguagem controlada prendia-se com o objectivo de facilitar a aprendizagem do inglês, tornando-o mais acessível a um falante

[9] British American Scientific International Commercial

não nativo e melhorando os níveis de compreensão para os falantes nativos (Schwitter, 1998:58). Consistia num léxico de proporções limitadas e de algumas regras que tratavam questões de morfologia, como a flexão e a formação de palavras. Não chegou, no entanto, a ser utilizado, pois revelou-se ineficaz, tal como constatado em (Nyberg *et al.*, 2003:245):

> "However, Basic English was never suitable for any practical purpose and therefore never used in the industry."

Ainda assim, apesar de ter sido posta de parte devido à sua inadequação prática, esta primeira linguagem controlada foi muito importante para o desenvolvimento de outras, servindo de ponto de partida para a criação de linguagens controladas que vieram a surgir mais tarde, nomeadamente, o Caterpillar Fundamental English (CFE), desenvolvido na década de 60 pela empresa Caterpillar[10].

Na base da criação do Caterpillar Fundamental English estava o propósito de reduzir os custos na produção de documentação técnica e o intuito de evitar a necessidade de tradução, pois a compreensão e a facilidade de leitura que ela proporcionava tornavam os textos acessíveis a engenheiros e mecânicos que não tinham o inglês como língua materna. O Caterpillar Fundamental English partilhava do vocabulário base do Basic English, que, no entanto, foi alargado para ir ao encontro das necessidades da empresa.

Entretanto, já na década de oitenta, a Caterpillar abandona o Caterpillar Fundamental English e desenvolve o Caterpillar Technical English (CTE). Esta última linguagem controlada estava orientada para a redacção de textos que se destinavam a ser traduzidos automaticamente. O Caterpillar Technical English esteve na origem de outras

[10] Mitte der 60er Jahre nahm die Firma Caterpillar Tractor Company (USA) das Konzept von BASIC English auf und wandte es auf einen engeren Diskursbereich an. Konkret wurden technische Handbücher mit Hilfe der Sprache Caterpillar Fundamental English (CFE) - einer Variante von BASIC English - produziert. (Schwitter, 1998)

Em meados da década de sessenta, a firma norte-americana Caterpillar Tractor Company serve-se do conceito do Basic English, aplicando-o a uma área mais restrita. Mais concretamente, produziu-se documentação técnica com a ajuda do Caterpillar Fundamental English, uma variante do Basic English. (Tradução minha)

linguagens controladas desenvolvidas por outras empresas, como o International Language for Serving and Maintanence (ILSAM), criado pela empresa Smart, AI.

O ILSAM está também na origem de várias linguagens controladas utilizadas pelas seguintes empresas Alcatel Bell, com o Controlled English; IBM, com o Easy English; Rank Xerox, com o Multilingual Customized English; Perkins Engines, com o Pace; Ericsson, com o Ericsson English (Dervisevic e Steensland, 2005:78).

A Scania, empresa sueca de fabrico de tractores, desenvolveu o Scania Swedish, que é hoje em dia a linguagem controlada utilizada na redacção de manuais de instrução desta empresa e é também a língua fonte utilizada na tradução automática da documentação da empresa (Almiqvist e Sagvall, 1996).

O KANT Controlled English foi especificamente desenvolvido para servir de *input* ao sistema de tradução automática KANT[11]. Este sistema destina-se exclusivamente à tradução de documentação técnica, daí que se tenha criado uma linguagem controlada para a redacção dessa documentação (Mitamura e Nyberg, 1995).

Ainda na década de setenta, mais exactamente em 1979, a Association of European Airlines (AEA) encarregou a Association Européenne des Constructeurs de Matériel Aérospatial (AECMA) de investigar sobre a compreensibilidade dos documentos de manutenção dos aviões. Na década de oitenta, mais concretamente em 1986, começa a usar-se o AECMA Simplified English. Actualmente, e depois da fusão, em 2004, entre a European Defense Industries Group (EDIG) e a Association of the European Space Industry (Eurospace) para a criação da Aerospace and Defense Industries Association of Europe (ASD), esta linguagem controlada passou a designar-se ASD Simplified Technical English e é hoje em dia a linguagem controlada em que se redige a documentação de manutenção da indústria aeronáutica. Esta linguagem controlada destina-se sobretudo ao utilizador humano e é também aquela que um maior número de empresas utiliza.

Outra linguagem controlada é o Attempto Controlled English (ACE), que foi criada a partir de uma investigação levada a cabo na Universidade de Zurique. Esta

[11] The KANT system (Knowledge-based, Accurate Natural-language Translation) has been primarily targeted towards the translation of technical documents written in CL. KANT has been developed for multilingual translations of heavy equipment documentation and is currently in production use for French and Spanish translations. (Mitamura, 1999)

linguagem controlada destina-se sobretudo a redigir especificações de requisitos para programas de computador (Schwitter, 1998:91). Em Fuchs e Schwitter (1995), justifica-se a aplicação do ACE na redacção de especificações de requisitos da seguinte forma:

> "Writing specifications for computer programs is not easy since one has to take into account the disparate conceptual worlds of the application domain and of software development. To bridge this conceptual gap we propose controlled natural language as a declarative and application-specific specification language. "

Segundo Wyner *et al.* (2009), faz-se uso de mais de quarenta linguagens controladas que cobrem línguas como o inglês, o esperanto, o francês, o alemão, o grego, o japonês, o mandarim, o espanhol e o sueco. Em Pool (2006), encontra-se referência a linguagens controladas para o chinês (Controlled Chinese), para o grego (Controlled Modern Greek), para o espanhol (interNOSTRUM Controlled Spanish), para o alemão (MULTILINT e Siemens-Dokumentationsdeutsch) e para o japonês (Plain Japanese). O quadro 3, apresentado a seguir, elenca um conjunto de empresas, bem como as linguagens controladas de que fazem uso:

Empresa	**Linguagem Controlada**
Alcatel	Controlled English Grammar (COGRAM)
Avaya	Avaya Controlled English (ACE)
Caterpillar	Caterpillar Technical English (CTE), Caterpillar Fundamental English (CFE)
Dassault Aerospace	Français Rationalisé
Ericsson	Ericsson English
General Motors (GM)	Controlled Automotive Service Language (CASL)
IBM	Easy English
Kodak	International Service Language
Nortel	Nortel Standard English (NSE)
Océ	Controlled English
Siemens	Siemens DokumentationsDeutsch

Scania	Scania Swedish
Sun Microsystems	Sun Controlled English

Quadro 3 Empresas e suas linguagens controladas

3.2 Virtualidades das linguagens controladas

> "if it were possible to dictate to people how they should write or speak, simply for the sake of making machine translation cheaper or easier, we could end up by making it more difficult for them to express themselves in their own language then it would be for them to learn a second language and use it." (Arthern citado por Gebuers, 1991)

Na subsecção anterior, evidenciaram-se as vantagens que as linguagens controladas oferecem. Ainda assim, apesar disso há quem aponte uma série de desvantagens.

A primeira desvantagem apontada diz respeito à redacção dos textos em linguagem controlada, dado que ao redactor é exigido que tenha formação específica para escrever em linguagem controlada. A formação pode ser morosa e dispendiosa, e pode haver relutância por parte dos autores/redactores em aprender a escrever em linguagem controlada (Devisevic e Steensland, 2005:62-63). Os custos que podem advir da implementação de uma linguagem controlada apenas se justificam pela dimensão da empresa que a vai usar.

O desenho e implementação de uma linguagem controlada requerem um esforço muito grande de colaboração entre entendidos em determinada área, os redactores de documentação técnica e os utilizadores. A importância desta colaboração justifica-se pelo facto de uma simplificação extrema poder excluir determinadas nuances de sentido que o redactor do documento técnico tenha querido transmitir (Kittredge, 2003: 443).

O próprio processo de escrita pode ser bastante lento e tem de ser auxiliado por ferramentas de apoio à redacção em linguagem controlada (*controlled language checkers*), que alertam o autor para uma palavra ou estrutura que não esteja conforme às regras da linguagem controlada. Estas ferramentas englobam verificadores de terminologia, de gramática, de estilo. Ao dispor do autor de textos em linguagens controladas estão também as memórias de autor (*authoring memory*), estas memórias funcionam como as memórias de tradução, tal como Brockmann (1997:10, *apud* Banjar) descreve:

> «the more controlled a source text, the more efficient these tools will be in the translation process. In the medium term, they will also be adapted for source-text authoring. This means that the writer will be able to re-use his text or her own material using an 'authoring memory', thus increasing consistency in the source language»

Como é óbvio, o autor tem de ter atenção redobrada em relação às regras, o que reduz a rapidez de escrita em cerca de 20% (Devisevic e Steensland, 2005: 62). No entanto, pode contrapor-se a esta ideia o facto de o trabalho de revisão do *output* da tradução tendo como língua de partida uma forma de linguagem controlada ser substancialmente reduzido.

Outra desvantagem apontada por alguns autores é a falta de criatividade que a escrita em linguagem controlada impõe. A simplicidade das estruturas permitidas pelas linguagens controladas pode levar os autores/redactores a não disporem de meios de expressão alternativos, já que a escrita em linguagem controlada é muitas vezes quase telegráfica. Isto pode também ser um obstáculo à fluência da escrita.

Toda esta argumentação perde peso, quando confrontada com as inúmeras vantagens que a escrita de documentação em linguagem controlada tem demonstrado. Se, por um lado, estas desvantagens se verificam, de facto, por outro lado, as vantagens tornam-se mais do que evidentes. Aliás, a investigação dedicada às linguagens controladas mostra bem o empenho e a confiança que inúmeras empresas depositam nelas, através do desenvolvimento de linguagens controladas próprias.

E se há quem argumente com o facto de as linguagens controladas deixarem pouco espaço para a criatividade, podemos contrapor a esta ideia as inúmeras vantagens

que a redução da ambiguidade pode ter para a tradução automática e para a compreensão dos textos por parte dos utilizadores humanos.

O argumento da falta de criatividade que as linguagens controladas impõem também se esbate quando pensamos no tipo de texto mais indicado para este tipo de escrita. O mesmo será dizer que a escrita em linguagem controlada se adequa sobretudo a textos técnicos. Ora, o texto técnico é extremamente específico, pelo que não requer grandes rasgos de criatividade, quer-se sim um texto inequívoco.

3.2.1 Impacto das linguagens controladas na tradução automática

Uma forma de melhorar o *output* da tradução automática é actuar sobre o próprio texto a traduzir de forma a adaptá-lo às características dos sistemas de tradução automática.

A pré-edição de um texto a traduzir resolve alguns problemas da tradução automática, por exemplo, ao eliminar as ambiguidades (de qualquer natureza), as palavras que o sistema possa não reconhecer e estruturas difíceis de processar computacionalmente.

Note-se que a pré-edição visa um *output* melhorado na tradução automática, não tem como objectivo, contudo, obter uma tradução perfeita.

No âmbito da tradução automática, as linguagens controladas são a forma mais radical de adaptar o texto de partida, uma vez que as alterações efectuadas são mais abrangentes do que numa fase de pré-edição. Estas podem abranger áreas como o léxico, a sintaxe, ou a semântica. O estilo, que engloba todos os fenómenos que não são passíveis de ser prescritos pelas regras da gramática, também entra no processo de controlo/simplificação.

Uma lista de regras para a escrita em linguagem controlada contém prescrições quanto ao vocabulário permitido para a redacção de textos. Permitido, porque o vocabulário é normalizado, para se conseguir uma coerência terminológica e lexical em todos os textos, e limita-se a um conjunto de palavras para evitar problemas de ambiguidade lexical.

A nível sintáctico, operam-se alterações importantes para o processamento computacional do texto. Assim, procura-se preterir determinadas construções em favor de uma simplificação das estruturas a utilizar. Estruturas problemáticas são, por exemplo, a coordenação, a subordinação, ou a coesão referencial.

A nível semântico, são de evitar estruturas e palavras que dêem origem a mais do que uma interpretação.

A nível estilístico, por exemplo, as enumerações não devem estar organizadas num corpo de texto, mas sim de forma tabular, em listas de itens. Veja-se, a seguir, um exemplo ilustrativo, retirado de Disborg (2007: 70). Em (A), temos um exemplo de uma enumeração feita em corpo de texto:

(A)

The firewall is the world's first key-upgradeable integrated security appliance. Its Intelligent Layered Security architecture delivers multiple layers of protection that work together to detect and block threats from attacking your network. Stateful firewall, VPN, intrusion prevention, application filtering, spam blocking, and content filtering are all integrated into single appliance and managed through a common interface.

Em (B), temos a mesma informação dada na forma de uma listagem, retirada da linguagem controlada Simplified Technical English:

(B)

The firewall has theses components:

- A stateful firewall
- A VPN
- A protection against instrusion
- A software program filter
- A spam filter
- A content filter.

A seguir, tome-se para análise a seguinte regra, mais uma vez recolhida do guia de escrita da linguagem controlada Simplified Technical English, retirado de Disborg (2007: 64):

(C) *Secção 2, Regra 2.1 Do not make clusters of more than three nouns.*

- Use hyphens to show the relationship between the most closely related words

- Explain the noun cluster. Then, if possible, use a shorter name after the initial explanation

Esta regra concerne a uma questão que se coloca à tradução automática. Sempre que se introduzam para tradução grupos de palavras sucessivas e interdependentes, os sistemas produzem resultados inaceitáveis. Por exemplo, atente-se na seguinte expressão:

(D) *Successful machine repairing tools*.

Esta é uma expressão que engloba mais do que uma palavra, o adjectivo *successful* e *machine reparing* modificam o nome *tools*. O Google traduz da seguinte forma:

(D) (b) *Máquina de sucesso reparar ferramentas.*

Uma forma de controlar esta expressão para a tradução automática seria a seguinte:

(D.1) successful tools for repairing machines

(D.1) obtém um melhor resultado, como se pode verificar pela tradução no Google:

(D.1) (b) instrumentos eficazes para a reparação de máquinas.

De uma forma geral, o resultado da tradução automática de textos redigidos em linguagem controlada é bastante aceitável, pelo que pode, muitas vezes, dispensar a pós-edição dos textos. Num sentido mais prático, o recurso às linguagens controladas prende-se com a redução dos custos da tradução, com o tempo gasto na tradução de grandes quantidades de texto e com a obtenção de *outputs* com bastante qualidade (Hutchins (1999, 1995), Mitamura (1999).

Em Akawa *et al.* (2007), analisa-se a relação entre as linguagens controladas, a qualidade da tradução automática e a necessidade de pós-edição, que se coloca nos seguintes termos:

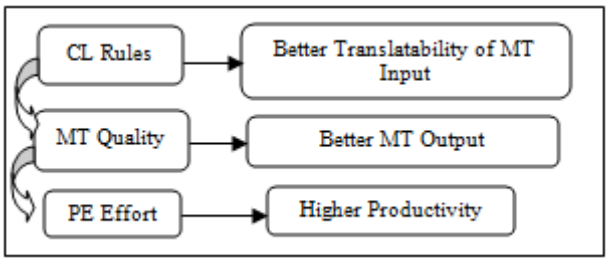

Figura 2 Linguagens controladas, qualidade do *output* em tradução automática e pós-edição (retirada de Akawa *et al.* (2007))

A análise parte da aplicação de um conjunto de regras ao *input* do sistema de tradução automática baseado em estatística MSR-MT da Microsoft. Os autores chegam à conclusão que a escrita em linguagem controlada tem implicações positivas na qualidade do *output* da tradução automática, como referem Aikawa *et al.* (2007):

> "For three of the four languages (Chinese, French and Dutch), the differences are statistically significant and support the hypothesis that applying CL rules to MT input has a positive effect on translation."

Na página electrónica www.controlledenglish.com encontram-se várias informações sobre a linguagem controlada Simplified Technical English, sendo dado a conhecer um caso concreto que ilustra bem o impacto das linguagens controladas na tradução automática. Desta forma, tendo para tradução para várias línguas, um manual com cerca de 300 páginas e cerca de 54000 palavras, a escrita em linguagem controlada permitiu uma redução em cerca de 15% no número de palavras e uma redução total de 45% nos custos de tradução.

4. Contributo para um Português Controlado

Neste capítulo, evidencia-se a utilidade das linguagens controladas na tradução automática. Parte-se de uma análise dos *outputs* de sistemas de tradução automática disponíveis em linha, com vista à captação dos principais problemas que se colocam à tradução automática, de forma a ilustrar o papel efectivo das linguagens controladas na resolução desses problemas. Com base nessa análise, propõem-se regras para controlar/simplificar esses fenómenos, tendo sempre em vista a melhoria dos *outputs.*

Em 4.1, descrevem-se alguns fenómenos que resultaram em *outputs* de tradução automática inaceitável. A partir dessa análise, os mesmos fenómenos são tratados do ponto de vista do seu controlo/simplificação. Subsequentemente, testam-se os resultados, através da tradução automática de texto a que se aplicaram os mecanismos de controlo/simplificação propostos.

Em 4.2, apresentam-se as regras de controlo/simplificação que resultaram da descrição apresentada em 4.1.

4.1 Tópicos de análise

4.1.1 Expressões verbais não atómicas vs. Expressões verbais atómicas

Nesta secção, descrevem-se os problemas que se põem à tradução automática de expressões verbais complexas do português europeu, digamos informalmente, que podem comutar com expressões verbais atómicas, como se ilustra:

(i) fazer (uma) investigação / investigar

(ii) ter receio /recear

(iii) fazer queixas / queixar-se

(iv) dar um presente a alguém / presentear alguém

(v)) dar uma contribuição / contribuir

(vi) ter influência / influenciar

Tomemos como exemplo (19), bem como as correspondentes traduções pelos sistemas Systran e Google, respectivamente:

(19) O cientista fez uma investigação sobre os sistemas de tradução automática.

(a) The scientist made an investigation on the systems of automatic translation.

(b) The scientist made a research on machine translation systems.

(19.1) O cientista investigou os sistemas de tradução automática

(a) The scientist researched the systems of automatic translation.

(b) The scientist investigated the automatic translation systems.

Na tradução da expressão complexa do exemplo (19), ambos os sistemas traduziram *fazer*, um verbo semanticamente fraco, por *to make*. A tradução da expressão verbal não atómica é aceitável, tal como a tradução das frases com as formas atómicas do verbo. Neste sentido, pelo facto de o *output* exibido pelo sistema Systran ser bastante adequado, podemos, então, concluir que ambas as construções podem ser utilizadas sem prejuízo para a tradução automática, uma vez que, numa fase de pós-edição, as alterações que teriam de ser feitas em (19) (a) e (19) (b) seriam mínimas.

Tomemos em conta agora a segunda construção. Vejamos a tradução automática da frase em (20), nos sistemas Systran e Google, respectivamente:

(20) A Maria tem receio de reprovar no exame.

(a) *The Maria has distrust to disapprove in the test.

(b) Mary is afraid to fail the test.

Como se pode verificar pelo *output* (20) (a), a tradução do Systran coloca problemas, o que não acontece com a tradução do Google. É necessário ter em conta que o Google é um sistema estatístico que inclui também uma base de dados composta por traduções, ao contrário do Systran. É provável que expressões semelhantes a (20) e

a respectiva tradução (20) (b) constem da sua base de dados e que o sistema recupere sempre a tradução.

O problema que se coloca com o Systran prende-se com o facto de o sistema não reconhecer a expressão verbal não atómica como uma unidade de sentido. Ou seja, o verbo *ter* e o nome *receio* formam uma unidade de sentido, que corresponde à forma atómica *recear*. Não reconhecendo esta construção como uma unidade de sentido, o sistema tradu-la palavra a palavra. Tal não constituiria um problema de maior, se o sistema não traduzisse *receio* por *distrust*, mas por *fear,* por exemplo.

Em (20.1), temos a frase com a forma atómica do verbo *recear*:

(20.1) A Maria receia reprovar no exame.

Veja-se a tradução automática desta frase nos sistemas Systran e Google, respectivamente:

(20.1) (a) *The Maria is afraid to disapprove in the test.

(20.1) (b) *Maria fears fail the test.

O Systran - (20.1) (a) - continua a traduzir *reprovar* por *disapprove*, já *recear* é traduzido pela forma perifrástica *to be afraid*. Na tradução da forma atómica do verbo, o resultado produzido pelo Google, em (20.1) (b), é pouco satisfatório, na medida em que *to* é omitido.

Ora, os *outputs* de ambos os sistemas, quer respeitem à expressão verbal não atómica quer à forma atómica, mostram que neste exemplo ambas as formulações colocam problemas à tradução automática.

Uma das hipóteses para explicar o desempenho do Systran prende-se, como já referido, com o facto de este não reconhecer a construção não atómica como uma unidade de sentido e consequentemente traduzir a expressão palavra a palavra. Por outro lado, este desempenho pode também dever-se ao facto de o sistema traduzir preferencialmente o verbo *reprovar* por *disapprove*. O verbo *disapprove* selecciona como argumento um objecto directo, como em (21):

(21) My father disapproves my choices.

No entanto este objecto directo pode de igual forma ser regido pela preposição *of*, como em:

(22) My father disapproves of my choices.

Tendo em conta estas características do verbo *disapprove*, o sintagma preposicionado *in the test* vem causar problemas, na medida em que o verbo em inglês não selecciona a preposição *in*. Este é um problema que pode estar relacionado com a informação especificada na entrada lexical do verbo, visto que todos os *outputs* apresentam como tradução para o verbo *reprovar* apenas *disapprove*, não regendo qualquer preposição.

Outra questão que se coloca é o facto de *exam* co-ocorrer na maior parte dos casos com *fail* e não com *disapprove*. Atente-se na seguinte frase e nas traduções dos sistemas Systran e Google, respectivamente:

(23) A Maria reprovou no exame.

(a) *The Maria disapproved in the examination.

(b) Mary failed the exam.

Voltando ao exemplo (20.1) – *A Maria receia reprovar no exame.* - e à tradução (20.1) (b) - **Maria fears fail the test.*- , o problema do Google prende-se apenas com a omissão de *to*. Tudo leva a crer que a frase foi traduzida em dois segmentos separados, por um lado *A Maria receia* e por outro *reprovar no exame*. O que pode dever-se ao facto de o Google dispor de memórias de tradução e, por essa razão, ter na sua base de dados segmentos de frases que são recuperados no processo de tradução. Neste caso, os dois segmentos foram recuperados e combinados para se obter a tradução em (20.1) (b). No entanto, o sistema não foi capaz de incluir a preposição *to*, que neste caso seria necessária para que o *output* fosse gramatical.

Numa conclusão provisória, pode referir-se que tanto as expressões não atómicas como as construções com a forma atómica colocam problemas à tradução automática. No entanto, como foi possível constatar pelos *outputs* acima apresentados, a utilização da forma atómica é a que oferece problemas mais fáceis de contornar numa fase de pós-edição, dado que, por exemplo, em (20.1) (b), é necessário apenas inserir *to.*

Atentemos noutro par expressão verbal não atómica/expressão verbal atómica. As frase em (24) e (24.1) são exemplos da ocorrência forma atómica, *queixar-se*, e da ocorrência forma não atómica, *fazer queixas*, respectivamente. (24) (a) e (24.1) (b) são as traduções dos sistemas Systran e Google, respectivamente.

(24) O João queixou-se à professora.

(24.1) O João fez queixas à professora.

(24) (a) *The João complained it the teacher.

(24) (b) The João made complaints to the teacher.

(24.1) (a) The John complained to the teacher.

(24.1) (b) The John made complaints to the teacher.

Comecemos pelos problemas que a forma atómica *queixar-se* e a forma não atómica *fazer queixas* colocam. Em ambos os casos, a construção envolve um complemento introduzido pela preposição *a*. No entanto, o problema da preposição apenas se coloca em (24) (a), no Systran, uma vez que, embora a contracção da preposição [*a* preposição + *a* artigo] pareça ser decomposta, apenas o determinante é traduzido para o seu equivalente em inglês – *the* – sendo a preposição omitida. No entanto, quando o clítico é deslocado para uma posição pré-verbal numa frase negativa, como no exemplo (25), o sistema reconhece a contracção da preposição com o artigo e tradu-la adequadamente, como se pode ver na respectiva tradução no Systran:

(25) O João não se queixou à professora

(a) The João did not complain to the teacher

Ainda no mesmo exemplo, outro problema se coloca: o clítico *se* é traduzido pelo pronome neutro *it*[12]. O problema que a forma atómica deste verbo (*queixar-se*)

[12] Analisam-se os casos de ambiguidade categorial da forma *se* noutra secção.

coloca é o facto de não ser possível omitir o clítico, ao contrário do que acontece com outros verbos, como por exemplo *casar*:

(26) A Maria casou-se ontem.

(27) A Maria casou ontem.

Assim sendo, a forma não atómica é a que conduz a melhores resultados na tradução automática, pelo que deve ser preferida em relação à forma atómica. Consequentemente, nos casos em que é possível a omissão do clítico, é preferível omitir o clítico.

Em (28) e (28.1), temos mais um caso para análise. Em (28), não temos uma expressão verbal não atómica, mas um caso de um verbo e os seus complementos. No entanto *dar um presente* pode comutar com o verbo *presentear*:

(28) Eu dei um presente à minha mãe.

(28.1) Eu presenteei a minha mãe.

(28) (a) I gave a present to my mother.

(28) (b) I gave a gift to my mother.

(28.1) (a) *I presenteei my mother.

(28.1) (b) I present to my mother.

Neste caso concreto, a frase em (28) não coloca quaisquer problemas a ambos os sistemas, como se pode ver tanto na tradução do Systran (28) (a), como na tradução do Google (28) (b).

O mesmo não acontece com a forma atómica (28.1), que coloca problemas a ambos os sistemas. O Systran, em (28.1) (a), não traduz o verbo *presentear*, e utiliza, na tradução, o verbo em português. Logo, o *output* do Systran é inaceitável. Refira-se que

este desempenho do sistema pode dever-se ao facto de o verbo ainda não constar do léxico do sistema.

Por sua vez, o Google - (28.1) (b) - tem um resultado aceitável. Em (28.1) (b), coloca-se ainda outra questão: a da forma verbal, que se relaciona com a falta de informação morfológica para indicar o tempo passado. Este *output* inaceitável pode dever-se ao facto de o Google não reunir na sua base de dados exemplos de traduções em que o verbo *presentear* esteja flexionado. Atente-se em outras frases em que o verbo *presentear* surge flexionado noutras formas. Na frase em (29), o verbo *presentear* está no Presente do Indicativo, na 1ª pessoa do singular. Vejamos o que acontece na tradução automática no Google:

(29) Eu presenteio o meu pai.

(a) *I hereby present my father.

O Google traduz *presentear* por *present* (*apresentar*), o que pode dever-se ao facto de *present* ser ambíguo, ao contrário de *presentear*. Vejamos um exemplo em que o verbo *presentear* está no Infinitivo e respectiva tradução no Google:

(30) Eu quero presentear a minha mãe com um livro.

(30) (a) I want to gift my mother with a book.

Neste caso, o infinitivo do verbo *presentear* não coloca problemas, o que vem sugerir a ideia de que talvez o sistema não disponha da frase em (28.1) na sua base de dados.

Em segundo lugar, com o verbo flexionado, em (28.1) (b), o sistema interpreta o determinante *a* como sendo a preposição *a*, traduzindo o determinante por *to*. Veja-se o que acontece quando o determinante é masculino, como na frase:

(31) *Eu presenteei o meu pai.*

(a) I present my father.

Neste caso, o problema da preposição não se coloca, na medida em que não há identidade formal entre o determinante *o* e a preposição *a*, contrariamente ao que acontece entre o determinante *a* e a preposição *a.*

Retomando os exemplos (28) e (28.1), as traduções de ambos os sistemas mostram, como já referido, que ambas as formas conduzem a resultados aceitáveis em tradução automática.

Analisemos agora as construções em (32). Ao contrário dos exemplos anteriores, excepto o exemplo (28) e (28.1), a expressão verbal em (32) não é uma expressão não atómica nos termos considerados anteriormente. Neste exemplo, temos um verbo – *dar* - e o seu objecto directo – *uma contribuição para a discussão*.

(32) O João deu uma contribuição para a discussão.

(32.1) O João contribuiu para a discussão.

Neste caso, em ambas as construções temos um complemento encabeçado pela preposição *para*, o que causa problemas na tradução automática no Systran, como se vê na tradução (32) (a). Mas (32.1) resulta num *output* aceitável no mesmo sistema.

As traduções do Google não apresentam qualquer problema, como se pode ver em (32) (b) e (32.1) (b).

(32) (a) *The João gave a contribution for the discussion.

(32.1) (a) The João contributed to the discussion.

(32) (b) The John gave a contribution to the discussion.

(32.1) (b) The John contributed to the discussion.

Desta forma, é conveniente utilizar a forma atómica do verbo, pois foi a que obteve melhores resultados na tradução automática em ambos os sistemas.

No que respeita à regência nominal[13], em (32) (a), coloca-se outro tipo de questão. A preposição que o nome *contribution* rege é *to*, logo a tradução do Systran não é aceitável. O problema pode prender-se com o facto de o Systran traduzir

[13] Muito sinteticamente, o que se entende por regência é a propriedade de algumas palavras seleccionarem preposições específicas. Verbos, substantivos e adjectivos seleccionam preposições – por exemplo, *assistir a*, *cheio de*, *talk to*, *want to*, etc. Por exemplo, o verbo *assistir* selecciona a preposição *a* e não outra qualquer – **assistir por*.

preferencialmente a preposição *para* por *for*, tendo em conta os testes efectuados neste sistema. O problema da inadequação de *for* como tradução de *para* não se coloca sempre, isto é, o facto de o sistema traduzir *para* por *for* não é um problema de maior em alguns casos, como por exemplo:

(33) A Maria partiu para Londres.

(34) A Maria comprou um presente para a Marta.

Veja-se a tradução destas duas frases no Systran:

(33) (a) The Maria left for London.

(34) (a) The Maria bought a present for Marta.

Nos exemplos acima apresentados, a tradução de *para* por *for* é adequada. Nestes casos, a tradução da preposição *para* por *for* corrobora a ideia de que *for* é a tradução preferencial, no Systran, para a preposição *para*.

Desta forma, no caso específico de (32) (a), a questão da inaceitabilidade na tradução da preposição relaciona-se com a regência do nome, ou seja, com o facto de *contribution* reger especificamente a preposição *to*. O nome *contribution* coloca sempre o problema de regência nas traduções do Systran, esteja ele integrado numa expressão verbal ou não. Veja-se os seguintes exemplos e as respectivas traduções no Systran:

(35) A Maria deu uma contribuição monetária para as vítimas.

(36) A contribuição para a discussão foi boa.

(35) (a) *The Maria gave a monetary contribution for the victims.

(36) (a) *The contribution for the discussion was good.

No caso da tradução do Google, em (32) (b), a questão da regência verbal não se coloca. Bem como nas restantes frases apresentadas que integram o nome *contribuição*, como se pode ver na respectiva tradução no Google.

(35) (b) Mary gave a monetary contribution to the victims.

(36) (b) The contribution to the discussion was good.

O facto de a tradução do Google ter um resultado positivo pode dever-se ao facto de este sistema possuir na sua base de dados exemplos de traduções em que o nome *contribution* surge com a preposição *to*.

Nos casos específicos de (32) e (32.1), a construção com a forma verbal atómica, em (32.1), é a mais recomendada para evitar problemas de regência verbal. Muitas vezes a forma atómica do verbo não coloca problemas de regência verbal, pois esta é normalmente um verbo que selecciona um objecto directo e este não é precedido de preposição.

Em (37), por exemplo, temos a forma não atómica (*ter influência*) que pode comutar com a forma atómica do verbo *influenciar*. Compare-se (37) com (37.1):

(37) O Pedro teve influência no negócio.

(37.1) O Pedro influenciou o negócio.

Neste caso, é indiferente utilizar qualquer uma das construções, pois, tanto com a expressão não atómica como a forma atómica se obtiveram resultados positivos na tradução automática, em ambos os sistemas.

Vejam-se os resultados das traduções nos sistemas, (37) (a) e (37.1) (a) para o Systran, e (37) (b) e (37.1) (b) para o Google.

(37) (a) Pedro had influence in the business.

(37.1) (a) Pedro influenced the business

(37) (b) The Peter was influential in the business.

(37.1) (b) The Pedro influenced the business.

Pode dizer-se que ambas as construções obtiveram um resultado adequado na tradução automática. No entanto, em (37) (b), o sistema parece «interpretar» o que é dito em (37). Semanticamente, o sistema transmite a informação que a frase de partida dá, mas não o faz nos mesmos termos. Este não é um problema de maior, uma vez que,

muitas vezes, na tradução humana se recorre também a transmitir o conteúdo sem usar a forma do texto de partida, sempre que isso seja possível, o que naturalmente acontece em consequência de diferenças estruturais entre as línguas.

Vejamos o que acontece quando introduzimos as seguintes alterações, mais concretamente, quando se insere um adjectivo (*decisiva*) para modificar ou um quantificador para quantificar o nome *influência*.

(38) O Pedro teve influência decisiva no negócio.

(39) O Pedro teve bastante influência no negócio.

(38) (b) Pedro had a decisive influence in the business.

(39) (b) Pedro had enough influence in the business.

A tradução obtida foi a desejada, ou melhor, a tradução da expressão verbal não foi igual a (37) (a), que não está, no entanto, em análise neste momento, pois estamos só a tratar a tradução do Google, em (37) (b).

Se considerarmos que o Google pode ter na sua base de dados a tradução (37.1) (b) para a frase (37.1), podemos concluir que (37.1) será sempre traduzida como em (37.1) (b).

A inserção do adjectivo, em (38), e do quantificador, em (39), resulta numa tradução diferente, pois vai alterar a frase que o sistema tem na sua base de dados.

A análise destes problemas visou mostrar as questões que se colocam na tradução automática de expressões verbais deste tipo. Isto é, construções não atómicas, que integram um verbo de suporte e um nome, adquirem sentido como uma unidade e, por isso, podem ser comutadas por formas atómicas de determinados verbos. A análise mostrou que os pares expressão não atómica/forma atómica do verbo não exibiram um comportamento uniforme, no que respeita à tradução pelos sistemas testados, e, nesta medida, não foi possível apresentar uma sugestão de simplificação/controlo uniformemente aplicável a todos os exemplos analisados.

4.1.2 Expressões verbais com verbos auxiliares aspectuais[14]

Nesta secção, analisam-se expressões verbais que integram um verbo auxiliar (verbo aspectual), uma preposição, regida pelo verbo aspectual, e um verbo principal no Infinitivo[15], como as que se propõem para análise a seguir.

Este tipo de construção revelou-se especialmente problemático para a tradução automática na medida em que os sistemas, em grande parte dos casos, não reconhecem estas estruturas como um complexo verbal e, em consequência, traduzem-nas palavra a palavra. Vejam-se as traduções para o exemplo (40), produzidas o Systran - (a) - e para o Google - (b):

(40) Ela acabou por comprar o casaco.

(a) *It finished to buy the coat.

(b) She ended up buying the coat.

A frase em (40) denota um evento que antecedido por um momento de incerteza ou dúvida ou de um obstáculo à realização desse evento. No entanto, nada impede a realização desse evento. *Ela acabou por comprar o casaco* denota que a decisão tomada pelo sujeito do predicado verbal foi precedida por um momento de incerteza, por exemplo. Este complexo verbal é composto pelo verbo *acabar por + verbo*.

Já o complexo verbal da frase em (41) é composto pelo verbo aspectual *acabar de + verbo* e denota uma acção que teve lugar num passado recente, ou seja, no momento exactamente anterior ao momento da enunciação.

(41) O João acabou de entrar na sala.

(a) *The João finished to enter in the room.

(b) The John just walked into the room.

[14] Na *Gramática da Língua Portuguesa* (2003:145-46), designam-se por *operadores aspectuais* formas como as analisadas no âmbito do presente trabalho. São elas *estar a*, *começar a*, *continuar a*, *acabar de*, *andar a + Infinitivo*. Neste sentido, justifica-se a afirmação: «Dizer que estas construções são operadores significa que se assume uma perspectiva dinâmica em que ocorre uma conversão de um determinado tipo de situação num outro, através de uma operação de transição (ou de transformação).» Esta designação será também adoptada no âmbito do presente trabalho, embora seja preferida uma designação mais geral, como expressões que integram verbos aspectuais.

A frase em (41) é parafraseável por *o João entrou agora mesmo na sala*, uma vez que *acabar de* «opera sobre processos e processos culminados e a leitura final é a de culminação ou, em alguns casos, de processo culminado, […]» (Mateus *et al.*, 2003: 150).

Relativamente aos resultados de tradução produzidos pelos sistemas, constatam-se vários problemas. Vejam-se as frases em (40) e (41) e as respectivas traduções no Systran, em oposição às traduções do Google. O Systran traduz as duas frases, (40) e (41), da mesma forma, o que não é aceitável, uma vez que as frases têm significados diferentes.

Comparando os *outputs* de ambos os sistemas, o Google é aquele que apresenta os resultados mais aceitáveis. Em (40) (b), traduz a expressão do exemplo (40) de forma fluente, através de um *phrasal verb*, seguido de *to buy (comprar)* na forma *–ing*. É provável que da sua base de dados já constasse a expressão complexa em (40) e a respectiva tradução com a seguinte estrutura *end up + -ing form*. Se se considerar que *acabar por* pode ter uma leitura modal, é interessante ver que o Google denota esse mesmo sentido. Confronte-se o exemplo em (40) com o exemplo em (42) e a respectiva tradução no Google:

(42) Ela acabará por comprar o casaco.

(b) She will eventually buy the coat.

Quanto ao Systran, em (40) (a), traduz o verbo *acabar* pelo seu sentido básico (*finish*), não tendo em conta que este está integrado num complexo verbal, não captando o seu significado. Acresce ainda que o verbo no Infinitivo, *comprar*, é traduzido pelo Infinitivo com *to (to buy)* o que nos leva a colocar duas hipóteses. Por um lado, o sistema terá este desempenho, por não reconhecer ou «ignorar» a preposição *por* e traduzir apenas o verbo no infinitivo. Por outro lado, é possível que o sistema reconheça a preposição *por* e a traduza pela preposição *to*. Para verificar estas duas hipóteses, testaram-se os sistemas com outros exemplos em que a preposição *por* está integrada em sintagmas preposicionais com diferentes valores sintáctico-semânticos[16].

[16]Veja-se o resultado dos testes no sistema Systran:
i. Faço isto por ti.
(a) I make this for it.

Os resultados obtidos revelam que quando a preposição vem seguida de um verbo no Infinitivo, a mesma é ignorada e o verbo no Infinitivo é traduzido cpm a preposição *to*, o que confirmaria a primeira hipótese. Quando a preposição introduz o agente da passiva, ela é traduzida por *by*. E finalmente, quando o sintagma preposicional tem o valor de razão/motivo (*Faço isto por ele* ou *por dez euros*), a preposição é traduzida por *for*. Conclui-se que, quando a preposição vem seguida de um verbo, esta é sempre ignorada.

No que concerne ao exemplo (41), o Google, em (41) (b), é o que apresenta, mais uma vez, os resultados mais aceitáveis. Em termos concretos, a expressão complexa *acabar de + verbo no Infinitivo* é reconhecida e traduzida pelo seu equivalente em inglês americano *just + Pretérito*. Note-se que, em inglês britânico, esta construção seria traduzida por *have just + Particípio Passado*. A preferência pela variante americana pode dever-se ao facto de o sistema (Google) ser americano e os *corpora* que compõem a sua base de dados estarem maioritariamente escritos na variante do inglês americano. Em (41) (a), o Systran exibe um desempenho idêntico ao que teve em (40) (a), ou seja, traduz a expressão complexa palavra a palavra, dando origem a um *output* semanticamente incongruente. A principal razão para a inaceitabilidade deste *output* prende-se com o facto de o verbo *enter* denotar uma acção pontual que é semanticamente incompatível com o sentido básico do verbo *finish*.

Retomando a análise do exemplo (40) (a), importa ressaltar que parece ser o verbo aspectual, o verbo *acabar* com a devida regência, a colocar problemas para a tradução automática, pelo menos para o Systran. O Systran traduz palavra a palavra ambas as expressões com verbo aspectual, tanto (40) como (41). Este desempenho é inadequado, na medida em que as frases em (40) e (41) veiculam diferentes sentidos. Senão veja-se: a tradução em (40) (a) tem um sentido diferente do sentido do exemplo (40). A frase em (40) (a), **It finished to buy the coat*, não atentando à questão do pronome *it*, tem o significado de que o sujeito da frase acaba de realizar a acção.

---

ii. Ele vendeu a casa por dez euros.
(a) It sold the house for ten euros.
iii. Ele foi preso por roubar uma jóia.
(a)It was imprisoned to rob a jewel.
iv. Ele foi apanhado por dois polícias.
(a) It was apanhado by two police.

E se as traduções do Google, que recorreu a *phrasal verbs*, foram aceitáveis; pode dizer-se, para o efeito, que o verbo *acabar*, regendo as preposições *por* e *de*, como em (40) e (41) funciona quase como um *phrasal verb*, daí o sentido de *acabar por* e *acabar de* não ser o sentido das traduções do Systran mas sim do Google.

Em (43), temos uma expressão com a seguinte estrutura: *andar a + verbo*:

(43) Ela anda a aprender chinês.

(a) It walks to learn Chinese.

(b) She walks to learn Chinese.

Esta expressão verbal denota uma acção com valor de duração, que teve início num dado momento no passado, decorre num momento presente, e que pode prolongar-se para além do presente. «Este operador aspectual apresenta algumas semelhanças com *estar a + Infinitivo*, nomeadamente em relação aos tipos aspectuais de predicados base sobre os quais opera, isto é, eventos e estados faseáveis. No entanto, a leitura final diverge, pois nesta construção obtém-se um estado habitual ou frequentativo.» (Mateus *et al.*, 2003:150) Ou seja, a frase em (43), Ela anda a aprender chinês, pode ser parafraseada por pela construção que integra um verbo aspectual, Ela está a aprender chinês, pois ambas as construções veiculam um aspecto imperfectivo

Em (43), o verbo *andar* causa problemas, uma vez que na construção complexa que estamos a tratar (*andar a + verbo*), *andar* não denota *caminhar*. Na construção em (43), *andar* é um verbo auxiliar, e não um verbo pleno como em:

(44) Ele anda na praia todas as manhãs.

Comparando os resultados das traduções, pode dizer-se que a expressão com verbo aspectual (43) é traduzida inadequadamente, tanto pelo Systran, em (43) (a), como pelo Google, em (43) (b), dado que nenhum deles apresenta a estrutura equivalente, e a mais adequada em inglês - *She is learning Chinese.* Os resultados inaceitáveis de tradução em (43) (a) e (43) (b) prendem-se com o facto de a tradução equivalente de *andar* na construção portuguesa ser o verbo *to be* na construção inglesa. O facto de o verbo *andar* corresponder ao verbo *to be* na estrutura inglesa, que equivale

à expressão complexa portuguesa, coloca problemas, na medida em que os sistemas não reconhecem a construção complexa, traduzindo os seus componentes literalmente.

Nas traduções efectuadas por ambos os sistemas, verificou-se que, tanto o Systran como o Google, traduzem a preposição *a*, que faz parte da construção com verbo aspectual, pela preposição inglesa *to*. À semelhança do que aconteceu nos exemplos (40) e (41), os sistemas parecem exibir o mesmo procedimento neste exemplo. Por um lado, os sistemas podem não reconhecer ou podem ignorar a preposição, traduzindo o verbo *aprender* no Infinitivo com *to*. Por outro lado, é possível que os sistemas reconheçam a preposição *a* e a traduzam por *to*.

O desempenho inadequado exibido pelo Systran pode dever-se ao facto de, por um lado, no dicionário do sistema, na entrada lexical do verbo *andar* não haver especificações que dêem conta do seu uso numa construção como a analisada em (43), e, por consequência, o verbo *andar* é traduzido como verbo pleno em inglês *walk*. Por outro lado, é possível que o sistema não tenha na sua gramática a estrutura complexa do exemplo (43). A corroborar esta hipótese está o resultado aceitável obtido na tradução da expressão semanticamente equivalente (*Ela está aprendendo chinês.)*, na variedade brasileira do português. É importante referir que, para a tradução em língua portuguesa, variedade usada no Systran é o português do Brasil, daí que a frase com a estrutura *estar + gerúndio*, uma estrutura mais comum na variedade brasileira do português, tenha um *output* aceitável, em oposição à construção com verbo *andar*, na qualidade de verbo aspectual, muito mais comum no português europeu.

No que concerne ao Google, talvez o sistema disponha, nos seus *corpora*, da construção complexa que integra o operador aspectual *estar a + verbo* (*Ela está a aprender chinês.)*, em muitos mais casos do que da construção ilustrada em (43) - *Ela anda a aprender chinês. -,* o que explica que o sistema traduza preferencialmente *andar* por *walk*. No possuindo exemplos da estrutura com o verbo *andar*, o sistema não a pode reconhecer e, consequentemente, acaba por traduzir o verbo *andar* na qualidade de verbo pleno (*walk*) e não na qualidade de verbo aspectual inserido numa estrutura como a analisada.

A frase em (45) integra uma expressão complexa com a seguinte estrutura: *começar por + verbo*.

(45) Ela começou por dizer o poema mais recente.

(a) *It started to say the poem most recent.

(b) *She begun by saying the poem later.

Esta construção com verbo aspectual denota o ponto de partida de uma acção. O objecto directo desta frase inclui um modificador no grau superlativo relativo de superioridade (cf. *o poema mais recente).* E embora não constitua o objecto de análise desta secção, sendo aqui superficialmente tratada, importa referir o desempenho de ambos os sistemas relativamente à estrutura de comparação, que poderia ser mais aceitável se a mesma fosse objecto de controlo no âmbito de uma linguagem controlada. Esta questão poderia constituir um tópico de análise mais profunda neste trabalho mas, pelas suas dimensões, não poderá ser tratado com a profundidade que merece.

Passando à análise dos *outputs* dos sistemas, e em termos globais, o desempenho do Systran é menos satisfatório relativamente ao desempenho do Google. Enquanto o Systran, em (45) (a), faz uma tradução literal da estrutura optando pelo verbo *start*, que não é a melhor escolha, o Google reconhece a expressão complexa, traduzindo-a adequadamente, procedendo a uma melhor selecção no que diz respeito ao auxiliar aspectual - *begin*. À semelhança do que tem acontecido com os exemplos analisados até agora, o Systran ignora a preposição *por* e traduz o verbo no Infinitivo precedido de *to*.

No que concerne ao problema da tradução do objecto directo do verbo, *o poema mais recente*, acima referido, cabe aqui dizer que o desempenho de ambos os sistemas não é satisfatório uma vez que a tradução do objecto directo é feita de forma segmentada. Este desempenho dever-se-á ao facto de o constituinte nominal integrar um modificador com grau explícito. Incapazes de interpretar a estrutura no seu todo, os sistemas segmentam-na em duas partes. Por um lado, interpretam o determinante e o nome como fazendo parte de um mesmo sintagma nominal e por outro, interpretam o modificador, *mais recente*, como um segmento, traduzindo-os separadamente. Consequentemente, o adjectivo é colocado em posição pós-nominal na tradução do Systran, quando deveria surgir em posição pré-nominal, visto ser esta a posição canónica do adjectivo em inglês numa relação de modificação. Ao colocar o adjectivo em posição pós-nominal, o sistema estabelece uma relação de predicação. Por outro

lado, o Google interpreta e traduz o segundo segmento como sendo um advérbio no grau comparativo (*later*).

Neste caso concreto, uma forma de controlar a estrutura de comparação, visando obter um *output* de melhor qualidade seria colocar o adjectivo em português numa posição pré-nominal, como no exemplo (46). Contudo, há que referir que esta estratégia não se pode alargar a todos os casos de modificação, uma vez que há adjectivos que adquirem um significado diferente dependendo da posição que ocupam.

(46) Ela começou por dizer o mais recente poema.

Neste contexto, o Systran apresenta um resultado mais aceitável no que diz respeito à estrutura de comparação, relativamente à tradução (45) (a.). Veja-se então o resultado da tradução no sistema:

(46) (a) It started to say the most recent poem.

O Google exibe também um desempenho aceitável, apresentando um *output* como (46) (b):

(46) (b) She started by saying the latest poem.

No entanto, desta vez o Google opta pelo verbo *start* para traduzir o verbo *começar* na construção com verbo aspectual, o que pode indicar que os verbos *start/begin* são comutáveis.

A construção em análise em (47) tem a estrutura *continuar a + verbo* e denota uma acção de continuidade que decorre num intervalo de tempo que inclui o presente.

(47) Ela continua a aprender chinês.

(a) It continues to learn chess.

(b) She continues to learn chess.

A frase em (47) não coloca qualquer problema aos sistemas. Verifica-se que ambos os sistemas têm traduções aceitáveis, o que pode dever-se ao facto de a estrutura em inglês ser muito próxima da estrutura em português. Ou seja, na expressão verbal

em (47), por exemplo, o verbo *continuar* tem como equivalente directo em inglês o verbo *continue*.

Na análise das traduções de ambos os sistemas, é importante referir o desempenho do Systran na tradução da preposição *a*. Aparentemente, o Systran tem o mesmo desempenho na tradução da preposição *a* nesta estrutura e na tradução das preposições nas expressões verbais analisadas até aqui, ou seja, o Systran traduz, mais uma vez, *a* por *to*. No entanto, torna-se impossível perceber se o sistema ignora a preposição e traduz o verbo *aprender* no Infinitivo precedido de *to*, como tem sido sugerido até aqui ou se o sistema reconhece a preposição e a traduz de forma aceitável. No decorrer da análise desta construção e tendo em conta o desempenho do Systran na tradução das estruturas que integram um verbo aspectual é concebível ponderar uma outra hipótese. Como já foi referido anteriormente, o verbo *continuar* tem equivalência directa no verbo inglês *continue*. Ora, a preposição *to* é a que o verbo *continue* rege, logo o sistema pode ter traduzido a preposição de forma aceitável por esta razão.

Em (48), temos uma construção com a seguinte estrutura: *estar a + verbo*.

(48) Ela está a rir.

(a) It is to laugh.

(b) She is laughing.

É através desta construção que se obtém o Progressivo em português europeu, na medida em que a acção é perspectivada como estando a decorrer. «A esta característica podemos ainda associar a de duração e a de incompletude, pois se uma eventualidade está no seu decurso, é natural que tenha duração e que também não esteja completa, ou não tenha atingido o seu ponto terminal.» (Mateus *et al.*, 2003:146). Esta estrutura é também tradicionalmente designada forma perifrástica.

O Systran, em (48) (a), oferece uma tradução inadequada, ao contrário do Google, em (48) (b), que obtém um *output* perfeitamente aceitável. Uma questão que é importante referir é a tradução que o Systran faz da preposição *a*. Mais uma vez, parece plausível dizer o que foi dito para os exemplos acima apresentados, pois o sistema apresenta o mesmo desempenho exibido na tradução das estruturas dos exemplos

anteriores. O Systran parece mais uma vez não reconhecer a preposição na estrutura com verbo aspectual, traduzindo o verbo *rir* no infinitivo precedido de *to*. O Google, por sua vez, reconhece a construção e tradu-la pela forma progressiva em inglês *be* + *-ing form.*

O facto de o Systran apresentar uma tradução inadequada da estrutura em (48) pode ser explicado pela seguinte razão. Em conformidade com o que tem vindo a acontecer até aqui, é possível que o sistema não possua, na sua gramática, especificações quanto ao uso da construção com verbo aspectual do exemplo (48). Mais uma vez, convém dizer que a construção apresentada em (48) é mais produtiva na variedade europeia do português, pelo que o Systran, fazendo uso da variedade brasileira a usada no sistema, traduz de forma inaceitável a estrutura verbal. Veja-se o resultado da tradução no Systran da estrutura frequente na variedade do português do Brasil:

(49) Ela está rindo
(49) (a) It is laughing.

O Google, por sua vez, traduz a construção com verbo aspectual em (48) de forma aceitável. Pode dizer-se que, devido ao facto de o Google ser um sistema baseado em estatística, e tendo em conta que o sistema faz uso de toda a Internet para a recolha de exemplos de traduções, é bem provável que a expressão complexa em (48) possa constar dos *corpora* que o sistema tem ao seu dispor.

Numa breve conclusão, pode dizer-se que a análise mostrou que este tipo de construção coloca problemas à tradução automática. O Google é o sistema que consegue um *output* mais aceitável na maior parte dos casos, o que pode dever-se ao facto de este sistema dispor de uma base de dados onde estão guardadas traduções e, muito provavelmente possuir nos seus *corpora* exemplos de traduções com as construções com verbo aspectual em análise. Já o Systran obtém resultados inaceitáveis na maior parte dos casos.

Sugere-se então substituir expressões deste tipo por formas verbais atómicas que mantenham o significado que as expressões em questão denotam. Esta proposta parece não estar em conformidade com a natureza das construções que integram verbos aspectuais, propostas para análise. De facto, a utilização de uma expressão verbal

complexa que integre um verbo aspectual não equivale completamente à utilização de uma forma verbal atómica. Por exemplo, retomando uma estrutura como (50), pode concluir-se que não veicula exactamente a mesma informação que a forma verbal atómica *canta* encerra em si, como em (51).

(50) Ela está a cantar.
(51) Ela canta.

Dito isto, torna-se importante explicar a ideia proposta. No âmbito de uma linguagem controlada direccionada para a tradução automática, ou seja, que visa eliminar problemas de processamento computacional/tradução de determinadas estruturas, a proposta de comutar uma estrutura que integre um *verbo aspectual + preposição + Infinitivo de um verbo principal* por uma forma verbal atómica faz todo o sentido.

Veja-se então para a frase em (40) - *Ela acabou por comprar o casaco.* - a substituição da expressão verbal complexa com um verbo aspectual pela forma atómica do verbo *comprar* no Pretérito Perfeito e a respectiva tradução nos sistemas, (a) para o Systran e (b) para o Google, como até aqui:

(52) Ela comprou o casaco.
(a) It bought the coat.
(b) She bought the coat.

Em (52), o processo de simplificação/controlo consistiu na substituição da estrutura com verbo aspectual pela forma atómica do verbo principal (*comprar*), de forma a conseguir que os sistemas tivessem um desempenho satisfatório. Atentando nas traduções bastante aceitáveis de ambos os sistemas, (52) (a), para o Systran, e (52) (b), para o Google, conclui-se que o recurso à forma atómica pode ser a melhor opção de controlo desta construção. O factor preponderante para esta melhoria deve-se ao facto de (52) ser uma frase simples, com ordem SVO, e, nesta medida, não oferecer problemas aos sistemas. No entanto, é importante atentar no sentido veiculado tanto pela forma com o verbo aspectual como pela forma do verbo principal (*comprar*) no Pretérito Perfeito (*comprou*). A estrutura com verbo aspectual denota um evento efectivamente concluído, mas que foi antecedida por um momento de incerteza ou dúvida ou por um obstáculo, mas que ainda assim foi realizada. O mesmo não pode ser

dito para a forma verbal no Pretérito Perfeito (*comprou*), na medida em que apenas denota uma acção efectivamente concluída. Logo, tornar-se-ia necessário dar mais informação contextual, de forma a colmatar o sentido da forma atómica, como, por exemplo, no fragmento:

(53) A Maria tinha dúvidas, mas ela comprou o casaco.

Em (53), temos uma frase que ilustra as alterações sugeridas, ainda que se tenha optado pela interpretação de dúvida para a construção aspectual *acabar por*. No entanto, esta também poderia denotar um obstáculo à realização do evento – *A Maria não tinha dinheiro, mas ela acabou por comprar o casaco.* É importante referir que o *input* da tradução não tem de ser perfeito, tem apenas de servir o objectivo principal de obter uma tradução automática aceitável.

Neste seguimento, as traduções dos sistemas mostram que a substituição da expressão verbal pela forma atómica do verbo resulta num *output* aceitável para ambos os sistemas.[17]

Em (54), as alterações efectuadas prenderam-se igualmente com a substituição da estrutura em análise pela forma atómica do verbo principal (*entrar*) no Pretérito Perfeito.

(54) O João entrou agora na sala.
(a) The João entered now in the room.
(b) John has now entered the room.

Retomando o que foi dito para o exemplo em (40), a construção *acabar de* + *verbo* veicula a ideia de uma acção que teve lugar num passado recente, ou seja, no momento exactamente anterior ao momento da enunciação. Neste sentido, refira-se que a forma atómica do verbo principal (*entrar*), individualmente, não veicula a mesma informação que a forma com verbo aspectual. O Pretérito Perfeito «é claramente um tempo do passado, embora não seja perfectivo na medida em que não determina na maior parte dos casos a existência de um estado consequente. É, no entanto, sempre terminativo, isto é, marca um momento em que um estado ou um evento terminou

[17] Refira-se que o facto de o Systran traduzir os pronomes pessoas da terceira pessoa do singular pelo pronome expletivo *it* não está em análise nesta secção. A questão é abordada na secção que trata o sujeito nulo.

[…].» (Mateus *et al.*, 2003:156). Desta forma, torna-se necessária a introdução do advérbio *agora* para expressar a ideia de que a acção terminou num tempo recente, como no exemplo (52).

Vejamos mais uma proposta de controlo/simplificação em (55):

(55) Ela aprende chinês.
(a) It learns Chinese.
(b) She learns Chinese.

Como foi verificado na análise do exemplo (43) - *Ela anda a aprender chinês.* -, o operador aspectual *andar a*, em conjugação com o Infinitivo do verbo principal, apresenta semelhanças com *estar a + infinitivo*, e embora o sistema reconheça com maior facilidade este último operador, o processo de simplificação/controlo não poderá passar por esta substituição. Ainda que *estar a* e *andar a* veiculem um aspecto imperfectivo o desempenho dos sistemas não se revela mais satisfatório com esta substituição.

Em consequência, a melhor opção de simplificação/controlo encontrada foi a substituição da construção com verbo aspectual pela forma do verbo principal (*aprender*) no Presente do Indicativo. Como se pode verificar em (55), o resultado obtido na tradução automática é bastante aceitável. Tal como o operador aspectual *andar a*, o recurso ao Presente do Indicativo mantém «a leitura de estado habitual, construído com base numa ocorrência indeterminada de eventos do mesmo tipo que têm lugar num intervalo de tempo não delimitado, mas que inclui o tempo da enunciação.» (Mateus *et al.*, 2003: 144).

Em (45) - Ela começou por dizer o poema mais recente -, a expressão verbal *começar por + verbo (dizer)* veicula dois tipos de informação. Por um lado, a informação que o verbo aspectual (*começar*) expressa e que nos dá conta do início de um evento. Evento esse que é denotado pelo verbo principal (*dizer*), por outro lado. No processo de simplificação, o verbo aspectual (*começar*) foi substituído pelo advérbio *inicialmente*, num primeiro teste, e pela expressão *no início*, num segundo teste, de forma a manter a informação que o verbo *começar* veicula. É assim possível simplificar a frase, o que resulta numa tradução bastante aceitável por parte dos sistemas, como se pode verificar em (56) (a) e (56) (b):

(56) Inicialmente ela disse o poema mais recente.

(56.1) No início, ela disse o poema mais recente.

(a) Initially it said the poem most recent.

(b) Initially she said the poem later.

(a) At the beginning, it said the poem most recent.

(b) At first, she said the poem later.

A expressão verbal em (57) não colocou quaisquer problemas na tradução automática por parte de ambos os sistemas e, nesta medida, nem sequer foi considerada no processo de simplificação. Uma possível justificação para o desempenho aceitável de ambos os sistemas pode ser o facto de a expressão verbal complexa *continuar a + verbo (continuar)* encontrar correspondência directa no verbo inglês *continue to*.

Em (58), apresenta-se mais um exemplo de uma operação de controlo/simplificação:

(58) Ela ri.

(a) It laughs.

(b) She laughs.

Na sequência da análise efectuada para a expressão verbal em (48) - *Ela está a rir.* -, e à semelhança do que foi dito para o Presente no exemplo em (56), optou-se pela substituição da expressão que integra um verbo aspectual pelo Presente do Indicativo.

A diferença entre os verbos estabelece-se em termos sintácticos. *Aprender* selecciona objecto directo, ao passo que *rir* não. É importante também salientar que o verbo *aprender* e o verbo *rir* denotam eventos que se projectam no tempo com durações diferentes. Enquanto o verbo *aprender* denota um processo que pode estender-se durante um período de tempo alargado e indeterminado, já o verbo *rir* denota uma acção que decorre por um período de tempo menos alargado do que *aprender*. Neste sentido, a sintaxe e a semântica de ambos os verbos condiciona o processo de simplificação/controlo. E, por essa razão, sugerem-se duas estratégias de simplificação/controlo: por um lado, o recurso a uma expressão adverbial de tempo, como por exemplo *neste momento*, por outro o recurso à variedade do português do

Brasil e dialectos do Alentejo e Algarve, ou seja, *estar* + *gerúndio*, já que esta construção se mostrou viável pelos testes nos sistemas. Veja-se os testes tendo em conta estas duas estratégias:

(59) Ela ri neste momento.
(60) Ela está rindo.

(a) It laughs at this moment.
(b) She laughs at the moment.

(a) It is laughing
(b) She is laughing.

As traduções dos exemplos simplificados também tiveram um *output* aceitável, daí que se sugira o uso das formas verbais atómicas como uma forma de evitar os problemas colocados pelas expressões que integram um verbo aspectual

4.1.3 Sujeito nulo

Nesta secção, aborda-se a questão do sujeito nulo em português em contraponto com a obrigatoriedade da realização do sujeito em inglês. O objectivo do estudo desta questão prende-se com as implicações que ela coloca na tradução automática, uma vez que, na maior parte dos casos, os sistemas aqui utilizados resolvem esta questão de forma pouco satisfatória.

Nesse sentido, torna-se importante pôr em relevo as diferenças fundamentais entre o português e o inglês. O português é uma língua de sujeito nulo, ou seja, admite a não realização do sujeito em frases finitas, contrariamente ao inglês, cuja realização do sujeito é obrigatória. Os sujeitos nulos ou subentendidos, em português, são legitimados por uma flexão verbal rica. Ou seja, o facto de em português a flexão verbal conter informação quanto ao número e pessoa, além do mais, permite que o sujeito seja facilmente identificado. A única excepção que deve ser referida é a identidade formal entre a 1ª e a 3ª pessoa de algumas formas verbais, como é o caso do Imperfeito do Indicativo. Veja-se o seguinte exemplo:

(61) Corria todas as manhãs à mesma hora.

A flexão da forma verbal *corria* não é suficiente para se perceber se o sujeito da frase é *eu* ou *ele/ela*. A desambiguação depende de informações adicionais do contexto linguístico ou situacional.

Na língua inglesa, pelo contrário, e em virtude de a flexão verbal ser bastante pobre, o sujeito tem de estar obrigatoriamente realizado. É exemplo disso a flexão verbal do Presente do Indicativo, constituída apenas por duas formas: a da 3ª pessoa do singular, que se distingue das restantes pessoas por conter o morfema *-s,* e a flexão do Pretérito Perfeito, que é idêntica para todas as pessoas na conjugação do verbo. Logo, na frase **Bought a book about gardening* torna-se impossível identificar o sujeito, daí que seja imperativa a realização do mesmo. A única excepção em inglês, no que diz respeito ao sujeito nulo, é o Imperativo, que admite o sujeito não realizado, como em *Come here, please*.

O «parâmetro do sujeito nulo» ou «parâmetro *pro-drop*» marca desta forma uma distinção entre línguas como o português, o italiano e o espanhol e línguas como o francês, o inglês e o alemão, tal como Chomsky (1981:241) refere:

> "The optimal assumption, hence the assumption that we will assume to be pending evidence to the contrary, is that there is a single parameter of core grammar – the "pro-drop parameter" – that distinguishes Italian-type languages from French-type languages. When the parameter is set one way or another, the clustering of properties should follow."

As frases, de (62) a (65), que a seguir se propõem para análise, são exemplos ilustrativos da não realização do sujeito na língua portuguesa:

(62) Compraram o livro ontem.

(63) Há um acidente na estrada principal.

(64) Dizem que o desemprego vai aumentar.

(65) Diz-se que ele comprou uma casa no campo.

Em (62), o sujeito não realizado é facilmente recuperado a partir da flexão do verbo. Ou seja, a forma verbal *compraram* tem traços de pessoa e de número que permitem identificar o conteúdo do sujeito e saber que se trata do pronome pessoal da 3ª pessoa do singular (*eles/elas*) ou de uma expressão nominal equivalente em número e pessoa como *os estudantes*, por exemplo.

O exemplo (63) ilustra um caso de sujeito nulo expletivo. Em (64), o sujeito não realizado é um sujeito nulo «indeterminado» ou de referência arbitrária. Esta é uma divergência estrutural entre algumas línguas e tem implicações no processo de tradução automática, como ficará claro na tradução dos exemplos.

Na frase em (65), o pronome clítico *se* constitui mais uma estratégia para exprimir o sujeito indeterminado ou de referência arbitrária. Isto é, a indeterminação do pronome clítico *se* não permite recuperar o sujeito. Muito sinteticamente, uma questão que o pronome clítico coloca prende-se com o facto de este ser formalmente idêntico a

outros pronomes, tais como os reflexos e recíprocos, e ao *se* apassivante, o que vai ter implicações na tradução automática.

O exemplo em (66) ilustra um caso de inversão de sujeito.

(66) Foram à manifestação muitos trabalhadores.

O português e outras línguas românicas, com excepção do francês, admitem este fenómeno, enquanto as línguas germânicas, línguas de sujeito nulo, não admitem a posição pós-verbal do sujeito. O facto de este estar deslocado para uma posição pós-verbal deixa «vazio» o lugar canónico do sujeito. Apesar de este não poder ser considerado um caso de sujeito nulo, a topicalização à direita do sujeito vai dar origem a desempenhos menos adequados na tradução automática.

A seguir, apresentam-se as traduções dos exemplos de (62) a (65), nos sistemas de tradução automática. Mais uma vez, (a) para o Systran e (b) para o Google. As traduções dos exemplos mostram que os sistemas não processam de forma satisfatória o sujeito nulo e a inversão do sujeito. Assim, atente-se nos exemplos:

(62) (a) They had bought the book yesterday.

(b) *Bought the book yesterday.

(63) (a) *It has an accident in the main road.

(b) There is an accident on the highway.

(64) (a) They say that the unemployment goes to increase.

(b) They say that unemployment will increase.

(65) (a) One says that it bought a house in the country.

(b) It is said that he bought a house in the country.

Na tradução do exemplo (62) - *Compraram o livro ontem.* -, o Systran, em (62) (a), tem um desempenho inaceitável na tradução do tempo verbal, o que se deve, possivelmente, à identidade formal entre a forma verbal *compraram* no Pretérito Perfeito e a mesma forma verbal no Pretérito Mais-que-Perfeito Simples. Esta identidade explica a inadequação da tradução da forma verbal *compraram*. Em inglês, o

Past Perfect é incompatível com o advérbio *yesterday*, pois «o *Past Perfect* exprime sempre uma forma ou outra de passado em relação a outro passado». (Larreya *et al.*, 1998:34). É possível que o sistema não tenha informação quanto à incompatibilidade sintáctico-semântica entre o *Past Perfect* e o advérbio *yesterday*.

O Google, (62) (a), por sua vez, produz uma tradução adequada da forma verbal. No entanto, deixa o lugar de sujeito vazio na frase (62). No que concerne à tradução do tempo verbal, o Google tem um desempenho aceitável, o que, muito provavelmente, se deve ao facto de não ter conhecimento linguístico.

Não se pretende defender aqui que os sistemas baseados em estatística têm melhores desempenhos do que os sistemas baseados em regras, o que se pretende concluir é que o facto de o sistema não possuir conhecimento linguístico e, em consequência disso, não dispor de informação quanto à flexão do verbo *comprar* no Pretérito Perfeito (*eles compraram*) e no Pretérito Mais-que-Perfeito Simples (*eles compraram*), pode significar que o factor da identidade formal não se coloca neste caso. Por outro lado, é possível que o sistema não disponha, nos seus *corpora*, de exemplos em que o *Past Perfect* ocorra na mesma frase com o advérbio *yesterday*.

Da mesma forma, o facto de o sistema (Google) não ter conhecimento linguístico pode também explicar que o mesmo não reconheça a obrigatoriedade de preencher o lugar do sujeito em inglês, deixando o lugar vazio.

Na tradução do exemplo (63) - *Há um acidente na estrada principal.* -, os sistemas exibem desempenhos diferentes. Em (63) (a), o verbo *haver*, que na frase em (63) tem o sentido de *existir*, é traduzido pelo Systran no sentido de *possuir/ter*. Desta forma, o sistema flexiona o verbo *to have* na 3ª pessoa do singular, e, na falta de sujeito, o sistema atribui à forma verbal um pronome expletivo *it* para preencher o lugar.

É importante referir as diferenças entre a variedade europeia e a variedade brasileira do português no que diz respeito a esta questão. No português europeu, o verbo *haver* pode ter o significado de *existir*. Pelo contrário, na variedade brasileira do português, o verbo *ter* ocorre muito mais frequentemente com o significado de 'existir', como no seguinte exemplo *Tem muito bandido nessa cidade*.

Mais uma vez, é necessário referir que a variedade do português utilizada nos sistemas de tradução automática, e para o caso, no Systran, é a brasileira, o que explica muitas traduções inaceitáveis quando o que se pretende do sistema é que este traduza para a variedade europeia do português. No entanto, não deixa de ser interessante ver o desempenho do sistema Systran quando a direcção da tradução é do inglês para o português. Veja-se o que acontece no seguinte exemplo:

(67) (i) There is an accident in the main road.

(i) (a) Há um acidente na estrada principal.

Neste caso, a tradução do Systran é adequada. Será provável que o sistema disponha de especificações quanto a esta questão quando o sentido da tradução é do inglês para o português. Por outro lado, é de notar que o desempenho do sistema é bastante mais aceitável quando a direcção da tradução é do inglês para português, na maior parte dos casos, o que pode dever-se ao facto de a tradução neste sentido estar muito mais desenvolvida e estudada.

É de notar, como se tem verificado nas traduções do Systran de outros exemplos ao longo deste trabalho que, na maior parte das situações, na falta de sujeito, o Systran preenche o lugar com um pronome expletivo *it* em inglês, o que pode dever-se ao facto de, sendo expletivo, este pronome ser semanticamente vazio e, na falta de informação concreta quanto ao sujeito na frase em português, no caso de um sujeito nulo, por exemplo, a tradução pelo *it* expletivo ser a melhor solução.

O Google, em (63) (b), tem um desempenho mais aceitável, na medida em que reconhece o verbo *haver* na acepção de *existir*, e tradu-lo pela forma equivalente em inglês *there to be*. Muito provavelmente, o sistema tem este desempenho por possuir nos seus *corpora* exemplos em que a probabilidade de *haver* (*há*) ser traduzido por *there to be* ser muito elevada.

Em (64) - *Dizem que o desemprego vai aumentar.* -, a questão do sujeito nulo com interpretação indefinida ou arbitrária é resolvida por ambos os sistemas da mesma forma. Tanto o Systran, em (64) (a), como o Google, em (64) (b), atribuem como sujeito do verbo *say* o pronome pessoal *they*. Na falta de sujeito, os sistemas preenchem o lugar

com o pronome pessoal *they*. Ambos os sistemas têm este desempenho pois a forma verbal em português (*dizem*) legitima-o pelos traços de pessoa e número.

Neste sentido, a tradução de ambos os sistemas é aceitável. No entanto, em termos pragmáticos, *they say* não veicula a mesma indefinitude que a construção com o sujeito nulo indefinido ou de interpretação arbitrária em português veicula. O facto de o pronome pessoal *they* ser co-referente ou anafórico pode remeter para uma entidade referida em outro momento do discurso ou do texto. Isto é, a expressão passiva *it is said* tem um sentido mais indefinido e é mais adequada para veicular o sentido do sujeito nulo da forma verbal *dizem*.

Em (65) - *Diz-se que ele comprou uma casa no campo.* -, ambos os sistemas têm desempenhos diferentes no que concerne à tradução da construção *diz-se que…* e do pronome impessoal *se*. Neste exemplo, o pronome clítico *se* «absorve a informação de caso nominativo» (Mateus, *et al.*, 2003: 445). Desta forma, o Systran, em (65) (a), traduz o pronome indefinido *se* pelo pronome *one*, que é indefinido quanto ao género, o que veicula o carácter indefinido da frase em português. A tradução é aceitável mas dá um carácter mais formal à frase. O Google, em (65) (a), por sua vez, traduz a expressão *diz-se que* pela expressão passiva *it is said that*. Estatisticamente, a expressão passiva é muito mais comum em inglês, talvez também por ser mais informal, daí que o Google apresente esta tradução.

Na tradução do exemplo em (65) pelo Systran, coloca-se a questão do pronome *ele* na frase subordinada. O sistema traduz o pronome da 3ª pessoa pelo pronome expletivo *it* em inglês. A par da tradução de um sujeito nulo pelo pronome expletivo *it*, este é um desempenho característico deste sistema. O sistema parece não ter especificações quanto à tradução dos pronomes pessoais da 3ª pessoa do singular (*ele/ela*) em português ou parece ter uma especificação muito exacta quanto à tradução dos pronomes pelo expletivo *it*.

Em (66) - *Foram à manifestação muitos trabalhadores.* -, a inversão do sujeito coloca problemas à tradução automática em ambos os sistemas.

(66) (a) *They were to the manifestation many workers.

(b) *The demonstration were many workers.

(66) (a) *They were to the manifestation many workers.

(b) *The demonstration were many workers.

Em (66) (a), o Systran preenche o lugar canónico do sujeito na estrutura da frase, com pronome *they*, sem ter em conta o sujeito da frase. O Google, em (66) (b), preenche o lugar do sujeito com o substantivo que constitui o complemento locativo da frase, *manifestação*. É provável que o sistema não disponha na sua base de dados de exemplos de frases em que o verbo *ir* (*foram*) ocorra numa posição inicial, ou seja, a probabilidade de ocorrência de um verbo no início da frase é nula.

4.1.3.1 Para o controlo do sujeito nulo em frases simples

Como se verificou na análise dos exemplos acima apresentados, o sujeito nulo no português europeu constitui um problema na tradução automática para inglês. De uma forma, geral, a realização do sujeito constituiria uma solução para o problema do sujeito nulo.

Numa frase como (62) - *Compraram o livro ontem.* -, o sujeito poderia ser realizado por uma expressão pronominal (um pronome pessoal como *eles/elas*) ou por uma expressão nominal (como *os estudantes*). Vejam-se os resultados da tradução de frases com o sujeito realizado, em ambos os sistemas:

(68) Eles compraram o livro ontem.

(a)They had bought the book yesterday.

(b) They bought the book yesterday.

(69) Os estudantes compraram o livro ontem.

(a) The students had bought the book yesterday.

(b) The students bought the book yesterday.

Os sistemas exibem desempenhos aceitáveis como resultado das alterações efectuadas. No que concerne à expressão pronominal, em (68), as traduções exibidas por ambos os sistemas são aceitáveis. Tanto o Systran, em (68) (a), como o Google, em (68) (b), reconhecem a expressão pronominal, o pronome pessoal da 3ª pessoa do plural *eles*, como sujeito e traduzem-na de forma aceitável[18].

O Systran - (69) (a) - reconhece a expressão nominal em (69) e tradu-la de forma aceitável, o que pode dever-se ao facto de o sistema traduzir palavra a palavra. O Google - (69) (b) -, por sua vez, também produz uma tradução aceitável.

Ainda assim, tanto em (68) como em (69), a questão do tempo verbal mantém-se na tradução do Systran. Isto é, o sistema continuar a exibir a tradução inaceitável efectuada para o exemplo (62). É escusado, mais uma vez, enunciar aquelas que se consideram as razões para tal desempenho, uma vez que tais razões já foram especificadas para o exemplo (62).

No que diz respeito à frase em (63) - *Há um acidente na estrada principal.* -, deve dizer-se que não é possível apresentar uma solução tão imediata quanto a solução sugerida para a frase (62). A questão do sujeito nulo num enunciado como o enunciado com o verbo *haver* na acepção de 'existir', requereria uma intervenção no próprio sistema e não no *input* da tradução, o que, neste caso, se revela muito importante no Systran. Uma intervenção como a que se sugere aqui teria de consistir numa especificação do uso do verbo *haver* com o sentido de 'existir', bem como da sua estrutura.

Ora, na impossibilidade de qualquer intervenção nos sistemas, na medida em que o presente trabalho prevê somente que se operem alterações no texto que serve de *input* à tradução, sugere-se que a reescrita da frase em (63) por uma paráfrase que transmita o sentido, como se ilustra no exemplo (70):

(70) Um acidente aconteceu na estrada principal.

[18] O mesmo não aconteceria se se tratasse do pronome pessoal na 3ª pessoa do singular, tal como será referido ainda neste tópico.

A frase em (70) expressa o mesmo sentido de *Há um acidente na estrada principal,* no entanto, a ver pelos desempenhos dos sistemas, não coloca quaisquer problemas:

(70) (a) An accident happened in the main road.

(b) An accident happened on the main road.

Esta opção de controlo apenas se aplica a casos como (70). Tendo em conta que, na língua portuguesa, também é possível a realização do sujeito com o verbo *haver*, ainda que em contextos muito restritos, como *Ele há cada uma!*, testou-se a realização do sujeito com este verbo. Numa frase como (71), testou-se a realização do sujeito como se ilustra em (72):

(71) Há quatro copos na mesa.

(72) Quatro copos há na mesa.

Os resultados obtidos nos sistemas são bastante interessantes:

(72) (a) Four cups have in the table.

(b) Four glasses are on the table.

Ambos os sistemas têm desempenhos bastante satisfatórios, tanto que uma possível opção de controlo poderá passar por esta solução.

No que concerne aos sujeitos nulos em (64) e (65), coloca-se a mesma questão que se coloca em (63), ou seja, não há uma solução satisfatória, o que se prende sobretudo com a diferença estrutural entre as línguas.

O caso da inversão do sujeito, como em (66), pode ser resolvida se o sujeito estiver na sua posição canónica, ou seja, numa posição pré-verbal.

Pode concluir-se que esta é uma questão difícil de contornar. A problematização da tradução automática dos diferentes sujeitos nulos apresentados para ilustrar a questão do sujeito nulo não permitiu chegar-se a uma proposta de controlo/simplificação unívoca.

4.1.3.2 O sujeito nulo em estruturas coordenadas e de subordinação

Em estruturas coordenadas do tipo de (73), é obrigatória a omissão do sujeito no segundo termo da estrutura coordenada quando os sujeitos são co-referentes e se trata da 3ª pessoa, a não ser que sobre este recaia um acento de intensidade. Se ambos os sujeitos forem de 3ª pessoa e tiverem uma referência disjunta, a realização do sujeito é obrigatória.

(73) A Maria comprou os sapatos e pagou em dinheiro.

(73.1) A Maria comprou os sapatos e ela pagou em dinheiro.

Tendo em consideração estes dados e o facto de a não realização do sujeito ter implicações na tradução automática do português para o inglês, como já evidenciado, testou-se a viabilidade da realização do sujeito em qualquer um dos casos. Veja-se então a tradução da frase em (73.1), em que o sujeito está realizado no segundo termo da estrutura coordenada:

(73.1) (a) The Maria bought the shoes and it she paid in money.

(b) Mary bought the shoes and she paid cash.

As traduções do exemplo (73.1) mostram que a realização do sujeito dá origem a um *output* adequado, uma vez que, como é sabido, o inglês não é uma língua de sujeito nulo. Uma frase como (73.1) coloca problemas no que respeita à interpretação, na medida em que, como já referido, o sujeito da frase que constitui o segundo membro da coordenação só com acento de intensidade é interpretado como co-referente com o sujeito da outra frase. E a interpretação disjunta não está disponível na frase de partida. Contudo, embora ambígua, em termos de interpretação do antecedente do pronome, a frase do *ouput* tem como interpretação preferencial aquela em que há co-referência, i. e., a interpretação esperada. Assim sendo, os problemas do *input* são irrelevantes, justificando-se a manipulação proposta.

Analisa-se agora a omissão do sujeito em frases subordinadas, uma vez que este tipo de frases partilha características com as estruturas coordenadas acima consideradas. Tomemos as seguintes frases para análise:

(74) A Maria afirmou que desconhecia o convite do João.

(75) O Pedro disse que viajaria pela Europa, se tivesse dinheiro.

Temos aqui dois casos de subordinação que implicam vários casos de sujeito não realizado. Na primeira frase, o sujeito da subordinada, sendo nulo, é interpretado como co-referente com o sujeito da subordinante.

Na segunda frase, o verbo *dizer* tem como complemento uma frase completiva, com sujeito nulo, co-referente com o sujeito da subordinante. A frase inclui ainda uma condicional, igualmente com sujeito nulo e co-referente com o da subordinante.

Veja-se a tradução das frases em questão, (74) (a) e (75) (a) para o Systran e (74) (b) e (75) (b) para o Google:

(74) (a) *The Maria affirmed that the invitation of the João ignored.

(b) Maria said it was unaware that the invitation of John.

(75) (a) *Pedro said that he would travel for the Europe, had money.

(b) The Peter said that he would travel to Europe if he had money.

A tradução da primeira frase pelo Systran apresenta vários problemas, que se devem ao sujeito não realizado. O facto de o sujeito não estar realizado vai fazer com que o sistema «procure» um sujeito, e fá-lo atribuindo o segmento *the invitation of the João* como sujeito à forma verbal *ignored*. Veja-se o resultado da realização do sujeito, como na frase (74), e a respectiva tradução no Systran e no Google:

(76) A Maria afirmou que o Pedro desconhecia o convite do João.

(a) The Maria affirmed that Pedro ignored the invitation of the João.
(b) Mary said that Peter was unaware of the invitation of John.

O resultado do teste mostra que a realização do sujeito não coloca os problemas vistos na tradução (74) (a). No caso da tradução do Google, os resultados são mais

aceitáveis. Mais uma vez se verifica que a realização do sujeito é muito importante para a obtenção de um *output* aceitável na tradução automática.

Na segunda frase, o Systran não atribui sujeito à frase subordinada, omite também a conjunção *if*. Vejamos o que acontece, para o Systran, quando o sujeito está realizado na condicional:

(77) O Pedro disse que viajaria pela Europa, se ele tivesse dinheiro.

(a) Pedro said that he would travel for the Europe, if it had money.

A introdução do pronome *ele* não é a melhor solução, devido à questão do pronome expletivo, que analisaremos mais à frente. No entanto, o sistema deixa de omitir a conjunção *if*. O Systran traduz preferencialmente um sujeito não realizado por um *it* expletivo, isto é, na maior parte das vezes em que o *input* da tradução tem um sujeito que não está realizado, o sistema atribui-lhe um *it* expletivo para preencher o lugar do sujeito não realizado. É importante referir que este problema parece acontecer apenas com a 3ª pessoa do singular. Veja-se o que acontece se alterarmos a pessoa para a 1ª pessoa do plural:

(78) Nós afirmámos que desconhecíamos o convite do João.
(79) Nós dissemos que viajaríamos pela Europa, se tivéssemos dinheiro.

Veja-se o resultado da tradução no Systran:

(78) (a) We affirmed that we ignored the invitation of the João.
(79) (a) We said that we would travel for the Europe, if we had money.

Como se pode ver nestes exemplos, a primeira pessoa do plural não causa os mesmos problemas.

Vejamos ainda outros exemplos de completivas com sujeito nulo. Tome-se para análise o seguinte exemplo:

(80) O primeiro-ministro afirmou que vai rectificar o Orçamento de Estado.

Esta é uma frase subordinada completiva finita. O verbo *afirmar* introduz uma oração completiva encabeçada pelo complementador *que* e com sujeito não realizado, o que faz com que, em contexto apropriado, a frase possa ter uma das seguintes interpretações:

(A) O primeiro-ministro afirmou que [o seu ministro] vai rectificar o Orçamento de estado.

(B) O primeiro-ministro afirmou que [ele próprio] vai rectificar o Orçamento de estado.

No primeiro caso, a interpretação dos sujeitos é disjunta. No segundo caso, a interpretação é de co-referência. Também aqui a interpretação do sujeito da frase completiva nestes casos coloca problemas à tradução automática. Vajam-se as traduções da frase em (80) para ambos os sistemas:

(a) *The Prime Minister affirmed that it goes will rectify the Budget of state.

(b) *The Prime Minister said it will amend the budget of state.

Nestes exemplos, como nos anteriores, o sujeito não realizado nas orações completivas, coloca problemas à tradução para inglês, na medida em que os sistemas preenchem o lugar do sujeito com o pronome expletivo *it*.

Centremo-nos agora na interpretação com referência disjuntiva. Neste caso, o sujeito é normalmente realizado. E, não sendo, na versão controlada deveria sê-lo. Veja-se o exemplo e respectiva tradução nos sistemas:

(81) O primeiro-ministro disse que o ministro das finanças ia rectificar o orçamento de estado.

(a) The Prime Minister said that the minister of the finances went will rectify the Budget of State.

(b) The prime minister said the finance minister would amend the state budget.

Ao longo da análise do tópico do sujeito nulo, tem-se defendido a realização do sujeito. Atente-se, no entanto, no facto de que os sistemas exibiram traduções aceitáveis quando o *input* envolvia frases subordinadas. Nesta medida, a intervenção no sentido de um controlo/simplificação das estruturas pode ser dispensada. Tratando-se de estruturas de coordenação, todavia, a realização do sujeito no segundo termo da estrutura obteve resultados positivos, pelo que é aconselhável.

4.1.4 Alternâncias verbais

Nesta secção, analisam-se casos de alternâncias verbais. Por alternância verbal entendem-se as diferentes construções sintácticas em que um dado verbo e os seus argumentos podem projectar-se.

No âmbito do presente trabalho, analisam-se verbos de alternância causativa. As frases a seguir apresentadas, são exemplos da variante causativa e da variante incoativa, esta última seleccionando o clítico *-se*:

(81) O Pedro partiu o vidro.

(81.1) O vidro partiu-se.

(82) O inimigo afundou o navio.

(82.1) O navio afundou-se.

(83) A Maria fechou a porta.

(83.1) A porta fechou-se.

Em (81), apresenta-se a variante causativa da alternância em questão. De um modo muito geral, sintacticamente, esta é uma frase simples que respeita a ordem canónica dos constituintes em português, a ordem SVO. Semanticamente, a frase tem um AGENTE (*O Pedro*) e um TEMA (*o vidro*).

Em (81.1), apresenta-se a variante incoativa da mesma alternância. Sintacticamente, esta variante apresenta a característica de ter no lugar do sujeito o elemento que na outra variante ocupa a posição de objecto directo. O verbo ocorre na chamada forma pronominal, sendo que o clítico não é interpretável como reflexo.

O mesmo pode ser dito para os outros pares de exemplos. No entanto, vejamos o desempenho dos sistemas na tradução dos exemplos, pois cada exemplo parece influenciar o desempenho dos sistemas de forma interessante. A razão da escolha deste

tipo de verbos é a presença do clítico *–se*, que pode causar problemas de ambiguidade, pois o *se* é uma forma ambígua.

A seguir apresentam-se as traduções dos exemplos nos dois sistemas utilizados, (a) para o Systran, e (b) para o Google:

(81) (a) Pedro broke the glass.

(b) Peter broke the glass.

(81.1) (a) The glass broke.

(b) The glass broke.

A tradução do primeiro caso de alternância, (81) para a variante causativa e (81.1) para a variante incoativa, não apresenta quaisquer problemas. Ambos os sistemas têm um desempenho aceitável.

A variante causativa da alternância em questão em (82) é traduzida de forma aceitável tanto pelo Systran (a) como pelo Google (b). A variante incoativa em (82.1) é traduzida de forma aceitável pelo Google (82.1) (b).

(82) (a) The enemy sank the boat.

(b) The enemy sank the boat.

(82.1) (a) The ship disappeared.

(b) The ship sank.

No entanto, o Systran apresenta uma tradução questionável, na medida em que traduz o verbo reflexo *afundar-se* pelo verbo *to disappear*. Embora, semanticamente, o verbo *to disappear* possa ser aceite como tradução do verbo reflexo *afundar-se*, uma vez que é provável que o navio tenha desaparecido, ou seja, tenha deixado de ser visto, quando se afundou. Contudo, este é um desempenho muito rebuscado para um sistema

de tradução automática, pelo que se torna importante perceber o porquê de o Systran exibir este desempenho.

Traduziram-se, no Systran, os verbos *afundar* e *afundar-se* isoladamente, isto é, sem estarem integrados numa frase. O resultado desta tradução no sistema mostrou que este tem especificações no seu léxico que dão conta de que a tradução preferencial para o verbo *afundar* (uso transitivo) é *to sink*, ao passo que a tradução preferencial para *afundar-se* (uso intransitivo) é *to disappear*. Conclui-se então que o Systran não dispõe, no seu léxico, do verbo *to sink* no seu uso intransitivo.

Atente-se agora, na tradução da alternância em (83) - *A Maria fechou a porta.* - e da variante incoativa da mesma alternância, em (83.1) - *A porta fechou-se.* -:

(83) (a) The Maria closed the door.

(b) Mary closed the door.

(83.1) (a) The door was closed.

(b) The door closed.

Em (83), a tradução da variante causativa por ambos os sistemas não apresenta, mais uma vez, quaisquer problemas. Já a variante incoativa, em (83.1), coloca problemas ao Systran, problemas esses que se devem às características da forma *se*, por esta ser ambígua. Isto é, a forma *se* pode pertencer a diferentes categorias e ter diferentes funções sintácticas, pode ser um pronome reflexo, pode ser um pronome recíproco, pode ser um clítico argumental de referência arbitrária (*se* nominativo), pode ser um clítico com estatuto argumental e funcional (*se* passivo), pode ser um clítico com comportamento de afixo derivacional (clítico ergativo/anticausativo), pode ser um clítico com conteúdo semântico e morfo-sintáctico (clítico inerente) e pode ser uma conjunção condicional. O Systran, em (83.1) (a), interpreta o clítico da variante incoativa como sendo o clítico passivo numa construção passiva.

O Google exibiu traduções aceitáveis para ambas as variantes dos exemplos (83) e (83.1).

Em alguns casos de verbos passíveis deste tipo de alternância, o clítico não tem de estar expresso e, por essa razão, testou-se a omissão do clítico na variante incoativa[19]:

(84) O vidro partiu.

(85) O navio afundou.

(86) A porta fechou.

Veja-se agora o resultado das traduções destes exemplos nos dois sistemas:

(84) (a)The glass left.

(b) The glass broke.

(85) (a) The ship sank.

(b) The ship sank.

(86) (a) The door closed.

(b) The door closed.

Como se verifica, com a variante causativa, os sistemas têm melhores resultados do que com a variante incoativa. Nos casos em que a variante causativa não pedir obrigatoriamente o clítico *–se*, pode usar-se a variante incoativa, como se observou nos exemplos (84), (85) e (86).

Em (84) (a), por uma relação de homonímia entre a forma verbal do verbo *partir* no Pretérito Perfeito (*partiu*), no sentido de 'deixar um lugar para se dirigir a outro', e a forma verbal *partir* no Pretérito Perfeito (*partiu*), no sentido de 'quebrar', o Systran traduz o verbo *partir* pelo verbo *to leave* (*left*). Se com o verbo partir na variante causativa sem clítico se coloca este problema, o mesmo não acontece com a variante causativa com clítico, como pode ver-se na tradução (81.1) (a). Naturalmente que o

[19] A omissão ou realização do clítico na variante incoativa suscita diferentes opiniões quanto à questão da aceitabilidade.

clítico permite desambiguar o verbo *partir*. Neste contexto, poder-se-ia optar pela utilização do verbo *quebrar*, numa tentativa de contornar a polissemia de *partir*.

Analise-se agora o desempenho dos sistemas na tradução da variante causativa, quando a direcção da tradução é do inglês para o português:

(87) The glass broke.

(a) O vidro quebrou.

(b) O vidro quebrou.

Neste exemplo, o desempenho de ambos os sistemas é semelhante, ambos omitem o clítico *–se*. No caso do Systran, em (87) (a), é possível que o facto de o sistema se basear na variedade brasileira do português tenha tido peso na tradução, uma vez que, na variedade brasileira, o recurso ao clítico na variante incoativa, neste tipo de alternância, não é frequente. No caso do Google, em (87) (b), a tradução pode ser explicada da seguinte forma: o sistema tem, na sua base de dados, a tradução para a frase em (87) e recupera essa tradução e na falta de tradução para o clítico *–se*, o sistema deixa essa posição vazia. Veja-se agora a tradução do exemplo (88):

(88) The ship sank.

(a) O navio afundou-se.

(b) O navio afundou.

Na tradução do exemplo (88), o Systran, em (88) (a), tem um desempenho aceitável. É interessante ver que, neste caso, o sistema traduz o clítico *–se*. E, neste seguimento, a explicação sugerida para (87) (a) deixa de fazer sentido, tendo em conta o desempenho do sistema em (88) (a), na medida em que não é provável que, a haver conhecimento linguístico para (87), o mesmo não seja utilizado na tradução de (88). Para o desempenho do Google em (87) (b), aplica-se o mesmo que foi dito para o exemplo (87) (b).

Em (89), coloca-se um problema de ambiguidade lexical que se deve à homonímia entre a forma do adjectivo *closed* e a forma do *Simple Past* do verbo *close* (*closed*). Veja-se o resultado das traduções em ambos os sistemas:

(89) The door closed.

(a) A porta fechado.

(b) A porta fechada.

Ambos os sistemas exibem um desempenho pouco satisfatório no que diz respeito a esta questão. No entanto, o Google, em (89) (b), ao interpretar a forma *closed* como um adjectivo, e ao colocá-la numa relação de modificação com o nome *porta*, resolve de forma satisfatória a concordância entre o adjectivo (atribuindo-lhe traços do feminino, singular) e o nome feminino singular. Pelo contrário, o Systran atribui traços do masculino, singular ao adjectivo.

Na sequência da análise efectuada no que concerne às alternâncias verbais, pode dizer-se que a questão se coloca da seguinte forma: relativamente à variante causativa, os sistemas exibem um desempenho aceitável na tradução de todas as frases que se propuseram para análise. Desta forma, defende-se o uso desta variante na escrita em português controlado.

4.1.5 Modalidade e modo

A expressão da modalidade coloca problemas tanto à tradução humana como à tradução automática, na medida em que «os conceitos modais podem ser expressos nas línguas naturais através de uma grande variedade de formas» (Mateus *et al.*, 2003:245). Assim, a modalidade pode ser veiculada através de verbos modais (*poder*, *dever*), através de outro tipo de verbos (*saber*, *crer*, *permitir*, entre outros), através de advérbio de frase (*provavelmente*, *possivelmente*), através de adjectivos (*provável*), ou através de determinados tempos verbais (Imperfeito, Futuro, Condicional.

Nesta secção, trata-se a tradução do verbo *poder* e a formas de exprimir a modalidade epistémica de uma forma mais produtiva na tradução automática. No que concerne à tradução automática do verbo *poder* para inglês, o problema que se coloca deriva do facto de os sistemas não conseguirem fazer a opção de tradução correcta entre o modal *can* e o modal *may*.

Na língua inglesa, os modais *can* e *may* veiculam diferentes tipos de modalidade. *Can* veicula informação modal de possibilidade material (aptidão), possibilidade moral (permissão), valores directivos (sugestão, pedido, censura), característica ocasional e possibilidade lógica. O verbo modal *may* veicula informação modal de permissão, possibilidade lógica e concessão, sugestão e censura (Larreya, 1998). Em ambos os casos, verificou-se, nos testes efectuados, que o verbo *poder* é a tradução preferencial para os sistemas.

A seguir, apresentam-se exemplos de frases, em que o verbo *poder* veicula diferentes tipos de informação, e as respectivas traduções no Systran (a) e no Google (b):

(90) Ele pode correr cinco quilómetros sem se cansar.
(91) Para falares com a Ana, podes telefonar depois das aulas.
(92) O Rui pode sair já.
(93) O João pode ter chegado há dez minutos

Em (90), o verbo *poder* expressa a ideia de capacidade física. Isto é, 'ele tem a capacidade física de correr cinco quilómetros sem se cansar'. Esta informação é

veiculada, em inglês, pelo verbo modal *can*. Na tradução automática da frase em (90), os sistemas exibem um desempenho aceitável. Vejam-se então os resultados:

(90) (a) It can run five kilometers without getting tired.

(b) He can run five miles without tiring.

Em (91), o verbo *poder* veicula uma ideia de possibilidade, ou seja, parafraseando, 'Para falar com a Ana, tens a possibilidade de telefonar depois das aulas'. Na tradução deste exemplo, o Google opta pelo verbo modal *can* para traduzir o verbo *poder*, já o Systran nem sequer reconhece a forma verbal *podes* e tem um desempenho inadequado, como se pode ver na tradução (91) (a).

(91) (a) *For speech with Ana, you pods to telephone after the lessons.

(b) To speak with Anna, you can call after school.

Em (92), o verbo *poder* exprime a ideia de permissão. Neste caso, o verbo *poder* já tem uma interpretação de verbo modal, ao contrário do que acontecia em (85), em que *poder* era um verbo pleno.

(92) (a) Rui can leave already.

(b) Rui can leave now.

A última frase merece uma atenção especial. Esta veicula a modalidade epistémica, que se relaciona com a expressão da incerteza e a probabilidade, pelo que a frase pode significar que é provável que o João tenha chegado há dez minutos, mas não é certo. De uma forma geral, os sistemas não descodificam correctamente o verbo *poder* na variante da modalidade epistémica. O mesmo não acontece com o verbo *poder* nos outros exemplos (veja-se acima), excepto em (91), em que o Systran nem sequer reconhece a forma verbal. As traduções do Google são bastante aceitáveis no que diz respeito à tradução do verbo *poder* e à tradução da modalidade no exemplo (93):

(93) (a) *The João can have fond has ten minutes.

(b) The John may have arrived ten minutes ago.

Vejamos novamente o exemplo que ilustra a modalidade epistémica - O João pode ter chegado há dez minutos. - e a respectiva tradução no Systran (93) (b). O que é dito na frase é que é provável ou possível que o João tenha chegado há dez minutos, mas não é certo. A tradução do Systran (93) (a) não é aceitável, uma vez que, na maior parte dos casos, o verbo *can* não transmite a mesma ideia de probabilidade, mas sim de aptidão. Atente-se ainda no facto de a forma do Particípio Passado, *chegado*, causar problemas de ambiguidade lexical, pelo que o sistema a traduz pelo adjectivo inglês *fond*, que significa *amado*, *acarinhado*.

Os casos em que *can* exprime o sentido de possibilidade lógica estão limitados por restrições que se relacionam com os tempos verbais. Por exemplo, na frase *He could win the election*, o verbo modal no Pretérito Perfeito expressa a ideia de possibilidade lógica ou probabilidade, 'ele poderia ganhar as eleições'. No entanto, na frase no Presente *He can win the election*, o sentido é de possibilidade material (Larreya *et al.*, 1998), ou seja, 'ele pode/é capaz de ganhar as eleições'.

A probabilidade expressa na frase em (93) é essencialmente traduzida pelo modal *may*. Logo, a tradução do Google (93) (b) é aceitável, ao contrário da tradução do Systran (93) (a).

Existem, no entanto, outras formas de exprimir a modalidade. Como pode ver-se nos exemplos que se seguem, a modalidade pode ser expressa por advérbios, como *provavelmente*; ou por expressões predicativas como *é possível*. Repare-se que a tradução automática do Systran para cada um dos exemplos mostra um *output* melhorado em relação a (93) (a):

(94) Provavelmente o João chegou há dez minutos.

(a) Probably the João arrived has ten minutes.

Ou:

(95) É possível que o João tenha chegado há dez minutos.

(a) It is possible that the João has arrived has ten minutes.

Desta forma, e apesar da agramaticalidade dos *outputs*, que se prende com outras questões que não a que se tomou para análise, refira-se que o processo de controlo/simplificação da questão da modalidade teve como resultado um *output* aceitável.

4.1.6 Uso de determinantes com nomes próprios

Nesta secção, analisa-se o uso de determinantes em construções de sintagmas nominais, em que o núcleo do sintagma é um nome próprio. Atente-se nos seguintes exemplos:

(96) O Pedro comeu o bolo.

(97) O Ivo comeu o bolo.

(98) O João comeu o bolo.

(99) O Rui comeu o bolo.

Nestas frases, temos dois sintagmas nominais com determinantes. Contudo, o sintagma nominal sujeito tem como núcleo um nome próprio (*Pedro, Ivo, João, Rui*) e o sintagma nominal objecto directo tem como núcleo um nome comum (*bolo*, para todos os exemplos).

O principal problema que se coloca é o facto de o inglês não fazer o mesmo uso dos determinantes, quando o núcleo do sintagma nominal é um nome próprio. A partir das traduções efectuadas, verificou-se que por exemplo o Systran não respeita essa regra, na maior parte dos casos, e que, por exemplo, o Google, apesar de obter resultados mais aceitáveis, ainda exibe um comportamento menos aceitável em alguns casos. Veja-se a tradução no Systran (a) e no Google (b) das frases de (96) a (99):

(96) (a) Pedro ate the cake.

(b) *The Peter ate the cake.

(97) (a) *The Ivo ate the cake.

(b) *The Ivo ate the cake.

(98) (a) *The João ate the cake.

(b) *The John ate the cake.

(99) (a) Rui ate the cake.

(b) *The Rui ate the cake.

O sistema traduz o determinante que antecede os nomes próprios, o que não é aceitável em inglês. As frases destes exemplos convergem no facto de os nomes próprios utilizados serem, até certo ponto, distantes dos nomes próprios correspondentes em inglês. Com distante queremos referir o facto de graficamente serem bastante diferentes, mesmo sendo reconhecidos e até traduzidos pelo sistema. Atente-se no facto de o sistema introduzir o determinante antes dos nomes próprios, mesmo nos casos em que traduz o nome próprio. Atente-se na tradução das restantes frases, em que apenas mudámos o nome próprio para um nome em português graficamente muito próximo ou até semelhante ao nome próprio inglês (exploração preliminar de uma ideia de P. Marrafa, c. p.):

(94) A Maria colheu as flores

(95) O David colheu as flores.

(96) A Sara colheu as flores.

(97) A Ana colheu as flores.

E as traduções dos sistemas correspondentes:

(a) *The Maria harvested the flowers

(a) *The David harvested the flowers.

(a) *The Sara harvested the flowers.

(a) *Ana harvested the flowers.

(b) Mary gathered the flowers

(b) David picked the flowers.

(b) Sara picked flowers.

(b) Ana picked flowers.

Nas traduções do Systran, persiste o problema da colocação do determinante antes de um nome próprio. Em quatro frases para análise o sistema apenas omitiu o determinante em uma. Tal pode dever-se ao facto de o sistema traduzir de forma muito literal. Repare-se que desta vez, nas traduções do Google, o sistema não colocou o determinante, eliminando assim a agramaticalidade. Tal parece dever-se ao facto de a grafia dos nomes próprios usados nestes exemplos ser igual ou muito próxima da grafia dos nomes próprios correspondentes em inglês.

Os resultados da tradução mostram que a omissão do determinante em português, resulta na omissão do determinante correspondente em inglês, logo, o *output* da tradução é aceitável. Nos testes efectuados, os sistemas exibiram um comportamento semelhante na tradução na maior parte dos sintagmas nominais cujo núcleo é um nome próprio, ou seja, ambos os sistemas realizaram o determinante na língua inglesa, que, no entanto, não permite determinantes a preceder nomes próprios.

Testaram-se os sistemas com nomes próprios com grafia muito próxima e até igual nas duas línguas. O Systran continuou a exibir um desempenho pouco aceitável ao traduzir os determinantes. O Google, por sua vez, obteve, neste teste, resultados satisfatórios. Num último teste aos sistemas, omitiu-se o determinante. Os resultados das traduções foram bastante aceitáveis, tanto para o Systran como para o Google. Desta forma sugere-se a omissão dos determinantes em sintagmas nominais cujo núcleo é um nome próprio.

4.2 Regras em Português Controlado

Nesta secção visa-se a elaboração de uma lista de regras que possam nortear a escrita em português controlado. Estas regras constituem um fragmento de um futuro português controlado e, embora não abranjam um grande número de fenómenos linguísticos problemáticos para a tradução automática, centram-se, ainda assim, em questões bastante importantes.

Esta lista de regras está dividida em secções (numeradas de 1 a 6) que correspondem aos fenómenos linguísticos abordados. Por sua vez, essas secções estarão subdivididas para indicar diferentes regras dentro do mesmo fenómeno linguístico.

Lista de Regras

Secção 1: Expressões verbais complexas que integram um verbo de suporte

1.1 No caso de a expressão verbal complexa e a forma atómica possuírem equivalência directa entre o português e o inglês, ambas podem ser utilizadas.

1.2 No caso de os verbos seleccionarem um clítico *–se*, podendo comutar com uma forma não atómica, deve preferir-se a forma não atómica.

1.3 Nos casos de predicados não atómicos – V+SN, por exemplo – equivalentes a predicados atómicos – V – deve preferir-se a forma não atómica.

Secção 2: Expressões verbais complexas com verbo auxiliar aspectual

2.1 Nos casos das expressões verbais complexas com auxiliares aspectuais, deve preferir-se o uso de uma forma verbal atómica que veicule o sentido da expressão verbal, recorrendo ao verbo principal da expressão. Para completar o sentido, pode fazer-se uso de advérbios, expressões adverbiais e até paráfrases. No caso específico da

expressão verbal *estar a + verbo*, é aconselhada a sua substituição pela expressão *gerúndio + infinitivo*.

Secção 3: Sujeito nulo

3.1 O sujeito deve estar realizado sempre que possível. Nos casos de coordenação, o sujeito deve estar expresso no segundo termo da estrutura. A obrigatoriedade da realização do sujeito numa estrutura de subordinação não é tão relevante.

Secção 4: Alternâncias verbais

4.1 Nos casos de alternâncias causativas/incoativas, deve preferir-se a variante causativa.

Secção 5: Modalidade e Modo

5.1 Na expressão da modalidade, deve-se recorrer a expressões adverbiais que veiculem informação modal, preterindo o uso do verbo *poder*.

Secção 6: Uso de determinantes com nomes próprios

6.1 Nos casos em que um sintagma nominal integre um nome próprio precedido de um determinante deve omitir-se o determinante.

5. Conclusão

O presente trabalho teve como objectivo último contribuir para a criação de uma linguagem controlada para a língua portuguesa: o Português Controlado. Este objectivo justifica-se pela falta de uma proposta conhecida neste âmbito. O presente trabalho não pretendeu ser um estudo exaustivo no sentido de se construir uma linguagem controlada. Pelo contrário, tal como se deixa claro no título, quis-se apenas contribuir para a criação de um fragmento de uma linguagem controlada para o português. Crê-se, contudo, que esta dissertação poderá servir de base para um projecto futuro na área das linguagens controladas. O trabalho desenvolvido dá um contributo para a investigação em língua portuguesa no âmbito da Engenharia da Linguagem.

No segundo capítulo do presente trabalho (secção 2.2.3), procedeu-se à sistematização de alguns problemas que a linguagem natural coloca ao seu processamento computacional. Essa sistematização incidiu nas seguintes questões: ambiguidade (lexical e estrutural), usos metafóricos da linguagem, anáfora, subordinação e combinatórias lexicais. A abordagem levada a cabo neste capítulo serviu para se introduzir a questão que viria a ser o cerne do quarto capítulo. No terceiro capítulo, aborda-se a temática das linguagens controladas. No quarto capítulo, desenvolve-se a problemática das dificuldades que a linguagem natural coloca ao processamento automático da linguagem natural, com o objectivo último de tornar o *output* em tradução automática adequado, através da escrita em linguagem controlada.

Ora, no âmbito da análise efectuada, registaram-se algumas características que os sistemas exibiram nas traduções efectuadas. O Systran revelou bastantes problemas na resolução de ambiguidades e na tradução de enunciados complexos. Este sistema produziu, na maior parte dos casos, traduções literais que tiveram implicações na aceitabilidade dos *outputs*, detectou-se, ainda, uma falta de informação gramatical e de informação nas entradas lexicais. Verificou-se ainda que o sistema utiliza a variedade brasileira do português.

Por outro lado, o Google tem um desempenho que pode considerar-se característico de um sistema baseado em estatística. Por exemplo, o sistema traduziu de forma relativamente aceitável uma frase declarativa com sujeito nulo do português para o inglês, sem atribuir sujeito à frase inglesa, pois este não estava presente no *input* e o

sistema, na falta de informação gramatical quanto à obrigatoriedade de realização do sujeito, deixou o lugar do sujeito vazio. De uma forma geral, na maior parte dos casos, o Google foi o sistema que teve resultados mais satisfatórios.

Tendo em conta estas características exibidas pelos sistemas, a linguagem controlada que se propõe no âmbito do presente trabalho, centrou-se no facto de os sistemas terem melhores desempenhos quanto mais simples for o *input*. Isto é, os sistemas têm mais dificuldades em traduzir enunciados ambíguos e enunciados complexos. Tome-se como exemplo a seguinte frase: *The cat ate the fish*. Estruturas como esta são facilmente interpretáveis e processáveis em sistemas de tradução automática. Este desempenho mostrou-se especialmente característico do sistema Systran.

Por razões óbvias, tal como já foi dito, este trabalho não constitui um estudo exaustivo, pelo que muitos tópicos ficaram por abordar. Tal como é referido no capítulo 4, na secção 4.1.2, a questão do controlo dos graus dos adjectivos é um desses temas. Na secção referida, traça-se uma breve linha de orientação que poderia guiar o desenvolvimento de uma sistematização mais ampla nesta área. Este é apenas um exemplo. No entanto, muitos outros poderiam ser tidos em consideração.

No seguimento do presente estudo sugere-se o desenvolvimento do presente fragmento de linguagem controlada no sentido de trazer mais línguas de chegada para o processo de tradução. Propõe-se ainda, para uma investigação futura, a possibilidade de se colocar a tradução automática e as linguagens controladas ao serviço de um novo tipo de texto: o texto jornalístico. O texto jornalístico é um texto tipicamente não técnico mas que, ainda assim, defende-se nesta conclusão, pode ser controlado. Tal como já foi dito, as linguagens controladas têm-se cingido ao domínio da linguagem técnica, quer seja na indústria, na forma de documentação técnica; na Aeronáutica; na Medicina, no Direito, etc. Desconhece-se uma linguagem controlada para outro tipo de texto que não o texto técnico ou científico, muito menos uma linguagem controlada para o texto jornalístico. Desta forma, poder-se-ia colocar as linguagens controladas ao serviço de outro tipo de público.

Referências

Almqvist, I. & Sågvall, A. (1996). *Defining ScaniaSwedish – a Controlled Language for Truck Maintenance.* http://stp.ling.uu.se/~corpora/scania/ash961.html (consultado em Janeiro de 2009)

Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., Sadler, L. (1994) *Machine Translation: an Introductory Guide*, NCC Blackwell, London.

Banjar, S. Y. (2001) *Controlled Language and Machine Translation* Bulletin of the Faculty of Arts, Vol. 17.

Cabré, M. Teresa, (1999) *Traducción y termonología: un espacio de encuentro ineludible* In *LA TERMINOLOGÍA Y COMUNICACIÓN Elementos para una teoría de base comunicativa y otros artículos*, Universitat Pompeu Fabra, Barcelona.

Castilla, A. *et al.* (2005) *Machine translation on the medical domain: the role of BLEU/NIST and METEOR in a controlled vocabulary setting*, MT Summit X, Phuket, Thailand, September 13-15, 2005, Conference Proceedings: the tenth Machine Translation Summit; pp.47-54. www.mt-archive.info/MTS-2005-Castilla.pdf (consultado em Janeiro de 2009)

Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press.

Coutinho Silva, C. (2008) *Complementos e modificadores preposicionais do nome: o caso das preposições de e a*. Dissertação de Mestrado. Universidade do Porto

Cunha, C. & L. F. Lindley Cintra (1984) *Nova Gramática do Português Contemporâneo,* Lisboa: Ed. Sá da Costa.

Devisevic, D. & Steensland, H. (2005) *Controlled language in software user documentation*. Tese de Mestrado. Universidade de Linköpings www.diva-portal.org/diva/getDocument?urn_nbn_se_liu_diva-4637-1__fulltext.pdf (consultado em Outubro de 2008)

Cruse, D. A. (1995) *Lexical Semantics*, Cambridge Textbooks in Linguistics Cambridge University Press, Cambridge.

Freigang, K. H. (2001) *A Tradução Automática: Passado, Presente e Futuro*, Universidade de Saarlandes, Saarbruecken. (http://cvc.instituto-camoes.pt/tradumatica/rev0/freigangPT.html (consultado em Julho, 2010)

Fuchs, N. e Schwitter, R. (1995) *Attempto Controlled Natural Language for Requirements Specifications*, 7th ILPS 95 Workshop on Logic Programming Environments, Portland, Oregon.

Gonçalves, Maria Fernandes (1994) *Para uma redefinição do parâmetro do sujeito nulo.* Tese de Mestrado. Universidade de Lisboa.

Gonçalves, V. e Carrapatoso, E. (2009) *Web Semântica e Cérebro Global juntos por uma boa causa,* EDUSER: Revista de Educação, Vol. 1 (1). https://www.eduser.ipb.pt/index.php/eduser/article/viewFile/14/5 (consultado em Setembro 2010)

Guimarães, M. J. (2008) *Interculturalidade na Produção Textual* XI Seminário de Tradução Científica e Técnica em Língua Portuguesa.

Hutchins, W. J. (1986) *Machine Translation: Past, Present, Future.* Elis Horwood, Chichester.

Hutchins e Somers (1992) *An Introduction to Machine Translation* Academic Press, Londres.

Hutchins, W. J. (1999) *The Development and Use of Machine Translation Systems and Computer-based Tools*, International Symposium on Machine Translation and Computer Language Information Processing, 26-28 June 1999, Beijing, China

Hutchins, J. (2003) *Machine translation: general overview*, in Mitkov, R (*ed.) The Oxford Handbook of Computational Linguistics* Oxford University Press, Oxford.

Hutchins, J.(2007) *Milestones in the history of machine translation.* www.hutchinsweb.me.uk/SUSU-2007-1-ppt.pdf

Kaji, H. (1999) *Controlled Languages for Machine Translation: State of the Art*, MT Summit VII.

Kittredge, R. J. (2003) *Sublanguages and controlled languages* in Mitkov, R (*ed.) The Oxford Handbook of Computational Linguistics* Oxford University Press, Oxford

Larreya, P. *et al.* (1998) *Gramática da Língua Inglesa*, Trad. Cristina Oliveira, Bertrand Editora, Venda Nova.

Mador-haim, S. *et al.*, (2006) *Controlled language for Geographical Information Systems*, Proceedings of the Fifth International Workshop on Inference in Computational Semantics Queries. www.cs.technion.ac.il/~winter/papers/nli-gis.pdf (consultado em Agosto 2009)

Marrafa, P. (1993) *Predicação secundária e predicados complexos em português: Análise e modelização.* Dissertação de Doutoramento. Universidade de Lisboa.

Marrafa, P. (2004) *Computação de ambiguidades sintácticas. Evidências em favor dos modelos baseados em conhecimento linguístico.* in Cognito 2(1), 1-10

Marrafa, P. (2006) *A Tradução na Sociedade de Informação.* http://dtil.unilat.org/tercer_seminario/actas/marrafa_pt.htm

Melero i Nogués M. (2006) *La recerca lingüística en la TA*, Revista Tradumàtica – Traducció i Tecnologies de la informació i la Comunicació 04 : Traducció Automàtica. http://www.fti.uab.cat/tradumatica/revista (consultadp em Maio 2010)

Mira Mateus *et al.*, (2003) *Gramática da Língua Portuguesa*, 5ª edição, Lisboa, Caminho.

Mitamura, T. e Nyberg, E. (1995) *Controlled English for Knowledge-based MT: Experience with Kant System*, in Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95). http://www.mt-archive.info/TMI-1995-Mitamura.pdf (consultado em Junho 2009)

Mitkov, R *ed.* (2003) *The Oxford Handbook of Computational Linguistics* Oxford University Press, Oxford.

Nasr, A. (1998) *A Linguistic Framework for Controlled Language Systems*, in Proceedings of the Second International Workshop on Controlled Language Applications. http://pageperso.lif.univ-mrs.fr/~alexis.nasr/Rech/Publi/claw98.pdf (consultado em Maio 2009)

Nyberg, E., *et al.* (2003). *Controlled language for authoring and translation*, in *Computers and Translation: A translator's guide.* John Benjamins Publishing Company, Amsterdam, The Netherlands.

Pace G. *et al.* (2009), *A controlled language for the specification of contracts.* http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-448/paper20.pdf

Pool, J. (2006) *Can controlled languages scale to the web?*, CLAW 2006 at AMTA 2006: 5th International Workshop on Controlled Language Applications.

Pulmann, S. (1996) *Controlled Language for Knowledge Representation,* in: Proceedings of the First International Workshop on Controlled Language Applications, Leuven, Belgium.

Pustejovsky, J. (1995) *The Generative Lexicon*, Massachussetts.

Shiffman *et al.,*(2009) *Controlled Natural Language for Clinical Practice Guidelines.* CNL 2009 attempto.ifi.uzh.ch/site/pubs/papers/cnl2009_shiffman.pdf (consultado em Junho 2009)

Somers, H. (2003) *Machine translation: latest developments,* in Mitkov, R *ed. The Oxford Handbook of Computational Linguistics* Oxford University Press, Oxford.

Spaggiari L.,*et al*. (2005) *A controlled language at Airbus*, in Machine Translation, Controlled Languages and Specialised Languages, edited by Sylviane Cardey, Peter Greenfield, Séverine Vienney. Lingvisticæ Investigationes, tome XXVIII, John Benjamins, 107-122

Specia, L., Machado Rino, L. H. (2002) *Introdução aos Métodos e Paradigmas de tradução Automática*. Universidade Estadual Paulista.

www.dc.ufscar.br/~lucia/TechRep/NILCTR0204-SpeciaRino.pdf (consultado em junho 2009)

Schwitter, R. (1998) *Kontrolliertes Englisch für Anforderungsspezifikationen*, Tese de Doutoramento. Universidade de Zurique.

van der Eijk, P. (1998). *Controlled languages in technical documentation*. Elsnews: The Newsletter of the European Network in Language and Speech, 7(1). Information Technology Research Institute, University of Brighton.

Vassiliou, M. *et al.* (2003), *Evaluating Specifications for Controlled Greek.* http://www.mt-archive.info/CLT-2003-Vassiliou.pdf (consultado em Junho 2010)

Vieira, R. & Strube de Lima, V. L. (2001) *Lingüística computacional: princípios e aplicações*. In: Martins, A. (org) "JAIA – Jornadas de Atualização em Inteligência Artificial". Fortaleza – CE.

www.controlledenglish.com (consultado em Junho de 2010)

Wyner A. *et al.* (2009) *On Controlled Natural Language: Properties and Prospects.* Workshop on Controlled Natural Languages, Merettino. http://wyner.info/research/Papers/CNLP&P.pdf (consultado em Junho 2010)